

AUTOMATIC SEGMENTATION OF BRAIN STRUCTURES IN MAGNETIC RESONANCE IMAGES USING DEEP LEARNING TECHNIQUES

Kaisar Kushibar

Per citar o enllaçar aquest document:
Para citar o enlazar este documento:
Use this url to cite or link to this publication:
<http://hdl.handle.net/10803/670766>

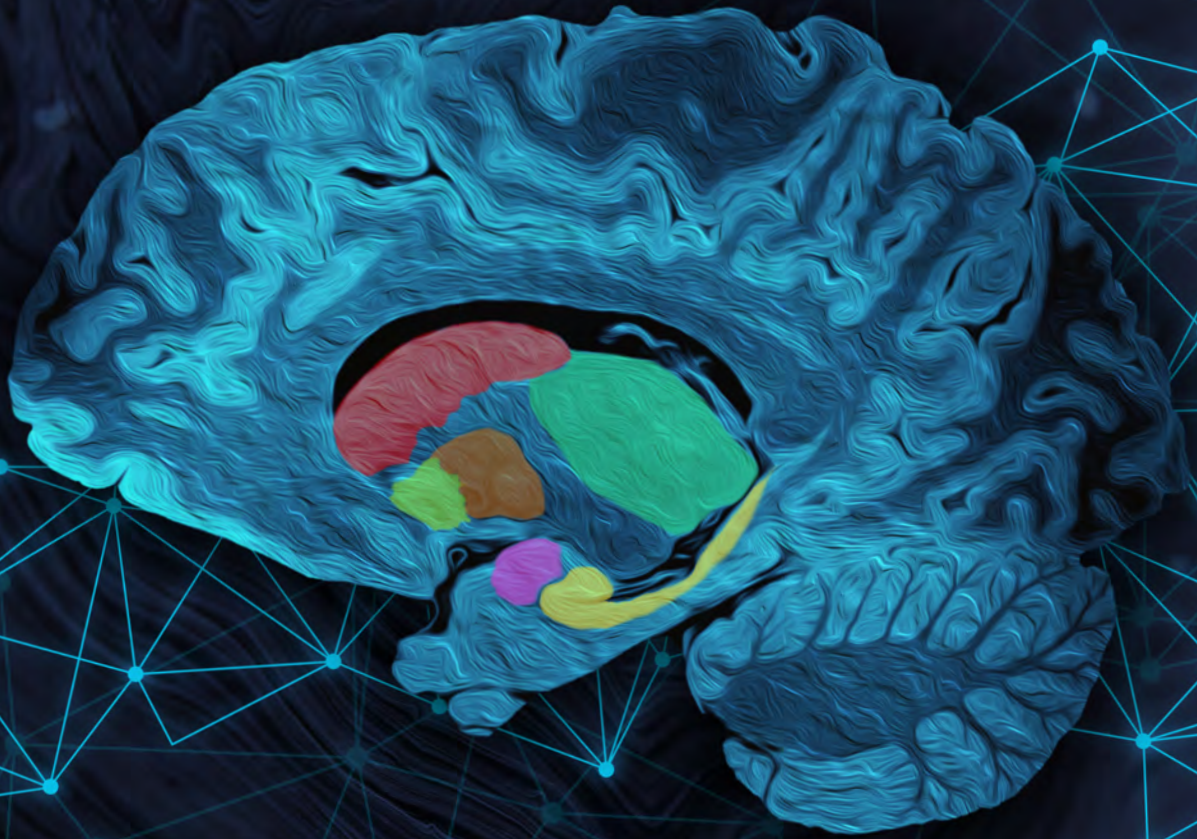


<http://creativecommons.org/licenses/by-nc-sa/4.0/deed.ca>

Aquesta obra està subjecta a una llicència Creative Commons Reconeixement-
NoComercial-CompartirIgual

Esta obra está bajo una licencia Creative Commons Reconocimiento-NoComercial-
CompartirIgual

This work is licensed under a Creative Commons Attribution-NonCommercial-
ShareAlike licence



PhD THESIS

AUTOMATIC SEGMENTATION OF BRAIN STRUCTURES IN MRI USING DEEP LEARNING TECHNIQUES

Kaisar Kushibar
2020



DOCTORAL THESIS

**Automatic segmentation of brain
structures in magnetic resonance images
using deep learning techniques**

Kaisar Kushibar

2020

DOCTORAL PROGRAM in TECHNOLOGY

Supervised by:

Dr. Sergi Valverde

Dr. Arnau Oliver

Dr. Xavier Lladó

Presented to obtain the degree of Doctor of Philosophy at the
University of Girona

*The Brain – is wider than the Sky –
For – put them side by side –
The one the other will contain
With ease – and you – beside –*

*The Brain is deeper than the sea –
For – hold them – Blue to Blue –
The one the other will absorb –
As sponges – Buckets – do –*

*The Brain is just the weight of God –
For – Heft them – Pound for Pound –
And they will differ – if they do –
As Syllable from Sound –*

Emily Dickinson, 1862

Acknowledgements

First and foremost, I thank God, the merciful and passionate, for providing me with strength, patience and opportunity to complete this PhD study. It would be have been impossible without Him to sustain mental health to move forward and cope with the inevitable stress that comes in research.

There cannot be a “one-man-army” in doing research and this PhD thesis is no different. At the end of my journey through the ups and downs of research, I would like to acknowledge individuals who made this work possible.

First of all, I would like to thank my supervisors, Xavier Lladó, Arnau Oliver and Sergi Valverde for their guidance and support throughout my research studies. Their knowledge and advice has influenced me a lot in the way I perceive the philosophy of research and its meaning in practice. It would not have been possible for me to reach this point without their continuous help and motivation. Their outstanding experience and professionalism will always be an example which I will do my best to achieve. I also would like to thank them for giving me the opportunity to be a teaching assistant for the MAIA master program, which has been a tremendous experience.

I would like to give special thanks to my colleagues at ViCOROB: Mostafa Salem, Sandra González-Vilà, Mariano Cabezas, Basel Alyafi, Albert Clerigues, Jose Bernal and Richa Agarwal. They were always open for discussions and I would like to thank them all for the inspiration they gave me, and also for their help and support.

I am grateful for the Catalan government for providing me with a scholarship to do a research in the University of Girona. It is incredibly delightful not only to know but also experience the government’s understanding on the importance of research and welcoming the foreigners alongside with their own citizens. Indeed, knowledge does not have any borders. Moreover, I was fortunate to be one of the VIBOT students in the Erasmus Mundus Master program, which helped me to realise my passion to research and was a great platform to dive into pursuing PhD. Therefore, I do not want to miss the chance to thank the selection committee and EU for giving me an opportunity to be a scholarship student in this unique program.

Academia is not only research and there are people behind, whose hard work keeps the engine of research running smoothly. I would like to thank Anna Ferrarons, Aina Roldán, Mireia Frigola and others from administrative staff for their help and patience. Moreover, thank you for your effort in organising team building and cultural events which were enjoyable experience. As an international researcher, I would like to give special thanks to Eugenia Paradedda from the Human Resources department, who was incredibly helpful with all my visa processes.

No journey is pleasant without friends alongside. I would like to thank my friends in Kazakhstan: Aziz, Aidos, Abdurakhman, Elzhan, Kuandyk, Madiyar, Mereke and Omirzhan, for all the fun and hospitality they showed me during my holiday visits. Moreover, I thank my friends in Girona: Layl, Nika, Dani and Carla for the incredible times we shared together. It is hard to say goodbye, but I hope we will meet again!

Last but not least, all the things I have done not only would be impossible but also would not matter at all without the unconditional support and love of my family. There is nothing in this world that can repay the sacrifices of my parents. Nevertheless, I thank you and hope I can keep the smiles on your faces for a long time. Moreover, I thank my brother and sister who showed extreme support and comfort throughout my entire life. Their encouragement was especially crucial when staying abroad. I wish to you and your families prosperity, health and happiness.

Publications

The presented thesis is a compendium of the following research articles:

- **K. Kushibar**, S. Valverde, S. González-Vilà, J. Bernal, M. Cabezas, A. Oliver, and X. Lladó. “Automated sub-cortical brain structure segmentation combining spatial and deep convolutional features”. *Medical Image Analysis*, 48, pp-177-186, 2018. Quality index: [JCR CSAI IF: 8.880, Q1(5/133)]
- **K. Kushibar**, S. Valverde, S. González-Vilà, J. Bernal, M. Cabezas, A. Oliver, and X. Lladó. “Supervised domain adaptation for automatic sub-cortical brain structure segmentation with minimal user interaction”. *Nature. Scientific Reports*, 9:6742, 2019. Quality index: [JCR MS IF: 4.011, Q1(15/69)]
- **K. Kushibar**, M. Salem, S. Valverde, À. Rovira, A. Oliver, and X. Lladó. “Unsupervised domain adaptation in deep learning for brain magnetic resonance image segmentation”. Submitted to *Knowledge-Based Systems*. 2020. Quality index: [JCR CSAI IF: 5.101, Q1(17/133)]

The rest of publications and conferences related with this PhD thesis are the following:

Journals

- J. Bernal, **K. Kushibar**, D. Sileshi A. S. Valverde, A. Oliver, R. Marti, X. Lladó. “Deep convolutional neural networks for brain image analysis on magnetic resonance imaging: a review”. *Artificial Intelligence in Medicine*, 95, pp 64-81, 2019. Quality index: [JCR MI IF: 3.574, Q1(5/26)]
- J. Bernal, **K. Kushibar**, M. Cabezas, S. Valverde, A. Oliver and X. Lladó. “Quantitative analysis of patch-based fully convolutional neural networks for tissue segmentation on brain magnetic resonance imaging”. *IEEE Access*, 7, pp.89986-90002, 2019. Quality index: [JCR CSIS IF: 4.098, Q1(23/155)]

- M. Rakic, M. Cabezas, **K. Kushibar**, A. Oliver, X. Lladó. “Improving the detection of autism spectrum disorder by combining structural and functional MRI information”. *NeuroImage: Clinical*, 25, p.102181, 2020. Quality index: [JCR N IF: 3.943, Q1(3/14)]

Conferences

- S. Valverde, M.Cabezas, J. Bernal, **K. Kushibar**, S. González-Villà, M. Salem, J. Salvi, A. Oliver, Xavier Lladó. “White matter hyperintensities segmentation using a cascade of three convolutional neural networks”. MICCAI Grand Challenge on White Matter Hyperintensities Segmentation. MICCAI 2017. Quebec, Canada. 14 September 2017.
- J. Bernal, **K. Kushibar**, S. Valverde, M.Cabezas, S. González-Villà, M. Salem, J. Salvi, A. Oliver, Xavier Lladó. “Six-month infant brain tissue segmentation using three dimensional fully convolutional neural networks and pseudo-labelling”. MICCAI Grand Challenge on 6-month infant brain MRI segmentation iSeg-2017. MICCAI 2017. Quebec, Canada. 14 September 2017.
- M. Cabezas, S. Valverde, S. González-Villà, A. Clèrigues, **K. Kushibar**, M. Salem, J. Bernal, A. Oliver, J. Salvi and X. Lladó. “Survival prediction using ensemble tumor segmentation and transfer learning”. Multimodal Brain Tumor Segmentation Challenge 2018 (BRATS) in Medical Imaging. MICCAI Workshop, 2018.
- J. Bernal, M. Salem, **K. Kushibar**, A. Clèrigues, S. Valverde, M. Cabezas, S. González-Villà, J. Salvi, A. Oliver, and X. Lladó. “MR Brain segmentation using an ensemble of multi-path u-shaped convolutional neural networks and tissue segmetnation priors”. MR Brain tissue segmentation Challenge in Medical Imaging. MICCAI Workshop, 2018.
- A. Clèrigues, S. Valverde, J. Bernal, **K. Kushibar**, M. Cabezas, A. Oliver, and X. Lladó. “Ensemble of convolution neural networks for acute stroke anatomy differentiation”. Ischiemic Stroke Lesion Segmentation (ISLES) in Medical Imaging. MICCAI Workshop, 2018.
- J. Bernal, **K. Kushibar**, S. Valverde, M.Cabezas, M. Salem, J. Salvi, A. Oliver, Xavier Lladó. “Six-month infant brain magnetic resonance image tissue segmentation using multi-atlas segmentation with joint label fusion and convolutional neural networks”. MICCAI Grand Challenge on 6-month infant brain MRI segmentation iSeg-2019. MICCAI 2019. China, 2019.

- M. Cabezas, S. Valverde, A. Clèrigues, M. Salem, **K. Kushibar**, J. Bernal, A. Oliver, J. Salvi and X. Lladó. “Brain tumour segmentation and prediction via CNNs”. Multimodal Brain Tumor Segmentation Challenge 2019 (BRATS) in Medical Imaging. MICCAI Workshop, China, 2019.

Book chapter

- J. Bernal, **K. Kushibar**, A. Clèrigues, A. Oliver, X. Lladó. “Deep Learning in Biology and Medicine”. First three authors contributed equally to the chapter titled “Deep learning in Medical Imaging”. *To be published*. 2020.

Acronyms

ANN Artificial Neural Network
ADHD Attention Deficit Hyper-activity Disorder
BET Brain Extraction Tool
CNN Convolutional Neural Network
CSF Cerebrospinal Fluid
EDSS Expanded Disability Status Scale
FLAIR Fluid Attenuated Inversion Recovery
FSL FMRIB Software Library
GM Gray Matter
IBSR Internet Brain Segmentation Repository
LST Lesion Segmentation Toolbox
MNI Montreal National Institute
MRI Magnetic Resonance Image
MS Multiple Sclerosis
T1-w T1-weighted
T2-w T2-weighted
TE Time to Echo
TR Repetition Time
UDA Unsupervised Domain Adaptation
WM White Matter
WMH White Matter Hyperintensities

Contents

Abstract	xv
Resum	xvii
Resumen	xix
1 Introduction	1
1.1 Research context	1
1.1.1 Sub-cortical brain structures	1
1.1.2 Magnetic Resonance Imaging	3
1.1.3 Segmentation of the sub-cortical structures	5
1.1.4 Deep learning	7
1.2 Research background	12
1.3 Objectives	13
1.4 Document structure	14
2 Automated sub-cortical brain structure segmentation combining spatial and deep convolutional features	17
3 Supervised domain adaptation for automatic sub-cortical brain structure segmentation with minimal user interaction	29
4 Unsupervised domain adaptation in deep learning for brain magnetic resonance image segmentation	47
5 Results and discussions	59

5.1	Network architecture	59
5.1.1	Implicit convolutional features	59
5.1.2	Explicit spatial features	60
5.1.3	Sample selection	61
5.2	Knowledge transfer for domain adaptation	62
5.2.1	Number of training images and trainable parameters	62
5.2.2	Domain adaptation and image standardisation	63
5.3	Unsupervised domain adaptation	64
5.3.1	Applications of domain adaptation	64
5.3.2	Effect of histogram loss	65
6	Conclusions and future work	67
6.1	Summary and contributions	67
6.2	Future work	70

List of Figures

1.1	Sub-cortical brain structures.	2
1.2	MRI volume and orthogonal views	4
1.3	MRI sequences.	4
1.4	A generic CNN example	8
1.5	Number of publications in peer-reviewed journals.	11
1.6	Graphical structure of the thesis.	15

List of Tables

1.1	Clinical applications.	3
-----	--------------------------------	---

Abstract

The sub-cortical brain structures are located beneath the cerebral cortex and include the thalamus, caudate, putamen, pallidum, hippocampus, amygdala, and accumbens structures. These bilateral structures – symmetrically located within the left and right hemispheres – are involved in systematic activities such as emotion, pleasure, memory and hormone production. Their deviations in volume are associated with different neurological diseases such as Alzheimer’s disorder, bipolar disorder or multiple sclerosis, among others. Manual segmentation of these structures is a time-consuming task and suffers from rater inter- and intra-variability. Therefore, developing automated methods for accurately segmenting the sub-cortical brain structures is important and it is an active research area.

This PhD thesis focuses on the development of deep learning based methods for accurate segmentation of the sub-cortical brain structures from Magnetic Resonance Images (MRI). This goal has been carried out in several stages. In the first stage, we have proposed a 2.5D – i.e. three 2D orthogonal planes of a 3D volume – Convolutional Neural Network (CNN) architecture that combines convolutional and spatial features. Additionally, we proposed a selective sample selection technique from structure boundaries. The experimental results demonstrated the effectiveness of the proposed approach in accurately segmenting all the sub-cortical brain structures and has shown state-of-the-art performance on well-known publicly available datasets – Multi-Atlas Labelling Grand Challenge MICCAI 2012 and Internet Brain Segmentation Repository (IBSR).

In general, CNN’s performance drops when tested with images from a different domain than the training set. This problem is referred to as the domain shift effect, where a network trained with MRI volumes with the same properties does not generalise to other images with different properties such as scanner type, imaging protocol and resolution. In the next stage of this PhD, we addressed this problem of domain adaptation using a supervised transfer learning strategy. We showed the effect of domain shift on the performance of deep learning methods and how our proposal not only resolves this issue but also drastically reduces the number of training images and trainable parameters of the network.

In the following stage, we developed a new approach that did not even require any manually annotated images for domain adaptation – an unsupervised deep learning method. In this approach, we proposed to employ a histogram loss to minimise the effect caused by domain differences directly within the convolutional layers of the CNN. This approach showed significant ($p < 0.001$) improvements from baseline segmentation results, where no domain adaptation was applied, and showed comparable results to unsupervised FIRST method. In order to show its robustness, we extended our unsupervised domain adaptation method for segmenting brain white matter hyperintensities. The experimental results showed that adapting the network in an unsupervised manner improved the baseline and outperformed traditional unsupervised methods used for this task.

All these completed stages paved the way for achieving an accurate and robust automated deep learning based method for segmenting all the sub-cortical brain structures. Moreover, this PhD thesis has been part of research frameworks within the projects of the ViCOROB group and different collaborating hospital centres. Furthermore, the methods developed during this PhD thesis were compiled into a toolbox and made publicly available for the research community.

Resum

Aquesta tesi doctoral es centra en el desenvolupament de mètodes basats en aprenentatge profund (*Deep Learning*) per a la segmentació precisa de les estructures cerebrals subcorticals en imatges de ressonància magnètica (RM). Aquestes estructures es troben sota el còrtex cerebral i inclouen el tàlem, el caudat, el putamen, el pallidum, l'hipocamp, l'amígdala i l'accumbens. Són estructures bilaterals, situades simètricament dins dels hemisferis esquerre i dret, i participen en activitats com l'emoció, el plaer, la memòria i la producció d'hormones. Les seves desviacions de volum s'associen a diferents malalties neurològiques com l'Alzheimer, el trastorn bipolar o l'esclerosi múltiple, entre d'altres. Tanmateix, la segmentació manual d'aquestes estructures és una tasca molt costosa i depèn en gran manera de l'expert que la realitza. Per tant, és important desenvolupar mètodes automatitzats robustos que permetin segmentar amb precisió aquestes estructures cerebrals subcorticals.

Aquesta tesi es compon de diferents contribucions que van des d'una nova proposta de segmentació basada en l'aprenentatge profund, fins a propostes per millorar la seva robustesa per a diversos reptes, com ara el problema del canvi de domini. Així doncs, primerament vam proposar una nova arquitectura de xarxa neuronal convolucional (CNN) 2.5D (és a dir, que analitza els tres plans ortogonals 2D d'un volum 3D) que combina característiques convolutives i espacials. A més, vam utilitzar una tècnica de sampleig de mostres selectives a partir dels contorns de les estructures a analitzar. Els resultats experimentals van demostrar l'efectivitat del plantejament proposat per segmentar amb precisió totes les estructures cerebrals subcorticals i van demostrar un rendiment molt satisfactori en els conjunts de dades públics i coneguts internacionalment com els *Multi-Atlas Labelling Grand Challenge MICCAI 2012* i *Internet Brain Segmentation Repository (IBSR)*.

En general, el rendiment de les xarxes CNN disminueix quan s'utilitza amb imatges d'un domini diferent del conjunt d'entrenament. Aquest problema es coneix com l'efecte del canvi de domini, on una xarxa entrenada amb volums de RM amb unes propietats determinades, com ara el tipus d'escàner, el protocol d'imatges i la resolució, no generalitza a altres imatges amb propietats diferents. Així doncs, en la següent contribució d'aquesta tesi, vam abordar aquest problema d'adaptació del domini mitjançant una estratègia supervisada d'aprenentatge de transferència. Els

experiments realitzats van mostrar l'efecte del canvi de domini en el rendiment dels mètodes d'aprenentatge profund i com l'aprenentatge de transferència proposat, no només resol aquest problema, sinó que també permet reduir dràsticament el nombre d'imatges d'entrenament i de paràmetres de la xarxa.

En l'etapa final, vam desenvolupar un nou enfocament per a l'adaptació del domini que no requeria cap imatge anotada manualment: un mètode d'aprenentatge profund totalment no supervisat. En la nova proposta es va emprar un anàlisi dels histogrames dins les pròpies capes convolucionals de la xarxa CNN per tal de minimitzar l'efecte causat per les diferències de domini directament. Aquesta contribució va mostrar millores significatives respecte dels resultats de segmentació basal, on no es va aplicar cap adaptació al domini, i obtenint resultats comparables al mètode FIRST. Per tal de demostrar la solidesa del mètode, vam testejar el seu ús en una aplicació diferent, la segmentació d'hiperintensitats dins la substància blanca cerebral. Els resultats experimentals van demostrar que l'adaptació no supervisada de la xarxa va millorar la proposta sense adaptació de domini i també el mètode no supervisat LST.

Les contribucions desenvolupades en aquesta tesi van obrir el camí per assolir un mètode automatitzat, precís i robust, basat en l'aprenentatge profund, per tal de segmentar les estructures cerebrals subcorticals. Aquesta tesi doctoral s'ha emmarcat en diferents projectes de recerca del grup ViCOROB i ha comptat amb la col·laboració de centres hospitalaris. A més a més, tots els mètodes desenvolupats durant aquesta tesi doctoral es van recopilar en una toolbox pública disponible per la comunitat científica.

Resumen

Esta tesis doctoral se centra en el desarrollo de métodos basados en el aprendizaje profundo (*Deep Learning*) para la segmentación precisa de las estructuras cerebrales subcorticales en imágenes de resonancia magnética (RM). Estas estructuras se encuentran bajo el córtex cerebral e incluyen el tálamo, el caudado, el putamen, el pallidum, el hipocampo, la amígdala y el accumbens. Son estructuras bilaterales, situadas simétricamente dentro de los hemisferios izquierdo y derecho, y participan en actividades como la emoción, el placer, la memoria y la producción de hormonas. Sus desviaciones de volumen se asocian a diferentes enfermedades neurológicas como el Alzheimer, el trastorno bipolar o la esclerosis múltiple, entre otros. Sin embargo, la segmentación manual de estas estructuras es una tarea muy costosa y depende en gran medida del experto que la realiza. Por lo tanto, es importante desarrollar métodos automatizados que permitan segmentar con precisión y robustez estas estructuras cerebrales subcorticales.

Esta tesis se compone de diferentes contribuciones que van desde una nueva propuesta de segmentación basada en el aprendizaje profundo, hasta propuestas para mejorar su robustez para varios retos, como el problema del cambio de dominio. Así pues, primeramente propusimos una nueva arquitectura de red neuronal convolucional (CNN) 2.5D (es decir, que analiza los tres planos ortogonales 2D de un volumen 3D) que combina características convolucionales y espaciales. Además, utilizamos una técnica de muestreo de muestras selectivas a partir de los contornos de las estructuras a analizar. Los resultados experimentales demostraron la efectividad del planteamiento propuesto para segmentar con precisión todas las estructuras cerebrales subcorticales y demostraron un rendimiento muy satisfactorio en los conjuntos de datos públicos y conocidos internacionalmente como el *Multi-Atlas Labelling Grand Challenge MICCAI 2.012* y el *Internet Brain Segmentation Repository (IBSR)*.

En general, el rendimiento de las redes CNN disminuye cuando se utiliza con imágenes de un dominio diferente del conjunto de entrenamiento. Este problema se conoce como el efecto del cambio de dominio, donde una red entrenada con volúmenes de RM con unas propiedades determinadas, tales como el tipo de escáner, el protocolo de imágenes y la resolución, no generaliza a otras imágenes con

propiedades diferentes. Así pues, en la siguiente contribución de esta tesis, abordamos este problema de adaptación del dominio mediante una estrategia supervisada de aprendizaje de transferencia. Los experimentos realizados mostraron el efecto del cambio de dominio en el rendimiento de los métodos de aprendizaje profundo y como el aprendizaje de transferencia propuesto, no sólo resuelve este problema, sino que también permite reducir drásticamente el número de imágenes de entrenamiento y de parámetros de la red.

En la etapa final, desarrollamos un nuevo enfoque para la adaptación del dominio que no requería ninguna imagen anotada manualmente: un método de aprendizaje profundo totalmente no supervisado. En la nueva propuesta se empleó un análisis de los histogramas dentro de las propias capas convolucionales de la red CNN para minimizar el efecto causado por las diferencias de dominio directamente. Esta contribución mostró mejoras significativas respecto de los resultados de segmentación basal, donde no se aplicó ninguna adaptación al dominio, obteniendo resultados comparables al método FIRST. Con el fin de demostrar la solidez del método, testeamos su uso en una aplicación diferente, la segmentación de hiperintensidades en la sustancia blanca cerebral. Los resultados experimentales demostraron que la adaptación no supervisada de la red mejoró la propuesta sin adaptación de dominio y también el método no supervisado LST.

Las contribuciones desarrolladas en esta tesis abrieron el camino para lograr un método automatizado, preciso y robusto, basado en el aprendizaje profundo, con el fin de segmentar las estructuras cerebrales subcorticales. Esta tesis doctoral se ha enmarcado en diferentes proyectos de investigación del grupo ViCOROB y ha contado con la colaboración de centros hospitalarios. Además, todos los métodos desarrollados durante esta tesis doctoral se recopilaron en una toolbox pública para su uso en la comunidad científica.

Chapter 1

Introduction

In this chapter, we describe the research context for the determined area of investigation of this PhD, justifying the relevance of the topic from the contextual point of view. Moreover, we list the projects that have been contributed to by the outcomes of the research accomplished in this thesis and introduce the pre-defined objectives. Then, the structure with brief descriptions for the remaining chapters are shown to give the reader a comprehensive outline of the thesis.

1.1 Research context

1.1.1 Sub-cortical brain structures

Segmentation of brain images in Magnetic Resonance Images (MRI) are addressed in various applications in medical imaging. Depending on the target of interest, segmentation in brain MRI could be performed to delineate brain regions by tissue type as grey matter (GM), white matter (WM), cerebrospinal fluid (CSF) [1], anatomical structures [2], or disease specific features caused by a neurological condition such as multiple sclerosis [3] or brain tumour [4]. In general, segmentation is performed as an initial step for further interpretation, for example, it could be immediately used for volumetric or shape analysis, or as a partial feature for a different study such as detection [5] and prediction [6]. Therefore, segmentation has been an exceptionally predominating research topic in brain imaging, which is also one of the biggest areas of investigation in the medical imaging field.

One of the essential research directions in medical brain imaging is segmentation of the sub-cortical brain structures. These deep grey-matter structures – the thalamus, putamen, caudate, pallidum, hippocampus, amygdala, accumbens (see Figure 1.1) – are located beneath the cerebral cortex and involved in complex ac-

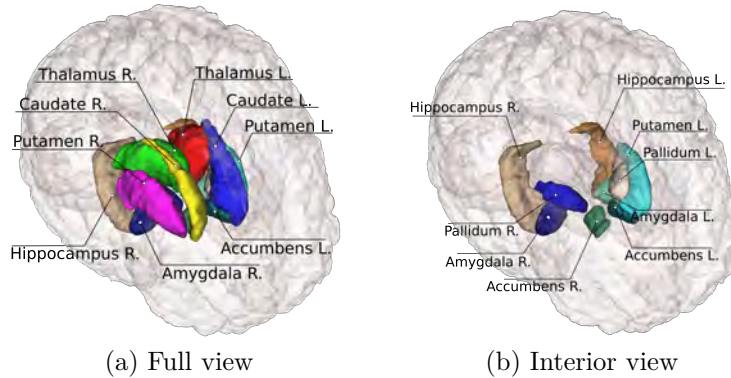


Figure 1.1: Sub-cortical brain structures. (a) Full view of the sub-cortical structures in the brain; (b) Interior structures that are not visible in the full view.

tivities such as memory, emotion, pleasure and hormone production.

Table 1.1 shows the clinical applications related to the sub-cortical brain structures analysis. As can be seen, the most common abnormalities for the diseases are related with morphological changes of the structures. Moreover, since these structures are situated in both hemispheres, changes in volumes relative to its bilateral counterpart can also be the biomarkers for neurological abnormalities such as Autism and Attention Deficit Hyper-activity Disorder (ADHD). Therefore, distinguishing the left and right parts of the structures is a prerequisite for automated segmentation methods.

Table 1.1 also shows that some of the brain structures are related with multiple diseases and also some neurological conditions are linked to several sub-cortical structures. Accordingly, in mood and behavioural studies, most, if not all the structures are taken into account. For instance, Schuetze et al. [18] investigated the shape and volume alterations in thalamus, putamen and pallidum structures in Autistic patients. Moreover, Hartberg et al. [34] reported that there are differences as well as similarities in sub-cortical structure relationships between patients with schizophrenia or bipolar disorder and healthy individuals. Moreover, Lamar et al. [19] showed that structural changes are associated with cardiovascular risk factors and Alzheimer's dementia. Taking this all into account, segmentation of the sub-cortical structures is essential in analysis of various neurological conditions, not only individually but also in their entirety.

Table 1.1: Clinical applications. Brain structure abnormalities associated with various diseases.

