Structural Medical Image Analyses using Consistent Volume and Surface Image Processing

By

Yuankai Huo

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Electrical Engineering

May 11, 2018

Nashville, Tennessee

Approved:

Bennett A. Landman, Ph.D.

Benoit M. Dawant, Ph.D.

Richard Alan Peters, Ph.D.

Hakmook Kang, Ph.D.

Richard G. Abramson, M.D.

Warren D. Taylor, M.D.

ACKNOWLEDGEMENTS

The four year's Ph.D. study and research experience has been the most excited journey in my life so far. When looking back this adventure, I have so many people to appreciate. First, I want to thank my wife Ge Liu, who has been a major source of my power and my precious since we started dating in 2005. I could not even have enough courage to start this journey without you, I LOVE YOU! I am grateful to my family, my father Yuejin Huo, my mother Jianya Zhang. You are the most selfless parents in the world and I feel so proud to be your son. I am grateful to my father in law Zhiyong Liu, my mother in law Xiaoxia Sun. You raised such a wonderful daughter and I appreciate you allow her to merry with me. I would like to thank my two daughters, Jessica Huo and Lillian Huo, who make me happy and peaceful. You make me understand the broader meaning of my life and thank you to be my and your mom's "en ai bao bei".

I want to thank my Ph.D. advisor Bennett Landman. You changed my life and I cannot be luckier than working with you. I appreciate you not only forgive the stupid mistakes I have done during my first two years, but also encourage me to be a better Yuankai. You are one of the few people that can match my highest rating for a human, "make sense", which looks easy but actually incredible difficult to be.

I want to thank the folks at Vanderbilt University who have helped me. Zhoubing Xu, thank you for taking care of me during my Ph.D. You are such a humble, smart and nice friend, which make everyone likes you. Shunxing Bao, thank you for helping me for both my career and personal life. You have been one of my best friend at Nashville. I am grateful to my lab folks, Andrew Asman, Andrew Plassard, Robert Harrigan, Benjamin Yvernault, Stephen Damon, Justin Blaber, Shikha Chaganti, Prasanna Parvathaneni, Peijun Hu, Camilo Bermudez, Vishwesh Nath, Allison Hainline, Colin Hansen, Hyeonsoo Moon, Sandra González-Villà, Ilwoo Lyu. I am grateful and thank you for your help.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

Chapter I. Introduction

1. Overview

Medical imaging refers to the technologies of creating visual representation of the interior of human body for scientific research and clinical analysis. Different imaging technologies (modalities) provide different properties, which enables us to investigate human body using particular field of view (FOV) and image contrast [2]. The history of medical imaging can be traced back to the discovery of X-ray in 1895 by Wilhelm Conrad Roentgen, who took the first X-ray on his wife's hand [3]. Since then, many milestones have been made to enable new modalities and devices that we are performing currently (e.g., Ultrasound, Computed Tomography (CT), Magnetic Resonance Imaging (MRI), etc.) [2].

Two-dimensional (2D) or three-dimensional (3D) medical images are the major outcomes from medical imaging techniques. Based on such images, clinical practitioners can make diagnoses by visually investigating the medical images, which relies heavily on the experts' experiences. To provide more information for clinical diagnoses and enable the scientific research, quantitative metrics are extracted from the qualitative images, which results in the research field called medical image analysis (MIA) [4]. To derive quantitative information from medical images, the expert manual delineation has been regarded as the "gold standard" due to the high reliability [5]. However, the manual delineation is resource and time consuming even with the advanced image-guided interactive tools [6]. Therefore, the fully-automated medical image processing is appealing for extracting quantitative metrics from qualitative medical images.

Medical image analysis is an interdisciplinary field of engineering, computer vision, mathematics, data science and medicine, which focuses on the computational analysis of the acquired medical image rather than image acquisition (medical imaging) [4]. The computational methods in MIA can be categorized to two parts (Figure I.1). The first part is the image processing (Figure I.1a), which uses mathematical and computational models to extract quantitative information or metrics from medical images. Representative image processing approaches are preprocessing [7], registration [8], segmentation [5], surface reconstruction [9], etc. The second part is called data analyses (Figure I.1b), which investigates and

understands the hidden regularities behind the metrics extracted by image processing. The common data analyses approaches include statistical analysis [10], visualization [11], modality specific computing [7], etc.

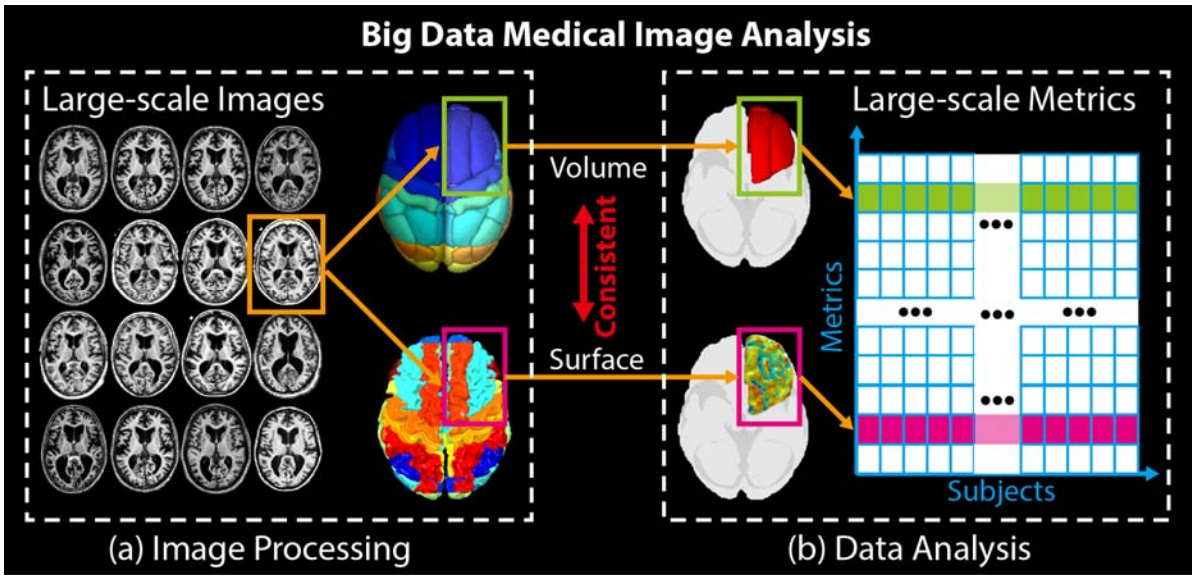


Figure I.1. The principle of Big Data Medical Image Analysis, which contains (1) large-scale image processing, and (2) large-scale data analysis. The focus of the dissertation is to provide a Big Data medical image analysis solution, which including large-scale image processing methods, consistent segmentation and surface reconstruction, inter-subject variation control, and large-scale data analysis. Then, we deploy the entire pipeline to understand the lifespan brain aging as an example.

Historically, the medical image analysis on structural images was limited to a small-scale cohort (e.g., <500 images), whose images were collected from a single scanner (site) (e.g., [12-23]). The rationales of using a small cohort are that (1) it is difficult for a single lab to collect a large-scale cohort (e.g., >5000 images) considering the time and resource consumption. (2) There are the difficulties in data sharing and collaborations between different institutes (e.g., need for approval from institutional review board (IRB)). (3) The image quality and homogeneity are easier to control by using small-scale image cohort collected from a single scanner.

In the past decade, advancements in data sharing and robust processing have made available considerable quantities of brain images all over the world, which has been changing the way of performing medical image analysis to the Big Data (large-scale) fashion [23, 24]. The recent special issue of 20th

anniversary of the Medical Image Analysis journal (MedIA) demonstrates this challenge and opportunity in the first paragraph of "Future Directions" chapter "Big data is becoming a reality with very large scale imaging projects underway or planned. This new scale of data is enabling the solution of challenging problems where the simplicity of methods can offset by the quantity of data available. There are very exciting opportunities at the interface of MIA and the field of Medical Informatics; however there a very few people currently working in both areas."[25].

The large-scale medical images are typically collected from multiple sites, which leads to the greater inter-subject variations than traditional small-scale cohorts. For instance, it is important to rectify the inter-subject variations in acquisitions, scanning protocols, scanner differences, population variations etc. in Big Data image analysis. Existing efforts on reconciling such variations are to (1) standardize the format and of data sharing[26], (2) perform meta-analysis using more data [27-29], (3) propose advances in image processing algorithms [30, 31]. However, given the fact that "The development of large-scale medical image analysis algorithms has lagged greatly behind the increasing quality (and complexity) of medical images and the imaging modalities themselves" [23], there is an urgent demand to develop large-scale image processing frameworks for the robust and timely medical image analysis [23].

Herein, new image processing methods and data mining approaches, compatible for the large-scale scenario, are required to (1) reduce the computational time for large-scale image processing, (2) achieve robust and consistent volume and surface metrics, (3) reconcile inter-subject variations for large cohorts, (4) perform large-scale data analysis using the metrics derived from image processing, and (5) to be robust for variations on intensities and contrasts for multi-site scans.

Beside the methodological challenges, applying large-scale medical image analysis techniques on investigating clinical and research problems leaves many rooms for researchers to fill in. Recent works have demonstrated the advantage of conducting large-scale medical image analysis in understanding prevalent human disorders [32], brain connectivity [33], psychiatric disorder [34] etc. Yet, only limited works have been conducted on investigating lifespan human brain aging, an essential topic in neurological research and clinical investigation, using Big Data medical images. Historically, age-related changes have

been studied in detail for specific age ranges (e.g., early childhood, teen, young adults, elderly, etc.) or more sparsely sampled for wider considerations of lifetime. Contemporaneous advancements [23, 24] in data sharing have made considerable quantities of brain images available from normal, healthy populations, which enable availability of the Big Data for investigating lifespan human aging.

Another interesting application of performing large-scale image processing methods is to explore the anatomies of abdomen organs. For instance, accurate non-invasive spleen volumetric size estimation plays an essential role in splenomegaly diagnosis and scientific studies [35]. Ultrasound [36-38] and computerized tomography (CT) [39-41] have been widely used in the spleen segmentation, yet, limited studies have been applied to magnetic resonance imaging (MRI) [42-44]. A major challenge of automated MRI spleen segmentation is that the absolute intensity of MRI is not in a quantitative scale like the Hounsfield Units (HU) in CT. Another challenge is that the relative intensity contrasts of abdominal tissues are in large variation using the different contrast mechanisms (e.g., T1-weighted (T1w), T2-weighted (T2w), proton density (PD), etc.). Such challenges hinder frequently used CT segmentation methods, which depend on absolute intensity scales, to be applied on the large-scale MR cohorts directly. Another direction is to model pyelocalyceal anatomy for the kidney, which can also influence the success rate of various treatment modalities of kidney stone. The traditional methods of deriving such quantitative measurements have relied on 2-dimensional images of a 3-dimensional system as well as manual delineations, which are both cumbersome and potentially inaccurate during treatment planning.

Herein, we present several new methods to address key technical challenges in large-scale medical image analyses and integrated such methods to investigate lifespan brain aging and abdominal image evaluation.

2. Challenges in Large-scale Image Analysis

The increasing demands of imaging-based diagnosis and rapid developments of the advanced medical imaging techniques lead to the rapid growth of imaging data produced by hospitals and institutes [23]. Only in the past decade, the worldwide clinical and scientific collaboration has provided hundreds of

terabytes of data, which has been made publicly available [24]. The dramatic increasing in the volume and dimension of the medical images results in the challenges of image storage, processing and analysis [23, 24]. However, new clinical and scientific opportunities are arising to explore the valuable information from the large-scale data [23, 24]. Ideally, the automated medical image analysis algorithms are the key to extract biomarkers (biometrics) efficiently and robustly [23, 24]. However, since traditional medical image analysis techniques historically designed for smaller cohorts, new challenges emerge when deploying the existing methods under large-scale scenarios [23-25]. This situation leads to the high demands of novel medical imaging processing and data analyses algorithms, which are able to deal with the unprecedented large-scale datasets [23-25].

2.1. Large-scale Brain Image Processing

Image segmentation and surface reconstruction are two essential methods in large-scale brain image processing. Image segmentation is a computational procedure that assign a distinct label for every voxel in the digital medical images [5]. The representative image segmentation approaches include, but not limit to, threshold based segmentation [45], C-means clustering [46], deformable models [47], graph cuts [48], shape model [49], appearance model [50], learning based model [51], atlas-based segmentation [52-54] etc. Using image segmentation, we are able to derive volume based metrics (e.g., volume size, shape, momentum etc.) of each ROI. Surface reconstruction is another fundamental image processing approach, whose aim is to reconstruct the surfaces of different ROIs based on segmentation and deformable model. The typical surface reconstruction tools include, FreeSurfer [55], CRUISE [56], BrainVISA [57] etc. From the surface reconstruction, the surface based metrics (e.g., surface area, thickness, curvature etc.) are derived.

In large-scale image processing, we not only want to achieve the higher sensitivity from each individual subject compared with traditional image processing, but also want to achieve higher robustness of segmentation and surface reconstruction across the large-cohort. Historically, the image segmentation and cortical reconstruction are typically conducted independently, which may lead to inconsistent metrics

from two procedures. Such spatial inconsistences can hinder the simultaneous usages of volume and surface features in large-scale data analyses. There are limited reports of methods [58-60] for consistent whole brain volumetric segmentation and cortical surface reconstruction.

Another challenge in large-scale medical image analysis is the large inter-subject variations. Unlike the traditional small-scale image analysis, whose variations are typically well controlled by an individual institute. Larger inter-subject variations need to be controlled, at least alleviated, in the large-scale scenarios. To control the inter-subject variations, the total intracranial volume (TICV) has been widely used as a covariate in brain volumetric analyses [61-67]. Compared with whole brain volume (WBV) [68], TICV is often preferred since it provides an estimation of premorbid brain size [69, 70]. Historically, the existing methods performed TICV estimation only used a single affine registration. To reconcile the large inter-subject variability, Commowick et al. proposed to build a personal specific anatomical atlas for head and neck [71]. However, this framework cannot be directly applied to establish probabilistic atlases since each probabilistic atlas is averaged from a group of segmentations.

2.2. Large-scale Image Analysis

Image processing provides large-scale measurements/features (e.g., volume, surface, TICV) from big medical image cohorts [72]. Then data analysis used such measurements to explore the hidden regularities behind the images, which is related to data mining [23-25, 73]. The next challenge is to explore the large-scale metrics by either developing new or adapting existing computational and statistical models. However, traditional image analysis methods can yield less optimal performance for the large-scale challenge. Taking the lifespan brain volume trajectory as an example, prevalent analysis approaches have had difficulties addressing (1) complex volumetric developments on the large cohort across the life time (e.g., beyond cubic age trends), (2) accounting for confound effects, and (3) maintaining an analysis framework consistent with the general linear model (GLM) approach pervasive in neuroscience.

2.3. Computational Efficiency

For the traditional atlas-based segmentation methods, the Big Data also bring the considerable

issues such as higher demands on computational resources and time. To alleviate the computational complexities, learning based algorithms have been successfully employed to speed up the labeling process including, but not limited to, SVMs [74-76], random forest[77, 78], artificial neural networks [74, 79], logistic LASSO [80] and boosting [75]. Unfortunately, the previous learning-based schemes are mostly limited to single anatomical region segmentation rather than whole brain. When applied on whole brain, the computational expensive non-rigid registration is typically required to alleviate large inter-subject variation.

2.4. Large-variations for the Abdomen

The last challenge is that most of the prevalent medical image analysis approaches are historically designed for neuro images, which hinders us to apply such methods (e.g., preprocessing, registration, multi-atlas label fusion) to abdomen directly [81]. A major reason is that the abdomen has greater heterogeneity than the brain. Moreover, the locations of abdominal viscera for same subject can change obviously between two scans. For inter-subject variations, the heterogeneity is even greater. For instance, the spleen size of splenomegaly cohort varies from 368 cubic centimeter (cc) to 5670 cc reported by [82].

3. Context for Advancing Large-scale Image Processing

Hundreds of secondarily derived biomarkers and biometrics can be extracted from a single medical image using advanced medical image processing methods, which allows the researchers to explore the hidden spatial and temporal relationships from large-scale dataset. We first introduce the multi-atlas segmentation (MAS) theory, then present two new techniques based on multi-atlas principle: (1) large-scale multi-atlas learner fusion (reduces the computational time), and (2) consistent multi-atlas segmentation and surface reconstruction (provides consistent volume and surface). Then, to reconcile the inter-subject variations, the data-driven probabilistic atlas and total intracranial volume estimation methods are introduced.

3.1. Multi-atlas Segmentation

Among segmentation methods, atlas-based segmentation is one of the most prominent families, which uses a pairing of structural MR scans and corresponding manual segmentation. In atlas-based segmentation models, an existing dataset (atlas) is spatially transferred to a previously unseen target image through deformable registration. Single-atlas segmentation has been successfully applied to some applications [83-85]. Yet, more recent approaches employ a multi-atlas paradigm as the de facto standard atlas-based segmentation framework [86, 87]. In multi-atlas segmentation, the typical framework is: (1) a set of labeled atlases are non-rigidly registered to a target image [8, 88-90], and (2) the resulting label conflicts are resolved using label fusion [87, 91-99].

The most prevalent multi-atlas label fusing theory has been developed to model the spatial relationships between atlases and targets in 3D scenarios. To improve the performance of 4D MAS for longitudinal data, we propose a novel longitudinal label fusion theory, called 4D joint label fusion (4DJLF) to incorporate the probabilistic model of temporal performance of atlases to the voting-based fusion.

3.2. Multi-atlas Learner Fusion

One major concern of applying multi-atlas segmentation framework on Big Data is the computational complexity as it typically takes over 24 hours for more than ten non-rigid registrations and the following multi-atlas label fusion. To decrease overall computational complexity, new approaches have emerged to minimize registration time. One of the most common methods is the atlas selection [97, 100-102], which reduces the times of registration by keeping the most representative atlases. Another direction is to use the learning based scheme, which grasps the non-local correspondences offline [91-93, 103, 104]. Advanced by the large-scale images, we present multi-atlas learner fusion (MLF), a framework for replicating the robust and accurate multi-atlas segmentation model, while dramatically lessening the computational burden.

3.3. Consistent Multi-atlas Volume and Surface Computing

Whole brain volume segmentation and cortical reconstruction has been typically considered as

independent processing in neuroimaging [105-110]. As a result, such spatial inconsistences can further hinder the consistent brain morphometry analyses. There are limited reports of methods for consistent whole brain volumetric segmentation and cortical surface reconstruction [58-60, 106, 111]. In this dissertation, we presented the multi-atlas CRUISE (MaCRUISE) method to achieve consistent whole brain segmentation and cortical surface reconstruction.

3.4. Big Data Driven Probabilistic Atlas

Probabilistic atlases are essential in understanding the spatial variation of brain anatomy, in visualization, and data processing. However, inter-subject variability is normally greater than inter-group variability, which hinders group-based atlases to capture individual variation. Advanced by large-scale training images, we presented a large-scale data-driven framework to learn a dictionary of the whole brain probabilistic atlases (132 regions) from 1888 heterogeneous 3D MRI training images.

3.5. Total Intracranial Volume Estimation

TICV is a widely used metric to reconcile inter-subject variations in neuro imaging, which is estimated by the volume inside the brain cranium including gray matter (GM), white matter (WM), cerebrospinal fluid (CSF) and meninges [112]. To derive accurate TICV estimation from brain MRI scan, a number of approaches have been developed and evaluated [113-122] [106] [117]. However, none of them estimate TICV by counting the voxels inside skull, which is a natural way of calculating TICV. In this dissertation, we present a multi-atlas based TICV estimation method using Non-Local Spatial STAPLE (NLSS) which is more accurate than previous methods and consistent with whole brain multi-atlas segmentation.

4. Large-scale Data Analysis

The large-scale data analysis has been broadly applied to medical research and healthcare in past decades, which enables us to establish the correlations between qualitative data (e.g., demographic data), quantitative medical records (e.g., laboratory values) [123], and diseases [124]. Different from the medical

records, the large-scale image data analysis has not been widely investigated due the high degree of freedom in big image cohorts.

4.1. Large-scale Multi-site Cohorts

The maturation of medical imaging technologies as well as the image sharing and storage approaches provide the opportunity to deploy large-scale analysis on medical images. Investigating fundamental diseases using multi-scale images [125], as well as multi-site images [126] have been recognized during the past decade. The National Institutes of Health (NIH) National Database of Autism Research (NDAR) ([127], https://ndar.nih.gov/) is a database of understanding the autism disease. The National Institute on Aging's (NIA) Baltimore Longitudinal Study of Aging (BLSA) ([128, 129], https://www.blsa.nih.gov/) is a clinical research programs of understanding aging and aging-related diseases. The collections of functional MRI (fMRI) have been publicly available on both task-based fMRI from OpenfMRI project ([130], https://openfmri.org/) and resting-state fMRI from "1000 Functional Connectomes" project (fcon_1000) ([131], http://fcon_1000.projects.nitrc.org/). Other publicly available cohorts include Information eXtraction from Images (IXI), Open Access Series on Imaging Studies (OASIS) [132] and Multi-Modal MRI Reproducibility Resource (MMMRR)[133].

4.2. Large Inter-subject Variation

For a single study, the medical imaging data may not face the difficulties using existing processing algorithms and statistical method. However, as data sets from different studies, populations and sites are amassed into a large-scale cohort, considerable challenges emerge. For instance, it is challenging of rectifying the inter-subject variations in acquisitions, scanning protocols, scanner differences, population variations etc. Recent efforts on reconciling such variations are to (1) standardize the format and of data sharing[26], (2) perform meta-analysis using more data [27-29], (3) propose advances processing algorithms [30, 31]. However, if any, common solutions are well accepted to perform image analysis by rectifying such variations on large-scale image cohorts [24].

4.3. Lifespan Brain Aging

In the past decade, many efforts have been made of performing Big Data medical image analysis in understanding, but not limited to, Parkinson's disease [32], brain connectivity [33], psychiatric disorder [34]. However, few, if any, works have been done on investigating the lifespan aging, the development of brain structures across lifespan, which is a key topic in understanding neuro-development. Herein, investigating lifespan aging on human brain is an appealing application of integrating the new Big Data medical image processing and analysis approaches. In this dissertation, we propose to investigate the lifespan human brain aging on more than 5,000 MR structural images.

5. Robust Multi-model Abdomen Image Processing

5.1. Atlas-based Splenomegaly Segmentation

Splenomegaly is an abnormal enlargement of the spleen, which is associated with liver disease, infection and cancer [134]. Accurate non-invasive spleen volumetric size estimation plays an essential role in splenomegaly diagnosis and scientific studies [35]. Spleen segmentation using Ultrasound [36-38] and computerized tomography (CT) [39-41] have been used as the major imaging techniques in quantifying spleen size [135, 136]. However, the MRI has not been widely used as the absolute intensity of MRI is not in a quantitative scale like the Hounsfield Units (HU) in CT. Another challenge is that the relative intensity contrasts of abdominal tissues are in large variation using the different contrast mechanisms (e.g., T1-weighted (T1w), T2-weighted (T2w), proton density (PD), etc.). In this dissertation, we propose to use multi-atlas segmentation framework with Bayesian atlas selection and surface constraint on robust multi-contrast MRI spleen segmentation for splenomegaly.

5.2. Deep Learning Based Splenomegaly Segmentation

In recent years, deep learning methods have shown their superior performance on automatic spleen segmentation compared with traditional methods [137]. However, the existing deep learning methods are typically deployed on CT images with normal size spleen (e.g., spleen size < 500 cubic centimeter (cc)).

When dealing with splenomegaly MRI segmentation (e.g., spleen size > 500 cc), we need to deal with large inhomogeneity on intensities of clinical acquired MR and large variations on shape and size of spleen for splenomegaly patients [138]. Recently, global convolutional network (GCN) have shown advantages in sematic segmentation on natural images with large variations by using larger convolutional kernels [139]. Meanwhile, adversarial networks have been proven able to refine the semantic segmentation results [140]. In this dissertation, we propose a new Splenomegaly Segmentation Network (SSNet) to perform the splenomegaly MRI segmentation under the image-to-image framework with the end-to-end training. In SSNet, the GCN is used as the generator while the conditional adversarial network (cGAN) is employed as the discriminator [18].

One major limitation of DCNN methods is that we typically have to manually trace a new set of training data when segmenting new organs or new imaging modalities. For instance, a DCNN trained with normal spleens was not able to capture spatial variations of splenomegaly. Image synthesis has been used to segment images for one modality from another [141-144], yet, paired images were typically required for traditional methods. Recently, two stage methods have been proposed to use cycle generative adversarial networks (CycleGAN) [145] to synthesize training images for a target modality[146, 147]. Then, these efforts trained a segmentation network independently using synthetic images [148]. However, these two independent stages did not use the complementary information between synthesis and segmentation. Herein, we proposed a novel end-to-end synthesis and segmentation network (EssNet) to achieve the unpaired MRI to CT image synthesis and CT splenomegaly segmentation simultaneously without using manual labels on CT.

5.3. Characterization of Pyelocalyceal Anatomy for Kidney

Nephrolithiasis is a costly and prevalent disease that is associated with significant morbidity including pain, infection, and kidney injury. While surgical treatment of kidney stones is generally based on size and quality of the stones, studies have suggested that specific characteristics of pyelocaliceal anatomy, such as the lower pole infundibulopelvic angle, can also influence the success rate of various

treatment modalities [3, 4]. However, the traditional methods of deriving such quantitative measurements have relied on 2-dimensional images of a 3-dimensional system as well as manual delineations, which are both cumbersome and potentially inaccurate during treatment planning [3, 5, 6]. In this dissertation, we propose a novel non-invasive framework that automatically achieves a tree structure of the renal collecting system using CT urograms, allowing for 3-dimensional characterization of the pyelocaliceal anatomy.

6. Contributions

The primary contributions are as follows. In Chapter II we present an efficient whole brain segmentation approach by learning features from large-scale MRI data. In Chapter III we present a novel multi-atlas CRUISE (MaCRUISE) method to combine the multi-atlas whole brain segmentation with brain cortical surface reconstruction. In Chapter IV we present a surface parcellation method to parcellate reconstructed whole brain surfaces to detailed cortical labels. Chapter V presents a novel data-driven method to establish a target image specified probabilistic atlas from large-scale cohorts. Chapter VI presents a novel simultaneous total intracranial volume (TICV) and posterior fossa volume (PFV) segmentation algorithm to achieve better performance than baseline methods. Chapter VII explore the life-span brain volume trajectories on whole brain, network, and region levels on more than 5,000 multi-site MRI brain volumes. The volumetric features were obtained using multi-atlas segmentation and a novel covariate-adjusted restricted cubic spline regression method was proposed to model the non-linear trajectory curves. In Chapter VIII we extend the multi-atlas label theory from 3D to 4D by considering the spatial temporal performance of registered atlases for longitudinal scenario. In Chapter IX we present a novel atlas-selection based segmentation method to perform MRI splenomegaly segmentation. We further leverage the splenomegaly segmentation accuracy by combing deep convolutional neural network and adversarial network in Chapter X. Chapter XI present a synthesis learning based segmentation method to perform splenomegaly segmentation on CT without having CT ground truth. In Chapter XII, we revisit the pyelocalyceal anatomy in management of kidney stone using 3D segmentation methods. Finally, we conclude in Chapter XIII by summarizing contributions and possible future directions.

6.1. Contributions on Brain

- We proposed the MLF framework cuts the runtime on a modern computer from 36 hours down to 3-8 minutes, which accelerate the multi-atlas segmentation on large-scale image. It explores the possibilities and limitations of designing fast whole brain segmentation methods on large-scale training images.
- We designed and implemented MaCRUISE to achieve consistent whole brain segmentation and cortical surfaces. Using MaCRUISE, we achieve 132 volume labels and 98 surface labels from a single T1-weighted (T1w) MRI scan by integrating previous separated multi-atlas segmentation theory and surface reconstruction theory.
- We present a data-driven framework to build a personal specific probabilistic atlas under the large-scale data scheme.
- We proposed a robust TICV estimation method using multi-atlas label fusion, which has been shown to be more accurate than previous methods. We created a set of TICV brain atlases to be publicly available for our community.
- We proposed to use C-RCS regression method within a multi-site cross-sectional framework and revisit the brain volumetry problem using more than 5,000 MR images.
- We proposed 4DJLF under the general label fusion framework by simultaneously incorporating the spatial and temporal covariance on all longitudinal time points, which is a longitudinal generalization of a leading joint label fusion method (JLF) that has proven adaptable to a wide variety of applications.

6.2. Contributions on Abdomen

- We performed the first study on multi-model MRI splenomegaly segmentation with multi-atlas segmentation as well as deep convolutional neural network.
- We compared different strategies for multi-atlas splenomegaly segmentation and proposed the novel L-SIMPLE multi-atlas framework.

- We proposed the SSNet to address spatial variations when segmenting extraordinarily large spleens. SSNet was designed based on the framework of image-to-image conditional generative adversarial networks.
- We introduced a novel end-to-end (EssNet) to achieve the unpaired MRI to CT image synthesis and CT splenomegaly segmentation simultaneously without using manual labels on CT.
- We proposed a novel non-invasive framework that automatically achieves a tree structure of the renal collecting system using computerized tomography (CT) urograms, allowing for 3-dimensional characterization of the pyelocaliceal anatomy.

6.3. Previous Publications

Many contributions of this dissertation have been previously published. A learning based fast multi-atlas segmentation method was introduced [149]. A consistent multi-atlas whole brain segmentation and surface reconstruction pipeline was proposed [150, 151]. A data-driven framework to build a personal specific probabilistic atlas was presented under the large-scale data scheme [152]. A robust method for automatic measurement of the TICV was introduced [153]. A new regression method was proposed to apply on large-scale neuroimages for understanding lifespan brain volumetry [150]. A longitudinal multi-atlas label fusion theory was presented [154]. Splenomegaly segmentation pipelines were proposed using multi-atlas segmentation [82], fully convolutional neural network [155], and synthesis learning [156]. A non-invasive framework was proposed to achieve a tree structure of the renal collecting system using CT [157].

Chapter II. Multi-atlas Learner Fusion: An efficient segmentation approach for large-scale data

1. Introduction

Magnetic resonance (MR) imaging of the brain is an essential diagnostic method in clinical investigation and an effective quantitate method in neurology and neurological research. To explore the complicated relationships between biological structure and clinical diagnosis as well as brain function, segmentation of anatomical structure on MR images has been widely used. Expert manual delineation [158, 159] has been regarded as "gold standard". However, since manual segmentation is extremely resource consuming, automatic methods have been proposed to get robust and accurate segmentation [52-54]. Atlas-based segmentation, which uses a pairing of structural MR scans and corresponding manual segmentation, is one of the most prominent approaches.

In atlas-based segmentation models, an existing dataset (atlas) is spatially transferred to a previously unseen target image through deformable registration. Single-atlas segmentation has been successfully applied on some applications [83-85]. Yet, more recent approaches employ a multi-atlas paradigm as the de facto standard atlas-based segmentation framework [86, 87]. In multi-atlas segmentation, the typical framework is: (1) a set of labeled atlases are non-rigidly registered to a target image [8, 88-90], and (2) the resulting label conflicts are resolved using label fusion [87, 91-99].

Recently, learning based multi-atlas segmentation has emerged from the multi-atlas segmentation as a new family of methods. One field of study deals with the generation of a template library based on the limit set of manual segmentation such as the LEAP algorithm [160] and the MAGeT Brain [161]. Other approaches used group-wise registration and iterative group-wise segmentation such as the MABMIS algorithm [162]. A new algorithm exploited the strengths of both label fusion and statistical classification to get more robust segmentations [163]. Meanwhile, the widely used supervised machine learning algorithms have also been successfully employed including, but not limited to, SVMs [74-76], random

forest[77, 78], artificial neural networks [74, 79], logistic LASSO [80] and boosting [75].

Unfortunately, this robustness of multi-atlas segmentation comes at the cost of computational complexity (CC) because both typical multi-atlas approaches and the learning based methods rely on expensive non-rigid registrations or non-local correspondences calculation. Concisely, we define these two types of computational complexity as (1) the computational complexity of conducting non-rigid registrations (CCNR), and (2) the computational complexity of capturing non-local correspondences (CCNC).

To decrease overall computational complexity without compromising segmentation quality, new approaches have emerged to minimize CCNR. One of the most common methods is the atlas selection [97, 100-102], which reduces the CCNR by keeping the most representative atlases. In recent years, researchers have even tried to eliminate the CCNR by employing non-local label fusion methods [91-93, 103, 104]. However, the reduction of CCNR is typically accompanied with the large increase of CCNC. To minimize the CCNC further, other researchers have attempted to use the learning based scheme, which grasps the non-local correspondences offline [74-79]. Once the model is well trained, it is able to be applied on the target image efficiently. However, these learning-based algorithms are still limited since the learning based models are applied and tested on homogenous small-scale dataset (typically less than 200 subjects from the same resource) without using a great deal of available heterogeneous data (from different resources e.g. different studies and scanners). As a result, the previous learning based schemes are mostly applied on segmenting single anatomical region or subcortical regions rather than whole brain. When applied on whole brain, non-rigid registration (high CCNR) is still essential to compensate the large inter-subject variation for the small size of the dataset.

In this chapter, to eliminate both CCNR and CCNC, we propose the multi-atlas learner fusion (MLF) framework. Due to the large amount of training atlases used in our framework, we are able to provide more candidates for atlas selection and a larger training pool during the learning step, which dramatically leads to reduction of the total computational complexity when segmenting a target image. Particularly, the MLF framework has the following important characteristics.

1. Efficient framework by using large-scale dataset. When the training dataset is large and representative enough (3464 images from 6 projects), the MLF framework is able to find the close trained AdaBoost learners ("close" means with the similar anatomy) for the target image. As a result, the MLF provides a high-speed learning based segmentation framework that only requires 3-8 min to segment a target image by totally eliminating the CCNR and CCNC.
2. The elements of the framework are designed for the large-scale scenario. The PCA is used for low-dimensional projection, which eliminates the computational expensive pairwise similarity measurements (typically required by manifold learning approaches) for thousands of training data (even on larger data sets). The AdaBoost, combined with decision tree, has proved to be an extremely successful in two-class classification (the case this chapter is investigating) and even described as the "best off-the-shelf classifier in the world" [164]. After the training procedure, 3464 AdaBoost learners were trained and a group of the closest learners (with smallest Euclidean distance on PCA low-dimensional space) were applied on each target image.
3. Application of whole brain segmentation. The framework is trained and applied on the whole brain segmentation (133 labels) which is much more complicated than segmenting single anatomical region or subcortical regions.

In the rest of the chapter, we propose a whole-brain (133 labels) multi-atlas segmentation framework using a large-scale data paradigm. Building on seminal works in machine learning (e.g., AdaBoost [165] and Principal Component Analysis (PCA)), we use a learning-based approach to emulate the accuracy of a premier multi-atlas segmentation framework while dramatically lessening the computational burden. Given a large collection of training data which was pre-processed using a state-of-the-art multi-atlas segmentation procedure, we: (1) construct a low-dimensional representation of our training data for computing neighborhood relationships and (2) optimize an AdaBoost classifier for each training image that maps a weak segmentation estimate (e.g., a majority vote of the local neighbors) to the

expensive, yet highly accurate, multi-atlas segmentation estimate. Thus, when a new target image needs to be segmented we simply need to (1) project the image into the low-dimensional space, (2) construct a weak initial segmentation, and (3) fuse the locally selected learners from the training phase. We refer to the algorithm as multi-atlas learner fusion (MLF) (Figure II.1).

2. Data and Pre-Processing

Herein, the complete data aggregates 7 unique datasets covering a wide range of demographics, ages, and neurological states (Table II.1). Data from 1000 Functional Connectome (fcon_1000)[166], Information eXtraction from Images (IXI), Open Access Series on Imaging Studies (OASIS)[132] and Multi-Modal MRI Reproducibility Resource (MMMRR)[133] are publicly available. The Baltimore Longitudinal Study on Aging (BLSA) is the study of aging whose data are collected by the National Institute of Aging [167]. The Deep Brain Stimulation (DBS) data is obtained from the DBS project at Vanderbilt University [168]. The Tennessee Twin Study (TTS) is an ongoing study that examines the health and wellbeing of twins born in Tennessee between 1984 and 1995 [169]. In total, a set of 3505 subjects was scanned resulting in a total of 3886 T1-weighted MR whole-brain volumes. For validation, the data was separated into three groups: training, testing, and reproducibility. First, the MMMRR dataset was used in its entirety as the reproducibility set as it consists of 21 subjects identically scanned twice. The remaining datasets were split 90%/10% into the training/testing cohorts. Note, all intra-subject scans were placed accordingly in the same training/testing group.

In addition, 50 MPRAGE images (from unique subjects) from OASIS dataset were manually labeled with 133 labels by NeuroMorphometrics with BrainCOLOR protocol [170]. Forty five images (from 50 MPRAGE images) were used as the original atlases in multi-atlas segmentation [132]. Meanwhile, 6 randomly selected images (from 45 MPRAGE images) were used for a simulation test. Lastly, the 5 unused images (from the 50 MPRAGE images) were used for an empirical evaluation.

For all 3,886 images, a state-of-the-art multi-atlas segmentation was performed. For consistency, all images were affinely registered [89] to the MNI305 atlas [171]. Practically, 10-20 atlases are sufficient

for a good multi-atlas segmentation [97]. Thus, based on our experience, for each image, the 15 closest atlases were selected (using a naïve PCA projection), pairwise registered [88, 89], fused [92, 172], and corrected through implicit error modeling [173]. On average, this process took 36 h on a modern computer.

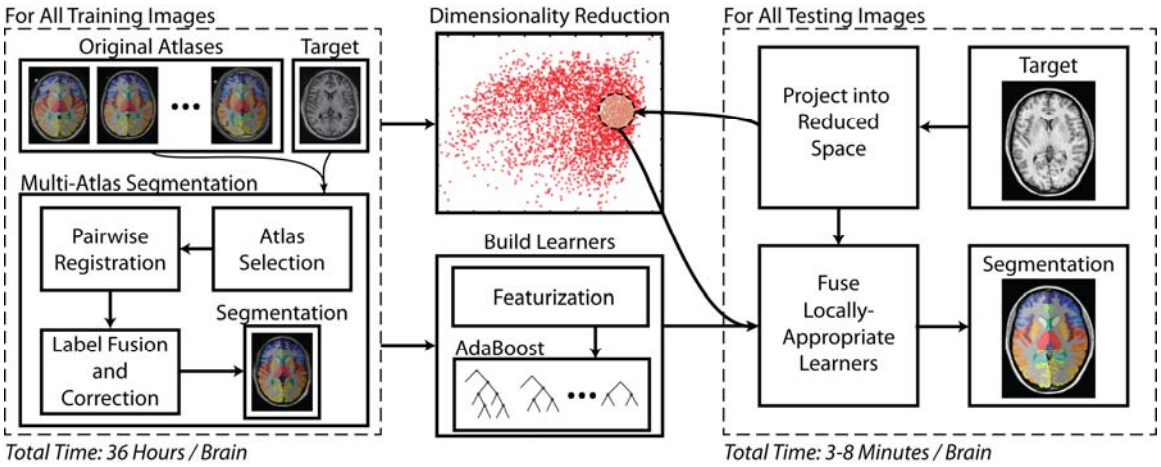


Figure II.1 Flowchart demonstrating the multi-atlas learner fusion (MLF) framework. A large collection of training images is processed offline using a typical multi-atlas segmentation pipeline. The dimensionality of the training images is then reduced, and learners are constructed to map a weak initial estimate to the multi-atlas segmentation. Finally, for a new testing image, the image needs to be projected into the low-dimensional space and the locally appropriate learners can be fused to efficiently and accurately estimate the final segmentation.

Table II.1. Data summary. Each value is represented by: number of subjects (number of images)

	Training	Testing	Repro.
1000 Functional Connectome (fcon_1000)[a]	1055 (1055)	117 (117)	
Baltimore Longitudinal Study on Aging (BLSA)	578 (883)	64 (94)	
Information eXtraction from Images (IXI)[b]	523 (523)	58 (58)	
Deep Brain Stimulation (DBS)	493 (493)	54 (54)	
*Open Access Series on Imaging Studies (OASIS)[c]	375 (392)	41 (44)	
Tennessee Twins Study (TTS)	113 (118)	13 (13)	
Multi-Modal MRI Reproducibility Resource (MMMRR)d			21 (42)
Total:	**3137 (3464)**	**347 (380)**	**21(42)**

a: https://www.nitrc.org/projects/fcon_1000/
b: http://www.oasis-brains.org/
c: http://biomedic.doc.ic.ac.uk/brain-development/
d: https://www.nitrc.org/projects/multimodal
**: With OASIS, 6 subjects are used for simulation and 5 subjects are used for empirical validation.*

Finally, for all 3464 training images, a low-dimensional representation was computed using PCA. Briefly, whole brain anatomical images were down-sampled to 2mm isotropic resolution and only the non-background voxels were used for the PCA analysis. Such voxels were extracted from a non-background mask whose probability of non-background is greater than 0.8. The non-background probability is

represented by a probabilistic map which is obtained by averaging the segmentations (set all non-background regions to 1 and background to 0) defined by the multi-atlas segmentation estimates. Local distances, the pairwise Euclidian distances between any two subjects on low-dimensional PCA domain, are computed using the projection weights onto the first 15 modes of variation (representing 15.33% of the total variation). Notice that the remaining variation (84.67%) might be introduced by the registration error and the large inter-subject variance of brain anatomy. The results of the pre-processing framework are summarized in Figure II.2.

3. Multi-Atlas Learner Fusion Theory

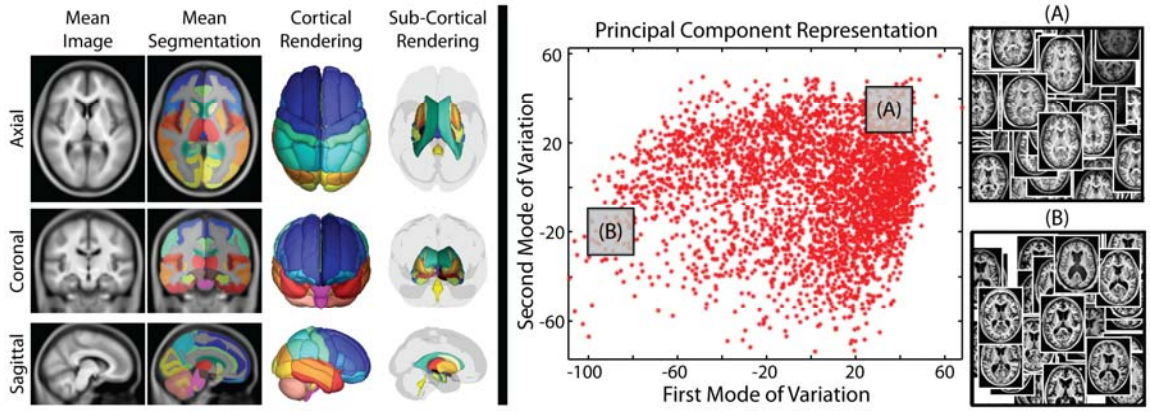


Figure II.2 Summary of the training data processed through multi-atlas segmentation and their corresponding representation in the estimated low-dimensional space. The inlays in (A) and (B) illustrate that the PCA distance metric leads to reasonable clustering of anatomical features.

The theory presented below builds on the foundation for learning-based error correction presented in [173]. For training image j, we assume that we are given (1) the target image, $\boldsymbol{I}_j \in \mathbb{R}^N$, (2) the initial weak segmentation, $\boldsymbol{\Psi}_j \in \boldsymbol{L}^N$, and (3) the multi-atlas segmentation, $\boldsymbol{\Omega}_j \in \boldsymbol{L}^N$, where N is the total number of voxels, and $\boldsymbol{L}$ is the set of possible labels (herein, $|\boldsymbol{L}| = 133$). As in [173], the AdaBoost training procedure is computed for all of the labels independently. For each label, let $\boldsymbol{B}_l$, such that $l \in \boldsymbol{L}$, be the collection of voxels for which any of the training images observe label l.

For the classifier, let the feature matrix be defined as $\boldsymbol{X}^l \in \mathbb{R}^{M \times F}$, such that each element, X^l_{mf}, is

the feature value for feature f at sample m and label l, where F is the number of features, and $M \leq |\boldsymbol{B}_l|$ is the number of samples (or voxels). For simplicity, we define the features at each sample the same way as [173]. Briefly, these consist of the voxel coordinates, the observed labels (i.e., all $\Psi_{ji}\ s.t.\ i \in \boldsymbol{R}_m$), the target intensities (i.e., all $\mathrm{I}_{ji}\ s.t.\ i \in \boldsymbol{R}_m$), and the corresponding spatial correlations where $\boldsymbol{R}_m$is the collection of voxels within the feature window defined for sample m (herein, a 5mm isotropic window centered at the current sample). This feature collection strategy results in a total number of features of $F = 1009$. Finally, we define the class vector as, $\boldsymbol{Y}^l \in \{-1, 1\}^M$, where $Y_m^l = 1$ if $\Omega_{jm} = l$, and $Y_m^l = -1$ otherwise.

For the AdaBoost training, let $\boldsymbol{D}_{jl}^{(t)} \in \mathbb{R}^M$, be the distribution of relative weights for all samples at iteration $t \leq T$ (where $D_{jlm}^{(0)} = \frac{1}{M}$ initially). The goal of the training process at iteration t is to optimize the weak learner, h_{jlt}, where $h_{jlt}[X_m^l] \in \{-1,1\}$

$$h_{jlt} = \arg\max_{h_{jlt}} \left| 0.5 - \sum_m D_{jlm}^{(t)} \left(1 - \delta\left(h_{jlt}[X_m^l], Y_m^l\right)\right) \right| \tag{2.1}$$

where, $\delta(\cdot,\cdot)$ is the Kronecker delta function. Note, herein, the weak learner in (1) is a decision tree and optimization of this learner is addressed later in the manuscript. Next, the weight associated with the current iteration, $\alpha_{jlt} \in \mathbb{R}$, is defined as

$$\alpha_{jlt} = \frac{1}{2} \ln \frac{1 - \sum_m D_{jlm}^{(t)} \left(1 - \delta\left(h_{jlt}[X_m^l], Y_m^l\right)\right)}{\sum_m D_{jlm}^{(t)} \left(1 - \delta\left(h_{jlt}[X_m^l], Y_m^l\right)\right)} \tag{2.2}$$

and the sample weight can be updated with

$$D_{jlm}^{(t+1)} = \frac{1}{Z} \exp\left(\alpha_{jlt} \delta\left(h_{jlt}[X_m^l], Y_m^l\right)\right) \tag{2.3}$$

where Z is a partition function ensuring that $\sum_m D_{jlm}^{(t+1)} = 1$. This process is then iterated until we have reached the desired number of iterations, T (herein, $T = 50$).

Once the training process has been performed on all training images, we can then approximate the desired multi-atlas segmentation through fusing the trained AdaBoost learners associated with the corresponding locally selected training images. If we let $\boldsymbol{J}$ be the set of selected training images, and $\boldsymbol{\Omega}^* \in$

$\boldsymbol{L}^N$ be the approximated multi-atlas segmentation, then Ω_i^* (i.e., the estimated label at voxel i) is computed

$$\Omega_i^* = \arg\max_{l \in \boldsymbol{L}} \sum_{j \in \boldsymbol{J}} \sum_t \alpha_{jlt} h_{jlt}[X_i^l] \tag{2.4}$$

where the feature matrix, $\boldsymbol{X}$, is defined in exactly the same way for the testing image as it was previously defined for the training images.

4. Methods and Results

Throughout, all segmentation comparisons are assessed with the mean Dice Similarity Coefficient (DSC) [174] across the 132 non-background labels, and all claims of statistical significance are made using a Wilcoxon signed rank test ($p < 0.01$) [175]. In Figure II.1 to Figure II.6, the DSC values were calculated in MNI305 space. To compare the label fusion results (in MNI305 space) with the manually labels images (in original space), in Figs. 7 and 8, the DSC values were calculated in original space by affinely transferred the label fusion results to each subject's original space. Here, the 4x4 affine matrices were the inverse matrices which were generated during the affine registration in preprocessing.

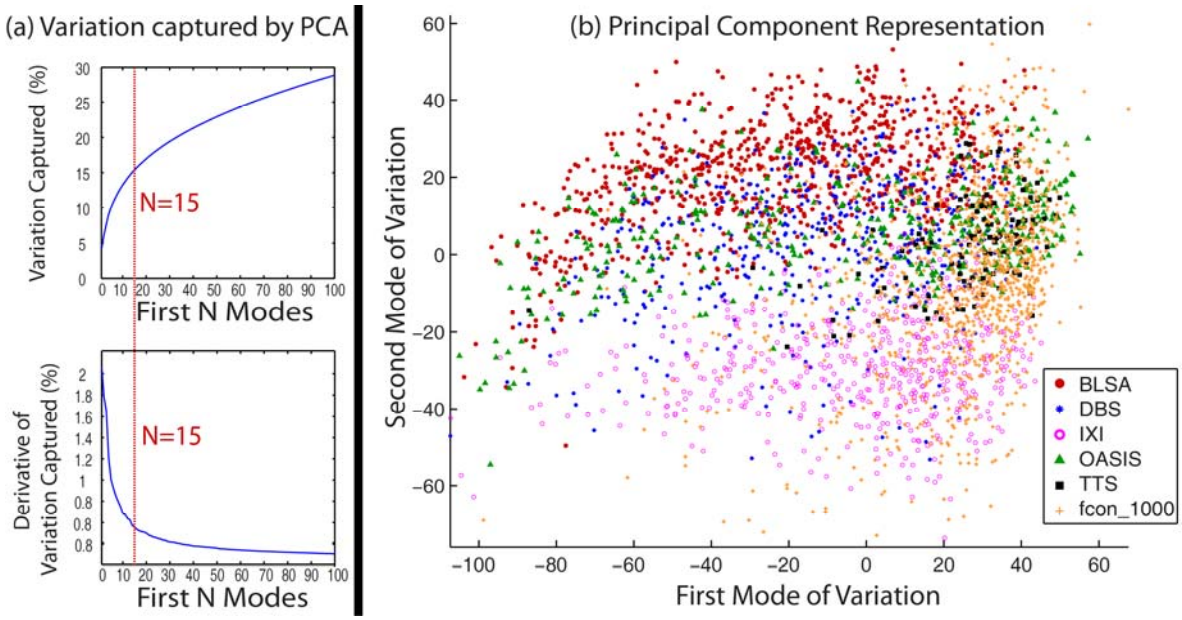


Figure II.3 (a) Total variation captured by first N modes from the PCA projection. The upper left figure shows the total variation captured by first N modes from the PCA. It is got from the percentage of the cumulated sum of the first N eigenvalues among all eigenvalues. The lower left figure shows the derivative of the upper left figure. (b) Coordinate embedding of 3464 training dataset from 6 projects. The first two modes in the PCA low-dimensional space are shown.

4.1. low-dimensional representation

For the large-scale framework, it is time-consuming to find the closest leaners by calculating the similarity measurements between every testing subject and 3464 training images in the original image space. Thus, the low-dimensional representation is used for computational efficiency. In the MLF framework, we need to find the close (anatomically similar) trained learners for a target image by a low-dimensional representation of high-dimensional MRI image data. Linear models such as principle component analysis (PCA) [176] and Multidimensional Scaling (MDS) [177] have been widely used to address this problem. In recent years, non-linear manifold learning algorithms like Isomap [178], Laplacian Eigenmaps [179] and Local Linear Embedding (LLE) [180] have also been successfully used in addressing the low-dimensional projection [160, 181-183]. However, the typical non-linear methods require the computational expensive pairwise similarity measurements which is a heavy burden for datasets with thousands, or more, 3D images. Thus, to accommodate the large-scale scheme, the PCA is employed in the MLF framework. The first 15 modes of variation in the PCA are used as the low-dimensional representation as it offers a practical / pragmatic choice that has shown stable performance for the MLF framework. The chosen number of components represents a balance between capturing more variations and avoiding over-fitting (Figure II.3a). However, we do not claim the optimality of the number of PCA low-dimensional representation from Figure II.3a. To validate the usage of PCA, the widely used Laplacian Eigenmaps method is also evaluated in this chaper. The comparisons are shown in the section 4.6.

The first two modes of variation for PCA applied to the 3464 training images are shown in Figure II.3b. As shown in the figure, the training images are densely distributed in the Eigenspace. As a result, locally closer trained learners are able to be found for a target image by the large-scale framework than the small-scale framework. Moreover, the images from different studies distributed differently in the Eigenspace, which means these studies are not redundant. Thus, a more representative training dataset is provided by the heterogeneous datasets.

4.2. Parameter Optimization and Sensitivity

First, we optimize the number of locally selected atlases for the initial weak segmentation (via a majority vote). For optimization, the desired parameters are swept across an appropriate range for a random subset of 50 training images. The results can be seen in Figure II.4. The Dice similarity values in the Figure II.4 are computed by comparing the 50 segmentations from the AdaBoost classifier with the corresponding multi-atlas segmentations. For the initial majority vote accuracy (Figure II.4A), using too few (e.g., 5) or too many (e.g., all available training data) results in sub-optimal accuracy. Additionally, there is marginal return when increasing the number of selected atlases beyond 25. Thus, as computation time is of primary concern, the closest 25 training images were used for all subsequent analysis.

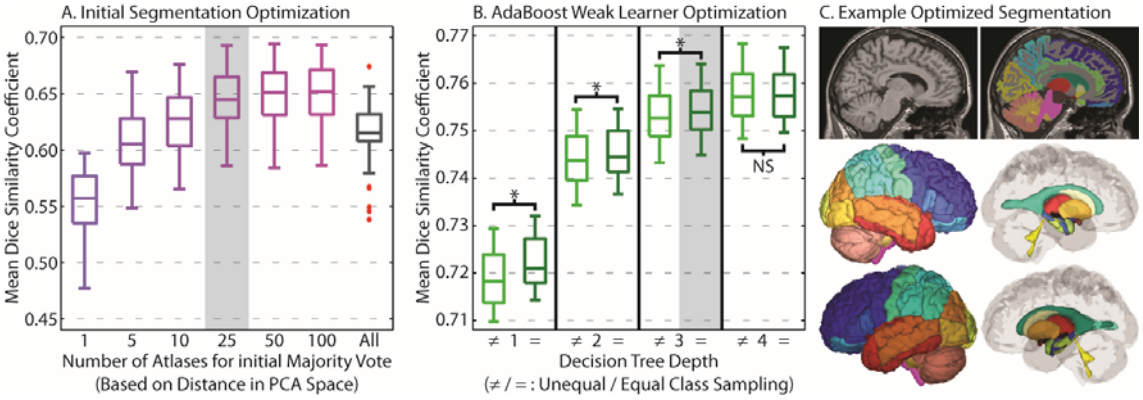


Figure II.4 Parameter optimization and sensitivity for the number of atlases fused for the initial majority vote (A), and the type of weak learner used for the AdaBoost classifiers (B). A representative segmentation using the optimized parameters can be seen in (C). Note, on (B), "*" indicates statistically significant difference, and "NS" indicates no significant difference.

Second, we optimize the weak learner (decision tree) used in AdaBoost classifier. The decision tree works as the weak learner h_{jlt} for the image j, label i and iteration t. At iteration t, The decision tree is built based on the Classification And Regression Tree (CART) method [184]. Each node can be split into two child and the splits are determined by the maximizing the classification rate[185, 186]. For the AdaBoost weak learner optimization (Figure II.4b), we consider decision trees with depths ranging from 1 (i.e., a "decision stump") to 4. Additionally, we consider two sampling methods, unequal and equal. For each label, the samples are the feature voxels from the training data (matrix X) and the corresponding true

values (matrix Y) within pre-calculated regional masks. Each regional mask extracts the voxels with probability larger than 0 from its regional probabilistic atlas, which is obtained by averaging the regional segmentations from all the 3464 training segmentation images. For unequal sampling, all available voxels within the mask were used for each label, regardless of the resulting class imbalance between the positive class ($Y = 1$) and negative class (Y =-1). For equal sampling, a random subset from the larger class was selected to enforce class balance (the same number of samples in positive and negative class). Here, it is evident that (1) increasing the decision tree depth improves training accuracy, and (2) equal class sampling provides a marginal, yet significant, improvement in segmentation accuracy. Given the marginal return and dramatic runtime increase of a depth 4 decision tree, a depth 3 decision tree with equal class sampling was used for all subsequent experiments.