Structure	Implied Disease	Abnormality
Thalamus	Multiple sclerosis	Atrophy [7]
	Alzheimer's	Atrophy [8]
	Schizophrenia	Non-conclusive volume difference [9]
	Parkinson	Reduced volume [10]
Caudate	Huntington's disease	Atrophy [11]
	Tourette syndrome	Reduced volume [12]
	Autism	Increased right volume [13]
	Attention deficit hyperactivity disorder	Reduced right volume [14]
	Fragile X syndrome	Increased volume [15]
Putamen	REM sleep behaviour disorder	Reduced volume [16]
	OCD	Volume enlargement [17]
Pallidum	Autism	Changes in shape [18]
	Hypertension	Changes in volume [19]
Hippocampus	Alzheimer's	Atrophy [20]
	Temporal lobe epilepsy	Asymmetric atrophy [21]
	Posttraumatic stress disorder	Reduced volume [22, 23]
	Major depression	Reduced volume [24]
	Schizophrenia	Reduced volume [25]
Bipolar disorder	Non-conclusive volume difference [26, 27]	
Amygdala	Schizophrenia	Reduced volume [28]
	Anxiety disorders	Reduced left volume [29]
	Bipolar disorder	Non-conclusive volume difference [25, 27, 30, 31]
Accumbens	Huntington's disease	Atrophy [32]
	Apathy in Parkinson's disease	Atrophy [33]

1.1.2 Magnetic Resonance Imaging

Magnetic Resonance Imaging (MRI) is widely used in medical practice and has become a standard tool for diagnosis, disease follow-up, treatment evaluation and brain development monitoring [35, 36, 37, 38, 2]. It is a preferred imaging technology for a range of clinical applications due to its non-intrusiveness, acquisition speed, provision of good contrast between tissues and painlessness. Moreover, the non-ionising property of MRI allows the patients to be examined multiple times in a short period of time. Additionally, MRI scans represent 3D volumes, which keep spatial alignment of the internal organs in a 3D space. The volumes are studied from different views, usually from three orthogonal views – axial, sagittal and coronal as shown in Figure 1.2.

Depending on the image acquisition parameters, specifically, Repetition Time

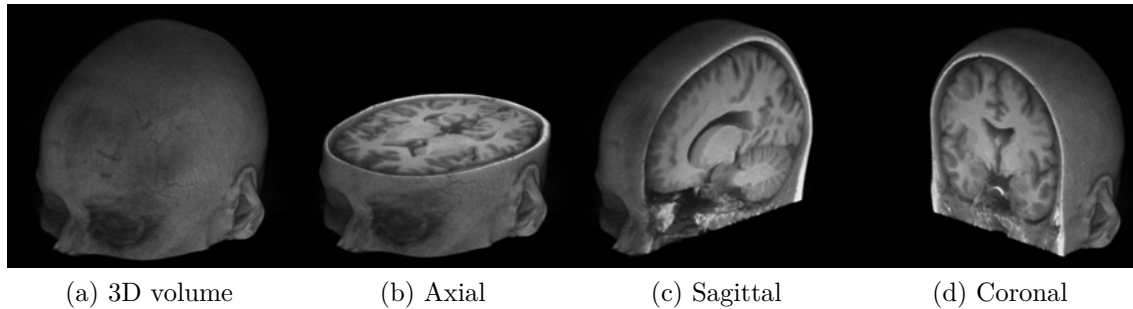


Figure 1.2: MRI volume in 3D and orthogonal views. (a) Full volume; (b) Axial view; (c) Sagittal view; (d) Coronal view.

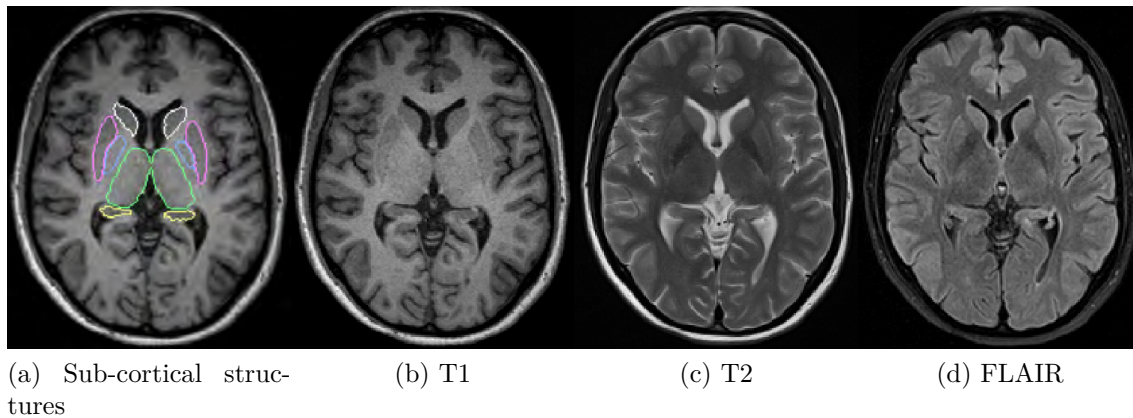


Figure 1.3: Different MRI Sequences. (a) T1-weighted with delineated sub-cortical structures; (b) T1-weighted; (c) T2-weighted; (d) FLAIR.

(TR) and Time to Echo (TE), different tissue contrasts are achieved in MRI scans. Short TR and TE produces images with dark CSF, light WM and grey GM, and this specification is called T1-weighted (T1-w) sequence (Figure 1.3b). In contrast to T1, by using longer TR and TE, one obtains images with bright CSF, dark-grey WM, and light GM, which is called T2-weighted (T2-w) sequence (Figure 1.3c). Another type of sequence, Fluid Attenuated Inversion Recovery (FLAIR) (Figure 1.3d), is similar to T2-w, but the TR and TE are much longer. In FLAIR sequences, the CSF, WM and GM appear dark, dark-grey and light-grey, respectively. There are other types of imaging sequences and different types of acquisition parameters that can be tuned to achieve certain image contrasts and characteristics [39].

Each of these sequences have their specific purposes, individually or together, because the tissue properties manifest in different ways. For instance, in multiple sclerosis diagnosis and monitoring, the lesions appear hyperintense (bright) in

T2 and FLAIR sequences, whereas in T1 images they appear hypointense (dark). Therefore, T2 and FLAIR sequences are the commonly used sequences in multiple sclerosis [35]. On the other hand, the brain structures are better distinguishable in the T1-weighted MR images as shown in Figure 1.3a. Accordingly, volumetric and shape analysis of the sub-cortical structures are done using the T1-w images.

Another important parameter in MRI scan acquisition is voxel spacing, which shows how much physical space is captured within a unit sample or voxel in a 3D volume. The spacing of pixels in a 2D plane of an MR image is referred as the resolution and the spacing between planes of a 3D volume is named as the slice thickness. The voxel spacing defines how sharp the details appear in the final output, thus, it is often preferred to have smaller voxel spacing (e.g. sub-millimetre). However, the smaller the voxel spacing and the slice thickness, the longer the acquisition time and the more slices to be acquired. Moreover, larger voxel spacing introduce more partial volume effect, where different tissue types appear blended together in a single voxel resulting blurred borders between different brain tissues. Accordingly, depending on the purpose of the scan, practitioners follow protocols that define a trade-off between the acquisition time and the image quality. For example, some common image resolutions include $1 \times 1 \times 1 \text{ mm}^3$ for structural T1-w imaging and $1 \times 1 \times 3 \text{ mm}^3$ for MS T2-lesion imaging.

1.1.3 Segmentation of the sub-cortical structures

Due to the importance of segmenting the sub-cortical structures in clinical practice, there is a demand for automated accurate segmentation methods because manually delineating all 14 structures in a 3D volume is a time-consuming and laborious task. Moreover, unlike the manual annotation of natural images, segmenting the brain structures has to be done by a trained neuroanatomist or technician. Additionally, the manual annotation is subject to inter- and intra-variability, which makes the segmentation outputs inconsistent. For example, the structure boundaries can be over or under segmented, meaning that quantification of their volumes will be different. This difference could change remarkably when the segmentation is done in images with low resolution. Taking this into account, we define some prerequisites for the automated segmentation methods:

- Accuracy – output has to correctly segment the target structures;
- Consistency – all segmentation outputs have to follow similar guidelines such as over or under segmentation;
- Robustness – the method has to perform similarly in different imaging domains, i.e. different acquisition setups such as voxel spacing, image dimensionality and different MRI scanner.

Over the past few years, several automated methods have been proposed for the sub-cortical brain structure segmentation. In the following sections, we describe some of the traditional approaches that are commonly used in medical practice as well as their degree of compliance with our defined requirements. Then, we introduce a new trending topic in medical image segmentation – the use of new Artificial Intelligence tool based on deep learning for improving image segmentation.

Traditional methods for structure segmentation

One of the well-known state-of-the-art automated methods for brain structure segmentation is FreeSurfer¹. This method is designed to segment brain MR images into 37 anatomical regions including all the sub-cortical structures [40]. The segmentation process is done in several steps: 1) image registration is applied to obtain a probabilistic atlas for the brain; 2) an anisotropic nonstationary Markov random field is used to assign labels for every voxel in an MR image; 3) post-processing is then applied to reclassify the neighbours of each label using a maximum likelihood classifier and removing small segmentation regions to avoid “islands” that are disconnected from the main segmentation region. FreeSurfer is highly dependent on registration, which makes the method sensitive to noise and motion artefacts. Although the method produces modest segmentation results in terms of accuracy compared to the state-of-the-art, it is consistent and robust to changes in the imaging protocol, voxel resolution and scanner.

Another well-known and commonly used method is FIRST [41] that is distributed with the FSL package². This method is based on an active shape and active appearance model, which is used to segment all the brain sub-cortical structures and the brainstem. As has been shown in [42], the performance of FIRST for images from different imaging domains is consistent, however, there is still room for improvement in terms of accuracy. Moreover, due to the powerful representation power of the active shape models, FIRST is highly robust to changes in image acquisition parameters. In fact, because of this property, FIRST always produces a segmentation output, even when the image has extreme imaging artefacts such as motion or noise. Nevertheless, the accuracy in this situation would be questionable, which is true for any other automated method.

There exist some other traditional methods that are based on multi-atlas segmentation strategies that have to be mentioned. In general, atlas based segmentation methods share common steps that include registering multiple manually segmented images to a target image. Then, deformed atlases are combined using label fusion to obtain the final segmentation for the target. Most common label fusion methods are

¹<https://surfer.nmr.mgh.harvard.edu>

²<https://fsl.fmrib.ox.ac.uk/fsl/fslwiki>

based on weighted voting, where each atlas is assigned a weight – either globally [43] or locally [44] – depending on its similarity to the target, thus, larger weight contributes more to the voting. However, the atlas to target similarities are obtained individually, which in turn may result in multiple atlases producing similar label errors. One of the top performing multi-atlas based methods for brain structure segmentation is PICSLS [45], which is also based on weighted label fusion technique. Unlike other common weighted voting strategies, it considers the possibility of a bias where multiple atlases may produce correlated segmentation errors. It is solved by building a pairwise uncertainty model between atlases that estimates the probability of each pair producing the same error. Then, instead of computing atlas weights individually, the similarity measure is derived between the target and each pair of atlases. One of the main drawbacks of this approach is the computational cost that includes multiple atlas registration, weight estimation and label fusion. This method was the winner of the MICCAI 2012 multi-atlas labelling grand challenge and has been used as a comparison standard in one of our proposals (Chapter 2) along with the previously mentioned tools FreeSurfer and FIRST.

Another traditional technique that has been successfully applied for brain structure segmentation includes non-local label fusion. In this type of approach, the label assignment is based on weights defined by similarities of a target voxel and all the atlas voxels in its neighbourhood. Due to the fact that it explores the neighbourhood of each voxel, the registration does not need to be precise, hence, it is possible to perform only linear registration instead of non-rigid deformation. An example of a non-local label fusion method includes the work of Coupé et al. [46], which introduced atlas preselection approach as an additional step, where the best atlases are selected depending on their similarity to target. Then, for each voxel, all dissimilar atlas voxels are discarded that do not contribute to the label fusion. The remaining voxels contribute to the weighting depending on their intensity similarity (luminance and contrast).

1.1.4 Deep learning

In general, deep learning is a part of machine learning algorithms that engage Artificial Neural Networks (ANNs). The ANN is an architectural paradigm that has layers of interconnected neurons – hence, the name neural network. Unlike the traditional machine learning methods that use hand-crafted features to train a classifier such as Support Vector Machine or Random Forest, ANNs learn these features directly from the training images [47]. Thus, making the features more relevant and specific to a particular task.

A type of ANN, Convolutional Neural Networks (CNNs), is an effective way of tackling imaging problems as the convolution operation is a more natural approach

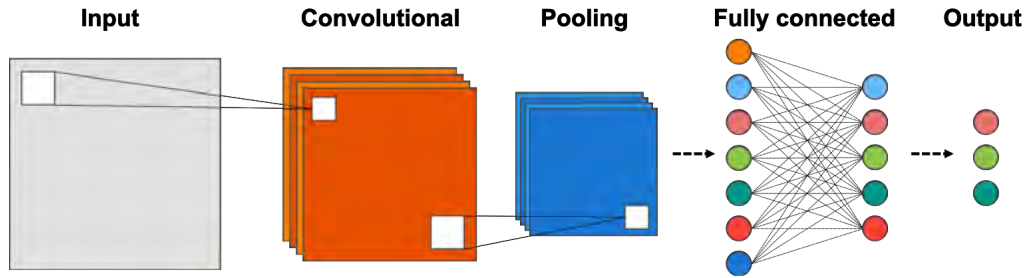


Figure 1.4: An example of a generic CNN illustrating the main building blocks that include: input, convolutional, pooling, fully connected and output layers.

for processing 2D and 3D signals. In this scenario, the network consists of convolutional layers that filter inputs with learned kernels. These kernels act in a similar way as the Gaussian for smoothing or Laplacian for edge detection but the difference is that they are created (learned) during the training process to produce an optimised set of features for the analysed problem. Sequential property of the convolutional layers – i.e. the output of one is an input to the next – allows the networks to model complicated non-linear functions and extract complex features from the inputs. There are multiple number of kernels per each layer that allows the network to capture various features at different levels. In fact, when the outputs (or activation maps) of convolutional layers of a trained CNN are observed, the early layers represent low level features as edges and blobs, whereas the deep layers produced more composite features. Figure 1.4 illustrates a generic network architecture that consists of the main building blocks of a CNN. The following paragraphs describe the main components for a CNN individually.

Input layer is where the network is fed with data. Depending on the architecture, the input data can be of two or more dimensions.

Convolutional layer consists of a user defined number of kernels. The kernel size is usually much smaller than the input, which reduces the number of connections and decreases the computational cost. This property of CNN is referred to as *sparse connectivity* [48]. Moreover, the input is filtered by a kernel using a predefined stride, which implies that one kernel is useful in all locations. This property, known as *parameter sharing* [48], allows the network to have less number of parameters. Additionally, convolutions are *shift equivariant*, meaning that one learned kernel remains effective in detecting features that are located in different spatial locations. Note that convolutions are not equivariant to scaling and rotation.

Pooling layer is usually applied after a convolutional layer and used to summarise the activation maps into a smaller representation. There are different types of pooling operations such as 1) max-pooling, where a neighbourhood of the activation map is replaced by the maximum element within that neighbourhood; and

2) average-pooling, where the area is replaced by the mean of its content. Pooling layers reduce the number of parameters and decrease memory and computational costs. Sometimes, pooling layers are substituted by convolutional layers that have stride size of two, which also reduce dimensionality of the input and at the same time produce activation maps.

Flattening layer is a commonly used element to reshape the 2D or 3D activation maps into one-dimensional representation to be further processed by the fully connected layers.

Fully connected layers are used to mine the features extracted by the previous layers and can be sequentially connected to each other producing a dense network.

Global Average Pooling layer (GAP) has been introduced to replace the flatten layer that is also used to obtain one-dimensional output. The difference from flattening is that instead of only reshaping the input it computes the mean value for each input feature map and then passes them to the following layers. Sometimes, the GAP layer replaces all the fully connected layers in the network, thus, reducing the computational cost and providing better generalisation performance as the fully connected layers are more prone to overfitting.

Output layer is the same as the fully connected layer but it represents the final output. The number of neurons in this layer is equal to the number of classes defined in the problem and shows which class the input belongs. The output can be single or dense, where the former produces one label per input and the latter produces a dense classification for a neighbourhood. Dense classification is achieved by replacing the fully connected and the output layers by convolutional layers with kernel sizes of one. This type of CNN is referred to as Fully Convolutional Neural Network (F-CNN). Depending on the problem, one or the other, CNN or F-CNN, could be more beneficial to achieve a better result.

Activation functions is an important piece that adds non-linearity property to neural networks. The outputs of each layer, both convolutional and fully connected, are passed through an activation function that applies transformation that convert input signals in different ways. There are different types of activation functions such as *sigmoid*, *tanh*, *softmax* and *Rectified Linear Unit (ReLU)*. For example, the sigmoid function transforms its input to a value between 0 and 1. Although the sigmoid and tanh functions are widely used, they suffer from the *vanishing gradient* problem, which occurs in large network architectures where the gradients of these functions become extremely low and make it difficult to which direction the weights should be updated. On the other hand, the ReLU does not have the vanishing gradient problem due to its formulation defined as $ReLU(x) = \max(0, x)$, which trims all the negative signals from the input, and its gradient is either 0 or 1. Nowadays the ReLU and its different formulations such as Parametric ReLU and

LeakyReLU, are one of the commonly used activation functions in literature.

Loss function is defined to determine the performance of the network and optimised during the training process. By formulating the loss function, one establishes a goal that shows what output should be achieved for a given input. Some of the common loss functions are *mean squared error* and *mean absolute error*, which are often used for regression, *categorical cross-entropy* used for classification. One of the common loss functions is the *Dice Similarity Coefficient* loss, which is widely used in F-CNNs for segmentation.

Back-propagation is the workhorse of learning in neural networks. Network training is done in main two steps: 1) forward pass – the input is forward propagated from the input to the output layer; and 2) backward pass – the cost from the loss function is back-propagated from the output to all the layers of the network. During the backward pass, proportional errors are obtained to define how much each parameter of the network is contributing to the final cost. Then, their corresponding gradients are calculated with respect to the loss function to define in which direction each parameter should be adjusted. This procedure is applied repeatedly to achieve optimal kernel and fully connected layer weights to represent important features for the given task [49].

There are different types of **optimisation** algorithms used during the back-propagation to make the training process more stable. For example, in *stochastic gradient descent (SGD)* the weights are updated for each sample at a time, whereas a version of SGD, *mini-batch gradient descent* applies weight updates for a portion of the whole training set and makes smoother parameter changes.

Regularisation techniques are used to improve the training process and avoid overfitting. *Dropout* is one of the well-known regularisation techniques, where random fraction of the connections are ignored during training. In doing so, the network becomes less prone to overfitting. There are other types of regularisation, such as weight or kernel regularisers that are used to apply penalties on layer parameters.

These are the main building blocks of CNNs, where new developments and proposals are mostly built upon.

Deep learning in medical imaging and its challenges

The success of deep learning methods in various tasks of computer vision and natural language processing has also influenced the field of medical image processing. As shown in Figure 1.5, over the past few years, the number of published research papers in medical imaging that use deep learning has increased drastically. The dominance of the deep learning approaches over traditional methods have become obvious as more and more teams have been employing such methods in the inter-



Figure 1.5: Number of publications in peer-reviewed journals from 2012 to 2020 related with deep learning in medical imaging. The numbers were attained using the google scholar service with keywords “medical imaging” and “deep learning”. Also, “-arxiv” and “-biorxiv” was used to exclude popular pre-prints. Queried: 25 April 2020.

national medical imaging challenges. For example, Bakas et al. [50] showed that in the recent years, deep learning approaches were becoming more popular and also the winning approaches for brain tumour segmentation.

Despite the fact that deep learning has been remarkably successful in natural image processing, its pinnacle has not been reached yet in medical image analysis. There are some challenges of deep learning applications that are specific to the medical image segmentation. One of the main challenges is the lack of available data with ground truth. As has been mentioned before, manual segmentation masks have to be done by trained experts and it is a time demanding work. Neural networks need a lot of training data to perform well on a given problem. Having more trainable parameters empowers the network with more representational capability. However, it becomes more data demanding and failure to do so results in overfitting of the network. The scarcity of medical images with ground truths makes networks to have less parameters and shallower than the deep architectures used in natural image processing.

Achieving a competent generalisation is another challenge in the application of deep learning in medical imaging. Variations in acquisition protocols in MRI cause differences in intensity and contrast even for the same type of sequence. Moreover, MRI machine vendors use their own proprietary reconstruction algorithms, which make images acquired in different scanners to have differences in appearance. Also, not all MRI scans have the same resolution and slice thickness. Although it does not cause any complications to a human expert, these differences are costly to auto-

mated methods, in particular, for deep learning approaches. An adequately trained CNN using a set of MRI images with the same characteristics cannot perform similarly when tested on a different image from a different scanner and protocol. This difficulty is known as the *domain shift problem* and it is a new and active research topic.

Another challenge is the class imbalance, which is immensely exhibited in the sub-cortical structure segmentation problem. The ideal scenario in neural network training is when the number of samples per class is approximately equal – i.e. balanced set. However, it is often not applicable in practice, where a training set has samples per class that are drastically varying in quantity. Training with imbalanced set causes the network to overfit to the class with a larger sample pool. Volumes of the sub-cortical structures vary drastically: the average difference between the largest (thalamus) and the smallest (accumbens) structures is $\approx 8.5 \text{ cm}^3$. This class imbalance makes segmentation of the sub-cortical structures more challenging.

1.2 Research background

The Computer Vision and Robotics (ViCOROB) institute, operating within the University of Girona, was established in 1996 and since then has been working on medical image analysis and robotics fields. The main focus of the medical imaging team in the early stages was aimed in the segmentation and registration of mammogram images. In 2009, the group started its collaboration with several medical experts in multiple sclerosis (MS) to develop new tools for brain MRI analysis that could be transferred for clinical use. The prior experience and effort of the team working on the new line of research widened the investigation framework which currently includes image pre-processing, registration, segmentation of MS, WMH and stroke lesions, tissue segmentation, atrophy analysis, new MS lesion detection, and brain structure segmentation.

The research line of this PhD is done within the framework of the following projects:

- [2015 - 2017] NICOLE: “Herramientas de neuroimagen para mejorar el diagnóstico y el seguimiento clínico de los pacientes con Esclerosis Múltiple”. Awarded in 2014 by the Spanish call Retos de investigación 2014. Ref: TIN2014- 55710-R.
- [2015 - 2019] BiomarkEM.cat: “New technologies applied to clinical practice for obtaining biomarkers of atrophy and lesions in magnetic resonance images of patients with multiple sclerosis”. Awarded in 2015 by the Fundació la Marató de TV3.

- [2016 - 2019] wASSABI: “Automatic brain Structures Segmentation As potential imaging BIomarkers”. Awarded in 2016 by the Ministerio de ciencia y tecnologia. Ref: TIN2015-73563-JIN
- [2018 - 2020] EVOLUTION: “Predictive models for multiple sclerosis using brain magnetic resonance imaging biomarkers”. Awarded in 2017 by Ministerio de ciencia y tecnologia. RETOS 2017. Ref: DPI2017-86696-R.

1.3 Objectives

Develop new deep learning based methods for brain structure segmentation and move towards domain adaptation As part of the NICOLE, BiomarkEM.cat, EVOLUTION and wASSABI projects, the goal of this thesis is described as the following.

To develop novel, automated methods for automatic brain structure segmentation in MRI using deep learning and development of domain adaptation strategies to overcome the repercussions caused by the domain shift.

In order to fulfil the prerequisites defined for automated methods for sub-cortical structure segmentation – accuracy, consistency and robustness – the planning for the thesis has been divided into different sub-objectives that tackle each requirement step-by-step. Accordingly, to successfully carry out the main target of the thesis the following sub-objectives were determined:

- To develop a novel method for segmentation of the sub-cortical brain structures using deep learning that complies with the first essential requirement – *accuracy*.
- To improve the *consistency* and *robustness* aspects of the proposal described in the previous sub-objective using supervised domain adaptation techniques such as transfer learning.
- To minimise the manual effort to maintain the *consistency* and *robustness* of the segmentation method using unsupervised domain adaptation approach.
- To evaluate the proposed algorithms using the international publicly available and in-house datasets. Also, to compare with the state-of-the-art methods from the literature.

Throughout the advancement of the settled objectives, as a team that strives for reproducibility of the findings in the research, we made the source codes publicly available to the medical imaging community.

1.4 Document structure

This thesis is done as the compendium of two Q1 SCR journal publications and one submission to a Q1 SCR journal that cover chapters from 2 to 4. The remaining part of the thesis are structured as shown below. Also, a graphical presentation of the document structure linking all the chapter is illustrated in Figure 1.6.

- **Chapter 2. Automated sub-cortical brain structure segmentation combining spatial and deep convolutional features.** In this chapter, we present our proposed deep learning method for brain structure segmentation. This chapter is based on the paper published in the *Medical Image Analysis* journal in 2018.
- **Chapter 3. Supervised domain adaptation for automatic sub-cortical brain structure segmentation with minimal user interaction.** This chapter is a continuation from the previously proposed method to solve the domain shift problem in MR Images with different imaging characteristics by reducing the supervision requirement. This chapter is based on the paper published in the *Nature: Scientific Reports* journal in 2019.
- **Chapter 4. Unsupervised domain adaptation in deep learning for brain magnetic resonance image segmentation.** In this chapter, we present another way of overcoming the domain shift problem based on an unsupervised domain adaptation approach. This work is also a continuation for the previous chapter, however, compared to the transfer learning, this method does not require manually annotated labels. The work shown in this chapter has been submitted to the *Artificial Intelligence in Medicine* journal in 2020.
- **Chapter 5. Results and discussion.** This chapter discusses the overall results and key findings obtained during this PhD thesis.
- **Chapter 6. Conclusions and future work.** In this chapter, the main conclusions based on the contributions are presented. Moreover, we discuss on possible future research directions.

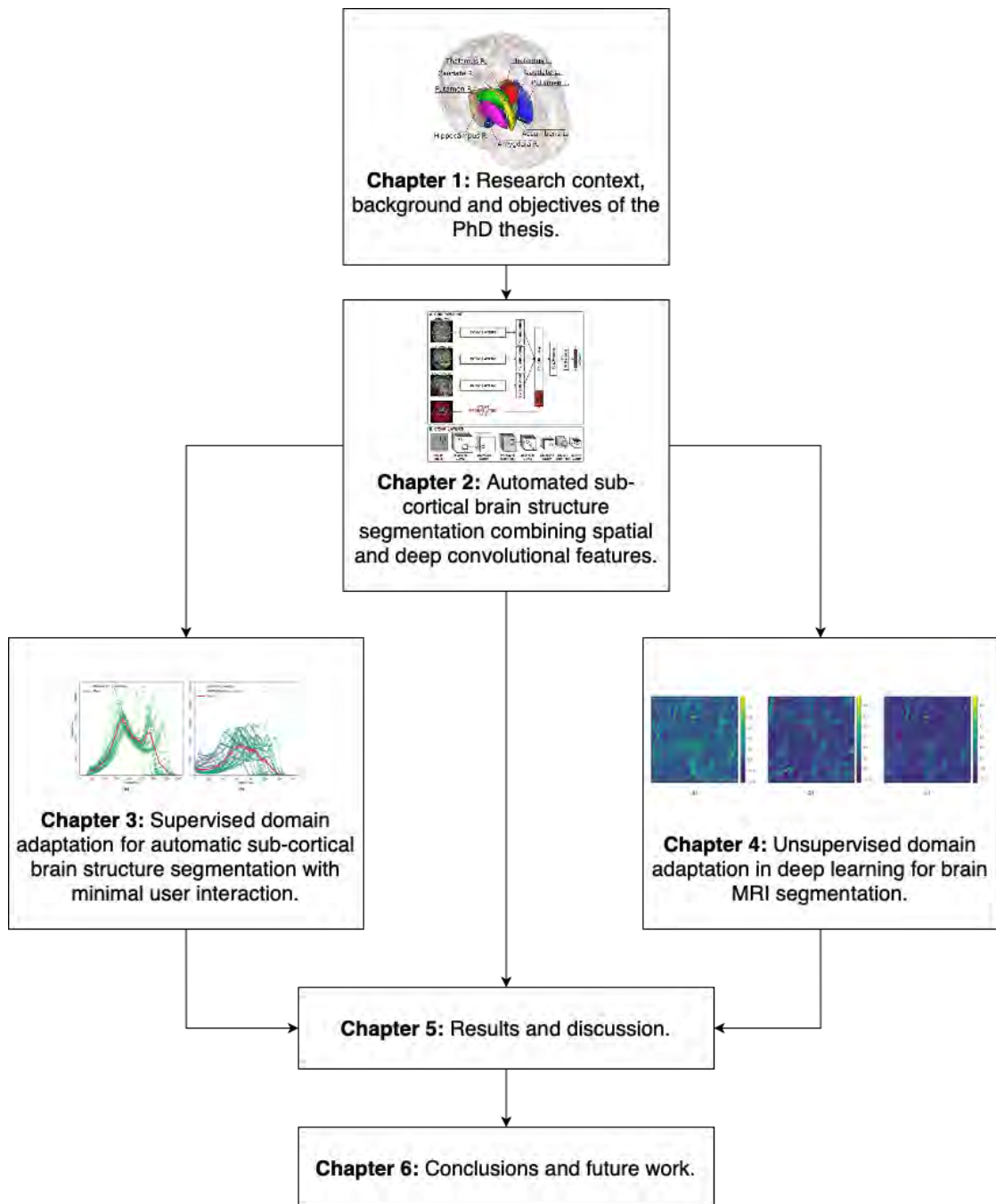


Figure 1.6: Structure of the PhD thesis. Chapter 1 gives an introduction to the topic of brain structure segmentation and deep learning. Chapters 2 to 4 introduce the main contributions published or submitted to international journals. Discussion on the results and important aspects of the contributions are given in Chapter 5. Chapter 6 concludes the work done in this PhD thesis and discusses possible improvements and directions for future work. The links between the chapters illustrate the flow and progression of the thesis.