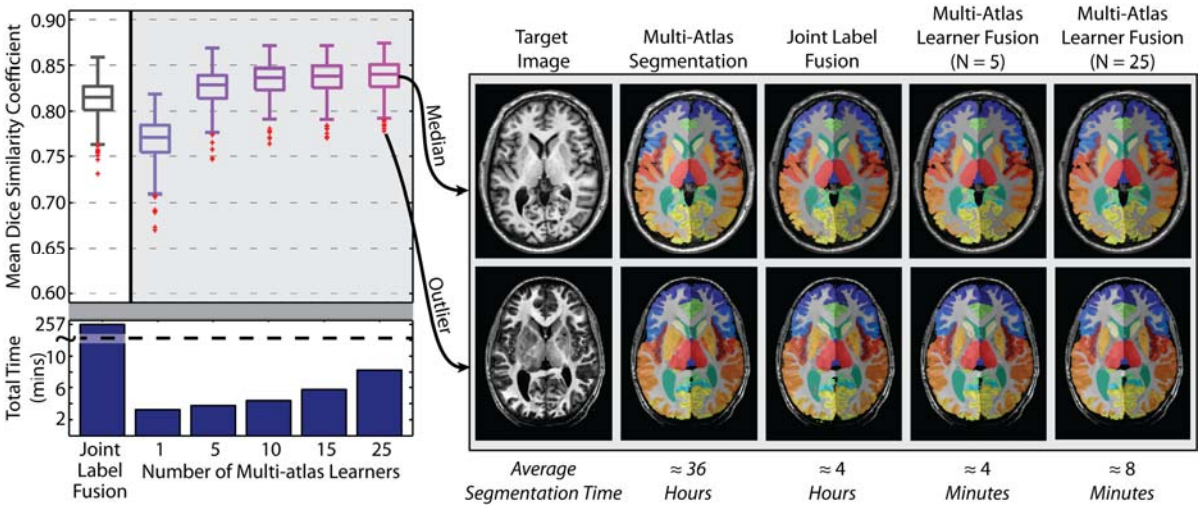


Figure II.5 Mean accuracy assessment for the defined testing data using the multi-atlas segmentation estimate as a "silver standard". The results demonstrate (1) the MLF framework provides a dramatic decrease in total segmentation time, (2) increasing the number of fused learners has valuable benefits in terms of segmentation accuracy, and (3) fusing more than 5 local learners the MLF framework provides substantial and significant accuracy benefits over the joint label fusion baseline.

4.3. Testing Data Accuracy and Assessment

Next, we quantify our ability to replicate the expensive multi-atlas segmentation result using the MLF framework. Using the multi-atlas segmentation estimate on our testing data (380 images) as a "silver

standard" we applied the MLF framework with varying numbers of local learners (from 1 to 25). The "silver standard" is the multi-atlas segmentations using both rigid and non-rigid registration [88, 89] and Non-local Spatial Staple label fusion [92]. As a benchmark, we consider fusing the 25 nearest training images using the premier joint label fusion (JLF) algorithm [91]. More specifically, the multi-atlas segmentation uses typical "non-rigid registration + fusion" framework to (1) generate the training images, and (2) demonstrate the state-of-the-art multi-atlas segmentation performance with non-local information. Once we get the trained framework, the MLF only requires an affine registration when applying new subjects to the trained AdaBoost learners. To compare with the MLF, the benchmark JLF also uses "affine registration + fusion" framework, which guarantees the MLF and the JLF are in the exactly same condition except the label fusion. The results of this experiment across the 380 testing images (Figure II.5) demonstrate: (1) increasing the number of local learners results in an improved ability to replicate the multi-atlas segmentation result, (2) using at least 5 learners results in significant and substantial improvement over the JLF benchmark, and (3) increasing the number of learners from 1 to 25 increases the total segmentation time from approximately 3 min to approximately 8 min – which remains a speedup of ≈30x over the JLF benchmark and ≈270x over the multi-atlas framework (shown in Table II.2). In Table II.2, we show the time consumed by registration and label fusion as well as the total time required for each framework. For multi-atlas segmentation, 15 non-rigid registrations were conducted for each testing subject. However, for the JLF and MLF, only 1 affine registration was required since all the training data and the trained AdaBoost learners had already been aligned to MNI space. The qualitative results support the quantitative accuracy analysis for both the worst and median cases from the testing set.

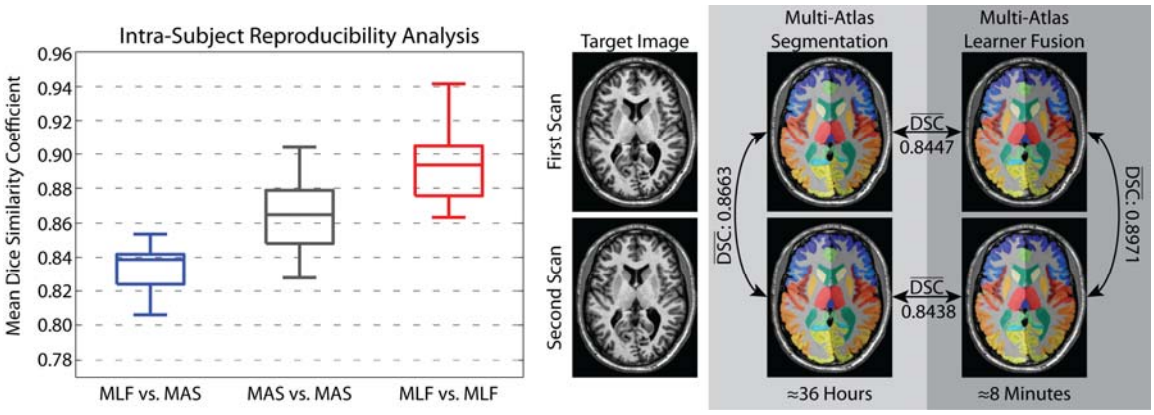


Figure II.6 Reproducibility analysis on the MMMRR dataset. Note, (1) the MLF similarity to the multi-atlas segmentation result approaches the intra-subject reproducibility for multi-atlas segmentation, and (2) MLF is significantly more reproducible than multi-atlas segmentation on this dataset.

4.4. Reproducibility Data Accuracy and Assessment

Then, we assess the reproducibility of the MLF framework using the MMMRR dataset (see Table II.1). Within this dataset, all 21 subjects were scanned twice with exactly the same scanning parameters. All subjects are healthy without history of neurological disease. This dataset is intended to be a resource for statisticians and imaging scientists to quantify the reproducibility of their imaging methods using data available from a generic session at 3T. The intra-subject reproducibility was assessed by comparing the mean DSC for: (1) the MLF result vs. the corresponding multi-atlas result, (2) the intra-subject multi-atlas estimates, and (3) the intra-subject MLF framework estimates. The results (Figure II.6) demonstrate: (1) the MLF similarity to the multi-atlas segmentation result approaches the intra-subject reproducibility for multi-atlas segmentation, and (2) MLF is significantly more reproducible than multi-atlas segmentation with a mean intra-subject DSC improvement of 0.0288.

4.5. Efficacy of Large-scale Data Model

Next, we compare the efficacy of the large-scale data model with a small-scale model via a simulation. The purpose of doing simulation is to compare the performance using large-scale heterogeneous dataset with using small-scare homogenous dataset. Moreover, more independent training and testing datasets can be generated from the limited number of available truth atlases (with manual segmentations).

To generate simulated data, we randomly selected 6 subjects from 45 atlases and divided them to 3 training subjects and 3 testing subjects. Then, a deformation was applied on the 3 training subjects to generate 90 deformed images and labels (30 for each subject) by sampling a sixth-order Chebyshev polynomial with random coefficients [92]. In these 90 image-label pairs, 10 were used as atlases in multi-atlas segmentation for three label fusion algorithms: (1) majority vote (MV), (2) Spatial Staple (SS) [172] and Non-local Spatial Staple (NLSS) [92] while the rest 80 were used as training data for the MLF framework. Note that the multi-atlas segmentation (MV, SS and NLSS) uses the non-local registration while the MLF framework does not. Lastly, the 3 testing subjects were deformed to 27 testing images using the same method as 3 training subjects.

After getting the simulated data, we (1) applied multi-atlas segmentation algorithms on 27 simulated testing images using 10 simulated atlases, (2) trained the MLF framework by 80 simulated training image-label pairs and tested the MLF framework by 27 simulated testing images, and (3) evaluated the large-scale data model by running the 27 simulated testing images under the MLF framework which was trained by 3464 images (see Table II.1). When testing the large-scale data model, for each testing subject, the same subject in large-scale training dataset was excluded to keep the testing procedure unbiased. The results (Figure II.7) show: (1) increasing the number of training data from 80 to 3464 results in significant improvement on the DSC, (2) with small-scale training data, the MLF framework performs worse than any of multi-atlas segmentation algorithms (MV, SS and NLSS), (3) with large-scale training data, the MLF framework (with 25 learners) provides significant improvement not only over the small-scale model but also over MV and SS, (4) the MLF framework with 25 learners performs less accurately than NLSS since the MLF framework does not use the non-local information which NLSS used. Therefore, the large-scale data model improves the performance of the MLF framework and achieves acceptable accuracy.

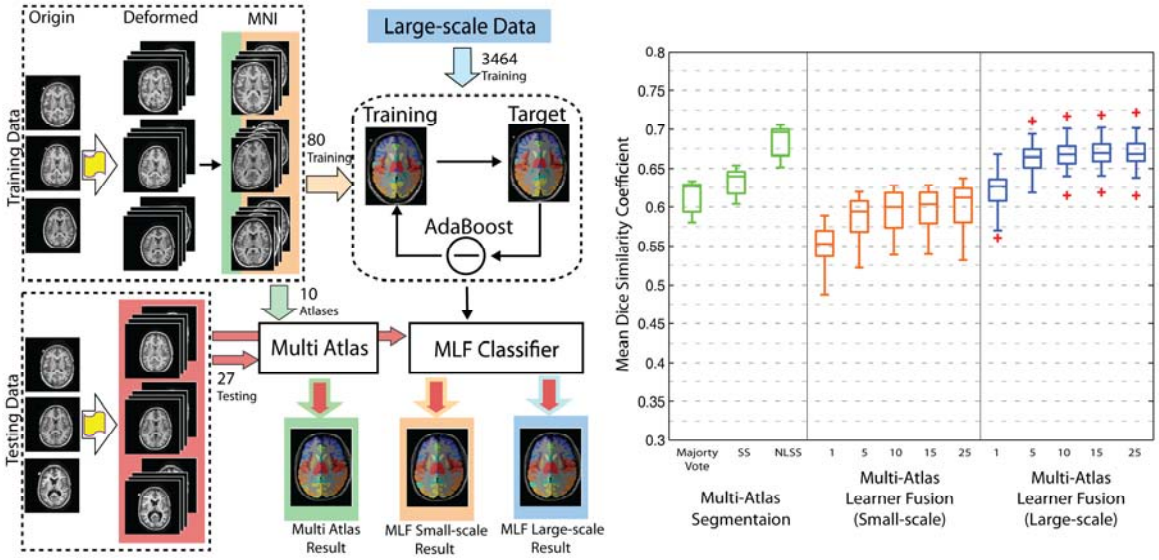


Figure II.7 Summary of the simulation and results. The flowchart shows the framework of the simulation: (1) 3 images were deformed to 90 simulated images and converted to MNI space by affine registration. (2) 10 of them were used as atlases for multi-atlas segmentation while 80 of them were used as training data for the MLF framework. (3) 3 images were deformed to 27 testing images for comparing the Multi-Atlas segmentation, small-scale model and big data model. The results demonstrate (1) the performance of the MLF framework is significantly improved when using big data model (3464 training images) and (2) the MLF framework under big data model provides the better performance than MV and SS even without using non-local information.

Table II.2 Runtime of each method on an Intel Xeon W3550 4 Core CPU (64 bit Ubuntu Linux 12.04)

Methods	Time consumed		
	Registration	Fusion	Total
Multi-Atlas segmentation (with MV)	≈ 22 h	≈ 5 min	**≈ 22 h**
Multi-Atlas segmentation (with SS)	≈ 22 h	≈ 2 h	**≈ 24 h**
Multi-Atlas segmentation (with NLSS)	≈ 22 h	≈ 14 h	**≈ 36 h**
Joint Label Fusion framework	≈ 2 min	≈ 4 h	**≈ 4 h**
Multi-Atlas Learner Fusion framework (with 5 learners)	≈ 2 min	≈ 2 min	**≈ 4 min**
Multi-Atlas Learner Fusion framework (with 25 learners)	≈ 2 min	≈ 6 min	**≈ 8 min**

4.6. Empirical Validation

Lastly, we compare the performance of MLF framework with state-of-the-art multi-atlas segmentation algorithms by an empirical validation. To conduct the empirical validation, we employed 5 manually labeled subjects (with the same protocol as atlases but have not been used as atlases) from the 50 MPRAGE images as unbiased testing data. Note that these were obtained from the human raters after the

conclusion of the algorithm training and development process. Since the testing dataset has the size n=5, all claims of statistical significance in this section are made using a Wilcoxon signed rank test ($p < 0.05$) which is the smallest significant level for n=5 [175].

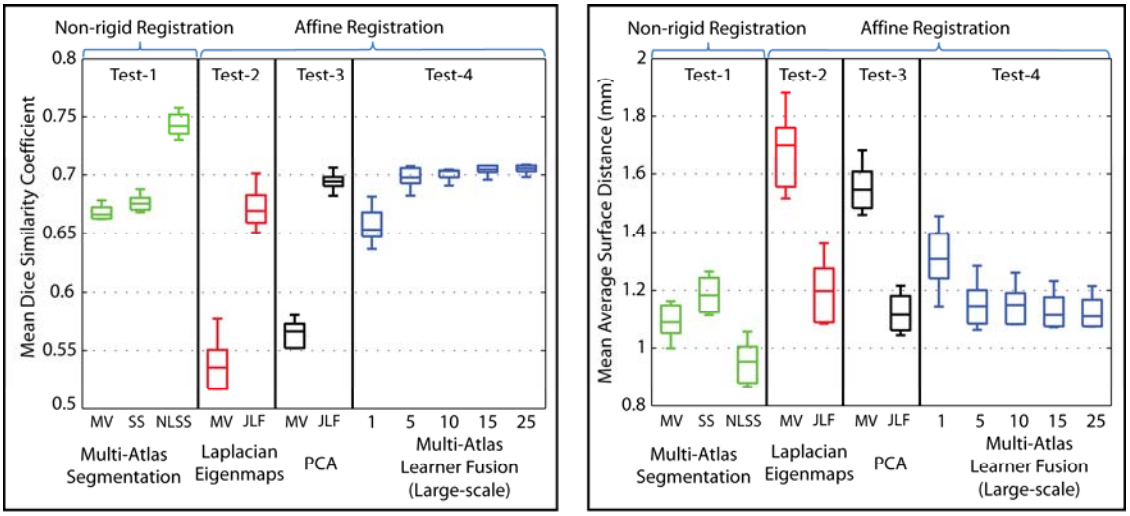
Briefly, we conducted four types of analyses called Test-1, Test-2, Test-3 and Test-4. In Test-1, the multi-atlas segmentation pipeline is applied to 5 MPRAGE images with different label fusion algorithms: MV, SS and NLSS (use 15 atlases from 45 MPRAGE images). In Test-2, the 25 nearest training images were selected by Laplacian Eigenmaps and then fused by the majority vote and JLF algorithm. Test-3 is the same as Test-2 except using the PCA for low-dimensional projection. Lastly, Test-4 applied the MLF framework with varying numbers of local learners (from 1 to 25). Note that Test-2, Test-3 and Test-4 use the same 3464 training images.

Overall, Test-1 has the highest CCNR among 4 groups. Test-2 is employed to compare the non-linear low-dimensional projection with the PCA used in Test-3. Test-3 serves as the benchmark to evaluate the performance of the MLF framework in Test-4.

While providing a speedup of ≈30x over the JLF benchmark (Test-3) and ≈270x over the multi-atlas framework (Test-1), the segmentation quality of MLF framework (Test-4) is comparable with other methods. Dice similarity is used as the main metric of segmentation quality (Figure II.8a). Meanwhile, the average surface distance (ASD) is used as a supplementary metrics (Figure II.8b). Figure II.9 compares different methods (same as Figure II.8) by showing the same axial slice from one subject in the testing dataset. Here, we discuss the Dice similarity first.

1. Test-1 vs Test-4. We compare the MLF framework (Test-4) with three non-rigid registration based multi-atlas segmentation algorithms, MV, SS and NLSS (Test-1). The mean Dice similarity coefficients of the MLF framework (with 25 learners) are significantly higher than MV and SS. Meanwhile. Meanwhile, as shown in the simulation, the MLF framework with 25 learners performs less accurately than NLSS, which uses both non-rigid registration (high CCNR) and non-local correspondence (high CCNC). The results demonstrate the MLF framework (without CCNR and CCNC) provides significant

improvement on Dice similarity over MV and SS (high CCNR) without using time-consuming non-rigid registration algorithms.



(a) Mean Dice Similarity Coefficient (b) Mean Average Surface Distance

Figure II.8 Results of empirical evaluation. The results indicate without using non-local information, the MLF framework (large-scale) provides better performance than two multi-atlas segmentation algorithms (MV and SS) and has comparable performance as the JLF benchmark. Note that, the multi-atlas segmentation used "non-rigid registration + fusion" framework while the JLF and the MLF used "affine registration + fusion" framework.

2. Test-3 vs Test-4. The comparison is conducted between the MLF framework (Test-4) and two benchmarks, majority vote and JLF (Test-3) which both use the same affine registration. Notice that the majority vote here is applied on the 25 atlases selected from 3464 training data (without CCNR). It is different from the majority vote in the multi-segmentations, which fuse the 15 non-rigid registered manual segmentations (in Test-1 with high CCNR). The MLF framework has significantly higher Dice similarity than the majority vote benchmark and has statistically indistinguishable Dice values comparing with the JLF benchmark. It proves that the MLF framework (without CCNR and CCNC), significantly outperforms the majority vote benchmark (without CCNR and CCNC) with the similar computational complexity. In addition, it has the comparable performance of JLF benchmark, which requires high CCNC to find non-local correspondences.

3. Test-2 vs Test-3, we compare the non-linear manifold learning method (Test-2) with the PCA method (Test-3) used in the MLF framework. The dataset used in Laplacian Eigenmaps is exactly the same

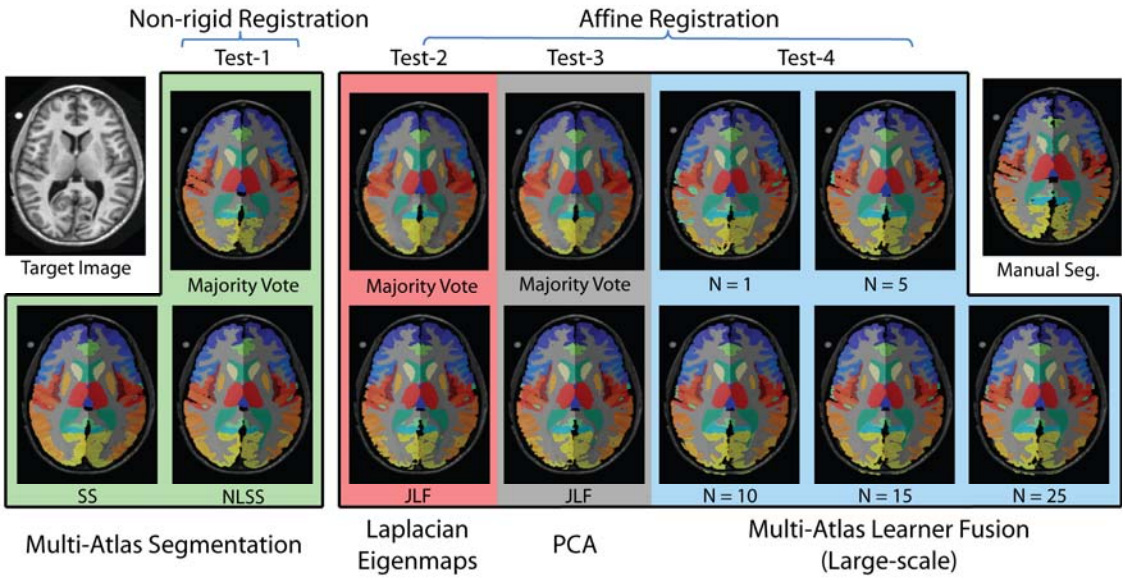


Figure II.9 Example for one subject, which corresponds to the different methods in Figure II.8. The anatomical and the manual segmentation of the target image are also provided.

as the one used for the PCA described in former sections. The closest subjects are selected based on the Euclidian distance of first 15 features in the Laplacian Eigenmaps. The Laplacian Eigenmaps is generated from the pairwise similarity measurements (normalized mutual information) between whole brain anatomical images. The results show that PCA performs significantly better than Laplacian Eigenmaps, which validates the usage of the PCA scheme. Even as validated, we do not claim any optimality of the PCA projection. Investigation into alternative low-dimensional projection methods could provide improvements.

The average surface distance (ASD) measurement repeats the finding in the Dice similarity except (1) the smaller value is better for ASD, which is different from the Dice similarity, and (2) the mean ASD is not significantly smaller than MV in multi-atlas segmentation. However, it is still better than the SS. The similar results from the surface distance provide a more robust comparison than using Dice similarity only.

To summarize, (1) the empirical validation repeats the results in the simulation, (2) the MLF framework (without CCNR and CCNC) outperforms MV and SS in Dice similarity coefficients without using non-rigid registration (high CCNR), (3) the MLF framework has comparable performance as JLF benchmark without using resource consuming non-local correspondences (high CCNC), and (4) PCA and

the Laplacian Eigenmaps have similar performance and PCA is a valid method under large-scale scenario.

5. Discussion and Conclusion

We present multi-atlas learner fusion (MLF), a framework for replicating the robust and accurate multi-atlas segmentation model, while dramatically lessening the computational burden. Using a training set of 3464 images, we estimate a low-dimensional representation of brain anatomy for selecting nearest appropriate example images, and build AdaBoost learners that map weak initial segmentations to the more accurate multi-atlas segmentation result. By completely bypassing the deformable atlas-target registrations, the MLF framework, cuts the runtime on a modern computer from 36 h down to 3-8 min – a speedup that could be further enhanced through GPU-based optimization. Specifically, we: (1) describe a technique for optimizing the initial segmentation and the AdaBoost learning parameters (Figure II.4), (2) quantify the ability to replicate the multi-atlas result with mean DSC of approximately 0.85 on a testing set of 380 images (Figure II.5), (3) demonstrate accuracies that are approaching the intra-subject multi-atlas reproducibility on a separate reproducibility dataset, and show significant increases in MLF reproducibility (Figure II.6), (4) show the advantage of large-scale data model by comparing small-scale training data with large-scale training data (Figure II.7), and (5) indicate the performance of MLF is better than MV and SS and is comparable to state-of-the-art multi-atlas segmentation algorithm (the JLF framework) without using non-local information (Figure II.8).

The results show the advantages of using large-scale data. Compared with the MLF framework under small-scale, the large-scale scheme improves the segmentation accuracy significantly. Compared with other state-of-the-arts multi-atlas segmentation methods, the MLF framework (without CCNR and CCNC) outperforms the typical multi-atlas frameworks (MV and SS) without using the resource consuming non-rigid registrations (high CCNR). Meanwhile, the MLF framework has comparable performance with the JLF benchmark (high CCNC). As a result, the MLF framework surpasses the expensive CCNR and CCNC, which speeds up the segmentation to 3-8 min without compromising on segmentation accuracy. With the availability of more training data (even the big data) the performance of the learning based large-

scale framework could be further enhanced.

In the interest of brevity, all of our comparisons have been against the standard pairwise registration framework for multi-atlas segmentation, and have not included the more recent advancements in groupwise registration (e.g., [162]). The primary reason for not directly including this comparison is: (1) groupwise registration is still a very active area of continuing research, and (2) the MLF framework is, in its essence, a machine learning perspective on the groupwise registration model. Meanwhile, since the simulated data and empirical data were manually labeled by the same protocol (BrainColor), the effect of inter-protocol comparison has not been discussed in this chaper.

The MLF framework is designed for the large-scale scenario so it does not perform well on small-scale dataset such as the 80 training dataset in the simulation. Meanwhile, although outside the scope of this chapter, applying the MLF framework on other applications (e.g., spinal cord segmentation and abdominal organ segmentation) would be interesting research topics in the future. As the soft tissues structures are not well constrained by bone and tend to exhibit higher inter-individual variation, we cannot make the conclusion that the proposed method is able to be applied on abdomen organ segmentation directly. However, this learning based large-scale processing framework might trigger new methods in organ segmentation with more representative training images and more powerful registration and label fusion tools for whole abdomen.

In the end, while the MLF framework shows great promise for rapid and accurate multi-atlas segmentation, there are certainly areas for which further investigation is warranted. Namely, first, we used a naïve PCA projection to model the neighborhood relationships between the training images. The proposed method is an open framework, which is able to incorporate with other algorithms. For example, the PCA and the AdaBoost algorithms could be replaced by any other low-dimensional projection methods and other two-class classifiers. More recent advancements in the manifold learning literature (e.g., [183]) present fascinating opportunities for more accurately modeling these relationships. Second, while highly successful, we do not claim any optimality of our AdaBoost-based learners. Investigation into alternative classification techniques (e.g., [187]could provide valuable improvements in segmentation modeling

without dramatically altering the MLF framework.

Chapter III. Consistent Cortical Reconstruction and Multi-atlas Brain Segmentation

1. Introduction

Whole brain segmentation and cortical surface reconstruction are two essential automatic techniques for quantitatively investigating Magnetic resonance (MR) images [50, 53, 188-190]. MR images provide morphometric measurements such as region of interest volume [191-194], cortical thickness [105, 195, 196], and surface area [197, 198] using either manual delineation or automatic medical image processing methods [199, 200]. Manual investigation is extremely resource consuming, so validated automatic methods [52-54] are overwhelmingly preferred.

Atlas-based segmentation assigns tissue labels to the voxels of unlabeled images using a pairing of an anatomical MR image and a corresponding manual segmentation [201]. The pair of images is commonly referred as an atlas. Initially, labels were transferred from a single atlas to a target by image registration [83-85]. However, single-atlas segmentation has difficulty capturing large inter-subject anatomical variation [202]. As reviewed in [203] the de facto standard atlas-based segmentation paradigm, has become to use multiple atlases and carry out label combination [86, 87, 91-99, 203].

Cortical reconstruction, the localization and representation of human cortical surfaces, is another widely used automatic technique in neuroscience [105-110]. Cortical reconstruction has been key to surface based registration [204-208], cortical labeling [209-211], population-based probabilistic atlas generation [212], and surface based morphometry [213, 214].

Spatial inconsistences that can hinder further brain morphometry analyses might develop because brain segmentation and cortical reconstruction are typically conducted separately. There are limited reports of methods for consistent whole brain volumetric segmentation and cortical surface reconstruction [58-60]. FreeSurfer is a well-known method for whole brain segmentation and cortical reconstruction that has been widely accepted as the de facto standard of brain segmentation [59, 106, 111]. FreeSurfer first automatically

labels whole brain image volumes as gray matter (GM), white matter (WM), cerebrospinal fluid (CSF), and subcortical regions by combining a Markov random field (MRF) and probabilistic atlases into a Bayesian framework [193, 215, 216]. Then, an outer (or pial) surface is reconstructed based on the GM/CSF boundaries while an inner surface is reconstructed based on the GM/WM interface [106]. Finally, the cortical GM regions are labeled based on a surface parcellation that forces the cortical segmentations to be consistent with the surfaces [209, 217]. However, since the latter steps strongly rely on the former steps in this "segmentation to surface reconstruction to parcellation" strategy, the cortical parcellation fails when the segmentations and surfaces are reconstructed incorrectly. FreeSurfer has yielded inaccurate whole brain segmentations and cortical surfaces in older adults typically with larger ventricles. When this happens, the resulting surface reconstruction and parcellation are inaccurate.

Cortical surface measurements from FreeSurfer have been evaluated against manual measurements in Alzheimer's disease [218] and post-mortem histologic measurements [219]. In both cases, FreeSurfer surface estimates showed a high level of correspondence with the manual estimates. Thus, alternative cortical surface algorithms should be consistent with FreeSurfer as long as FreeSurfer operates as intended. Substantial differences would indicate a failure of either FreeSurfer or the novel method. FreeSurfer is not the only approach for segmenting cortical surfaces. Cortical Reconstruction using Implicit Surface Evolution (CRUISE) [58, 220, 221] is a well-validated method that reconstructs consistent cortical surfaces and fuzzy segmentation [222-224].

In this chapter, we propose a novel "multi-atlas segmentation to surface" method called Multi-atlas Cortical Reconstruction Using Implicit Surface Evolution (MaCRUISE). MaCRUISE simultaneously obtains 133 volumetric labels from a single multi-atlas segmentation and achieves volume consistent and robust cortical surfaces based on the same segmentation. Multi-atlas segmentation is performed with Non-local Spatial Staple (NLSS) [92, 172]. The main contribution of this work is to integrate cortical reconstruction and multi-atlas segmentation. Specifically: (1) MaCRUISE obtains self-consistent whole brain multi-atlas segmentation (133 labels) and cortical surfaces without compromising surface accuracy. (2) MaCRUISE achieves more accurate volumetric segmentations than a traditional multi-atlas framework.

(3) While both deriving consistent whole brain segmentations and cortical surfaces, MaCRUISE is comparable in accuracy to FreeSurfer while achieving greater robustness across an elderly population. Notably, we do not seek to "outperform" FreeSurfer or CRUISE in terms of absolutely accuracy for cases in which these methods work as designed since they have both been extensively validated with respect to human expertise.

This work extends previous conference work [225]. Herein, we present a more complete description of the MaCRUISE and a more thorough analysis of the performance on an extended dataset. Additionally, we introduce MaCRUISE+ (by extending MaCRUISE using the CRUISE+ approach [220]) as a method to reconstruct accurate cortical surfaces and volumetric segmentations when multiple sclerosis (MS) lesions are present.

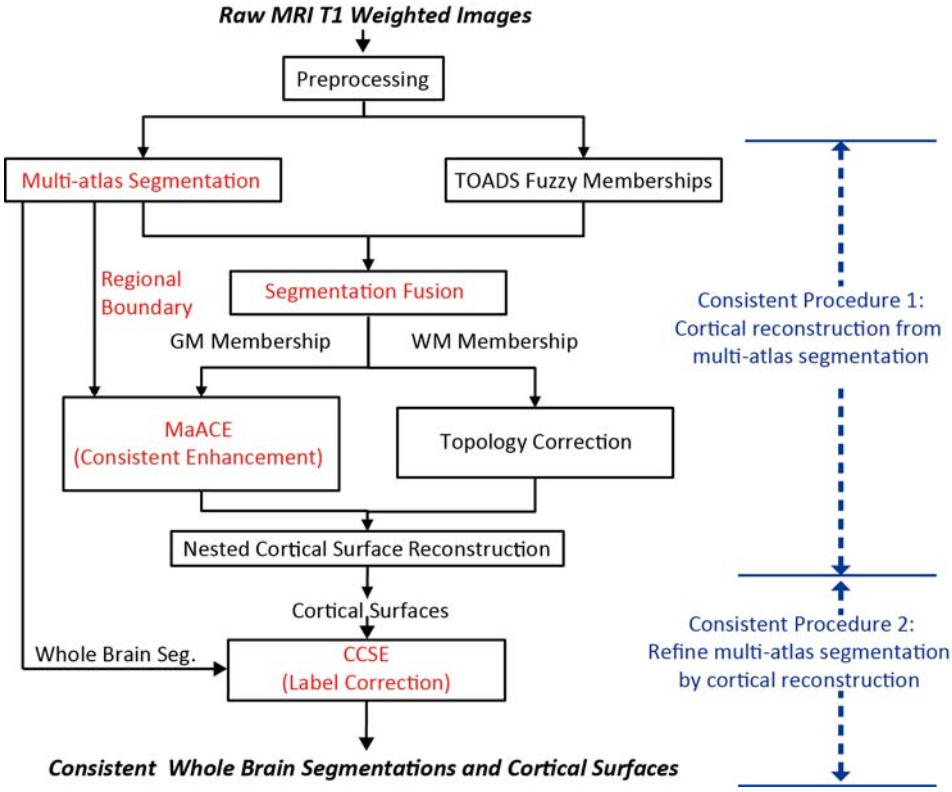


Figure III.1 Block diagram of MaCRUISE. Black text indicates the steps in original CRUISE while red text indicates the additional steps in MaCRUISE.

2. Theory and implementation

MaCRUISE is a method that produces consistent multi-atlas segmentations and cortical reconstruction from T1-weighted MR images (Figure III.1). First, cortical surfaces are reconstructed based on estimated tissue class memberships and multi-atlas boundary information. Second, multi-atlas segmentations are refined by the reconstructed cortical surfaces.

2.1. Preprocessing

Images are bias corrected with N4 [226] prior to being used as inputs for multi-atlas segmentation. The bias corrected images are skull stripped with SPECTRE [227] and processed by dura stripping [220] in preparation for TOAD.

2.2. Segmentation

2.2.1. Multi-atlas segmentation

Multi-atlas segmentation is performed with 45 MPRAGE images from the Open Access Series on Imaging Studies (OASIS) dataset [132]. The images are expertly delineated using 133 labels (132 brain regions and 1 background) according to the BrainCOLOR protocol [170]. All of the 45 OASIS atlases are available from Neuromorphometrics Inc. (http://www.neuromorphometrics.com/) and 35 of the atlases are freely available from the MICCAI 2012 Grand Challenge and Workshop on Multi-Atlas Labeling [228] (https://masi.vuse.vanderbilt.edu/workshop2012/).

Briefly, each target image is first affinely registered [89] to the MNI305 atlas [171]. Following [92, 149], the 15 closest atlases for each target image are selected from the 45 OASIS atlases using PCA projection. The 15 selected atlases are non-rigidly registered to the target image [88] and non-local spatial staple label fusion (NLSS) [92, 172] is used to combine the labels from each atlas to the target image. For non-rigid registration, we use symmetric image normalization (SyN), with a cross correlation similarity metric convergence threshold of 10^{-9} and convergence window size of 15, provided by the Advanced Normalization Tools (ANTs) software [88]. After multi-atlas labeling, each voxel in the brain is assigned

to one of the 133 labels in the BrainCOLOR protocol.

To assist with the cortical reconstruction framework in CRUISE, all cortical GM labels are combined into a single GM segmentation (M_{GM}). All WM labels and several subcortical labels (nucleus accumbens, amygdala, lateral ventricle, pallidum, putamen, thalamus, and ventral diencephalon) are combined into a single "pseudo-WM" segmentation (M_{WM}). The "pseudo-WM" subcortical labels are used to define M_{WM} to mimic the CRUISE "Autofill" procedure [58]. Finally, M_{GM}, M_{WM}, and the remaining subcortical labels (hippocampus, amygdala, basal forebrain, and inferior lateral ventricle) are grouped together to form a cerebrum segmentation M_{Cerebrum} (Figure III.2).

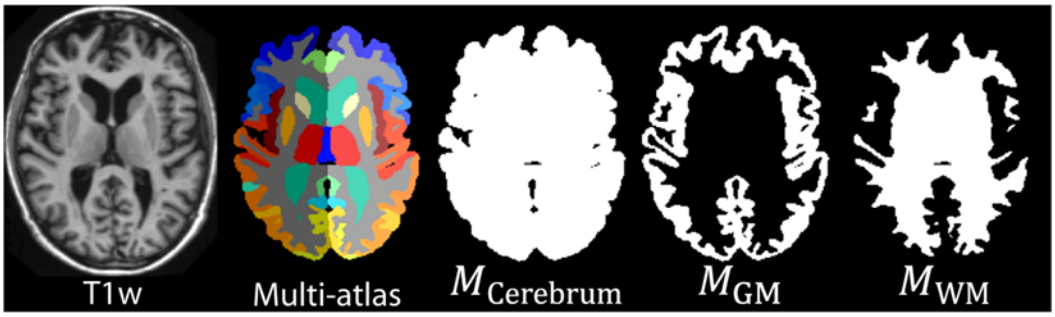


Figure III.2 Results from NLSS multi-atlas segmentation. From the multi-atlas segmentation, we derive cerebrum segmentation, GM segmentation and WM segmentation.

2.2.2. Memberships from TOADS

A straightforward way of reconstructing consistent cortical surfaces based on the multi-atlas segmentation is to establish surfaces on NLSS's GM/WM hard segmentation directly (NLSS+CRUISE). However, the atlases are manually labeled based on the expert defined protocol, so objective bias occurs. Moreover, the surface reconstruction suffers from the partial volume effect (PVE) in NLSS's hard segmentation. As shown in Figure III.3, independent application of CRUISE after NLSS (NLSS+CRUISE) does not yield accurate surfaces.

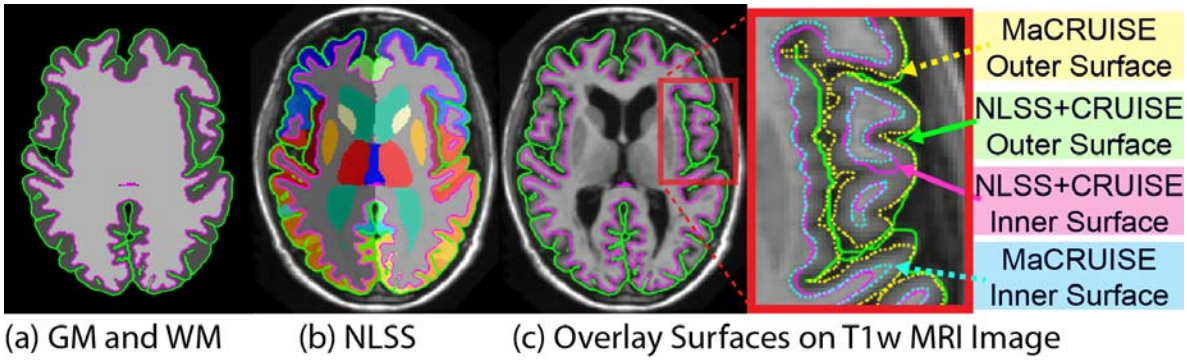


Figure III.3 Here we present the differences and challenges in directly applying multi-atlas hard segmentation to cortical reconstruction. ("NLSS+CRUISE"). (a) shows cortical reconstruction based on GM and WM segmentation using CRUISE. (b) shows the consistent surfaces with NLSS multi-atlas. (c) shows that the outer surface (green) and inner surface (magenta) from NLSS+CRUISE are inaccurate on enlarged 2D overlay (red rectangle). The dotted surfaces indicate the improvements by using the proposed MaCRUISE method

To address the bias and PVE in multi-atlas segmentation, fuzzy memberships are introduced in MaCRUISE using TOADS [224]. TOADS conducts fuzzy segmentation on skull and dura stripped T1 volumetric MR images by combining topological and statistical atlases. Finally, robust memberships μ_T of GM, WM, and CSF ($\mu_T{}^i_{\mathrm{GM}}$, $\mu_T{}^i_{\mathrm{WM}}$, and $\mu_T{}^i_{\mathrm{CSF}}$) for each voxel i are derived from TOADS.

2.2.3. Segmentation fusion

Multi-atlas hard segmentations are combined with the TOADS memberships to obtain fused GM, WM, and CSF memberships (μ^i_{GM}, μ^i_{WM}, and μ^i_{CSF}) for each voxel (Figure III.4). The combination consists of four stages.

Stage I assigns TOADS membership values within multi-atlas cerebrum segmentations.

$$\mu^i_{\mathrm{WM}} = \mu_T{}^i_{\mathrm{WM}}, \mu^i_{\mathrm{GM}} = \mu_T{}^i_{\mathrm{GM}} \text{ and } \mu^i_{\mathrm{CSF}} = \mu_T{}^i_{\mathrm{CSF}} \quad \text{if } M^i_{\mathrm{Cerebrum}} == 1 \tag{3.1}$$

This stage initializes the membership value from the TOADS fuzzy membership function within the multi-atlas cerebrum segmentation M_{Cerebrum}.

Stage II eliminates all the memberships outside the multi-atlas cerebrum segmentations.

$$\mu^i_{\mathrm{WM}} = 0, \mu^i_{\mathrm{GM}} = 0 \text{ and } \mu^i_{\mathrm{CSF}} = 0 \qquad \text{if } M^i_{\mathrm{Cerebrum}} == 0 \tag{3.2}$$

This step not only restricts outer boundaries of brain tissues by cleaning up the remaining dura and skull but also removes the cerebellum and brain stem by multi-atlas segmentations. This replaces the cerebellum and brain stem removal step in TOADS.

Stage III fills in the WM using the multi-atlas WM segmentation, which serves as an approximation of the inner cortical volume.

$$\mu^i_{\mathrm{WM}} = 1, \mu^i_{\mathrm{GM}} = 0 \text{ and } \mu^i_{\mathrm{CSF}} = 0 \qquad \text{if } M^i_{\mathrm{WM}} == 1 \tag{3.3}$$

This stage plays a similar role as the "Autofill" procedure in CRUISE, which modifies the WM segmentation by filling the ventricles and subcortical GM structures (e.g., putamen, caudate nucleus, thalamus, hypothalamus).

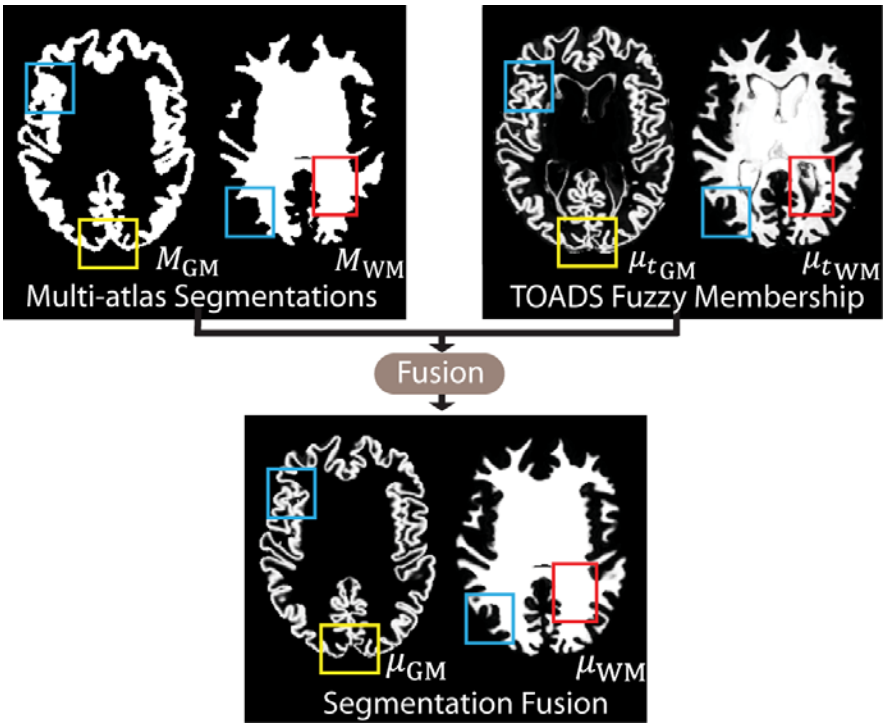


Figure III.4 Refined segmentations are obtained from segmentation fusion with the following characteristics: (1) PVE issues in NLSS multi-atlas segmentation are resolved (blue rectangles), (2) the fused segmentations have WM labels consistent with TOADS (red rectangles), and (3) non-cerebrum tissues are cleaned by the multi-atlas segmentation (yellow rectangles).

Stage IV corrects the inaccurate skull-stripping for the voxels whose $\mu_{t_{\mathrm{GM}}}^{i}$ are extremely small — that is, smaller than a constant C within M_{GM}^{i}. C is empirically set to 0.001 as the default value in MaCRUISE.

$$\mu_{\mathrm{WM}}^{i} = 0, \mu_{\mathrm{GM}}^{i} = 1 \text{ and } \mu_{\mathrm{CSF}}^{i} = 0 \qquad \text{if } M_{\mathrm{GM}}^{i} == 1 \text{ and } \mu_{t_{\mathrm{GM}}}^{i} < C \tag{3.4}$$

In other words, if voxel i is labeled as GM in the multi-atlas segmentation but also has an extremely small GM membership value in TOADS, then we trust the multi-atlas segmentation and set the membership value to 1 because this typically happens when skull-stripping fails.

After conducting the previous four stages sequentially, we obtain a fused segmentation that (1) is restricted to multi-atlas cerebrum segmentation, (2) addresses PVE by assigning fuzzy membership values inside the multi-atlas GM hard segmentation, (3) has robust WM filling using multi-atlas WM and subcortical segmentation, and (4) fixes incorrect GM membership values that result from inaccurate skull-stripping.

2.3. Cortical reconstruction

2.3.1. Multi-atlas anatomically consistent GM enhancement

Although the PVE in GM segmentation is addressed by segmentation fusion, the GM membership function in tight sulci is still obscured or even undetectable because the GM cortex is "back to back" in tight sulcal regions. To detect these sulci, one family of approaches applies cortical thickness constraints to estimate their locations [105, 229]. Another approach called Anatomically Consistent Enhancement (ACE) [56, 58] edits the GM membership values by creating a thin separation between sulcal GM banks based on evidence of the presence of CSF. However, ACE might not be able to detect tight sulci when the presence of CSF is not well captured by TOADS, especially when the contrast between GM and CSF is low. Moreover, the spatial location of sulci from ACE might not be consistent with the multi-atlas segmentation.

To force the estimated sulci to be consistent with multi-atlas segmentation, a hierarchical method called Multi-atlas Anatomically Consistent GM Enhancement (MaACE) is proposed to assign multi-atlas

cortical boundaries with the highest priority while estimating the sulci locations (Figure III.5). MaACE generalizes ACE for consistency by solving for $T(\mathrm{x})$ in the following Eikonal equation [58, 230]:

$$\begin{aligned} F(\mathrm{x})\|\nabla T(\mathrm{x})\| &= 1 \text{ in } \Omega \\ T(\mathrm{x}) &= 0 \text{ for } \mathrm{x} \in \Gamma \end{aligned} \tag{3.5}$$

where $T(\mathrm{x})$ is the weighted distance function for spatial 3-D position x. $F(\mathrm{x})$ is a speed function (defined below) and Γ is the location of the interface between GM and WM (0.5-isosurface). $T(\mathrm{x})$ can be computed using the fast matching method [230]. If $F(\mathrm{x})$ is equal to one everywhere, then $T(\mathrm{x})$ is the Euclidean distance from the GM/WM interface and the estimated sulci will be located at the midpoint between the gyral banks. The ACE approach defines $F(\mathrm{x})$ to be a spatial varying function that depends on the CSF membership values at x:

$$F(\mathrm{x}) = 1 - 0.9\mu_{\mathrm{CSF}}(\mathrm{x}) \tag{3.6}$$

where the $\mu_{\mathrm{CSF}}(\mathrm{x})$ is the CSF membership function and the 0.9 is an empirical coefficient. In this case, $T(\mathrm{x})$ can be regarded as the time it takes for a wave front starting from the GM/WM interface to reach x where the speed of the wave front will slow down in the CSF.

Since different cortical labels are separated mainly by sulci location in the BrainCOLOR protocol, cortical boundary locations in the multi-atlas segmentation are used as additional evidence of sulci in MaACE. We combine the boundary information to ACE and specify $F(\mathrm{x})$ as:

$$F(x) = 1 - 0.9\big(max\{\mu_{\mathrm{CSF}}(x), \mu_{\mathrm{boundary}}(x)\}\big) \tag{3.7}$$

where $\mu_{\mathrm{boundary}}(x)$ represents the boundary information in multi-atlas segmentations for which $\mu_{\mathrm{boundary}}(\mathrm{x}) = 1$. This is when x is at the "boundary" of cortical labels. The boundary is defined as any cortical voxel that (1) detects two or more different cortical labels among its 26 connections, and (2) does not detect WM labels among its 26 connections.

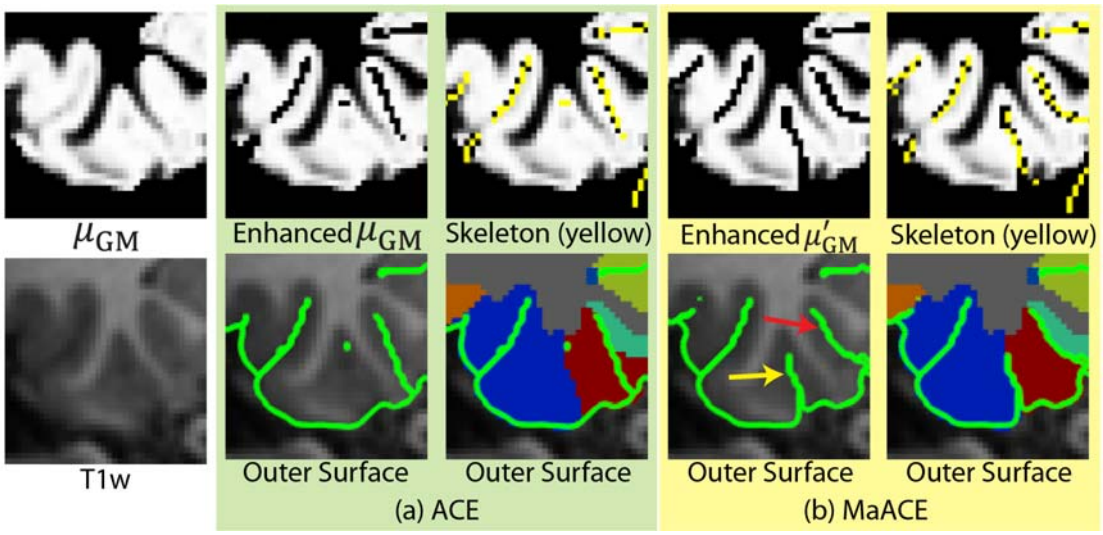


Figure III.5 MaACE compared with the ACE method, (1) MaACE is able to detect sulci in the outer surface that are not detected by ACE, particularly when CSF evidence is not visible (yellow arrow in b). (2) MaACE also forces sulci locations to be consistent with multi-atlas segmentation at the boundaries of cortical labels (red arrow in b). This figure also shows the enhanced GM membership and skeleton from ACE and MaACE (top row).

When using boundary information, MaACE detects additional sulci locations, which are not detected by ACE (yellow arrow in Figure III.5). Meanwhile, MaACE forces the sulci location to be consistent with multi-atlas segmentation (red arrow in Figure III.5b). The benefits of , which are a generalized form of Eq. 3.7, can be understood by considering its action in specific cases:

Case I: $F(\mathrm{x})$ becomes 0.1 when the multi-atlas segmentation boundaries exist with certainty — i.e., $\mu_{\mathrm{boundary}}(\mathrm{x}) = 1$. This step forces the estimated sulci to be consistent with the sulci definition in multi-atlas segmentation no matter if CSF evidence exists or not.

Case II: When CSF exists and multi-atlas segmentation boundaries do not (i.e., $\mu_{\mathrm{CSF}} > \mu_{\mathrm{boundary}}(\mathrm{x})$), then $F(\mathrm{x})$ becomes formula (6) which is conventional ACE. It forces the estimated sulci to be consistent with the evidence of CSF.

Case III: If we do not have evidence from either multi-atlas boundaries or CSF (i.e., $\mu_{\mathrm{boundary}}(\mathrm{x}) = \mu_{\mathrm{CSF}} = 0$), then $F(\mathrm{x})$ becomes a constant speed 1 and the sulci are located at the midpoint between sulcal banks (as in conventional ACE).

Using (7) and applying the fast matching method (FMM) starting from the GM/WM interface, segmentation consistent sulci are obtained from the "shocks" — that is, where the wave fronts hit each other [230]. In FMM, new values of $T(\mathrm{x})$ are obtained by solving quadratic equations using $F(\mathrm{x})$ and finite forward and backward difference approximation of $\nabla T(\mathrm{x})$. The ACE framework [56, 58] indicates that if an additional centered finite difference approximation $\nabla_c T(\mathrm{x})$ is conducted on the FMM derived $T(\mathrm{x})$, values of $F(\mathrm{x})\| \nabla_c T(\mathrm{x})\|$ are much smaller than 1. As a result, the set of shock points are obtained by applying a constant threshold Q.

$$S = \{x | F(\mathrm{x})\|\nabla_c T(\mathrm{x})\| \leq Q \text{ and } T(x) > 1\} \tag{3.8}$$

The threshold Q is smaller than 1 and empirically set to 0.85. Use of the constraint $T(x) > 1$ guarantees that the estimated sulci are only found outside the GM/WM surface and at a distance of 1 mm or greater from the GM/WM surface.

The final estimated sulci locations are obtained by conducting a thinning morphological operation on S to obtain its skeleton (which is centered on S and is only one voxel thick). After obtaining this skeleton, the GM membership function is modified as follows.

$$\mu'_{\mathrm{GM}}(x) = \begin{cases} F(\mathrm{x})\|\nabla_c T(\mathrm{x})\| \cdot \mu_{\mathrm{GM}}(\mathrm{x}) & \text{if x is on skeleton} \\ \mu_{\mathrm{GM}}(\mathrm{x}) & \text{otherwise} \end{cases} \tag{3.9}$$

2.3.2. Topology-perserving deformable cortical reconstruction

Three cortical surfaces — inner, central, and outer — are reconstructed with subvoxel accuracy by using the Topology-preserving Geometric Deformable surface Model (TGDM). First, the filled WM membership function is refined by a topology correction step to remove holes and handles. Then, an inner surface is reconstructed using the topological corrected WM membership values [56, 222]. A GVF force [231], a curvature force, and a regional pressure force are applied to push the inner surface from the GM/WM interface to the pial surface using the TGDM level set approach. The GVF force is generated by the MaACE-corrected GM membership function. The regional pressure force guarantees that the central surface is located within the cortical segmentations. Finally, using the central surface as the initial surface,

the outer surface is found using another TGDM step controlled by a curvature force and the MaACE-corrected GM membership function [58, 107, 230, 232, 233]. The TGDM method used is the same as that used in the original CRUISE algorithm.

2.4. Cortical consistent segmentation editing

Despite efforts to maintain consistency between the various sources of information, inconsistent voxels still remain at this stage (Figure III.6). We introduce the Cortical Consistent Segmentation Editing (CCSE) method to ensure that the multi-atlas segmentation is consistent with the cortical surfaces that have been reconstructed using TGDM. CCSE allows us to define what is "consistent" in a quantitative manner using two coefficients: an inner surface consistency coefficient α and an outer surface consistency coefficient β.