Chapter 2

Automated sub-cortical brain structure segmentation combining spatial and deep convolutional features

In this chapter, we present our proposed deep learning method for automatic segmentation of sub-cortical brain structures on MRI. This approach uses explicit spatial features defined by atlas probabilities and implicit features extracted by the convolutional neural network. The explicit features are integrated within a CNN and guides the network to overcome some regions with intensity irregularities and brain abnormalities. Moreover, the proposed sample selection technique – negative samples from structure boundaries – showed to be effective in better delineating the structures borders. The method has shown state-of-the-art performance in two well-known and publicly available datasets – MICCAI 2012 Challenge and IBSR 18. The proposal has been published in the following paper:

<p>Paper published in the Medical Image Analysis (MIA) OPEN ACCESS Volume: 48, Pages: 177–186, Published: June 2018 DOI: 10.1016/j.media.2018.06.006 JCR CSAI IF 8.880, Q1(5/133)</p>
--



Contents lists available at ScienceDirect

Medical Image Analysis

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

Automated sub-cortical brain structure segmentation combining spatial and deep convolutional features



Kaisar Kushibar^{1,*}, Sergi Valverde¹, Sandra González-Villà, Jose Bernal, Mariano Cabezas, Arnau Oliver, Xavier Lladó

Institute of Computer Vision and Robotics, University of Girona, Ed. P-IV, Campus Montilivi, Girona, 17003, Spain

ARTICLE INFO

Article history:

Received 26 September 2017

Revised 1 March 2018

Accepted 9 June 2018

Available online 15 June 2018

Keywords:

Brain

MRI

Sub-cortical structures

Segmentation

Convolutional neural networks

ABSTRACT

Sub-cortical brain structure segmentation in Magnetic Resonance Images (MRI) has attracted the interest of the research community for a long time as morphological changes in these structures are related to different neurodegenerative disorders. However, manual segmentation of these structures can be tedious and prone to variability, highlighting the need for robust automated segmentation methods. In this paper, we present a novel convolutional neural network based approach for accurate segmentation of the sub-cortical brain structures that combines both convolutional and prior spatial features for improving the segmentation accuracy. In order to increase the accuracy of the automated segmentation, we propose to train the network using a restricted sample selection to force the network to learn the most difficult parts of the structures. We evaluate the accuracy of the proposed method on the public MICCAI 2012 challenge and IBSR 18 datasets, comparing it with different traditional and deep learning state-of-the-art methods. On the MICCAI 2012 dataset, our method shows an excellent performance comparable to the best participant strategy on the challenge, while performing significantly better than state-of-the-art techniques such as FreeSurfer and FIRST. On the IBSR 18 dataset, our method also exhibits a significant increase in the performance with respect to not only FreeSurfer and FIRST, but also comparable or better results than other recent deep learning approaches. Moreover, our experiments show that both the addition of the spatial priors and the restricted sampling strategy have a significant effect on the accuracy of the proposed method. In order to encourage the reproducibility and the use of the proposed method, a public version of our approach is available to download for the neuroimaging community.

© 2018 The Authors. Published by Elsevier B.V.

This is an open access article under the CC BY-NC-ND license.

(<http://creativecommons.org/licenses/by-nc-nd/4.0/>)

1. Introduction

Brain structure segmentation in Magnetic Resonance Images (MRI) is one of the major interests in medical practice due to its various applications, including pre-operative evaluation and surgical planning, radiotherapy treatment planning, longitudinal monitoring for disease progression or remission (Kikinis et al., 1996; Phillips et al., 2015; Pitiot et al., 2004), etc. The sub-cortical structures (i.e. thalamus, caudate, putamen, pallidum, hippocampus, amygdala, and accumbens) have attracted the interest of the

research community for a long time, since their morphological changes are frequently associated with psychiatric and neurodegenerative disorders and could be used as biomarkers of some diseases (Debernard et al., 2015; Mak et al., 2014). Therefore, segmentation of sub-cortical brain structures in MRI for quantitative analysis has a major clinical application. However, manual segmentation of MRI is extremely time consuming and hardly reproducible due to inter- and intra- variability among operators, highlighting the need for automated accurate segmentation methods.

Recently, González-Villà et al. (2016), reviewed different approaches for brain structure segmentation in MRI. One of the commonly used automatic brain structure segmentation tools in medical practice is FreeSurfer,² which uses non-linear registration and an atlas-based segmentation approach (Fischl et al., 2002). Another classical approach, also popular in the medical community,

* Corresponding author.

E-mail addresses: kaisar.kushibar@udg.edu (K. Kushibar), sergio.valverde@udg.edu (S. Valverde), sgonzalez@eia.udg.edu (S. González-Villà), jose.bernal@udg.edu (J. Bernal), mariano.cabezas@udg.edu (M. Cabezas), aoliver@eia.udg.edu (A. Oliver), xavier.llado@udg.edu (X. Lladó).

¹ These authors contributed equally to this work.

² <https://surfer.nmr.mgh.harvard.edu/>.

is the method proposed by Patenaude et al. (2011) – FIRST, which is included into the publicly available software FSL.³ This method uses the principles of Active Shape (Cootes et al., 1995) and Active Appearance Models (Cootes et al., 2001) that are put within a Bayesian framework, allowing to use the probabilistic relationship between shape and intensity to its full extent.

In recent years, deep learning methods, in particular, Convolutional Neural Networks (CNN), have demonstrated a state-of-the-art performance in many computer vision tasks such as visual object detection, classification and segmentation (Krizhevsky et al., 2012; He et al., 2016; Szegedy et al., 2015; Girshick et al., 2014). Unlike handcrafted features, CNN methods learn from observed data (LeCun et al., 1998) making relevant features to a specific task. Naturally, CNNs are also becoming a popular technique applied in medical image analysis. There have been many advances in the application of deep learning in medical imaging such as expert-level performance in skin cancer classification (Esteva et al., 2017), high rate detecting cancer metastases (Liu et al., 2017), Alzheimer's disease classification (Sarraf and Tofghi, 2016), and spotting early signs of autism (Hazlett et al., 2017).

Some CNN methods have also been proposed for brain structure segmentation. One of the common ways used in the literature is patch-based segmentation, where patches of a certain size are extracted around each voxel and classified using a CNN. Application of 2D, 3D, 2.5D patches (patches from the three orthogonal views of an MRI volume) and their combinations including multi-scale patches can be found in the literature for brain structure segmentation (Brébisson and Montana, 2015; Bao and Chung, 2016; Milletari, 2017; Mehta et al., 2017). Combining patches of different views and dimensions is done in a multi-path manner, where CNNs consist of different branches corresponding to each patch type, i.e. parallel interconnected processing modules analyze each of the inputs. In contrast to patch-based CNNs, fully convolutional neural networks (FCNN) produce segmentation for a neighborhood of an input patch (Long et al., 2015). Shakeri et al. (2016) adapted the work of Chen et al. (2016) for semantic segmentation of natural images using FCNN. Moreover, 3D FCNNs, which segment a 3D neighborhood of an input patch at once, have been investigated by Dolz et al. (2018) and Wachinger et al. (2018). Although FCNNs show improvement in segmentation speed due to parallel segmentation of several voxels, they suffer from a high number of parameters in the network in comparison with patch-based CNNs.

It is common to apply post-processing methods to refine the final segmentation output. Inference of CNN-priors and statistical models such as Markov Random Fields and Conditional Random Fields (Lafferty et al., 2001) were used in the experiments of Brébisson and Montana (2015), Shakeri et al. (2016), and Wachinger et al. (2018). A modified Random Walker based segmentation refinement has been also proposed by Bao and Chung (2016).

Apart from implicit information that is provided by the extracted patches from MRI volumes, explicit characteristics distinguishing spatial consistency have been studied. Brébisson and Montana (2015) included distances to centroids to their networks. Wachinger et al. (2018) used the Euclidean and spectral coordinates computed from eigenfunctions of a Laplace-Beltrami operator of a solid 3D brain mask, to provide a distinctive perception of spatial location for every voxel. These kinds of features provide additional spatial information, however, extracting these explicit features from an unannotated MRI volume requires some preliminary operations to be attended (e.g. repetitive training of the network to compute initial segmentation mask).

From the reviewed literature, we have observed that most of the current deep learning approaches for sub-cortical brain struc-

ture segmentation focus on segmenting only the large sub-cortical structures (thalamus, caudate, putamen, pallidum). However, other important small structures (i.e. hippocampus, amygdala, accumbens), which are used for examining neurological disorders such as schizophrenia (Altshuler et al., 1998; Lawrie et al., 2003), anxiety disorder (Milham et al., 2005), bipolar disorder (Altshuler et al., 1998), Alzheimer (Fox et al., 1996), etc., are not considered. These small structures have smaller volume – hence, lower number of samples – compared to the other larger structures, which hinders training deep learning strategies and makes the segmentation task more challenging. In this paper, we present our approach for segmenting the sub-cortical structures: a new 2.5D CNN architecture – i.e., the three orthogonal views of a 3D volume – that incorporates probabilistic atlases as spatial features. Although probabilistic atlases have been used before in deep learning methods (Ghafoorian et al., 2017), they have never been applied for segmenting the sub-cortical brain structures. Within our research, unlike most of the existing deep learning approaches, we address segmenting all the sub-cortical structures, including the smallest ones. To the best of our knowledge, this is the first deep learning method incorporating atlas probabilities into a CNN for sub-cortical brain structure segmentation. Moreover, we propose a particular sample selection technique, which allows the neural network to learn to segment the most difficult areas of the structures in the images, and also show its importance in achieving higher accuracy. We test the proposed strategy in two well-known datasets: MICCAI 2012⁴ (Landman and Warfield, 2012) and IBSR 18⁵; and compare our results with the classical and recent CNN strategies for brain structure segmentation. Additionally, we make our method publicly available for the community, accessible online at https://github.com/NIC-VICOROB/sub-cortical_segmentation.

2. Method

2.1. Input features

In our method, we employ 2.5D patches to incorporate information from three orthogonal views of a 3D volume. In our case, each patch has a size of 32×32 pixels. Although 3D patches may provide more information of surroundings for the voxel that is being classified, they are computationally and memory expensive. Thus, by using 2.5D patches, we approximate the information that is provided by a 3D patch in computational time and memory efficient manner.

Along with the appearance based features provided by the T1-w MRI, we employ spatial features extracted from a structural probabilistic atlas. In our experiments, we used the well-known Harvard-Oxford (Caviness et al., 1996) atlas template in MNI152 space distributed with the FSL package,⁶ which has been built using 47 young adult healthy brains. In our method, first, T1-w image of the considered datasets using a block matching approach (Ourselin et al., 2000). Then, non-linear registration of the atlas template to subject volume is applied using fast free-form deformation method (Modat et al., 2010). The deformation field obtained after the registration is used to move the probabilistic atlas into the subject space. Registration processes have been carried out using the well known and publicly available tool NiftyReg.⁷ Afterwards, vectors of size 15, corresponding to seven anatomical structures with left and right parts separately and background, were

⁴ <https://masi.vuse.vanderbilt.edu/workshop2012>.

⁵ <https://www.nitrc.org/projects/ibsr>.

⁶ <https://fsl.fmrib.ox.ac.uk/fsl/fslwiki>.

⁷ <http://cmictig.cs.ucl.ac.uk/wiki/index.php/NiftyReg>.

³ <https://fsl.fmrib.ox.ac.uk/fsl/fslwiki>.

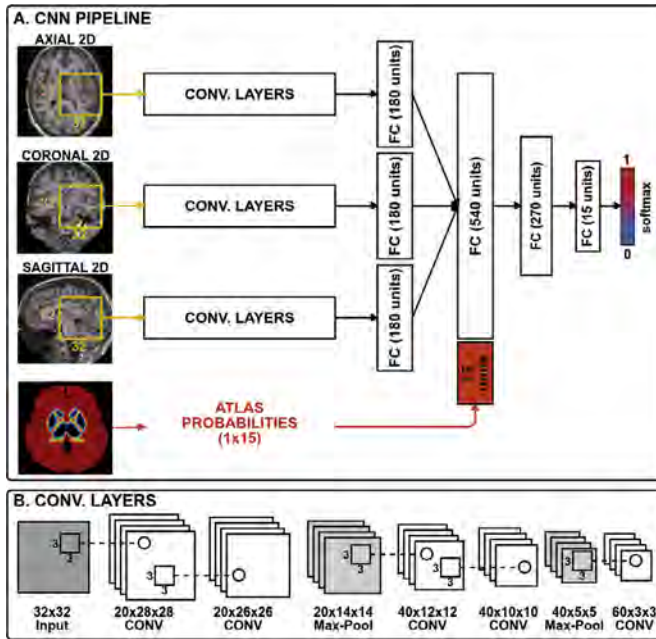


Fig. 1. The proposed 2.5D CNN architecture has three convolutional branches and a branch for spatial prior. 2D patches of size 32×32 pixels are extracted from three orthogonal views of a 3D volume. Spatial prior branch accepts a vector of size 15 with atlas probabilities for each of the 14 structures and background.

extracted from probabilistic atlas for every voxel and used as an input feature to train the network.

2.2. CNN architecture

Fig. 1 illustrates our proposed CNN architecture. It consists of three branches to process the patches extracted from axial, coronal, and sagittal views of a 3D volume, and one branch corresponding to the spatial priors. The branch for the spatial prior accepts a vector of size 15 with atlas probabilities for each structure and the background. The first three branches have the same organization of convolutional and max-pooling layers as shown in **Fig. 1(B)**. All the feature maps of the convolutional layers are passed through the Rectified Linear Unit (ReLU) activation function (Glorot et al., 2011). For all the convolutional layers, kernels of size 3×3 are set to make the CNN deep without losing in performance and bursting the number of parameters as it has been studied in Simonyan and Zisserman (2014). Then, the outputs of the convolutional layers are flattened and followed by fully connected (FC) layers with 180 units each. Next, FC layers of each branch including atlas proba-

bilities are fully connected to two consecutive FC layers with 540 and 270 units. The final classification layer has 15 units with the softmax activation function.

The atlas probabilities provide the network with spatial information, i.e. likelihood of an input patch belonging to one of the 14 classes or background. This information can be added either as additional input sequences (i.e. as additional channels to T1-w image patches) or later in the fully connected layers. However, when working with a high number of classes, the former way of atlas incorporation becomes impractical in terms of training/testing time due to an increase in number of trainable parameters of the network as well as a vast increase in memory usage. Accordingly, we use the latter approach, where we provide a vector of size 15 with each element corresponding to the central pixel's probability of belonging to one of the classes, which is fused with the output of the first fully connected layer after the convolutional part of the network.

2.3. CNN training

For training our network, we extract 2.5D patches from the training set and using the provided ground truth labels we optimize the kernel and fully connected layer unit weights based on the loss function. In the proposed network we employ the categorical cross-entropy loss function, which is minimized using the Adam (Kingma and Ba, 2014) optimization method. This technique automatically controls the learning rate and uses moving averages of the parameters, which allows the step size to be effectively large and converges to optimal step size without tuning it manually.

When training a CNN, it is important to take into account how the training samples are extracted from an image. Random selection of certain number of samples from an image is one of the common techniques in the literature. However, when it comes to the segmentation of the sub-cortical structures, the background (negative) samples turn out to be dispersed in the subject volume. Hence, it would lead to imperfect segmentation results on the borders of the structures, which are the most delicate areas to process due to the low contrast between the structure and the background. Therefore, we propose to extract the negative samples only from the structure boundaries as shown in **Fig. 2**. In doing so, we force the network to learn only from the structure boundaries and dismiss other parts of the background.

The training sample selection is performed as follows: from all the available training images, we first select the positive samples from all the voxels from the 14 sub-cortical structures. Then, the same number of negative samples are randomly selected from the structure boundaries within five voxel distance, forming a balanced dataset of sub-cortical and boundary voxels. More details about

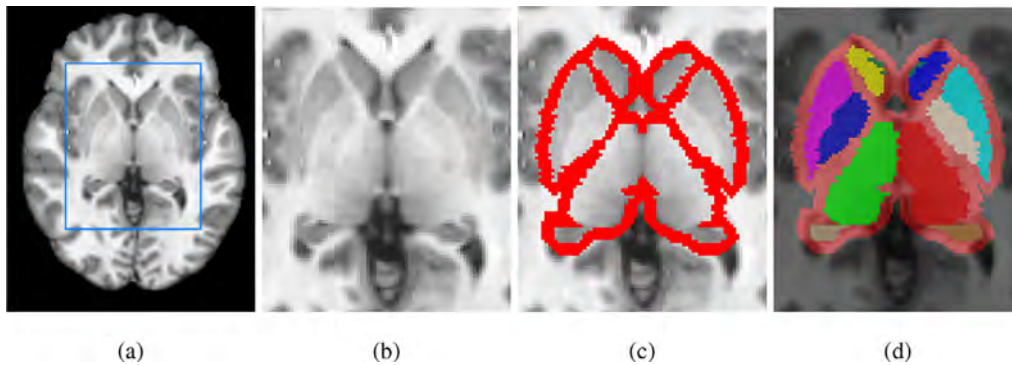


Fig. 2. Negative sample selection from the boundaries of the target structures. (a) T1-w image with a rectangle representing the ROI; (b) T1-w ROI; (c) structure boundaries; (d) ground truth labels with boundaries.

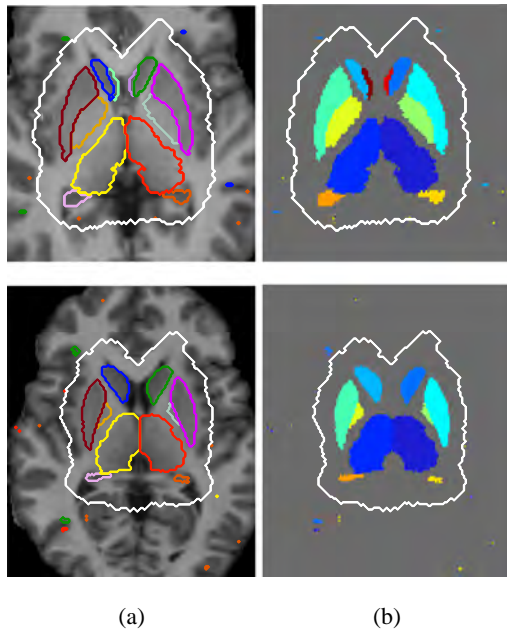


Fig. 3. Two different examples of segmentation outputs without using ROI before post-processing. Columns: a) T1-w image and segmentation result; b) Segmentation output on solid background for better visualization of spurious outputs. ROIs are delineated in white.

batch size and number of epochs of the training process for the selected datasets will be given in Section 3.

2.4. CNN testing

To perform the segmentation of a new image volume, we extract all the patches from the image and predict class label probabilities using the trained CNN. Then, we assign a label corresponding to the maximum a posteriori probability for the central pixel of each input patch. Notice that knowing the order of the patch extraction is important to be able to reconstruct the final segmentation output. We also take advantage of the location of the sub-cortical structures, which are located in the central part of the brain. Due to the knowledge provided by the atlases, regions of interest (ROI) are automatically defined for all the subject volumes to achieve faster training and testing speeds.

Since the network has been trained with the negative samples extracted only from the structure boundaries, it produces spurious outputs in unseen areas of the background when segmenting a testing volume. In order to overcome this issue, we apply a post-processing step, where for each class only the region with the biggest volume within the ROI is preserved. For such post-processing, it is important to make sure that the volume and location of the misclassified regions are not larger than the volumes of any of the structures nor adjacent to the structure boundaries. When segmenting a new image, we send only ROI as an input to the network. In doing so, we ensure that the misclassified voxels have small size, as most of the input patches correspond to the sub-cortical area. Moreover, since the network is well trained to classify the boundaries of the structures, there will be no misclassified voxels adjacent to the structure boundaries. Fig. 3 illustrates examples, when all the patches were set as input to the network. As it can be observed, the background is well defined around the structure borders, and most of the spurious outputs appear outside the ROI.

2.5. Implementation and technical details

The proposed method has been implemented in the Python language,⁸ using Lasagne⁹ and Theano¹⁰ (Bergstra et al., 2011) libraries. All experiments have been run on a GNU/Linux machine box running Ubuntu 16.04, with 32 GB RAM memory. CNN training has been carried out on a single TITAN-X GPU (NVIDIA corp, United States) with 12 GB RAM memory. The proposed method is currently available for downloading at our research website.¹¹

3. Results

This section presents the results obtained by the proposed method on two datasets. The first dataset is the one provided in the MICCAI Multi-Atlas Labeling challenge¹² (Landman and Warfield, 2012) and the second is a publicly available dataset from the Internet Brain Segmentation Repository¹³ (IBSR). Details of these datasets and the corresponding results will be given in Sections 3.2 and 3.3 respectively.

3.1. Evaluation measures

For evaluating the proposed method, we selected two metrics that are commonly used in the literature. These are overlap and spatial distance-based metrics, which show similarity and discrepancy of automatic and manual segmentations. The first measurement is Dice Similarity Coefficient (DSC) (Dice, 1945) defined as the following for automatic segmentation A and manual segmentation B :

$$DSC(A, B) = \frac{2|A \cap B|}{|A| + |B|}. \quad (1)$$

DSC measures the overlap of the segmentation with the ground truth on a scale between 0 and 1, where the former shows no overlap and the latter represents 100% overlap with the ground truth.

For the spatial distance based metric, Hausdorff Distance (HD) is used in our experiments. This metric is defined as a function of the Euclidean distances between the voxels of A and B as:

$$HD(A, B) = \max(h(A, B), h(B, A)), \quad (2)$$

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

In other words, HD is the maximum distance from all the minimum distances between boundaries of segmentation and boundaries of the ground truth.

Similarly to Wachinger et al. (2018), we used Wilcoxon signed-rank test to test the statistical significance of: 1) the differences in DSC and HD between our and state-of-the-art methods; and 2) the effect of using spatial features and the proposed sample selection technique.

3.2. MICCAI 2012 dataset

This dataset consists of 35 T1-w MRI volumes split into 15 cases for training and 20 cases for testing. Manually segmented ground truth for each image is available as well, which contains 134 structures overall. In our experiments, we extracted 14 classes corresponding to seven sub-cortical structures with left and right parts separately. All the subject volumes have even voxel spacing of 1 mm^3 with a size of 256×256 voxels in axial, sagittal, and coronal views respectively.

⁸ <https://www.python.org/>.

⁹ <http://lasagne.readthedocs.io>.

¹⁰ <http://deeplearning.net/software/theano/>.

¹¹ https://github.com/NIC-VICOROB/sub-cortical_segmentation.

¹² <https://masi.vuse.vanderbilt.edu/workshop2012>.

¹³ <https://www.nitrc.org/projects/ibsr>.

Table 1

MICCAI 2012 dataset results. Mean DSC \pm standard deviation and HD \pm standard deviation values for each structure obtained using FreeSurfer, FIRST, PICSL, and our method. Structure acronyms are: left thalamus (Tha.L), right thalamus (Tha.R), left caudate (Cau.L), right caudate (Cau.R), left putamen (Put.L), right putamen (Put.R), left pallidum (Pal.L), right pallidum (Pal.R), left hippocampus (Hip.L), right hippocampus (Hip.R), left amygdala (Amy.L), right amygdala (Amy.R), left accumbens (Acc.L), right accumbens (Acc.R) and average value (Avg.). Highest DSC and HD values for each structure are shown in bold.

Method	FreeSurfer Fischl (2012)		FIRST Patenaude et al. (2011)		PICSL Wang and Yushkevich (2013)		Our method	
	DSC	HD	DSC	HD	DSC	HD	DSC	HD
Tha.L	0.830 \pm 0.018	4.94 \pm 1.01	0.889 \pm 0.018	4.65 \pm 0.90	0.920 \pm 0.013	3.22 \pm 0.99	0.921 \pm 0.018	3.39 \pm 1.13
Tha.R	0.849 \pm 0.021	4.76 \pm 0.75	0.890 \pm 0.017	4.39 \pm 0.92	0.924 \pm 0.008	3.11 \pm 0.79	0.920 \pm 0.016	3.31 \pm 1.01
Cau.L	0.808 \pm 0.079	9.89 \pm 3.09	0.797 \pm 0.046	3.56 \pm 1.30	0.885 \pm 0.074	3.44 \pm 1.89	0.894 \pm 0.071	3.32 \pm 2.00
Cau.R	0.801 \pm 0.042	10.39 \pm 3.09	0.837 \pm 0.117	4.16 \pm 1.37	0.887 \pm 0.065	3.60 \pm 1.67	0.892 \pm 0.057	3.51 \pm 1.67
Put.L	0.771 \pm 0.039	6.31 \pm 1.09	0.860 \pm 0.060	3.79 \pm 1.76	0.909 \pm 0.042	3.07 \pm 1.40	0.916 \pm 0.023	2.63 \pm 1.09
Put.R	0.799 \pm 0.026	5.85 \pm 0.84	0.876 \pm 0.080	3.26 \pm 1.23	0.908 \pm 0.046	2.91 \pm 1.41	0.914 \pm 0.031	2.75 \pm 0.99
Pal.L	0.693 \pm 0.189	3.89 \pm 1.07	0.815 \pm 0.088	2.89 \pm 0.71	0.873 \pm 0.032	2.52 \pm 0.54	0.843 \pm 0.101	2.38 \pm 0.76
Pal.R	0.792 \pm 0.085	3.45 \pm 0.98	0.799 \pm 0.060	3.18 \pm 0.93	0.874 \pm 0.047	2.49 \pm 0.59	0.861 \pm 0.049	2.59 \pm 0.61
Hip.L	0.784 \pm 0.054	6.35 \pm 1.87	0.809 \pm 0.022	5.49 \pm 1.66	0.871 \pm 0.024	4.34 \pm 1.66	0.876 \pm 0.020	4.48 \pm 2.02
Hip.R	0.794 \pm 0.025	6.19 \pm 1.59	0.810 \pm 0.140	4.80 \pm 1.66	0.869 \pm 0.022	4.01 \pm 1.45	0.879 \pm 0.020	3.76 \pm 1.23
Amy.L	0.585 \pm 0.064	5.05 \pm 0.97	0.721 \pm 0.053	3.54 \pm 0.72	0.832 \pm 0.026	2.44 \pm 0.29	0.833 \pm 0.032	2.39 \pm 0.39
Amy.R	0.576 \pm 0.076	5.43 \pm 0.90	0.707 \pm 0.054	4.11 \pm 0.75	0.812 \pm 0.033	2.72 \pm 0.50	0.821 \pm 0.027	2.72 \pm 0.69
Acc.L	0.630 \pm 0.055	4.28 \pm 1.11	0.699 \pm 0.089	6.81 \pm 8.76	0.790 \pm 0.050	2.57 \pm 0.67	0.799 \pm 0.052	2.39 \pm 0.64
Acc.R	0.443 \pm 0.065	5.47 \pm 1.02	0.678 \pm 0.081	3.93 \pm 1.75	0.783 \pm 0.058	2.65 \pm 0.76	0.791 \pm 0.067	2.54 \pm 0.65
Avg.	0.725 \pm 0.137	5.87 \pm 2.48	0.799 \pm 0.094	4.18 \pm 2.76	0.867 \pm 0.061	3.08 \pm 1.27	0.869 \pm 0.064	3.01 \pm 1.30

3.2.1. Experimental details

Skull-stripping was applied to extract the brain and cut out other parts appearing in the MRI such as eyes, skull, skin, and fat using the BET algorithm ([Smith, 2002](#)). The spatial intensity variations on the MRI volumes were corrected using a bias field correction algorithm – N4ITK ([Tustison et al., 2010](#)), which is included in the publicly available ITK¹⁴ toolkit. Both preprocessing methods were run with default parameters.

In our experiments, we trained a single model using the available training set of 15 images, while we tested the other 20 images as provided in the original MICCAI 2012 Challenge. From the training set, we extracted around 1.5M (750K of sub-cortical voxels and 750K of boundary voxels) sample patches of size 32×32 pixels from three orthogonal views, where around 1.1M (75%) were used for training and 400K samples for validation (25%). The extracted patches were passed to the network for training in batches of size 128. The network was set to train for 200 epochs, yet, we applied early stopping of the training process to prevent over-fitting. The training process was automatically terminated when the validation accuracy did not increase after 20 epochs.

3.2.2. Comparison with other available methods

The performance of the proposed approach is compared with widely used tools in medical practice – FreeSurfer and FIRST. We also compared the performance of our method with the one of PICSL ([Wang and Yushkevich, 2013](#)) method, which is a multi-atlas based segmentation strategy that uses joint fusion technique with corrective learning. PICSL was the winner of the MICCAI 2012 Challenge for brain structure segmentation and still shows the best results on this dataset. We used the default parameters for the methods of FreeSurfer and FIRST to produce segmentation masks for the testing volumes. Accordingly, the training and testing split matches the configuration we used for evaluating the proposed method. We have to note that, with this dataset, there were no individually reported numerical results for each of the sub-cortical structure in other CNN based approaches.

3.2.3. Results

[Table 1](#) shows overall and per structure mean DSC and HD values on the MICCAI 2012 dataset. According to the results, our method showed significantly ($p < 0.001$) higher DSC of 0.869 than FIRST and FreeSurfer which yielded 0.799 and 0.725 overall mean

DSC, respectively. Moreover, as it can be observed, the HD values showed similar behavior as DSC, where the proposed approach significantly outperformed both of these methods ($p < 0.001$), in average, with a reduction of 1.17 mm and 2.86 mm with respect to FIRST and FreeSurfer. Also, the DSC and HD results of our method with respect to FreeSurfer and FIRST were significantly higher for all the structures individually. Our method did not show a significant difference in comparison with PICSL in terms of DSC ($p > 0.05$), having similar mean of 0.867 and 0.869 for PICSL and our method, respectively. However, there was a significant improvement for the left caudate, right putamen, right hippocampus, and left accumbens structures ($p < 0.05$). The average HD values of our approach and PICSL also confirmed previous DSC numbers, but no significant increase per structure was observed. [Fig. 4](#) shows a qualitative comparison of segmentation outputs from FreeSurfer, FIRST, PICSL, and our method. As it can be observed, FreeSurfer provided less accurate segmentation output with coarse structure boundaries. FIRST produced smooth segmentation on the borders, however, the overlap between the ground truth was poor. Our method's segmentation output was similar to the one of PICSL's and both of the methods had consistent structure boundaries, which were not far from the ground truth. [Fig. 5c](#) depicts an example of low DSC score (0.61) produced by our method for the right caudate structure. As it can be seen from the T1-w image, the intensities above the caudate structure are similar to the ones of the actual structure region defined by the manual segmentation ([Fig. 5a](#)). This irregularity led to an apparent atlas registration error, where a region outside the structure was defined with high atlas probabilities ([Fig. 5b](#)). Even though our network takes both – the intensities and the atlas probabilities – into account, these kinds of pathological cases may lead to inaccurate segmentation results. However, this is also a common issue for other methods as seen in [Fig. 5d](#), where an atlas based method (PICSL) also fails in accurately segmenting this structure.

Apart from having similar results to the best performing method on this dataset, our strategy gained a good improvement in training and segmentation times. According to [Landman and Warfield \(2012\)](#), PICSL took 330 CPU hours for training 138 classifiers used for correcting systematic errors. Reported segmentation time of PICSL with optimal parameters was more than 50 minutes per subject volume ([Wang and Yushkevich, 2013](#)). In comparison with the above, the execution time of our CNN strategy was around 8 hours for training and less than 5 min for testing, including the atlas registration.

¹⁴ <https://itk.org/>.

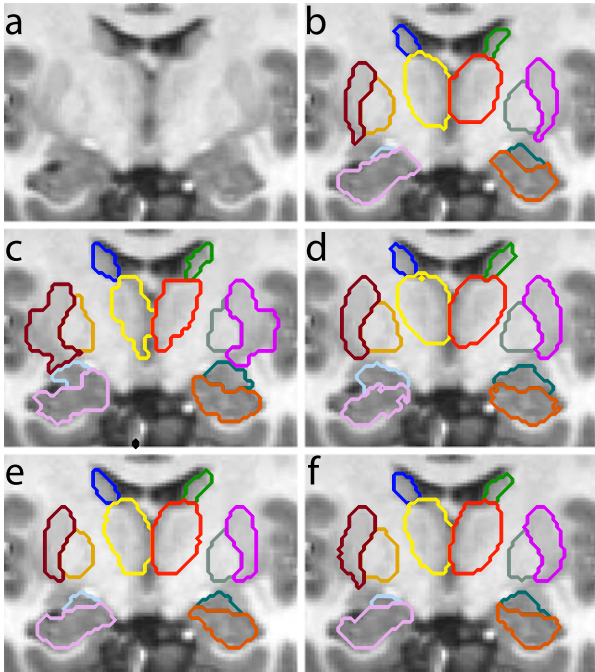


Fig. 4. Qualitative comparison of segmentation outputs obtained by FreeSurfer, FIRST, PICSL, and our method on MICCAI 2012 dataset. a) T1-w image; b) Ground truth; c) FreeSurfer; d) FIRST; e) PICSL; f) Our method. Visible structures on coronal view: thalamus, caudate, pallidum, putamen, hippocampus, and amygdala.

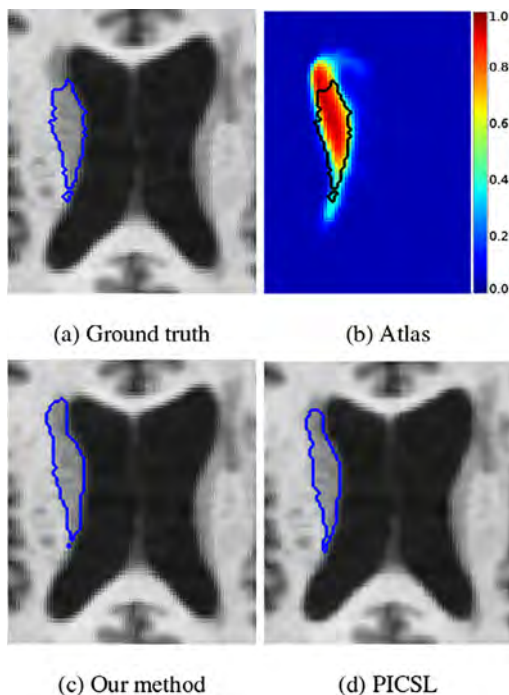


Fig. 5. Example of a segmentation result with a low DSC value for the right caudate structure. a) manual segmentation; b) probabilistic atlas with overlaid manual segmentation shown in black; c) our method; d) PICSL.

3.3. IBSR 18 dataset

This dataset consists of 18 T1-w subject volumes with manually segmented ground truth with 32 classes. Similarly to the MICCAI 2012 dataset, we extracted 14 classes corresponding to seven sub-cortical brain structures with left and right parts separately. The subject volumes of this dataset

have dimensionality of $256 \times 256 \times 128$ and different voxel spacings: $0.84 \times 0.84 \times 1.5 \text{ mm}^3$, $0.94 \times 0.94 \times 1.5 \text{ mm}^3$, and $1.00 \times 1.00 \times 1.5 \text{ mm}^3$. Images in this dataset have lower contrast and resolution in comparison with the MICCAI 2012 dataset, which makes the segmentation task even more challenging.

3.3.1. Experimental details

For the experiments with this dataset, we followed the same preprocessing steps as done with the MICCAI 2012 dataset, which included skull-stripping and bias field correction. Since there was no training and testing split on this dataset, we performed our experiments using a leave-one-subject-out cross-validation scheme. For each 17-1 fold, we extracted around 1.1M patches from each of the three orthogonal views, divided into 825K (75%) training and 220K (25%) validation sets. Each model was trained for 200 epochs applying also early stopping policy in the training process after 20 epochs.

3.3.2. Comparison with other available methods

For this dataset, our results will be compared against: 1) to the commonly used FreeSurfer and FIRST methods including the statistical significance test, since the evaluation values for each subject volume were computed by us using the corresponding tools; and 2) to recent CNN approaches of Shakeri et al. (2016), Mehta et al. (2017) (BrainSegNet), Bao and Chung (2016) (MS-CNN), and Dolz et al. (2018). The results for the recent methods were taken from their corresponding papers exactly as they have been reported. We have to mention that most of the CNN based methods report results only for a specific group of sub-cortical structures, but do not show or consider the results for the other, yet important, sub-cortical structures. Note also that the comparison on HD metric is present only for FreeSurfer, FIRST and our method, but not for other considered methods because most of the approaches do not report HD values.

3.3.3. Results

Table 2 shows the mean DSC and HD values for each of the evaluated methods. Our method showed a better performance in comparison to both FreeSurfer and FIRST methods for all the sub-cortical structures. The overall DSC mean of our method was significantly higher than both of the methods ($p < 0.001$), with mean DSC of 0.740, 0.808, and 0.843 for FreeSurfer, FIRST and the proposed strategy, respectively. In terms of HD values, our method showed overall mean of 4.49, whereas FreeSurfer and FIRST yielded 5.21 and 4.50, respectively. The proposed strategy significantly outperformed FreeSurfer with ($p < 0.001$), however the difference with FIRST was not significant ($p > 0.05$). As shown in Table 2, FreeSurfer performed worst for almost all the structures, while FIRST and our method showed similar performance. On both thalamus structures, our method showed lowest score in comparison with the other methods, however it yielded better HD for the small structures like amygdala, accumbens, and hippocampus. In general, HD metric is very sensitive to outliers, hence, a few misclassified voxels can cause considerable reduction in performance as seen in the results for the thalamus structure in our method.

Compared to other CNNs, our approach outperformed the method proposed by Shakeri et al. (DSC = 0.808) on the eight evaluated structures. Similarly, the performance of the proposed approach was also superior on the six structures evaluated in the work of Mehta et al. (DSC = 0.841). Further, we compare our method with MS-CNN, which has reported average DSC values for six structures for left and right parts together (overall DSC = 0.807). Our method's mean DSC on these structures was 0.859, which was higher than the result of MS-CNN (0.807) and yielded higher DSC scores for all the structures. Finally, when compared with the work of Dolz et al., our method showed a compa-

Table 2

Comparison of our method with the state-of-the-art methods as well as previous CNN approaches on IBSR dataset in terms of DSC, HD, and standard deviation. Structure acronyms are: left thalamus (Tha.L), right thalamus (Tha.R), left caudate (Cau.L), right caudate (Cau.R), left putamen (Put.L), right putamen (Put.R), left pallidum (Pal.L), right pallidum (Pal.R), left hippocampus (Hip.L), right hippocampus (Hip.R), left amygdala (Amy.L), right amygdala (Amy.R), left accumbens (Acc.L), right accumbens (Acc.R). “–” represents no results were reported on corresponding structure. The average (Avg.) values show mean DSC for the presented structure DSC scores. Highest DSC and HD values for each structure are shown in bold.

Method	FreeSurfer		FIRST		Shakeri	BrainSegNet	MS-CNN	Dolz	Our method	
	DSC	HD	DSC	HD					DSC	DSC
Tha.L	0.815 ± 0.056	5.367 ± 1.168	0.893 ± 0.017	3.819 ± 0.850	0.866 ± 0.023	0.88 ± 0.050	0.889	0.92	0.910 ± 0.014	7.159 ± 0.402
Tha.R	0.864 ± 0.022	4.471 ± 1.245	0.885 ± 0.012	4.273 ± 1.137	0.874 ± 0.021	0.90 ± 0.029			0.914 ± 0.016	7.256 ± 0.571
Cau.L	0.796 ± 0.050	6.435 ± 1.939	0.783 ± 0.044	4.128 ± 1.575	0.778 ± 0.053	0.86 ± 0.047	0.849	0.91	0.896 ± 0.018	4.054 ± 1.412
Cau.R	0.809 ± 0.048	8.201 ± 2.443	0.870 ± 0.027	3.687 ± 0.791	0.783 ± 0.068	0.88 ± 0.048			0.896 ± 0.020	4.153 ± 1.061
Put.L	0.789 ± 0.038	5.310 ± 0.923	0.869 ± 0.020	4.421 ± 1.185	0.838 ± 0.026	0.91 ± 0.022	0.875	0.90	0.900 ± 0.014	5.216 ± 1.788
Put.R	0.829 ± 0.031	4.716 ± 1.189	0.880 ± 0.010	4.725 ± 1.814	0.824 ± 0.039	0.91 ± 0.023			0.904 ± 0.012	4.577 ± 0.410
Pal.L	0.632 ± 0.171	4.652 ± 1.294	0.810 ± 0.033	3.477 ± 0.572	0.763 ± 0.031	0.81 ± 0.089	0.787	0.86	0.825 ± 0.050	3.849 ± 0.574
Pal.R	0.774 ± 0.032	3.966 ± 0.793	0.809 ± 0.037	3.990 ± 1.075	0.736 ± 0.055	0.83 ± 0.086			0.829 ± 0.046	3.700 ± 0.576
Hip.L	0.760 ± 0.036	5.787 ± 1.264	0.806 ± 0.023	5.571 ± 1.592	–	0.81 ± 0.065	0.788	–	0.851 ± 0.024	4.177 ± 1.087
Hip.R	0.767 ± 0.060	5.615 ± 1.600	0.817 ± 0.023	4.349 ± 0.984	–	0.83 ± 0.071			0.851 ± 0.024	4.124 ± 0.824
Amy.L	0.661 ± 0.069	5.521 ± 1.517	0.742 ± 0.064	4.648 ± 1.950	–	0.76 ± 0.087	0.654	–	0.763 ± 0.052	4.326 ± 0.822
Amy.R	0.690 ± 0.067	4.720 ± 1.553	0.757 ± 0.062	4.402 ± 1.493	–	0.71 ± 0.087			0.768 ± 0.058	4.292 ± 1.064
Acc.L	0.604 ± 0.071	3.634 ± 0.783	0.684 ± 0.098	7.770 ± 8.803	–	–	–	–	0.744 ± 0.053	3.026 ± 0.676
Acc.R	0.574 ± 0.074	4.507 ± 1.077	0.703 ± 0.076	3.733 ± 1.482	–	–	–	–	0.752 ± 0.047	2.995 ± 0.609
Avg.	0.740 ± 0.110	5.207 ± 1.761	0.808 ± 0.080	4.499 ± 2.810	0.808 ± 0.063	0.841 ± 0.064	0.807	0.898	0.843 ± 0.071	4.493 ± 1.533

Table 3

Effect of spatial features and the proposed sample selection technique. MICCAI 2012 dataset. Random sampling – method without using the sample selection from boundaries (including the spatial priors). No atlas – method without incorporating atlas priors (using the sampling technique). Final method – proposed method that includes both the spatial features and the sampling technique. Structure acronyms are: left thalamus (Tha.L), right thalamus (Tha.R), left caudate (Cau.L), right caudate (Cau.R), left putamen (Put.L), right putamen (Put.R), left pallidum (Pal.L), right pallidum (Pal.R), left hippocampus (Hip.L), right hippocampus (Hip.R), left amygdala (Amy.L), right amygdala (Amy.R), left accumbens (Acc.L), right accumbens (Acc.R). The values with an asterisk (*) indicate that the final method obtained significantly higher results than that of the strategy without atlas priors. Highest DSC values for each structure are shown in bold.

Method	Random sampling	No atlas	Final method
Tha.L	0.860 ± 0.013	0.911 ± 0.024	0.921 ± 0.017*
Tha.R	0.862 ± 0.014	0.917 ± 0.017	0.920 ± 0.016
Cau.L	0.831 ± 0.067	0.880 ± 0.103	0.894 ± 0.071*
Cau.R	0.834 ± 0.048	0.864 ± 0.131	0.892 ± 0.057
Put.L	0.871 ± 0.024	0.900 ± 0.073	0.916 ± 0.023*
Put.R	0.872 ± 0.027	0.913 ± 0.029	0.914 ± 0.031
Pal.L	0.784 ± 0.040	0.852 ± 0.086	0.843 ± 0.101
Pal.R	0.775 ± 0.057	0.833 ± 0.099	0.861 ± 0.049*
Hip.L	0.778 ± 0.034	0.871 ± 0.019	0.876 ± 0.020*
Hip.R	0.770 ± 0.026	0.876 ± 0.018	0.879 ± 0.020*
Amy.L	0.709 ± 0.025	0.824 ± 0.037	0.833 ± 0.032*
Amy.R	0.716 ± 0.054	0.819 ± 0.035	0.821 ± 0.027
Acc.L	0.744 ± 0.060	0.796 ± 0.052	0.799 ± 0.052
Acc.R	0.689 ± 0.091	0.753 ± 0.106	0.791 ± 0.067*
Avg.	0.792 ± 0.076	0.858 ± 0.083	0.869 ± 0.064*

able performance, although this last work showed slightly higher averaged DSC values for the four biggest structures.

3.4. Effect of the spatial priors

We ran experiments using the proposed method with and without spatial priors to determine the effect of using such features to the segmentation performance on both datasets. For this experiment, we analyzed the results in terms of DSC on the MICCAI 2012 dataset. We did not present the results of this experiment for the IBSR 18 dataset for simplicity, since it produced a similar outcome. In order to test our network without the spatial features, we modified the architecture (Fig. 1) by removing the branch of atlas probabilities and keeping only three branches of convolutional layers.

Table 3, shows DSC results of our method with random sampling, without using spatial features, and the final method. Inclu-

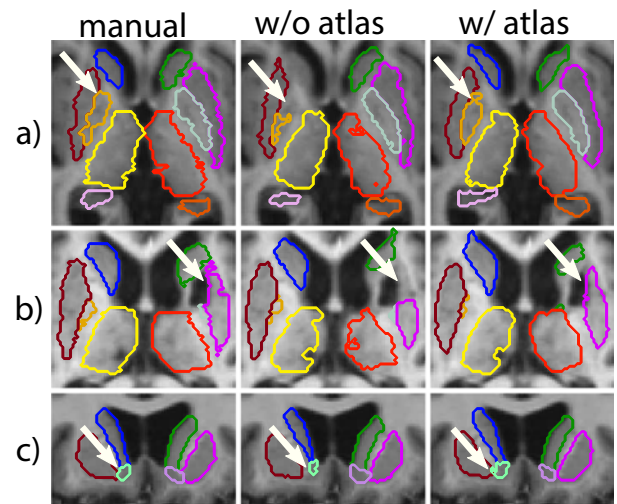


Fig. 6. Comparison of segmentation outputs for difficult areas of the (a) pallidum, (b) putamen, and (c) accumbens structures in some of the images from MICCAI 2012 dataset using the proposed method with and without the spatial priors. Regions of remarkable improvement when employing the atlas priors are indicated with arrows.

sion of the spatial features significantly improved the overall DSC ($p < 0.001$), as well as the results for almost all the structures. The segmentation difference can be seen from Fig. 6, where difficult areas of the pallidum, putamen, and accumbens structures were segmented better by the method that comprised the spatial features. Hence, the spatial priors helped to overcome difficult areas, producing more accurate segmentation for some images that had intensity and shape irregularities that could not be observed in any of the training images. Although the spatial priors are effective to overcome these sort of issues, it could be misleading in certain cases, where the irregularity is extremely large - as shown in Fig. 6b, where a hole is present in the left pallidum structure. The final method obtained lower score for the left pallidum on this subject volume, which downgraded the average DSC for this structure (Table 3). On the other hand, our method without the spatial priors segmented this area better than the final approach, however, the overall difference was not significant ($p = 0.101$).

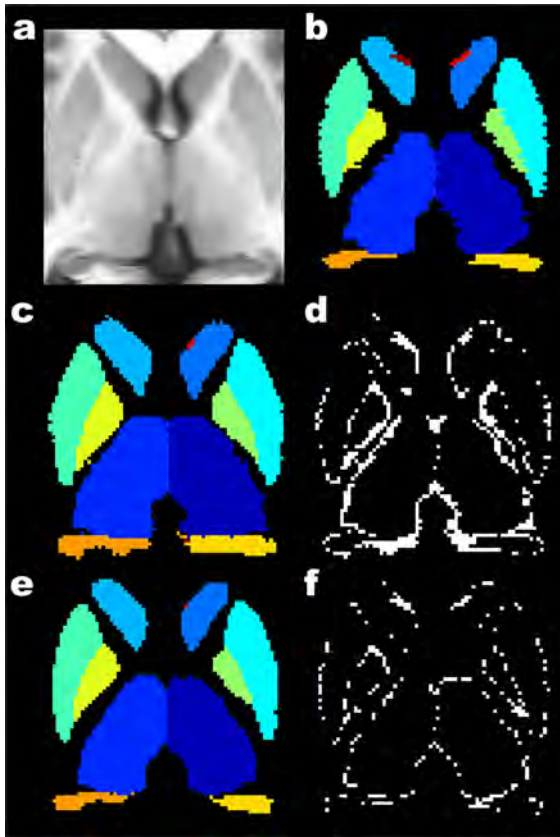


Fig. 7. Illustration of misclassification occurrence on borders. MICCAI 2012 dataset. (a, b) T1-w image and manual segmentation; (c, d) segmentation using random sample selection and difference from ground truth; (e, f) segmentation using the sample selection from borders and difference from ground truth.

3.5. Effect of sample selection

In this section, we show the effect of sample selection from structure boundaries using the MICCAI 2012 dataset. For this experiment, random sample selection from all the brain tissues has been used for training the network. For every epoch, we extracted the same number of voxels (1.5M), split equally into the sub-cortical structures (750K) and background (750K). Here, background voxels were randomly selected from whole brain volume, instead of selecting only from structure boundaries (see Fig. 2d). The network was again trained for 200 epochs using the same configuration. Spatial features were also included in training.

Table 3 shows the results corresponding to this experiment. Mean DSC obtained with our network without using the sample selection technique was 0.792 compared to 0.869 of the final approach. Accordingly, the proposed sample selection technique significantly improved the network's performance in average as well as for each of the structures ($p < 0.001$). Fig. 7 illustrates the segmentation results produced by our final approach and without applying sampling from borders. As it can be seen from the difference between ground truth and segmentation masks, the final strategy produced better segmentation on the boundaries than random sample selection method. In fact, the difference of our segmentation and the ground truth mask was not substantial, but only a few voxels. We also can observe that the intensities on the border voxels of the structures are mostly confounding. Therefore, assigning these voxels to the structure or background is highly dependent on ground truth.

4. Discussion

In this paper, we have proposed a fully automated 2.5D patch-based CNN approach that combines both convolutional and a priori spatial features for accurate segmentation of the sub-cortical brain structures. In our approach, a structural sub-cortical atlas has been registered into the image space to extract the spatial probability of each voxel, and, later, fused with the extracted convolutional features in the fully connected layers. The inclusion of the spatial information increases the execution time by adding atlas registration. However, it allows us to filter out misclassified regions that have bigger size than the actual structures in the segmentation output, which may appear in unobserved areas (i.e. not included in the training phase) of the brain as a consequence of applying restricted sampling. As seen in all the experiments, the addition of the spatial priors and the restricted sampling strategy have a significant effect on the accuracy of the proposed method, outperforming or showing a comparable performance to both classic as well as other novel deep learning approaches for segmenting the sub-cortical structures.

Compared to other state-of-the-art techniques such as FreeSurfer and FIRST, the spatial agreement of the proposed method with the manual segmentation is clearly higher in all evaluated datasets. As seen in other radiological tasks, this reinforces the effectiveness of CNN techniques when manual expert annotations are available. On the MICCAI 2012 dataset, our method shows an excellent performance, slightly over-performing the best challenge participant strategy – PICSL. Although not directly evaluated, our method clearly reduces the training and inference time. However, it has to be noted that most of the execution time of PICSL is due to highly computational registration processes which were carried out on CPU, while our method relies on GPU processors to speed-up training. Other CNN methods have also been evaluated on the MICCAI 2012 database (Wachinger et al., 2018; Mehta et al., 2017). However, these works do not report exact evaluation values for sub-cortical structures, hence, no direct comparison can be established.

In contrast, different CNN methods that have been evaluated using the IBSR 18 dataset have reported exact numerical values. When compared to other CNN approaches, our method also showed a significant increase in the performance with respect to most of them, and a comparable behavior with the method proposed by Dolz et al. However, as seen in Section 3.3, previous studies do not always deal with all sub-cortical structures, restricting a more detailed comparison with respect to our proposal. Additionally, the training methodology also differed among the strategies. In this aspect, although all our experiments were carried out using the leave-one-out approach, we also repeated our IBSR 18 experiments using a six-fold (15 training and three testing) validation strategy to perform a fair comparison with some of the considered methods. The complete results of the six-fold validation strategy were not depicted in the paper for simplicity, but, our network achieved similar results with only 0.005 of difference in DSC with respect to the leave-one-out strategy, showing the robustness of the proposed approach to changes in the number of training images.

According to the experimental results, employing the spatial features to the CNN significantly improved the performance of the network. The atlas priors showed to be useful in guiding the network when segmenting the difficult areas. As we have seen in Section 3.4, CNN that leveraged the spatial priors coped with these intensity based difficulties. Accordingly, by providing the atlas probabilities, we make sure that the anatomical shape and structure are taken into account before assigning a label to a voxel. Since the sub-cortical structures follow the similar anatomical structure in all patients, the inclusion of the spatial features

makes the segmentation approach more robust to irregularities in intensity based features obtained from T1-w images by providing additional location-based information. Despite being prone to the inherent errors in image registration and not showing as much DSC improvements as in border-selective sampling (Table 3), the addition of these a priori spatial class probabilities, or other explicit fused problem-specific information, may have other direct benefits such as reduction of the effect of low contrast, poor resolution, presence of noise, and artifacts close to the structure boundaries. Some examples of improvements in this regard were illustrated in Fig. 6.

Our results also showed the importance of sampling and class balancing in the training process. By feeding the network with only the most difficult negative samples, we ensured that useful samples were used in the training process. When compared to the rest of CNN approaches, our method without restricted sampling yielded a similar performance to other methods such as the one of Shakeri et al. (2016) and MS-CNN (Bao and Chung, 2016) even if trained on the same conditions, which highlights the effectiveness of the used sampling strategy. As a counterpart, these kind of approaches tend to generate false positive regions outside the sub-cortical space, due to the lack of contextual spatial information of the whole brain. Within our proposal, we took advantage of the already computed spatial priors to reduce the segmentation to only a region of interest containing the sub-cortical structures, which reduced remarkably the inference time. Remaining false positive voxels were then post-processed by maintaining only the biggest region for each class.

Our study comprises some limitations. Although our analysis shows that incorporating a-priori atlas information is effective on segmentation of the sub-cortical structures, there is room for further analysis of this approach in other brain segmentation tasks. Furthermore, the addition of atlas probabilities requires nonlinear registration, which may be tedious and prone to errors if applied on extreme cases such as advanced pathological subjects with a high degree of atrophy. Additionally, the extrapolation of our sample selection technique to other more general brain segmentation tasks should also be studied. As part of supervised training strategies, the accuracy of CNN methods tend to decrease significantly in other image domains (i.e. different MRI scanner, image protocol, etc.) than the ones used for training. Nevertheless, there is still a little evidence of the capability of CNN methods in radiological tasks with small or none datasets, which highlights the need of further studying this issue to increase the accuracy of such approaches. With no more evidence in this field, FIRST may be more appropriate in these scenarios when few or no training data is available. Another constraint involves the applicability of the proposed method on datasets of images with neurological diseases comprising, for instance, white matter lesions, which affect brain structure segmentation (González-Villà et al., 2017).

5. Conclusion

In this paper, we have presented a novel CNN based deep learning approach for accurate and robust segmentation of the sub-cortical brain structures that combines both convolutional and prior spatial features for improving the segmentation accuracy. In order to increase the accuracy of the classifier, we have proposed to train the network using a restricted sample selection to force the network to learn the most difficult parts of the structures. As seen from all the experiments carried out on the public MIC-CAI 2012 and IBSR 18 datasets, the addition of the spatial priors and the restricted sampling strategy have a significant impact on the effectiveness of the proposed method, outperforming or showing a comparable performance to state-of-the-art methods such as FreeSurfer, FIRST and different recently proposed CNN approaches.

In order to encourage the reproducibility and the use of the proposed method, a public version is available to download for the neuroimaging community at our research website.

Acknowledgments

Kaisar Kushibar and Jose Bernal hold FI-DGR2017 grant from the Catalan Government with reference numbers 2017FL_B00372 and 2017FL_B00476, respectively. This work has been partially supported by La Fundació la Marató de TV3, by Retos de Investigación TIN2014-55710-R, TIN2015-73563-JIN, and DPI2017-86696-R from the Ministerio de Ciencia y Tecnología. The authors gratefully acknowledge the support of the NVIDIA Corporation with their donation of the TITAN-X PASCAL GPU used in this research.

References

- Altschuler, L.L., Bartzokis, G., Grieder, T., Curran, J., Mintz, J., 1998. Amygdala enlargement in bipolar disorder and hippocampal reduction in schizophrenia: an MRI study demonstrating neuroanatomic specificity. *Arch. Gen. Psychiatry* 55, 663–664.
- Bao, S., Chung, A.C., 2016. Multi-scale structured CNN with label consistency for brain MR image segmentation. *Comput. Methods Biomech. Biomed. Eng.* 1–5.
- Bergstra, J., Breuleux, O., Lamblin, P., Pascanu, R., Delalleau, O., Desjardins, G., Goodfellow, I., Bergeron, A., Bengio, Y., Kaelbling, P., 2011. Theano: deep learning on GPUs with python. *J. Mach. Learn. Res.* 1, 1–48.
- Brébisson, A., Montana, G., 2015. Deep neural networks for anatomical brain segmentation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 20–28.
- Caviness V. S., Jr, Meyer, J., Makris, N., Kennedy, D.N., 1996. MRI-based topographic parcellation of human neocortex: an anatomically specified method with estimate of reliability. *J. Cognit. Neurosci.* 8, 566–587.
- Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A. L., 2016. Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. [arXiv:1606.00915](https://arxiv.org/abs/1606.00915).
- Cootes, T.F., Edwards, G.J., Taylor, C.J., 2001. Active appearance models. *IEEE Trans. Pattern Anal. Mach. Intell.* 23, 681–685.
- Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J., 1995. Active shape models—their training and application. *Comput. Vision Image Understanding* 61, 38–59.
- Debernard, L., Melzer, T.R., Alla, S., Eagle, J., Van Stockum, S., Graham, C., Osborne, J.R., Dalrymple-Alford, J.C., Miller, D.H., Mason, D.F., 2015. Deep grey matter MRI abnormalities and cognitive function in relapsing-remitting multiple sclerosis. *Psychiatry Res.* 234, 352–361.
- Dice, L.R., 1945. Measures of the amount of ecologic association between species. *Ecology* 26, 297–302.
- Dolz, J., Desrosiers, C., Ayed, I.B., 2018. 3D fully convolutional networks for subcortical segmentation in MRI: a large-scale study. *Neuroimage* 170, 456–470.
- Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S., 2017. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542, 115–118.
- Fischl, B., 2012. FreeSurfer. *Neuroimage* 62, 774–781.
- Fischl, B., Salat, D.H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., Van Der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., et al., 2002. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron* 33, 341–355.
- Fox, N., Warrington, E., Freeborough, P., Hartikainen, P., Kennedy, A., Stevens, J., Rossor, M.N., 1996. Presymptomatic hippocampal atrophy in alzheimer's disease: a longitudinal MRI study. *Brain* 119, 2001–2007.
- Ghafoorian, M., Karssemeijer, N., Heskes, T., van Uden, I.W., Sanchez, C.I., Litjens, G., de Leeuw, F.E., van Ginneken, B., Marchiori, E., Platel, B., 2017. Location sensitive deep convolutional neural networks for segmentation of white matter hyperintensities. *Sci. Rep.* 7, 5110.
- Girshick, R., Donahue, J., Darrell, T., Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 580–587.
- Glorot, X., Bordes, A., Bengio, Y., 2011. Deep sparse rectifier neural networks. In: *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 315–323.
- González-Villà, S., Oliver, A., Valverde, S., Wang, L., Zwiggelaar, R., Lladó, X., 2016. A review on brain structures segmentation in magnetic resonance imaging. *Artif. Intell. Med.* 73, 45–69.
- González-Villà, S., Valverde, S., Cabezas, M., Pareto, D., Vilanova, J.C., Ramió-Torrentà, L., Rovira, A., Oliver, A., Lladó, X., 2017. Evaluating the effect of multiple sclerosis lesions on automatic brain structure segmentation. *Neuroimage* 15, 228–238.
- Hazlett, H.C., Gu, H., Munsell, B.C., Kim, S.H., Styner, M., Wolff, J.J., Elison, J.T., Swanson, M.R., Zhu, H., Botteron, K.N., et al., 2017. Early brain development in infants at high risk for autism spectrum disorder. *Nature* 542, 348–351.
- He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778.