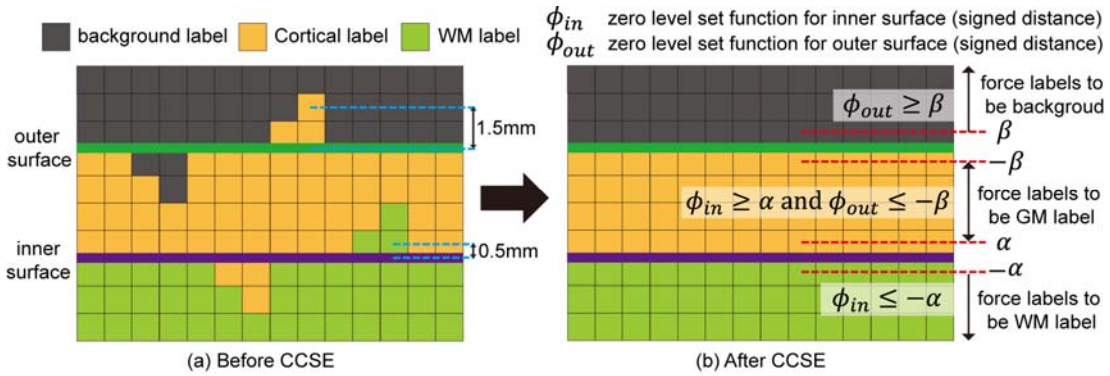


Figure III.6 The CCSE step corrects the inaccurate cortical labels to background or WM, if they are located outside of the outer surfaces or inside the inner surfaces, respectively. Meanwhile, CCSE adjusts the incorrect volume-wise labels to be cortical labels for voxels between inner and outer surfaces. The distances between voxels and surfaces are provided by the zero set level set functions ϕ_{in} and ϕ_{out}. The level of consistency is quantitatively controlled by two consistent coefficients, the inner surface consistent coefficient (α) and the outer surface consistent coefficient (β).

Let ϕ_{in} and ϕ_{out} be the level set functions for the inner and outer cortical surfaces reconstructed using TGDM. We can use these functions together with α and β to correct the labels produced by multi-atlas segmentation in the following way: (1) If a voxel is not labeled as background but it is more than β mm outside the outer surface, then we label it as background. (2) If a voxel is more than α mm inside the inner surface but has a background or cortical label that should be outside the GM/WM interface then it is

relabeled as WM. (3) In between the outer and inner surfaces, all voxels should be given cortical labels. If a voxel is incorrectly labeled, then it is marked as "needs label" and it is relabeled as one of the 98 cortical labels in the BrainCOLOR protocol using an iterative strategy described in [234]. Briefly in each iteration, the remaining "needs label" voxels are filled by the most commonly occurring cortical labels around its 26 connections. This procedure is performed iteratively until all "needs label" voxels are relabeled or no more voxels could be reached. (4) If a voxel is both on skeleton and its ϕ_{out} is $0 \leq \phi_{out} \leq \varepsilon$, we keep the original label and the ε is empirically set to 0.05 mm in MaCRUISE so that the labels with tight "back to back" sulcal surfaces (< 1 voxel width) are not over corrected.

Although the estimated cortical surfaces have subvoxel accuracy (since they are produced using a connectivity consistent marching cubes algorithm), the multi-atlas segmentation result only has voxel accuracy. This means that distances to the surfaces are reported with subvoxel accuracy but volumetric labels are restricted to the accuracy of the voxels. Since most T1w MR images (obtained for clinical and research purposes) have resolutions on the order of 1mm, it makes sense to choose α and β to be 0.5 mm so that voxels that cover about half of the cortex are given cortical labels. Therefore, both α and β are set to 0.5 mm for the remainder of this manuscript where the sensitivity of the algorithm regarding α and β is explored. Note that in the software implementation, users are free to choose alternative values for both α and β.

2.5. Extension to handle WM lesions with MaCRUISE+

We introduce a variation on MaCRUISE called MaCRUISE+, which incorporates the CRUISE+ method into the MaCRUISE framework. CRUISE+ [220] accurately and automatically reconstructs cortical surfaces when WM lesions are present, which commonly occurs in patients with multiple sclerosis. As with CRUISE+, MaCRUISE+ uses both Fluid Attenuated Inversion Recovery (FLAIR) T2-weighted (T2w) images and T1w images together with the Lesion-TOADS algorithm [235] in place of the TOADS algorithm. Lesion-TOADS estimates fuzzy membership functions for GM, WM, CSF, and the WM lesions. MaCRUISE+ uses the WM mask generated by Lesion-TOADS to remove inaccurate multi-atlas cortical

boundaries within the WM lesions. The other steps of MaCRUISE+ are identical to those of MaCRUISE.

3. Methods and Results

Validation of the MaCRUISE and MaCRUISE+ methods was performed with four distinct datasets and experiments. First, absolute surface accuracy of MaCRUISE was compared with the reference methods on a public database of expertly traced cortical surface points for control subjects. Second, absolute surface accuracy of MaCRUISE+ was compared with the reference methods on a public database of expertly traced cortical surface points for multiple sclerosis patients. Third, absolute volumetric accuracy of MaCRUISE was compared with the reference methods on an available (for purchase) database of expertly labeled whole brain volumes. Fourth, the robustness of MaCRUISE was assessed relative to the reference methods on a database of older healthy subjects. All validation datasets were obtained from different individuals other than the atlases used to construct the MaCRUISE and MaCRUISE+ methods.

3.1. Landmark based surface validation on healthy data

3.1.1. Data

The first experiment used a publicly available dataset consisting of five healthy subjects (age range: 30–49) [133] with Magnetization Prepared RApid Gradient Echo (MPRAGE) T1-weighted images acquired in the sagittal orientation (resolution=$1.0 \times 1.0 \times 1.2\ \mathrm{mm}^3$; FOV=$240 \times 204 \times 256\ \mathrm{mm}^3$). In prior work [220], two human raters placed 420 landmarks on both outer and inner surfaces of each subject at the calcarine fissure, cingulate gyrus, central sulcus, parieto-occipital sulcus, superior frontal gyrus, superior temporal gyrus, and Sylvian fissure. The landmarks were made on sulcal fundi, sulcal banks, and gyral crowns with floating point precision. For FreeSurfer, the T1w input images were interpolated to its optimal resolution ($1.0 \times 1.0 \times 1.0\ \mathrm{mm}^3$) using the default setting. For CRUISE, the recommended voxel resolution for optimal performance ($0.8 \times 0.8 \times 0.8\ \mathrm{mm}^3$) was used.

3.1.2. Experiment and results

Each of the methods (FreeSurfer, CRUISE, NLSS+CRUISE, and MaCRUISE) was run in an automated manner on each of the 5 datasets. Accuracy was assessed by computing the absolute surface errors (distance from surfaces to landmarks). Briefly, NLSS+CRUISE errors were larger than FreeSurfer and CRUISE, and the surface errors of MaCRUISE were comparable to those of both FreeSurfer and CRUISE. Table III.2statistically evaluates the differences in **Error! Reference source not found.**by conducting paired t-tests and Cohen's d effect size [236] analyses. Note that small p-value might indicate a significant effort of a magnitude that is not clinically relevant, so we rely on both metrics to interpret differences. Figure III.7shows the reconstructed inner and outer surfaces from one subject in the first experiment.

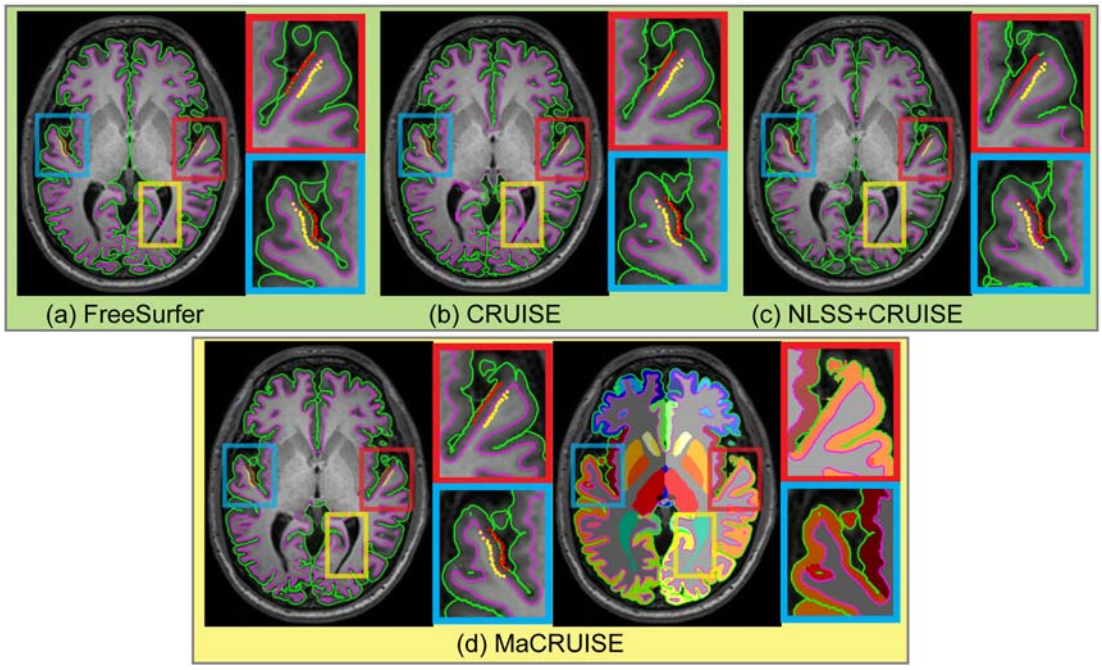


Figure III.7 Inner and outer surfaces are shown for different methods for a healthy subject. The red and yellow dots in blue and red rectangles are the manual outer and inner surface landmarks, respectively. FreeSurfer and CRUISE are two benchmark methods that achieve accurate surfaces. Note, NLSS+CRUISE does not reconstruct accurate surfaces. Using MaCRUISE, we obtain consistent cortical surfaces and whole brain multi-atlas segmentations. MaCRUISE generates accurate surfaces at lateral ventricles as well as highlighted in yellow rectangles.

Table III.1 Absolute surface errors on subjects with healthy anatomy with MaCRUISE (mean ± standard deviation in mm).

		FreeSurfer	CRUISE	NLSS+CRUISE	MaCRUISE
Optimal resolution*		1x1x1mm^3	0.8x0.8x0.8 mm^3	0.8x0.8x0.8 mm^3	0.8x0.8x0.8 mm^3
Rater A	Outer Surface	0.524 ± 0.372	0.486 ± 0.413	0.880 ± 0.755	0.518 ± 0.414
	Inner Surface	0.460 ± 0.371	0.540 ± 0.429	0.799 ± 0.758	0.544 ± 0.431
Rater B	Outer Surface	0.434 ± 0.369	0.613 ± 0.546	1.050 ± 0.889	0.585 ± 0.464
	Inner Surface	0.432 ± 0.362	0.542 ± 0.483	0.913 ± 0.961	0.544 ± 0.482

* We resampled the original images to either 1x1x1 mm^3 or 0.8x0.8x0.8 mm^3 prior to running the different methods. The best results and their corresponding resolutions are reported in this table.

Table III.2 Paired t-test and effect size analyses on absolute surface errors for landmarks with MaCRUISE.

		Rater A		Rater B	
		p value	Cohen's d*	p value	Cohen's d
NLSS+CRUISE vs. FreeSurfer	Outer Surface	<0.001	**0.598**	<0.001	**0.905**
	Inner Surface	<0.001	**0.567**	<0.001	**0.662**
NLSS+CRUISE vs. CRUISE	Outer Surface	<0.001	**0.648**	<0.001	**0.592**
	Inner Surface	<0.001	0.419	<0.001	0.488
MaCRUISE vs. FreeSurfer	Outer Surface	0.541	0.015	<0.001	0.361
	Inner Surface	<0.001	0.209	<0.001	0.262
MaCRUISE vs. CRUISE	Outer Surface	<0.001	0.078	<0.001	0.055
	Inner Surface	<0.001	0.009	0.145	0.003

*Cohen's d score is defined as "trivial" (d<0.2), "small effect" (0.2≤d<0.5), "medium effect" (0.5≤d<0.8), or "large effect" (d≥0.8). The bold d value numbers indicate the "medium" or "large" effect. Double underline indicates the significantly superior methods (p<0.001 and d≥0.5), while the dotted underline indicates a lack of evidence for systematic differences (p>0.05 or d<0.5). Single underline indicates the significantly superior methods (p<0.001 and d≥0.5) from at least one rater.

3.2. Landmark based surface validation on MaCRUISE+

3.2.1. Data

The second experiment used five publicly available MS subjects, consisting of four female subjects and one male subject with a mean age of 48.4 years (range: 40–59) with both MPRAGE and FLAIR images [220]. In prior work [220], the images were annotated in both healthy cortical regions and near lesions. The MPRAGE T1w images were acquired in the sagittal orientation with resolution $1.0 \times 1.0 \times 1.2\ \mathrm{mm}^3$. The FLAIR T2w images were acquired in the sagittal orientation but at resolution $0.83 \times 0.83 \times 2.2\ \mathrm{mm}^3$. All datasets were isotropically interpolated to $0.83 \times 0.83 \times 0.83\ \mathrm{mm}^3$ [220]. Two human raters labeled 420 landmarks per surface for each MS subject in approximately the same regions of interest (ROI) as described in but not near any WM lesions. To evaluate the surface reconstruction performance near WM lesions, five additional ROIs were specified to be near WM lesions. The original two raters and a third human rater each marked 50 landmarks for each MS image. As a result, a total of 2100 landmarks for healthy anatomy and 250 landmarks for cortex near WM lesions were used to evaluate the performance.

3.2.2. Experiment and results

Each of the methods (FreeSurfer, CRUISE+, and MaCRUISE+) was run in an automated manner on each of the 5 datasets. As an additional baseline comparison, FreeSurfer was run with the same lesion mask as used by MaCRUISE and MaCRUISE+, which was generated by Lesion-TOADS (referred as Corrected FreeSurfer*). Note that since the spatial resolution of the data was already 0.83 mm isotropic to match the highest resolution FLAIR data, all methods used the same data resolution.

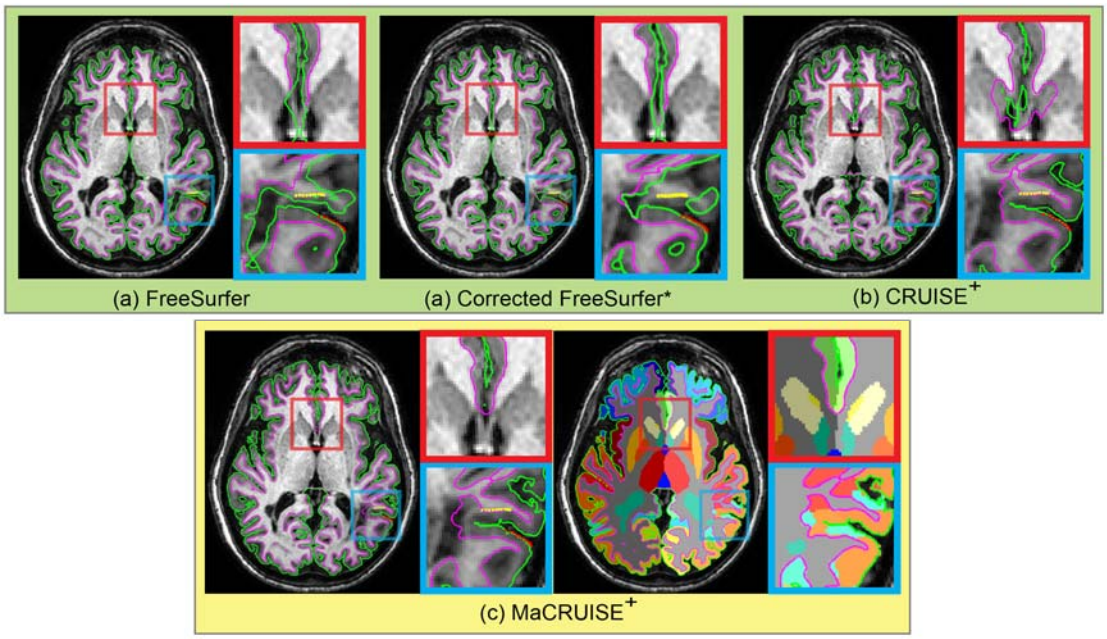


Figure III.8 Inner and outer surfaces are shown for each method for an MS subject. Red and yellow dots in blue and red rectangles are the manual outer and inner surface landmarks, respectively, near WM lesions. Based on the landmarks, CRUISE+ and MaCRUISE+ achieve more accurate surfaces than FreeSurfer and lesion corrected FreeSurfer*. Note that the corrected FreeSurfer* uses the same lesion mask as CRUISE+ and MaCRUISE+, which is generated by Lesion-TOADS. From (c), MaCRUISE+ achieves consistent cortical surfaces and whole brain segmentations that CRUISE+ does not.

Accuracy was assessed by computing the absolute surface errors (distance from surfaces to landmarks) as shown in **Error! Reference source not found.**and using paired t-test and effect size analyses as shown in **Error! Reference source not found.** with the same approach as MaCRUISE. The underlined annotations in Table III.4 Paired t-test and effect size analyses on absolute surface errors for landmarks with healthy anatomy and WM lesions with MaCRUISE^{+} indicate the superior methods ($p<0.001$ and $d\geq0.5$) (definition found in Table III.2). Figure III.8 shows the reconstructed inner and outer surfaces from one MS subject with landmarks near WM lesions.

Table III.3 Absolute surface errors with healthy anatomy and WM lesions with MaCRUISE$^+$ (mean ± standard deviation in mm).

Landmarks			FreeSurfer	Corrected FreeSurfer*	CRUISE$^+$	MaCRUISE$^+$
Healthy Anatomy	Rater A	Outer Surface	0.445 ± 0.394	0.433 ± 0.339	0.509 ± 0.490	0.529 ± 0.496
		Inner Surface	0.572 ± 0.471	0.511 ± 0.420	0.482 ± 0.455	0.485 ± 0.510
	Rater B	Outer Surface	0.778 ± 1.605	0.600 ± 0.976	0.518 ± 0.539	0.624 ± 0.698
		Inner Surface	0.423 ± 0.326	0.411 ± 0.302	0.368 ± 0.340	0.390 ± 0.354
Near WM Lesions	Rater A	Outer Surface	0.858 ± 1.588	0.679 ± 1.009	0.551 ± 0.566	0.670 ± 0.808
		Inner Surface	0.536 ± 0.488	0.494 ± 0.407	0.337 ± 0.283	0.368 ± 0.293
	Rater B	Outer Surface	0.874 ± 1.498	0.735 ± 1.043	0.589 ± 0.599	0.682 ± 0.755
		Inner Surface	0.476 ± 0.564	0.446 ± 0.551	0.425 ± 0.315	0.387 ± 0.340
	Rater C	Outer Surface	1.028 ± 1.270	0.874 ± 0.878	0.641 ± 0.553	0.705 ± 0.699
		Inner Surface	0.707 ± 0.530	0.696 ± 0.588	0.410 ± 0.293	0.447 ± 0.322

*FreeSurfer after correction with the WM lesion masks generated by Lesion-TOADS.

Table III.4 Paired t-test and effect size analyses on absolute surface errors for landmarks with healthy anatomy and WM lesions with MaCRUISE$^+$

Landmarks			Rater A		Rater B		Rater C	
			P value	Cohen's d	P value	Cohen's d	P value	Cohen's d
Healthy Anatomy	MaCRUISE$^+$ vs.FreeSurfer	Outer Surface	<0.001	0.186	<0.001	0.124		
		Inner Surface	<0.001	0.177	<0.001	0.098		
	MaCRUISE$^+$ vs. Corrected FreeSurfer*	Outer Surface	<0.001	0.226	0.037	0.028		
		Inner Surface	<0.001	0.057	0.002	0.065		
	MaCRUISE$^+$ vs.CRUISE$^+$	Outer Surface	<0.001	0.040	<0.001	0.170		
		Inner Surface	0.330	0.006	0.145	0.063		
Near WM Lesions	MaCRUISE$^+$ vs.FreeSurfer	Outer Surface	0.022	0.149	0.011	0.162	<0.001	0.315
		Inner Surface	<0.001	0.417	0.024	0.191	<0.001	**0.594**
	MaCRUISE$^+$ vs. Corrected FreeSurfer*	Outer Surface	0.870	0.010	0.249	0.059	<0.001	0.213
		Inner Surface	<0.001	0.355	0.147	0.128	<0.001	**0.526**
	MaCRUISE$^+$ vs.CRUISE$^+$	Outer Surface	<0.001	0.170	<0.001	0.137	0.003	0.103
		Inner Surface	0.013	0.110	0.005	0.115	0.002	0.121

*FreeSurfer after correction with the WM lesion masks generated by Lesion-TOADS.
Please see Table 2 for a description of effect size with Cohen's d score.

3.3. Segmentation Accuracy

3.3.1. Data

The accuracy of CCSE corrected segmentation was quantitatively evaluated with five MR volumetric images (MPRAGE T1w images with resolution$1.0 \times 1.0 \times 1.0\ mm^3$) from the OASIS dataset [132]. The images were independently labeled by an expert anatomist at Neuromorphometrics Inc. (http://www.neuromorphometrics.com/). The labeling protocols and procedures were the same as with the 45 atlases used in NLSS framework. However, the original 45 atlases have been available and used for several years of algorithm development. We felt that there existed a possibility that the performance could be over-tuned on these datasets, so the five images were retrieved at a later time and were distinct from the original 45 atlases. This approach avoided any unintentional bias that could have been present in a standard

cross-validation analysis.

3.3.2. Experiment and results

Each of the methods (JLF, NLSS, and MaCRUISE with CCSE) was run in an automated manner on each of the 5 datasets. The accuracy of NLSS and CCSE corrected segmentations were evaluated by calculating the Dice values with respect to manual segmentations. To determine statistical differences, we used a Wilcoxon signed rank test on the averaged Dice values (132 labels) for each subject with a sample size n=5 for each test. Moreover, we calculated the averaged Dice values on all cortical labels (98 labels) and WM labels (2 labels). The p value used was 0.05, which is the smallest feasible significance level of n = 5. [175]. We evaluated the sensitivity of MaCRUISE to the consistency coefficients α and β by sweeping them independently from 0 mm to 1 mm with 0.05 mm intervals and re-running all subjects with MaCRUISE for a total of 441 parameter combinations on 5 subjects. As an additional comparison for volumetric accuracy, joint label fusion (JLF) [91] was applied to the registered atlases with its default setting on the same data.

MaCRUISE improved segmentations over the entire range of consistency coefficients α and β (Figure III.9). The largest improvement averaged over all labels was more than 0.013 Dice (at $\alpha = 0.2$ mm and $\beta = 0.2$ mm). Note that CCSE is based on cortical surfaces, so the largest benefits were seen in cortical labels, while the WM labels were only affected by the inner surface consistency coefficient (α). The lower row of Figure III.9 shows a box plot of Dice improvements for $\alpha = 0.2$ mm, $\beta = 0.2$ mm and $\alpha = 0.5$ mm, $\beta = 0.5$ mm. These two sets of coefficients represent those with the largest improvements in this dataset and those that were selected as default values used in MaCRUISE respectively. Both sets of box plots reveal that the Dice values are significantly improved compared to NLSS. Surprisingly, even though the Dice values of WM were already above 0.9, they were improved by nearly 0.03 in the case of $\alpha = 0.2$ mm and $\beta = 0.2$ mm. The use of the default values sacrifices approximately 0.01–0.03 in Dice value over the optimal values. The median Dice values of CCSE were greater than those of JLF, and the CCSE achieved significant better performance than JLF in the case of $\alpha = 0.5$ mm and $\beta = 0.5$ mm.

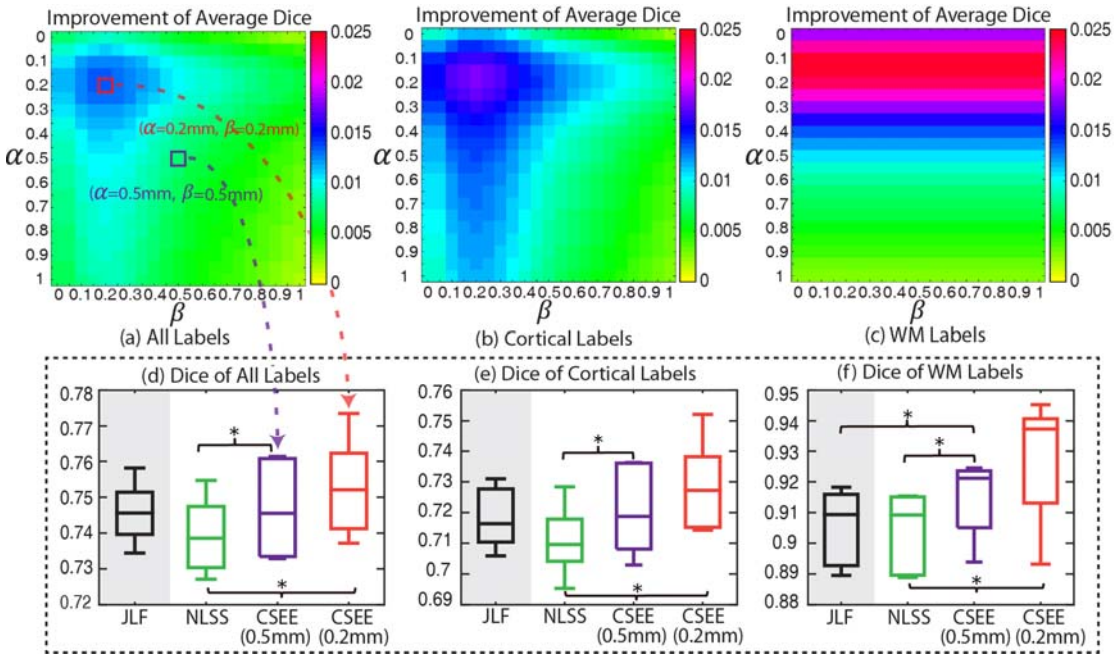


Figure III.9 This figure shows the sensitivity MaCRUISE has to α and β by varying them between 0 mm to 1 mm with 0.05 mm intervals. The upper row shows average Dice improvement from NLSS to CSEE in MaCRUISE. (a) The method has maximum improvement when $\alpha = 0.2$ mm and $\beta = 0.2$ mm. (b) The cortical labels follow a similar trend. (c) WM labels are only affected by the inner surface consistent coefficient α. (d) The box plot shows the largest Dice improvements of all 132 labels from this dataset ($\alpha = 0.2$ mm, $\beta = 0.2$ mm) compared to the default values in MaCRUISE ($\alpha = 0.5$ mm, $\beta = 0.5$ mm). (e) and (f) demonstrates the improvements of all 98 cortical labels and 2 WM labels respectively. We compare our approaches with the state-of-the-art JLF method as well. "*" indicates statistically significant difference.

3.4. Robustness of consistent cortical surfaces and segmentations

3.4.1. Data

We conducted a quantitative and qualitative robustness test on images of 200 control volunteers (100 M/ 100 F, ages 60.3 to 92.1, mean age 77.6). MPRAGE T1w MR volumetric images were collected as part of the Baltimore Longitudinal Study of Aging (BLSA) study, which is a study of aging operated by the National Institute on Aging [128, 237].

3.4.2. Experiment and results

Each of the methods (FreeSurfer, CRUISE, and MaCRUISE) was run in an automated manner on each of the 200 datasets. Average surface distance (ASD) and correlation analyses were conducted to

evaluate the global performance and consistency between MaCRUISE and the benchmarks (CRUISE and FreeSurfer). The number of global failures (outliers) was used as the robustness metric. First, the surface distance [196] between MaCRUISE and the benchmarks was examined to detect outliers. The artificial surface regions that separate the two hemispheres in FreeSurfer were excluded from the ASD measurement since MaCRUISE and CRUISE do not have such surfaces. Second, the segmentations of the lateral ventricles were examined to identify additional failures.

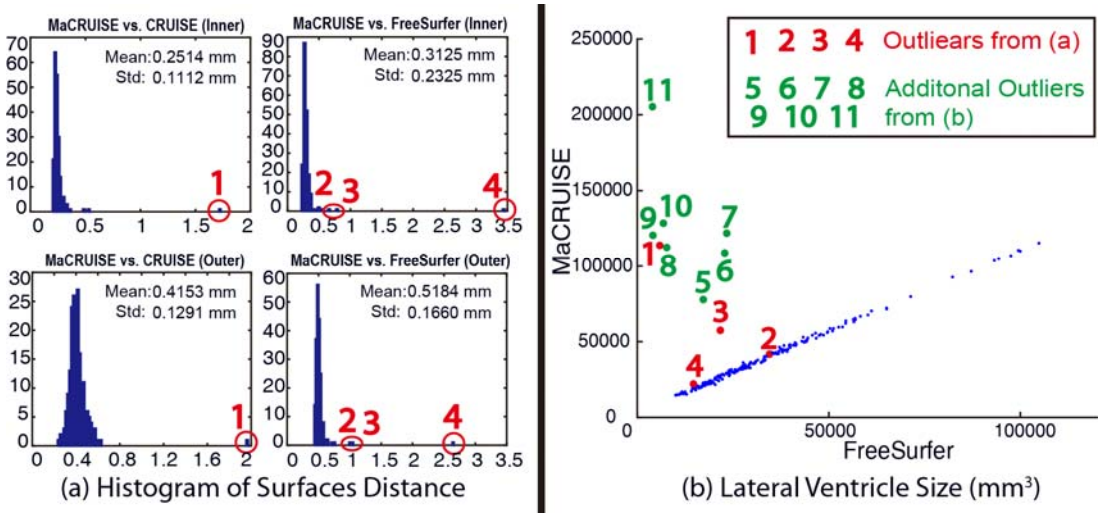


Figure III.10 This figure shows the average surface distance (ASD) between different methods and the correlation of lateral ventricle size for the population of elderly subjects. (a) The ASD between MaCRUISE with CRUISE and FreeSurfer is less than 0.5 mm in most cases, but four outliers are found. (b) The size of lateral ventricle is plotted using FreeSurfer and MaCRUISE which identified seven more outliers. A total of 11 inconsistent outliers are detected where failures occured in one of the methods. We note that FreeSurfer systematically estimates smaller ventricle size than MaCRUISE in the outliers.

Mean ASD between MaCRUISE and the benchmark algorithms are generally around or smaller than 0.5 mm (Figure III.10a). However, there are four images (marked using red numbers 1 through 4) that are located outside of a margin of 2.5 standard deviations. These large surface distances indicate that at least one of the methods failed with these images. For the ventricle volumes, a strong linear correlation was found except in seven outlier volumes (marked using green numbers 4 through 11) (Figure III.10b). Thus, a total of 11 failed volumes were automatically detected. The segmentations and surfaces of the failures for these subjects are shown in Figure III.11 (red outliers) and Figure III.12 (green outliers). The global failures

(in the red rectangles) occur in all 11 volumes for FreeSurfer and in two volumes for CRUISE. In contrast, we do not find any global failures from MaCRUISE. Therefore, none of the 11 failures are attributable to MaCRUISE. To complete the analysis, we visually inspected the surfaces and segmentations for the remaining 189 volumes and did not find any global failures for either MaCRUISE or the benchmark algorithms.

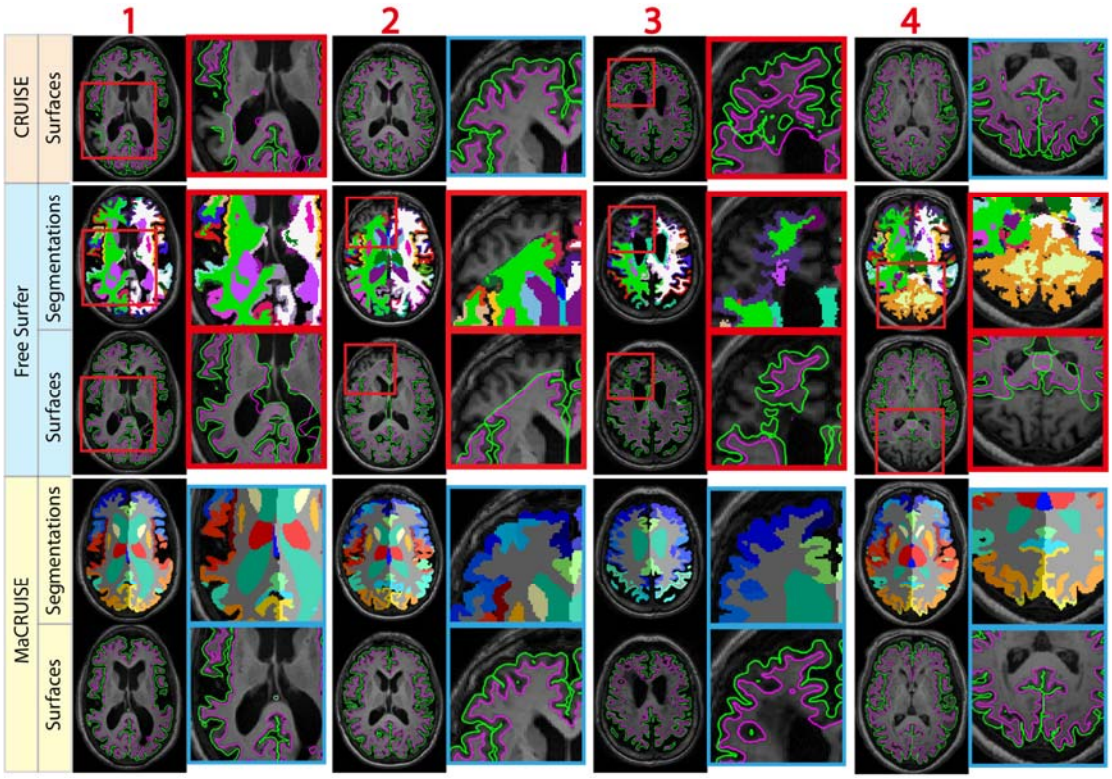


Figure III.11 The four outliers from surface distance analysis are shown. Both whole brain segmentations and cortical surfaces on axial slices are provided. The areas in red rectangles show the global failures in FreeSurfer whereas MaCRUISE did not exhibit any such failures.

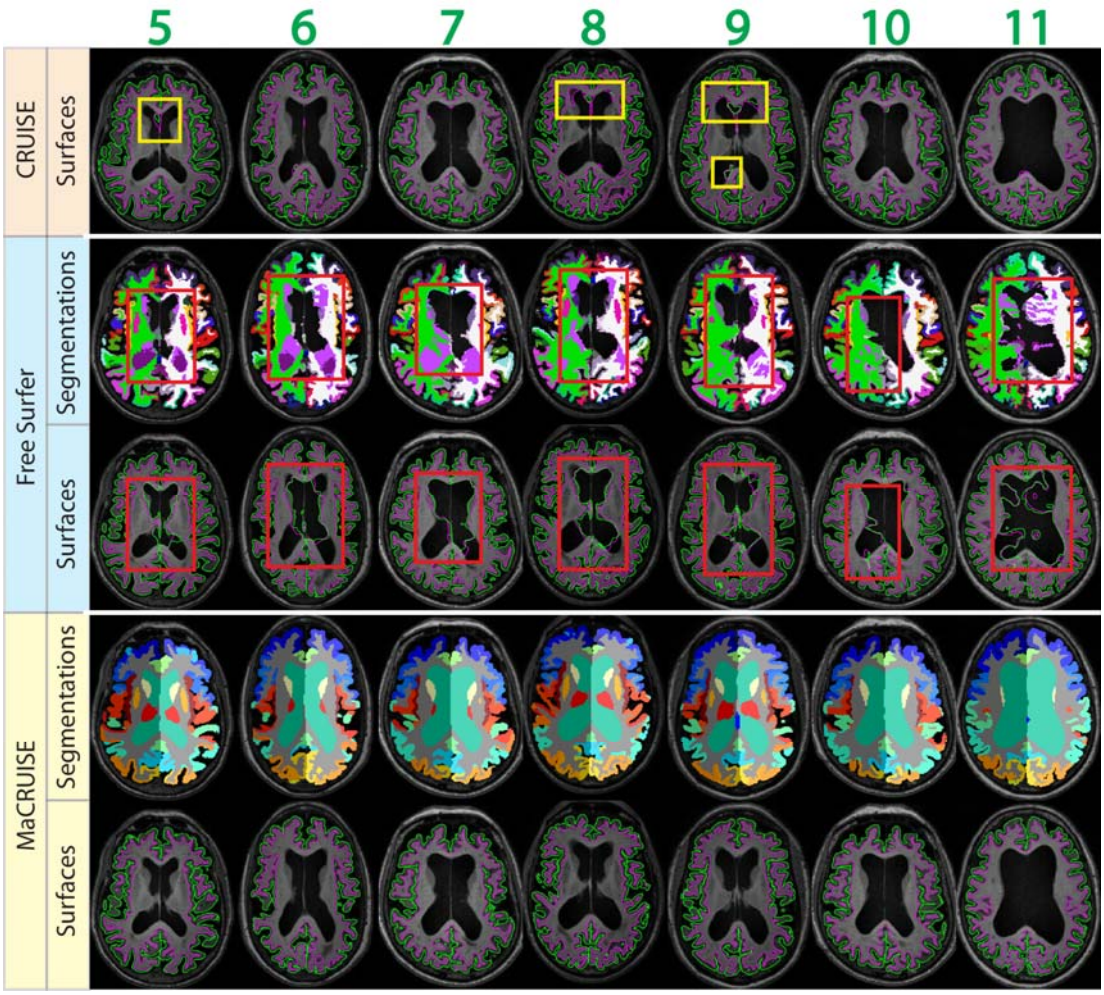


Figure III.12 The seven outliers from inconsistent lateral ventricle size are shown. Both whole brain segmentations and cortical surfaces on axial slices are provided. The areas in red rectangles show the global failures while the areas in yellow rectangles show the local inaccurate surfaces. MaCRUISE did not exhibit such failures in any images.

4. Discussion

MaCRUISE is an open framework that allows users to replace NLSS and TOADS with other multi-atlas or fuzzy segmentation approaches. MaCRUISE^{+} is an example of replacing TOADS with Lesion-TOADS (also available as open source), which incorporates the MaCRUISE framework in the case of pathology. MaCRUISE and MaCRUISE^{+} are publicly available as open source software through the JIST software package (http://www.nitrc.org/projects/jist/) [238, 239]. MaCRUISE is also implemented as a

plugin called “PlugInsMaCRUISE” in CRUISE software. The source code is available using CVS access (www.nitrc.org:/cvsroot/toads-cruise).

While FreeSurfer has been widely used and is regarded as the *de facto* standard method for generating whole brain segmentations and cortical surface locations, FreeSurfer failed globally in about 5% of older adult populations from a BLSA sample dataset. Even though manual correction would probably address these failures, the time required for manual correction makes it undesirable. Compared with FreeSurfer, the proposed MaCRUISE achieves (1) greater robustness in older populations, and (2) comparable accuracy on normal healthy images. To the best of our knowledge, this is the first work that integrates multi-atlas segmentation into cortical reconstruction. Moreover, the perspective of using multi-atlas segmentation and the proposed consistency adjustment approaches could be integrated into FreeSurfer, which might improve the robustness of FreeSurfer on older adult populations. The statistical analyses using both paired t-test and effect size analyses indicate a lack of evidence for systematic differences ($p>0.05$ or $d<0.5$) between MaCRUISE and CRUISE when examining datasets on which CRUISE has been validated. Therefore, the proposed MaCRUISE achieves comparably accurate cortical surfaces compared with the CRUISE method while providing consistent whole brain segmentation when the original CRUISE method does not.

Consistency is another essential challenge in clinical and scientific analyses of MR brain images. MaCRUISE establishes consistent brain segmentation and cortical surfaces by combining multi-atlas segmentation with cortical reconstruction. However, the naive strategy of directly deploying CRUISE after NLSS (NLSS+CRUISE) did not yield accurate surfaces. With the specific contributions of this work (i.e., segmentation fusion, MaACE, and CCSE), MaCRUISE improves both surface and volumetric accuracy (Figure III.9).

While MaCRUISE shows great promise for consistent multi-atlas segmentation and cortical reconstruction, there are certain areas that warrant further investigation. We used NLSS framework as the multi-atlas segmentation algorithm and employed TOADS as the fuzzy segmentation approach. Since the two different segmentation methods are conducted independently, inconsistency exists between their

segmentations. To reconcile the inconsistency, the segmentation fusion method is used in the MaCRUISE. While this is highly successful, we do not claim optimality of using NLSS and TOADS. Using other multi-atlas or fuzzy segmentation methods might yield a better performance when establishing consistent multi-atlas segmentation and cortical reconstruction. Since the proposed method is an open framework, users are encouraged to explore methods other than NLSS and TOADS freely. Recently, [240] indicated that the Advanced Normalization Tools (ANTs) based framework achieved a higher predictive performance than FreeSurfer by evaluating thickness-based prediction of age and gender. Such analyses of predictive power are relevant, but depend heavily on the population context. For example, a method that exaggerated aging effects would have greater power to detect aging, but could be less accurate in an absolute sense and potentially less useful when aging is not an effect of interest. Examining predictive power of MaCRUISE versus other approaches would be a valuable direction for further investigation.

There are potential drawbacks in the presented MaCRUISE approach. First, the robustness of multi-atlas segmentation framework comes at the cost of computational complexity from both expensive non-rigid registration and non-local correspondences calculations. Empirically, MaCRUISE typically takes approximately 38 h. This is broken up into NLSS framework ($\approx$36 h), TOADS segmentations ($\approx$1 h) and cortical reconstruction ($\approx$1 h) on a single core of an Intel Xeon W3550 4 Core CPU (64 bit Ubuntu Linux 14.04). As a result, MaCRUISE has much greater time complexity than CRUISE (<2 h) or FreeSurfer (<15 h) on the same machine. Recently, a learning based multi-atlas framework called multi-atlas learner fusion (MLF) has been proposed to reduce the time that multi-atlas segmentation requires to less than 10 min [149]. Replacing the NLSS by MLF would be a promising way of reducing the total computing time of MaCRUISE to less than 3 h. Second, both the multi-atlas segmentation and TOADS results are functions of the imaging sequence and are thus biased based on the sequence [241]. Contrast synthesis may become an important approach to ensure performance across imaging sequences e.g., following [242]. Third, the cortical surfaces derived between subjects do not have pre-defined correspondence, which necessitates surface and/or image registration. Finally, we do not claim the optimality of the number of atlases used in

the experiments. Fifteen atlases were chosen based on our previous experience with the collection of 45 available atlases [92, 149]. Users may wish to optimize the number of atlases for their application via cross-validation or bootstrapping [203].

5. Conclusion

Herein, we introduced MaCRUISE, a novel consistent whole brain segmentation and cortical surface reconstruction approach using multi-atlas segmentation. MaCRUISE achieved greater robustness on T1w MRI images from older adults than FreeSurfer without compromising on the accuracy of normal healthy images. MaCRUISE achieves significantly greater volumetric accuracy than solely using NLSS multi-atlas segmentation. MaCRUISE+ established consistent cortical surfaces and volumetric segmentations for images with WM lesions.

From landmark based surface validation, we demonstrated that MaCRUISE achieved consistent whole brain multi-atlas segmentation and cortical reconstruction (Figure III.7) without compromising accuracy (**Error! Reference source not found.** and Table III.2) since the differences between MaCRUISE and the benchmark algorithms are either "trivial" ($d<0.2$) or "small effect" ($d<0.5$). MaCRUISE^{+} was similarly accurate (**Error! Reference source not found.** and Table III.4) and provided consistent whole brain segmentations (Figure III.8). MaCRUISE allows users to control the consistency level between whole brain segmentations and reconstructed surfaces using the consistency coefficients α and β. The refined segmentations achieved robust improvements on a wide range of different $\alpha's$ and $\beta's$ (0 mm to 1 mm) compared to NLSS (Figure III.9). Finally, by evaluation of gross failures on a collection of 200 volumetric images from older adults, MaCRUISE is more robust to errors in surface segmentation than CRUISE or FreeSurfer (Figure III.10, Figure III.11 and Figure III.12). In all cases, MaCRUISE achieved consistent segmentations of the cortical surface and all brain labels, which was not the case for either CRUISE or FreeSurfer.

Chapter IV. Improved Stability of Whole Brain Surface Parcellation with Multi-atlas Segmentation

1. Introduction

Mapping the anatomical and functional relationships in the human brain is essential for image-based brain mapping. Detailed and consistent whole brain volume segmentation and surface parcellation provide the tools to establish such relationship by classifying the brain tissue and cortex into different functional regions. Many previous efforts have been proposed to perform the whole brain segmentation or surface parcellation; however, only few works provided consistent whole brain segmentation and surface parcellation [58-60]. FreeSurfer has been widely accepted as the *de facto* standard for consistent whole brain segmentation and surface parcellation using "surface-to-volume" strategy [59, 106, 111]. Recently, another "volume-to-surface" approach called multi-atlas cortical reconstruction using implicit surface evolution (MaCRUISE) was proposed to establish the consistent and robust whole brain segmentation and showed its advantages in certain aspects [225, 243]. MaCRUISE combined the multi-atlas segmentation (MAS) [153] with the Cortical Reconstruction using Implicit Surface Evolution (CRUISE) surface reconstruction [58] to achieve the consistent volume segmentation and cortical surfaces. Although it performed detailed volume segmentation (with 132 labels) and reconstructed consistent cortical surfaces, the MaCRUISE approach did not provide the cortical surface parcellatio. To understand the human anatomical and functional relationships, more regional features from cortical surfaces (e.g., area, thickness, curvature) are appealing to quantify brain anatomy for population analyses [150, 244, 245]. This work is motivated by the previous learning based surface parcellation methods [210, 246-248].

Herein, we extend the MaCRUISE method to MaCRUISE surface parcellation (MaCRUISEsp) by developing the volume segmentation based surface parcellation (VSBSP) and topological correction functionalities (Figure IV.1). MaCRUISEsp has following advantages: (1) The parcellated central surface (located inside the gray matter) was provided along with the traditional inner surface (white matter surface)

and outer surface (pial surface). The parcellated surfaces have been used in a recent gray matter based DTI mapping method [249]. (2) 98 cortical labels were provided by MaCRUISEsp for inner, outer and central surfaces respectively. To validate the method, 42 T1-weighted (T1w) MR volumes (21 scan-rescan longitudinal pairs from Kirby21 dataset [250]) were used. The proposed method achieved 0.94 on median Dice similarity coefficient (DSC) for central surface parcellation and superior performance on inner surface parcellation compared with FreeSurfer.

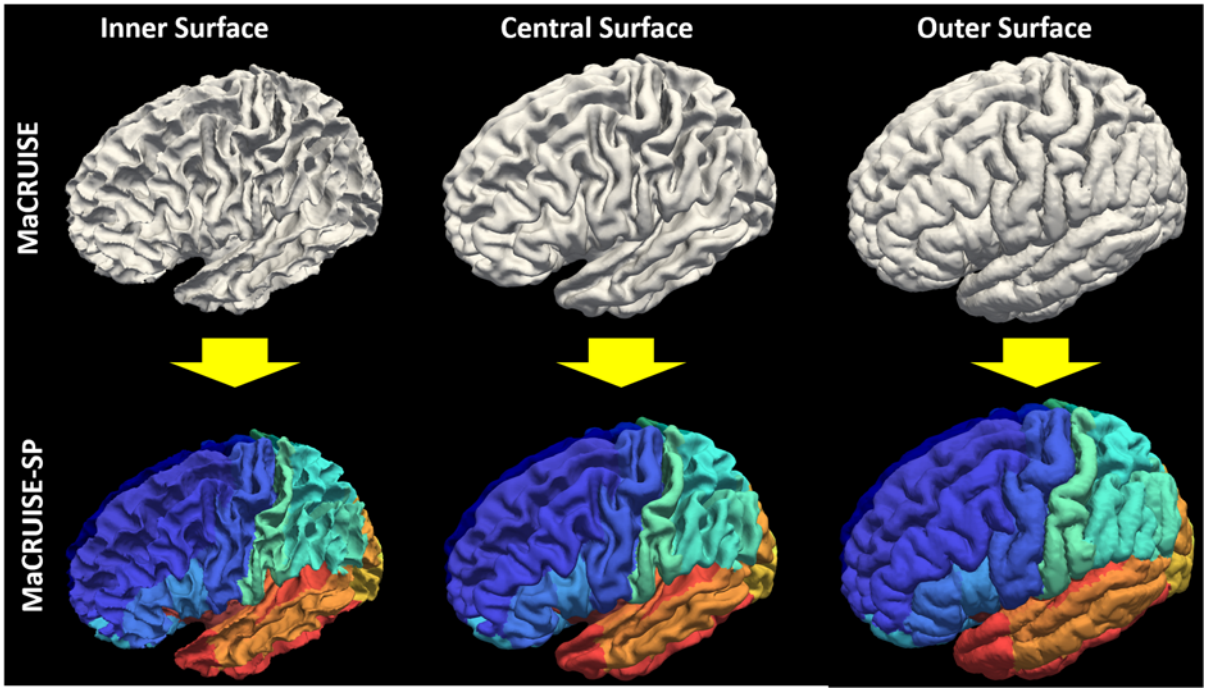


Figure IV.1 The motivation of MaCRUISEsp was to provide quantitative surface labels for MaCRUISE surfaces.

2. Method

2.1. Multi-atlas Segmentation based Surface Reconstruction

The input image of the entire processing pipeline was a single T1w brain magnetic resonance image (MRI). First, non-local spatial STAPLE (NLSS) multi-atlas segmentation framework was used to achieve whole brain segmentation[153]. Then, the MaCRUISE approach was deployed on the target image to obtain consistent whole brain segmentation and cortical surface reconstructions [225, 243]. From MaCRUISE, the

inner, central, and outer surfaces were reconstructed, which were spatial consistent with volumetric segmentation (Figure IV.2).

2.2. Volume Segmentation Based Surface Parcellation

The central surface was parcellated from the whole brain volumetric segmentation. Briefly, we propagate volume labels to the central surface using the nearest label projection. For each vertex on the surface, the corresponding volumetric cortical label was assigned as the label of such vertex. This process was performed on all vertices to get the entire central surface parcellated. Since the central surface were bounded in the gray matter (GM), each vertex on the central surface were assigned a cortical label (rather than white matter or background labels). The BrainCOLOR atlas/protocol [170] was used in the proposed MaCRUISEsp framework to parcellate each surface to 98 cortical labels.

2.3. Topological Correction

In the BrainCOLOR protocol, each label represented a brain region with one connected component (OCC). However, after propagating the volumetric labels to surfaces, the OCC was not always ensured due to the topological mismatch. Therefore, the topological correction (TC) step was introduced to ensure each surface label to be an OCC. First, we detect the number of components of each label using "trimesh2" software (http://gfx.cs.princeton.edu/proj/trimesh2/). Then, all components on the surfaces (except the largest one) were marked as "need to fix". After repeating the previous steps for all labels, we marked all non OCC vertices as "need to fix" and fixed all of them using an iterative nearest neighbor filling strategy described in [234]. In each iteration, the remaining "needs to fix" vertices were filled by the most commonly occurring surface labels around their neighbor as the following equation:

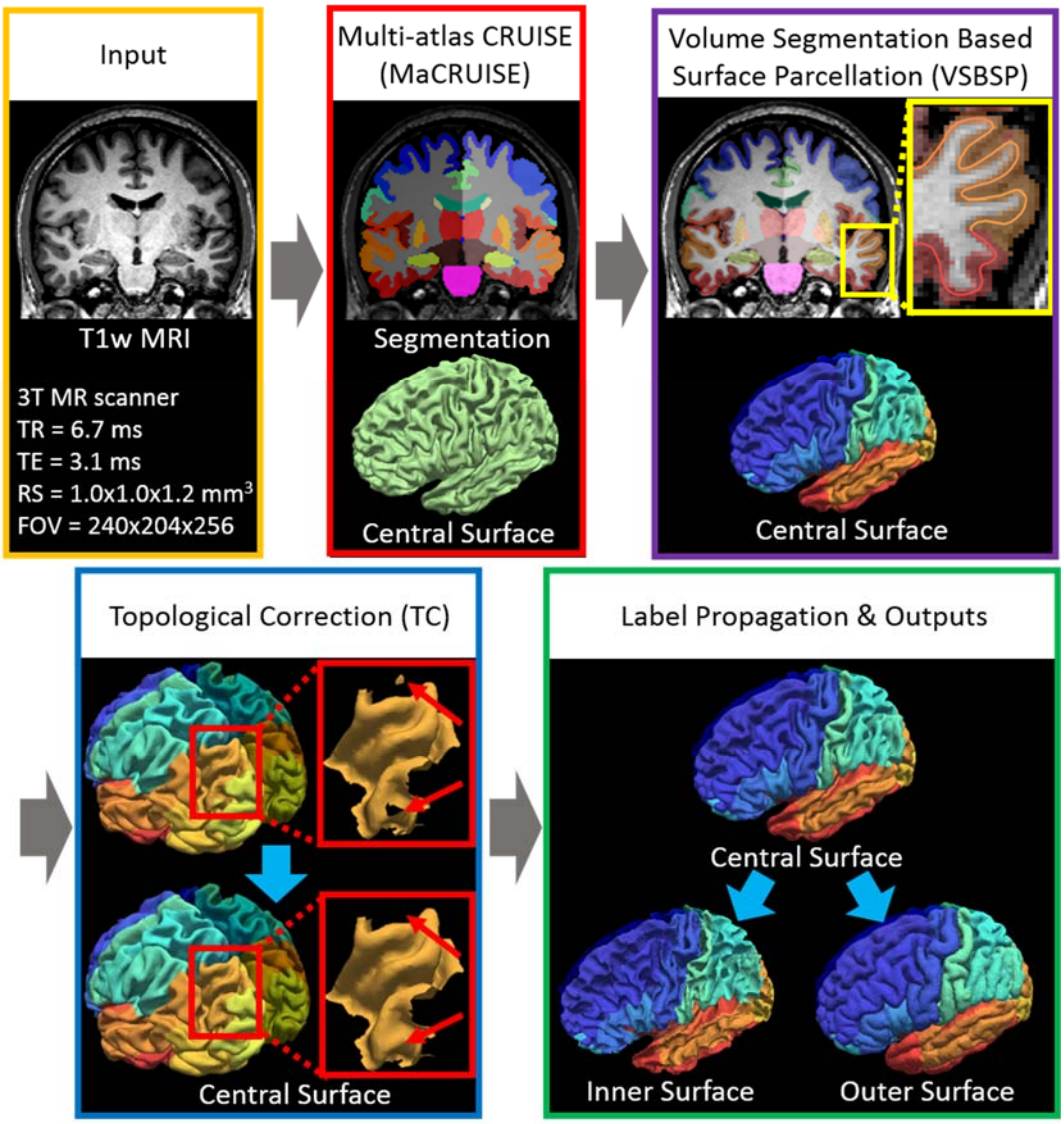


Figure IV.2 Work flow of MaCRUISEsp. (1) MaCRUISE was deployed on a single T1w MRI volume to achieve consistent whole brain segmentations and cortical surfaces (inner, central and outer). (2) Surface parcellation was performed on central surface using volume segmentation based surface parcellation (VSBSP). (3) The topological correction is conducted to ensure the one connected component (OCC) for each surface region. (4) The inner and outer surfaces were parcellated on by propagating the labels from central surfaces. Finally, 98 cortical labels were assigned for each surface.

$$\widehat{L_i} = \underset{n}{\operatorname{argmax}} \sum_{k \in \gamma(i)} (L_k == n), \qquad n \in [1,2,3 \dots, N] \tag{4.1}$$

where k indicates the indices of the labeled voxels (with the label L_k) around the unlabeled voxel i. The n represents all possible N cortical labels. After the topological correction, the central surface was corrected to OCC for each label.

2.4. Surface Label Propagation

After previous steps, the central surface was parcellated and corrected. Then, the inner and outer surfaces were parcellated by propagating the labels from the central surface. For each vertex on inner (or outer) surface, the label was propagated from another vertex on central surface, who had the smallest Euclidean distance to inner (or outer) vertex. To handle the label propagation on the back to back cortical surfaces with narrow sulcus, central vertices outside (inside) the normal plane of the vertices on the inner (outer) surfaces were considered in the distance calculation. Particularly, the nearest searching was restricted by the normal half of the plane that perpendicular to the norl direction.