- Kikinis, R., Shenton, M.E., Iosifescu, D.V., McCarter, R.W., Saiviroonporn, P., Hokama, H.H., Robatino, A., Metcalf, D., Wible, C.G., Portas, C.M., et al., 1996. A digital brain atlas for surgical planning, model-driven segmentation, and teaching. *IEEE Trans. Visual. Comput.Graph.* 2, 232–241.
- Kingma, D. P., Ba, J., 2014. Adam: a method for stochastic optimization. arXiv:1412.6980.
- Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. In: *Advances in Neural Information Processing Systems*, pp. 1097–1105.
- Lafferty, J., McCallum, A., Pereira, F., et al., 2001. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: *Proceedings of the Eighteenth International Conference on Machine Learning*, vol. 1, pp. 282–289.
- Landman, B., Warfield, S., 2012. MICCAI 2012 workshop on multi-atlas labeling. In: *Medical Image Computing and Computer Assisted Intervention Conference*.
- Lawrie, S.M., Whalley, H.C., Job, D.E., Johnstone, E.C., 2003. Structural and functional abnormalities of the amygdala in schizophrenia. *Ann. New York Acad. Sci.* 985, 445–460.
- LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 2278–2324.
- Liu, Y., Gadepalli, K., Norouzi, M., Dahl, G. E., Kohlberger, T., Boyko, A., Venugopalan, S., Timofeev, A., Nelson, P. Q., Corrado, G. S., HIPP, J. D., Peng, L., Stumpe, M. C., 2017. Detecting cancer metastases on gigapixel pathology images. arXiv:1703.02442.
- Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3431–3440.
- Mak, E., Bergsland, N., Dwyer, M., Zivadinov, R., Kandiah, N., 2014. Subcortical atrophy is associated with cognitive impairment in mild parkinson disease: a combined investigation of volumetric changes, cortical thickness, and vertex-based shape analysis. *Am. J. Neuroradiol.* 35, 2257–2264.
- Mehta, R., Majumdar, A., Sivaswamy, J., 2017. BrainSegNet: a convolutional neural network architecture for automated segmentation of human brain structures. *J. Med. Imaging* 4, 024003–024003.
- Milham, M.P., Nugent, A.C., Drevets, W.C., Dickstein, D.S., Leibenluft, E., Ernst, M., Charney, D., Pine, D.S., 2005. Selective reduction in amygdala volume in pediatric anxiety disorders: a voxel-based morphometry investigation. *Biol. Psychiatry* 57, 961–966.
- Milletari, F., et al., 2017. Hough-CNN: deep learning for segmentation of deep brain regions in MRI and ultrasound. *Comput. Vision Image Understanding*.
- Modat, M., Ridgway, G.R., Taylor, Z.A., Lehmann, M., Barnes, J., Hawkes, D.J., Fox, N.C., Ourselin, S., 2010. Fast free-form deformation using graphics processing units. *Comput. Methods Programs Biomed.* 98, 278–284.
- Ourselin, S., Roche, A., Prima, S., Ayache, N., 2000. Block matching: a general framework to improve robustness of rigid registration of medical images. In: *MICCAI*, volume 1935. Springer, pp. 557–566.
- Patenaude, B., Smith, S.M., Kennedy, D.N., Jenkinson, M., 2011. A Bayesian model of shape and appearance for subcortical brain segmentation. *Neuroimage* 56, 907–922.
- Phillips, J.L., Batten, L.A., Tremblay, P., Aldosary, F., Blier, P., 2015. A prospective, longitudinal study of the effect of remission on cortical thickness and hippocampal volume in patients with treatment-resistant depression. *Int. J. Neuropsychopharmacol.* 18, pyv037.
- Pitiot, A., Delingette, H., Thompson, P.M., Ayache, N., 2004. Expert knowledge-guided segmentation system for brain MRI. *Neuroimage* 23, S85–S96.
- Sarraf, S., Tofghi, G., et al., 2016. DeepAD: Alzheimers Disease Classification via Deep Convolutional Neural Networks Using MRI and fMRI. bioRxiv, p. 070441.
- Shakeri, M., Tsogkas, S., Ferrante, E., Lippe, S., Kadoury, S., Paragios, N., Kokkinos, I., 2016. Sub-cortical brain structure segmentation using F-CNN's. In: *Biomedical Imaging (ISBI)*, 2016 IEEE 13th International Symposium on. IEEE, pp. 269–272.
- Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv:arXiv:1409.1556.
- Smith, S.M., 2002. Fast robust automated brain extraction. *Hum. Brain Mapp.* 17, 143–155.
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–9.
- Tustison, N.J., Avants, B.B., Cook, P.A., Zheng, Y., Egan, A., Yushkevich, P.A., Gee, J.C., 2010. N4ITK: improved n3 bias correction. *IEEE Trans. Med. Imaging* 29, 1310–1320.
- Wachinger, C., Reuter, M., Klein, T., 2018. DeepNAT: deep convolutional neural network for segmenting neuroanatomy. *Neuroimage* 170, 434–445.
- Wang, H., Yushkevich, P.A., 2013. Groupwise segmentation with multi-atlas joint label fusion. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 711–718.


Chapter 3

Supervised domain adaptation for automatic sub-cortical brain structure segmentation with minimal user interaction

Deep learning based methods require a considerable amount of labelled training data to perform well on the designated task. When segmenting medical images, a CNN trained with one set of images does not perform well on another set of scans that have different characteristics such as acquisition protocol or scanner type. This issue is known as the ‘domain-shift problem’ and in this chapter, we present our research on supervised domain adaptation using a transfer learning strategy. Using our method, we have achieved similar state-of-the-art results obtained with a full training when using only the half of the training images. Moreover, it was possible to significantly outperform the well-known traditional tools using only one or just a few training images. This proposal has been published in the following paper:




<p>Paper published in the Nature: Scientific Reports (Sci.Rep.) OPEN ACCESS Volume: 9, Number: 1, Pages: 1–15, Published: May 2019 DOI: 10.1038/s41598-019-43299-z JCR MS IF 4.011, Q1(15/69)</p>

SCIENTIFIC REPORTS



OPEN

Supervised Domain Adaptation for Automatic Sub-cortical Brain Structure Segmentation with Minimal User Interaction

Kaisar Kushibar , Sergi Valverde, Sandra González-Villà , Jose Bernal , Mariano Cabezas, Arnau Oliver & Xavier Lladó

In recent years, some convolutional neural networks (CNNs) have been proposed to segment sub-cortical brain structures from magnetic resonance images (MRIs). Although these methods provide accurate segmentation, there is a reproducibility issue regarding segmenting MRI volumes from different image domains – e.g., differences in protocol, scanner, and intensity profile. Thus, the network must be retrained from scratch to perform similarly in different imaging domains, limiting the applicability of such methods in clinical settings. In this paper, we employ the transfer learning strategy to solve the domain shift problem. We reduced the number of training images by leveraging the knowledge obtained by a pretrained network, and improved the training speed by reducing the number of trainable parameters of the CNN. We tested our method on two publicly available datasets – MICCAI 2012 and IBSR – and compared them with a commonly used approach: FIRST. Our method showed similar results to those obtained by a fully trained CNN, and our method used a remarkably smaller number of images from the target domain. Moreover, training the network with only one image from MICCAI 2012 and three images from IBSR datasets was sufficient to significantly outperform FIRST with ($p < 0.001$) and ($p < 0.05$), respectively.

Structural and morphological changes in brain structures are often associated with different neurodegenerative disorders such as bipolar disorder¹, Alzheimer's², schizophrenia³, Parkinson's disease⁴, and multiple sclerosis⁵. Many of these neurological abnormalities are usually diagnosed with careful analysis of the structural, T1-weighted (T1-w) magnetic resonance images (MRIs). Analysis of the sub-cortical structures – located beneath the cerebral cortex and including thalamus, caudate, putamen, pallidum, hippocampus, amygdala, and accumbens – is very important. Their deviations in volume over time are considered as biomarkers of the aforementioned diseases and are used for pre-operative evaluation and surgical planning⁶, longitudinal monitoring for disease progression or remission^{7,8}.

Providing an accurate automated segmentation for the sub-cortical structures is very important because manually labelling an MRI volume is a time-consuming and tedious task⁹. Well-known, commonly used tools such as FIRST¹⁰ and FreeSurfer¹¹ are available unsupervised methods. However, the advancement in computational technologies, such as graphics processing units (GPUs), has brought a new way to tackle the problem of image classification and segmentation using deep learning techniques, particularly, convolutional neural networks (CNNs). These approaches showed better results in many computer vision tasks such as image classification¹², object recognition¹³ and segmentation¹⁴, than the traditional unsupervised methods that leverage hand-crafted features because the CNN features are learned directly from the training images¹⁵.

In recent years, deep learning has become a popular approach in medical imaging and brain MRI analysis^{16,17}. Some methods based on deep learning strategies have also been proposed for brain structure segmentation^{18–20}. The results of these approaches were promising; however, these methods were trained and tested in the same image domain – i.e., the same protocol, same MRI scanner, resolution and image contrast – and their behaviour with image domain change has not been evaluated. This is a common issue in supervised approaches, where a

Institute of Computer Vision and Robotics, University of Girona. Ed. P-IV, Campus Montilivi, University of Girona, 17003, Girona, Spain. Correspondence and requests for materials should be addressed to K.K. (email: kaisar.kushibar@udg.edu)

change in image domain causes an unexpected outcome because such methods learn a data distribution solely from the training set, making the supervised model less generalisable. Moreover, obtaining a well-trained deep CNN model requires a vast amount of training data with ground truth, which currently is a major issue in the medical image analysis field. Therefore, the traditional unsupervised methods – FIRST and FreeSurfer – are still the preferred tools of choice.

Few studies have proposed different methods to overcome the domain shift difficulty in medical images. These methods are often referred to as domain adaptation methods that tackle the problem of domain shift in medical images and address the broader statistical issue of out-of-sample generalisation in deep learning. One of the recent proposals includes domain adaptation using adversarial networks²¹. In this approach, a network contains an additional domain discriminator branch, which penalises the network when the features extracted from two different domains are distinct. In doing so, the network is forced to learn more domain invariant features. The loss computation with the discriminator does not require ground truth segmentation; therefore, adversarial domain adaptation could be carried out in an unsupervised manner for the target dataset. However, such approaches require a subtle parameter tuning and training of the network from scratch for different target domains. Another way to address the domain shift problem is via transfer learning, where the weights of an already trained network are fine tuned to adapt to a new target domain. This is inspired by the early convolutional layers capturing similar low-level features such as edges, curves and blobs. Accordingly, by performing an additional training on a smaller target dataset, it is possible to fine tune only some of the deep layers of the network that represent higher level features. Recent studies²² have shown that transfer learning and fine tuning decreases the training time drastically while demanding fewer training samples than that of full training. Additionally, the behaviour of the transfer learning strategy with different set of parameters has been recently analysed, indicating the effectiveness of this approach over full training for brain white matter hyperintensity segmentation²³. In their work, the authors²³ investigated two datasets containing fluid-attenuated inversion recovery (FLAIR) and magnetisation-prepared rapid gradient-echo (MPRAGE). The images of the two datasets were acquired with the same scanner and protocol, except for FLAIR images that had different image resolutions.

In this paper, we investigated the transfer learning and fine tuning strategies for domain adaptation on MRI volumes acquired with different scanners and protocols to segment sub-cortical brain structures. In our experiments, we employed a state-of-the-art deep learning based method that combines spatial and deep convolutional features for sub-cortical structure segmentation¹⁹. Within our study, we demonstrated the effect of domain shift on the neural network's performance and analysed an adequate number of MRI volumes to adapt the CNN to a new domain and outperform traditional unsupervised methods. Because the sub-cortical structures drastically varied in their volumes, structure-wise performance after domain adaptation with different number of training images was also evaluated. Additionally, we performed an experiment to show the applicability of this study in real-case scenarios by accelerating the initial manual segmentation. Also, the training and testing time complexities were evaluated to examine how transfer learning could speed up the segmentation process compared with a fully trained neural network. Moreover, to encourage the reproducibility of our results, we made the source code used in training, transfer learning, Dice similarity coefficient (DSC), and statistical test calculations publicly available. Additionally, the manually corrected segmentation masks used in our experiments and label mappings for the MICCAI 2012 and IBSR datasets were made available for the community at https://github.com/NIC-VICOROB/sub-cortical_segmentation.

Materials

Datasets. In this work, we used two well-known, publicly available datasets – Internet Brain Segmentation Repository (IBSR) and MICCAI Multi-Atlas Labelling challenge (MICCAI 2012)²⁴. More details on these datasets and their domain differences are provided in the following sections.

MICCAI 2012 dataset. The MICCAI 2012 dataset contains 35 images in total, which are split into 15 training and 20 testing image volumes according to the Multi-Atlas Labelling challenge. The 20 testing images were always used only for testing purposes and were never included in the training or validation processes. All images have a $1 \times 1 \times 1 \text{ mm}^3$ resolution and image size of $256 \times 256 \times 256$. Additionally, all image volumes in this dataset were acquired using the same MRI scanner – SIEMENS (1.5T). They were provided with manually annotated ground truth masks for 134 structures. We extracted 14 classes corresponding to the seven sub-cortical structures with left and right parts each.

IBSR. The IBSR dataset consists of 18 images with an image size of $256 \times 256 \times 128$ and three different resolutions: $0.84 \times 0.84 \times 1.5 \text{ mm}^3$, $0.94 \times 0.94 \times 1.5 \text{ mm}^3$ and $1 \times 1 \times 1.5 \text{ mm}^3$. The subject volumes of the IBSR dataset were obtained using two different MRI scanners: GE (1.5T) and SIEMENS (1.5T). Manually segmented ground truths for 43 different structures are provided²⁵, and we extracted the 14 labels corresponding to the sub-cortical structures for our experiments. The MR brain images in this dataset and their manual segmentations were provided by the Center for Morphometric Analysis at Massachusetts General Hospital. Additionally, this dataset is a part of the Child and Adolescent NeuroDevelopment Initiative²⁶ (CANDI) and was provided under the Creative Commons: Attribute license²⁷.

Domain comparison. Proper selection of the datasets for the transfer learning experiments is crucial because the domain difference should be present to confirm the method's robustness. As seen above in the selected datasets' details, they differ in resolution and in MRI scanner type. The intensity distribution only in the brain area (i.e., skull-stripped) of these datasets also varies in terms of their profile (Fig. 1). The maximum intensity in the MICCAI 2012 volume reaches up to ≈ 2000 and that for the IBSR image is ≈ 140 . This behaviour in intensity distribution is observed among the subjects in both datasets. Because the contrast and the intensity values of the

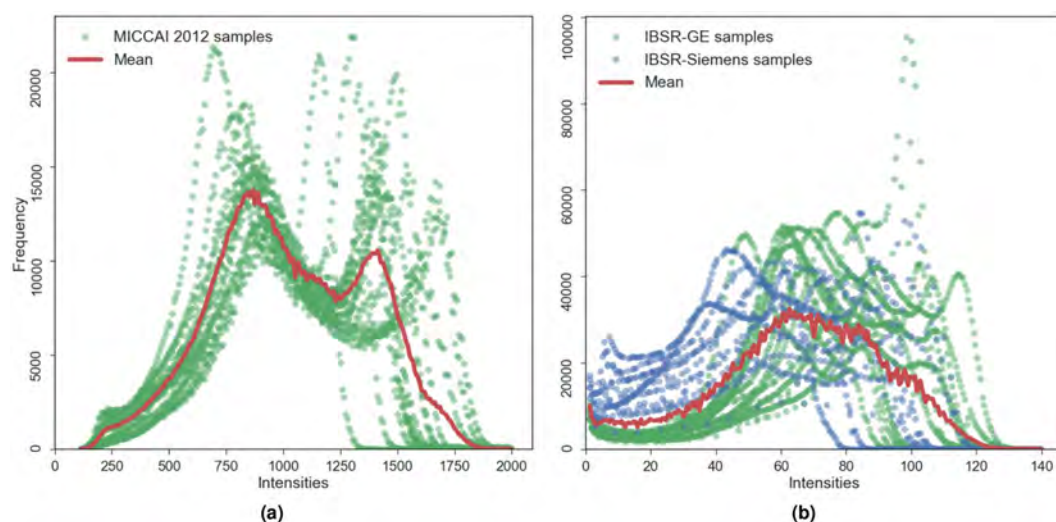


Figure 1. Intensity distributions in the brain area of the MRI volumes of MICCAI 2012 (a) and IBSR (b) datasets. The mean histogram is shown in solid red, and the intensity distributions of all images are shown in green for MICCAI 2012 and IBSR-GE datasets, and blue for IBSR-Siemens.

structures of interest are represented from different distributions, pre-processing techniques such as contrast enhancement and histogram equalisation applied for each image individually cannot compensate for the imaging protocol differences between the datasets (e.g., image resolution). However, MRI volume standardisation across datasets using histogram matching could be an interesting line of research and is also analysed in this paper. Considering these variations, the datasets of interest perfectly fit the challenge of the domain shift problem.

Moreover, there is within-group variability of the intensity distributions of the subject volumes in the IBSR dataset among the three different image resolution groups due to the different MRI machines and various magnitudes of the partial volume effect. As shown in Fig. 1, the individual intensity distributions of the images in IBSR dataset also vary drastically, whereas the MRI volumes of MICCAI 2012 follow a similar profile. This intensity distribution variability in IBSR images makes the domain adaptation more challenging.

Methods

CNN architecture. The CNN architecture used in our experiments is shown in Fig. 2 and consists of three paths to process 2D patches of size 32×32 . Each path is equipped with five convolutional layers, which are followed by a fully connected layer. The outputs of these paths are concatenated together with an additional 15 units corresponding to atlas probabilities. Finally, two fully connected layers are used to mine and classify the produced output by the preceding layers. Three 2D patches are extracted for every voxel from the axial, sagittal and coronal views of a 3D volume, making 2.5D patch samples. Next, each orthogonal 2D patch of the 2.5D sample is inputted into the three paths of the CNN. Although full 3D patches contain more surrounding information for a voxel, it is more memory demanding than using 2D patches. Therefore, employing 2.5D patches is a good trade-off between memory and contextual information for the network.

Network training. To train the network, all samples were extracted from the 14 sub-cortical structures, and the background (negative) samples were selected only from the structure boundaries, which were obtained by dilating the ground truth by five voxels. Extracting the negative samples in this way allows the network to learn the most difficult areas of the region of interest that correspond to the structure borders¹⁹. Next, the atlas probabilities for 14 structures and the background are extracted, corresponding to all training samples and making a vector of size 15. These probabilities provide the network with spatial information and guide it to overcome intensity-based difficulties in some MRI volumes such as imaging artefacts and small tissue changes¹⁹. All the extracted samples were randomly split into training and validation sets with 75% and 25% proportions, respectively.

Once the training samples were extracted along with their atlas probabilities, the training of the network was performed in batches of 128 for 200 epochs. An early stopping policy was defined with patience 20 – i.e., the training stops if no increase was observed in the validation accuracy for 20 consecutive epochs. Optimisation was conducted for the categorical cross-entropy loss function using the Adam²⁸ optimisation method with an initial learning rate of 10^{-2} .

Transfer learning. The transfer learning and fine tuning procedures were performed as follows. First, the network is fully trained from scratch as described above with one dataset, referred to as the source. Next, all the convolutional layers were frozen, and the weights of the last classification layer were reset. Accordingly, when the network is trained with the images from another dataset, referred to as a target, the weights of the convolutional layers are not updated. Additionally, the fully connected layers of the network were fine tuned – i.e., the weights were adjusted to better fit the new domain. We trained the softmax layer from scratch because it was used to

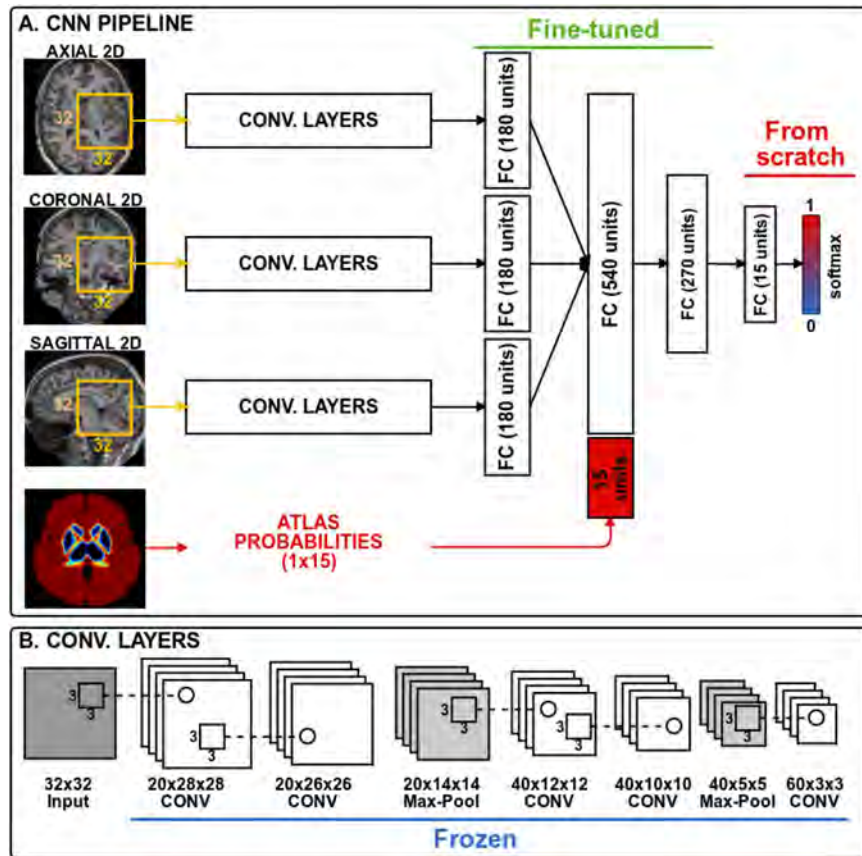


Figure 2. CNN architecture. The weights of all convolutional layers are frozen during transfer learning. The fully connected layers are fine-tuned and the last classification layer is trained from scratch.

classify the extracted features. Note that the initial learning rate was reduced to 10^{-4} during transfer learning to avoid rapid weight updates because most of the trainable parameters in the network were frozen.

Network testing. To test a trained model, all 2.5D patches and corresponding atlas vectors were extracted from an MRI volume. Because the sub-cortical structures are located in the central part of the brain, the patches were obtained from a region of interest (ROI) defined by a mask from the dilated atlas probabilities. This helps to increase the processing speed and avoid false positives around the sub-cortical region. The network was well trained to classify the background only around the structures; therefore, some misclassified voxels may appear under the ROI. Those voxels were removed by keeping only the largest volume for each class.

Image pre-processing. Before extracting patch samples from the MRI volumes to train the network, we performed some commonly used pre-processing steps:

- *Brain extraction* – i.e., removing non-brain structures, such as the eyes and skull, from an MRI volume. Compared with the previous study on sub-cortical structure segmentation¹⁹, we used the ROBEX (v1.2) tool²⁹ instead of BET (fsl-v5.0)³⁰. This is due to the robustness of the former method over the latter. Additionally, ROBEX does not require any parameter tuning compared with BET.
- *Atlas registration* – we performed non-linear registration of a template MRI with a probabilistic atlas to all images in the selected datasets. The probabilistic atlases were used in the network as explicit spatial information, which helps to improve the segmentation accuracy¹⁹. In this study, we used the well-known Harvard-Oxford probabilistic atlas³¹ distributed with the FSL tool (<http://www.fmrib.ox.ac.uk/fsl>). The non-linear registration of the atlas template to the subject volume was applied using the fast free-form deformation method³² that was implemented in the NiftyReg tool (<http://cmictig.cs.ucl.ac.uk/wiki/index.php/NiftyReg>).
- *Intensity normalisation* – as the maximum intensity values from both of the datasets differ drastically, all the subject volume intensities are normalised to have a zero mean and unit variance before training and testing the pipeline.

Technical details. The network was implemented using the Keras³³ deep learning library. The DSC scores were obtained using the Nipype³⁴ data processing framework. The statistical tests were performed using the SciPy python package³⁵. The Nibabel³⁶ python package was used to read and write the medical imaging files.

Experiments and evaluation. We trained our network with one dataset as the source and applied transfer learning with the other set as a target. Next, we repeated the same experiment but changed the datasets in the opposite order to show the method's robustness. The target training MRI volumes, randomly chosen in previous iterations, are kept in the next iteration of transfer learning to have an unbiased estimate on changes in the results. For the sake of brevity, we compared our results with those of FIRST. According to our previous study¹⁹, FIRST showed better results than FreeSurfer to segment the sub-cortical structures in both selected datasets, and comparison of our method only with the former method is sufficient. The following experiments were carried out using the two selected datasets to evaluate the performance of domain adaptation using transfer learning.

- *From IBSR to MICCAI 2012* – All images from IBSR were used as the source, and domain adaptation was performed for the target MICCAI 2012 dataset. The results for the MICCAI 2012 dataset are shown for the 20 testing cases, and images for transfer learning were randomly selected from the 15 training cases.
- *From MICCAI 2012 to IBSR* – All images from MICCAI 2012 were used as the source, and the network was adapted to the target IBSR images. Because the IBSR dataset does not have training and testing splits, for each iteration of transfer learning, we randomly selected the corresponding number of MRI volumes and repeated the iteration with different images – e.g., a single iteration using one training image takes two steps: (1) training once with one image; (2) then training again with a different image to obtain an overall score for all 18 cases.
- *Grouping by MRI scanner* – The images of the IBSR dataset were split into two sets depending on the MRI scanner manufacturer for the IBSR-GE (12 images) and IBSR-Siemens (6 images) groups. The MICCAI 2012 dataset was used as the source, while the two new groups were set as targets. Due to the smaller number of images in the IBSR-Siemens group, transfer learning was performed for three iterations and the source dataset remained the same.
- *Corrected FIRST segmentation as ground truth* – Transfer learning was applied using the manually corrected segmentation outputs from FIRST as the ground truth. This shows how the initial domain adaptation of the network could be accelerated and avoid ground truth preparation from scratch. Additionally, the results of the first two experiments were also compared with two different approaches that could be used to tackle the domain shift problem:
- *Standardised (normalised) images* – The intensities of all images in both datasets were standardised to the mean histogram of the MICCAI 2012 using the two-stage method of Nyúl³⁷. This experiment was performed to evaluate how intensity normalisation would affect the network's performance because it has been shown to be effective for three classical image segmentation algorithms³⁸.
- *Mixed datasets* – The network was trained from scratch using a mixed dataset containing normalised images of all subjects from source and iteratively added target volumes. This is to compare transfer learning with the performance of the network when there is more variability in the training data distribution.

Furthermore, the execution times regarding the neural network's training and testing were evaluated to show how transfer learning can accelerate the convergence of the CNN.

We used the Dice similarity coefficient (DSC) to quantitatively analyse the results of our experiments. This metric evaluates the overlap of the automatic segmentation mask over the manually segmented ground truth. The DSC value varies between 0 (non-overlap) and 1 (full overlap). Because the larger structures contribute to the average DSC more than the smaller ones, penalised DSC by the inverse of the volume structure³⁹ could be used to evaluate the performance of the method. In doing so, different segmentation approaches can be compared without volume bias. Therefore, we also used the weighted DSC to compare and verify the results of different experimental setups for consistency. Additionally, in our results, we showed the DSC values for each structure independently to demonstrate the evolution of the method's performance along the iterations of transfer learning. Moreover, we used the pairwise non-parametric Wilcoxon signed-rank test (two-sided) to compare the statistical significance of our results with respect to the state-of-the-art tools. The results were considered significant for ($p < 0.05$).

Results

From IBSR to MICCAI 2012. In this section, the results for MICCAI 2012 are shown when the IBSR dataset was used as the source. Initially, the network was fully trained with all the images from IBSR, and transfer learning iterations were performed by incrementing the target training set's size at a time. Table 1 summarises the DSC scores for FIRST, the results from full training, and the results when using transfer learning for the MICCAI 2012 dataset. Fine tuning the network with only one image drastically increased the performance of the network, significantly outperforming FIRST ($p < 0.001$). Incrementally adding an image volume to the target training set gradually improved the overall DSC score from 0.834 up to 0.860 ($p < 0.001$). Additionally, it was possible to obtain a result similar to the fully trained network using only half of the training images. The domain shift effect on the trained network could be clearly seen from the results where no transfer learning was applied: an extreme performance drop compared with the fully trained network was observed in both, overall and structure-wise scores. These low scores were caused by confounding segmentation outputs of the network, where left and right parts for some structures were swapped.

Improvement in the DSC values for each structure when increasing the number of target training images can be seen in Fig. 3(a). The highest DSC scores were achieved for the large structures such as the thalamus and putamen. Additionally, the difference in DSC when using one or seven images for training was not high, indicating that one image was adequate to obtain accurate segmentation for these structures. Interestingly, substantial improvement could be achieved for the smallest structures such as the amygdala and accumbens when the number of training images for transfer learning was increased.

Str.	FIRST	FT	No TL	TL 1	TL 2	TL 3	TL 4	TL 5	TL 6	TL 7
Tha.L	0.889 ± 0.017	0.920 ± 0.017	0.529 ± 0.251	0.905 ± 0.014	0.904 ± 0.017	0.908 ± 0.017	0.909 ± 0.015	0.910 ± 0.015	0.912 ± 0.015	0.913 ± 0.015
Tha.R	0.890 ± 0.018	0.924 ± 0.016	0.418 ± 0.220	0.903 ± 0.012	0.908 ± 0.013	0.912 ± 0.011	0.914 ± 0.011	0.914 ± 0.012	0.912 ± 0.014	0.917 ± 0.012
Cau.L	0.797 ± 0.117	0.885 ± 0.071	0.694 ± 0.090	0.861 ± 0.066	0.867 ± 0.062	0.874 ± 0.063	0.875 ± 0.063	0.878 ± 0.061	0.880 ± 0.062	0.884 ± 0.060
Cau.R	0.837 ± 0.046	0.887 ± 0.057	0.774 ± 0.050	0.870 ± 0.049	0.874 ± 0.051	0.877 ± 0.050	0.883 ± 0.053	0.881 ± 0.053	0.886 ± 0.052	0.885 ± 0.053
Put.L	0.860 ± 0.080	0.909 ± 0.023	0.884 ± 0.023	0.903 ± 0.023	0.910 ± 0.024	0.911 ± 0.025	0.913 ± 0.023	0.914 ± 0.023	0.914 ± 0.023	0.915 ± 0.024
Put.R	0.876 ± 0.060	0.908 ± 0.031	0.884 ± 0.018	0.906 ± 0.024	0.910 ± 0.023	0.912 ± 0.023	0.913 ± 0.023	0.915 ± 0.024	0.913 ± 0.025	0.915 ± 0.024
Pal.L	0.815 ± 0.060	0.873 ± 0.101	0.374 ± 0.269	0.842 ± 0.032	0.856 ± 0.028	0.861 ± 0.028	0.862 ± 0.024	0.865 ± 0.023	0.866 ± 0.024	0.866 ± 0.024
Pal.R	0.799 ± 0.088	0.874 ± 0.049	0.111 ± 0.181	0.839 ± 0.043	0.850 ± 0.041	0.853 ± 0.043	0.857 ± 0.044	0.856 ± 0.048	0.858 ± 0.050	0.862 ± 0.045
Hip.L	0.809 ± 0.014	0.871 ± 0.020	0.808 ± 0.021	0.825 ± 0.034	0.835 ± 0.033	0.846 ± 0.026	0.849 ± 0.026	0.851 ± 0.024	0.854 ± 0.027	0.856 ± 0.026
Hip.R	0.810 ± 0.022	0.869 ± 0.020	0.822 ± 0.019	0.845 ± 0.019	0.841 ± 0.027	0.854 ± 0.020	0.854 ± 0.019	0.858 ± 0.020	0.861 ± 0.020	0.863 ± 0.020
Amy.L	0.721 ± 0.054	0.832 ± 0.032	0.669 ± 0.043	0.740 ± 0.043	0.777 ± 0.032	0.800 ± 0.031	0.809 ± 0.029	0.810 ± 0.030	0.812 ± 0.029	0.812 ± 0.033
Amy.R	0.707 ± 0.052	0.812 ± 0.027	0.613 ± 0.056	0.739 ± 0.047	0.750 ± 0.046	0.766 ± 0.044	0.774 ± 0.049	0.784 ± 0.040	0.788 ± 0.038	0.789 ± 0.042
Acc.L	0.699 ± 0.081	0.790 ± 0.052	0.693 ± 0.055	0.769 ± 0.050	0.779 ± 0.050	0.784 ± 0.055	0.788 ± 0.049	0.786 ± 0.050	0.789 ± 0.041	0.792 ± 0.046
Acc.R	0.678 ± 0.089	0.783 ± 0.067	0.238 ± 0.203	0.722 ± 0.093	0.729 ± 0.091	0.759 ± 0.091	0.763 ± 0.087	0.771 ± 0.090	0.774 ± 0.084	0.765 ± 0.084
Avg.	0.799 ± 0.094	0.867 ± 0.064	0.608 ± 0.272	0.834 ± 0.077	0.842 ± 0.073	0.851 ± 0.068	0.854 ± 0.066	0.857 ± 0.065	0.859 ± 0.063	0.860 ± 0.064
wAvg.	0.706 ± 0.061	0.803 ± 0.040	0.474 ± 0.083	0.754 ± 0.047	0.766 ± 0.046	0.783 ± 0.048	0.788 ± 0.046	0.791 ± 0.047	0.794 ± 0.042	0.792 ± 0.044

Table 1. From IBSR to MICCAI 2012. Mean ± standard deviation DSC values for FIRST, full training and transfer learning with an incremental number of training images. FT (Full training) – the network is trained from scratch with the MICCAI 2012 dataset. TL X – transfer learning with X number of target volumes. No TL – tested directly on the model trained with the IBSR dataset. The structure acronyms are as follows: left thalamus (Tha.L), right thalamus (Tha.R), left caudate (Cau.L), right caudate (Cau.R), left putamen (Put.L), right putamen (Put.R), left pallidum (Pal.L), right pallidum (Pal.R), left hippocampus (Hip.L), right hippocampus (Hip.R), left amygdala (Amy.L), right amygdala (Amy.R), left accumbens (Acc.L), right accumbens (Acc.R), average value (Avg.) and weighted average DSC (wAvg.).