3. Experiments

3.1. Data

42 T1w MPRAGE MRI volumes (21 scan-rescan patients) from Kirby21 dataset [250] were used in the empirical validation to evaluate the reproducibility of the proposed MaCRUISEsp framework. The cohort consists of 11 male and 10 female patients, were collected from 3T Philips Achieva scanner with parameters: TR = 6.7 ms, TE=3.1 ms, resolution (RS) = $1.0 \times 1.0 \times 1.2\text{mm}^3$ and the field of view (FOV) = $240 \times 204 \times 256\text{mm}$.

3.2. Experiments

The MaCRUISEsp pipeline (Figure IV.3) was deployed on the dataset. Then the Dice similarity coefficient (DSC) was calculated on the parcellated scan-rescan whole brain surfaces. Briefly, each rescan surface was registered to the scan surface using rigid registration. Then the correspondence of vertices on the paired surfaces were established using the closest point matching. Finally, the DSC was derived by dividing the number of matched vertices by the average number of the vertices on the registered scan-rescan surfaces. The Wilcoxon signed rank test [175] was used for statistical analyses. All claims of statistically significance in this paper are made using the Wilcoxon signed rank test for $p < 0.01$.

4. Results

The qualitative results (Figure IV.3) as well as the quantitative results (Figure IV.4) on the registered scan-rescan surfaces were demonstrated. In Figure IV.4, the reproducibility results on inner and outer surfaces using FreeSurfer Destrieux 2009 atlas were employed as the baseline performance. Note that in FreeSurfer, the Destrieux atlas has fewer labels (75 labels) on surfaces compared with the BrainCOLOR atlas (98 labels) in MaCRUISEsp framework, which would bias FreeSurfer toward larger ROIs and higher DSC. The Pearson correlation results (surface area and cortical thickness) across 21 scan-rescan pairs for all cortical labels were provided in the Figure IV.5.

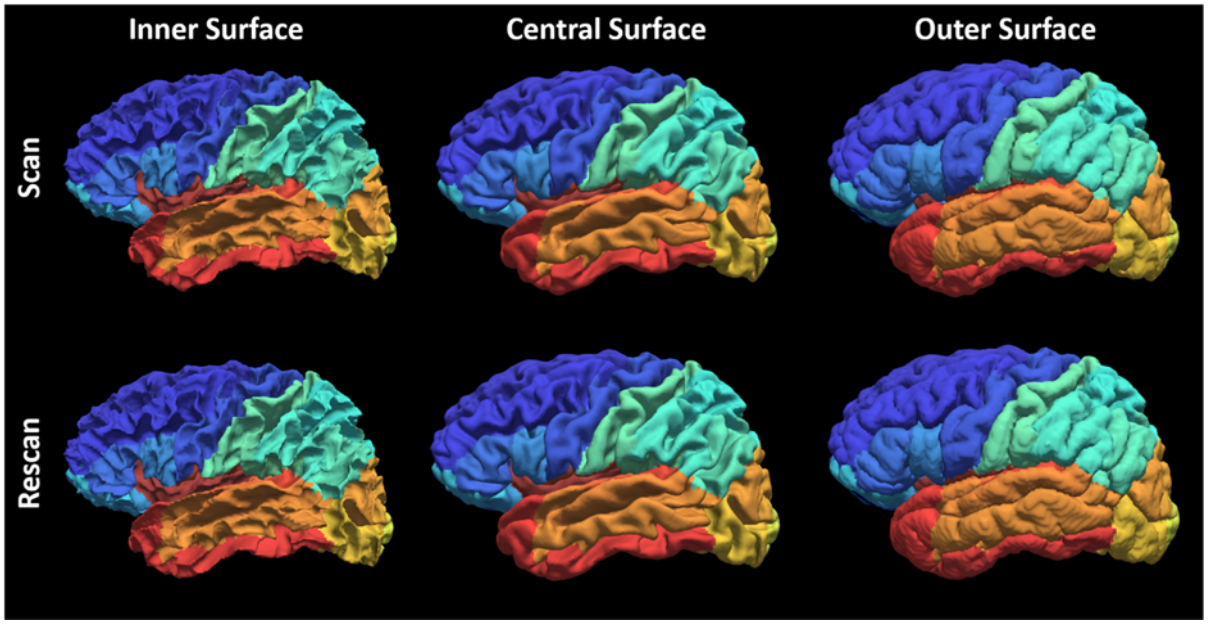


Figure IV.3 Qualitative reproducibility results on the surface parcellation between a randomly selected scan-rescan patient using MaCRUISEsp.

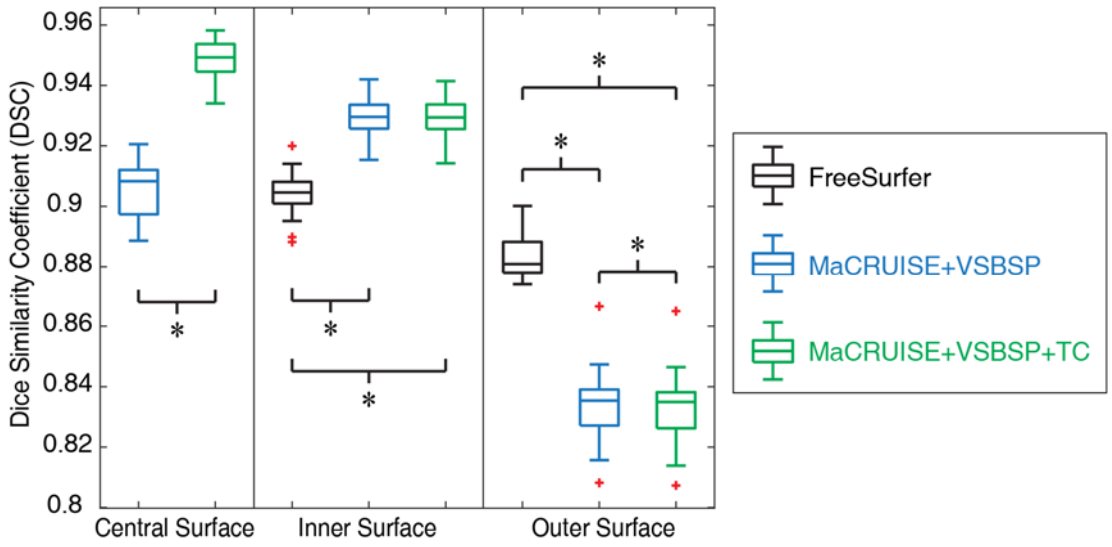


Figure IV.4 Quantitative segmentations results on the surface parcellation for the entire Kirby21 cohort. The reproducibility on inner and outer surfaces using FreeSurfer's Destrieux 2009 atlas (75 labels) were employed as the baseline. The MaCRUISE+VSBSP method as well as the MaCRUISEsp (MaCRUISE+VSBSP+TC) method using BrainCOLOR atlas (98 labels) were presented. The symbol "*" indicated the differences are significant for the Wilcoxon signed rank test for $p < 0.01$.

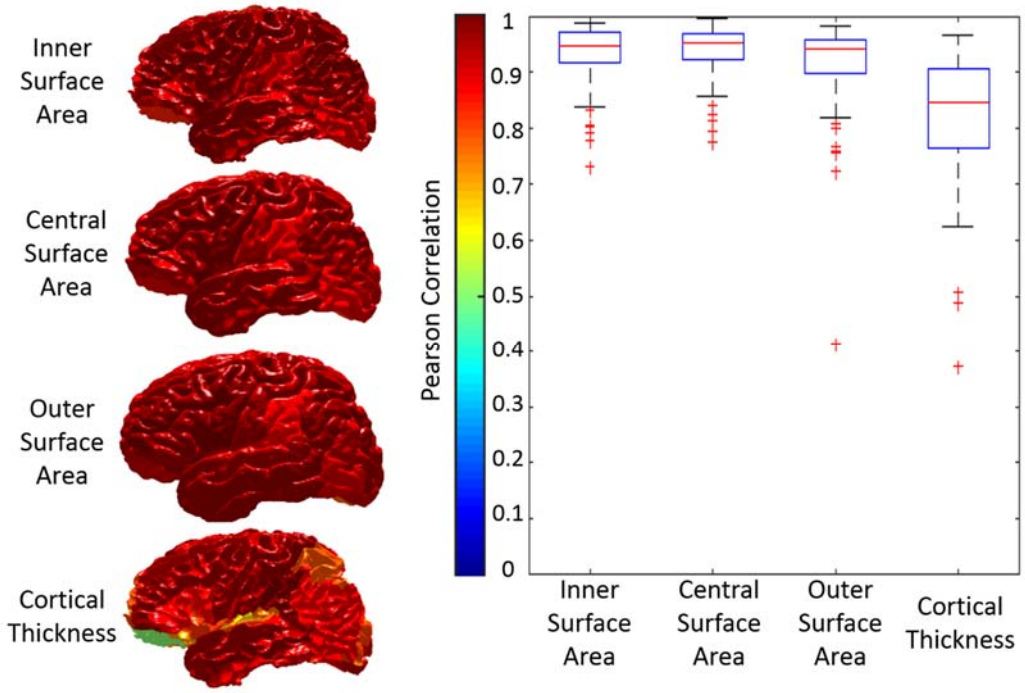


Figure IV.5 The reproducibility of surface metrics (surface area and cortical thickness) were shown. The Pearson correlation values for four metrics on each label were shown in the left panel. The color of each label corresponds to the Pearson correlation value showed in the color bar. Then, the qualitative results of all labels were shown as the boxplot in the right panel.

5. conclusion and Discussion

We present the MaCRUISEsp method for the whole brain surface parcellation. From the experimental results, the “volume-to-surface” strategy with topological correction provided us 0.95, 0.92 and 0.85 median DSC for central surface, inner surface and outer surface respectively (Figure IV.4). The results showed that the MaCRUISEsp provided the central surface parcellation, which not was typically provided by FreeSurfer. With topological correction, the MaCRUISEsp obtained the generally better reproducibility than without using topological correction. The proposed methods achieved significantly higher reproducibility than FreeSurfer on inner surface parcellation while the FreeSurfer achieved significantly higher reproducibility than the proposed methods on the outer surface parcellation. Note that the comparison was made in the situation that more labels were provided by MaCRUISEsp (98 labels) compared with FreeSurfer (75 labels). For a more thoughtful analysis, the reproducibility on the surface metrics were provided in Figure IV.5. Qualitatively, the results from proposed methods were encouraging, but are not directly comparable to FreeSurfer as the two approaches use different definitions of cortical labels.

Chapter V. Data-driven Probabilistic Atlases Capture Whole-brain Individual Variation

1. Introduction

Probabilistic atlases play important roles in understanding the spatial variation of brain anatomy, in visualization, and in the processing of data. The basic framework of making probabilistic atlases is to bring the image data from the selected subjects into an atlas space by rigid or non-rigid registration [251]. Then, probabilistic maps are generated by averaging the segmentations of regions from a specific group of subjects with similar demographic data, such as age, sex and from the same site. However, the inter-subject variability is normally larger than the inter-group variability, which causes the group-based scheme to fail to capture a great deal of individual variation.

To overcome the large inter-subject variability, Commowick et al. proposed the "Frankenstein's creature paradigm" to build a personal specific anatomical atlas for head and neck region [71]. The paradigm first selected regional anatomical atlases based on a training database then merged them together into a complete atlas. However, this framework cannot be directly applied on making probabilistic atlases since each probabilistic atlas is averaged from a group of segmentations. Moreover, compared with the 105 CT images used as the database in Commowick's framework, we employ 2349 heterogeneous MRI images in our framework.

In this chapter, we propose a large-scale data-driven framework to learn a dictionary of the whole brain probabilistic atlases (132 regions) from 1888 heterogeneous 3D MRI training images. The novel contributions of this chapter are (1) providing a new data-driven perspective of making whole brain probabilistic atlas, (2) generating the more accurate personal specific probabilistic atlases by using the large-scale data from different groups and even different sites, and (3) achieving low computational cost of applying the learned dictionary on new subjects.

2. Data

The complete dataset aggregates 9 datasets (7 are publicly available), with a total 2349 MRI T1w 3D images obtained from healthy subjects. The 2349 images are divided to 1888 training datasets and 431 testing datasets based on the site and demographic information. The entire 1888 training images are used to train the data-driven framework called "Training Set 1888". A subset of 720 training images called "Training Set 720" are employed to generate group atlases. The 431 testing datasets are selected with at least 15 available withheld subjects in each group.

3. Methods

The proposed data-driven framework consists of two main portions. First, a dictionary is learned by the training data (Figure V.1). Second, the learned dictionary is applied to a new subject by affine alignment to MNI space (Figure V.2).

3.1. Get Regional Segmentations and Point Distribution Model

All 720 training subjects were first affinely registered [89] to the MNI305 atlas [171]. Then, a state-of-the-art multi-atlas segmentation (including atlases selection, pairwise registration [88], label fusion [92] and error correction [173]) was performed on each subject. 45 MPRAGE images from OASIS dataset were used as original atlases which are manually labeled with 133 labels (132 brain regions and 1 background) by the BrainCOLOR protocol [170]. Here, we define S_i as the whole brain segmentations with 133 labels and the $i \in \{1,2 \dots,720\}$ represent different subjects.

Then, a mean segmentation $\bar{S}$ is generated from all $\{S_i\}_{i=1,2,\dots,720}$ by majority vote label fusion. Since the $\bar{S}$ is smooth, it is a good template of making surface meshes for 132 regions. When the meshes are generated, the vertices $\bar{V}^k$ on the mean segmentation $\bar{S}$ can be propagated to individual segmentations [252]. We non-rigidly register each S_i to $\bar{S}$ and get the diffeomorphism $\phi_i(\cdot)$ [88]. The inverse transformation $\phi_i^{-1}(\cdot)$ is used to propagate the $\bar{V}^k$ back to individual vertices V_i^k (Figure V.1).

3.2. Clustering

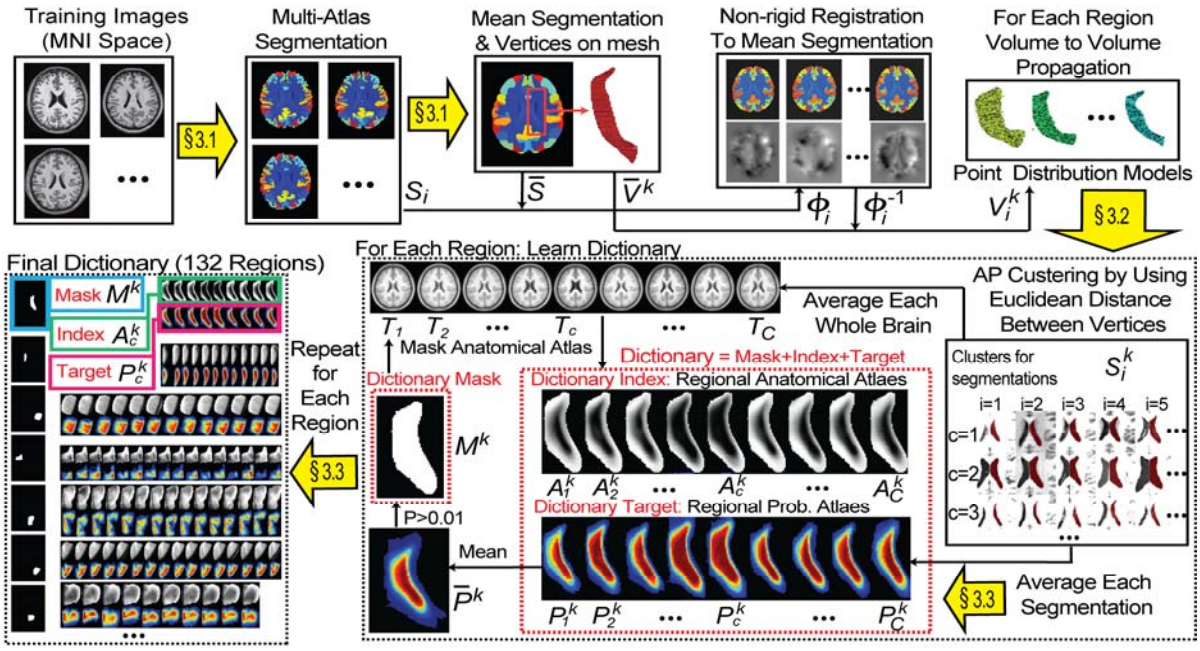


Figure V.1 Flowchart of training a data-driven dictionary of whole brain probabilistic atlas.

The Affinity Propagation (AP) clustering method [253] was used to cluster the similar segmentations by using the V_i^k as features. The advantage of AP clustering is it can adaptively cluster the samples into a number of clusters without providing the number of clusters. For region k, the negative mean Euclidian distance $d^k(i,j)$ between vertices V_i^k and V_j^k is used as the similarity measurement for AP clustering,

$$d^k(i,j) = -\frac{1}{M_k}\sum_{m=1}^{M_k}\left\|v_{i,m}^k - v_{j,m}^k\right\|^2 \tag{5.1}$$

where the $v_{i,m}^k$ and $v_{j,m}^k$ are the m^{th} vertex in the vertices V_i^k and V_j^k. M_k is the size of the vertices V_i^k or V_j^k. Typically, 7~20 reliable clusters are generated for each region.

3.3. Learn Dictionary

3.3.1. For One Region

The regional anatomical atlases A_c^k are the "dictionary index" and the regional probabilistic atlases

P_c^k corresponding "dictionary target" (red rectangular in Figure V.1). First, the regional probabilistic atlases P_c^k for the cluster c is obtained by averaging the segmentations that belong to that cluster.

$$P_c^k = \frac{1}{L_c}\sum S_i^k, \quad T_c = \frac{1}{L_c}\sum I_i, \quad all\ i \in cluster\ c \tag{5.2}$$

where S_i^k is the segmentation of region k from subject i and L_c is the number of segmentations in the cluster c. The anatomical atlases for each cluster are found by (2) and I_i is the whole brain anatomical image from subject i.

However, as shown in Figure V.1, each T_c is a whole brain anatomical atlas rather than a regional anatomical atlas for region k. So, we need to extract the target area for region k by a reasonable mask M^k.

To get the mask M^k, we (1) average all $\{P_c^k\}_{c=1,2,\ldots,C}$ to $\bar{P}^k$ (2) obtain the M^k by setting the threshold $\bar{P}^k > 0.01$. The obtained mask will be much larger than any individual segmentation, which covers the potential spatial locations of region k.

Finally, we apply the mask M^k on every T_c to get a regional anatomical atlas A_c^k

$$A_c^k = T_c \circ M^k \tag{5.3}$$

The masked A_c^k is corresponding to the regional probabilistic atlas P_c^k.

3.3.2. For Whole Brain

We repeat the "For One Region" steps 132 times (for all regions except background) to get the whole brain dictionary as shown in the lower left part of Figure V.2.

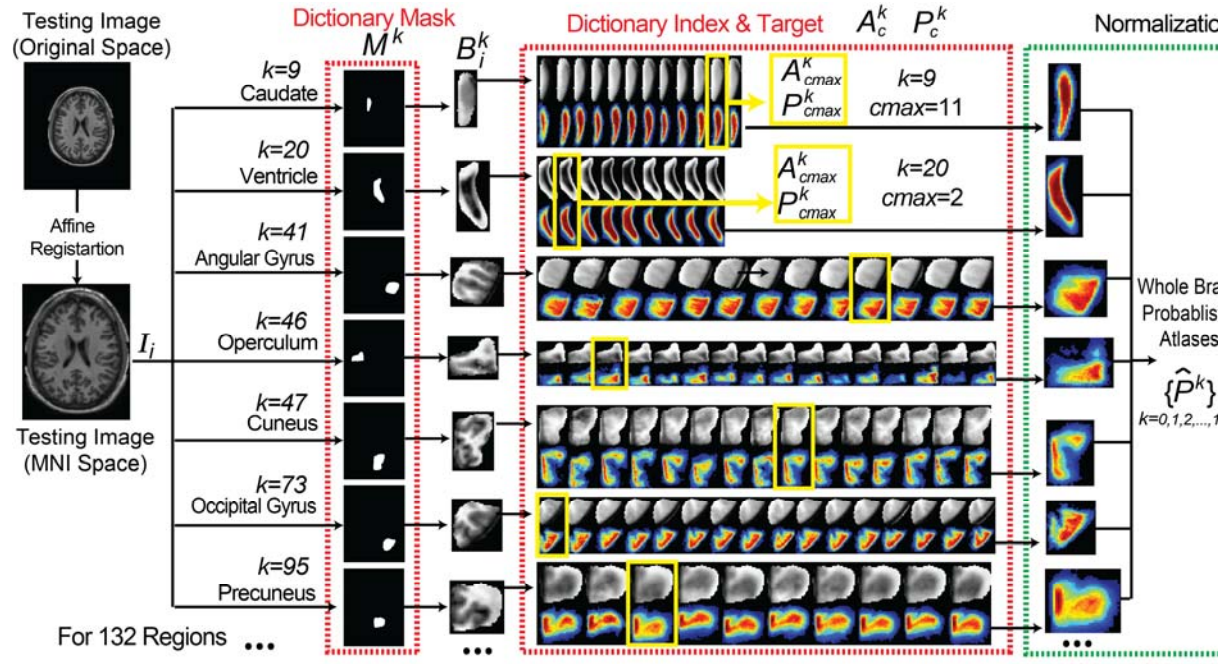


Figure V.2 Flowchart of applying the dictionary to customize a probabilistic atlas for a new subject.

3.4. Apply Dictionary on New Subjects

To efficiently establish an individual whole brain probabilistic atlases, each target subject is affinely aligned [89] to the MNI305 atlas to get I_i (Figure V.2). Then, the regional intensity B_i^k can be masked out by

$$B_i^k = I_i \circ M^k \tag{5.4}$$

By comparing the B_i^k to our learned dictionary, the index can be obtained by finding the most correlated regional anatomical atlas A_c^k. The correlation metrics used here is the Pearson correlation. Once the index c_{max} is found, the corresponding $P_{c_{max}}^k$ is chosen as the regional probabilistic atlas for the new subject.

$$c_{max}^k = \arg\max_c \ corr\left(A_c^k, B_i^k\right), \qquad c \in \{1,2,\dots,C\} \tag{5.5}$$

Repeating equations (4) and (5) for all regions, we find the 132 most correlated regional probabilistic atlases for the new subject.

3.5. Normalize to Whole Brain Atlas

Since the regional probabilistic atlases were chosen independently, the total probability for a voxel might be larger or smaller than 1. To normalize them to a complete set of whole brain probabilistic atlases, we employed a whole brain tissue probabilistic mask M^t from 1888 training image which contains the voxels with tissue probability greater than 0.05. For each voxel (x, y, z) within the mask M^t, the 132 regional probabilistic atlases are normalized to 1; otherwise we keep it untouched.

$$\hat{P}^k(x,y,z) = \begin{cases} \dfrac{P^k_{c^k_{max}}(x,y,z)}{\sum_{k=1}^{132} P^k_{c^k_{max}}(x,y,z)} & x,y,z \in brain\ mask\ M^t \\ P^k_{c^k_{max}}(x,y,z) & otherwise \end{cases} \tag{5.6}$$

Last, the probability of background $\hat{P}^0(x, y, z)$ is obtained by

$$\hat{P}^0(x,y,z) = 1 - \sum_{k=1}^{132} \hat{P}^k(x,y,z) \tag{5.7}$$

The set of $\{\hat{P}^k(x, y, z)\}_{k=0,1,2,\ldots,132}$ is the normalized data-driven whole brain probabilistic atlases for the new subject. For each voxel in the whole brain probabilistic atlases, the total probability of 132 labels and background is 1.

4. Experimental Results

Two metrics are employed in the experiments. First, the Jensen-Shannon (JS) divergence is used to assess the spatial similarity between the probabilistic atlases and the target segmentations for each testing subject [254]. Here, the "target segmentations" means the multi-atlas segregations for the withheld testing images and the manual segmentations for the OASIS images. The smaller JS divergence value is, the more similar the two spatial distributions are. So, smaller is the better for JS.

Second, to compare the different probabilistic atlases more intuitively, we apply "naive segmentation" on whole brain by choosing labels with the highest probability for each voxel. Notice that we are not providing a novel segmentation algorithm. Instead, we compare the spatial accuracy of different probabilistic atlases by using the naïve segmentation since this approach is entirely depending on the

probability. Then, the Dice similarity measures the overlaps between the naive segmentations and the target segmentations.

All statistical significance tests are made using a Wilcoxon signed rank test ($p<0.01$). Creating a whole brain probabilistic atlas for a new subject can be done with 1 rigid registration and 12 seconds of CPU time (Xeon W3520 2.67GHz).

4.1. Evaluation by Withheld Testing Data

Figure V.3 and Figure V.4 show the results by using withheld testing subjects. The green boxplots represent the average JS or Dice values by applying the probabilistic atlases from all the other 17 group atlases for one testing subject. The blue, red and orange boxplots show the JS or Dice values by using the corresponding group probabilistic atlases, data-driven probabilistic atlases from Training Set 720 and from Training Set 1888.

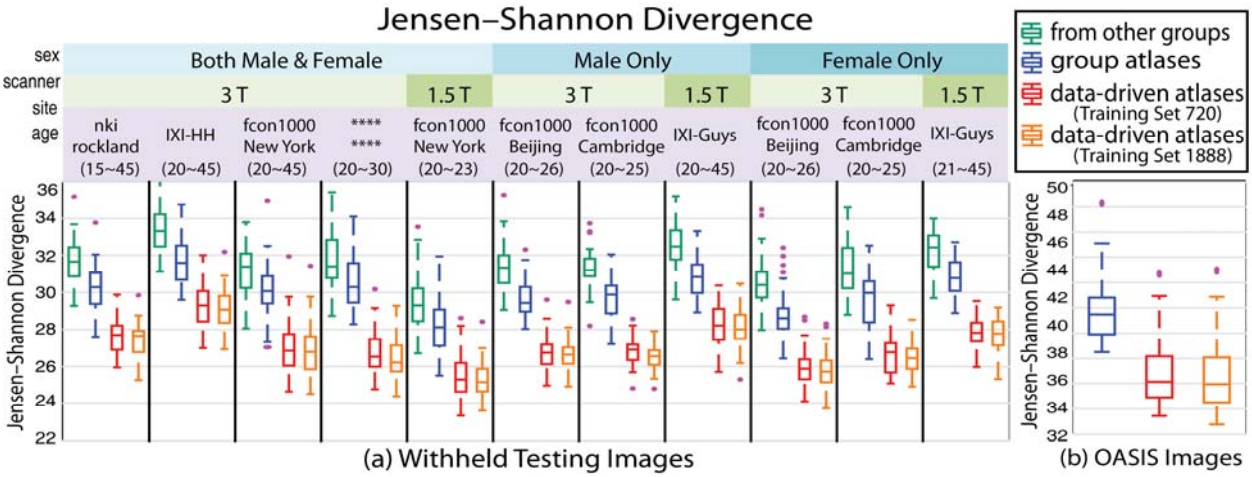


Figure V.3 Jensen-Shannon divergence. The comparisons of JS divergence for different atlases are all significantly different for both withheld and OASIS testing images.

Figure V.3 and Figure V.4 demonstrate that the data-driven atlases match the target segmentations significantly better than the traditional group based atlases with the significantly smallest JS divergence and greatest Dice values while the atlases from other groups perform the worst. Moreover, for the data-driven atlases with two different numbers of training images, the large-scale Training Set 1888 performs significant better than Training Set 720 for both JS divergence and Dice similarities.

To conclude, (1) the group based atlases perform significantly better than the atlases from other groups which demonstrates the group-based framework is able to control the inter-group variability; (2) our proposed data-driven framework produced the more accurate probabilistic atlases than group based atlases by capturing the individual variance; (3) by using the large-scale training data, the performance of data-driven framework is improved significantly.

Evaluation by OASIS Data

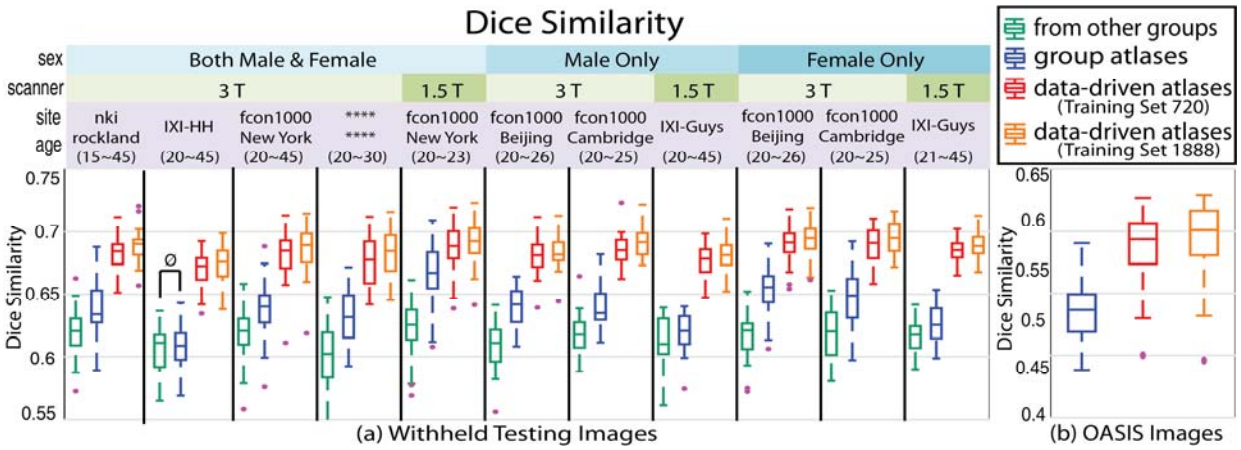


Figure V.4 Dice similarity. The comparisons of Dice value for different atlases are all significant for both withheld and OASIS testing images except the IXI-HH group marked by "Ø".

45 subjects from OASIS dataset with manual segmentations are used for 44 leave one tests. The data-driven probabilistic atlases are obtained from the learned dictionary. The right hand panel of results in Figure V.3 and Figure V.4 show that the results of manual segmentations repeat the finding in previous section

Moreover, we show one testing subject (slice z = 75 in MNI space from 3D image) from the OASIS dataset in Figure V.5. By comparing with the manual segmentations for 6 regions, it shows that the data-driven atlases match the true segmentations more accurately than the group atlases. Moreover, the large-scale Training Set 1888 matches the manual segmentation better than the smaller Training Set 720.

5. Discussion

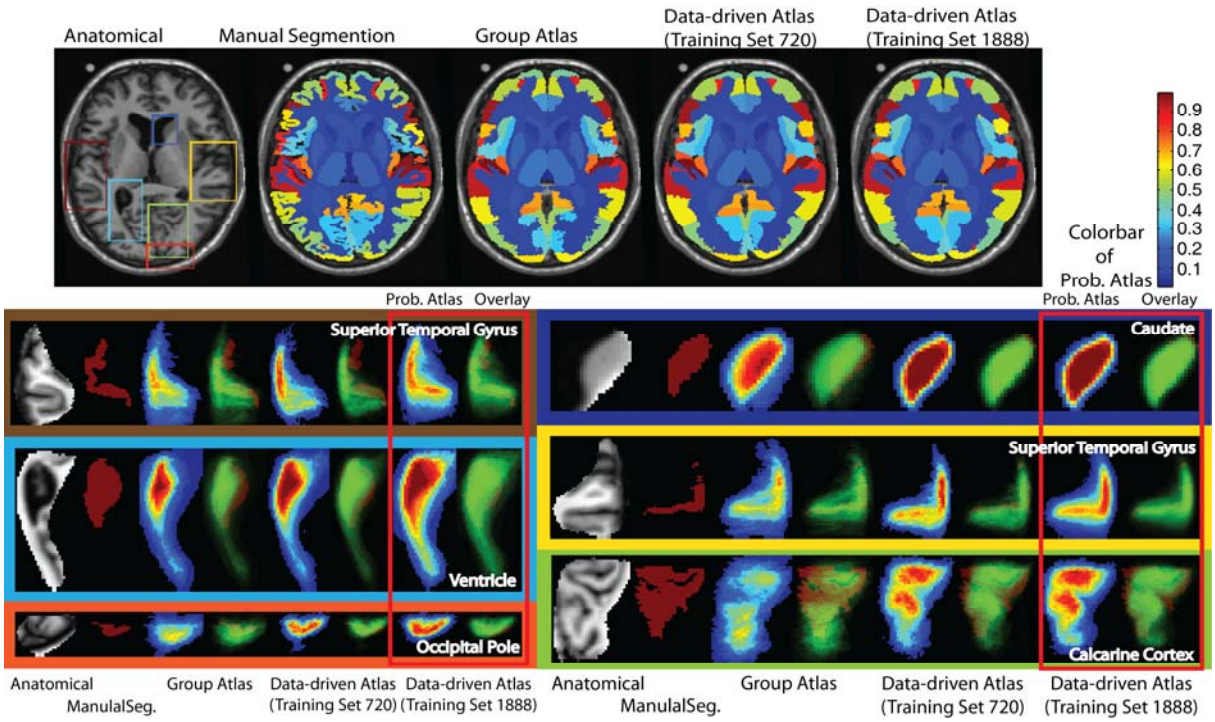


Figure V.5 One testing subject from OASIS dataset. Top row shows the anatomical image, manual segmentation, highest probability segmentations using the group probabilistic atlases, Training Set 720 and Training Set 1888. The lower rows show the details of 6 regions. For each region, from left to right are: anatomical image, manual segmentation, probabilistic atlases generated by different methods and their overlays on manual segmentations.

We present a data-driven framework to learn a dictionary of whole brain probabilistic atlases and apply it on newly seen subjects to achieve accurate individualized whole brain probabilistic atlases. This framework (1) provides a new perspective of using data-driven scheme rather than the traditional group based methods, (2) uses the large-scale heterogeneous data to achieve more personal specific probabilistic atlases than using the single-group and single-site data by capturing the individual variation (3) demonstrates the advantages of using large-scale scheme in generating personal probabilistic atlases compared with the smaller size of training images, and (4) only requires one affine registration and Pearson correlations for applying the learned dictionary on a new subject which achieves low computational cost.

Due to the higher accuracy and low computational cost, the proposed data-driven personal specific probabilistic atlases are able to replace the traditional group based atlases when used as the priors in many medical image processing algorithms and applications.

Chapter VI. Simultaneous Total Intracranial Volume and Posterior Fossa Volume Estimation using Multi-atlas Label Fusion

1. Introduction

Total intracranial volume (TICV), the volume inside the brain cranium, is the total volume of gray matter (GM), white matter (WM), cerebrospinal fluid (CSF) and meninges [112]. In volumetric analyses, many inter-subject differences can be explained by differences in head size [61]. To reduce variability, TICV has been widely used as a covariate in regional and whole brain volumetric analyses [61-67]. Compared with whole brain volume (WBV) [68], TICV is often preferred since it provides an estimation of premorbid brain size [69, 70].

Manual delineation of the cranial vault is the gold standard for measuring TICV from magnetic resonance (MR) images [63]. However, this labor-intensive and time-consuming procedure is impractical on large cohort. As a result, automatic TICV estimation methods are appealing. One family of methods directly applies the automatic skull-stripping techniques to TICV estimation for particular imaging modalities. In MRI, skull is dark while CSF is bright in some modalities (e.g., T2-weighted (T2w) and proton density (PD)). Therefore, the brighter CSF and brain tissues are able to be segmented from the darker skull using skull-stripping, and the total volume of the CSF and brain tissues are used as TICV. For instance, the brain extraction tool (BET) and the brain surface extractor (BSE) achieved accurate TICV estimation using PD images [255]. However, both skull and CSF are dark in other modalities (e.g., T1-weighted (T1w)), in which the skull-stripping techniques typically yield less accurate TICV estimations because of the low contrast between the CSF and skull. To derive accurate TICV estimation on such MR modalities, a number of approaches have been developed and evaluated [113-122]. Among these methods, three of the most prevalent are integrated in FreeSurfer (FS) [106], FMRIB Software Library (FSL) [117], and Statistical Parametric Mapping (SPM12). In FreeSurfer, the estimated TIV (eTIV) tool estimates TICV by investigating the affine transformation between target image and template [116]. The idea is that the TICV

volume is correlated with the determinant of the transform matrix (called “scaling factor”), which aligns a target image with a template. SIENAX, part of FSL, also provides a volumetric scaling factor as a normalization for head size [256]. This scaling factor is the determinant of scaling matrix from affine registration, which rescales the target image's skull to the template’s skull [257]. Therefore, FreeSurfer and FSL do not provide explicit skull/CSF boundaries (SCB) when estimating TICV. SPM provides two different approaches for TICV estimation (e.g., implemented in SPM5 and SPM8). The first approach, called the reverse brain mask (RBM) method, non-rigidly registers a TICV mask from template space to individual space [120, 258]. The second approach accumulates the tissue probabilities of GM, WM, and CSF in standard space using the “New Segment” toolbox [113, 259]. The first approach provides a TICV mask in individual space, however the second method produces more accurate TICV estimations [260]. More recently, the newly released SPM12 provides a new “Tissue Volumes” toolbox, which combines the advantages from two previous approaches in a unified framework [261]. As a result, SPM12 achieves superior TICV estimations compared with previous SPM versions [261]. However, the TICV value and the related SCB are provided in standard space by SPM12 rather than in individual space. Extra efforts from the user side are required if the users want to achieve consistent TICV value and SCB in individual space.

FreeSurfer, FSL and SPM12 are three of the most well validated and widely accepted TICV estimation software packages. However, none of them estimate TICV by counting the voxels inside skull (or SCB), which is a natural way of calculating TICV. The reason is that it is difficult to obtain adequate intensity contrast between skull and CSF in MR T1-weighted (T1w) images (assuming that the thickness of dura is negligible). To obtain the SCB, multispectral MR data (e.g., T2-weighted (T2w), proton density (PD)), with more clear skull evidence, have been combined with T1w images in TICV estimation [63, 115, 120, 122]. However, it is still essential to measure TICV with explicit SCB using a single T1w image since: (1) T2w and PD images are not available in all datasets and T1w images are commonly available structural MR sequences. (2) TICV estimation with SCB not only leads to a natural way of obtaining TICV (count voxels inside skull) but also allows us to calculate sub-region volumes, e.g., posterior fossa volume (PFV),

which is essential in investigating cerebellum development, e.g., [262-264].

TICV estimation using STAPLE label fusion [96] has been proposed to derive SCB using a single T1w image [265]. However, the STAPLE label fusion algorithm has shown limitations [266], which have led to extensions of STAPLE [92, 172, 267-274]. Recently, an improved method called Non-local Spatial STAPLE (NLSS) label fusion, a combination of Spatial STAPLE [172] and Non-local STAPLE (NLS) [92], has shown advantages over STAPLE, Spatial STAPLE and NLS in brain segmentation [149, 151, 272, 275, 276], optic nerve segmentation [277-280] and spinal cord segmentation [281]. Therefore, using NLSS in TICV estimation is promising as it takes both spatial varying performance and non-local intensity correspondence into account. Although the NLSS method has been successfully applied in different applications, its mathematical derivation has not been published yet, which hinders other researchers seeking to implement and use NLSS methods.

In this chapter, we proposed to use NLSS approach to estimate TICV and PFV simultaneously from a single MR T1w image. The main contributions of this work are: (1) TICV and PFV are simultaneously obtained with explicit SCB. (2) We develop TICV and PFV labels for 45 images of the widely used OASIS dataset under BrainCOLOR protocol [170, 228] and make a subset freely available online (https://www.nitrc.org/frs/?group_id=385). (3) This the first journal appearance of NLSS method with detailed mathematical derivation. In the multi-atlas segmentation framework, the pairs of T1w images and TICV labels (atlases) are essential [203]. Normally, atlases are obtained by labor-intensive manual tracing. However, since skull has much higher Hounsfield unit (HU) than other brain tissues [282], we speed up the atlas generation using a semi-manual strategy to obtain TICV and PFV labels using a dataset with 20 paired MR and CT images. Then, the TICV and PFV labels are propagated to the BrainCOLOR atlases [170, 228] by deploying NLSS multi-atlas segmentation. From leave-one-out evaluations and reproducibility analyses, the NLSS TICV estimation method demonstrates its advantages compared with FreeSurfer, FSL, SPM12 and a previously proposed STAPLE TICV estimation approach. The new TICV and PFV labels in OASIS BrainCOLOR atlases provide acceptable performance, which enables simultaneous whole brain segmentation as well as TICV and PFV estimation without conducting additional time-consuming non-

rigid registrations. Moreover, NLSS tool is publically available as open source software through the JIST software package (http://www.nitrc.org/projects/jist/) [238, 239].

2. Theory

The derivation of NLSS closely follows Spatial STAPLE [172] and NLS [92], which use Expectation-Maximization (EM) framework [283, 284]. The majority of the derivations of STAPLE), Spatial STAPLE, and Non-Local STAPLE are left to their original works, but they are described briefly in this work. The notation follows STAPLE [96].

2.1. Problem Definition

A target gray-level image with N voxels is represented as $I \in \mathbb{R}^{N\times 1}$. The corresponding latent true segmentation for the target image is given by $T \in \{0,1,\dots,L-1\}^{N\times 1}$, where $\{0,1,\dots,L-1\}$ represents L possible labels for a given voxel i ($i \in \{1,2,\dots,N\}$). Since T is unknown, the labels for the target image are estimated using R pairs of atlases with intensity values $A \in \mathbb{R}^{N\times R}$ and label decisions $D \in \{0,1,\dots,L-1\}^{N\times R}$. In STAPLE family of approaches, the label fusion problem is regarded as a probabilistic estimation of hidden true segmentation based on the performance of multiple atlases. The performance parameter $\theta_{js's}$ indicates the probability that observed label is s' given that the true label is s for atlas j ($j \in \{1,2,\dots,R\}$). All $\theta_{js's}$, can be written as a matrix $\theta \in [0,1]^{R\times L\times L}$, called performance parameters. The "[0,1]" indicates each $\theta_{js's}$ satisfies $0 \le \theta_{js's} \le 1$.

2.2. STAPLE

The full derivation of the STAPLE algorithm is available in [96]. Briefly, the goal of STAPLE is to select the performance parameters θ, such that they maximize the complete log-likelihood

$$\hat{\theta} = \arg\max_{\theta} \ln f(D|\theta) \tag{6.1}$$

corresponding to the observed atlases D and the unobserved latent true labels T. Since T is not available, the performance parameters are estimated through EM framework. In the E-step, the weight

variables $W^{(k)} \in [0,1]^{L\times N}$ are derived from $\theta^{(k)}$, where $W_{si}^{(k)}$ represents the probability that the true label of voxel i is s at iteration k given $W_{si}^{(k)} \equiv f(T_i = s|D, \theta^{(k)})$. Applying Bayes' rule and the assumed conditional independence between atlases, $W^{(k)}$ for a particular voxel and label is given by

$$W_{si}^{(k)} = \frac{f(T_i = s)\prod_j f\left(D_{ij} = s'|T_i = s, \theta_j^{(k)}\right)}{\sum_n f(T_i = n)\prod_j f\left(D_{ij} = s'|T_i = n, \theta_j^{(k)}\right)} \tag{6.2}$$

where $f(T_i = s)$ is the prior probability that label s is the true label at i and will be discussed later in this chapter. n represents all existing labels while j represents all atlases. Using $\theta_{js's}^{(k)}$ as the simplified expression of $f(D_{ij} = s'|T_i = s, \theta_j^{(k)})$, the $W_{si}^{(k)}$ can be rewritten as

$$W_{si}^{(k)} = \frac{f(T_i = s)\prod_j \theta_{js's}^{(k)}}{\sum_n f(T_i = n)\prod_j \theta_{js'n}^{(k)}} \tag{6.3}$$

The denominator is the partition function to force $\sum_s W_{si}^{(k)} = 1$.

Following the derivation of [96], the M-Step maximizes performance parameters at the iteration $k+1$ as

$$\theta_j^{(k+1)} = \arg\max_{\theta_j} \sum_i E\left[\ln f\left(D_{ij}\ \middle|T_i,\ \theta_j\right)\middle|\ D, \theta_j^{(k)}\right] \tag{6.4}$$

which can be solved as

$$\theta_{js's}^{(k+1)} = \frac{\sum_{i:D_{ij=s'}} W_{si}^{(k)}}{\sum_i W_{si}^{(k)}} \tag{6.5}$$

where $\theta_{js's} \geq 0$ and $\sum_{s'} \theta_{js's} = 1$. This process iteratively solves for the true data likelihood in the E-Step and updates the performance parameters in the M-Step.

2.3. Spatial STAPLE

Spatial STAPLE (SS) is an extension of the STAPLE algorithm where the performance parameters, θ, are calculated at each voxel [172]. The parameters are given by $\theta \in [0,1]^{R\times N\times L\times L}$, which correspond to performance parameters defined voxel-wise instead of globally. As a result, the E-step in Spatial STAPLE

is given by

$$W_{si}^{(k)} = \frac{f(T_i = s)\prod_j \theta_{jis's}^{(k)}}{\sum_n f(T_i = n)\prod_j \theta_{jis'n}^{(k)}} \tag{6.6}$$

which incorporates the spatially varying performance. $\theta_{jis's}^{(k)}$ is the simplified expression of $f(D_{ij} = s'|T_i = s, \theta_{ji}^{(k)})$. The M-step follows the derivation of STAPLE, but since the degrees of freedom are a factor of N higher than STAPLE, two extra extensions are included to account for the increased complexity. First, the performance parameters are binned over small pooling regions instead of a strictly voxel-wise derivation. Following [172], this is implemented by defining spatial pooling regions B, where B_i is the index of the bin which voxel i is contained in. Second, the performance is augmented by a non-parametric prior $\theta_j^{(0)}$ on the performance following [172] and [270]. This augmentation improves the stability of the performance parameters. Thus the M-Step is given by

$$\theta_{jis's}^{(k+1)} = \frac{\lambda_{ijs'}\theta_{js's}^{(0)} + \sum_{i' \in B_i : D_{i'j} = s'} W_{si'}^{(k)}}{\lambda_{ijs'}\sum_s \theta_{js's}^{(0)} + \sum_{i' \in B_i} W_{si'}^{(k)}} \tag{6.7}$$

where $\sum_{s'} \theta_{jis's} = 1$. $\lambda_{ijs'}$ is a weighting parameter depends on the size of pooling region B, which balances the prior and the updated probability. We derive $\lambda_{ijs'}$ using the same definition as [172].

2.4. Non-Local STAPLE

Non-local STAPLE (NLS) incorporates the image intensity from both the atlas images $A \in \mathbb{R}^{N \times R}$ and target image $I \in \mathbb{R}^{N \times 1}$ into the STAPLE framework using a patch-based non-local correspondence manner [92]. Patch-based non-local correspondence was initially introduced to account for registration inaccuracy [93]. NLS incorporates patch-based non-local correspondence into the STAPLE framework as follows. The E-Step is given by

$$W_{si}^{(k)} = \frac{f(T_i = s)\prod_j \sum_{i' \in \mathcal{N}(i)} \theta_{js's}^{(k)} \alpha_{ji'i}}{\sum_n f(T_i = n)\prod_j \sum_{i' \in \mathcal{N}(i)} \theta_{js'n}^{(k)} \alpha_{ji'i}} \tag{6.8}$$

$\mathcal{N}(i)$ is a search neighborhood around voxel i and $\alpha_{ji'i}$ is the non-local weighting between voxel

i in the target image at voxel i' on the jth atlas, within the search parameter $\mathcal{N}(i)$. $\alpha_{ji'i}$ is given by

$$\alpha_{ji'i} = \frac{1}{Z_\alpha} \exp\left(-\frac{\left\| \wp(A_{i'j}) - \wp(I_i) \right\|_2^2}{2\sigma_i^2} \right) \exp\left(-\frac{\mathcal{E}_{i'i}^2}{2\sigma_d^2} \right) \tag{6.9}$$

where $\wp(\cdot)$ is the set of intensities within its patch neighborhood. In this definition, $\left\| \wp(A_{i'j}) - \wp(I_i) \right\|_2^2$ is the L2-norm between the atlas patch centered at i' and the target patch centered at i, $\mathcal{E}_{i'i}^2$ is the Euclidean distance in physical space between i and i', σ_i and σ_d are the standard deviations for the intensity and distance weights respectively, and Z_a normalizes α to be a valid probability distribution for each atlas and target voxel. The M-Step for Non-Local STAPLE is

$$\theta_{js's}^{(k+1)} = \frac{\sum_i \left(\sum_{i' \in \mathcal{N}(i): D_{i'j} = s'} \alpha_{ji'i} \right) W_{si}^{(k)}}{\sum_i W_{si}^{(k)}} \tag{6.10}$$

which follows the original M-Step of STAPLE while incorporating non-local correspondence.

2.5. Non-local Spatial STAPLE

The Non-local Spatial STAPLE (NLSS) algorithm follows directly from the derivations of Spatial STAPLE and Non-local STAPLE. The NLSS algorithm defines the following performance level function

$$f(D, A | T, I, \theta) \tag{6.11}$$

In the NLSS algorithm, θ is spatially varying as in Spatial STAPLE and non-local correspondence is used to account for registration errors.

2.5.1. NLSS E-Step

The E-Step of NLSS follows similar to the E-Step of STAPLE. First, Bayes' rule is applied as

$$W_{si}^{(k)} = \frac{f(T_i = s) f(D, A | T_i = s, I, \theta)}{\sum_n f(T_i = n) f(D, A | T_i = n, I, \theta)} \tag{6.12}$$

Following the expansions, this becomes

$$W_{si}^{(k)} = \frac{f(T_i = s) \prod_j \sum_{i' \in \mathcal{N}(i)} \theta_{jis's}^{(k)} \alpha_{ji'i}}{\sum_n f(T_i = n) \prod_j \sum_{i' \in \mathcal{N}(i)} \theta_{jis'n}^{(k)} \alpha_{ji'i}} \tag{6.13}$$

This derivation incorporates both the spatially varying performance parameters derived in Spatial STAPLE and the non-local correspondence derived in Non-local STAPLE.

2.5.2. NLSS M-Step

In M-step of NLSS, the previously calculated $W_{si}^{(k)}$ is used to update $\theta_{ji}^{(k+1)}$ by maximizing the expectation of the log likelihood function as

$$\begin{aligned}\theta_{ji}^{(k+1)} &= \underset{\theta_{ji}}{\operatorname{argmax}} \sum_{i' \in B_i} E\left[\ln f\left(D_j, A_j | T_{i'}, I_{i'}, \theta_{ji}\right) | D, A, I, \theta^{(k)}\right] \\ &= \underset{\theta_{ji}}{\operatorname{argmax}} \sum_{i' \in B_i} \sum_{s} W_{si'}^{(k)} \ln f\left(D_j, A_j | T_{i'} = s, I_{i'}, \theta_{ji}\right) \\ &= \underset{\theta_{ji}}{\operatorname{argmax}} \sum_{i' \in B_i} \sum_{s} W_{si'}^{(k)} \ln \left(\sum_{i'' \in \mathcal{N}(i'): D_{i''j} = s'} \theta_{jis's} \, \alpha_{ji''i'} \right)\end{aligned} \tag{6.14}$$

Using a Lagrange λ Multiplier [285] with constrain $\sum_{s'} \theta_{jis's} = 1$ and setting the derivative equal to zero this becomes

$$0 = \frac{\partial}{\partial \theta_{jin'n}} \left[\sum_{i' \in B_i} \sum_{s} W_{si'}^{(k)} \ln \left(\sum_{i'' \in \mathcal{N}(i'): D_{i''j} = s'} \theta_{jis's} \, \alpha_{ji''i'} \right) + \lambda \sum_{s'} \theta_{jis's}^{(k+1)} \right] \tag{6.15}$$

This equation can be solved as

$$\theta_{jis's}^{(k+1)} = \frac{\sum_{i' \in B_i} \left(\sum_{i'' \in \mathcal{N}(i'): D_{i''j} = s'} \alpha_{ji''i'} \right) W_{si'}^{(k)}}{\sum_{i' \in B_i} W_{si'}^{(k)}} \tag{6.16}$$

Like Spatial STAPLE, the same whole-image implicit prior $\theta_{js's}^{(0)}$ is introduced for computational and stability concerns [172]. The prior can be derived from a number of approaches (e.g. STAPLE [96], Majority Vote, Locally Weighted Vote [95], etc.). In our NLSS implementation, the Majority Vote while ignoring "consensus voxels" (i.e., voxels where all raters agree) is employed as default method [86]. This method ignores the consensus voxels when constructing the performance level parameters. Then, the final stable version of Eq. 5.16 is reformulated to

$$\theta_{jis's}^{(k+1)} = \frac{\lambda_{ijs'}\theta_{js's}^{(0)} + \sum_{i'\in B_i}\left(\sum_{i''\in\mathcal{N}(i'):D_{i''j}=s'}\alpha_{ji''i'}\right)W_{si'}^{(k)}}{\lambda_{ijs'}\sum_s\theta_{js's}^{(0)} + \sum_{i'\in B_i}W_{si'}^{(k)}} \tag{6.17}$$

where $\lambda_{ijs'}$ is a weighting parameter depends on the size of pooling region B_i, which balances the prior and the updated probability. We derive $\lambda_{ijs'}$ and B_i using the same way as [172].

Notice that the Eq.6.16 is the theoretical expression of M-step in the EM framework while the Eq.6.17 is an approximate maximizer for computational and stability concerns. The implementations of both cases have been provided in the publically available open-source code, which enable the users to switch from each other by controlling $\lambda_{ijs'}$. In practice, the Eq.6.17 typically provides better performance than Eq.6.16. Therefore, the implementation of Eq.6.17 is the default setting in NLSS open-source code.

2.5.3. Initialization, parameters and detection of convergence

The voxelwise prior $f(T_i = s)$ in NLSS is initialized using the weak log-odds majority vote [95]. The performance parameters are typically initialized assuming each atlas has high performance as

$$\theta_{jis's} = \begin{cases} 0.95 & \text{if } s = s' \\ \dfrac{0.05}{L-1} & \text{else} \end{cases} \tag{6.18}$$

The search neighborhood $\mathcal{N}(\cdot)$ and the patch neighborhood $\wp(\cdot)$ are the two key parameters in non-local search model. In all presented experiments, the search neighborhood $\mathcal{N}(\cdot)$ was set to $7 \times 7 \times 7$ voxels search window centered at a target voxel while the patch neighborhood $\wp(\cdot)$ was empirically set to $3 \times 3 \times 3$ voxels. The two standard deviation parameters σ_i and σ_d in Eq.6.7 were empirically set to 0.1 and 1.5 respectively. The algorithm is iterated until the trace of the difference of confusions matrices between iterations is small, typically less than 10^{-4}.

3. Method

This section first introduces a semi-manual method to establish atlases with TICV and PFV labels Second, the multi-atlas segmentation framework using NLSS label fusion is demonstrated. Third, the procedure of generating TICV and PFV labels for the BrainCOLOR (BC) atlases is introduced. Last, the

statistical analysis methods used in this work are introduced.

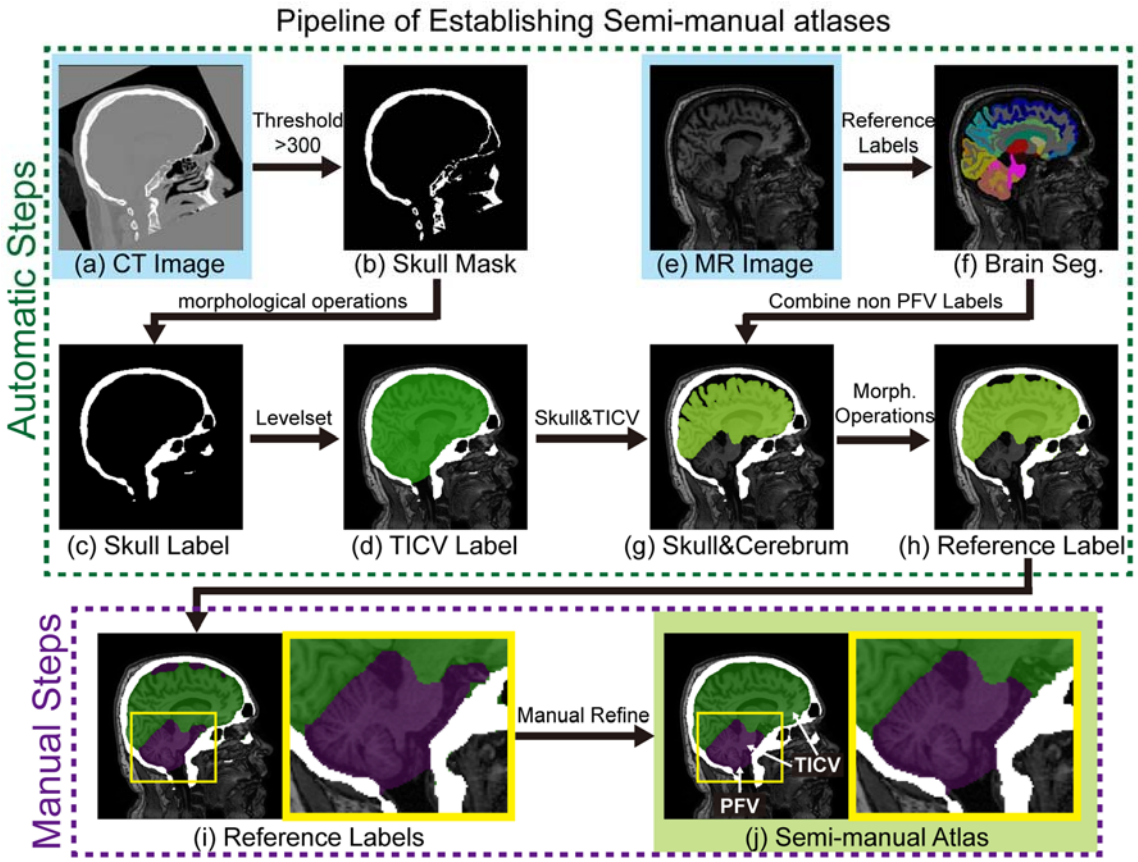


Figure VI.1 Semi-manual pipeline of establishing atlases. First, the TICV label is obtained by applying a threshold, morphological operations and the level set method on CT images. Then, the TICV label is propagated to MR image space and the reference PFV label are provided by merging TICV label and the automatic whole brain segmentation. Finally, the semi-manual atlases are obtained by conducting manual refinement on the reference labels.

3.1. Semi-manual Segmentations and Semi-manual Atlases

We start by automatic skull labeling using CT images, then obtain TICV labels (voxels inside brain skull), and finally propagate labels to MR images using rigid registration. The procedure of automatically generating TICV atlas (Figure VI.1) is inspired by the recent work [121]. Briefly, each CT image is aligned to MR image using rigid registration [89] (Figure VI.1a). Then, the skull masks are obtained from CT images, whose voxel values are greater than 300 HU [286] (Figure VI.1b). Then, a 3D closing morphological operation (a dilation followed by an erosion) followed by neck removal [287] is applied on the skull mask to obtain the binary skull label. The closing morphological operation fills the holes in the

skull, and the inner side of the filled skull provides the SCB (Figure VI.1c).

The TICV segmentation is the region inside the SCB. However, the SCB is not a closed surface (e.g., the foramen magnum in the occipital bone). Such opening regions make it difficult to derive the TICV segmentation by only using morphological operations. To deal with the opening regions automatically, Topology-preserving Geometric Deformable Model (TGDM) [288] with gradient vector flow (GVF) field [231] is employed. The Standard Geometric Deformable Model (SGDM) has been widely used in image segmentation due to its parameterization independence and ease of implementation. However, topological flexibility of SGDM is not always desired in medical image segmentation especially when the number of components has been known and must be preserved. Based on our anatomical prior knowledge, the TICV segmentation should only contain one component (one contour surface). Therefore, the TGDM framework is employed to keep such topology. In its implementation, the level set contour of TGDM is moved by the gradient vector flow (GVF) field [231]. The advantage of GVF field is that it forces the contour towards skull and has close to zero force at the opening regions. We also apply a curvature force [288] to keep the surface smooth at the opening regions. Using TGDM, the non-skull voxels inside zero level set are labeled as TICV segmentation. Such segmentation has a smooth boundary at the opening regions. By copying the labels from the registered CT images voxel-by-voxel, we obtain skull and TICV labels on MR images (Figure VI.1d).

Then, we label posterior fossa within the TICV labels. Instead of doing complete manual delineation, a rough automatic segmentation is provided as the reference labels to accelerate the procedure. Briefly, we start with a NLSS multi-atlas segmentation to obtain the whole brain segmentations (133 labels) for each MR image under BrainCOLOR protocol [170, 228] (Figure VI.1f). Then, we group the cerebrum regions (above tentorium cerebelli) together, which excludes the CSF and tissues in posterior fossa tissues (cerebellum and brainstem) (Figure VI.1g). A closing morphological operation is conducted to obtain the reference labels (Figure VI.1h and j), which indicates the rough location of posterior fossa. Finally, a manual refinement step is conducted by an experienced graduate student to correct the inaccurate voxels in the reference labels and obtain the final PFV labels (Figure VI.1j). Using this semi-manual pipeline, we

obtain the 20 atlases consist of both T1w images and labels (posterior fossa, cerebrum and background). The TICV is the sum of posterior fossa and cerebrum.

3.2. NLSS Multi-atlas framework

We use a canonical multi-atlas segmentation framework which contains registration, atlas selection, label propagation and label fusion [203]. Briefly, the target image is first corrected by a N4 bias field correction [226] and then affinely registered [89] to the MNI305 atlas [171]. Practically, using 10–20 atlases are sufficient to achieve accurate whole brain segmentation [97]. Empirically, the 15 closest atlases with smallest Euclidian distance to the target image on PCA manifold are chosen if total number of available atlases is greater than 15 [149]. Then, the 15 selected atlases are then non-rigidly registered to the target image [88]. For non-rigid registration, we use symmetric image normalization (SyN), with a cross correlation similarity metric convergence threshold of 10^{-9} and convergence window size of 15, provided by the Advanced Normalization Tools (ANTs) software [88]. Finally, the proposed NLSS label fusion is used to combine the labels from each atlas to the target image. After multi-atlas labeling, each voxel is assigned to one of the labels.

3.3. TICV and PFV labels for OASIS BrainCOLOR atlases

Using the semi-manual strategy, Researchers are able to reconstruct semi-manual atlases using their own data. However, the paired MR and CT images are not typically available, especially when people want to derive both TICV and PFV labels as well as whole brain segmentation simultaneously (e.g., 133 labels in BrainCOLOR protocol). Therefore, we propagate the TICV and PFV labels from semi-manual atlases to the BrianCOLOR atlases [170, 228], which consist of 45 OASIS images [132]. We have made a subset of new BrainCOLOR atlases freely available online to facilitate the community.

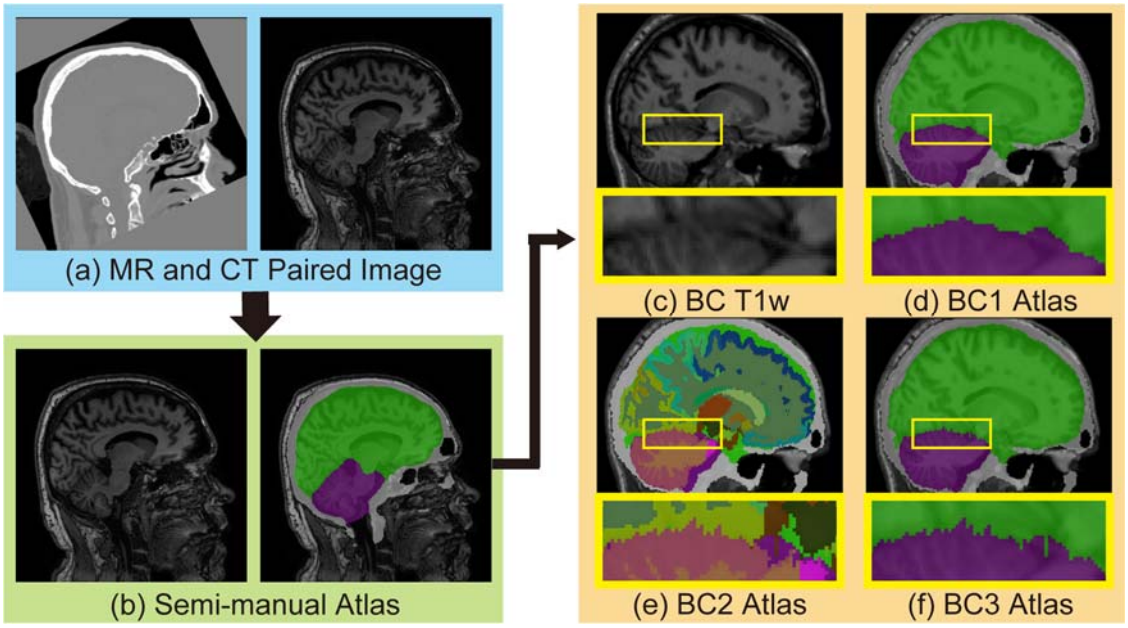


Figure VI.2 BC1, BC2 and BC3 atlases are obtained by adding TICV and PFV labels. (a) 20 paired MR-CT images are used to generate (b) semi-manual atlases. Then the NLSS multi-atlas segmentation is conducted on (c) T1w images 45 OASIS images in BrainCOLOR (BC) atlases to achieve TICV and PFV labels. (d) The first automatic segmentation results are referred as BC1 atlases. (e) Then the original 133 labels from BC are merged with BC1 atlases by keeping the BC labels if conflictions happen. The merged BC2 atlases contain 136 labels including the TICV, PFV and BC labels. (f) The 136 labels are merged back to 4 labels to resolve conflicts and form the BC3 atlases. A subset of BC2 atlases have been made freely available online to facilitate other researchers. We compare the performance of BC1, BC2 and BC3 atlases as well as semi-manual atlases.

Briefly, the semi-manual atlases (Figure VI.2b) are employed to segment 45 OASIS T1w images using the NLSS multi-atlas segmentation (Figure VI.2c). Then, the TICV and PFV labels are derived for the OASIS dataset, which are referred as BrainCOLOR1 (BC1) atlases. Then, the BrainCOLOR2 (BC2) atlases are derived by combining TICV and PFV labels with 133 original labels in BrainCOLOR atlases. Note that if the original manual labels conflict with the TICV or PFV definition in BC1 atlases, we keep the original labels in BC2. Finally, The BrainCOLOR3 (BC3) atlases are obtained by merging the TICV and PFV labels in BC2 atlases.

3.4. Statistical Analysis

In this chapter, we conduct several types of volumetric analyses between FreeSurfer, FSL, SPM12 and multi-atlas approaches. To evaluate the volumetric similarity between the automatic methods and semi-

manual segmentations, the absolute volume similarity (ASIM) (a ratio from 0 to 1, higher is better) is employed as:

$$\text{ASIM} = 1 - \frac{|V_1 - V_2|}{0.5(V_1 + V_2)} \tag{6.19}$$

However, the ASIM only compares the similarity of volume sizes without reflecting the spatial information especially the accuracy of SCB. For instance, the segmentations that have similar amounts of volume may have large differences in spatial appearance and location. Therefore, the widely used Dice coefficient (Dice) is employed as:

$$\text{Dice} = \frac{2|A \cap B|}{|A| + |B|} \tag{6.20}$$

where A and B represent any two binary volumes that need to be compared and $|\cdot|$ represents the volume of regions. Dice values evaluate the overlap between regions A and B which takes both volumetric and spatial information into account. Moreover, the mean surface distance (MSD) between A and B is also employed to measure the average surface distance between binary volumes.

The reproducibility is another important aspect of evaluating TICV estimation. In this chapter, we assess the reproducibility of TICV estimation using a test-retest strategy, which compares the TICV and PFV measurements between two consequential scans from the same subject. To capture this difference, the absolute volume difference (ADIFF) (a ratio from 0 to 1, lower is better) is used as:

$$\text{ADIFF} = \frac{|V_1 - V_2|}{0.5(V_1 + V_2)} \tag{6.21}$$

After obtaining the previous metrics, the Wilcoxon signed rank test [175] is used for statistical analyses. All claims of statistically significance in this chapter are made using the Wilcoxon signed rank test for $p < 0.05$.

4. Data and results

4.1. Accuracy Test

Twenty subjects, with both MR and CT images from the deep-brain stimulation (DBS) project,

were employed to evaluate the accuracy of TICV and PFV estimation. The MR images were 3D T1w volumes with 256×256×190 voxels, which have 1×1×1 mm resolution. The CT images were acquired with pixel size = 0.49 mm, slice thickness = 0.625 mm and FOV = 250×250×190 mm. From these paired MR-CT images, 20 semi-manual atlases (MR T1w images and labels) were generated using the semi-manual method. Note that the CT images were only used in generating semi-manual atlases, but were not used in the evaluations.

First, FreeSurfer (FS), FSL, and SPM12 were deployed on the 20 T1w MR images to estimate the TICV results. Then, the NLSS multi-atlas framework was deployed on the same dataset using leave-one-out strategy. In each leave-one-out test, other 19 atlases were used as candidate atlases, which ensured the independence to the testing image. The linear relationship between the estimated TICV results and true TICV volumes (semi-manual atlases) were evaluated by linear regressions (Figure VI.3). The linear relationship between the estimated TICV results and with the true TICV volumes (semi-manual atlases) were evaluated by linear regressions (Figure VI.3). The R^2 coefficient of determination was provided to indicate how strong the linearity was between measurements, where the higher R^2 indicated the stronger linearity. From the results, the NLSS TICV estimation achieved the largest R^2 values (R^2=0.970) to the semi-manual segmentations while FSL had the lowest R^2. NLSS TICV estimation also had R^2=0.942 to FreeSurfer and R^2=0.956 to SPM12. The lower right box plot indicated the ASIM scores for different methods compared with semi-manual segmentation. NLSS TICV had significant higher ASIM scores than FreeSurfer and SPM12. The ASIM score for FSL was not shown since it only provided scaling factors rather than volumetric values.

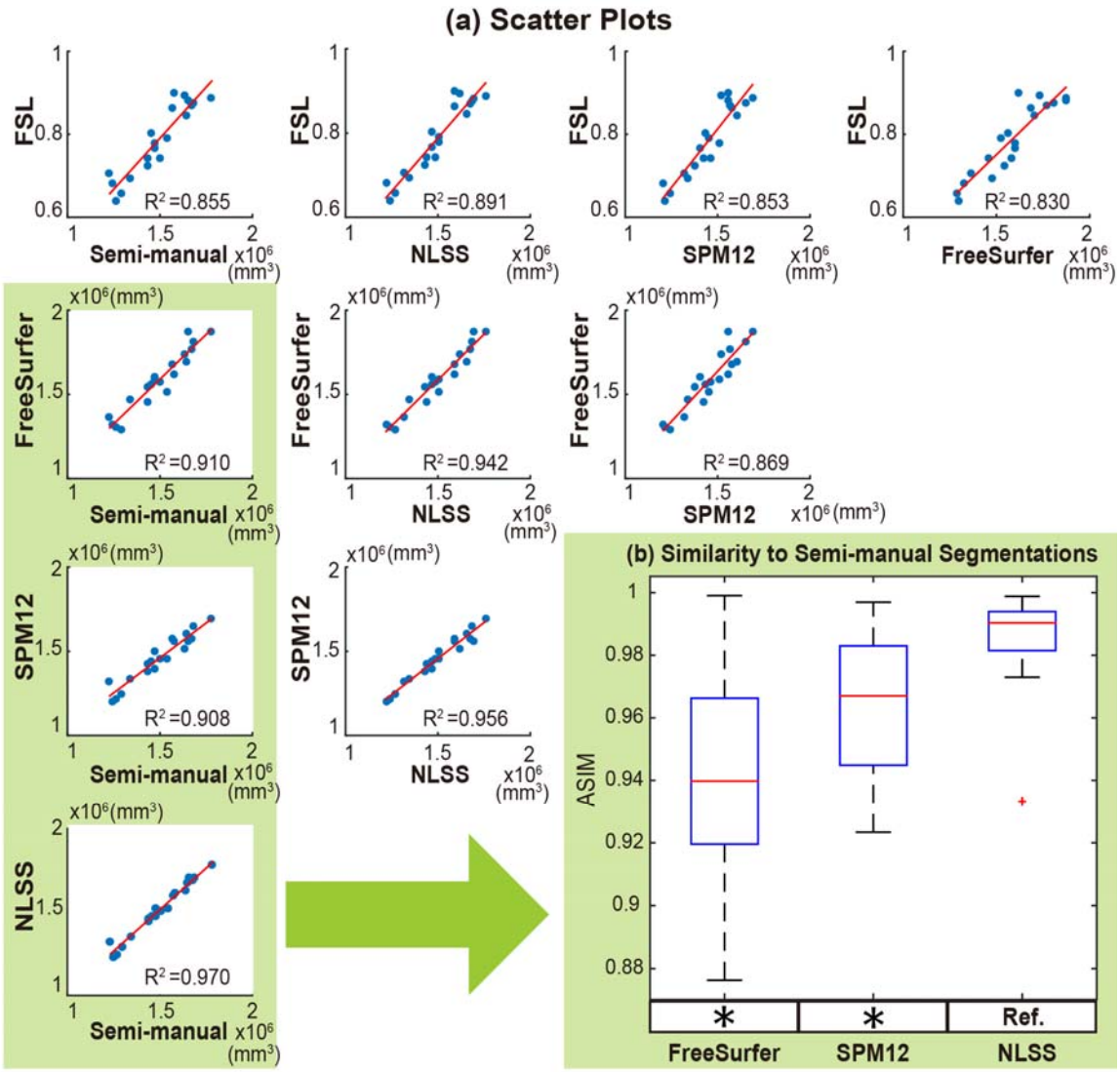


Figure VI.3 (a) Scatter plots comparing FreeSurfer, FSL, SPM12 and NLSS on TICV estimation. In the first column, different automatic methods are compared with semi-manual segmentations by plotting the TICV volumes with a red line of best fit and NLSS method using semi-manual atlases achieves latest $R^2 = 0.970$. The remaining columns show the scatter plots between automatic methods. NLSS method still achieves large R^2 values compared with FreeSurfer, FSL and SPM12. (b) Box plot of ASIM values between FreeSurfer, SPM12 and NLSS with Semi-manual segmentations. The proposed NLSS ("Ref.") method achieves significantly higher ("∗") ASIM scores than FreeSurfer and SPM12. Since FSL only provides scaling factors rather than TICV volumes, it does not have units in (a) and not shown in (b).

Second, NLSS TICV estimation was compared with the previously proposed STAPLE TICV estimation [265]. For more complete analyses, we also compared the NLSS estimation with other label fusion approaches such as majority vote (MV), Spatial STAPLE, NLS and joint label fusion (JLF) (Table

VI.1 and Table VI.2) using the semi-manual atlases. The JLF (Wang et al., PAMI 2013) approach is the state-of-the-art label fusion method using non-local intensity similarity. In each leave-one-out analysis, the BC1, BC2 and BC3 atlases (on 45 OASIS images) were also generated from the 19 semi-manual atlases. Then these intermediate atlases were also deployed on the target image and their accuracies were compared with semi-manual atlases using the same NLSS multi-atlas framework.

Table VI.1 shown four different metrics of evaluating the accuracy of different TICV measurement approaches: (1) Intraclass correlation (ICC) and Pearson Correlation were used to measure the correlation between different methods and semi-manual segmentations. The two-way random single measures was used as the ICC model [289]. (2) The ASIM values were used to show the accuracy of TICV volumetric estimation. (3) Dice similarity coefficients were employed to take the spatial information into account upon the ASIM metric. (4) MSD values were also derived to measure the average surface distance between binary segmentations. From Table VI.1, the family of multi-atlas segmentations (MV, STAPLE, SS, NLS, JLF and NLSS) obtained higher correlation coefficients than the prevalent FreeSurfer, FSL and SPM12 approaches. The multi-atlas approaches achieved higher mean and smaller standard deviation (std) on ASIM metric. Within the multi-atlas family, when using the same semi-manual atlases, the NLSS TICV estimation achieved higher scores on correlation coefficients, mean ASIM and mean Dice than previously proposed STABLE TICV estimation. Meanwhile, it had the smaller mean MSD and the lower standard deviation than the STAPLE method. The NLSS estimation was significantly superior to MV, Spatial STAPLE, NLS on both TICV (Table VI.1) and PFV (Table VI.2). The NLSS and JLF had advantages on PFV and TICV respectively. However, the differences between NLSS and JLF were not statistically significant. When comparing the performance between different atlases, the BC1, BC2 and BC3 atlases performed worse than the semi-manual atlases on correlation coefficients, mean ASIM, mean Dice and mean MSD. However, the correlation coefficients and the mean ASIM values of using BC1, BC2 and BC3 atlases were still higher than FreeSurfer, FSL, and SPM12.

Figure VI.4, Figure VI.5 and Figure VI.6to 6 show the box plots and the statistical results using Wilcoxon signed rank test. In each figure, the statistical analyses were conducted between the NLSS method

using semi-manual atlases (marked as reference "Ref.") with other approaches or different atlases. If the difference was statistically significant, we marked the method with "*" symbol. Otherwise, we marked the method with not significant "N.S.". Figure VI.4 shows the ASIM values, which only considered volumetric results for both TICV and PFV segmentations. For TICV estimation, the ASIM of NLSS (semi-manual atlases) was significantly higher than FreeSurfer, SPM12, STAPLE, Spatial STAPLE and NLS. For PFV estimation, the ASIM of NLSS (semi-manual atlases) was significantly higher than STAPLE, Spatial STAPLE, and NLS. The different performance between NLSS and JLF are not statistically significant. Using the same NLSS method with different atlases, the semi-manual atlases performed significantly better than BC1, BC2 and BC3 atlases in both TICV and PFV volumetric estimation.

It is also important to note how the improved accuracy is able to be translated into clinical research benefits. We evaluated the statistical power of detecting a group difference between two simulated clinical cohorts using two-sample t-test at significant level 0.05.

Figure VI.5 employed the Dice similarity coefficients as the metric, which took both volumetric and spatial information into account. Since the TICV and PFV segmentations were not provided by the default processing in FreeSurfer, FSL, and SPM12, we conducted statistical analyses within the multi-atlas family. For both TICV and PFV segmentations, the NLSS using semi-manual atlases achieved the significant higher Dice values than MV, STAPLE, Spatial STAPLE, and NLS. The semi-manual atlases also achieved significant higher Dice values than the BC1, BC2 or BC3 atlases. Figure VI.6 reflected the statistical analyses on MSD. Again, NLSS using semi-manual atlases had the smaller MSD compared with MV, STAPLE, Spatial STAPLE, and NLS. The performance between NLSS and JLF in Figure VI.5 and Figure VI.6 are not statistically significant. To visually check the findings in Figure VI.5 and Figure VI.6, Figure VI.7 shows the qualitative performance of different methods on the same subject. The surfaces of the semi-manual segmentations, which used as reference results, were remarked as red contours. The area of positive error (estimate larger than reference) was the area with green and purple color outside the contours while the negative error (estimate smaller than reference) was colored as white.

Table VI.1 Accuracy test results of TICV

Atlases		Does not use atlases					Semi-manual				BC1	BC2	BC3
Methods		FS	FSL	SPM12	MV	STAPLE	SS	NLS	JLF	NLSS	NLSS	NLSS	NLSS
Corr.	Pearson	0.954	0.923	0.953	0.959	0.957	0.960	0.969	**0.985**	**0.985**	0.965	0.963	0.964
	ICC	0.836	N/A	0.916	0.961	0.936	0.957	0.964	**0.985**	**0.985**	0.942	0.964	0.907
ASIM	μ	0.941	N/A	0.964	0.976	0.971	0.977	0.979	**0.986**	**0.986**	0.972	0.978	0.961
	σ	0.032	N/A	0.023	0.022	0.0285	0.024	0.024	**0.014**	0.015	0.026	0.020	0.028
Dice	μ	N/A	N/A	N/A	0.977	0.975	0.977	0.979	**0.983**	**0.983**	0.975	0.975	0.970
	σ	N/A	N/A	N/A	0.008	0.01	0.008	0.008	**0.005**	0.006	0.008	0.006	0.009
MSD	μ	N/A	N/A	N/A	0.968	1.058	0.984	0.888	**0.725**	0.743	1.106	1.112	1.245
(mm)	σ	N/A	N/A	N/A	0.268	0.374	0.301	0.294	**0.184**	0.197	0.306	0.244	0.326

"Corr." means correlation analyses. The bold values indicate the best performance. "N/A" means the values are not available since (1) FSL only provides scaling factors rather than TICV volumes, (2) FreeSurfer (FS) and SPM12 (SPM) do not generate hard TICV segmentation in individual space during the standard default processing.

Note: "SS" is Spatial STAPLE. "μ" is the mean and "σ" is the standard deviation.

Table VI.2 Accuracy test results of PFV

Atlases		Semi-manual						BC1	BC2	BC3
Methods		MV	STAPLE	SS	NLS	JLF	NLSS	NLSS	NLSS	NLSS
Corr.	Pearson	0.947	0.934	0.949	0.963	**0.979**	0.976	0.958	0.958	0.958
	ICC	0.944	0.818	0.945	0.953	0.971	**0.975**	0.919	0.951	0.888
ASI	μ	0.975	0.940	0.973	0.974	0.982	**0.984**	0.963	0.973	0.953
M	σ	0.023	0.029	0.021	0.018	0.017	**0.016**	0.02	0.018	0.02
Dice	μ	0.960	0.951	0.959	0.964	0.968	**0.968**	0.955	0.954	0.954
	σ	0.008	0.011	0.007	0.007	**0.006**	**0.006**	**0.006**	**0.006**	**0.006**
MSD	μ	0.847	1.011	0.858	0.767	0.689	**0.675**	0.946	0.933	0.969
(mm)	σ	0.15	0.214	0.14	0.126	0.120	**0.107**	0.121	0.118	0.132

Please see Table VI.1 for the descriptions of abbreviations.

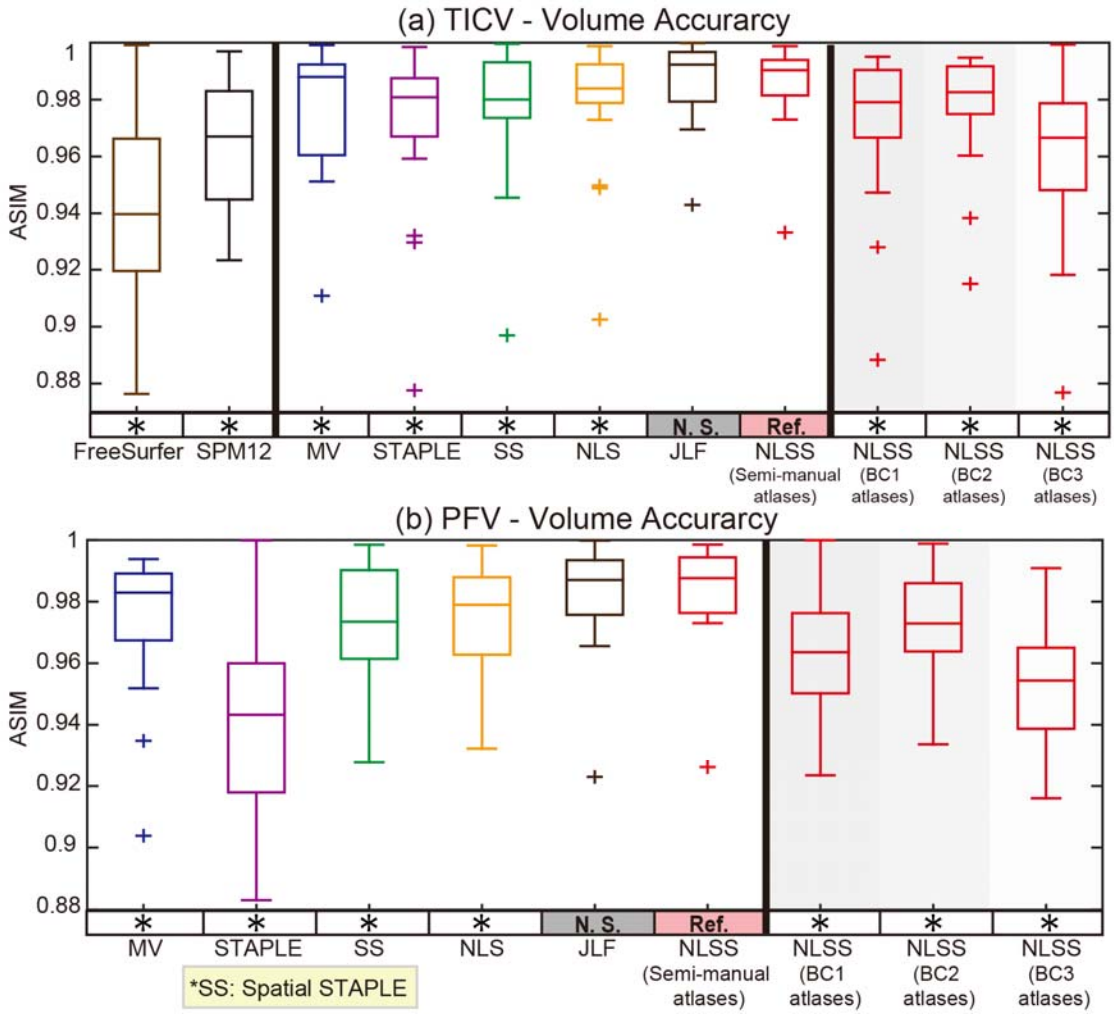


Figure VI.4 Box plots and statistical results on volume accuracy. The statistical analyses were conducted between the proposed NLSS TICV estimation using semi-manual atlases (marked as reference "Ref.") with other approaches or different atlases. If the difference was statistically significant, we marked the other method with "*" symbol. Otherwise, we marked it as "N.S."

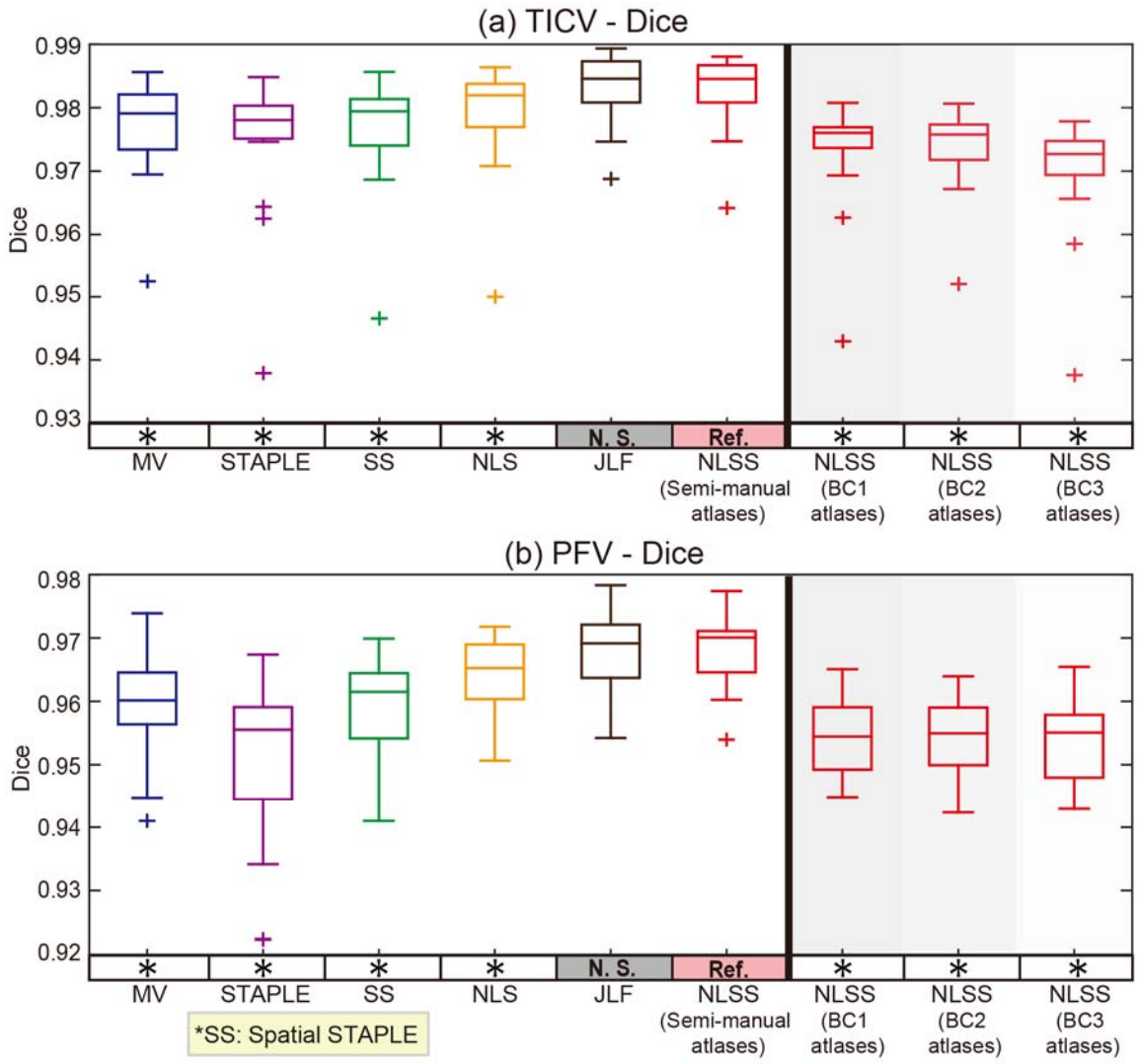


Figure VI.5 Box plots and statistical results on Dice coefficients. The statistical analyses were conducted between the proposed NLSS TICV estimation using semi-manual atlases (marked as reference "Ref.") with other approaches or different atlases. If the difference was statistically significant, we marked the other method with "*" symbol. Otherwise, we marked it as "N.S."

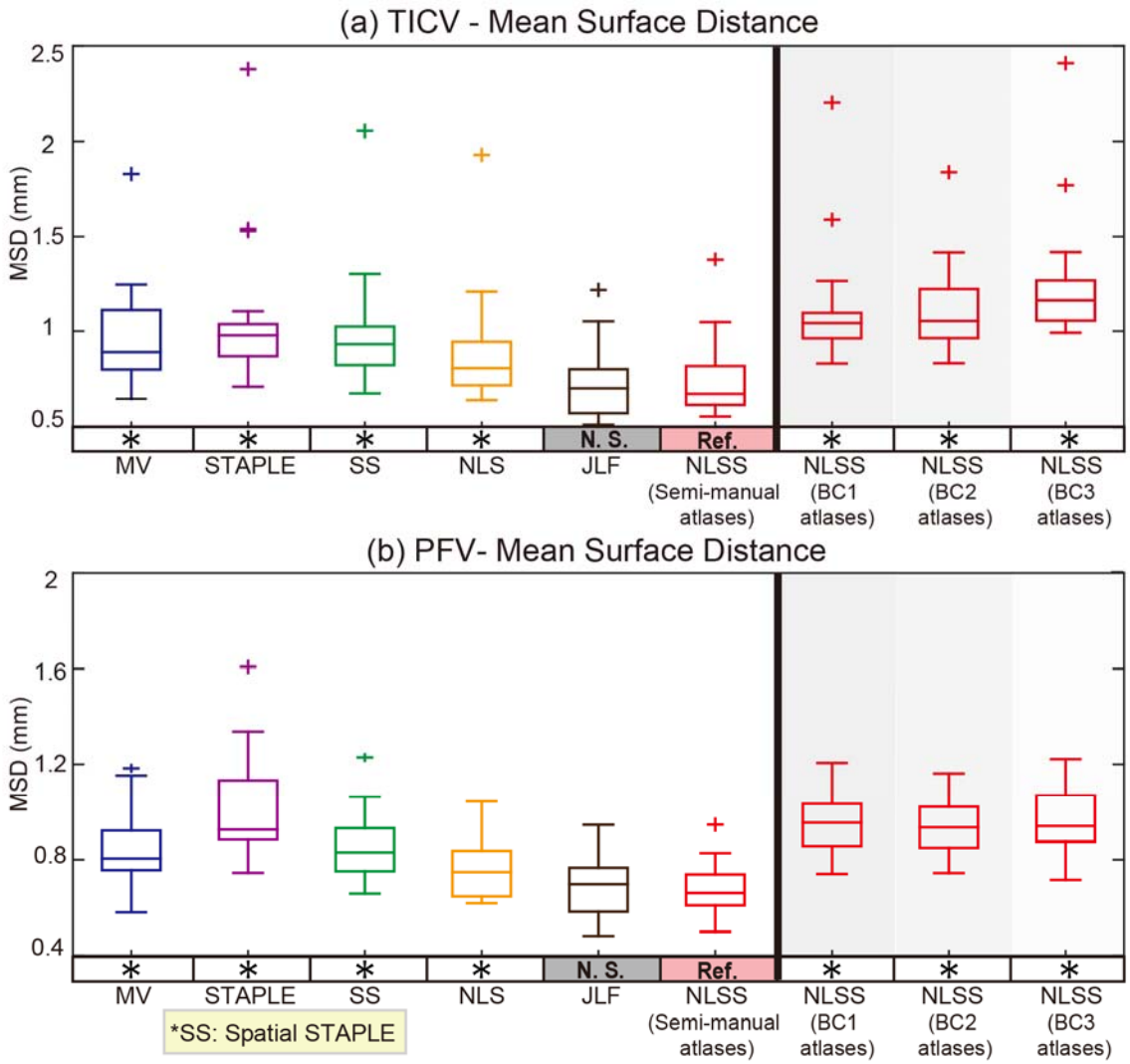


Figure VI.6 Qualitative results comparing multi-atlas segmentation methods with semi-manual segmentation. The red contours represent the spatial location of the semi-manual segmentation. The white color indicates the negative error, in which the estimated segmentation is smaller than the semi-manual reference. The green and purple color outside the red contours indicates the positive error, in which the estimated segmentation is larger than reference.

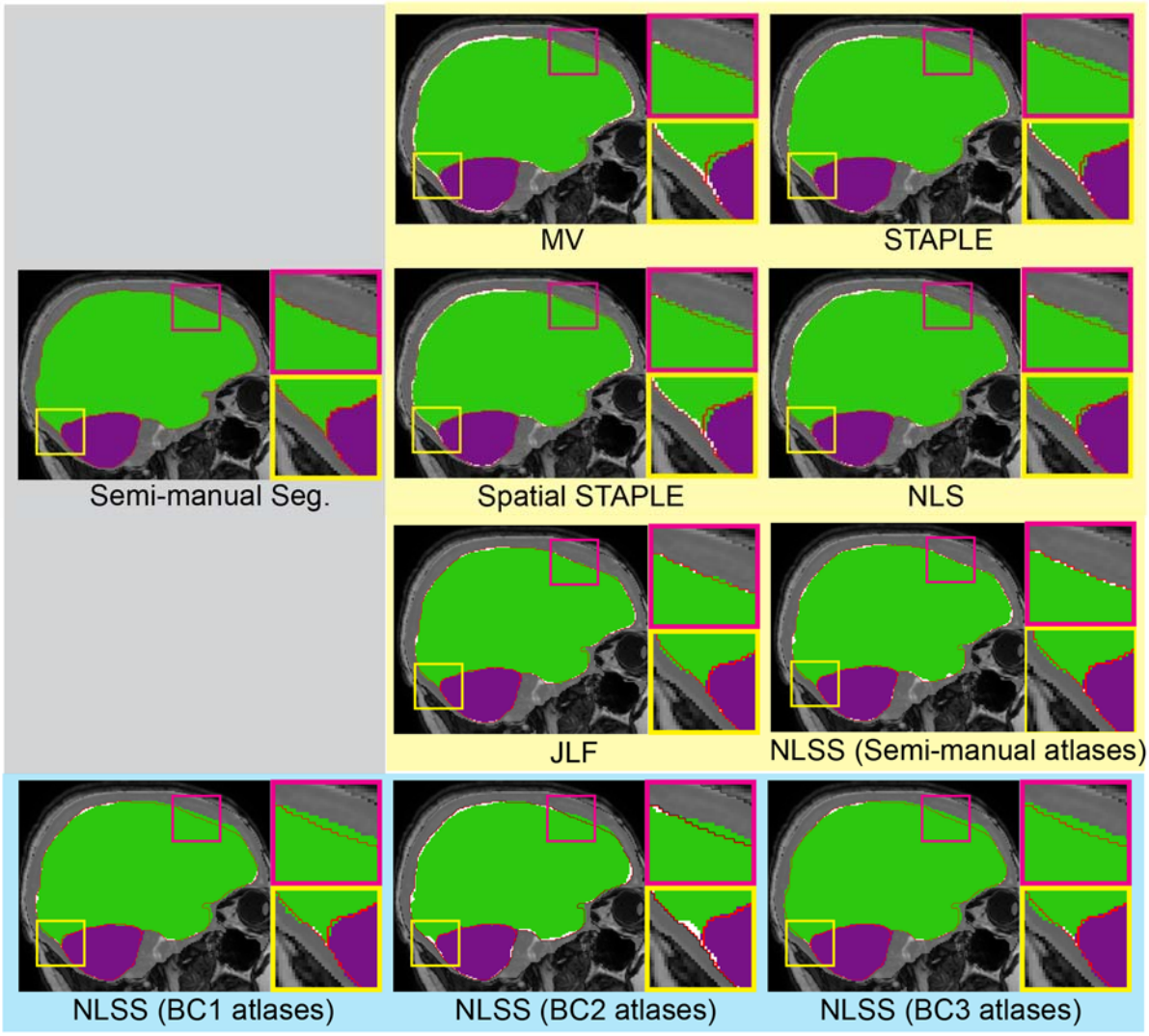


Figure VI.7 Qualitative results comparing multi-atlas segmentation methods with semi-manual segmentation. The red contours represent the spatial location of the semi-manual segmentation. The white color indicates the negative error, in which the estimated segmentation is smaller than the semi-manual reference. The green and purple color outside the red contours indicates the positive error, in which the estimated segmentation is larger than reference.

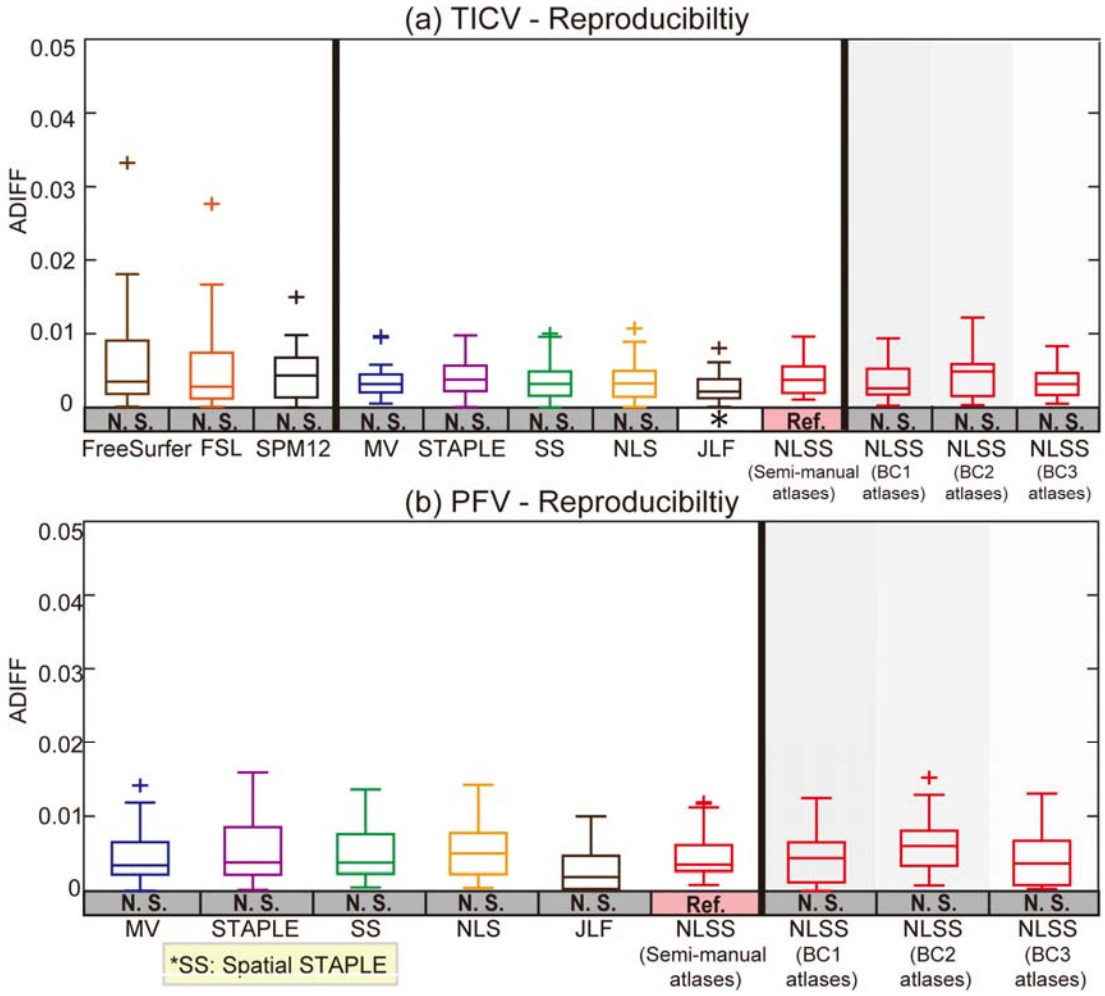


Figure VI.8 Volumetric reproducibility analysis of different approaches on scan-rescan T1w images. For all methods, inconsistency of TICV estimation between two scans on the same subject is less than 2%. The statistical analyses were conducted between the proposed NLSS TICV estimation using semi-manual atlases (marked as reference "Ref.") with other approaches or different atlases. If the difference was statistically significant, we marked the other method with "*" symbol. Otherwise, we marked it as "N.S."

4.2. Reproducibility Test

We employed the publicly available Kirby21 dataset (https://www.nitrc.org/projects/multimodal), which consisted of scan-rescan images on 21 subjects [133]. Each subject had two scans with multispectral MR data (e.g., MPRAGE, FLAIR, DIT etc.) and we used 42 T1w MPRAGE images (with 1×1×1.2 mm resolution over an FOV of 240×204×256 mm) in this reproducibility test. Ideally, the TICV and PFV estimations between two scans from the same subject should be close to each other.

Figure VI.8 demonstrated the reproducibility of different methods on the same 21 pairs of scan-rescan T1w images. We used the ADIFF metric to reflect the ratio of the different volume in the total volume. The results indicated that for both TICV and PFV estimations, all methods achieved small ADIFF values (mostly smaller than 2%).

4.3. Sensitivity of Non-local Search Parameters

In NLSS, the search neighborhood $\mathcal{N}(\cdot)$ and the patch neighborhood $\wp(\cdot)$ are the two essential parameters of controlling the non-local search range and the size of patch. Figure VI.9 demonstrates the Sensitivity to NLSS non-local search parameters: (a) the sensitivity of search neighborhood $\mathcal{N}(\cdot)$, and (b) the sensitivity of patch neighborhood $\wp(\cdot)$. The $\mathcal{N}(\cdot)$ and $\wp(\cdot)$ are evaluated using six different sizes of dimensions: 1×1×1, 3×3×3, 5×5×5, 7×7×7, 9×9×9, and 11×11×11 (voxels). The Dice and MSD are provided for both TICV and PFV estimation using NLSS multi-atlas segmentation framework. Gray outlines indicate the values use in the experiments of this chapter.

5. Conclusion and Discussion

This chapter proposes the simultaneous TICV and PFV estimation framework using multi-atlas label fusion. Using the NLSS multi-atlas framework, we are able to obtain accurate TICV and PFV estimation simultaneously with explicit boundary between skull and CSF. The mathematical derivation is provided for NLSS. The performance of the proposed method was compared with prevalent FreeSurfer, FSL, and SPM12 methods and the previously proposed STAPLE based TICV estimation. For more complete analyses, the NLSS method is also compared with MV, Spatial STAPLE, NLS and JLF.

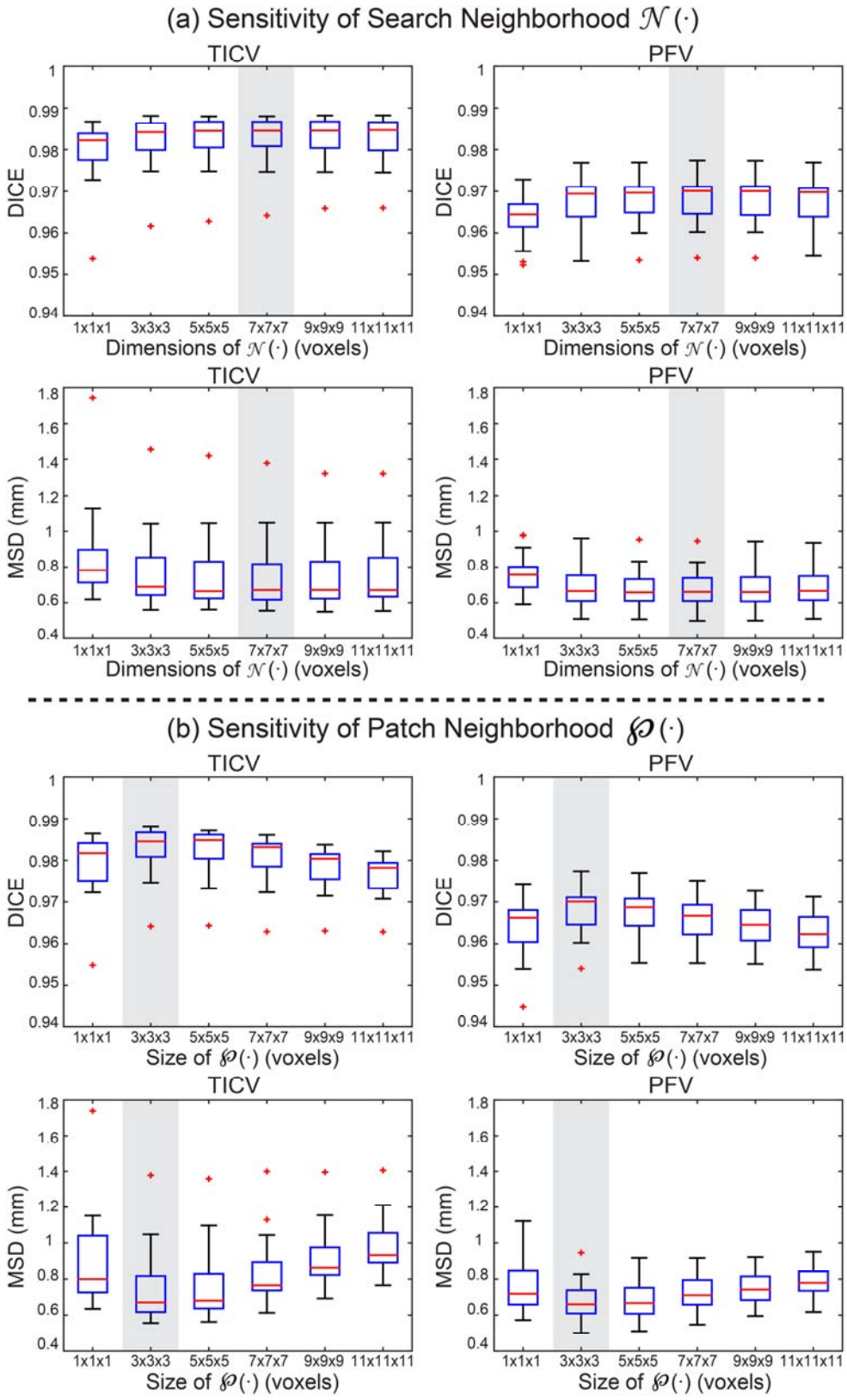


Figure VI.9 Sensitivity to NLSS non-local search parameters.

Compared with the FreeSurfer, FSL, SPM12, the proposed NLSS approach achieves significant superior performance in TICV estimation with highest correlation coefficients, mean ASIM, mean Dice and lowest mean MSD (Table VI.1 and Table VI.2, Figure VI.3). Compared with other label-fusion methods (Figure VI.4, 5 and 6): (1) NLSS approach achieves statistical better performance in simultaneous TICV and PFV estimation than the previously proposed STAPLE method [265]. (2) NLSS approach achieves statistical superior performance than MV, Spatial STAPLE and NLS). (3) For ASIM, Dice and MSD, the differences between NLSS and JLF are not statistically significant, which means NLSS and JLF are comparable accurate in TICV and PFV estimation. From Table VI.1 and Table VI.2, the JLF has overall better measurements in TICV estimation, while the NLSS has better measurements in PFV estimation. From Figure VI.8, all methods achieve high reproducibility (ADIFF<0.2). JLF method achieves statistical smaller ADIFF score than NLSS method on TICV estimation. Overall considering all results, JLF is superior on TICV side while NLSS is superior on PFV side when conducting the simultaneous TICV and PFV estimation.

The accuracy and reproducibility are the two essential aspects when evaluating the performance of TICV estimation. FreeSurfer, FSL and SPM12 achieves high reproducibility demonstrates that the affine registration and tissue segmentation used in the three methods are reproducible. The superior accuracy and high reproducibility indicate that the multi-atlas based approaches do not compromise on reproducibility while providing more accurate estimations. The multi-atlas labeling approaches not only provide more accurate TICV estimation but also estimates PFV simultaneously (which is not available in FreeSurfer, SPM12 and FSL). The PFV is essential in investigating the clinical conditions of the cerebellum [262-264]. The continuing investigation of this work would be on the relationship between the accuracy of TICV estimations and the power of detecting differences between empirical datasets. For instance, we could evaluate the statistical power of detecting the differences of particular metrics (corrected by TICV) between patients and controls using different TICV estimation methods.

We provide new TICV and PFV labels on the widely used 45 OASIS images using BrainCOLOR protocol. The new atlases enable simultaneous BrainCOLOR, TICV and PFV segmentation from only one

set of time-consuming non-rigid registration. To evaluate the performance of the new BC1, BC2 and BC3 atlases, we compared them with semi-manual atlases using the same NLSS framework. Using these intermediate atlases, we lost less than 2% of accuracy from ASIM and Dice score and increased the MSD to less than 0.5 mm compared with directly using semi-manual atlases. However, the performances of BC1, BC2 and BC3 atlases are still better than FreeSurfer, FSL and SPM12 (Table VI.1). Since the BC2 atlases have included original BrainCOLOR labels, we provide these BC2 atlases freely available online to facilitate other researchers (https://www.nitrc.org/frs/?group_id=385). The T1w MR images of the same OASIS images for BC2 atlases are available via subscription from Neuromorphometrics Inc. (http://www.neuromorphometrics.com/) and a subset of them are freely available from MICCAI 2012 Grand Challenge and Workshop on Multi-Atlas Labeling [228] (https://masi.vuse.vanderbilt.edu/workshop2012/).

The semi-manual atlas generation method may be applied on other datasets if paired MR and CT images are available. The rigid registration is used to align CT and MRI images in this study. The registration performance might be affected if huge neck/jaw movements happen in either modality. For such cases, applying a brain mask (masking out neck and jaw) before registration would address the movement issue. The proposed NLSS multi-atlas segmentation framework is flexible in terms of incorporating other regions of interest during TICV estimation. For example, recently, multi-atlas labeling has been used to label brain skull on CT-MRI datasets [290]. In TICV and posterior fossa estimation, we only interested in the accuracy of the inner skull boundary, so we did not seek to fully characterize the cranium. However, it would be interesting to simultaneously provide TICV, PFV and skull labels in the future. The TICV estimation using multi-atlas segmentation is computationally more expensive than using FreeSurfer, FSL and SPM since multiple non-rigid registrations ($\approx$1.5 hours per registration) are conducted for a target image. However, the total length of running time can be reduced by running such independent registrations in parallel. Moreover, the computed registration can be used for other purpose (e.g. segmenting other brain structure, morphometry, manifold learning etc.).