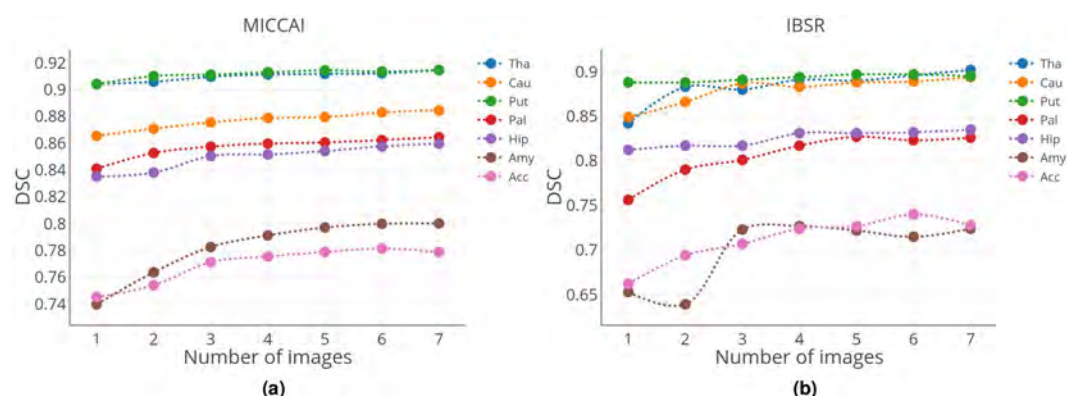


Figure 3. Change in the average DSC scores per structure with the increasing number of training images (left and right parts of all structures are averaged) for (a) MICCAI 2012 and (b) IBSR datasets as targets. Structure acronyms are thalamus (Tha), caudate (Cau), putamen (Put), pallidum (Pal), hippocampus (Hip), amygdala (Amy), and accumbens (Acc).

Moreover, we observed that transfer learning actually helped to leverage previously acquired knowledge from the source dataset. Figure 4(a) illustrates the average DSC results for seven iterations of transfer learning with original images, training from scratch, transfer learning with standardised images, mixed set training, and the results of FIRST, full training, and the results of testing standardised images without transfer learning. The network performed worse in terms of the overall average DSC when trained from scratch with the same number of training images as that of transfer learning. Intensity normalisation definitely helped to improve the performance of the network when it was trained with the source and directly tested on the target without adapting the network to the new domain. As shown in Fig. 5(a), the standardised image segmentation was better, improving the average DSC from 0.608 to 0.787 for the MICCAI 2012 dataset ($p < 0.001$). However, no substantial improvement was observed when the standardised images were used for transfer learning. Additionally, using seven target training volumes was significantly less using the standardised images, with a DSC of 0.842 compared with 0.860 using the original images ($p < 0.001$). Additionally, transfer learning showed a better performance than that of the mixed dataset results. As illustrated in Fig. 4(a), the DSC values obtained by transfer learning was significantly higher for all iterations ($p < 0.001$). The results in the first three iterations were similar, slightly increasing from 0.806 to

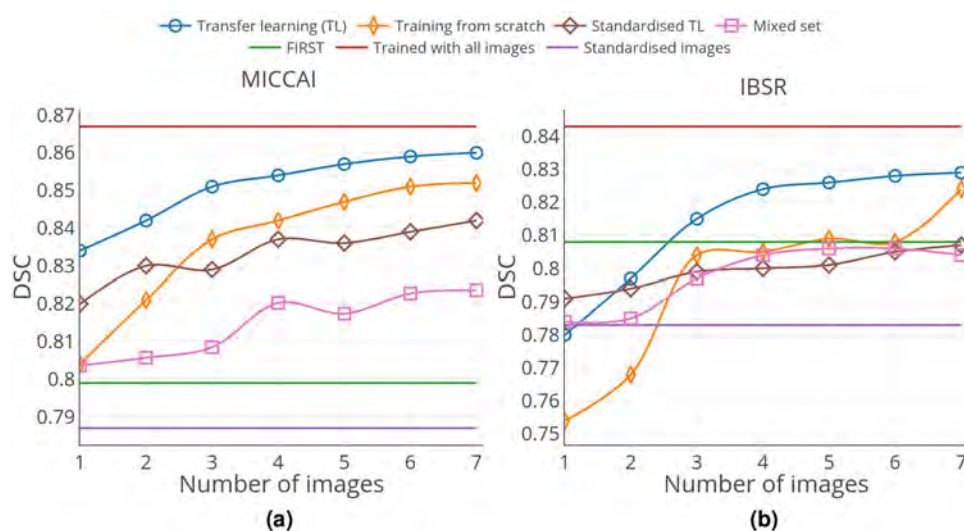


Figure 4. Overall average DSC results for (a) MICCAI 2012 and (b) IBSR datasets. The results are shown for seven iterations of transfer learning with original images, training from scratch, transfer learning with standardised images, and mixed set training. The horizontal lines correspond to the results of FIRST, full training, and the results of testing standardised images without transfer learning. The training volumes in each iteration for all cases are the same.

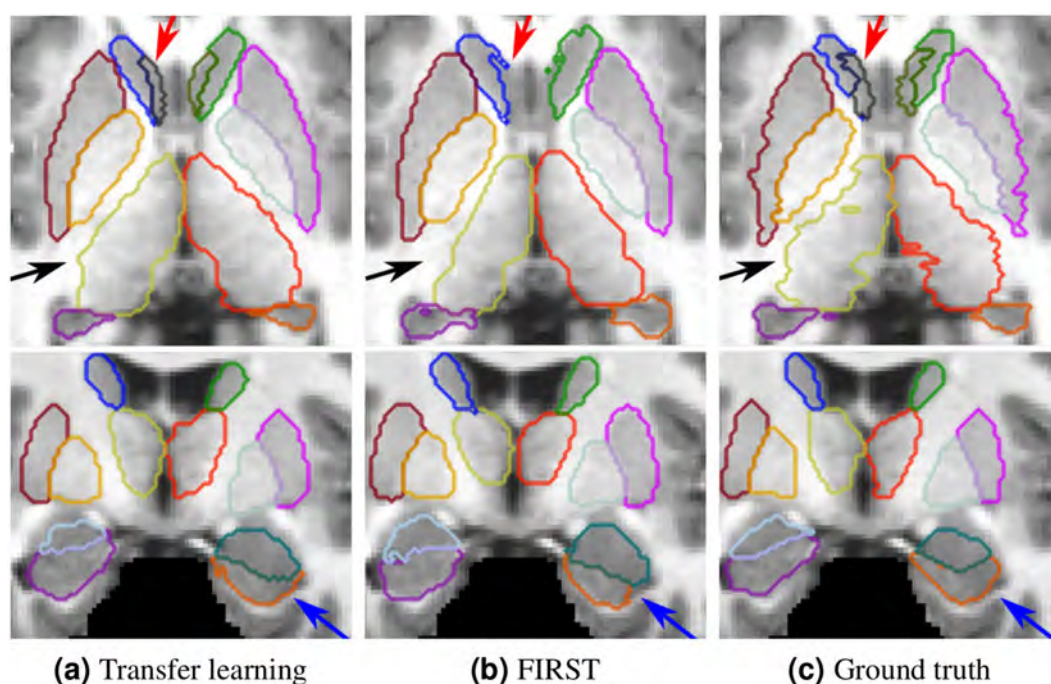


Figure 5. Qualitative results for the MICCAI 2012 dataset. (a) Transfer learning with one image; (b) segmentation result from FIRST; (c) ground truth. The results in the top row shown in axial and bottom ones shown in coronal views. The arrows indicate the following structures: red → accumbens, black → thalamus, blue → hippocampus.

0.808, and a significant increase starting from the fourth iteration, reaching 0.820 ($p < 0.001$). However, the performance did not improve any further. The average DSC values for each structure for the image standardisation and mixed set experiments are included in Supplementary Tables S1 and S3.

Some qualitative results for the MICCAI 2012 dataset are shown in Fig. 5, where transfer learning with a single image produced similar segmentation outputs to the ground truth. FIRST failed to properly segment the smallest structure, the accumbens (pointed with red arrow). Additionally, our method produced better segmentation for the thalamus structure (pointed with black arrow). Better segmentation using our method can also be observed

Str.	FIRST	LOO	No TL	TL 1	TL 2	TL 3	TL 4	TL 5	TL 6	TL 7
Tha.L	0.893 ± 0.017	0.910 ± 0.014	0.128 ± 0.226	0.847 ± 0.068	0.887 ± 0.017	0.882 ± 0.026	0.889 ± 0.024	0.890 ± 0.025	0.893 ± 0.016	0.898 ± 0.014
Tha.R	0.885 ± 0.012	0.914 ± 0.016	0.081 ± 0.173	0.837 ± 0.101	0.879 ± 0.031	0.877 ± 0.038	0.892 ± 0.030	0.890 ± 0.029	0.899 ± 0.015	0.906 ± 0.012
Cau.L	0.783 ± 0.044	0.896 ± 0.018	0.440 ± 0.290	0.857 ± 0.030	0.872 ± 0.031	0.890 ± 0.021	0.883 ± 0.022	0.887 ± 0.023	0.890 ± 0.025	0.894 ± 0.018
Cau.R	0.870 ± 0.027	0.896 ± 0.020	0.455 ± 0.306	0.840 ± 0.040	0.861 ± 0.034	0.883 ± 0.020	0.883 ± 0.020	0.889 ± 0.019	0.889 ± 0.021	0.895 ± 0.020
Put.L	0.869 ± 0.020	0.900 ± 0.014	0.845 ± 0.036	0.890 ± 0.028	0.889 ± 0.018	0.891 ± 0.022	0.896 ± 0.020	0.896 ± 0.022	0.895 ± 0.024	0.896 ± 0.020
Put.R	0.880 ± 0.010	0.904 ± 0.012	0.839 ± 0.029	0.886 ± 0.037	0.887 ± 0.025	0.890 ± 0.026	0.893 ± 0.027	0.897 ± 0.022	0.899 ± 0.020	0.894 ± 0.027
Pal.L	0.810 ± 0.033	0.825 ± 0.050	0.651 ± 0.141	0.737 ± 0.092	0.797 ± 0.034	0.801 ± 0.071	0.820 ± 0.033	0.830 ± 0.036	0.824 ± 0.041	0.826 ± 0.040
Pal.R	0.809 ± 0.037	0.829 ± 0.046	0.437 ± 0.243	0.775 ± 0.091	0.784 ± 0.039	0.800 ± 0.054	0.813 ± 0.030	0.824 ± 0.028	0.822 ± 0.029	0.825 ± 0.031
Hip.L	0.806 ± 0.023	0.851 ± 0.024	0.700 ± 0.050	0.811 ± 0.033	0.814 ± 0.031	0.819 ± 0.030	0.831 ± 0.032	0.829 ± 0.033	0.831 ± 0.030	0.834 ± 0.029
Hip.R	0.817 ± 0.023	0.851 ± 0.024	0.716 ± 0.040	0.813 ± 0.033	0.820 ± 0.028	0.814 ± 0.038	0.832 ± 0.029	0.833 ± 0.031	0.833 ± 0.027	0.836 ± 0.028
Amy.L	0.742 ± 0.064	0.763 ± 0.052	0.505 ± 0.147	0.647 ± 0.096	0.654 ± 0.062	0.735 ± 0.052	0.735 ± 0.043	0.729 ± 0.059	0.728 ± 0.052	0.738 ± 0.055
Amy.R	0.757 ± 0.062	0.768 ± 0.058	0.449 ± 0.126	0.659 ± 0.075	0.625 ± 0.075	0.711 ± 0.056	0.719 ± 0.056	0.716 ± 0.058	0.701 ± 0.072	0.711 ± 0.080
Acc.L	0.684 ± 0.098	0.744 ± 0.053	0.576 ± 0.113	0.668 ± 0.104	0.702 ± 0.117	0.710 ± 0.083	0.728 ± 0.070	0.734 ± 0.066	0.747 ± 0.080	0.722 ± 0.102
Acc.R	0.703 ± 0.076	0.752 ± 0.047	0.429 ± 0.107	0.656 ± 0.084	0.685 ± 0.099	0.704 ± 0.066	0.721 ± 0.068	0.720 ± 0.084	0.733 ± 0.063	0.733 ± 0.068
Avg.	0.808 ± 0.080	0.843 ± 0.071	0.518 ± 0.276	0.780 ± 0.111	0.797 ± 0.105	0.815 ± 0.085	0.824 ± 0.078	0.826 ± 0.081	0.828 ± 0.081	0.829 ± 0.085
wAvg.	0.714 ± 0.066	0.762 ± 0.037	0.505 ± 0.085	0.676 ± 0.051	0.704 ± 0.048	0.722 ± 0.047	0.738 ± 0.052	0.741 ± 0.053	0.750 ± 0.051	0.742 ± 0.052

Table 2. From MICCAI 2012 to IBSR. Mean ± standard deviation DSC values for FIRST, full training and transfer learning with an incremental number of training images. LOO (Leave one out) – the results of leave one out cross validation. TL X – transfer learning with X number of target volumes. No TL – tested directly on the model trained with MICCAI 2012 dataset. The structure acronyms are as follows: left thalamus (Tha.L), right thalamus (Tha.R), left caudate (Cau.L), right caudate (Cau.R), left putamen (Put.L), right putamen (Put.R), left pallidum (Pal.L), right pallidum (Pal.R), left hippocampus (Hip.L), right hippocampus (Hip.R), left amygdala (Amy.L), right amygdala (Amy.R), left accumbens (Acc.L), right accumbens (Acc.R), average value (Avg.) and weighted average DSC (wAvg.).

for the hippocampus (pointed with blue arrow), where the curvature of the structure was preserved similar to the ground truth.

From MICCAI 2012 to IBSR. For this experiment, we fully trained the network using all 15 training images from the MICCAI 2012 training dataset. Next, similar to the previous case, several transfer learning iterations were made using IBSR image volumes as the target and compared with the results of FIRST. Table 2 shows the obtained DSC values for FIRST, fully trained network with IBSR images using leave-one-out cross validation, and transfer learning results using zero to seven images. Very low DSC scores with high standard deviation in the results in No TL (see Table 2) showed the effect of the domain shift problem once again, confirming that this issue is present in both ways. The results were significantly improved when applying transfer learning with only one image, yielding a DSC value of 0.78 ($p < 0.001$). However, it was not higher than that of FIRST due to the different intensity distributions in MRI volumes of the IBSR dataset (Fig. 1). The results obtained by training the network with two images were similar to FIRST and not statistically significant ($p > 0.05$). A significant growth in average was observed when selecting three random images with DSC, reaching up to 0.815 compared with 0.808 with FIRST's ($p < 0.05$). Increasing the number of target set images from four to seven resulted in similar DSC scores, slightly increasing from 0.824 to 0.829 ($p > 0.05$).

The structure-wise improvement after each iteration of transfer learning for the IBSR dataset when using MICCAI 2012 as the source is shown in Fig. 3(b). Because the IBSR dataset comprises MRI volumes with different intensity distributions, we observed slight fluctuations in DSC for some structures. More substantial improvements were observed for the smallest sub-cortical structures, such as the pallidum, amygdala and accumbens, when the number of training images increased. By contrast, the larger structures were more accurately segmented starting from the first iteration of transfer learning and slightly increased through all iterations.

Similar to the previous experiment, we observed a benefit of using transfer learning over training from scratch. As shown in Fig. 4(b), transfer learning obtained better overall DSC than the network trained from scratch for all the iterations. Image standardisation was also useful in the case when no domain adaptation was applied. As shown in Fig. 4(b), using normalisation improved the average DSC from 0.518 to 0.783 ($p < 0.001$). However, similar to the case with the MICCAI 2012 dataset, the improvement throughout the iterations of transfer learning using the standardised images was not considerable. Additionally, the DSC for the standardised images using seven target training subjects (0.807) was significantly lower than that for the original images (0.829) with $p < 0.001$. The comparison of transfer learning and network trained from scratch using the mixed set of normalised images is shown in Fig. 4(b). In the first iteration, mixed set training showed a similar overall DSC of 0.784 to transfer learning (0.780), but the difference was not significant ($p > 0.05$). However, with more images, transfer learning was always significantly better than mixed set training ($p < 0.001$). In the case of mixed set training, the average DSC was significantly improved when adding more images, resulting in the highest DSC of 0.806 (five added images) ($p < 0.001$). However, it showed no further improvements when more images were added to the

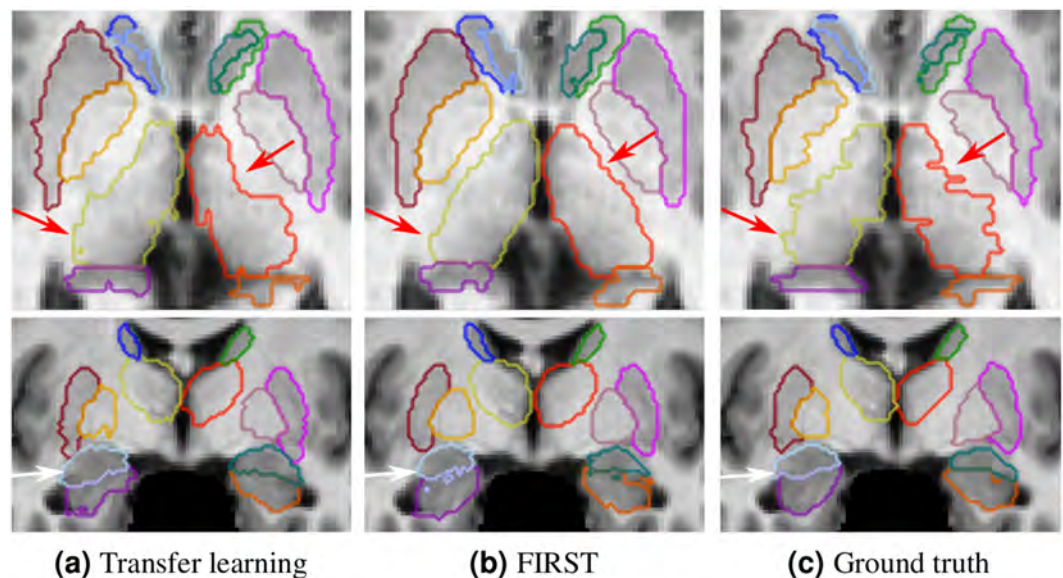


Figure 6. Qualitative results for the IBSR dataset. (a) Transfer learning with three images from three different intensity distribution groups; (b) segmentation result from FIRST; (c) ground truth. The results in the top row are shown in the axial view and those in the bottom are shown in the coronal views. The arrows indicate the following: red → left and right thalamus, white → amygdala and hippocampus structures.

mixed set, instead reaching a plateau. The structure-wise results for the image standardisation and mixed set experiments are presented in Supplementary Tables S2 and S4.

Figure 6 illustrates some qualitative results obtained for this experiment. We showed the results of transfer learning using three training images because they were significantly better than those of FIRST ($p < 0.05$). As shown in the first row in Fig. 6(a), the segmentation result using transfer learning for the thalamus structure (indicated with red arrows) was more similar to the ground truth (Fig. 6c) than that of FIRST. Moreover, some spurious outputs could be observed in the boundaries of the adjacent amygdala and hippocampus structures (indicated with a white arrow) for the FIRST segmentation.

Grouping by MRI scanner. Table 3 shows the results for the three iterations of transfer learning using the IBSR-GE group, the results of FIRST, training from scratch using leave-one-out cross-validation, and when no domain adaptation was applied. The average DSC of leave-one-out significantly outperformed FIRST ($p < 0.001$), whereas the transfer learning with two images yielded similar results, with an average DSC of 0.805 compared with 0.802 for FIRST. A significantly higher DSC of 0.814 ($p < 0.05$) could be achieved using only three images for domain adaptation than that of FIRST. The average DSC was very low when the images were tested directly on the network without domain adaptation, and applying transfer learning using one image improved the average DSC from 0.498 to 0.784; however, it was still lower than that of FIRST.

Table 4 shows the results for the IBSR-Siemens dataset using FIRST, training from scratch with leave-one-out cross-validation, no domain adaptation, and three iterations of transfer learning. The highest results were achieved using leave-one-out with an average DSC of 0.845 significantly outperforming FIRST at 0.818 ($p < 0.001$). The results of testing IBSR-Siemens images without transfer learning showed poor performance, as expected. The performance of the CNN was drastically improved using only one image for transfer learning, yielding a DSC of 0.826, slightly higher than that of FIRST but not statistically significant ($p = 0.08$). A significantly higher DSC than that of FIRST was obtained using two and three images for domain adaptation, reaching 0.840 and 0.846, respectively, with $p < 0.001$ for both cases. Moreover, the results of the third iteration of transfer learning were similar to the one of leave-one-out cross-validation with $p = 0.87$.

Corrected FIRST segmentation as ground truth. Although the obtained results of the previous experiments are promising, we understand that manually segmenting all 14 sub-cortical structures, even for one image, is time consuming compared with, for instance, brain lesion segmentation²³, which is a two-class problem. To overcome this issue, we also studied the use of the segmentation result of FIRST to train the network. FIRST provides a smooth unsupervised segmentation result; however, it does not perform well on small structures and structure boundaries. Therefore, in this experiment, we performed transfer learning using corrected FIRST segmentation outputs. The corrections included the following: removing outliers, filling holes and manually correcting some boundaries of the structures. We must note that the manual correction was performed by an operator who considered only the structures with good visual contrast from their surroundings. An example of a corrected segmentation for the left accumbens and left caudate structure boundaries is shown in Fig. 7. In this experiment, we used the IBSR dataset as the source, and two iterations of transfer learning with MICCAI 2012. The result when using one image with corrected labels for transfer learning was slightly higher than that of FIRST, yielding a DSC score of 0.805 ± 0.075 . However, the difference was not statistically significant ($p = 0.74$). Significant

Str.	FIRST	LOO	No TL	TL 1	TL 2	TL 3
Tha.L	0.894 ± 0.015	0.908 ± 0.014	0.092 ± 0.228	0.849 ± 0.039	0.873 ± 0.031	0.881 ± 0.031
Tha.R	0.882 ± 0.011	0.913 ± 0.017	0.122 ± 0.202	0.845 ± 0.081	0.881 ± 0.027	0.884 ± 0.028
Cau.L	0.771 ± 0.047	0.891 ± 0.017	0.329 ± 0.294	0.854 ± 0.039	0.867 ± 0.029	0.879 ± 0.020
Cau.R	0.806 ± 0.026	0.892 ± 0.022	0.351 ± 0.317	0.862 ± 0.034	0.882 ± 0.024	0.885 ± 0.026
Put.L	0.867 ± 0.023	0.896 ± 0.015	0.835 ± 0.038	0.875 ± 0.025	0.882 ± 0.018	0.887 ± 0.023
Put.R	0.883 ± 0.009	0.901 ± 0.012	0.831 ± 0.031	0.875 ± 0.031	0.883 ± 0.028	0.887 ± 0.024
Pal.L	0.802 ± 0.031	0.809 ± 0.052	0.652 ± 0.149	0.773 ± 0.074	0.798 ± 0.046	0.813 ± 0.033
Pal.R	0.809 ± 0.028	0.816 ± 0.049	0.491 ± 0.268	0.780 ± 0.070	0.782 ± 0.054	0.805 ± 0.047
Hip.L	0.804 ± 0.015	0.854 ± 0.021	0.693 ± 0.046	0.801 ± 0.034	0.812 ± 0.036	0.819 ± 0.031
Hip.R	0.812 ± 0.014	0.851 ± 0.022	0.711 ± 0.041	0.803 ± 0.03	0.814 ± 0.032	0.819 ± 0.037
Amy.L	0.745 ± 0.050	0.756 ± 0.045	0.464 ± 0.163	0.709 ± 0.046	0.719 ± 0.046	0.719 ± 0.062
Amy.R	0.758 ± 0.055	0.758 ± 0.058	0.443 ± 0.143	0.689 ± 0.068	0.736 ± 0.054	0.723 ± 0.059
Acc.L	0.655 ± 0.099	0.739 ± 0.059	0.563 ± 0.109	0.663 ± 0.099	0.669 ± 0.102	0.697 ± 0.065
Acc.R	0.691 ± 0.082	0.743 ± 0.048	0.392 ± 0.100	0.602 ± 0.080	0.673 ± 0.074	0.691 ± 0.061
Avg.	0.802 ± 0.083	0.838 ± 0.073	0.498 ± 0.284	0.784 ± 0.101	0.805 ± 0.089	0.814 ± 0.084
wAvg.	0.696 ± 0.069	0.754 ± 0.038	0.484 ± 0.078	0.657 ± 0.057	0.692 ± 0.063	0.712 ± 0.043

Table 3. From MICCAI 2012 to IBSR-GE. Mean ± standard deviation DSC values for FIRST, full training with leave-one-out cross-validation (LOO) and transfer learning with an incremental number of training images. TL X – transfer learning with X number of target volumes. S The structure acronyms are as follows: left thalamus (Tha.L), right thalamus (Tha.R), left caudate (Cau.L), right caudate (Cau.R), left putamen (Put.L), right putamen (Put.R), left pallidum (Pal.L), right pallidum (Pal.R), left hippocampus (Hip.L), right hippocampus (Hip.R), left amygdala (Amy.L), right amygdala (Amy.R), left accumbens (Acc.L), right accumbens (Acc.R), average value (Avg.) and weighted average DSC (wAvg.).

Str.	FIRST	LOO	No TL	TL 1	TL 2	TL 3
Tha.L	0.892 ± 0.022	0.914 ± 0.013	0.201 ± 0.222	0.888 ± 0.019	0.891 ± 0.016	0.900 ± 0.014
Tha.R	0.889 ± 0.014	0.916 ± 0.014	0.000 ± 0.000	0.894 ± 0.015	0.902 ± 0.015	0.904 ± 0.015
Cau.L	0.805 ± 0.028	0.906 ± 0.017	0.663 ± 0.075	0.896 ± 0.021	0.905 ± 0.014	0.905 ± 0.015
Cau.R	0.892 ± 0.016	0.903 ± 0.015	0.663 ± 0.141	0.893 ± 0.014	0.892 ± 0.023	0.885 ± 0.026
Put.L	0.872 ± 0.016	0.909 ± 0.006	0.866 ± 0.021	0.901 ± 0.014	0.902 ± 0.007	0.910 ± 0.009
Put.R	0.875 ± 0.011	0.908 ± 0.010	0.856 ± 0.018	0.905 ± 0.013	0.907 ± 0.011	0.908 ± 0.013
Pal.L	0.827 ± 0.034	0.857 ± 0.028	0.649 ± 0.135	0.866 ± 0.020	0.852 ± 0.043	0.862 ± 0.027
Pal.R	0.808 ± 0.055	0.857 ± 0.024	0.328 ± 0.150	0.836 ± 0.017	0.835 ± 0.037	0.840 ± 0.034
Hip.L	0.811 ± 0.036	0.843 ± 0.030	0.715 ± 0.059	0.818 ± 0.028	0.821 ± 0.034	0.837 ± 0.025
Hip.R	0.826 ± 0.034	0.850 ± 0.031	0.725 ± 0.040	0.812 ± 0.025	0.827 ± 0.029	0.846 ± 0.024
Amy.L	0.736 ± 0.090	0.778 ± 0.067	0.587 ± 0.048	0.751 ± 0.064	0.787 ± 0.048	0.780 ± 0.060
Amy.R	0.756 ± 0.080	0.787 ± 0.057	0.463 ± 0.091	0.725 ± 0.059	0.733 ± 0.072	0.760 ± 0.052
Acc.L	0.742 ± 0.069	0.754 ± 0.041	0.603 ± 0.128	0.705 ± 0.052	0.754 ± 0.041	0.757 ± 0.040
Acc.R	0.725 ± 0.063	0.769 ± 0.044	0.502 ± 0.086	0.668 ± 0.045	0.749 ± 0.057	0.750 ± 0.039
Avg.	0.818 ± 0.073	0.854 ± 0.065	0.559 ± 0.256	0.826 ± 0.085	0.840 ± 0.070	0.846 ± 0.066
wAvg.	0.747 ± 0.044	0.778 ± 0.029	0.548 ± 0.082	0.713 ± 0.028	0.766 ± 0.036	0.769 ± 0.030

Table 4. From MICCAI 2012 to IBSR-Siemens. Mean ± standard deviation DSC values for FIRST, full training with leave-one-out cross-validation (LOO) and transfer learning with an incremental number of training images. TL X – transfer learning with X number of target volumes. The structure acronyms are as follows: left thalamus (Tha.L), right thalamus (Tha.R), left caudate (Cau.L), right caudate (Cau.R), left putamen (Put.L), right putamen (Put.R), left pallidum (Pal.L), right pallidum (Pal.R), left hippocampus (Hip.L), right hippocampus (Hip.R), left amygdala (Amy.L), right amygdala (Amy.R), left accumbens (Acc.L), right accumbens (Acc.R), average value (Avg.) and weighted average DSC (wAvg.).

improvements could be achieved when using two corrected images, obtaining a DSC of 0.817 ± 0.075 ($p < 0.001$). This shows that the output segmentation from FIRST could be used as a starting point for transfer learning and avoid manual segmentation of all 14 structures from scratch.