Chapter VII. Mapping Lifetime Brain Volumetry with Covariate-Adjusted Restricted Cubic Spline Regression from Cross-sectional Multi-site MRI

1. Introduction

Brain volumetry across the lifespan is essential in neurological research and clinical investigation. Magnetic resonance imaging (MRI) allows for quantification of such changes, and consequent investigation of specific age ranges or more sparsely sampled lifetime data [1]. Contemporaneous advancements in data sharing have made considerable quantities of brain images available from normal, healthy populations. However, the regression models prevalent in volumetric mapping (e.g., liner, polynomial, non-parametric model, etc.) have had difficulty in modeling complex, cross-sectional large cohorts while accounting for confound effects.

This chapter proposes a novel multi-site cross-sectional framework using Covariate-adjusted Restricted Cubic Spline (C-RCS) regression to map brain volumetry on a large cohort (5111 MR 3D images) across the lifespan (4~98 years). The C-RCS extends the Restricted Cubic Spline [291, 292] by regressing out the confound effects in a general linear model (GLM) fashion. Multi-atlas segmentation is used to obtain whole brain volume (WBV) and 132 regional volumes. The regional volumes are further grouped to 15 networks of interest (NOIs). Then, structural covariance networks (SCNs), i.e. regions or networks that mature or decline together during developmental periods, are established based on NOIs using hierarchical clustering analysis (HCA). To validate the large-scale framework, confidence intervals (CI) are provided for both C-RCS regression and clustering from 10,000 bootstrap samples.

Table VII.1 Data summary of 5111 multi-site images.

Study Name	Website	Images	Sites
Baltimore Longitudinal Study of Aging (BLSA)	www.blsa.nih.gov	605	4
Cutting Pediatrics	vkc.mc.vanderbilt.edu/ebrl	586	2
Autism Brain Imaging Data Exchange (ABIDE)	fcon_1000.projects.nitrc.org/indi/abide	563	17
Information eXtraction from Images (IXI)	www.nitrc.org/projects/ixi_dataset	523	3
Attention Deficit Hyperactivity Disorder (ADHD200)	fcon_1000.projects.nitrc.org/indi/adhd200	949	8
National Database for Autism Research (NDAR)	ndar.nih.gov	328	6
Open Access Series on Imaging Study (OASIS)	www.oasis-brains.org	312	1
1000 Functional Connectome (fcon_1000)	fcon_1000.projects.nitrc.org	1102	22
Nathan Kline Institute Rockland (NKI_rockland)	fcon_1000.projects.nitrc.org/indi/enhanced	143	1

2. Methods

2.1. Extracting Volumetric Information

The complete cohort aggregates 9 datasets with a total 5111 MR T1w 3D images from normal healthy subjects (Table VII.1). 45 atlases are non-rigidly registered [88] to a target image and non-local spatial staple (NLSS) label fusion [98] is used to fuse the labels from each atlas to the target image using the BrainCOLOR protocol [170] (Figure VII.1). WBV and regional volume are then calculated by multiplying the volume of a single voxel by the number of labeled voxels in original image space. In total, 15 NOIs are defined by structural and functional covariance networks including visual, frontal, language, memory, motor, fusiform, basal ganglia (BG) and cerebellum (CB).

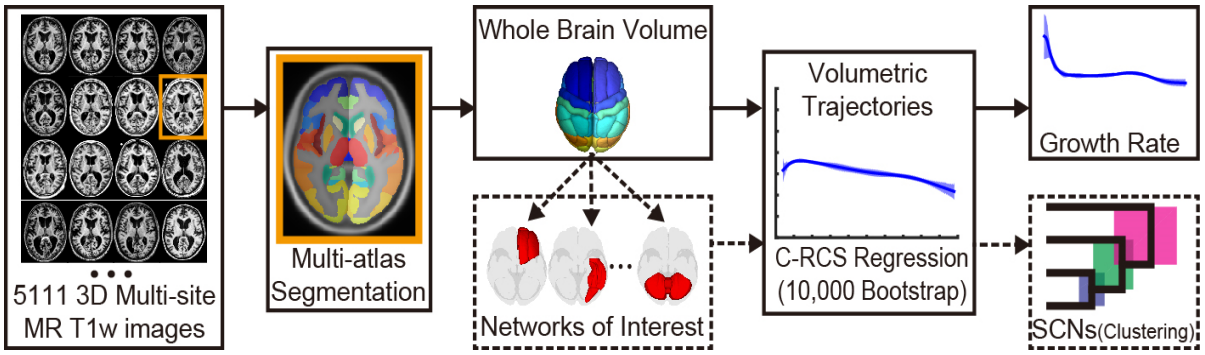


Figure VII.1 The large-scale cross-sectional framework on 5111 multi-site MR 3D images.

2.2. Covariate-Adjusted Restricted Cubic Spline (C-RCS)

We define x as the ages of all subjects and $S(x)$ as the corresponding brain volumes. In canonical nth degree spline regression, splines are used to model non-linear relationships between variables $S(x)$ and x by deciding the connections between K knots $(t_1 < t_2 < \cdots < t_K)$. In this work, such knots were determined based on previously identified developmental shifts [1], specifically corresponding with transitions between childhood (7-12), late adolescence (12-19), young adulthood (19-30), middle adulthood (30-55), older adulthood (55-75), and late life (75-90). Using the expression from Durrleman [291], the canonical nth degree spline function is defined as

$$S(x) = \sum_{j=0}^{n} \dot{\beta}_{oj} x^j + \sum_{i=1}^{K} \dot{\beta}_{in} (x - t_i)_+^n \tag{7.1}$$

where $(x - t_i)_+ = x - t_i,$ if $x > t_i$; $(x - t_i)_+ = 0,$ if $x \leq t_i$.

To regress out confound effects, new covariates $X_1', X_2', \ldots, X_c'$ (with coefficients $\beta_1', \beta_2', \ldots, \beta_c'$) are introduced to the nth degree spline regression

$$S(x) = \sum_{j=0}^{n} \dot{\beta}_{oj} x^j + \sum_{i=1}^{K} \dot{\beta}_{in} (x - t_i)_+^n + \sum_{u=0}^{C} \beta_u' X_u' \tag{7.2}$$

where C is the number of confound effects.

In the RCS regression, a linear constrain is introduced [291] to address the poor behavior of the cubic spline model in the tails ($x < t_1$and $x > t_K$)[293]. Using the same principle, C-RCS regression

extends the RCS regression ($n = 3$) and restricts the relationship between $S(x)$ and x to be a linear function in the tails. First, for $x < t_1$,

$$S(x) = \dot{\beta}_{00} + \dot{\beta}_{01}x + \dot{\beta}_{02}x^2 + \dot{\beta}_{03}x^3 + \dot{\beta}_{13} + \sum_{u=0}^{C} \beta'_u X'_u \tag{7.3}$$

where $\dot{\beta}_{02} = \dot{\beta}_{03} = 0$ ensures the linearity before the first knot. Second, for $x > t_K$,

$$S(x) = \dot{\beta}_{00} + \dot{\beta}_{01}x + \dot{\beta}_{13}(x - t_1)_+^3 + \cdots + \dot{\beta}_{K3}(x - t_K)_+^3 + \sum_{u=0}^{C} \beta'_u X'_u \tag{7.4}$$

To guarantee the linearity of C-RCS after the last knot, we expand the previous expression and force the coefficients of x^2 and x^3 to be zero. After expansion,

$$\begin{aligned} S(x) = &\left(\dot{\beta_{00}} + \dot{\beta}_{13}t_1^3 + \cdots + \dot{\beta}_{K3}t_K^3 + \sum_{u=0}^{C} \beta'_u X'_u\right) \\ &+ \left(\dot{\beta}_{01} + 3\dot{\beta}_{13}t_1^2 + \cdots + 3\dot{\beta}_{K3}t_K^2\right)x \\ &+ \left(3\dot{\beta}_{13}t_1 + 3\dot{\beta}_{23}t_2 + \cdots + 3\dot{\beta}_{K3}t_K\right)x^2 \\ &+ \left(3\dot{\beta}_{13} + 3\dot{\beta}_{23} + \cdots + 3\dot{\beta}_{K3}\right)x^3 \end{aligned} \tag{7.5}$$

As a result, linearity of $S(x)$ at $x > t_K$ implies that $\sum_{i=1}^{K} \dot{\beta}_{i3}t_i = 0$ and $\sum_{i=1}^{K} \dot{\beta}_{i3} = 0$. Following such restrictions, the $\dot{\beta}_{(K-1)3}$ and $\dot{\beta}_{K3}$ are derived as

$$\dot{\beta}_{(K-1)3} = -\frac{\sum_{i=1}^{K-2} \dot{\beta}_{i3}\,(t_K - t_i)}{t_K - t_{K-1}} \text{ and } \dot{\beta}_{K3} = \frac{\sum_{i=1}^{K-2} \dot{\beta}_{i3}\,(t_{K-1} - t_i)}{t_K - t_{K-1}} \tag{7.6}$$

and the complete C-RCS regression model is defined as

$$\begin{aligned} S(x) = \dot{\beta}_{00} + \dot{\beta}_{01}x + \sum_{i=1}^{K-2} \dot{\beta}_{i3}\,[&(x - t_i)_+^3 - \frac{t_K - t_i}{t_K - t_{K-1}}(x - t_{K-1})_+^3 \\ &+ \frac{t_{K-1} - t_i}{t_K - t_{K-1}}(x - t_K)_+^3] + \sum_{u=0}^{C} \beta'_u X'_u \end{aligned} \tag{7.7}$$

2.3. Regressing Out Confound Effects by C-RCS Regression in GLM Fashion

To adapt C-RCS regression in the GLM fashion, we redefine the coefficients $\beta_0, \beta_1, \beta_2, \ldots, \beta_{K-1}$ as

Harrell [292] where $\beta_0 = \dot{\beta}_{00}, \beta_1 = \dot{\beta}_{01}, \beta_2 = \dot{\beta}_{13}, \beta_3 = \dot{\beta}_{23}, \beta_4 = \dot{\beta}_{33}, \cdots, \beta_{K-1} = \dot{\beta}_{(K-2)3}$. Then, the C-RCS regression with confound effects becomes

$$S(x) = \beta_0 + \sum_{j=1}^{K-1} \beta_j X_j + \sum_{u=0}^{C} \beta'_u X'_u \tag{7.8}$$

where C is the number for all confound effects (X'_u). $X_1 = x$ and for $j = 2, \dots, K-1$

$$X_j = \left(x - t_{j-1}\right)_+^3 - \frac{t_K - t_{j-1}}{t_K - t_{K-1}}(x - t_{K-1})_+^3 + \frac{t_{K-1} - t_{j-1}}{t_K - t_{K-1}}(x - t_K)_+^3 \tag{7.9}$$

Then, the beta coefficients are solvable under GLM framework. Once $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_{K-1}$ are obtained, two linear assured terms $\hat{\beta}_K$ and $\hat{\beta}_{K+1}$ are estimated:

$$\hat{\beta}_K = \frac{\sum_{i=2}^{K-1} \hat{\beta}_i \, (t_{i-1} - t_K)}{t_K - t_{K-1}} \text{ and } \hat{\beta}_{K+1} = \frac{\sum_{i=2}^{K-1} \hat{\beta}_i \, (t_{i-1} - t_{K-1})}{t_{K-1} - t_K} \tag{7.10}$$

The final estimated volumetric trajectories $\hat{S}(x)$ can be fitted as

$$\hat{S}(x) = \hat{\beta}_0 + \sum_{j=1}^{K+1} \hat{\beta}_j (x - t_j)_+^3 + \sum_{u=0}^{C} \hat{\beta}'_u X'_u \tag{7.11}$$

In this work, gender, field strength and total intracranial volume (TICV) are employed as covariates X'_u. TICV values are calculated using SIENAX [257]. Field strength and TICV are used to regress out site effects rather than using site categories directly since the sites are highly correlated with the explanatory variable age.

2.4. SCNs and CI using Bootstrap Method

Using aforementioned C-RCS regression, the lifespan volumetric trajectories of WBV and 15 NOIs are obtained from 5111 images. Simultaneously, the piecewise volumetric trajectories within a particular age bin (between adjacent knots) of all 15 NOIs ($\hat{S}_i(x), i = 1,2, \dots,15$) are separated to establish SCNs dendrograms using HCA [294]. The distance metric D used in HCA is defined as $D = 1 - \text{corr}(\hat{S}_i(x), \hat{S}_j(x))$, $i, j \in [1,2, \dots,15]$ and $i \neq j$, where $\text{corr}(\cdot)$ is the Pearson's correlation between any two C-RCS fitted piecewise trajectories $\hat{S}_i(x)$ and $\hat{S}_j(x)$ in the same age bin.

The stability of proposed approaches is demonstrated by the CIs of C-RCS regression and SCNs using bootstrap method [295]. First, the 95% CIs of volumetric trajectories on WBV (Figure VII.2) and 15 NOIs (Figure VII.3) are derived by deploying C-RCS regression on 10,000 bootstrap samples. Then, the distances D between all pairs of clustered NOIs are derived using 15 (NOIs) ×10,000 (bootstrap) C-RCS fitted trajectories. Then, the 95% CIs are obtained for each pair of clustered NOIs and shown on six SCNs dendrograms (Figure VII.4). The average network distance (AND), the average distance between 15 NOIs for a dendrogram, can be calculated 10,000 times using bootstrap. The AND reflects the modularity of connections between all NOIs. We are able to see if the AND are significantly different during brain development periods by deploying the two-sample t-test on AND values (10,000/age bin) between age bins.

3. Results

Figure VII.2a shows the lifespan volumetric trajectories using C-RCS regression as well as the growth rate (volume change in percentage per year) of WBV when regressing out gender and field strength effects. Figure VII.2b indicates the C-RCS regression on the same dataset by adding TICV as an additional covariate. The cross sectional growth rate curve using C-RCS regression is compared with 40 previous longitudinal studies (19 are TICV corrected)[1], which are typically limited on smaller age ranges.

Using the same C-RCS model in Figure VII.2b, Figure VII.3 indicates the both lifespan and piecewise volumetric trajectories of 15 NOIs. In Figure VII.4, the piecewise volumetric trajectories of the 15 NOIs within each age bin are clustered using HCA and shown in one SCNs dendrogram.

Then, six SCNs dendrograms are obtained by repeating HCA on different age bins, which demonstrate the evolution of SCNs during different developmental periods. The ANDs between any two age bins in Figure VII.4 are statistically significant ($p<0.001$).

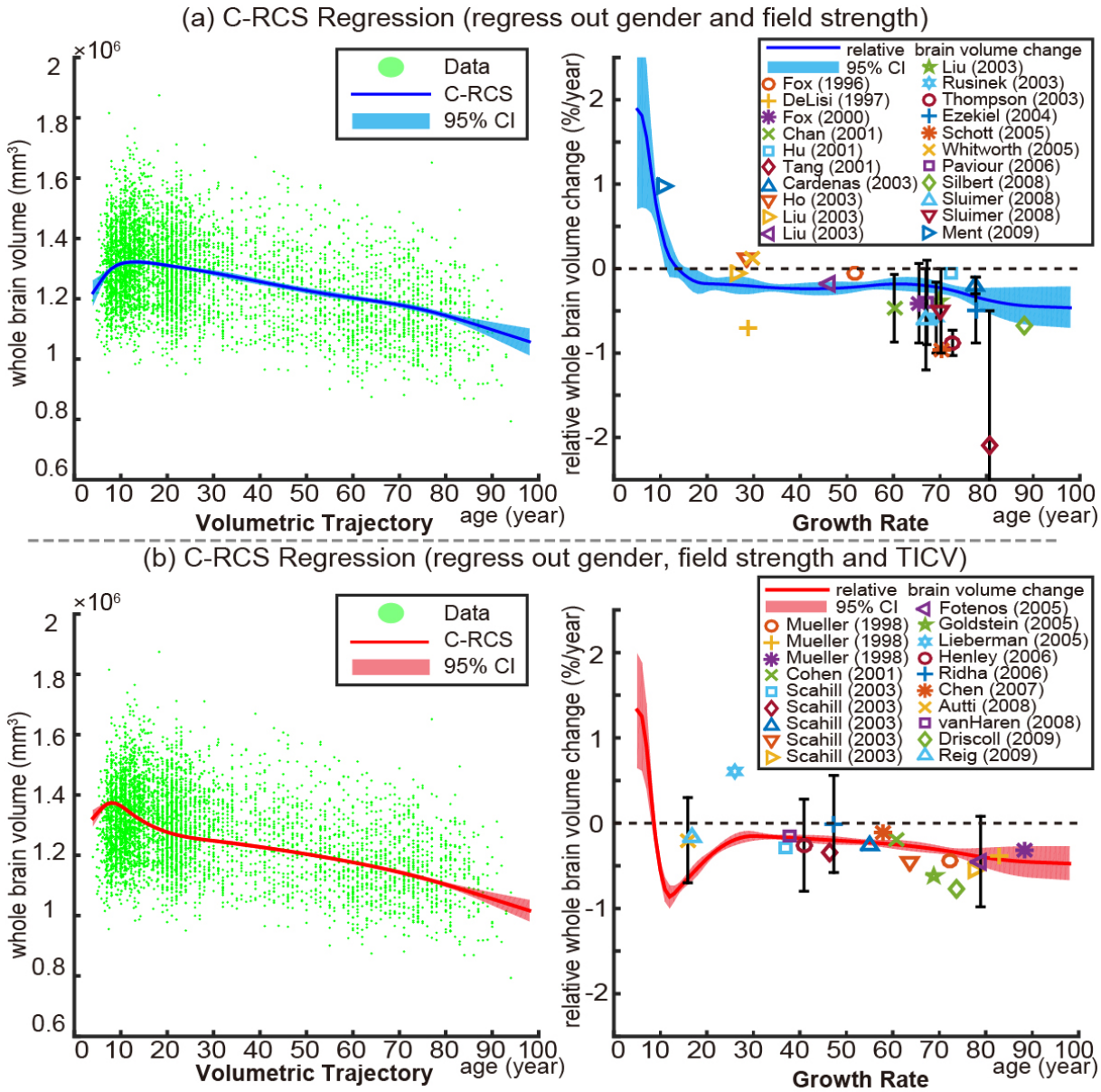


Figure VII.2 Volumetry and growth rate. The left plot in (a) shows the volumetric trajectory of whole brain volume (WBV) using C-RCS regression on 5111 MR images. The right figure in (a) indicates the growth rate curve, which shows volumetric change per year of the volumetric trajectory. In (b), C-RCS regression is deployed on the same dataset by additionally regressing out TICV. Our growth rate curves are compared with 40 previous longitudinal studies [1] on smaller cohorts (21 studies in (a) without regressing out TICV and 19 studies in (b) regressing out TICV). The standard deviations of previous studies are provided as black bars (if available). The 95% CIs in all plots are calculated from 10,000 bootstrap samples

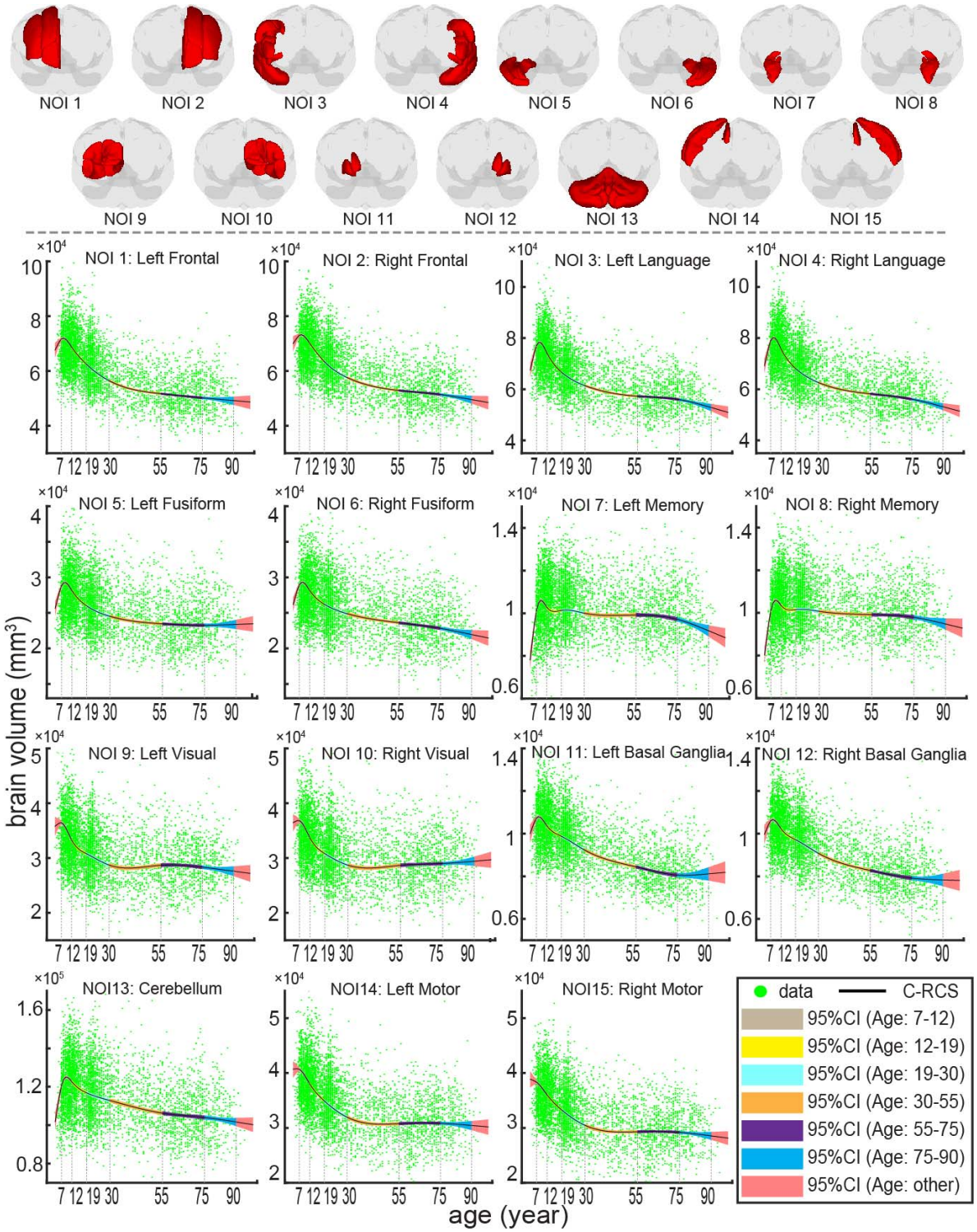


Figure VII.3 Lifespan trajectories of 15 NOIs are provided with 95% CI from 10,000 bootstrap samples. The upper 3D figures indicate the definition of NOIs (in red). The lower figures show the trajectories with CI using C-RCS regression method by regressing out gender, field strength and TICV (same model as Figure VII.2b). For each NOI, the piecewise CIs of six age bins are shown in different colors. The piecewise volumetric trajectories and CIs are separated by 7 knots in the lifespan C-RCS regression rather than conducting independent fittings. The volumetric trajectories on both sides of each NOI are derived separately except for CB.

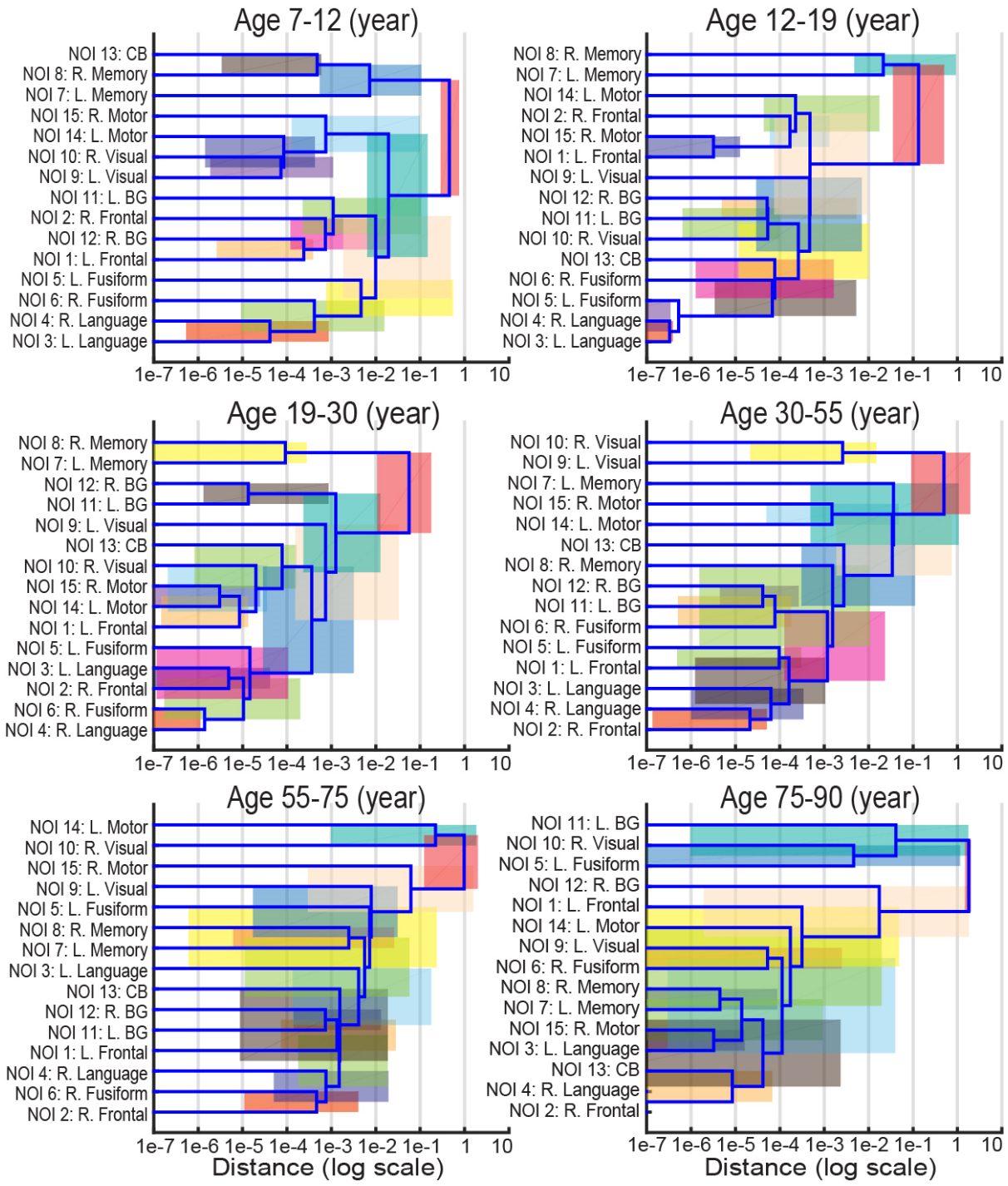


Figure VII.4 The six structural covariance networks (SCNs) dendrograms using hierarchical clustering analysis (HCA) indicate which NOIs develop together during different developmental periods (age bins). The distance on the x-axis is in log scale, which equals to one minus Pearson's correlation between two curves. The correlation between NOIs becomes stronger from right to left on the x-axis. The horizontal range of each colored rectangles indicates the 95% CI of distance from 10,000 bootstrap samples. Note that the colors are chosen for visualization purposes without quantitative meanings.

4. Conclusion and Discussion

This chapter proposes a large-scale cross-sectional framework to investigate life-time brain volumetry using C-RCS regression. C-RCS regression captures complex brain volumetric trajectories across the lifespan while regressing out confound effects in a GLM fashion. Hence, it can be used by researchers within a familiar context. The estimated volume trends are consistent with 40 previous smaller longitudinal studies. The stable estimation of volumetric trends for NOI (exhibited by narrow confidence bands) provides a basis for assessing patterns in brain changes through SCNs. Moreover, we demonstrate how to compute confidence intervals for SCNs and correlations between NOIs. The significant difference of AND indicates that the C-RCS regression detects the changes of average SCNs connections during the brain development.

Emerging "big data" studies need a regression that is able to capture the complicated lifespan brain development without unnecessarily sacrificing power. The proposed C-RCS regression is a such framework that addresses age-range analyses and varied neuroanatomical regions of interest. To the best of our knowledge, this is the first work that uses C-RCS to quantify temporal changes in SCNs using brain volumetry with a cross-sectional, multi-site paradigm. The challenge of using C-RCS method is that the knots should be defined properly. The software is freely available online[1].

[1] https://www.nitrc.org/frs/?group_id=385

Chapter VIII. 4D Multi-atlas Label Fusion using Longitudinal Images

1. Introduction

An essential challenge in volumetric (3D) image segmentation on longitudinal medical images is to ensure the temporal consistency while retaining sensitivity. The consistency of longitudinal segmentation is essential to control the "type I" false positive error while the sensitivity of longitudinal segmentation is important to control the "type II" false negative error. One wants to control both two types of errors when investigating clinical studies (e.g., understanding normal aging [128, 129]). Many efforts have been made to incorporate the temporal dimension into volumetric segmentation (4D) for the studies. One family of 4D methods is to control the longitudinal variations during pre/post-processing using 4D intensity filtering [296], 4D registration [297], or temporal mean template [298]. These methods control inter-subject variations between target images, which result in more consistent 3D segmentations. Another family of 4D methods is to incorporate the longitudinal variations within segmentation methods, such as 4D fuzzy C-means [299] or 4D graph-cuts [300]. In the past decade, multi-atlas segmentation (MAS) has been regarded as de facto standard segmentation method in 3D scenarios [203]. To improve the performance of 4D MAS for longitudinal data, several previous avenues have been explored. Li $et\ al.$ [301] proposed a MAS based 4D surface labeling approach, which minimized a spatial temporal energy function. However, the energy function is designed for using surface features (e.g., shape, cortical folding geometries etc.), which is limited to surface labeling. Guo $et\ al.$ [302] proposed a hierarchical feature learning approach to obtain common feature representations using longitudinal multi-modal (T1 and T2) images. However, the application is restricted on the availability of multi-modal longitudinal data. Wang $et\ al.$ [303] proposed a 4D label fusion method with temporal sparse representation technique, which was not limited by applications or modalities. However, this method (1) only considered two consecutive time points (t and t+1) in the temporal smooth term, and (2) assumed the l1-norm sparsity of fusion weights. When more than two longitudinal target images are available, the more comprehensive strategy is to consider the spatial smoothness on all time points simultaneously (Figure VIII.1). Moreover, in the general label fusion framework (without sparsity limitation), the voting based [91] and statistical fusion [92] have been successfully applied in 3D image

segmentations [203], which motivated this work proposed a general purpose 4D label fusion theory that simultaneously considers all available longitudinal images (time points) and can be adapted to different applications.

2. Theory

2.1. Model Definition

Let one target image be represented by $T_t, t \in [1,2,\dots,k]$. 4DJLF considers all available longitudinal target images, $\mathbf{T} = \{T_1, T_2, \dots, T_k\}$. First, all longitudinal target images are registered to the first-time point using rigid registration [89]. n pairs of atlases $\mathbf{A} = \{A_1, A_2, \dots, A_n\}$ are available in the MAS, where each pair consists of one intensity atlas and one label atlas. Then, we register the n intensity atlases to k longitudinal target images to achieve $m = n \times k$ registered pairs of atlases. For mathematical convenience, we concatenate all registered atlases (based on the sequence in **T**) to derive m registered intensity atlases set **I** and m registered label atlases set **S** as

$$\begin{aligned} \mathbf{I} &= \{I_1^{(1)}, \dots, I_n^{(1)}, I_{n+1}^{(2)}, \dots, I_{2n}^{(2)}, \ \cdots \ , I_{2n+1}^{(k)}, \dots, I_m^{(k)}\} \\ \mathbf{S} &= \{S_1^{(1)}, \dots, S_n^{(1)}, S_{n+1}^{(2)}, \dots, S_{2n}^{(2)}, \ \cdots \ , S_{2n+1}^{(k)}, \dots, S_m^{(k)}\} \end{aligned} \tag{8.1}$$

where the superscripts "$(\cdot)$" indicate to which target image that atlas was registered.

The k longitudinal target images provide m registered atlases, where each atlas correspond to one time point (target image). The rationale of boosting the registrations is to reconcile the registration errors and intensity inhomogeneity among **T** under the hypothesis that **T** are similar but not identical to each other. In the weighted voting framework, the consensus segmentation $\bar{S}$ for voxel x on t_{th} target image is

$$\bar{S}^t(x) = \sum_{i=1}^{m} w_i^t(x) S_i(x) = \mathbf{w}^t(x) \cdot \mathbf{S}(x) \tag{8.2}$$

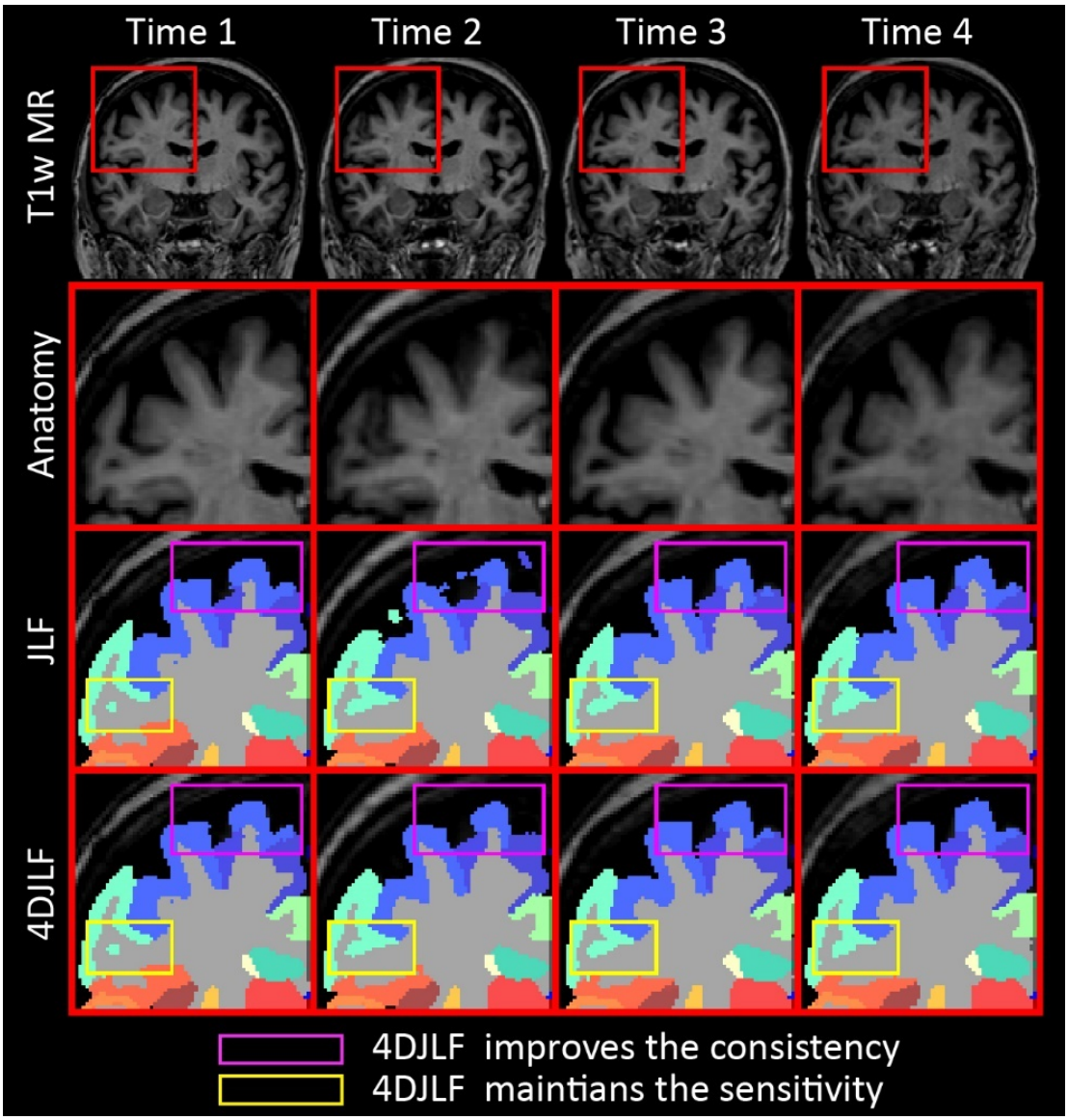


Figure VIII.1 An example of the inconsistency of 3D joint label fusion (JLF) segmentation across longitudinal multiple scans from the same subject. The 4DJLF is proposed to improve the consistency while maintain the sensitivity.

where $\mathbf{w}^t(x) = \{w_1^t(x), w_2^t(x), \ldots, w_m^t(x)\}$ are spatially varying weights restricted by $\sum_{i=1}^{m} w_i^k(x) = 1$. Adopting [91], the error $\delta_i^t(x)$ made by atlas S_i on t_{th} target image in the binary segmentation is

$$\delta_i^t(x) = S_T^t(x) - S_i(x) \tag{8.3}$$

where $S_T^t(x)$ is the hidden true segmentation. $\delta_i^t(x) = 0$ indicates the right decision is made, while $\delta_i^t(x) = -1$ or 1 means the wrong decision is made. Then, our purpose is to find a set of voting weights

$\mathbf{w}^t(x)$ for each target image T_t that minimize the total expected error between the automated labeled image $\bar{S}^k$ and hidden true label image S_T^t, given by the following energy function

$$\begin{aligned} & E_{\delta_1^t(x),\ldots,\delta_m^t(x)}\left[\left(S_T^t(x)-\bar{S}^t(x)\right)^2|\mathbf{T},\mathbf{I}\right]= \\ & = E_{\delta_1^t(x),\ldots,\delta_m^t(x)}\left[\left(\sum_{i=1}^m w_i^t(x)\,\delta_i^t(x)\right)^2|\mathbf{T},\mathbf{I}\right] \\ & = \sum_{i=1}^m\sum_{j=1}^m w_i^t(x)w_j^t(x)\,E_{\delta_i^t(x)\delta_j^t(x)}\left[\delta_i^t(x)\delta_j^t(x)|T_1,\ldots,T_k,I_1,\ldots,I_m\right] \\ & = \mathbf{w}_x^{t\,T}\boldsymbol{M}_x^t\mathbf{w}_x^t \end{aligned} \tag{8.4}$$

where $\mathbf{w}_x^{t\,T}$ is the transpose of vector $\mathbf{w}_x^t$ at voxel x. $\boldsymbol{M}_x^t$ is a $m \times m$ pairwise dependency matrix that

$$\boldsymbol{M}_x^t(i,j) = p(\delta_i^t(x)\delta_j^t(x)=1|T_1,\ldots,T_k,I_1,\ldots,I_m) \tag{8.5}$$

Finally, the estimated weights $\widehat{\mathbf{w}}_x^t$, which is our target, is derived by

$$\widehat{\mathbf{w}}_x^t = \arg\min_{\mathbf{w}_x^t} \mathbf{w}_x^{t\,T}(\boldsymbol{M}_x^t + \alpha\boldsymbol{I})\mathbf{w}_x^t \tag{8.6}$$

where α is a small positive constant (e.g., $\alpha = 0.1$ in the experiments) and $\boldsymbol{I}$ is a $m \times m$ diagonal matrix. The $\alpha\boldsymbol{I}$ is used to ensure the unique solution of $\widehat{\mathbf{w}}_x^t$.

2.2. JLF-Multi

As a baseline, we consider a simple temporal model (JLF-Multi)to performing the 4D label fusion. We assume that each target image in $\mathbf{T}$ contributes equally to the label fusion for target T_t. In this case, Eq. 8.5 is can be approximated as the following expression

$$\boldsymbol{M}_x^t(i,j) \propto \sum_{y\in B(x)} |T_t(y) - I_i(\mathcal{N}_i(y))| \cdot |T_t(y) - I_j(\mathcal{N}_j(y))| \tag{8.7}$$

where the Σ improves the spatial smoothness by adding multiple voxels y in a patch neighborhood $B(x)$ (e.g., $2 \times 2 \times 2$ by default), and the non-local patch searching is conducted within a search neighborhood $\mathcal{N}(y)$ (e.g., $3 \times 3 \times 3$ by default), which are both common practices in state-of-the-art label fusion methods [91, 92].

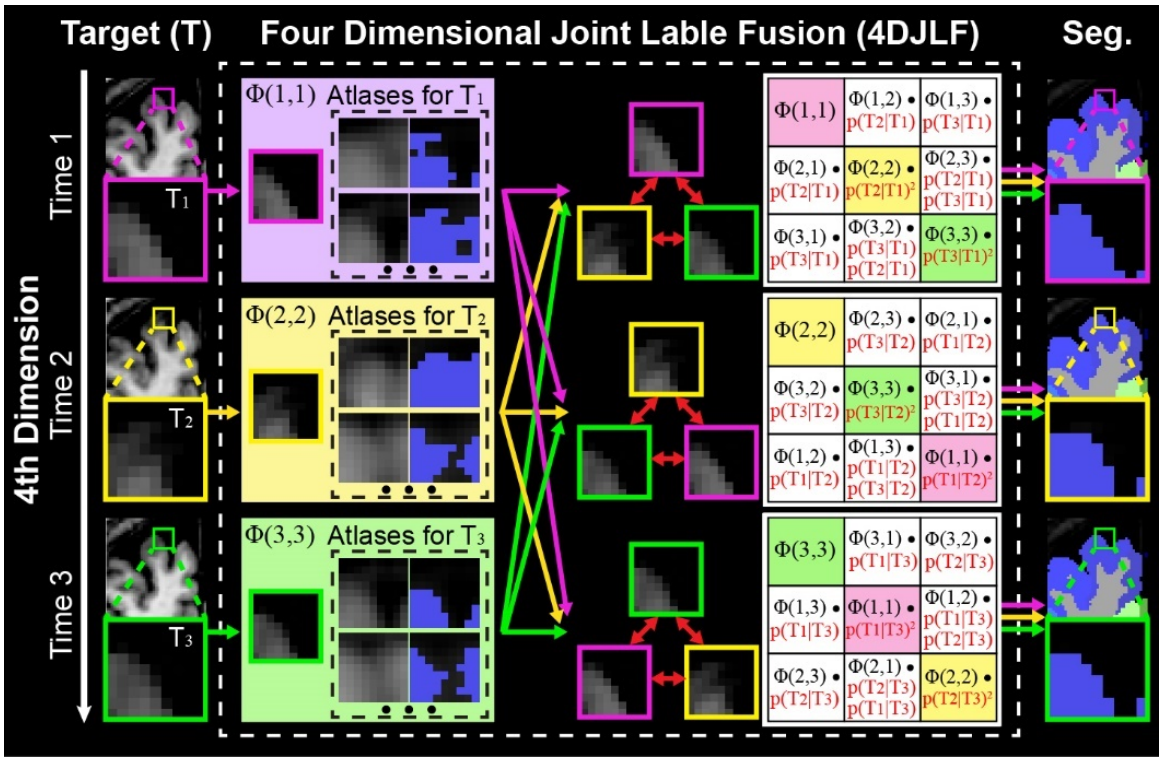


Figure VIII.2 The 4DJLF framework. First, the same set of atlases are registered to the longitudinal target images (3 time points in figure). Then, the Φ matrices are calculated using Eq. 8.13. Finally, the spatial temporal performance of all atlases are model by Eq. 8.14, which leads to the final segmentations ("Seg."). Note that the upper right 3×3 matrix is identical to Eq. 8.15. The original JLF estimates the block diagonal elements of the generalized covariance matrix (highlighted in magenta, green, and yellow) which would result in independent temporal estimates.

2.3. 4DJLF

In JLF-Multi, each longitudinal target image contributes equally to the 4D label fusion. However, this assumption is not always valid. Considering the case that if target images shown a sudden atrophy after a time point. The solution to keep sensitivity is that the label fusion on a target image with atrophy should trusts much more on the atlases (raters) after the atrophy happened. Herein, we propose the new dependency matrix $\ddot{\boldsymbol{M}}_{\boldsymbol{x}}^{\boldsymbol{t}}(i,j)$ by adaptively evaluating the longitudinal raters' performance on any target image patches using a probabilistic model

$$\ddot{\boldsymbol{M}}_{\boldsymbol{x}}^{\boldsymbol{t}}(i,j) = p\left(T_q(x), T_r(x)\middle|T_t(x)\right) \cdot \left(\sum_{y\in B(x)} \left|T_q(y) - I_i^{(q)}\left(\mathcal{N}_i(y)\right)\right| \cdot \left|T_r(y) - I_j^{(r)}\left(\mathcal{N}_j(y)\right)\right|\right) \tag{8.8}$$

where the new dependency matrix $\ddot{\boldsymbol{M}}_{\boldsymbol{x}}^{\boldsymbol{t}}(i,j)$ not only evaluates the similarity between atlases and target images but also considers the longitudinal similarities between target images. The "(q)" and "(r)" indicate which atlases that I_i and I_j were registered to and the value of q and r are able to be derived from Eq. 8.1. Then, probability of using the raters (atlases) from T_{q} and T_{r} given target T_t is modeled in a conditional probability

$$p\left(T_q(x), T_r(x)\middle|T_t(x)\right) = p\left(T_q(x)\middle|T_t(x)\right) \cdot p(T_r(x)|T_t(x)) \tag{8.9}$$

by assuming T_{q} and T_{r} are conditionally independent, we have

$$p\left(\left(T_q(x)\middle|T_t(x)\right)\right) = \exp\left(\beta \cdot \sum_{y\in B(x)} \frac{\left|T_q(y) - T_t(y)\right|}{\left|T_q(y) - I_i^{(q)}\left(\mathcal{N}_i(y)\right)\right|}\right) \tag{8.10}$$

$$p\left((T_r(x)|T_t(x))\right) = \exp\left(\beta \cdot \sum_{y\in B(x)} \frac{|T_r(y) - T_t(y)|}{\left|T_r(y) - I_j^{(r)}\left(\mathcal{N}_j(y)\right)\right|}\right) \tag{8.11}$$

where β is a sensitivity coefficient and is empirically set to 100 in the experiments.

2.4. Relationship between 4DJLF to JLF

The proposed 4DJLF theory is a generalization of JLF theory, which is not only designed to improve the reproducibility but also maintaining the sensitivity compared with JLF. If the β is set to an extreme large number, the $p\left(T_q(x), T_r(x)\middle|T_t(x)\right)$ will be extreme large for atlases from other time points, but still equals to 1 for the atlases from the target image itself. Therefore, the weights of the atlases from other time points will be infinitely close to zero and only the atlases registered to the target time T_t is considered. In that case, the 4DJLF is degenerated to JLF.

To visualize such relationship (as shown in Figure VIII.2), we redefine the right side of Eq. 8.8 as

the $\Gamma_x(i,j)$

$$\Gamma_x(i,j) = \sum_{y \in B(x)} \cdot \left| T_q(y) - I_i^{(q)}\left(\mathcal{N}_i(y)\right) \right| \cdot \left| T_r(y) - I_j^{(r)}\left(\mathcal{N}_j(y)\right) \right| \tag{8.12}$$

Then, we define a matrix $\Phi_{p,q}$ as the following

$$\Phi_x(q,r) = \begin{bmatrix} \Gamma_x(i',j') & \Gamma_x(i',j'+1) & \cdots & \Gamma_x(i',j'+k) \\ \Gamma_x(i'+1,j') & \Gamma_x(i'+1,j'+1) & \cdots & \Gamma_x(i'+1,j'+k) \\ & \vdots & \ddots & \vdots \\ \Gamma_x(i'+k,j') & \Gamma_x(i'+k,j'+1) & \cdots & \Gamma_x(i'+k,j'+k) \end{bmatrix} \tag{8.13}$$

where $i' = (q-1) \times k + 1$ and $j' = (r-1) \times k + 1$.

For simplify, we assume three longitudinal target images are used and the first time point is the target image (upper row in Figure VIII.2). We use Eq. 8.13 and rewrite the $p\left(\left(T_q(x)\middle|T_t(x)\right)\right)$ as $p_x\left(\frac{T_1}{T_1}\right)$ to visualize the $\ddot{\boldsymbol{M}}^{\boldsymbol{t}}$ at the first time point ($t = 1$ and the subscript x is omitted for simplicity).

$$\ddot{\boldsymbol{M}}^{\boldsymbol{1}} = \begin{bmatrix} \Phi(1,1)p\left(\frac{T_1}{T_1}\right)p\left(\frac{T_1}{T_1}\right) & \Phi(1,2)p\left(\frac{T_1}{T_1}\right)p\left(\frac{T_2}{T_1}\right) & \Phi(1,3)p\left(\frac{T_1}{T_1}\right)p\left(\frac{T_3}{T_1}\right) \\ \Phi(2,1)p\left(\frac{T_2}{T_1}\right)p\left(\frac{T_1}{T_1}\right) & \Phi(2,2)p\left(\frac{T_2}{T_1}\right)p\left(\frac{T_2}{T_1}\right) & \Phi(2,3)p\left(\frac{T_2}{T_1}\right)p\left(\frac{T_3}{T_1}\right) \\ \Phi(3,1)p\left(\frac{T_3}{T_1}\right)p\left(\frac{T_1}{T_1}\right) & \Phi(3,2)p\left(\frac{T_3}{T_1}\right)p\left(\frac{T_2}{T_1}\right) & \Phi(3,3)p\left(\frac{T_3}{T_1}\right)p\left(\frac{T_3}{T_1}\right) \end{bmatrix} \tag{8.14}$$

Since $p\left(\frac{T_1}{T_1}\right) = 1$, the $\ddot{\boldsymbol{M}}^{\boldsymbol{1}}$ is further simplified to

$$\ddot{\boldsymbol{M}}^{\boldsymbol{1}} = \begin{bmatrix} \Phi(1,1) & \Phi(1,2)p\left(\frac{T_2}{T_1}\right) & \Phi(1,3)p\left(\frac{T_3}{T_1}\right) \\ \Phi(2,1)p\left(\frac{T_2}{T_1}\right) & \Phi(2,2)p\left(\frac{T_2}{T_1}\right)^2 & \Phi(2,3)p\left(\frac{T_2}{T_1}\right)p\left(\frac{T_3}{T_1}\right) \\ \Phi(3,1)p\left(\frac{T_3}{T_1}\right) & \Phi(3,2)p\left(\frac{T_3}{T_1}\right)p\left(\frac{T_2}{T_1}\right) & \Phi(3,3)p\left(\frac{T_3}{T_1}\right)^2 \end{bmatrix} \tag{8.15}$$

where the $\ddot{\boldsymbol{M}}^{\boldsymbol{1}}$ is identical to the upper right matrix in Figure VIII.2. Here, note that $\Phi(1,1)$ is the same as the M_x matrix in JLF [91], which also demonstrates the relationship between 4DJLF and JLF.

3. Methods and Results

3.1. Data and Preprocessing

Six healthy subjects with total 21 longitudinal T1-weighted (T1w) MR scans (mean age 82.3, range:72.5~90.2) were randomly selected from were randomly selected from Baltimore Longitudinal Study of Aging (BLSA) [128]. Each image had 170×256×256 voxels with 1.2×1×1 mm resolution. 15 pairs of atlases containing both T1w and label images from BrainCOLOR (http://braincolor.mindboggle.info/protocols/) were employed. The intensity atlases had 1mm isotropic resolution and the label atlases contained 132 labels for entire brain. In order to evaluate the sensitivity, one randomly selected T1w image from a healthy subject (age 11) in ADHD-200 OHSU dataset [304] was used in the robustness test.

The 21 longitudinal target images were first affinely registered [89] to the MNI305 atlas [171].Then, the spatially aligned longitudinal atlases $\mathbf{T} = \{T_1, T_2, \dots, T_k\}$ were derived by rigidly registering each target image to the first time point. Then, 15 atlases were non-rigidly registered [88] to all target images to achieve the intensity and label atlases in Eq. 8.1 (performed $m = 15 \times 21$ non-rigid registrations). The same preprocessing was also deployed on the one AHDH-200 target image.

3.2. Reproducibility Experiment and Results

First, JLF approach were deployed on all 21 longitudinal target images independently using default parameters. The longitudinal reproducibility of JLF was evaluated by calculating the Dice similarity coefficients between all pairs of longitudinal images (Figure VIII.3a). Then JLF-multi and 4D JLF were deployed on the same dataset (using the same default parameters as JLF), whose Dice values between all pairs of longitudinal images were shown in Figure VIII.3b. To statistically compare the reproducibility between methods, Wilcoxon signed rank test and Cohen's d effect size analyses were performed between JLF-Multi vs. JLF and 4D JLF vs. JLF (Table VIII-1). The "*" indicated such difference satisfied (1) $p<0.01$ in Wilcoxon signed rank test, and (2) $d>0.1$ in effect size.

The temporal changes on volume sizes of whole brain, gray matter, white matter and ventricle for all target images were shown in Figure VIII.4. Figure VIII.5 provides two examples of quantitative results

from subject 2 and 5 in Figure VIII.4 respectively.

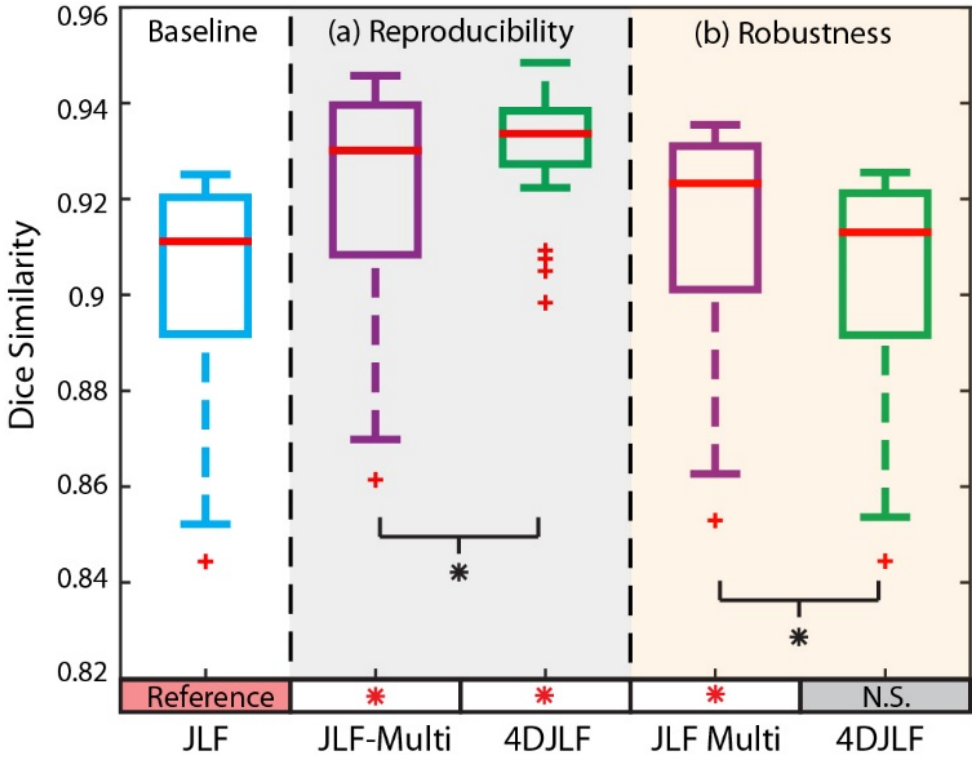


Figure VIII.3 Quantitative results. (a) The reproducibility experiment shown that the proposed 4DJLF had overall significantly better reproducibility than JLF and JLF-Multi. (b) The robustness test indicated that 4DJLF maintained the sensitivity as JLF, while JLF-Multi was not able to do so. The red "*" means the method satisfied both $p<0.01$ and effect size>0.1 compared with JLF ("Reference"), while the "N.S." means at least one was not satisfied. The black "*" means the difference between two methods satisfied both $p<0.01$ and effect size>0.1.

Table VIII-1 Quantitative Results of Reproducibility Experiment

		JLF	JLF-Multi	4DJLF
Dice	mean	0.9032	0.9213	**0.9311**
	std.	0.0221	0.0231	**0.0119**
Difference to JLF	p value	N/A	<0.01	<0.01
	Cohen's d	N/A	0.7983	1.5691

Table VIII-2 Quantitative Results of Robustness Test

		JLF	JLF-Multi	4DJLF
Dice	mean	0.9032	0.9138	0.9043
	std.	0.0221	0.0228	0.0224
Difference to JLF	p value	N/A	<0.01	<0.01
	Cohen's d	N/A	0.4703	**0.0463**

3.3. Robustness Test and Result

Second, a robustness test was conducted to evaluate the sensitivity of JLF, JLF-Multi and 4DJLF methods. In this experiment, we combined the previously mentioned ADHD-200 image to each target image to formed 21 dummy longitudinal pairs. This test simulated the large temporal variations since the two images in each pair were (1) independent (2) collected from different scanners, and (3) had at least 60 years' difference. Then, the 4D segmentation methods were deployed on such cases to see if the 4D methods can maintain the sensitivity compared with JLF. The Figure VIII.3b and Table VIII-2 indicated the 4DJLF had "trivial" changes on reproducibility (effect size <0.1) compared with JLF, while JLF-Multi had large differences compared with JLF.

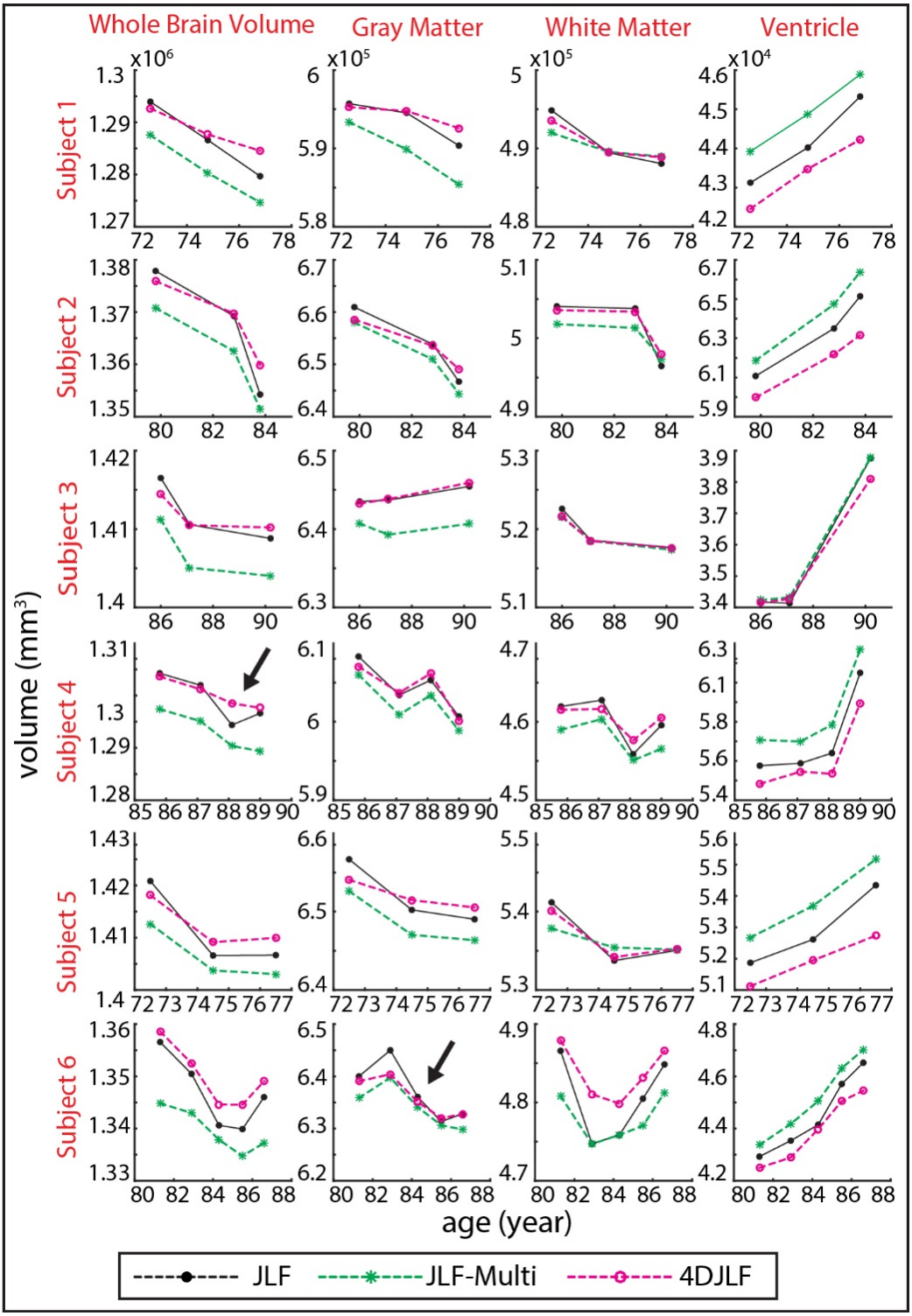


Figure VIII.4 This figure demonstrated the longitudinal changes of whole brain volume, gray matter volume, white matter volume and ventricle volume for all 6 subjects (21 time points). The black arrows indicated that the proposed 4DJLF reconciles some obvious temporal inconsistency by simultaneously considering all available longitudinal images.

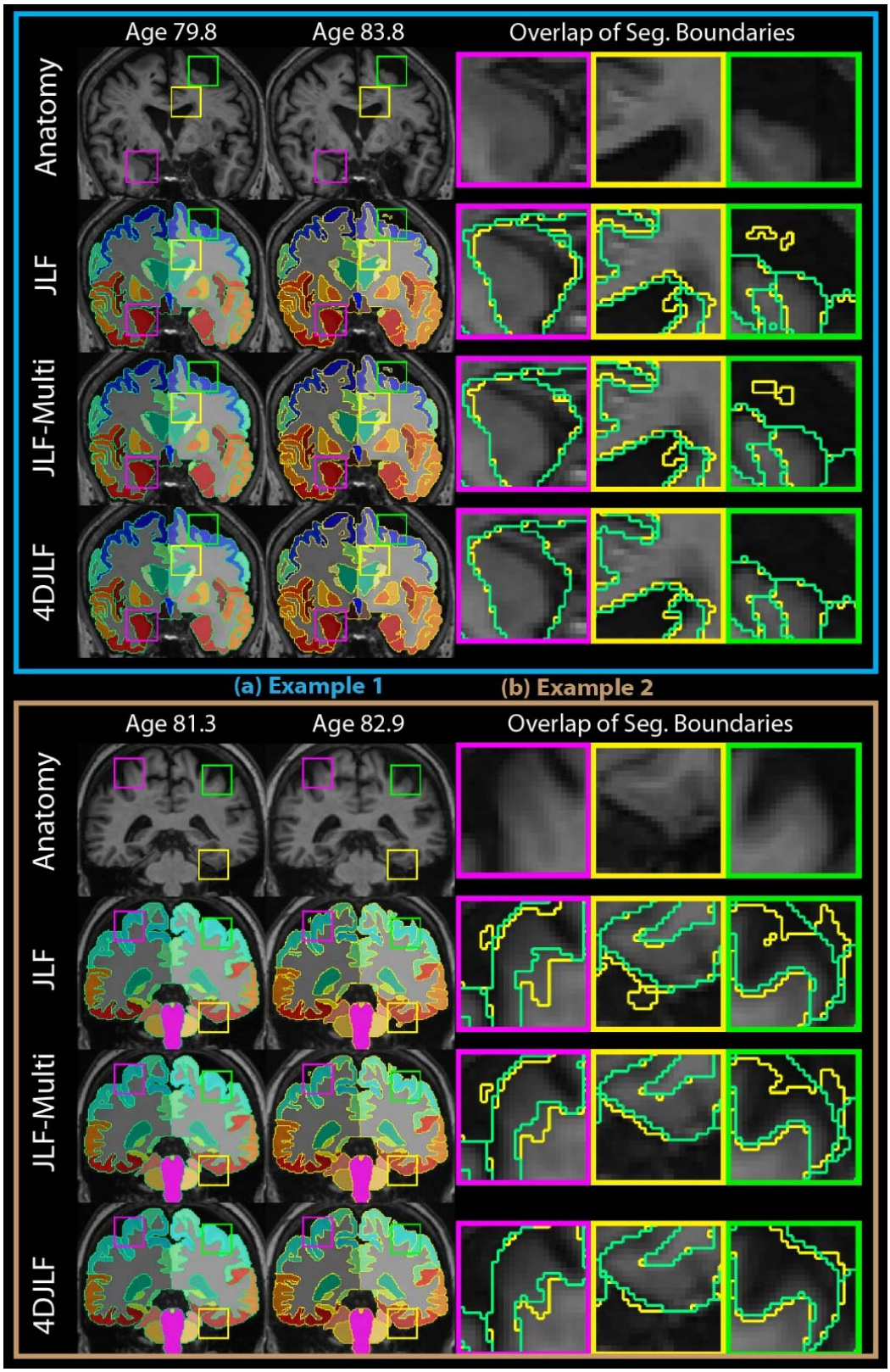


Figure VIII.5 Qualitative results of deploying longitudinal segmentation methods on two examples.

4. Conclusion and Discussion

Herein, we propose the 4DJLF multi-atlas label fusion strategy by modeling the spatial temporal performance of atlases. The proposed 4D theory incorporates the ideas from the two major families of label fusion theories (voting based fusion and statistical fusion) by generalizing the leading JLF label fusion method to a 4D manner. The results demonstrated that the proposed method was not only able to improve the longitudinal reproducibility (Figure VIII.3a, 4 and 5) but also reduces the segmentation errors compared with traditional 3D JLF (Figure VIII.5). Meanwhile, the 4DJLF did not significantly change the segmentation reproducibility when performing on dummy longitudinal pairs of images (Figure VIII.3b). Such result indicated that the 4DJLF was able to keep the sensitivity, while the naïve 4D-Multi was not. All experiments in this paper are able to be run in a modern Linux workstation (e.g. 12 core CPU, 8G memory). For a representative target image (with two other time points available), JLF consumed ≈ 1 hour, 3.7GB RAM using 15 registered atlases; JLF-Multi and 4DJLF consumed ≈ 3 hours, 5GB RAM using 45 registered atlases.

4DJLF demonstrates that temporal covariance matrices can be robustly and efficiently estimated within label fusion, and that these statistical properties can be used to improve MAS. There are multiple opportunities where this approach could be applied and should be investigated: (1) 4DJLF could be used for online consistency where each new volume is fused with 4-D while holding prior segmentation consistent. (2) This work proposed to use the naïve intensity similarity in the probabilistic model to evaluate the temporal performance of atlases. More advanced statistical label fusion model can be integrated to optimize the probabilistic model using maximum likelihood estimation (MLE) or maximum a posteriori probability (MAP). (3) The sparse representation idea could be introduced to the model reduce the computational time. (4) The approach is compatible to the previous efforts in longitudinal segmentation (e.g., 4D registration, 4D intensity normalization) and could be integrated into a full 4D pipeline. (5) The empirical validation is limited since we did not have ready access to manually labeled longitudinal whole brain image with detailed labels. More thorough investigation of longitudinal brain atlases will lead to better understandings of consistency, reproducibility, and accuracy.

Chapter IX. Robust Multi-contrast MRI Spleen Segmentation for Splenomegaly using Multi-atlas Segmentation

1. Introduction

Abnormal enlargement of the spleen, called splenomegaly [1], is a clinical finding in the patients with liver disease [2], cancer [3] and infection [4]. To quantify spleen enlargement, non-invasive spleen volume estimation approaches have been proposed using different imaging modalities (e.g., ultrasound [5-8], computed tomography (CT) [9-12], magnetic resonance imaging (MRI) [13, 14]). Slice-by-slice manual tracing on three-dimensional (3D) spleen volumes has been regarded as the gold standard of in vivo spleen size estimation [14]. However, the manual delineation is resource and time consuming, especially for large cohorts. To alleviate manual efforts and accelerate the spleen volume estimation, many endeavors haven been made. One direction is to replace 3D delineation with less time consuming one-dimensional (1D) manual measurements (e.g., splenic width, length, thickness) [7]. With 1D measurements, the whole spleen volume can be estimated using regression models. Another direction seeks to obtain 3D volumetric spleen segmentation automatically using medical image segmentation approaches [15]. Previous automatic spleen segmentation methods are typically able to be categorized by, but not limited to, shape/contour based models [16], intensity based models [17], graph cuts [18], learning based models [19], and atlas-based methods [20].

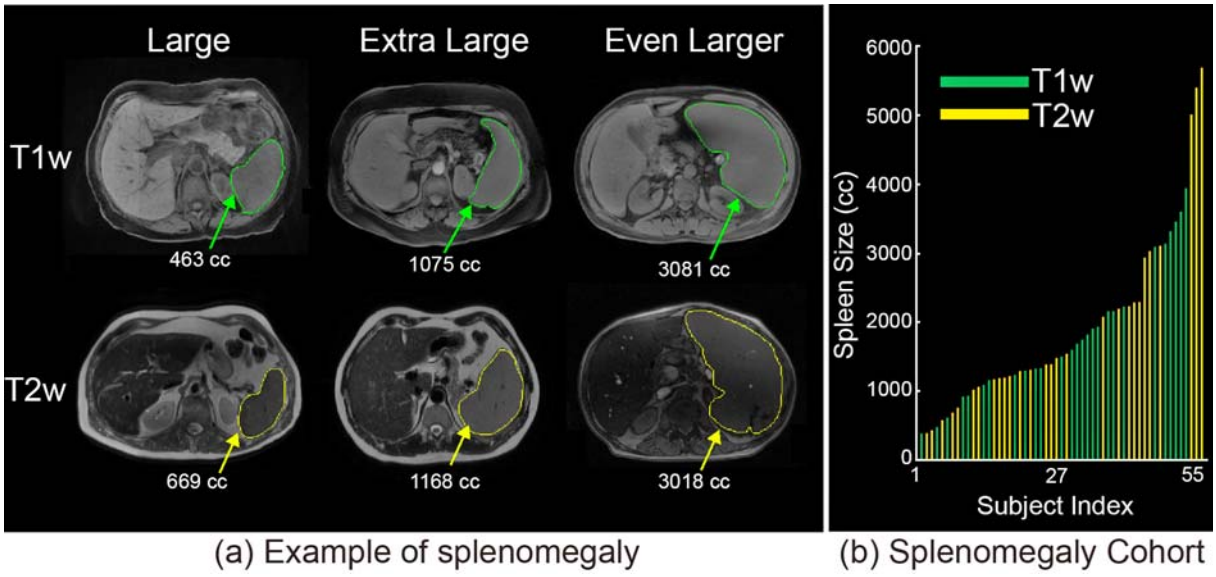


Figure IX.1 (a) presents heterogeneous sequences in clinical acquired abdominal MRI as well as the examples of splenomegaly spleens on MRI. (b) shows the spleen size and sequence type of all 55 MRI.

Most previous spleen segmentation methods were proposed using CT imaging since it has been used as the standard technique in abdominal imaging [7]. One of the essential benefits for medical imaging processing is that the image intensities in CT are the quantitative Hounsfield Unit (HU). The scaled intensity feature are essential in the learning based segmentation methods, such as discriminative models [21] and vantage point forests (V.P. Forests) [22]. In the past decades, MRI has been successfully used in clinical diagnosis and scientific investigations. Compared with CT, MRI eliminates the radiation risk for patients [23, 24], and the frequency of clinical abdominal MRI renders MRI based spleen volume estimation techniques an attractive target. However, the intensities in clinical acquired MRI are heterogeneous (Figure IX.1a) and without absolute scales, such as HU in CT. Therefore, the intensity based segmentation methods developed for CT cannot be directly applied on MRI. Relatively few spleen segmentation methods have been proposed for MRI. Behrad et al. proposed an MRI spleen segmentation method using neural networks and recursive watershed [19]. Farraher et al. achieved accurate spleen segmentation using a semi-automated dual-space clustering segmentation technique [25]. Wu et al. integrated Gabor texture features with snake post-processing for MRI spleen segmentation [26]. Pauly et al. proposed the supervised regression method

to perform the whole body segmentation on the particular MRI Dixon sequences [27]. The multi-atlas segmentation (MAS) method is regarded as state-of-the-art and has been deployed on various scenarios on both CT and MRI [28-35]. Yet, MAS has not been applied to spleen segmentation on clinically acquired splenomegaly MRI.

In this paper, (1) we evaluate the performance of Selective and Iterative Method for Performance Level Estimation (SIMPLE) atlas selection method [36] based on our previous efforts on CT spleen segmentation [31, 32]. (2) For the particular concerns for MRI clinical splenomegaly images, we propose the L-SIMPLE method to achieve the robust spleen segmentation using craniocaudal spleen length (L). To perform the evaluation and validation, 55 clinical acquired MRI volumes were examined, consisting of 28 T1-weighted (T1w) and 27 T2-weighted (T2w) scans (Figure IX.1b), which represented the two major contrast mechanisms in clinically acquired abdominal MRI.

This paper extends a previous conference paper [33] in the following ways. First, a more complete description of the different MAS methods is provided. Second, a graph cut based refinement is created to ensure the topological correctness. Third, more thorough analyses of using craniocaudal spleen length and graph cuts are demonstrated.

2. Methods

2.1. Multi-atlas Segmentation Framework

The general MAS framework consists of preprocessing, image registration, atlas selection, label propagation and multi-atlas label fusion (MLF) [30]. Briefly, first a target image was preprocessed using N4 bias field correction [37] and resampled to 1.5 mm isotropic voxel size using FMRIB's Linear Image Registration Tool (FLIRT) [38]. Second, each atlas image was sequentially affinely registered and non-rigidly registered using DEnsE Displacement Sampling (DEEDS) [39]. Registration accuracy is essential in the atlas based segmentation methods; DEEDS was chosen based on its superior performance in a relevant comparative evaluation [40]. Third, atlases selection is performed to address substantial

registration failures. Finally, MLF was conducted on the selected registered atlases using joint label fusion (JLF) [41]. In this paper, a substantial algorithmic focus is on designing and evaluating atlas selection methods (Figure IX.2).

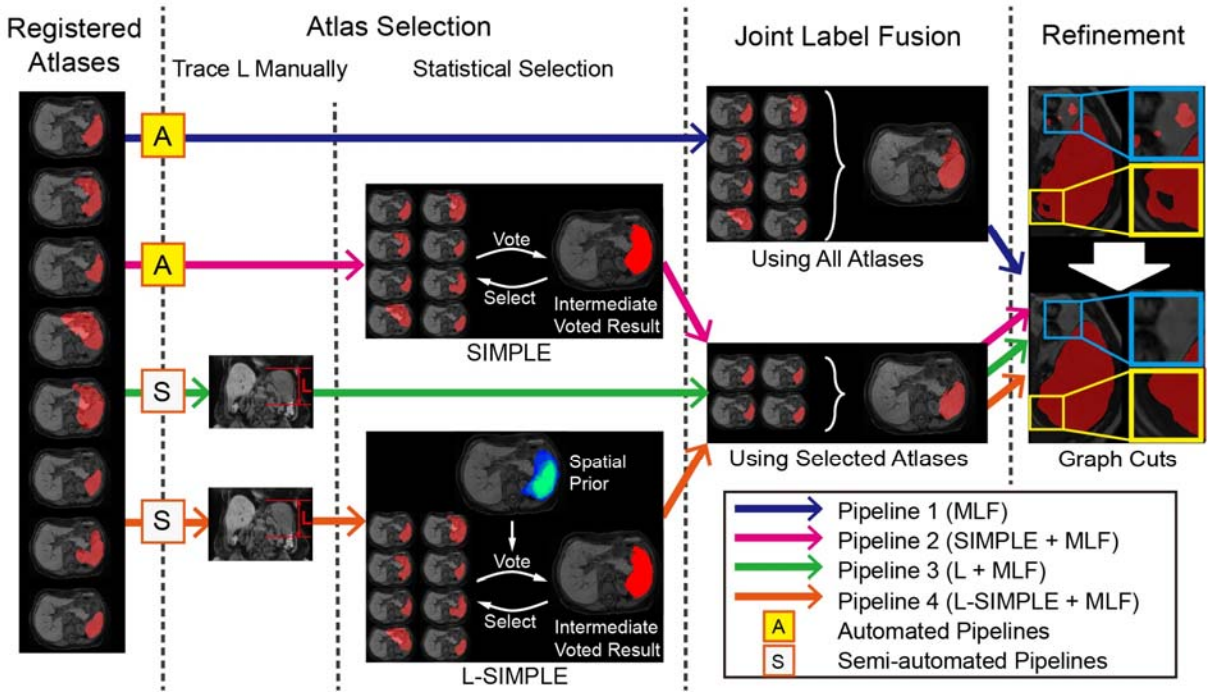


Figure IX.2 Multi-atlas labeling steps for each of the four pipelines. Pipeline 1 conducted multi-atlas label fusion (MLF) on all registered atlases without using atlas selection. Pipeline 2 employed the SIMPLE atlas selection method before performing MLF. Pipeline 3 used the craniocaudal spleen length (L) to guide the atlas selection. Pipeline 4 evaluated the proposed L-SIMPLE method, which integrated the feature L to the SIMPLE atlas selection under the Bayesian framework. For all pipelines, a post refinement procedure was included to ensure the topological correctness of the spleen segmentation (one connected component).

2.2. Automated Pipelines

Two automated pipelines (without manual intervention) were evaluated as shown in Figure IX.2.

Pipeline 1: Pipeline 1 consisted of a naïve strategy that excluded the atlas selection step in the MAS framework (Figure IX.2). Note that registration failures typically occur more frequently in abdominal registrations (Figure IX.3) compared with brain registrations. Therefore, using all registered atlas images might lead to inaccurate label fusion results (Figure IX.3; blue rectangles).

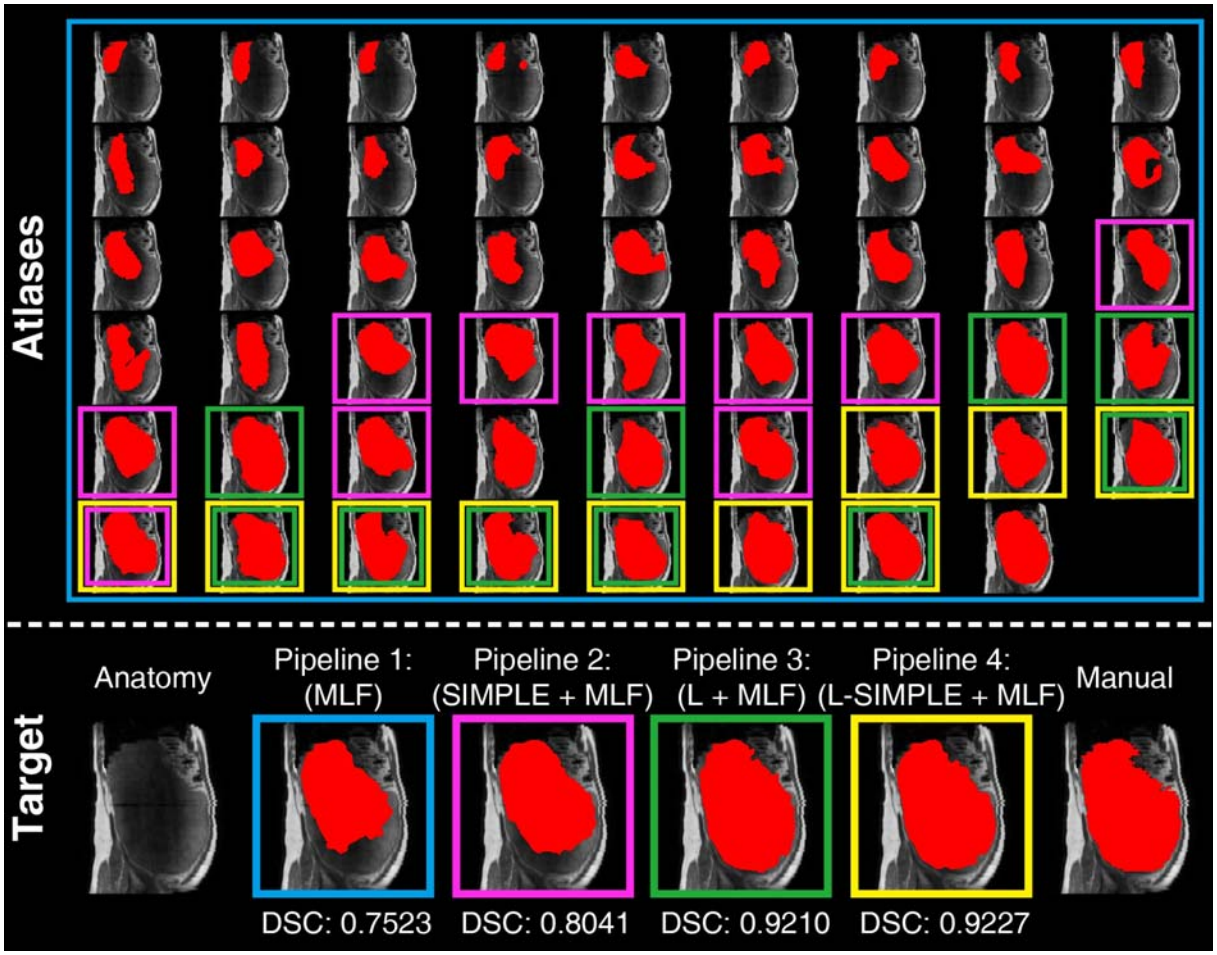


Figure IX.3 This figure presents an example of using different atlas selection strategies. The upper panel reflects the registration results of registering each atlas to the target image. The target image is shown as the left figure on the lower panel. The registered atlases are arranged based on the Dice similarity coefficient (DSC) to the target manual segmentation, whose DSC increased from top left to bottom right. Pipeline 1 (in blue rectangles) employed all registered atlases in the label fusion. Pipeline 2 (in pink rectangles) performed the atlas selection using SIMPLE method. Pipeline 3 (in green rectangles) used the craniocaudal spleen length (L) to guide the atlas selection. Pipeline 4 (in yellow rectangles) integrated L and SIMPLE to the proposed L-SIMPLE method under the Bayesian framework. In this example, Pipeline 4 chose the better atlas candidates (lower rows in upper panel) for the atlas selection, which achieved the highest DSC relative to the manual segmentation.

Pipeline 2: To alleviate registration failures, the Selective and Iterative Method for Performance Level Estimation (SIMPLE) method [36] was used in the atlas selection in Pipeline 2 (Figure IX.2). The SIMPLE method was proposed as a voting based label fusion method. In this work, SIMPLE was used in the similar way as a recent work [31], where SIMPLE has been applied to the atlas selection by iteratively evaluating the Dice similarity coefficient between intermediate segmentation and atlases.

2.3. Semi-automated Pipeline using craniocaudal spleen length

The SIMPLE atlas selection in Pipeline 2 only considered the registered atlas labels in an iterative atlas selection manner without taking the anatomical information from the intensity atlases into account. Therefore, although the SIMPLE method was able to achieve robust performance on most of the cases, it would not be able to select better atlas candidates when multiple registration failures occur in a similar fashion (pink rectangles in Figure IX.3). Therefore, we proposed to use craniocaudal spleen length (L) to guide the atlas selection (Pipeline 3 in Figure IX.2).

Pipeline 3: In clinical diagnosis of splenomegaly, one dimensional (1D) measurements had been used to estimate spleen volume efficiently. Following [32], the 1D craniocaudal spleen length (L) yielded 0.8613 Pearson correlation with ground truth on spleen volume estimation using ≈ 1 minute manual efforts. Therefore, the craniocaudal spleen length was employed in Pipeline 3 to guide the atlas selection. The craniocaudal spleen length was calculated by multiplying slice thickness by the numbers of visible slices on axial direction [7]. The number of visible slices is typically derived manually by experts [7]. In this study, since we had delineated the whole spleen for all volumes, we derived the numbers of visible slices automatically by subtracting the smallest axial slice number from the largest axial slice number that contained the spleen label. Then, atlas selection was deployed by choosing the ten atlases whose craniocaudal spleen length values were the closest to the target image.

2.4. Semi-automated Pipeline using L-SIMPLE

In Pipelines 2 and 3, the SIMPLE and craniocaudal spleen length (L) were used to conduct atlas selection respectively. In this paper, we propose the L-SIMPLE method, which employed the craniocaudal spleen length as a prior information to guide the SIMPLE atlas selection (Pipeline 4 in Figure IX.2).

Pipeline 4: In Pipeline 4, the L-SIMPLE method was proposed to perform the atlas selection by integrating the craniocaudal spleen length (L) with the SIMPLE approach under a Bayesian framework. A probabilistic map was obtained by averaging the ten registered spleen labels, whose craniocaudal spleen lengths were the closest to the target image. Then the probabilistic map served as a prior in L-SIMPLE to

guide the iterative atlas selection. The inputs of L-SIMPE were (1) The craniocaudal spleen lengths of the target image, and (2) registered spleen label atlases $\mathbf{A} = \{A_1, A_2, \dots, A_M\}$, where each A_j represented the jth label atlas in total M available atlases. The outputs of L-SIMPLE were N selected atlases $\mathbf{A}'$ for the following multi-atlas label fusion ($N \leq M$). The complete L-SIMPLE algorithm was:

Step 1) The $\mathbf{A}$ were used as all atlases initially. The spleen spatial prior $p(T)$ was obtained by averaging the r registered label atlases, whose craniocaudal spleen length had the smallest differences compared with target image's craniocaudal spleen length. $p(T = 1)$ was the probability prior map of the spleen (spleen label was 1), while $p(T = 0)$ was the probability prior map of non-spleen tissues as well as background.

Step 2) The iterative atlas selection strategy was performed. $\mathbf{A}^k$ represented the set of the remaining n^k atlases at iteration k. For each voxel i, the likelihood function of spleen was defined by

$$\begin{aligned} f\left(\mathbf{A}_i^k \middle| T_i = 1\right) &= \frac{1}{n} \sum_{j=1,2,\dots,n} A_{ji}^k \\ f\left(\mathbf{A}_i^k \middle| T_i = 0\right) &= 1 - \frac{1}{n} \sum_{j=1,2,\dots,n} A_{ji}^k \end{aligned} \tag{9.1}$$

Step 3) Using the prior in step 1 and likelihood function in step 2, the Bayesian posterior probability of spleen at voxel i was derived as

$$f\left(T_i = 1 \middle| \mathbf{A}_i^k\right) = \frac{f\left(\mathbf{A}_i^k \middle| T_i = 1\right) f(T_i = 1)}{f\left(\mathbf{A}_i^k\right)} \tag{9.2}$$

$$f\left(T_i = 0 \middle| \mathbf{A}_i^k\right) = \frac{f(\mathbf{A}_i^k | T_i = 0) f(T_i = 0)}{f(\mathbf{A}_i^k)}$$

Step 4) The intermediate spleen segmentation S at voxel i was obtained by

$$\begin{aligned} S_i &= 1, \quad \text{if } f\left(T_i = 1 \middle| \mathbf{A}_i^k\right) \geq f\left(T_i = 0 \middle| \mathbf{A}_i^k\right) \\ &= 0, \quad \text{if } f\left(T_i = 1 \middle| \mathbf{A}_i^k\right) < f\left(T_i = 0 \middle| \mathbf{A}_i^k\right) \end{aligned} \tag{9.3}$$

Step 5) The one-dimensional weight vector w was defined by the Dice similarity coefficient (DSC) between each A_j^k and S.

$$w_j = \mathrm{DSC}(A_j^k, S) \tag{9.4}$$

Step 6) For the k+1 iteration, the $\mathbf{A}^{k+1}$ was a subset of $\mathbf{A}^k$ by comparing w_j with mean ($\overline{w}$) and standard deviation (σ_w) of w.

$$\mathbf{A}^{k+1} = \{A_j^k\}, \quad \text{for} \quad j: \quad w_j > (\overline{w} - \sigma_w) \tag{9.5}$$

Step 7) If the n^{k+1} (size of $\mathbf{A}^{k+1}$) was less than the minimum number of atlases N (herein, 10) or $n^{k+1} = n^k$, the L-SIMPLE was terminated and $\mathbf{A}^k$ was returned as selected atlases. Otherwise, the method performed another iteration at step 2.

2.5. Refinement Using Graph Cuts

Since the MAS segmentation was conducted based on voxel wise voting, spleen topology (one connected component) was not guaranteed. Therefore, a post processing step using graph cuts was used to ensure the topological correctness of MAS spleen segmentation. The graph cuts method proposed in [31] was used in this work, which maximized the Markov random field (MRF) based energy function [42, 43].

3. Data

A clinical cohort containing 55 abdominal MRI volumes was acquired from 26 patients with splenomegaly. Eight patients were scanned one time, seven patients were scanned twice, while eleven patients were scanned three times. This cohort has two major features. First, the cohort was a multi-contrast dataset, which consists of 27 T1w and 28 T2w images. This dataset was used to evaluate the performance of the proposed methods on clinically acquired multi-contrast MRI images. Second, the cohort had large variations on spleen volume size for splenomegaly, varying from 368 cubic centimeter (cc) to 5670 cc. The mean spleen volume was 1881 cc while the standard deviation was 1219 cc.

The leave-one-subject-out strategy was employed for the empirical validation, which means that the 55 MRI image volumes were used as either atlases or target images in each leave-one-subject-out test. To achieve the 3D whole spleen labels on atlases, the manual delineation was obtained on every volume by an experienced rater. The whole spleen segmentation for each scan was traced slice-by-slice (axially).

4. Experiments and Results

The Wilcoxon signed rank test [44] was used for statistical analyses. All statements of statistical significance are made using the Wilcoxon signed rank test for $p<0.05$.

4.1. Validation the Rationale of Using L

4.1.1. Experiments

Fifty-five clinical scans were used to evaluate the rationale of using craniocaudal spleen length in atlas selection. We consecutively performed affine and rigid registration using DEEDS registration method [39] on all possible combinations between 55 image volumes. (1) Each image was used as a target image. (2) All the other available images except the target image's longitudinal scans were employed as moving images, which were then registered to the target image. This strategy was called "leave-one-subject-out", which means the longitudinal scans (three at maximum) for every target image were excluded from the atlases. Therefore, 52 to 54 atlases were used for each target image. (3) The affine transformation and non-rigid transformation field were applied on the spleen labels of source images. (4) The DSC values were calculated between source images and target. Finally, affine and non-rigid registrations were performed on 2890 pairs of source and target 3D volumes using 55 scans.

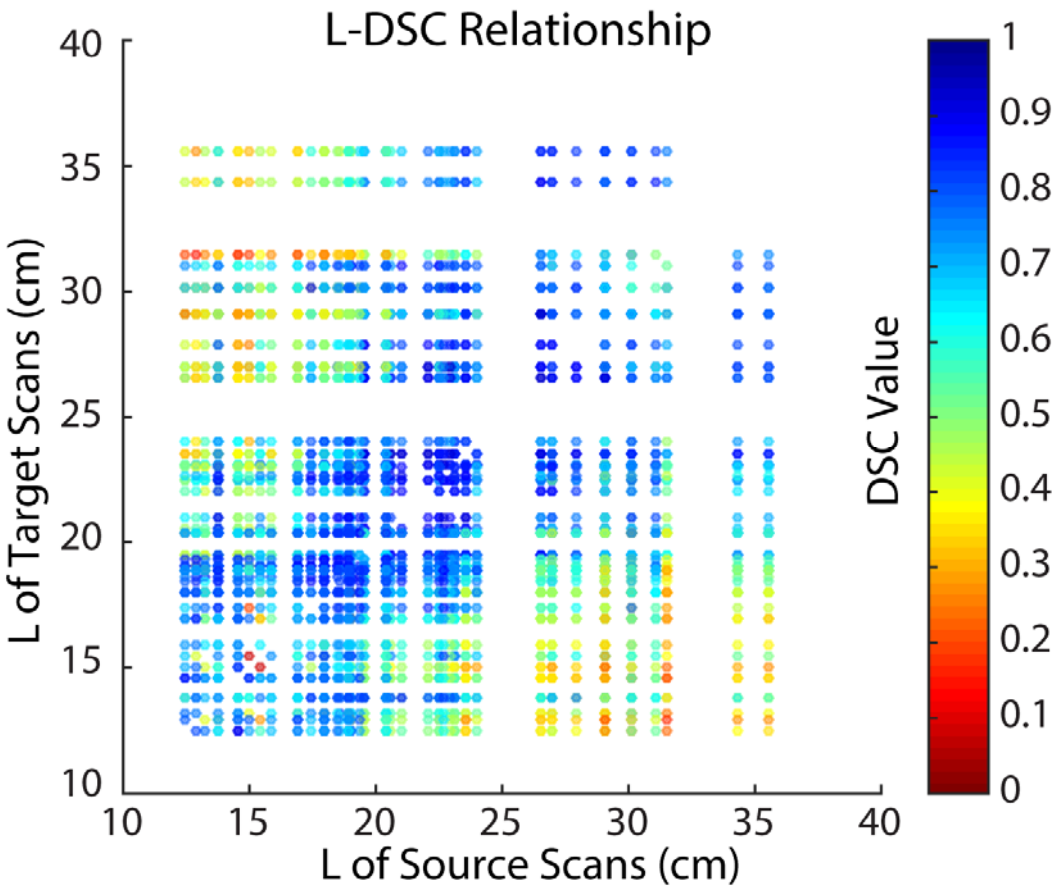


Figure IX.4 The scatter plot demonstrated that 2890 registrations have been performed on all possible combinations between 55 clinical acquired splenomegaly images. The coordinate of each dot corresponded to the craniocaudal spleen length (L) of the source and target scan of the registration. The color of each dot indicated the DSC value between the registered spleen label and the manual segmentation.

4.1.2. Results

The registrations were conducted on 2890 pairs of scans. In each pair, the craniocaudal spleen length of source and target scan were used as x and y coordinates in the Figure IX.4. The color of each dot indicated the DSC value between the registered source spleen label and target spleen label. From the scatter plot, the registrations between scans with similar craniocaudal spleen length typically achieved better performance on DSC.

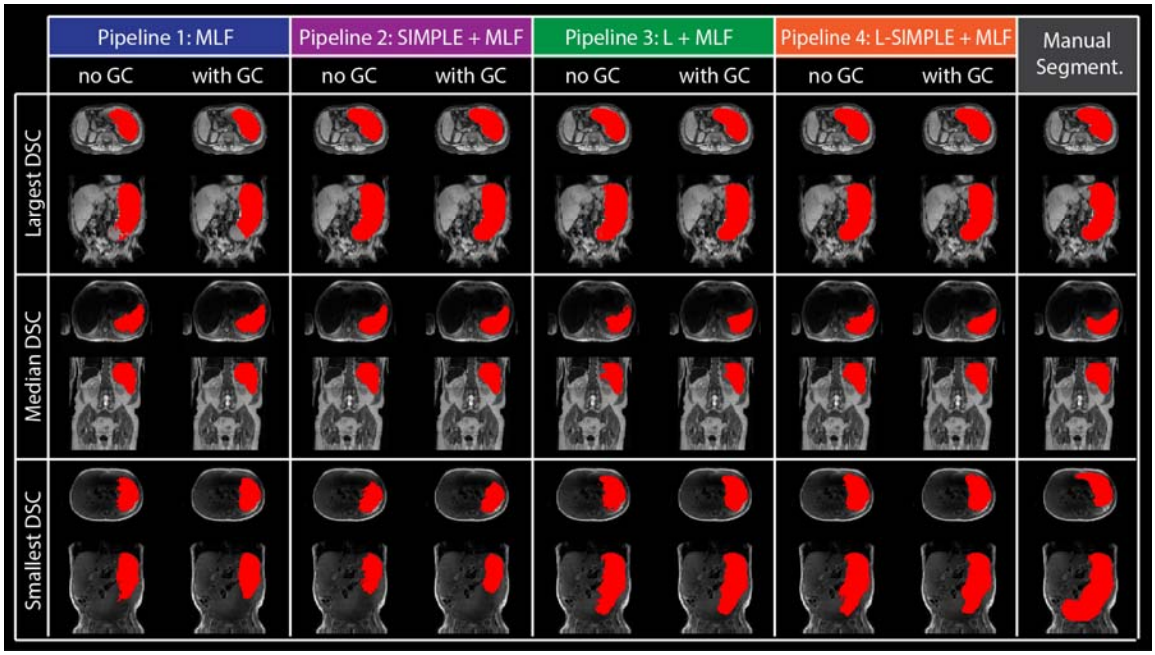


Figure IX.5 The qualitative results of four pipelines on the three subjects with largest, median and smallest DSC of Pipeline 4 with GC were shown with manual segmentation. For each pipeline, the "no GC" indicated the results without Graph Cuts while the "with GC" demonstrated the results with Graph Cuts.

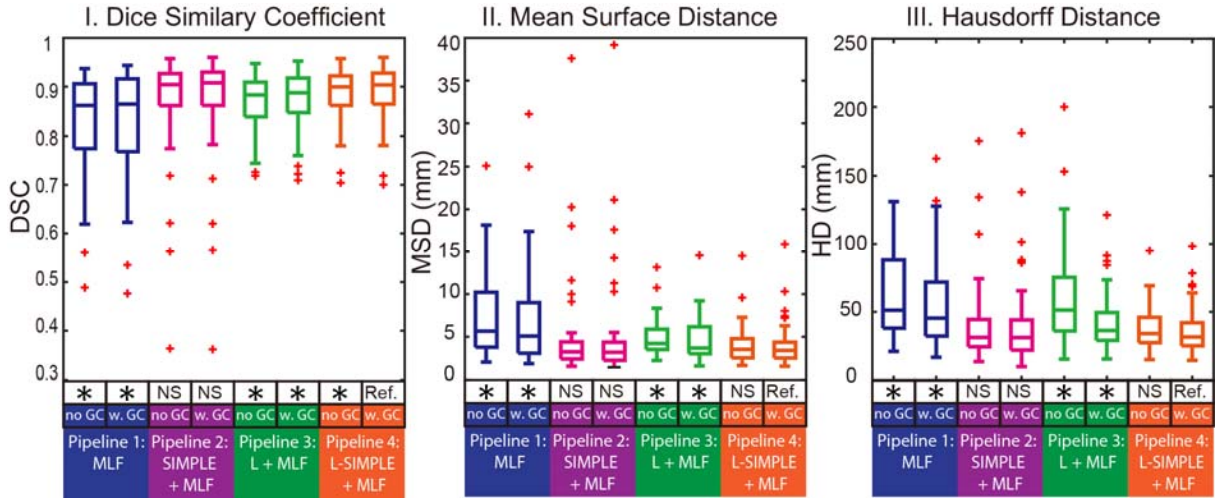


Figure IX.6 The quantitative results of four pipelines on Dice similarity coefficient (DSC), mean surface distance (MSD) as well as Hausdorff distance (HD) are shown in boxplots. The "no GC" indicated the results without Graph Cuts while the "w. GC" demonstrated the results with Graph Cuts. The statistical analyses were conducted between the proposed Pipeline 4 L-SIMPLE with Graph Cuts method (marked as reference "Ref.") with other approaches. Statistically significant, differences are marked with a "*" symbol. Non-significant differences are indicated with "N.S."

Table IX-1 Performance of Four Pipelines using All 55 Volumes in A Leave-one-subject-out approach.

Measurements		Automated					Semi-automated			
		V.P. Forests [25]	Pipeline 1		Pipeline 2		Pipeline 3		Pipeline 4	
			No GC	With GC	No GC	With GC	No GC	With GC	No GC	With GC
Dice similarity (DSC)	median	0.697	0.861	0.864	0.905	**0.908**	0.883	0.888	0.900	0.904
	mean±std	0.70 ±0.12	0.82 ±0.11	0.83 ±0.11	0.87 ±0.10	0.88 ±0.10	0.87 ±0.06	0.87 ±0.06	0.88 ±0.06	**0.89 ±0.06**
Mean surface distance (mm)	median	21.42	5.68	5.09	3.24	**3.19**	4.21	3.67	3.54	3.41
	mean±std	22.69 ±8.29	7.23 ±4.80	6.93 ±5.74	4.75 ±5.73	4.83 ±6.04	4.86 ±2.13	4.52 ±2.41	**3.96 ±2.17**	3.97 ±2.45
Hausdorff distance (mm)	median	123.64	51.19	45.44	31.29	**31.25**	51.31	36.45	34.10	31.65
	mean±std	135.2 ±48.8	61.4 ±29.6	53.4 ±31.3	39.7 ±28.4	39.6 ±30.9	61.8 ±35.6	42.6 ±20.9	38.3 ±15.8	**37.1 ±18.2**
	DSC<0.8	45	19	15	6	**5**	8	8	7	6

4.2. Validation on Four Pipelines

4.2.1. Experiments

The same 55 scans were used in the leave-one-subject-out validations on the four different pipelines respectively. The selection of atlases and target images was the same as section IV.A "Validation the Rationale of Using L". In these experiments, Pipeline 1 to 4 were deployed as atlas selection and label

fusion as Figure IX.2.

We also compared our pipelines with a recent learning based method called vantage point forest (V. P. Forests) [22]. The code was downloaded from the link in that paper. All the parameters were set to the default except the "num_labels". In this study, we set num_labels = 1 since we only had one spleen label.

4.2.2. Results

The qualitative results of four pipelines are demonstrated in Figure IX.5. The qualitative results of comparing the proposed Pipeline 4 with other method had been shown in Figure IX.6 and Table IX-1. The performance of graph cuts using DSC is significantly higher than without graph cuts refinement.

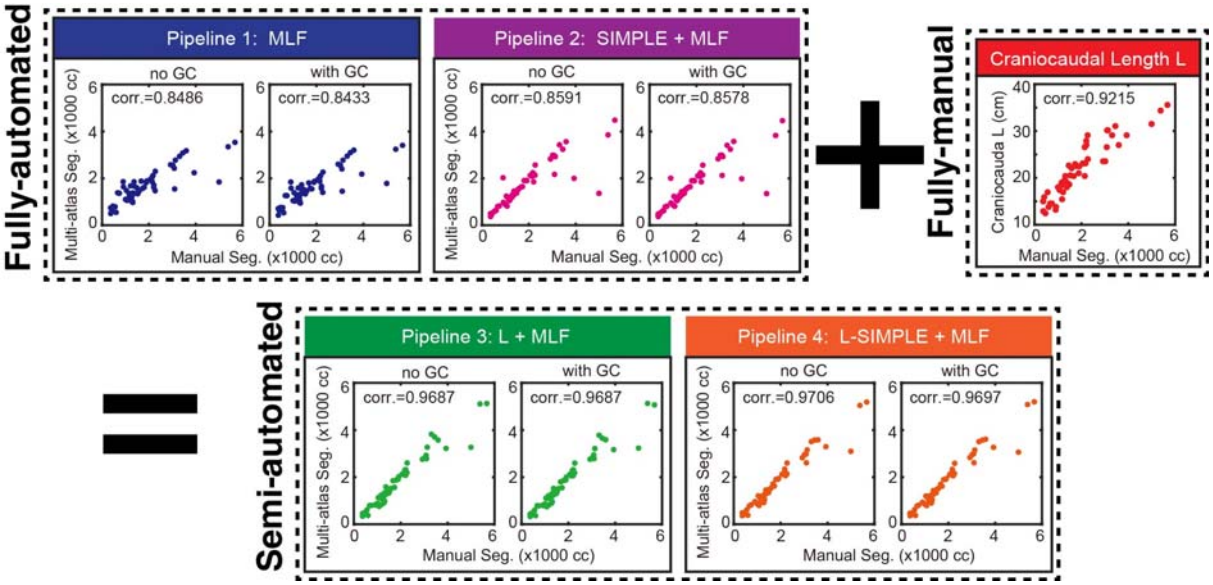


Figure IX.7 The correlation analyses between different pipelines with manual segmentation. The semi-automated pipelines achieved higher Pearson correlation values than fully-automated pipelines and fully-manual L measurements. The "+" and "=" indicated that the Pipeline 3 and 4 integrated the information derived from Pipeline 1 and 2 plus the craniocaudal spleen length (L). The "corr." reflected the Pearson correlation values. The "no GC" indicated the results without Graph Cuts while the "with GC" demonstrated the results with Graph Cuts.

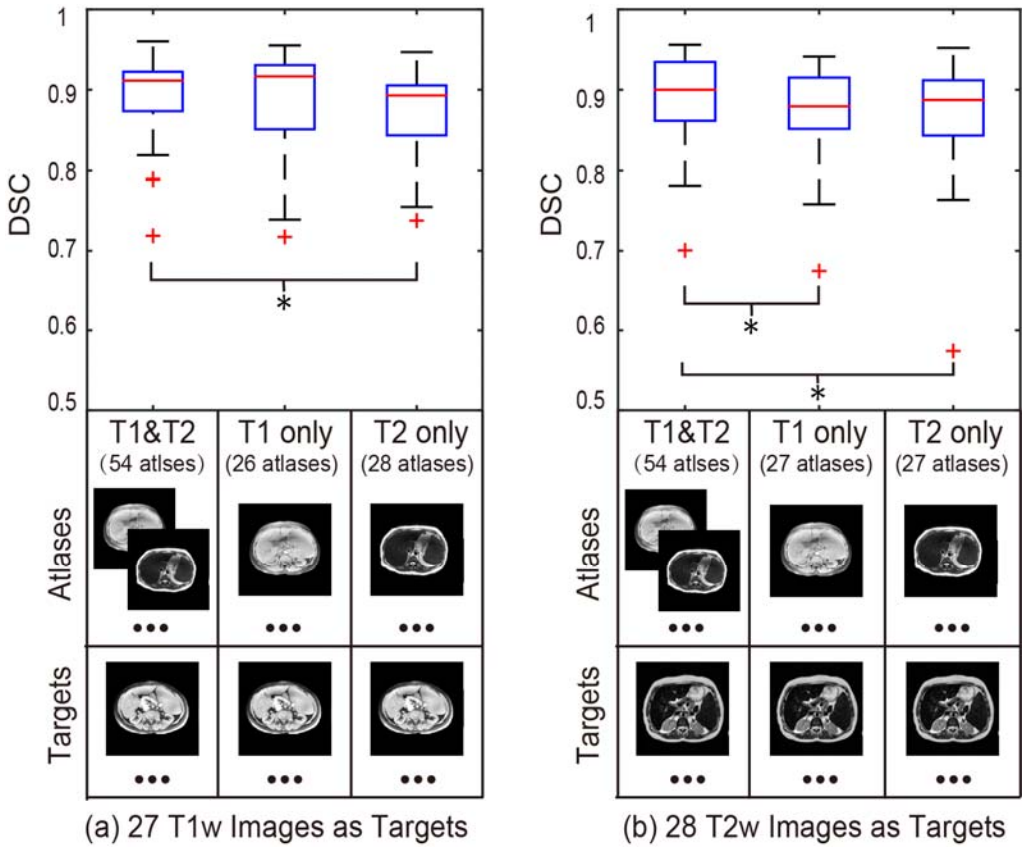


Figure IX.8 The sensistivty analyses of the proposed L-SIMPLE method on multi-contrast images. (a) demonstrates that using both T1w and T2w images as atlases achieved better performance than only using T1w or T2w atlases on segmenting T1w images. (b) shows that using both T1w and T2w images as atlases achieved better performance than only using T1w or T2w atlases. From (a) and (b), it is evident that the performance of using the same sequence on both atlases and targets did not yield a significant difference on DSC compared with using the different sequences for atlases and targets respectively. The "*" symbol indicates significant differences.

4.3. Sensitivity Analyses on Multi-Contrast Scenarios

4.3.1. Experiments

The multi-contrast images (e.g., T1w and T2w) in clinical acquired images were heterogeneous on both absolute intensity and intensity contrast. In this experiment, we explored the robustness of the MAS methods on the multi-contrast images. Moreover, we evaluate the performance of using (1) both T1w images as atlases and targets, (2) both T2w images as atlases and targets, (3) T1w images as atlases and T2w images as targets, and (4) T2w images as atlases and T1w images as targets.

4.3.2. Results

For T1w target images, using both T1w and T2w atlases achieved significantly higher DSC than using all T1w or T2w atlases. For T2w target images, using both T1w and T2w atlases achieved significantly higher DSC than using all T1w or T2w atlases. No significant differences were detected.

5. Discussion

Fully automated segmentation methods are commonly preferred over manual or semi-manual segmentation methods. Therefore, we evaluate the fully-automated Pipeline 1 and Pipeline 2. The results demonstrated that the Pipeline 2 was able to achieve 0.9 median DSC on spleen segmentation for splenomegaly. However, outliers (e.g., bad segmentations with DSC<0.7) were generated from the registration failures. Such poor cases were typically not desired in the clinical scenarios. To alleviate such failures, the 1D manual measurement L was introduced to form the Pipeline 3 and Pipeline 4. From the validations, the Pipeline 4 achieved more robust segmentations (Pearson correlation > 0.97) without sacrificing on segmentation accuracy (DSC>0.9) compared with Pipeline 2. Meanwhile, the number of worst cases (DSC<0.8, DSC<0.75 and DSC<0.7) were alleviated when introducing the L. Since manual efforts were still required in Pipeline 4, a meaningful future work would be automated craniocaudal spleen length estimation using machine learning and artificial intelligence.

In this work, four atlas selection strategies (none, automated, manual, semi-automated) have been evaluated. Other atlas selection methods could be used to further leverage the performance of the atlas based spleen segmentation. Craniocaudal length L can be used for spleen volume estimation directly using regression models (with 0.816 correlation to the true volume reported in [10]). The proposed pipelines not only achieved higher correlation scores but also provided the 3D volumetric segmentations that the regression was not able to. The computational time of registering one atlas to target image was typically <5 min in our experiments. The computational time would be further reduced when performing atlas selection (e.g., using the information from spleen length L). Another direction worth pursing using the spleen length L and its spatial information to initialize or leverage the image registration. In the future, the publicly

available dataset from VISCERAL Anatomy3 challenge could be used to evaluate the proposed method or new methods on abdominal organ segmentation [45].

6. Conclusion

In this paper, we have proposed the L-SIMPLE method and evaluated the performance of multi-atlas segmentation on clinical acquired MRI for splenomegaly patients. The rationale of introducing the manual measurement L was illustrated in Figure IX.4. Figure IX.5 and Figure IX.6 demonstrated that the fully automated Pipeline 2 (SIMPLE+MLF) and semi-automated Pipeline 4 (L-SIMPLE + MLF) both achieved DSC>0.9. By using the feature L, Pipeline 4 achieved 0.97 Pearson correlation with the manual segmentation (in Figure IX.7 and Table IX-1), which was better than either fully automated pipelines or only using the spleen length L. The performance of all the four pipelines were better than the V. P. Forests method, which shown the robustness of the proposed methods on the multi-contrast MRI segmentation. By using the prior from the manually traced L, the worst cases of the spleen volume estimations were alleviated as shown in Figure IX.6 and Table IX-1. The number of worst cases (DSC<0.8, DSC<0.75 and DSC<0.7) for the Pipeline 3 and 4 were less than Pipeline 1 and 2. Although the improvements on DSC, MSD, HD using the graph cuts refinement were not large compared with omitting refinement, the graph cut ensures the topological correctness of the final spleen segmentation (one connected component).

Figure IX.8 evaluated the sensitivity of the proposed method on multi-contrast scenarios. The results demonstrated that the proposed method yields consistent segmentation performance even if the contrast mechanism of atlases and targets are different (T1w and T2w). Meanwhile, using all available atlases, the performance of the segmentation was better than to pre-classify them to T1w atlases or T2w atlases.