Training and testing times. As shown in recent studies²², transfer learning allows the deep neural network to converge faster than a CNN trained from scratch. Our studies clearly confirmed this statement, by achieving much faster training time for the network. The average training time per epoch using a single training image was eight seconds, and it gradually increased when adding more images, reaching 63 seconds per epoch with seven

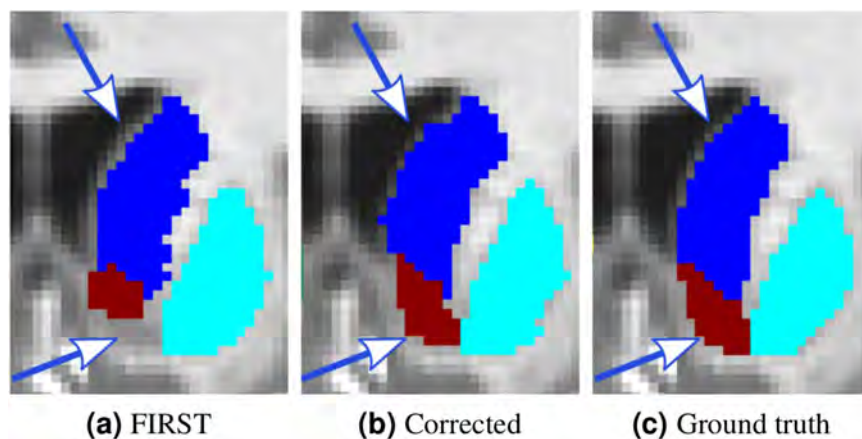


Figure 7. Corrected left accumbens and left caudate structures from the FIRST segmentation output. Coronal view: **(a)** FIRST segmentation; **(b)** corrected segmentation; **(c)** ground truth. Examples of some of the corrected areas are indicated with arrows.

training images. Comparison with full training, which took 832 seconds per epoch on average, the training time of transfer learning was less by two orders of magnitude when using one image and ten times less when using seven training images.

The testing time using our method was 1.3 minutes (run on GPU) + 3.7 minutes (atlas registration, run on CPU) per volume in average. On the other hand, FIRST took approximately 10 minutes to test one subject volume on average; however, this method does not require any training. All the experiments were run using a machine with a 3.40-GHz CPU clock and on a single TITAN-X GPU (NVIDIA corp, United States) with 12 GB of RAM memory.

Discussion

Our experiments showed that the weights of the convolutional layers trained with the source dataset could generalise the features extracted from the target set. However, the domain shift problem requires a fine tuning of the way these features are interpreted. Thus, in our approach, we updated the weights of the fully connected layers and trained the classification layer from scratch.

In the experimental results, we showed that similar intensity distributions within the dataset helped the network to better generalise and provide a more predictable outcome. This was observed when the MICCAI 2012 dataset was used as the target, where we could see a smooth increase in the structure-wise DSC (Fig. 3a) and the overall average DSC (Table 1) after each iteration of transfer learning. By contrast, when the IBSR dataset was used as the target, the within-group variability of the MRI volumes in the dataset affected the results of transfer learning. As observed earlier (Table 2), substantial improvement in average DSC occurred in the first three iterations of transfer learning, but the results for the subsequent iterations reached a plateau. This means that the added images through the fourth and seventh iterations did not provide the network with more useful information, making the learned weights of the CNN less general. The weights of the convolutional layers remained the same during transfer learning, indicating that they are not adapted to fit the new target domain. Therefore, when dealing with different intensity distributions, it is better to introduce more representative examples to make the fully connected layers better adapt to a new dataset. Additionally, one could notice that the results when using seven training images were not close to those using leave-one-out cross-validation. This behaviour was expected because in leave-one-out, 17 images were used to segment only one subject volume, which allowed the network to learn more variations present in this dataset. The within-group variability of the intensity distributions in the IBSR dataset showed unstable results for some structures during transfer learning (Fig. 3b). Additionally, random selection of the MRI volumes after every iteration added up to this behaviour because the outcome of the CNN relies on the descriptiveness of the training images. However, the overall trend showed an increase in DSC when more images were used for training.

Similar behaviour as in the first two experiments were observed when the images of the IBSR dataset were grouped into two sets by the MRI scanner type. As shown in Tables 3 and 4, for both new groups, transfer learning significantly outperformed FIRST using only three and two target training images for the IBSR-GE ($p < 0.05$) and IBSR-Siemens ($p < 0.001$) datasets, respectively. Ignoring the different number of images in both of these groups, it is interesting that the images of IBSR-GE were more difficult to segment for all considered segmentation approaches, including FIRST, deep learning, and domain adaptation. One cause may be the imaging artefacts present in some of the images from the IBSR-GE group. Figure 8 illustrates examples of the images from the two groups, indicating there are motion artefacts and more noise present in the IBSR-GE group than in the IBSR-Siemens images.

Comparison of transfer learning with training from scratch was also carried out to show the effectiveness of transferring knowledge. One could observe, for the case of IBSR (Fig. 4b), that the increase in DSC for transfer learning was smoother because the number of training images increased, whereas the results of the CNN trained from scratch showed a steep increase in the third iteration. This was due to the within-group variability of the

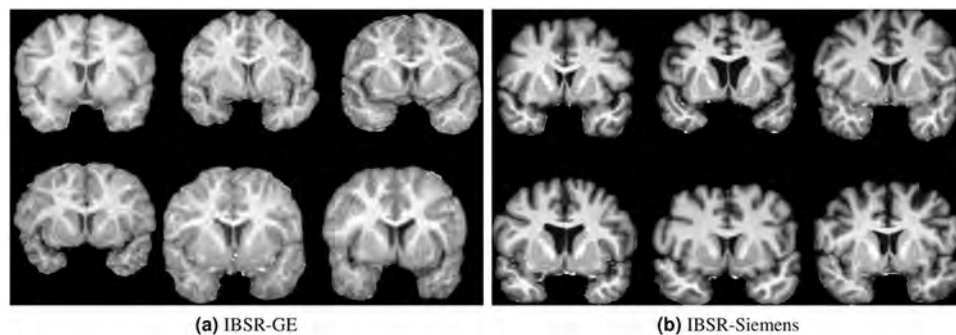


Figure 8. Illustration of some of the images from the (a) IBSR-GE and (b) IBSR-Siemens datasets.

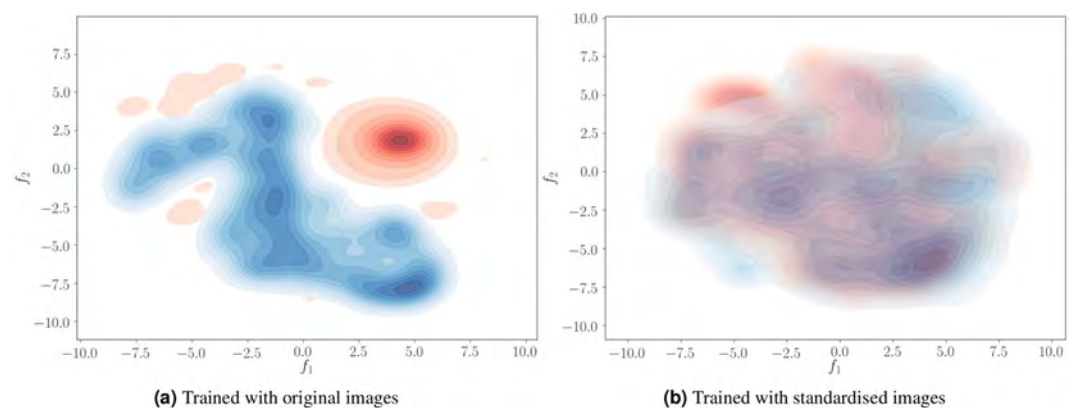


Figure 9. Visualisation of the concatenation layer outputs for the MICCAI 2012 (Red) and IBSR (Blue) datasets using original (a) and standardised images (b). 555-Dimensional feature vectors projected into 2-dimensional space using the t-SNE algorithm.

image volumes in this dataset, and the selected image at this iteration was more representative than the previous two MRI scans. Additionally, because the training images for both were the same for each iteration, transfer learning compensated for the unseen cases and produced better results. A similar comparison (Fig. 4a) for the case of MICCAI 2012 dataset showed a gradual increase for both, transfer learning and training from scratch, with the former yielding better results in all iterations. Once more, we saw that a similarity in the intensity distributions within the target dataset makes a considerable difference.

As shown in the results (Fig. 4), image standardisation for both datasets helped to achieve better results even without transfer learning. Interestingly, the outcome of transfer learning using one image with the original images was similar to that of normalised images without transfer learning. This means that transfer learning implicitly performs normalisation by adapting the weights of the fully connected layers to better interpret the features extracted by the convolutional layers. As shown in Fig. 9(a), the distributions of the feature vectors extracted from the concatenation layer for two datasets have a clear separation when the network is trained with the original images. By contrast, these distributions overlap when the network is trained with the standardised images (Fig. 9b). Accordingly, when no transfer learning is applied, the features extracted from the target images would be similar to the source, and the network produces better results even without initial training of the network. Although the image standardisation is helpful to directly use the network for segmenting images in other domains, it involves image intensity interpolation, which could be disadvantageous to obtain a better-adapted model using transfer learning.

Moreover, as shown in Fig. 4, training the network with a mixed set of images from different domains resulted in a slight increase in overall DSC when up to four images were added to the training set. However, there were no improvements observed with more images, but the average DSC stayed within a similar range. This behaviour of the network was caused by the interpolation in the IBSR image labels and the differences in the raters. According to these observations, domain adaptation using transfer learning would be a better choice than adding new images to the training process when there are only a few annotated images available in the target domain.

In the case of MICCAI 2012, all MRI volumes shared a similar intensity distribution, making the fully connected and classification layers of the network able well trained to overcome the results of FIRST using only one training image volume. On the other hand, the performance of the network also depends on the source dataset. This hypothesis arises due to the observation seen in the results with the IBSR dataset when MICCAI 2012 is used as the source. Because MICCAI 2012 images have a similar intensity profile, there are no filters in the convolutional layers that can consider discrepancies in intensity distribution. However, this assumption requires further

analysis and experiments that should involve other datasets with similar within-dataset dissimilarities as in IBSR. This has not been analysed due to a lack of existing datasets that have ground truth segmentation labels for the sub-cortical structures. Therefore, we considered it a limitation of this study, and more elaboration is needed with more datasets containing intensity distribution dissimilarities to quantitatively analyse their impact on the performance of transfer learning.

The results of the weighted and unweighted DSC values had the same trend over the transfer learning iterations. This demonstrates that our method was consistent in terms of segmenting all the structures regardless their different sizes. However, subtle differences were observed where a higher DSC for one method was less in the weighted DSC. This shows that average DSC score favours high accuracy in the larger structures, whereas the impact of the smaller structures is lesser to the overall result. This issue could bring unreliable results in comparing different methods, especially, for evaluating sub-cortical structure segmentation approaches where the imbalance among classes is considerably large. Although the conventional DSC metric is mostly used in the literature, we encourage using the weighted DSC to verify the robustness of methods. The weighted average DSC values for all the experiments could be found in Supplementary Fig. S1.

As illustrated in the qualitative results (Figs 5 and 6), transfer learning produced segmentation masks that were more similar to the ground truth than FIRST. Because FIRST is based on the active shape model strategy, it tries to preserve the structure boundaries to the mean structure shape defined in the method itself. Therefore, some structural variations in shape may decrease the performance of this method. Moreover, FIRST found it difficult to properly define the boundaries for the adjacent amygdala and hippocampus structures (Fig. 6) due to their similar intensity profiles with no differentiable separation between them. By contrast, our method relies not only on local information but also on the surrounding context, considering the brain structural shape in the region.

One of the goals of transfer learning is to adapt a network that performs well in a new domain using only a few images for training. However, we have performed the domain adaptation using more target training images to confirm that the network does not overfit. According to the experimental results, adding more images did not further improve the results but showed similar performance in all the iterations of transfer learning. Therefore, in this paper, we have shown the results of only seven iterations where the network has reached the point of stability. The results of transfer learning with more than seven images is shown in Supplementary Table S5.

Although transfer learning was shown to be effective to deal with the domain adaptation problem, it still requires at least one manually annotated image volume, limiting our method to be used out of the box. However, the initial manual segmentation could be carried out more quickly by correcting the segmentation output from FIRST as shown earlier. Once the neural network is adapted, it could then be applied continuously with no retraining needed. Another limitation of our approach, as in all deep learning methods in general, is the necessity for the computational power of GPU. The training and testing times of such approaches will grow when run on CPU. Nonetheless, more powerful and affordable GPUs are becoming available.

Conclusions

In this paper, we have demonstrated the application of transfer learning for sub-cortical structure segmentation to overcome the domain shift problem. In our experiments we have employed our previously proposed deep learning strategy that combines spatial and convolutional features. As shown in the results, we could achieve significantly better results than those of the well-known FIRST tool using one and three images for MICCAI 2012 ($p < 0.001$) and IBSR ($p < 0.05$) datasets, respectively. Accordingly, the transfer learning strategy is an excellent way to overcome the demand of deep learning methods for a large amount of data, especially in medical image analysis, where the ground truth availability is scarce. It allows us to use existing available datasets to bootstrap deep learning architectures and adapt the weights to fit to a new domain using much less training set than in full training.

We also showed that within-group variability in a dataset has an important effect on the network's generalisability, suggesting that domain adaptation performs better when target images share a similar intensity distribution. Transferring the knowledge obtained from one dataset to another actually helped to achieve better performance. This was confirmed in our experiments, where transfer learning yielded superior results over the network trained from scratch using the same number of training images. Moreover, transfer learning was shown to be a better choice than other alternative solutions to the domain shift problem such as standardising images and mixed dataset training. Additionally, as we have seen in the experimental results, great acceleration in the training speed of the network could be achieved. Furthermore, we have made the source code of the pipeline available to the research community (https://github.com/NIC-VICOROB/sub-cortical_segmentation).

References

1. Frazier, J. A. *et al.* Structural brain magnetic resonance imaging of limbic and thalamic volumes in pediatric bipolar disorder. *Am. J. Psychiatry* **162**, 1256–1265 (2005).
2. De Jong, L. *et al.* Strongly reduced volumes of putamen and thalamus in alzheimer's disease: an mri study. *Brain* **131**, 3277–3285 (2008).
3. Rimol, L. M. *et al.* Cortical thickness and subcortical volumes in schizophrenia and bipolar disorder. *Biol. psychiatry* **68**, 41–50 (2010).
4. Mak, E., Bergsland, N., Dwyer, M., Zivadinov, R. & Kandiah, N. Subcortical atrophy is associated with cognitive impairment in mild parkinson disease: a combined investigation of volumetric changes, cortical thickness, and vertex-based shape analysis. *Am. J. Neuroradiol.* **35**, 2257–2264 (2014).
5. Houtchens, M. *et al.* Thalamic atrophy and cognition in multiple sclerosis. *Neurol.* **69**, 1213–1223 (2007).
6. Kikinis, R. *et al.* A digital brain atlas for surgical planning, model-driven segmentation, and teaching. *IEEE Transactions on Vis. Comput. Graph.* **2**, 232–241 (1996).
7. Phillips, J. L., Batten, L. A., Tremblay, P., Aldosary, F. & Blier, P. A prospective, longitudinal study of the effect of remission on cortical thickness and hippocampal volume in patients with treatment-resistant depression. *Int. J. Neuropsychopharmacol.* **18**, py037 (2015).

8. Storelli, L. *et al.* Measurement of Whole-Brain and Gray Matter Atrophy in Multiple Sclerosis: Assessment with MR Imaging. *Radiol.* **172468** (2018).
9. González-Villà, S. *et al.* A review on brain structures segmentation in magnetic resonance imaging. *Artif. Intell. Medicine* **73**, 45–69 (2016).
10. Patenaude, B., Smith, S. M., Kennedy, D. N. & Jenkinson, M. A Bayesian model of shape and appearance for subcortical brain segmentation. *Neuroimage* **56**, 907–922 (2011).
11. Fischl, B. FreeSurfer. *Neuroimage* **62**, 774–781 (2012).
12. Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 1097–1105 (2012).
13. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778 (2016).
14. Girshick, R., Donahue, J., Darrell, T. & Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 580–587 (2014).
15. LeCun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **86**, 2278–2324 (1998).
16. Litjens, G. *et al.* A survey on deep learning in medical image analysis. *Med. Image Analysis* **42**, 60–88 (2017).
17. Bernal, J. *et al.* Deep convolutional neural networks for brain image analysis on magnetic resonance imaging: a review. *Artif. intelligence medicine*, <https://doi.org/10.1016/j.artmed.2018.08.008> (2018).
18. Dolz, J., Desrosiers, C. & Ayed, I. B. 3D fully convolutional networks for subcortical segmentation in MRI: A large-scale study. *NeuroImage* **170**, 456–470 (2018).
19. Kushibar, K. *et al.* Automated sub-cortical brain structure segmentation combining spatial and deep convolutional features. *Med. Image Analysis* **48**, 177–186 (2018).
20. Wachinger, C., Reuter, M. & Klein, T. Deepnat: Deep convolutional neural network for segmenting neuroanatomy. *NeuroImage* **170**, 434–445 (2018).
21. Kamnitsas, K. *et al.* Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. In *International Conference on Information Processing in Medical Imaging*, 597–609 (Springer, 2017).
22. Tajbakhsh, N. *et al.* Convolutional neural networks for medical image analysis: full training or fine tuning? *IEEE Transactions on Med. Imaging* **35**, 1299–1312 (2016).
23. Ghafoorian, M. *et al.* Transfer Learning for Domain Adaptation in MRI: Application in Brain Lesion Segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 516–524 (Springer, 2017).
24. Landman, B. & Warfield, S. MICCAI 2012 workshop on multi-atlas labeling. In *Medical Image Computing and Computer Assisted Intervention Conference* (2012).
25. Rohlfing, T. Image similarity and tissue overlaps as surrogates for image registration accuracy: widely used but unreliable. *IEEE transactions on medical imaging* **31**, 153–163 (2012).
26. Kennedy, D. N. *et al.* CANDIShare: a resource for pediatric neuroimaging data. *Neuroinformatics* **10**, 319–22 (2012).
27. About The Creative Commons Licenses. <http://creativecommons.org/about/licenses> (2019).
28. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. *ArXiv e-prints* 1412.6980 (2014).
29. Iglesias, J. E., Liu, C.-Y., Thompson, P. M. & Tu, Z. Robust brain extraction across datasets and comparison with publicly available methods. *IEEE Transactions on Med. Imaging* **30**, 1617–1634 (2011).
30. Smith, S. M. Fast robust automated brain extraction. *Hum. Brain Mapp.* **17**, 143–155 (2002).
31. Caviness, V. S. Jr., Meyer, J., Makris, N. & Kennedy, D. N. MRI-based topographic parcellation of human neocortex: an anatomically specified method with estimate of reliability. *J. Cogn. Neurosci.* **8**, 566–587 (1996).
32. Modat, M. *et al.* Fast free-form deformation using graphics processing units. *Comput. Methods Programs Biomed.* **98**, 278–284 (2010).
33. Chollet, F. *et al.* Keras. <https://keras.io> (2019).
34. Gorgolewski, K. J. *et al.* Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in Python. 0.13.1, <https://doi.org/10.5281/zenodo.581704> (2017).
35. Jones, E. *et al.* SciPy: Open source scientific tools for Python (2001).
36. Brett, M. *et al.* nibabel: 2.1.0. *Zenodo*, <https://doi.org/10.5281/zenodo.60808> (2016).
37. Nyúl, L. G., Udupa, J. K. & Zhang, X. New variants of a method of mri scale standardization. *IEEE transactions on medical imaging* **19**, 143–150 (2000).
38. Shah, M. *et al.* Evaluating intensity normalization on MRIs of human brain with multiple sclerosis. *Med. Image Analysis* **15**, 267–282 (2011).
39. Crum, W. R., Camara, O. & Hill, D. L. Generalized overlap measures for evaluation and validation in medical image analysis. *IEEE transactions on medical imaging* **25**, 1451–1461 (2006).

Acknowledgements

Kaiser Kushibar and Jose Bernal hold FI-DGR2017 grant from the Catalan Government with reference numbers 2017FI_B00372 and 2017FI_B00476, respectively. This work has been partially supported by La Fundació la Marató de TV3, by Retos de Investigación TIN2014-55710-R, TIN2015-73563-JIN, and DPI2017-86696-R from the Ministerio de Ciencia y Tecnología. The authors gratefully acknowledge the support of the NVIDIA Corporation with their donation of the TITAN-X PASCAL GPU used in this research. We also thank the two anonymous reviewers for their constructive comments which helped to improve the quality of this study.

Author Contributions

K.K. performed the experiments and wrote the manuscript. S.V., S.G., J.B. and M.C. provided comments and feedback on the paper and the results. A.O. and X.L. supervised the project. All authors reviewed the manuscript.

Additional Information

Supplementary information accompanies this paper at <https://doi.org/10.1038/s41598-019-43299-z>.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2019

Chapter 4

Unsupervised domain adaptation in deep learning for brain magnetic resonance image segmentation

In this chapter, we present our approach for unsupervised domain adaptation that minimises the differences in the CNN activation maps for two different imaging domains. The proposal has been evaluated in two different brain segmentation problems: 1) sub-cortical brain structure segmentation; and 2) brain white matter hyperintensities segmentation. For both of the problems, the proposed domain adaptation significantly improved the baseline models and showed similar or better performance than the traditional unsupervised segmentation tools. This work has been submitted to the following journal:

Submitted to the Knowledge-Based Systems journal (KBS) (Under Review) JCR CSAI IF: 5.101, Q1(17/133)
--

Chapter 5

Results and discussions

This thesis encompasses a natural progression of the sub-cortical brain structures segmentation problem starting from a network architecture proposal using full training, moving to a transfer learning and unsupervised approaches for the domain adaptation problem. In this chapter, we present a comprehensive discussion on the results and findings obtained in this thesis. The following sections provide a meta-analysis of the previous chapters, outlining the most important aspects of each proposal on completing the main objective.

5.1 Network architecture

5.1.1 Implicit convolutional features

In Chapter 2, we introduced our proposed deep learning method for accurate segmentation of the sub-cortical brain structures in MRI. The network architecture comprised of three convolutional branches that process patch samples representing the orthogonal – axial, coronal, and sagittal – views of MR images.

The convolutional features are extracted directly from the input patches making them implicit representations that were learned during training. The derived implicit features appear to be domain dependent, which become distinct for samples attained from different imaging domains. This has been illustrated in Chapter 3, where the projection of the convolutional features were clearly separated on the 2D plane. We addressed this issue in two different domain adaptation methods: 1) supervised transfer learning also in Chapter 3; and 2) unsupervised domain adaptation in Chapter 4. In the first case, we adapt the fully connected layers to interpret the implicit features in a new way. In doing so, we are agreeing with the convolutional layer outputs assuming that they are good representations for the input

patches and forcing the fully connected layers to adapt to it. Since the network has been pre-trained to solve a similar problem, we could obtain competent results even when freezing (i.e. not updating) the weights of all the convolutional layers, hence, optimising for the computational time and number of training images because the number of trainable parameters were reduced drastically. In the unsupervised domain adaptation case, the problem has been approached from a different viewpoint, where we do not agree with the implicit feature representation and force the convolutional layers to adapt to a new domain. By employing the proposed histogram loss, which is solely based on the activation maps, there was no need for the new target domain to have the ground truth masks, therefore being an unsupervised domain adaptation. Moreover, in UDA we are using the cross entropy loss as one of the terms which is computed using the available labelled source data. Since the extracted features from the source are changed to match the ones of the target during training, the classifier layer has to be retrained to interpret them in a new way and produce accurate segmentation results. Therefore, the cross entropy loss was included as one of the terms in the total loss function, which relearns how to classify the source feature maps after changing them.

5.1.2 Explicit spatial features

The convolutional layers in the network are followed by a concatenation of the convolutional features with the spatial features, which continued by fully connected and classification layers. In the proposed architecture, unlike the learned implicit features, the spatial features provide explicit information, which are extracted from a probabilistic atlas. The explicit features are domain independent and serve as an additional guide to the network to perform the segmentation task.

In the sub-cortical brain structure segmentation problem, employing additional explicit information is an effective way to overcome the lack of available training images. Accordingly, by using the spatial features, the network was able to reduce the diversity limitations in the training samples, where specific abnormalities or imaging artefacts were present only in a small subset of the whole training set. Moreover, since the explicit features are domain independent, their purpose was still suitable during the domain adaptation for both supervised and unsupervised methods.

There are some concerns that could be raised about the possible registration errors and their effect on the performance of the network. In the best case scenario, when the registration is done perfectly, the mask labels could be directly propagated to the target image to obtain a segmentation mask. However, we have to accommodate the registration errors by adding extra features or using more complex pipelines. Examples of such integration were done using multi-atlas techniques

as majority voting and label fusion. In our proposals, looking from an opposing perspective to deep learning, leveraging CNN based elements to the pipeline could be perceived as an addition to an atlas based segmentation. In this case, the CNN acts as an error correction mechanism to the atlas based segmentation. Moreover, the atlas probabilities provide advantageous contextual information about spatial location of an input sample. This is useful in terms of differentiating left and right sub-cortical structures as well as to which structure the input sample more likely belongs, even in presence of registration errors. Another example of contextual information would be the deployment of multi-scale [51] patch samples to provide global and local information to the network.

Note that the fully connected part of the CNN was able to find a balance between the implicit and explicit information without inclining towards one single input feature. In our network, the number of units in each fully connected layer decrease as it goes deeper. There is no specific guideline on selecting the number of units for the fully connected layers. However, one can optimise it for the number of training parameters without compromising the representative capacity of the network.

5.1.3 Sample selection

The selective sample selection technique proposed in this thesis, where the negatives were collected only from the structure boundaries, was applied for the full training, and both domain adaptation methods. As it was shown in the ablation study in Chapter 2, the proposed sample selection scheme showed significant ($p < 0.001$) improvements over random sample selection. This actually means that for the sub-cortical segmentation, the most important and difficult parts are in the structure boundaries.

As it was introduced in Chapter 1, differences in manual segmentation masks mostly occur on the structure boundaries, which could be some slight over- or under-segmentation depending on the rater. This suggests that when applying transfer learning, we are not only adapting the network to reduce the domain shift effect but also readjusting boundary delineation process to a new rater. However, when it comes to the unsupervised domain adaptation, we are unable to accommodate the segmentation to a different rater and optimising the network only for the reduction of the domain shift effect. This indicates that during unsupervised domain adaptation the network performs to the maximum of its pre-trained model capabilities. It was demonstrated in Chapter 4, where the periventricular WMH lesions were not segmented in the target images because the network was specifically trained to segment them as background in accordance with the source ground truth masks. In order to solve this issue, the network has to be retrained using transfer learning to accommodate the segmentation protocol differences. Since the operator variability

often occur systematically, an alternative approach would be to apply a correction method such as AdaBoost [52].

5.2 Knowledge transfer for domain adaptation

Currently, one of the biggest challenges in deep learning for medical image analysis is the domain difference in MR images. As it was covered in Chapter 3, transfer learning was an effective way for adapting the network to a new domain. We have discussed how in transfer learning the fully connected layers in the network are retrained to interpret different implicit features of a new domain. In this section, we discuss the concept of domain adaptation with transfer learning with more details.

5.2.1 Number of training images and trainable parameters

It is a common practice to employ pre-trained networks with weights trained with a large scale dataset – such as ImageNet [53], which consists of over 15 million labelled high-resolution natural images. This type of knowledge transfer is useful when there is not enough training data to train the network from scratch. Such pre-trained network models are used as feature extractors or fine-tuned with the idea that the early convolutional layers are kept frozen to reduce the number of trainable parameters. Using this type of pre-trained models could be insufficient for some tasks where the network architecture does not fit the problem and the extracted features do not have enough representative power. In order to improve their performance, new additional layers are introduced or more layers are retrained. However, this increases the number of trainable parameters and require more training images. In our method, we froze all the convolutional layers and fine-tuned only the fully connected part of the network. Since our pre-trained model weights were trained to solve the same problem, learning the implicit feature differences were easier, which allowed to significantly ($p < 0.001$) outperform the traditional method FIRST using only one and three training images for the MICCAI 2012 and IBSR datasets, respectively. Hence, adjusting only the weights of the densely connected layers were enough, although it still had some limitations.

We observed that applying transfer learning to the network with more images improved the results, but it reached a plateau when the number of training images increased further. This happens when the representative power of the trainable part of the network approaches to its maximum. This means that at some point, the implicit features extracted from the pre-trained convolutional layers cannot generalise to the new domain and it has its limits. Despite this fact, the original idea of transfer learning is to obtain better performance with a limited number of training images.

Therefore, our findings recommend to increase the number of trainable parameters of the network when more training images are available.

5.2.2 Domain adaptation and image standardisation

In Chapter 3, we compared the performance of the domain adaptation using transfer learning and image standardisation techniques. Image standardisation is an approach where intensity differences in MRI scans are transformed into a common space. In our comparison, we registered all the images into the MNI space and applied Nyúl histogram matching [54]. This transformation allowed to have 1 *mm* isotropic resolution and slice thickness for all the MR images and also minimised the differences in the intensities.

The initial segmentation DSC was improved to 0.787 in comparison to directly testing the target cases on a network trained with source dataset without image standardisation, which yielded DSC of 0.608. However, performing transfer learning using the transformed images did not show promising results. It required four training images to achieve the closest DSC of 0.837 to the DSC result of 0.834 with original images when using only one image. Moreover, the network reached the plateau faster and did not show higher rates of improvement when more training images were added to the target training set.

This behaviour was observed due to the interpolation in both the T1-w images and ground truths. The interpolation on T1-w images causes artefacts, which deteriorate the image quality, especially in the boundaries that are the most important areas in the images. The interpolation on ground truth occurs two times. First, when moving to MNI, where misalignment between the T1-w and ground truth mask occurs. This in turn affects how the network is trained. Second, when moving the segmentation output from MNI to the original space, which also reduces the segmentation accuracy.

Additionally, in the experiment with mixed set of images, it was also shown that transfer learning is preferred over increasing the diversity in the training dataset. Mixing images during training could be beneficial in terms of improving the generalising capability of the CNN. However, having manual segmentation masks from different raters for the images in the mix set is not recommended as they tend to add more noise and restrict the network from converging into the global minimum.

5.3 Unsupervised domain adaptation

Transfer learning approach is an extremely effective way for domain adaptation, however, manually segmenting the brain structures even for a single image requires a long time and effort, therefore, this approach could be unsuitable in an ad-hoc situation. In this section, we extend the discussion provided in Chapter 4, where we introduced our proposed method for unsupervised domain adaptation (UDA from now on), by merging its key aspects with the concepts of the previous proposals.

5.3.1 Applications of domain adaptation

We have demonstrated the application of UDA for two different segmentation problems: 1) sub-cortical brain structure segmentation; and 2) brain white matter hyperintensities (WMH) segmentation. The experimental results showed that UDA was useful to reduce the domain shift effect for both of the considered segmentation tasks. Moreover, it showed the network’s versatility in terms of applicability to different problems.

Regarding the problem specifications, WMH lesion segmentation is profoundly different than the brain structure segmentation problem. In healthy cases, the sub-cortical brain structures are similar for all subjects in terms of location. For instance, the amygdala structure is always adjacent from an anterior-superior location to the hippocampus structure. The differences in structures between two subjects are in their shape and volume. However, the general representative shape of the sub-cortical structures are the same. For example, the caudate structure always has the form of a curved droplet with the head and tail oriented from anterior to posterior, respectively. The segmentation method FIRST actually takes advantage of this property by using an active shape model algorithm, where the average shape model is fitted to the structures during the segmentation process. Moreover, our brain structure segmentation method also leverages this characteristic by employing spatial information in the form of a probabilistic atlas and the selective sample selection technique. However, in the WMH segmentation task, the lesions vary in volume, shape and location, which requires a different approach for solving the problem. Since we could not benefit from the sample selection method in this segmentation task, we used a cascaded training concept that has been previously used in Valverde et al. [55]. In the cascaded training scheme, the network is trained two times one after another to reduce the number of false positives. During the first training, we extracted all lesion voxels as positives and randomly selected an equal number of negative voxels within the brain area. Then, this first trained model used to produce initial segmentation masks. The second training was done in a similar way, but all the negative samples were selected from the wrong classified voxels

in the initial segmentation masks. Accordingly, the number of false positives were reduced and better WMH lesion segmentation was achieved.

5.3.2 Effect of histogram loss

In contrast to the transfer learning strategy, in our UDA method, the fully connected layers of the network do not agree with the changes in the implicit features extracted from the convolutional layers. Instead, they are forced to produce similar outputs to the target domain. Moreover, due to the required weight updates in the convolutional layers, unlike in the transfer learning approach, they were set trainable, i.e. not frozen.

The domain adaptation was achieved using the histogram loss, which was applied to the deep convolutional layers and all the fully connected layers of the network. The first two convolutional layers were not included because the activation maps of the early convolutional layers represent basic features as edges and blobs. Minimising the feature map differences for the target and source allowed the network to adapt to a new domain. As confirmed by the results, it was effectively improving the performance of the network from the baseline network model, which did not undergo the domain adaptation. The same behaviour was observed for both of the segmentation tasks.

For the sub-cortical brain structures segmentation, we were able to reach the performance of the unsupervised tool FIRST and for the WMH lesion segmentation our method outperformed the unsupervised LST method, effectively diminishing the performance decline of the baseline model. However, there are some factors that should be taken into account. First of all, the adapted model does not accommodate to the differences in rater. Meaning that although the domain shift effect has been reduced, the consistency of the segmentation mask in terms of over- or under-segmentation will be the same as the rater in the source dataset. Second, the performance of the adapted model is highly dependent on the pre-trained model. If the model's performance before using UDA is low, then it will remain low after the domain adaptation. This means that UDA cannot learn to better classify the input patches to the correct category and its performance is limited to the extent of the pre-trained model.

In our case, the second point above was not the main issue in adapting the network because the pre-trained model's performance is among the state-of-the-art. However, differences in raters between the datasets was one of the main causes that affected the DSC scores in the experimental results. In fact, this is a prevailing issue among unsupervised approaches and one of the reasons for the success of the supervised methods that learn rater specific aspects for a single dataset.

Furthermore, the histogram loss can be integrated to a network in an end-to-end trainable fashion without employing any additional complexities. An alternative approach for UDA is to employ adversarial training method [56, 57], where an extra branch is used to classify domain differences for the implicit features and penalise the network if they are distinct. In doing so, the CNN is enforced to learn domain invariant features. Although the concept of adversarial training is compelling, it comes with some drawbacks. First of all, one has to build another network architecture for the discriminator and optimise its number of parameters to the task. Second, the training could be unstable when either the discriminator or the feature extractor is trained faster than the other [58]. In order to avoid this, the parameters of the feature extractor and the discriminator networks have to be tuned. In contrast to the adversarial domain adaptation, in our method, the training process does not require any subtle parameter tuning.

Chapter 6

Conclusions and future work

In this chapter, we conclude the work accomplished during this PhD, focusing on the main contributions and takeaway messages. Moreover, we discuss on future directions as possible paths of improvements and further developments.

6.1 Summary and contributions

Overall, all the work done in this PhD was to achieve the main objective that included the development of an automated deep learning based method for sub-cortical brain structure segmentation that satisfied the following prerequisites – accuracy, consistency and robustness. In order to achieve this goal, the main objective was divided into several sub-tasks: 1) to develop a method for accurate segmentation of the sub-cortical brain structures; 2) to increase the method’s consistency and robustness using supervised domain adaptation; 3) to decrease manual effort of the method to maintain the consistency and robustness using unsupervised domain adaptation; 4) to validate the method using international and in-house datasets.

The contributions completed for each of the sub-tasks were published or submitted to high ranking international journals and also were contributed to the involved projects at ViCOROB research institute, such as NICOLE, BiomarkEM.cat, EVOLUTION and wASSABI. Moreover, this PhD thesis has been supported by the FI Catalan government grant.

In the following paragraphs, we provide the main conclusions and contributions of this PhD thesis.

- We proposed a novel deep learning method for segmenting all the sub-cortical structures from MR images. The proposed deep learning architecture processed 2.5D patches to extract convolutional features that were combined with

spatial features. The probabilistic atlas inputs provided the network with explicit spatial features that were directly integrated within the network and were trained to be interpreted by the fully connected layers. We have shown the effectiveness of the spatial features in guiding the CNN to segment the difficult areas within the brain that included intensity irregularities. Moreover, we proposed a new sample selective technique where the negatives were extracted from the structure boundaries. This background selection scheme allowed to significantly improve the segmentation performance of the network because the structure boundaries are the most difficult parts to classify. The proposed method showed the best state-of-the-art results in segmenting all the sub-cortical structures and was published in the top ranking journal in the area, *Medical Image Analysis* in June 2018 [JCR CSAI IF 8.880, Q1(5/133)].

- We analysed the effect of domain shift problem on the performance of our deep learning approach for sub-cortical brain structure segmentation. The resulting evidence showed that changes in MRI scanner, imaging protocol such as resolution, slice thickness, and contrast difference have a considerable impact on the performance of the network. A well trained CNN with one imaging domain cannot perform similarly when directly tested with MR images from a different domain. Therefore, we proposed a transfer learning strategy as a solution for adapting the CNN to overcome the domain shift effect. In our proposal, we achieved state-of-the-art results for adapting the network to a new domain by decreasing the number of trainable parameters of the network and remarkably reducing the required training images. Our results showed that using only one and three images from a new target domain were sufficient to significantly ($p < 0.001$) outperform the traditional state-of-the-art method FIRST [41]. Moreover, this study demonstrated that transfer learning was a preferred solution to image standardisation techniques such as histogram matching and mixed training the network with images from different domains. This work accomplished the consistency and robustness requirements of the main objective and was published in *Nature: Scientific Reports* journal in May 2019 [JCR MS IF 4.011, Q1(15/69)].
- The previous work of transfer learning was an effective and preferred way for domain adaptation because it adapts the network not only to a different imaging domain but also to a different rater. However, there is still requirement for at least one or few annotated training images, which still could be a laborious task to delineate all 14 sub-cortical structures. Therefore, in order to minimise the effort for maintaining the consistency and robustness features of our method, we proposed a novel unsupervised method for overcoming the domain shift problem in deep learning. In this proposal, we directly change the convolutional features while training to minimise the differences of the

implicit convolutional features between the original and new target domains. It was achieved by employing the histogram loss that was integrated to the network. The network was end-to-end trainable and did not require any exhaustive hyperparameter tuning. We have tested the proposed method in two different brain segmentation problems to show its robustness and applicability for diverse tasks: 1) sub-cortical brain structure segmentation; and 2) brain White Matter Hyperintensities (WMH) segmentation. For both of the problems, our unsupervised domain adaptation method showed similar or better results than the traditional state-of-the-art methods: 1) FIRST [41] for brain structure; and 2) LST [59] for WMH lesion segmentation. Moreover, it showed significantly ($p < 0.001$) better performance than the baseline network, where the results were obtained without applying domain adaptation. This work has been submitted to the *Knowledge-Based Systems* journal and it is currently under revision [JCR CSAI IF: 5.101, Q1(17/133)].

- Finally, all the research contributions were validated using well-known and publicly available international datasets, and also the in-house dataset with MRI scans and manual annotations provided by the Vall d’Hebron Hospital of Barcelona. For the sub-cortical brain structure segmentation, we used the International Grand Challenge and Workshop on Multi-Atlas Labelling [60] MICCAI 2012 and Internet Brain Segmentation Repository ¹. Additionally, the source-code and an easy-to-use application with integrated Graphical User Interface that incorporates all contributions of this PhD thesis are made publicly available for the research community at our research group github repository: https://github.com/NIC-VICOROB/sub-cortical_segmentation.

Moreover, throughout the period of this PhD fellowship, various collaborations have taken place with the other researches of the ViCOROB group. In particular, a review on deep learning in brain medical image analysis [61] and quantitative analysis of patch-based networks for brain tissue segmentation [62]. Moreover, contributions in the International MICCAI challenges were done for infant brain tissue segmentation and adult brain tissue segmentation with presence of multiple sclerosis lesions [63], where our proposed methods were ranked in top positions. Additionally, brain structure parcellation was also applied for autism spectrum disorder detection [5] with machine learning techniques combining structural and functional MRI information. As can be seen in Rakić et al. [5], the proposed automated brain structure segmentation pipeline is the first step for many disease analysis.

¹<https://www.nitrc.org/projects/ibsr>

6.2 Future work

There are some aspects that were not investigated or out of scope of this PhD thesis and left as future work. In this section, we describe possible research lines that could be continuations of this topic.

First of all, the immediate step to be taken is the performance analysis of the deep learning approaches in segmenting the sub-cortical structures in presence of multiple sclerosis or general white matter hyperintensity lesions. This could be an interesting study as it has been done for the traditional methods in González-Villà et al. [64], but was not explored for the deep learning based methods. The proposed unsupervised domain adaptation method done in this PhD thesis makes this evaluation possible because the experiments would be done for multiple sites.

In parallel to the previous point, we are currently performing an analysis of longitudinal changes in volumes of the sub-cortical structures of patients with multiple sclerosis. The goal of this study is to research their correlations with clinical test results such as Expanded Disability Status Scale (EDSS) and neuropsychological assessment outcomes that evaluate cognitive impairment, audio-visual attention, task switching and memory. The final goal is to develop a method for predicting disease progression over time, which is crucial in attending patients within the risk group.

Other steps could be done for improving the unsupervised domain adaptation by incorporating the adversarial domain adaptation together with our proposal. The adversarial loss could be put as an additional term to the histogram loss and by weighting them accordingly a more stable training process could be achieved. However, the rater differences between the datasets is one of the main reasons that make the unsupervised methods perform worse than supervised. This is due to the supervised methods' advantage that learn rater specific aspects for a single dataset. It would be desirable if there was a unified protocol for producing manual annotations that would make the unsupervised deep learning methods the preferred option to the traditional approaches as the former approach offers more flexibility in domain adaptation such as transfer learning.

Another direction that is linked to the domain adaptation would be data augmentation using the Generative Adversarial Networks (GAN) [65, 66]. By employing the GAN paradigm, we could generate training images for a new imaging domain without manual ground truths by using the masks of publicly available datasets. In this way, we could emulate the images of the target domain by having structural similarities obtained by the existing sub-cortical structure masks and domain information from the target images.

The final goal of these studies would be to provide applicable tools that could be used in the clinical practice. The accomplished work in this PhD thesis indeed

paves the way towards this goal and the described future work could solidify the concrete to support the use of Artificial Intelligence in Medical Image Analysis.

Bibliography

- [1] Lingraj Dora, Sanjay Agrawal, Rutuparna Panda, and Ajith Abraham. State-of-the-art methods for brain tissue segmentation: A review. *IEEE Reviews in Biomedical Engineering*, 10:235–249, 2017.
- [2] Sandra González-Villà, Arnau Oliver, Sergi Valverde, Liping Wang, Reyer Zwiggelaar, and Xavier Lladó. A review on brain structures segmentation in magnetic resonance imaging. *Artificial Intelligence in Medicine*, 73:45–69, 2016.
- [3] Antonios Danelakis, Theoharis Theoharis, and Dimitrios A Verganelakis. Survey of automated multiple sclerosis lesion segmentation techniques on magnetic resonance imaging. *Computerized Medical Imaging and Graphics*, 70:83–100, 2018.
- [4] Mina Ghaffari, Arcot Sowmya, and Ruth Oliver. Automated brain tumour Segmentation using multimodal brain scans, a survey based on models submitted to the BraTS 2012-18 challenges. *IEEE Reviews in Biomedical Engineering*, 13:156–168, 2019.
- [5] Mladen Rakić, Mariano Cabezas, Kaisar Kushibar, Arnau Oliver, and Xavier Lladó. Improving the detection of autism spectrum disorder by combining structural and functional mri information. *NeuroImage: Clinical*, 25:102181, 2020.
- [6] Leon Weninger, Oliver Rippel, Simon Koppers, and Dorit Merhof. Segmentation of brain tumors and patient survival prediction: Methods for the brats 2018 challenge. In *International MICCAI Brainlesion Workshop*, pages 3–12. Springer, 2018.
- [7] MK Houtchens, RHB Benedict, R Killiany, J Sharma, Z Jaisani, B Singh, B Weinstock-Guttman, CRG Guttmann, and R Bakshi. Thalamic atrophy and cognition in multiple sclerosis. *Neurology*, 69(12):1213–1223, 2007.
- [8] LW De Jong, K Van der Hiele, IM Veer, JJ Houwing, RGJ Westendorp, ELEM Bollen, PW De Bruin, HAM Middelkoop, MA Van Buchem, and J Van

- Der Grond. Strongly reduced volumes of putamen and thalamus in Alzheimer's disease: an MRI study. *Brain*, 131(12):3277–3285, 2008.
- [9] William Byne, Erin A Hazlett, Monte S Buchsbaum, and Eileen Kemether. The thalamus and schizophrenia: Current status of research. *Acta Neuropathologica*, 117(4):347–368, 2009.
- [10] SH Lee, SS Kim, WS Tae, SY Lee, JW Choi, SB Koh, and DY Kwon. Regional volume analysis of the Parkinson disease brain in early disease stage: Gray matter, white matter, striatum, and thalamus. *American Journal of Neuroradiology*, 32(4):682–687, 2011.
- [11] Elizabeth H Aylward, Ann Marie Codori, Adam Rosenblatt, Meeia Sherr, Jason Brandt, Oscar C Stine, Patrick E Barta, Godfrey D Pearlson, and Christopher A Ross. Rate of caudate atrophy in presymptomatic and symptomatic stages of Huntington's disease. *Movement Disorders*, 15(3):552–560, 2000.
- [12] Michael H Bloch, James F Leckman, Hongtu Zhu, and Bradley S Peterson. Caudate volumes in childhood predict symptom severity in adults with Tourette syndrome. *Neurology*, 65(8):1253–1258, 2005.
- [13] Eric Hollander, Evdokia Anagnostou, William Chaplin, Katherine Esposito, M. Mehmet Haznedar, Elizabeth Licalzi, Stacey Wasserman, Latha Soorya, and Monte Buchsbaum. Striatal volume on magnetic resonance imaging and repetitive behaviors in autism. *Biological Psychiatry*, 58(3):226–232, 2005.
- [14] Virginia Tremols, Anna Bielsa, Joan-Carles Soliva, Carol Raheb, Susanna Carmona, Josep Tomas, Joan-Domingo Gispert, Mariana Rovira, Jordi Fauquet, Adolf Tobeña, Antoni Bulbena, and Oscar Vilarroya. Differential abnormalities of the head and body of the caudate nucleus in attention deficit-hyperactivity disorder. *Psychiatry Research: Neuroimaging*, 163(3):270–278, 2008.
- [15] Stephan Eliez, Christine M. Blasey, Lisa S. Freund, Trevor Hastie, and Allan L. Reiss. Brain anatomy, gender and IQ in children and adolescents with fragile X syndrome. *Brain*, 124(8):1610–1618, 2001.
- [16] Timothy M Ellmore, Ashley J Hood, Richard J Castriotta, Erin F Stimming, Roger J Bick, and Mya C Schiess. Reduced volume of the putamen in rem sleep behavior disorder patients. *Parkinsonism & Related Disorders*, 16(10):645–649, 2010.
- [17] Yasutaka Kubota, Wataru Sato, Takanori Kochiyama, Shota Uono, Sayaka Yoshimura, Reiko Sawada, Morimitsu Sakihama, and Motomi Toichi. Putamen volume correlates with obsessive compulsive characteristics in healthy population. *Psychiatry Research: Neuroimaging*, 249:97–104, 2016.

- [18] Manuela Schuetze, Min Tae M Park, Ivy YK Cho, Frank P MacMaster, M Malar Chakravarty, and Signe L Bray. Morphological alterations in the thalamus, striatum, and pallidum in autism spectrum disorder. *Neuropsychopharmacology*, 41(11):2627–2637, 2016.
- [19] Melissa Lamar, Elizabeth A Boots, Konstantinos Arfanakis, Lisa L Barnes, and Julie A Schneider. Brain structural alterations common to cardiovascular disease risk factors and Alzheimer’s dementia. *Vascular Disease, Alzheimer’s Disease, and Mild Cognitive Impairment: Advancing an Integrated Approach*, page 241, 2020.
- [20] NC Fox, EK Warrington, PA Freeborough, P Hartikainen, AM Kennedy, JM Stevens, and Martin N Rossor. Presymptomatic hippocampal atrophy in Alzheimer’s disease. *Brain*, 119(6):2001–2007, 1996.
- [21] N. Bernasconi, S. Duchesne, A. Janke, J. Lerch, D.L. Collins, and A. Bernasconi. Whole-brain voxel-based statistical analysis of gray matter and white matter in temporal lobe epilepsy. *NeuroImage*, 23(2):717–723, 2004.
- [22] Gerardo Villarreal, Derek A Hamilton, Helen Petropoulos, Ira Driscoll, Laura M Rowland, Jaqueline A Griego, Piyadasa W Kodituwakku, Blaine L Hart, Rodrigo Escalona, and William M Brooks. Reduced hippocampal volume and total white matter volume in posttraumatic stress disorder. *Biological Psychiatry*, 52(2):119–125, 2002.
- [23] Noriyuki Kitayama, Viola Vaccarino, Michael Kutner, Paul Weiss, and J Douglas Bremner. Magnetic resonance imaging (MRI) measurement of hippocampal volume in posttraumatic stress disorder: A meta-analysis. *Journal of Affective Disorders*, 88(1):79–86, 2005.
- [24] J Douglas Bremner, Meena Narayan, Eric R Anderson, Lawrence H Staib, Helen L Miller, and Dennis S Charney. Hippocampal volume reduction in major depression. *The American Journal of Psychiatry*, 157(1):115–118, 2000.
- [25] Lori L Altshuler, George Bartzokis, Thomas Grieder, John Curran, and Jim Mintz. Amygdala enlargement in bipolar disorder and hippocampal reduction in schizophrenia: An MRI study demonstrating neuroanatomic specificity. *Archives of General Psychiatry*, 55(7):663–664, 1998.
- [26] SM Strakowski, MP Delbello, and CM Adler. The functional neuroanatomy of bipolar disorder: A review of neuroimaging findings. *Molecular Psychiatry*, 10(1):105–116, 2005.

- [27] Hilary P Blumberg, Joan Kaufman, Andrés Martin, Ronald Whiteman, Jane Hongyuan Zhang, John C Gore, Dennis S Charney, John H Krystal, and Bradley S Peterson. Amygdala and hippocampal volumes in adolescents and adults with bipolar disorder. *Archives of General Psychiatry*, 60(12):1201–1208, 2003.
- [28] Stephen M Lawrie, Heather C Whalley, Dominic E Job, and Eve C Johnstone. Structural and functional abnormalities of the amygdala in schizophrenia. *Annals of the New York Academy of Sciences*, 985(1):445–460, 2003.
- [29] Michael P Milham, Allison C Nugent, Wayne C Drevets, Daniel S Dickstein, Ellen Leibenluft, Monique Ernst, Dennis Charney, and Daniel S Pine. Selective reduction in amygdala volume in pediatric anxiety disorders: A voxel-based morphometry investigation. *Biological Psychiatry*, 57(9):961–966, 2005.
- [30] Lori L Altshuler, George Bartzokis, Tom Grieder, John Curran, Tanya Jimenez, Kristin Leight, Jeffery Wilkins, Robert Gerner, and Jim Mintz. An MRI study of temporal lobe structures in men with bipolar disorder or schizophrenia. *Biological Psychiatry*, 48(2):147–162, 2000.
- [31] Stephen M Strakowski, Melissa P DelBello, Kenji W Sax, Molly E Zimmerman, Paula K Shear, John M Hawkins, and Eric R Larson. Brain magnetic resonance imaging of structural abnormalities in bipolar disorder. *Archives of General Psychiatry*, 56(3):254–260, 1999.
- [32] Simon JA van den Bogaard, Eve M Dumas, Tanka P Acharya, Hans Johnson, Douglas R Langbehn, Rachael I Scahill, Sarah J Tabrizi, Mark A van Buchem, Jeroen van der Grond, Raymund AC Roos, et al. Early atrophy of pallidum and accumbens nucleus in Huntington’s disease. *Journal of Neurology*, 258(3):412–420, 2011.
- [33] Nicolas Carriere, Pierre Besson, Kathy Dujardin, Alain Duhamel, Luc Defebvre, Christine Delmaire, and David Devos. Apathy in Parkinson’s disease is associated with nucleus accumbens atrophy: a magnetic resonance imaging shape analysis. *Movement Disorders*, 29(7):897–903, 2014.
- [34] Cecilie B Hartberg, Kjetil Sundet, Lars M Rimol, Unn K Haukvik, Elisabeth H Lange, Ragnar Nesvåg, Ingrid Melle, Ole A Andreassen, and Ingrid Agartz. Subcortical brain volumes relate to neurocognition in schizophrenia and bipolar disorder and healthy controls. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 35(4):1122–1130, 2011.
- [35] Massimo Filippi, Maria A Rocca, Olga Ciccarelli, Nicola De Stefano, Nikos Evangelou, Ludwig Kappos, Àlex Rovira, Jaume Sastre-Garriga, Mar Tintoré,

- Jette L Frederiksen, et al. MRI criteria for the diagnosis of multiple sclerosis: MAGNIMS consensus guidelines. *The Lancet Neurology*, 15(3):292–303, 2016.
- [36] Cecilie Jacobsen, Jesper Hagemeyer, Kjell-Morten Myhr, Harald Nyland, Kirsten Lode, Niels Bergsland, Deepa P Ramasamy, Turi O Dalaker, Jan Petter Larsen, Elisabeth Farbu, et al. Brain atrophy and disability progression in multiple sclerosis patients: a 10-year follow-up study. *Journal of Neurology, Neurosurgery & Psychiatry*, pages jnnp–2013, 2014.
- [37] Nancy C Andreasen, Dawei Liu, Steven Ziebell, Anvi Vora, and Beng-Choon Ho. Relapse duration, treatment intensity, and brain tissue loss in schizophrenia: a prospective longitudinal mri study. *American Journal of Psychiatry*, 170(6):609–615, 2013.
- [38] Rebecca C Knickmeyer, Sylvain Gouttard, Chaeryon Kang, Dianne Evans, Kathy Wilber, J Keith Smith, Robert M Hamer, Weili Lin, Guido Gerig, and John H Gilmore. A structural mri study of human brain development from birth to 2 years. *Journal of Neuroscience*, 28(47):12176–12182, 2008.
- [39] Hanzhang Lu, Lidia M Nagae-Poetscher, Xavier Golay, Doris Lin, Martin Pomper, and Peter CM Van Zijl. Routine clinical brain mri sequences for use at 3.0 tesla. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 22(1):13–22, 2005.
- [40] Bruce Fischl, David H Salat, Evelina Busa, Marilyn Albert, Megan Dieterich, Christian Haselgrove, Andre Van Der Kouwe, Ron Killiany, David Kennedy, Shuna Klaveness, et al. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron*, 33(3):341–355, 2002.
- [41] Brian Patenaude, Stephen M Smith, David N Kennedy, and Mark Jenkinson. A Bayesian model of shape and appearance for subcortical brain segmentation. *NeuroImage*, 56(3):907–922, 2011.
- [42] Kaisar Kushibar, Sergi Valverde, Sandra González-Villà, Jose Bernal, Mariano Cabezas, Arnau Oliver, and Xavier Lladó. Supervised domain adaptation for automatic sub-cortical brain structure segmentation with minimal user interaction. *Scientific Reports*, 9(1):6742, 2019.
- [43] Rolf A Heckemann, Joseph V Hajnal, Paul Aljabar, Daniel Rueckert, and Alexander Hammers. Automatic anatomical brain mri segmentation combining label propagation and decision fusion. *NeuroImage*, 33(1):115–126, 2006.
- [44] Xabier Artaechevarria, Arrate Munoz-Barrutia, and Carlos Ortiz-de Solórzano. Combination strategies in multi-atlas image segmentation: application to brain mr data. *IEEE transactions on medical imaging*, 28(8):1266–1277, 2009.

-
- [45] Hongzhi Wang and Paul A Yushkevich. Groupwise segmentation with multi-atlas joint label fusion. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 711–718. Springer, 2013.
- [46] Pierrick Coupé, José V Manjón, Vladimir Fonov, Jens Pruessner, Montserrat Robles, and D Louis Collins. Patch-based segmentation using expert priors: Application to hippocampus and ventricle segmentation. *NeuroImage*, 54(2):940–954, 2011.
- [47] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.
- [48] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.
- [49] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *Cognitive Modeling*, 5(3):1, 1988.
- [50] Spyridon Bakas, Mauricio Reyes, Andras Jakab, Stefan Bauer, Markus Rempfler, Alessandro Crimi, Russell Takeshi Shinohara, Christoph Berger, Sung Min Ha, Martin Rozycki, et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. *arXiv preprint arXiv:1811.02629*, 2018.
- [51] Konstantinos Kamnitsas, Christian Ledig, Virginia FJ Newcombe, Joanna P Simpson, Andrew D Kane, David K Menon, Daniel Rueckert, and Ben Glocker. Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. *Medical image analysis*, 36:61–78, 2017.
- [52] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. In *European conference on computational learning theory*, pages 23–37. Springer, 1995.
- [53] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 2009.
- [54] László G Nyúl, Jayaram K Udupa, and Xuan Zhang. New variants of a method of MRI scale standardization. *IEEE Transactions on Medical Imaging*, 19(2):143–150, 2000.
- [55] Sergi Valverde, Mariano Cabezas, Eloy Roura, Sandra González-Villà, Deborah Pareto, Joan C Vilanova, Lluís Ramió-Torrentà, Àlex Rovira, Arnau Oliver, and Xavier Lladó. Improving automated multiple sclerosis lesion segmentation with

- a cascaded 3D convolutional neural network approach. *NeuroImage*, 155:159–168, 2017.
- [56] Konstantinos Kamnitsas, Christian Baumgartner, Christian Ledig, Virginia Newcombe, Joanna Simpson, Andrew Kane, David Menon, Aditya Nori, Antonio Criminisi, Daniel Rueckert, et al. Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. In *International Conference on Information Processing in Medical Imaging*, pages 597–609. Springer, 2017.
- [57] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. In *Domain Adaptation in Computer Vision Applications*, pages 189–209. Springer, 2017.
- [58] Kevin Roth, Aurelien Lucchi, Sebastian Nowozin, and Thomas Hofmann. Stabilizing training of generative adversarial networks through regularization. In *Advances in Neural Information Processing Systems*, pages 2018–2028, 2017.
- [59] Paul Schmidt, Christian Gaser, Milan Arsic, Dorothea Buck, Annette Förstler, Achim Berthele, Muna Hoshi, Rüdiger Ilg, Volker J Schmid, Claus Zimmer, et al. An automated tool for detection of FLAIR-hyperintense white-matter lesions in multiple sclerosis. *NeuroImage*, 59(4):3774–3783, 2012.
- [60] B Landman and S Warfield. MICCAI 2012 workshop on multi-atlas labeling. In *Medical Image Computing and Computer Assisted Intervention Conference*, 2012.
- [61] Jose Bernal, Kaisar Kushibar, Daniel S Asfaw, Sergi Valverde, Arnau Oliver, Robert Martí, and Xavier Lladó. Deep convolutional neural networks for brain image analysis on magnetic resonance imaging: a review. *Artificial Intelligence in Medicine*, 95:64–81, 2019.
- [62] Jose Bernal, Kaisar Kushibar, Mariano Cabezas, Sergi Valverde, Arnau Oliver, and Xavier Lladó. Quantitative analysis of patch-based fully convolutional neural networks for tissue segmentation on brain magnetic resonance imaging. *IEEE Access*, 7:89986–90002, 2019.
- [63] Jose Bernal, Mostafa Salem, Kaisar Kushibar, Albert Clèrigues, Sergi Valverde, Mariano Cabezas, Sandra González-Villà, Joaquim W Salvi, Arnau Oliver, and Xavier Lladó. MR brain segmentation using an ensemble of multi-path u-shaped convolutional neural networks and tissue segmentation priors. *Accessed: Feb, 20:2019*, 2018.

- [64] Sandra González-Vilà, Sergi Valverde, Mariano Cabezas, Deborah Pareto, Joan C Vilanova, Lluís Ramió-Torrentà, Àlex Rovira, Arnau Oliver, and Xavier Lladó. Evaluating the effect of multiple sclerosis lesions on automatic brain structure segmentation. *NeuroImage: Clinical*, 15:228–238, 2017.
- [65] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27, pages 2672–2680. Curran Associates, Inc., 2014.
- [66] Xin Yi, Ekta Walia, and Paul Babyn. Generative adversarial network in medical imaging: A review. *Medical Image Analysis*, page 101552, 2019.