Chapter X. Splenomegaly Segmentation using Global Convolutional Kernels and Conditional Generative Adversarial Networks

1. Introduction

Spleen volume estimation is essential in detecting splenomegaly (abnormal enlargement of the spleen), which is a clinical biomarker for spleen and liver diseases [35, 134]. Manual tracing on medical images has been regarded as gold standard of spleen volume estimation. To replace the tedious and time consuming manual delineation, many previous works have been proposed to perform automatic spleen segmentation on ultrasound [36-38], computed tomography (CT) [39-41, 325, 334] or magnetic resonance imaging (MRI) [135, 136, 138, 335]. In recent years, deep learning methods have shown their advantages on automatic spleen segmentation compared with traditional medical image processing methods [137]. However, the existing deep learning methods are typically deployed on CT images collected from healthy populations (e.g., spleen size < 500 cubic centimeter (cc)). When dealing with splenomegaly MRI segmentation (e.g., spleen size > 500 cc), we need to overcome two major challenges: (1) the large inhomogeneity on intensities of clinical acquired MR images (e.g., T1 weighted (T1w), T2 weighted (T2w) etc.), and (2) the large variations on shape and size of spleen for splenomegaly patients [138]. Recently, global convolutional network (GCN) have shown advantages in sematic segmentation on natural images with large variations by using larger convolutional kernels [139]. Meanwhile, adversarial networks have proven able to refine the semantic segmentation results [140].

In this paper, we propose a new Splenomegaly Segmentation Network (SSNet) to perform the splenomegaly MRI segmentation under the image-to-image framework with the end-to-end training. In SSNet, the GCN is used as the generator while the conditional adversarial network (cGAN) is employed as the discriminator [336]. To evaluate the performance of SSNet, the widely validated Unet [337] and GCN were employed as benchmark methods. Sixty clinical acquired MRI scans (32 T1w and 28 T2w) were used as the experimental cohort to test the robustness of the proposed SSNet on the multi-contrast scenario. The

experimental results demonstrated that the SSNet achieved more accurate and more robust segmentation performance compared with benchmark methods.

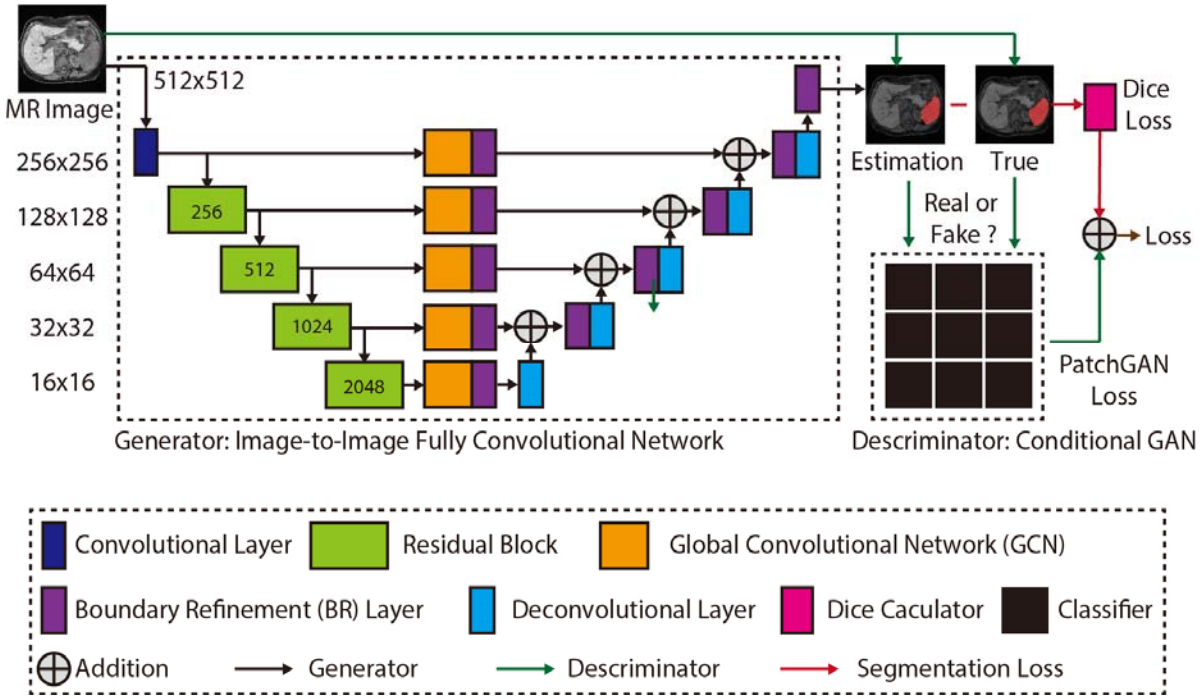


Figure X.1 The proposed network structure of the Splenomegaly Segmentation Net (SSNet). The number of channels of each encoder is shown in the green boxes, while the number of channels of each decoder is two. The image (or feature map) resolution for each level is shown on the left side of this figure.

2. Methods

The SSNet was designed under the GAN framework, which consisted of both a generator and discriminator (Figure X.1). In this section, we introduce each component in the SSNet.

2.1. Generator of SSNet

The GCN was employed as the generator in SSNet for the image-to-image segmentation, where the input and output images had the same resolution 512×512. Each training image was sent to a convolutional layer (kernel size = 1, channels = 64, stride = 2, padding = 3). Then, the "encoder" portion (left side of GCN) extracted the feature maps from the convolutional layer using four hierarchical residual blocks, which were the same as the ResNet [338]. Then, five GCN units [139] were used to transfer the

feature maps for each layer to two channels using the large convolutional kernels. The equivalent kernel size was the resolution of the feature map by assembling two 1D orthogonal kernels [139]. The new feature maps with large reception field were further sent to the boundary refinement layer that is defined in [139]. Next, the refined feature maps were added to the up-sampled feature maps from the "decoder" portion (right side of GCN). Finally, the added maps were further refined by boundary refinement layer and deconvolved to the final segmentations. In Figure X.1, the number of channels of each encoder was shown in the green boxes, while the number of channels of each decoder was two. The image resolution for each level was shown on the left side of Figure X.1.

2.2. Discriminator of SSNet

In SSNet, the conditional GAN (cGAN) was used to further refine the segmentation results in the end-to-end training[336]. Briefly, estimated segmentation, manual segmentation and input images were used under the conditional manner. For the true segmentation, the ground truth for the cGAN was "true." For the segmentation from the generator, the ground truth for the cGAN was "false." The PatchGAN [336] was used as the classifier for the cGAN, which was a compromise solution between classifying the whole image and classifying each pixel.

2.3. Loss Function and Optimization

The loss function of SSNet was defined as $\mathrm{Loss}_{\mathrm{SSNet}}$ in the following equation.

$$\mathrm{Loss}_{\mathrm{SSNet}} = \mathrm{Loss}_{\mathrm{Dice}} + \lambda \cdot \mathrm{Loss}_{\mathrm{GAN}} \tag{10.1}$$

$\mathrm{Loss}_{\mathrm{Dice}}$ represents the Dice loss, which was the negative Dice similarity coefficient (DSC) score between the segmentation from the generator and the manual segmentation. The $\mathrm{Loss}_{\mathrm{GAN}}$ indicated the GAN loss, which was the binary cross entropy (BCE) loss between the cGAN estimations and true classes. The λ was a constant value that decided the weights when adding the two losses. In our study, the λ was empirically set to 100. The Adam optimization [339] was used as the optimization function (learning rate = 0.00001).

3. Experiments

3.1. Data

We used 60 clinically acquired abdominal MRI scans (32 T1w / 28 T2w) from splenomegaly patients to evaluate the performance of different deep convolutional networks. Images were acquired after informed consent and the study was monitored by an approved institutional review board. The data accessed in this study was de-identified. Among the entire cohort, 45 scans (24 T1w / 21 T2w) were used as training data, while the remaining 15 scans (8 T1w / 7 T2w) were employed as independent validation data. For each scan, the MRI volume was resampled to $512 \times 512 \times 512$ resolution to obtain 512 axial, 512 coronal as well as 512 sagittal 2D images. The manual segmentations of spleens were traced by an experienced rater using the Medical Image Processing Analysis and Visualization (MIPAV) software [11]. From the manual segmentations, the minimum size of spleen is 368 cubic centimeter (cc), the maximum size is 5670 cc, the mean spleen volume is 1881 cc, and the standard deviation is 1219 cc.

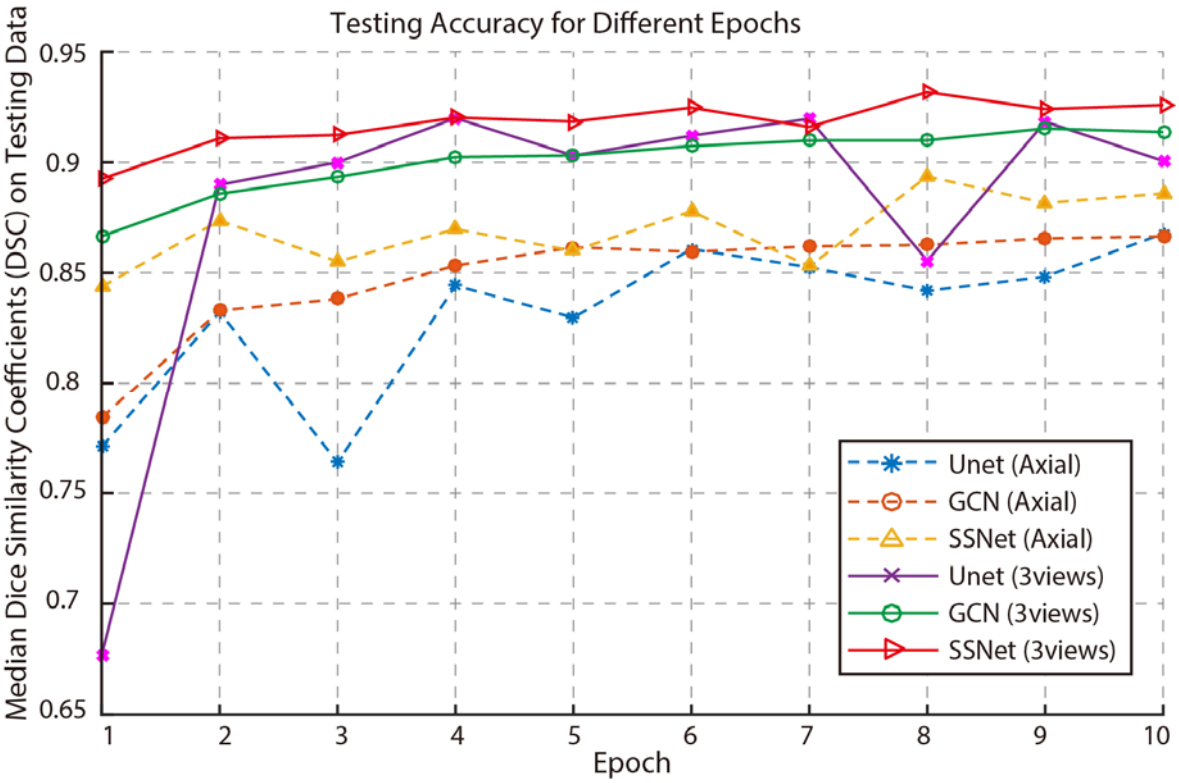


Figure X.2 The testing accuracy of different epochs was shown in this figure. The y axial indicated the mean Dice similarity coefficients (DSC) on all testing volumes, while the x axial presented the epoch number from one to ten. The dashed curves were the testing accuracy for the case that only axial images were used as training and testing images. The solid curves were the testing accuracy for the case that all axial, coronal and sagittal view images were used in both training and testing scenario.

3.2. Experiments

Two sets of the experiments were performed to compare the performance of the proposed SSNet with Unet and GCN benchmarks. Since it was a 2D segmentation problem, we used the ImageNet [340] pertained model as the initialization for each network when the pertained model was available. The first set of the experiments only used the axial images as both training and testing images. Then, the 3D volumetric spleen segmentations were derived by assembling the testing images slice by slice from the same testing scan. For the second set, all axial, coronal and sagittal view 2D images from the 45 resampled training scans were used to train three networks: (1) the first network (axial view network) was trained by all axial view images, (2) the second network (coronal view network) was trained by all coronal view images, and (3) the third network (sagittal view network) was trained by all sagittal view images. In the testing procedure, the

15 independent testing scans were used for an external validation. For each resampled testing scan, all axial view 2D images were segmented by the axial view network and then concatenated to a 3D segmentation. Similarly, 3D segmentations from coronal and sagittal views for such testing scan were obtained from the coronal view network and the sagittal view network. Finally, the three 3D segmentations (for each testing scan) were fused to one final segmentation by (1) merging three segmentations from different views to a single segmentation using "union" operation, (2) performing open morphological operations to smooth the boundaries, and (3) performing close morphological operations to fill the holes.

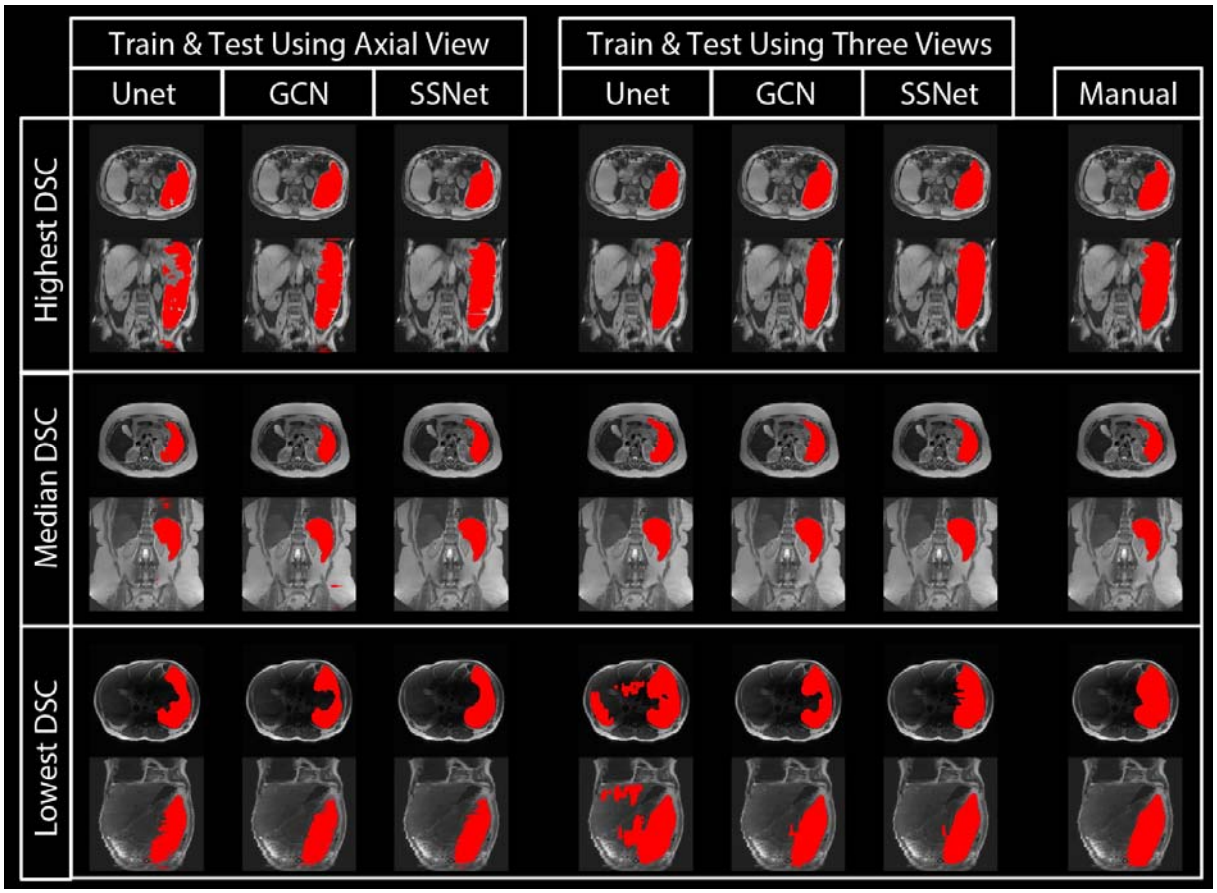


Figure X.3 The qualitative results of different methods. The segmentation results of Unet, GCN and SSNet on using (1) only axial 2D images, and (2) all axial, coronal and sagittal 2D images are shown in the figure for different columns. The manual segmentation results for the same subjects are presented as well. The results of three subjects were selected from the highest, median and lowest DSC from the SSNet's testing data.

3.3. Validation Metrics

The Dice similarity coefficient (DSC) values relative to the manual segmentation were used as the metrics to evaluate the performance of different segmentation methods. All statistical significance tests were made using a Wilcoxon signed rank test ($p<0.01$).

4. Results

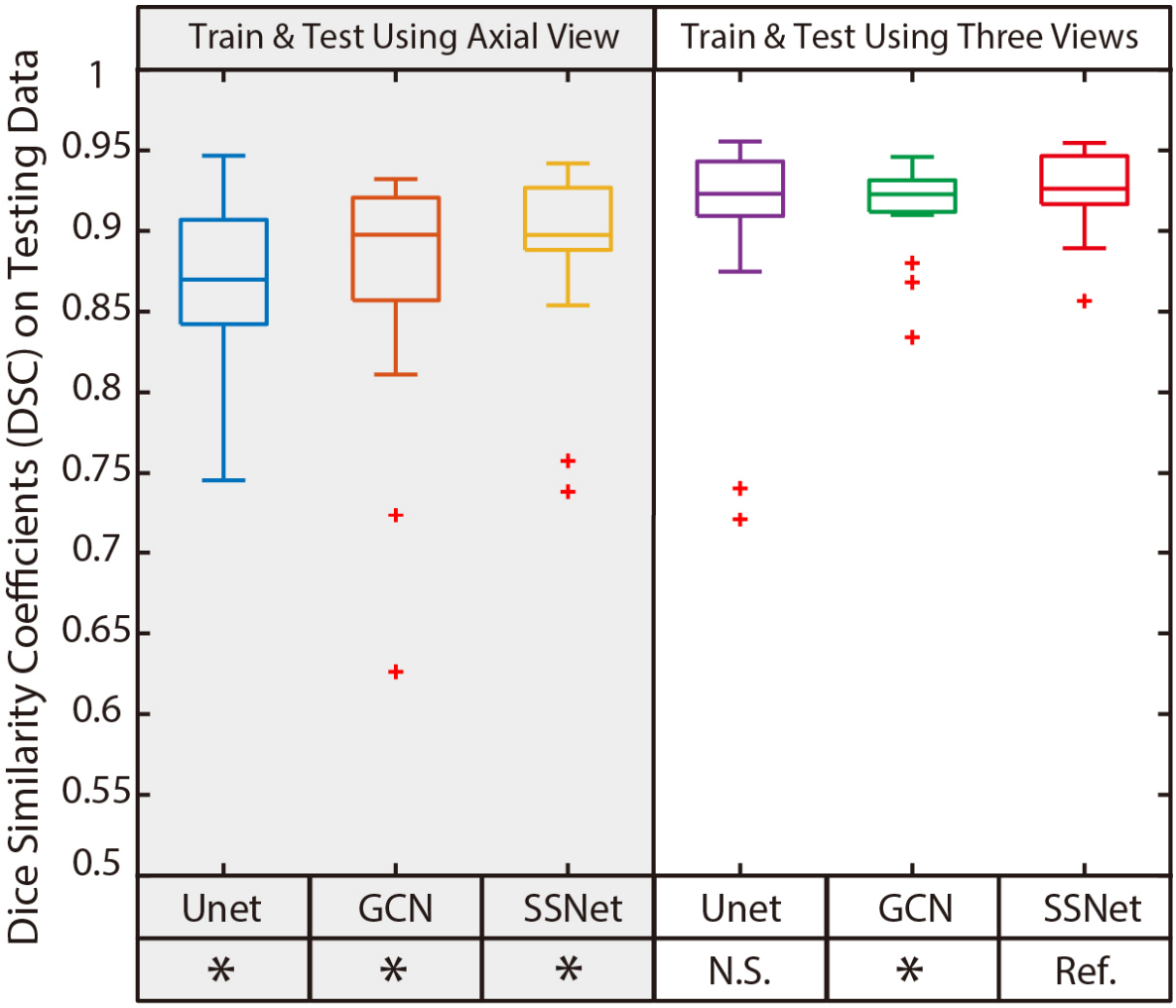


Figure X.4 The quantitative results of different methods. The box plots in left panel indicate the results of using only axial view images, while the right panel presents the results of using all axial, coronal and sagittal images as in both training and testing. The Wilcoxon signed rank tests were employed as statistical analyses, where "Ref." indicates the reference method. The "*" indicates the p<0.01 while the "NS" means not significant.

Figure X.2 presents the testing accuracy of different methods and experimental strategies as median DSC curves for ten epochs. The y axial indicated the mean Dice similarity coefficients (DSC) on all testing volumes, while the x axial presented the epoch number. The dashed curves were the testing accuracy for the case that only axial images were used astraining and testing images. The solid curves were the testing accuracy for the case that all axial, coronal and sagittal view images were used in both training and testing processing. From this figure, the mean testing accuracy plots were systematically increased when trained with more epochs. For most of the epochs, the proposed SSNet achieved more accurate testing results than GCN and Unet on both single view and multi-view training scenarios.

Figure X.3 presents the qualitative results of different deep learning methods along with the manual segmentation. The upper, middle and lower rows were corresponding to the subjects with highest, median and lowest DSC values of SSNet using three views. The segmentation results of Unet, GCN and SSNet on using (1) only axial 2D images, and (2) all axial, coronal and sagittal 2D images were shown in the figure for different columns. The manual segmentation results for the same subjects were presented as the right-most column. In Figure X.4 presents the quantitative results of different deep learning methods as box plots. All the other methods were compared with the proposed SSNet using three view images ("Ref."). The proposed method achieved significantly better DSC results ($p<0.01$) than methods with "*" except the one with "N.S.". The lowest DSC value of the SSNet is smaller than the benchmark methods. From Figure X.3 and Figure X.4, the GCN outperformed the Unet by capturing the large spatial variation for the splenomegaly segmentation. By adding GAN supervision, the proposed method not only alleviated the outliers but also achieved the higher median DSC (0.9262) and mean DSC (0.9260) compared with baseline methods. Meanwhile, using richer training data on three imaging views leveraged the segmentation performance for a significant margin.

5. conclusion and Discussion

We proposed the SSNet to perform the splenomegaly segmentation using MRI clinical acquired scans. Richer training data in the form of 2-D triplanar sections improved all methods, but SSNet remained

superior than GCN and had fewer outliers than Unet. From Figure X.2, the proposed SSNet achieved generally better performance on median DSC compared with benchmark methods on different epoch numbers. From Figure X.3 and Figure X.4, the SSNet was shown to achieve more accurate (higher median DSC) and more robust (higher lowest DSC) segmentation performance compared with benchmark results. The results also demonstrated that using all axial, coronal and sagittal images as both training and testing data consistently provided us better segmentation performance than using single axial view.

The major limitation of this work was that the segmentation was performed on the 2D images, which might lose the 3D spatial information. In the future, it would be worth exploring 3D deep neural networks to conduct the splenomegaly segmentation. Another interesting direction could be to integrate the clinical diagnostic information to the image segmentation using the attention models [341].

Chapter XI. Adversarial Synthesis Learning Enables Segmentation Without Target Modality Ground Truth

1. Introduction

Splenomegaly, the condition of having an abnormally large spleen (e.g., >500 cubic centimeter), is a biomarker for liver disease, infection and cancer. Previous automated methods have been proposed to perform segmentation on normal spleens [313, 322] and with splenomegaly [325, 342, 343]. Recently, deep convolutional neural network (DCNN) based methods have been used in splenomegaly and shown superior performance [155, 344]. However, one major limitation of deploying DCNN methods is that one typically has to manually trace a new set of training data when segmenting organs in a new imaging modality or segmenting abnormal organs from a new disease cohort. For instance, a DCNN trained with normal spleens was not able to capture the spatial variations of splenomegaly (Figure X.1). Therefore, a straightforward solution is to manually annotate a set of splenomegaly CT scans. However, manual tracing is resource intensive and potentially error prone.

Image synthesis has been used to segment images for one modality from another [141-144]. However, paired images were typically required for traditional methods. Recently, the cycle generative adversarial networks (CycleGAN) [145] provided an effective tool for inter-modality synthesis from unpaired images [146, 147]. Therefore, one could synthesize the training images and labels for splenomegaly patients and labels on one modality (e.g., MRI) while targeting another modality (e.g., CT). Upon such idea, Chartsias et al. [148] proposed an CT to MRI synthesis method using CycleGAN and trained another independent MRI segmentation network (called "Seg.") using the synthesized MRI images. Although still using manual labels for both modalities, this two stage framework (called "CycleGAN+Seg.") revealed a promising direction: segmentation was possible without ground truth in the target modality.

In this paper, we propose a novel end-to-end synthesis and segmentation network (EssNet) to

perform MRI to CT synthesis and CT splenomegaly segmentation simultaneously without using ground truth labels in CT. The EssNet was trained by unpaired MRI and CT scans and only used manual labels from MRI scans.

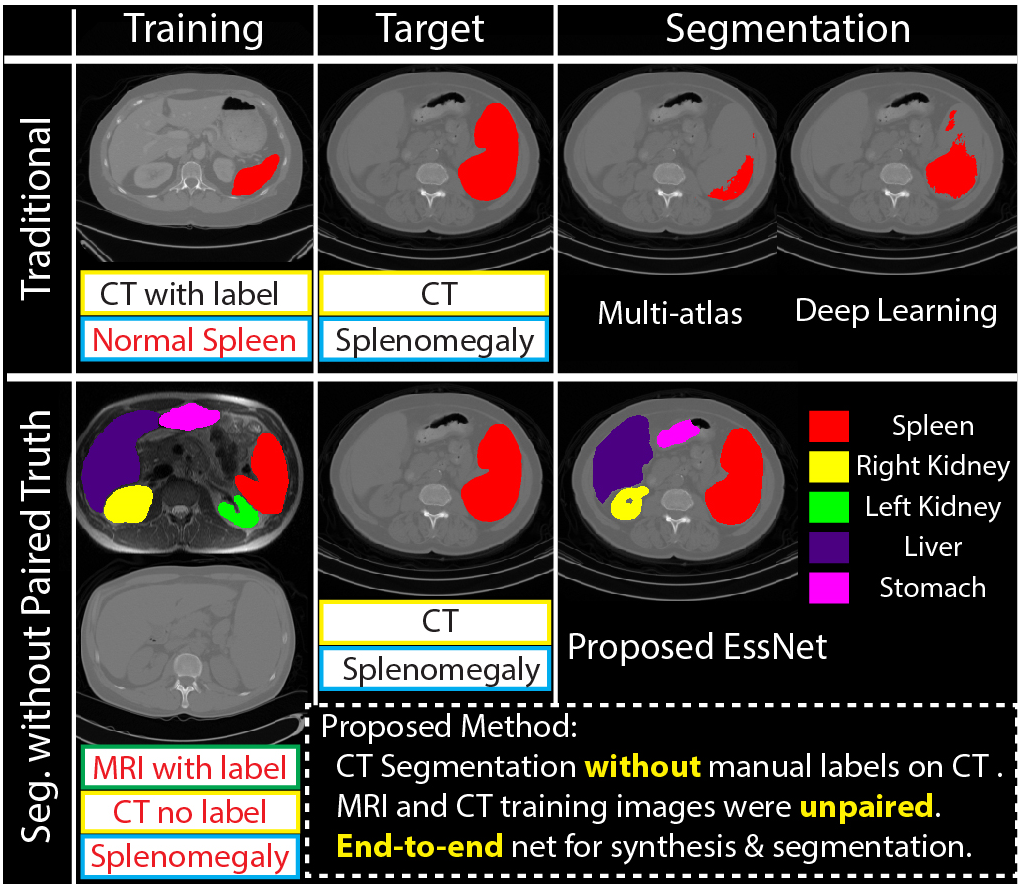


Figure XI.1 The upper row shown that carnonical methods trained by normal spleen failed in splenomegaly segmentation. The lower row shown that the proposed EssNet achieved splenomegaly segmentation from unpaired MRI and CT training images without using CT labels.

2. Data

Unpaired 60 whole abdomen MRI T2w scans and 19 whole abdomen CT with splenomegaly spleen were used as the experimental data, whose imaging parameters and demographic information were introduced in [344] and [325]. Six labels (spleen, left kidney, right kidney, liver, stomach and body) were manually delineated for each MRI [344], while one label (spleen) was manually traced for each CT scan [325]. Additional 75 whole abdomen CT scans with normal spleens [322] were used to train a baseline

DCNN method.

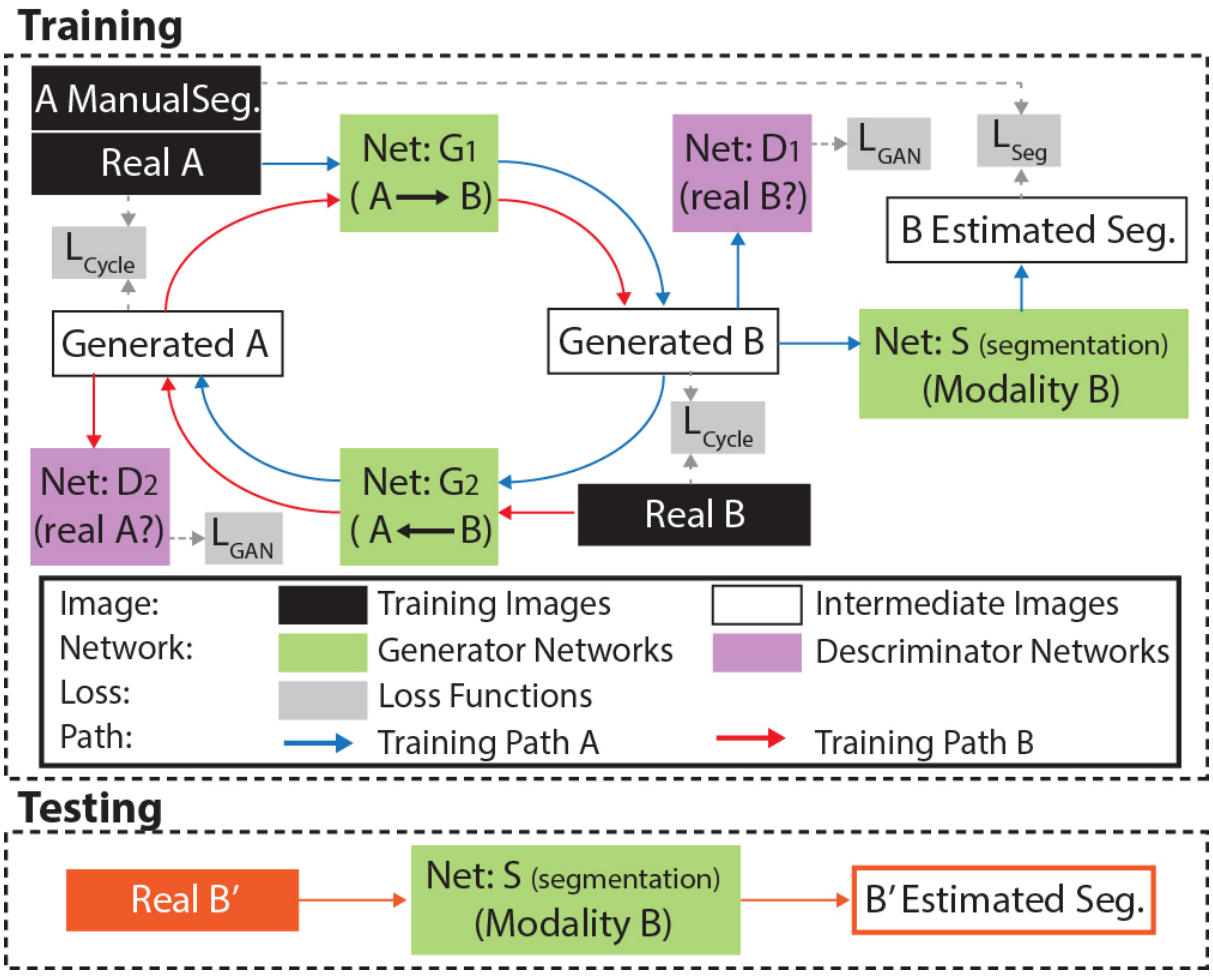


Figure XI.2 The left side was the CycleGAN synthesis subnet, where A was MRI and B was CT. G_1 and G_2 were the generators while D_1 and D_2 were discriminators. The right subnet was the segmentation subnet for an end-to-end training. Loss function were added to optimize the EssNet.

3. Method

The network structure of EssNet is shown in Figure XI.2**,** where"A" indicates MR images while "B" represents CT images. The 9 block ResNet (defined in [145, 345]) was used as the two generators (G_1 and G_2). G_1 synthesized an image x in modality A to the generated B image ($G_1(x)$), while G_2 synthesize an image y in modality B to the generated A image ($G_2(y)$). The PatchGAN (defined in [145, 336]) was employed as the two adversarial discriminators (D_1 and D_2). D_1 distinguished if the CT image was real or

generated, while D_2 determined for the MR image. When deploying such framework on unpaired A and B, two training paths (Path A and Path B) existed in forward cycles. The cycle synthesis subnet was basically the same as CycleGAN [145].

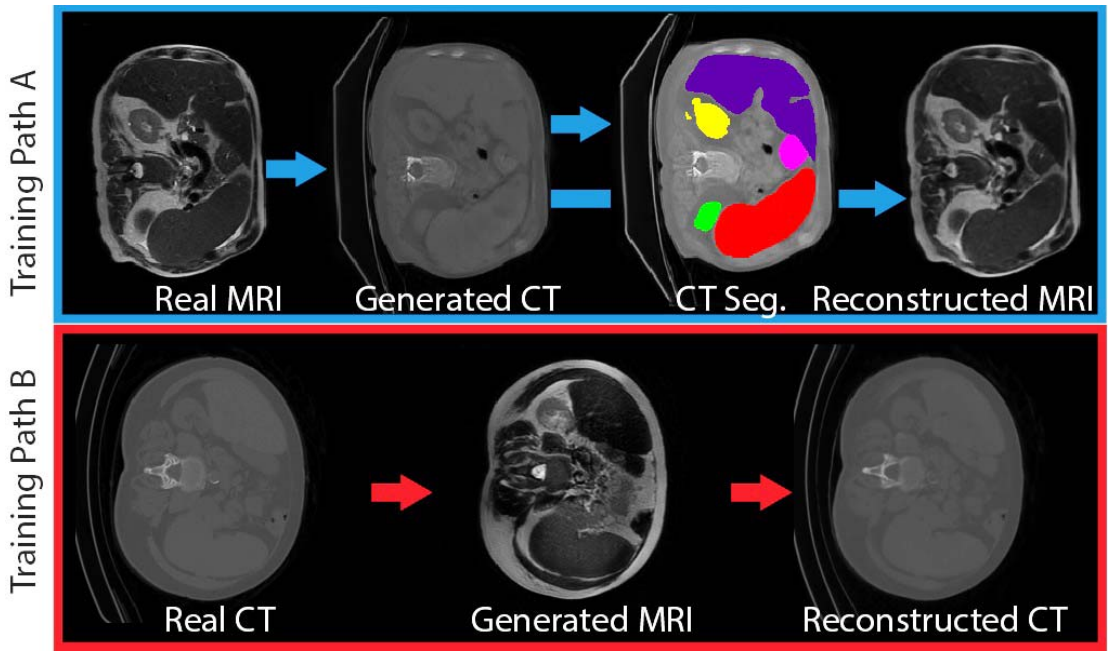


Figure XI.3 The qualitative results of the synthesized images and segmentations in training Path A and Path B.

Since the aim of the proposed EssNet was to perform end-to-end synthesis and segmentation. The segmentation network S was concatenated after G_1 directly as an additional forward branch in Path A. The 9 block ResNet [145, 345] were used as S, whose network structure was identical to G_1. Then, the estimated segmentation from generated B was derived.

Five loss functions were used to train the network. Two adversarial loss functions $\mathcal{L}_{\text{GAN}}$ were defined as

$$\begin{aligned}\mathcal{L}_{\text{GAN}}(G_1, D_1, A, B) &= E_{y\sim B}[\log D_1(y)] + E_{x\sim A}[\log(1 - D_1(G_1(x)))] \\ \mathcal{L}_{\text{GAN}}(G_2, D_2, B, A) &= E_{x\sim A}[\log D_2(x)] + E_{y\sim B}[\log(1 - D_2(G_2(y)))]\end{aligned} \tag{11.1}$$

Two cycle consistency loss $\mathcal{L}_{\text{cycle}}$ functions were used to compare the reconstructed images with real images.

$$\mathcal{L}_{cycle}(G_1, G_2, A) = E_{x\sim A}[\|G_2(G_1(x)) - x\|_1]$$
$$\mathcal{L}_{\text{cycle}}(G_2, G_1, B) = E_{y\sim B}[\|G_1(G_2(y)) - y\|_1] \quad (11.2)$$

The segmentation loss function was defined as

$$\mathcal{L}_{\text{seg}}(S, G_1, A) = -\sum_i m_i \cdot \log(S(G_1(x_i))) \quad (11.3)$$

where m was the manual labels for image x, i was the index of a pixel. Then, the total loss function was defined as

$$\mathcal{L}_{\text{total}} = \lambda_1 \cdot L_{\text{GAN}}(G_1, D_1, A, B) + \lambda_2 \cdot \mathcal{L}_{\text{GAN}}(G_2, D_2, B, A) + \lambda_3 \cdot \mathcal{L}_{\text{cycle}}(G_1, G_2, A) + \lambda_4 \cdot \mathcal{L}_{\text{cyle}}(G_2, G_1, B) + \lambda_5 \cdot \mathcal{L}_{\text{seg}}(S, G_1, A) \quad (11.4)$$

In this work, the lambdas were empirically set to $\lambda_1 = 1, \lambda_2 = 1, \lambda_3 = 10, \lambda_4 = 10, \lambda_5 = 1$. To minimize the $\mathcal{L}_{\text{total}}$, the Adam optimizer was used [145]. The examples of real, synthesized, reconstructed and segmentation images for Path A and Path B were shown in Figure XI.3.

In testing, only trained network S was used and B' represented the testing CT images. the Dice similarity coefficient (DSC) values between automated and manual segmentations were used as the metrics to evaluate the performance of different segmentation methods. All statistical significance tests were made using a Wilcoxon signed rank test ($p<0.05$).

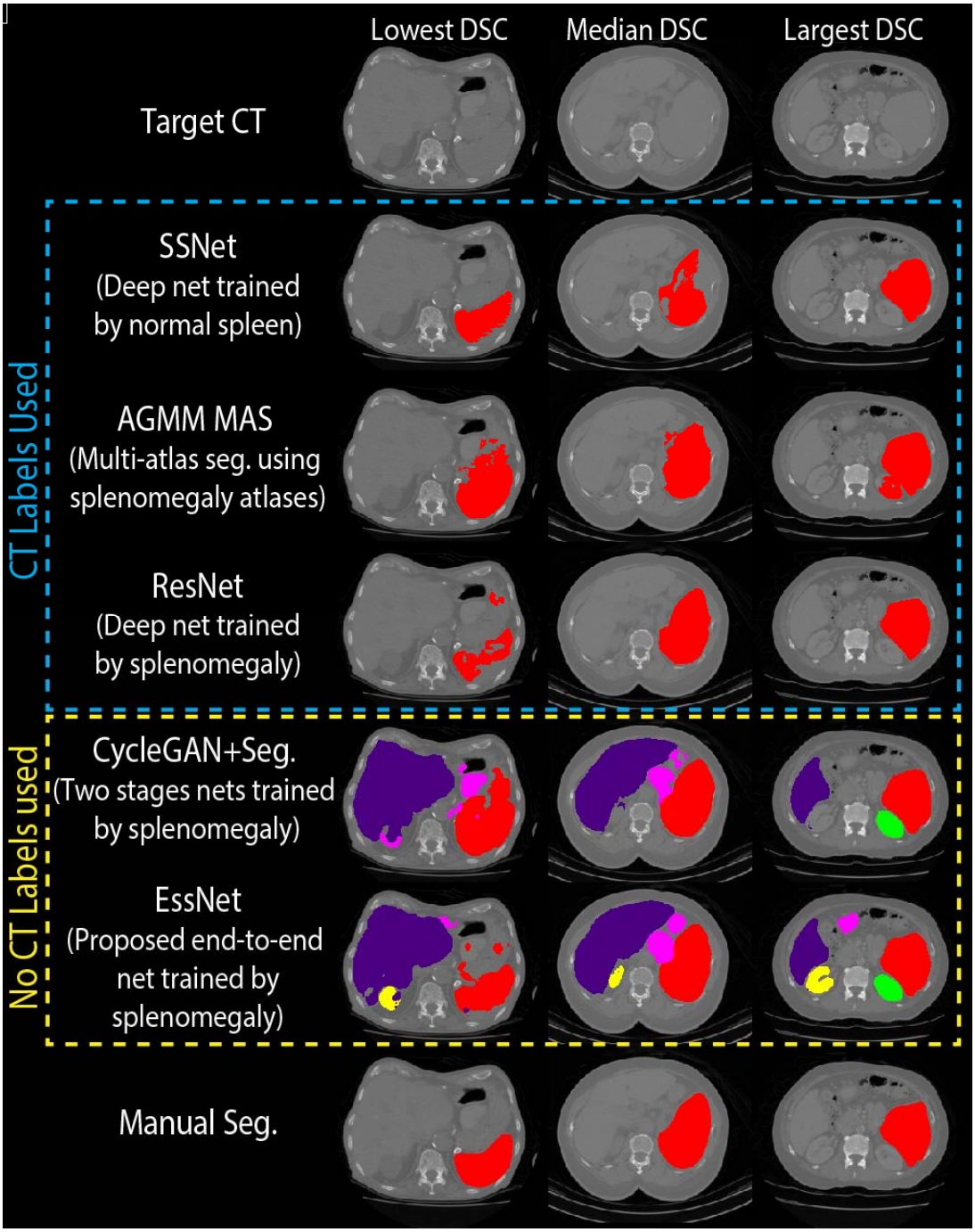


Figure XI.4 The qualitative results of (1) three canonical methods using CT manual labels in CT segmentation, and (2) CycleGAN+Seg. and the proposed EssNet methods without using CT manual labels. The splenomegaly CT labels were only used in validation and excluded from training for (2). Moreover, later methods not only performed spleen segmentation but also estimated labels for other organs, which were not provided by canonical methods when such labels were not available on CT.

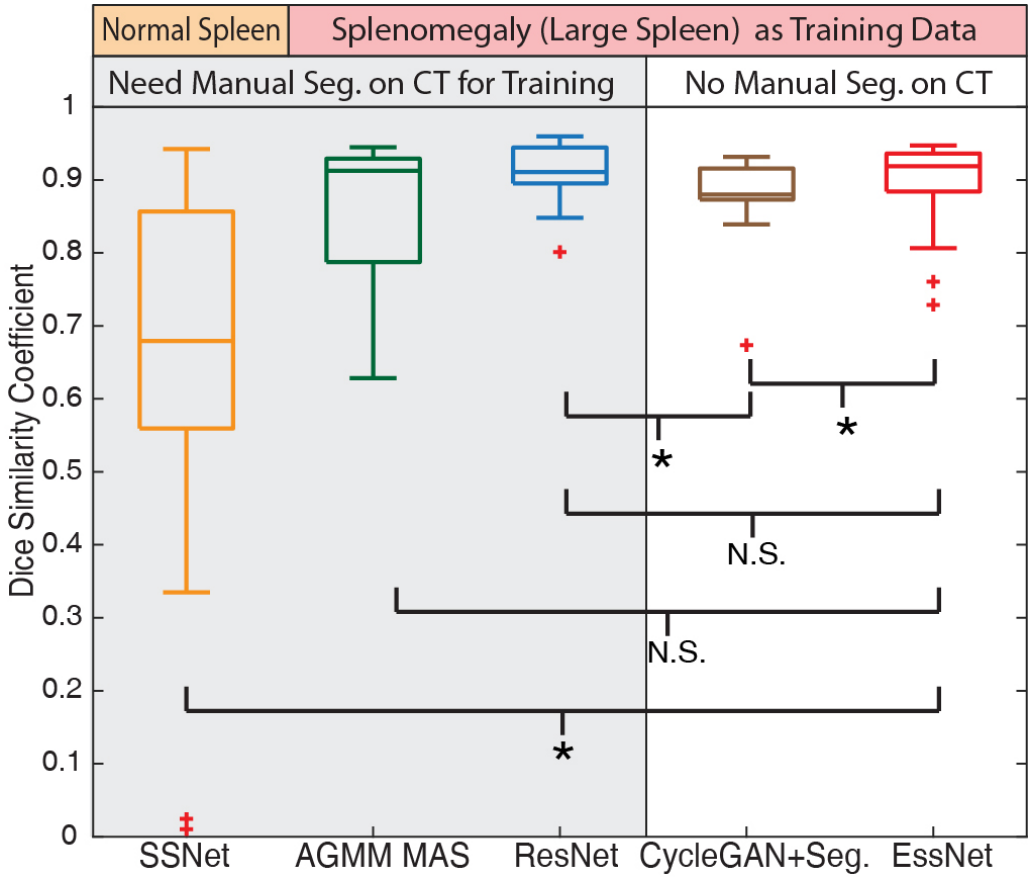


Figure XI.5 The boxplot results of all CT splenomegaly testing images, where "*" means the difference are significant at p<0.05, while "N.S." means not significant.

4. Results

The qualitative results of different methods on three subjects (lowest, median and highest DSC for the EssNet) were shown in the Figure XI.4. From the results, the EssNet was not only able to perform the spleen segmentation, but also estimated segmentations on liver, left kidney, right kidney and stomach. The quantitative results of different segmentation strategies on all CT scans were shown in the Figure XI.5 as a boxplot. The "*" indicates the difference were significant, while "N.S." means not significant.

5. Conclusion and Discussion

In this work, we proposed the end-to-end EssNet for simultaneous image synthesis and

segmentation. We demonstrate this approach on splenomegaly CT segmentation without using ground truth labels in CT. From Figure XI.3, the proposed end-to-end approach was able to achieve MRI to CT synthesis, CT to MRI synthesis, and the CT segmentation simultaneously. Figure XI.4 shown that the proposed method was not only able to obtain spleen segmentation but also estimate liver, kidney, stomach labels, while the canonical methods using CT data only were not able to when such labels were not available on CT. Figure XI.5 shown that the SSNet trained by normal spleen CT images was significantly worse than other methods. The proposed EssNet method was significantly better than the two stages CycleGAN+Seg. method. Without using CT labels, the EssNet achieved the comparable performance as the AGMM MAS and ResNet that used CT labels. On the contrary, the performance of CycleGAN+Seg. was significantly worse than ResNet.

This study opens the possibility of using EssNet to perform the segmentations on other modalities on which target labels are not known and paired inter-modality data are not available. An interesting limitation of this work is that the networks are 2-D (but assessed in 3-D) due to time and memory concerns. Either post processing for 3-D consistency or 3D EssNet would be a promising area.

Chapter XII. Automated characterization of pyelocalyceal anatomy using CT urograms in management of kidney stones

1. Introduction

Prevalence of kidney stone disease, or nephrolithiasis, has been rising over the last several decades and now affects approximately 1 in 11 individuals in the United States [1]. Most stones that do not spontaneously pass will require surgical treatment with ureteroscopy (retrograde endoscopy through the urethra and bladder), extracorporeal shock wave lithotripsy (stones fragmentation using noninvasive shock waves), percutaneous lithotripsy (endoscopy through 1 cm direct puncture into the kidney), or very rarely laparoscopic or open surgery. An efficient and effective choice of surgical approach is critical given the significant morbidity due to kidney stones, including pain, infection, and renal insufficiency, as well as associated costs, which were estimated to be over $5 billion in 2000 [2].

In determining an optimal operation, it is essential to consider anatomic factors and stone features as these affect treatment success rates [3, 4]. Prior studies correlating specific characteristics of the pyelocalyceal anatomy (kidney drainage or collecting system), such as the infundibulopelvic angle (IPA) (angle representing the lowest dependent portion of the drainage system), and stone-free rates after surgery have utilized 2-dimensional (2D) imaging studies to characterize the 3-dimensional (3D) urinary collecting system, a discrepancy that has led to conflicting data [3, 5, 6]. For example, the range of infundibulopelvic angles in patients using 2-dimensional retrograde pyelograms are not consistent with those measured from 3-dimensional resin casts of cadaver kidneys. Furthermore, many of these studies were performed with manual measurements of images taken during surgery, meaning the images are not available pre-operatively to actually aid in treatment planning. The above indicate a strong need for imaging-based 3D analysis of pyelocalyceal anatomy in order to achieve appropriate patient-specific preoperative planning and counseling. Abundant availability of computed tomography (CT) scans provides an ideal opportunity to develop algorithms for patient-specific computer-aided treatment guidance. In addition, this type of data at

a population level will be highly valuable in the development of novel devices for kidney stone surgery and general characterization of anatomy.

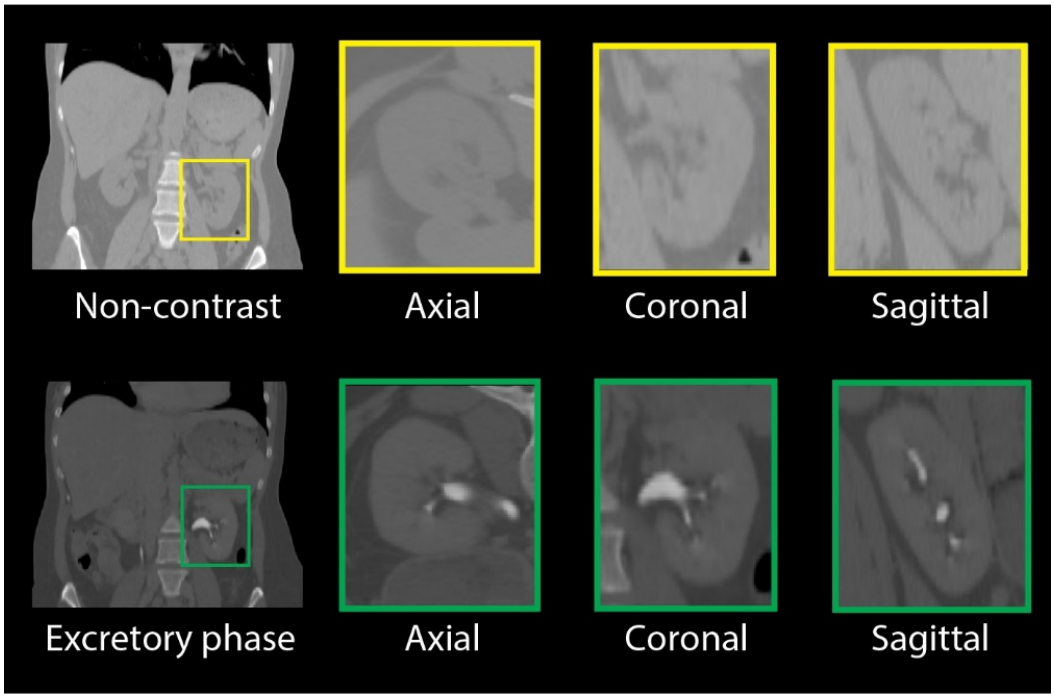


Figure XII.1 Top: Non-contrast CT with cropped images of the kidney in which pyelocalyceal system is not visualized. Bottom: Excretory phase of CT Urogram with cropped images of kidney and pye-localyceal anatomy illuminated during excretion of contrast by the kidneys.

In this feasibility study, we aimed to automatically identify the 3D structure of the renal collecting system anatomy in CT Urograms that could then be used to measure the IPA, a key feature previously identified as potentially correlating with success of a given surgical approach.

2. Methods

2.1. Patient Selection and Imaging

The Institutional Review Board approved this study with a waiver of informed consent. Electronic medical records were used to randomly identify patients who had a CT urogram for hematuria workup. Exclusion criteria included any treated or untreated kidney pathology including tumors, presence of kidney stones, anatomic variants, and chronic renal insufficiency as this affects the rate of contrast excretion.

Images were manually reviewed to confirm good image quality. All excretory phase sequences in this study were performed in the prone position (Excretory Phase in Figure VII.1) at an 8 minute delay per institutional protocol with 3mm axial reconstructions.

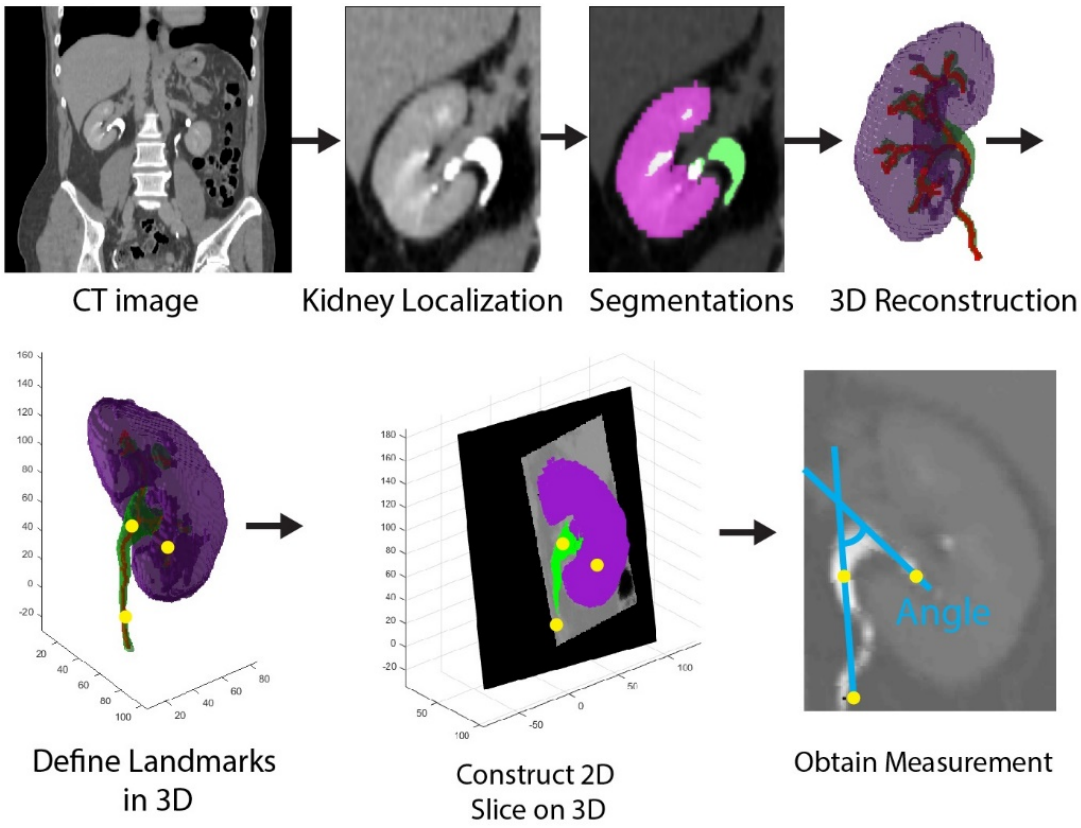


Figure XII.2 The workflow of the proposed framework. First, the whole kidney was localized and segment-ed using multi-atlas segmentation. Then the pyelocalyceal structure was segmented from a Gaussian Matured Model and the tree structure was subsequently derived. Key landmarks (yellow dots) were manually identified from the 3D reconstruction and tree structure to con-struct an oblique 4mm thick plane from which the IPA was measured.

2.2. Automated Localization and Segmentation of Whole Kidney

Figure XII.2 demonstrates the workflow of the proposed the algorithm. A SIMPLE context learning-based multi-atlas segmentation framework [7] was used to achieve whole kidney segmentation. To achieve the SIMPLE framework, 30 pairs of atlases (anatomical CT scans and corresponding labels) were obtained from MICCAI 2015 MeDiCAL challenges (https://www.synapse.org/#!Synapse:syn3193805/wiki/89480). Two sets of cropped atlases were then formed based on kidney locations (30 pairs each for the left and right kidneys). The atlases were manually

cropped by an experienced rater using MIPAV software [8]. Next, the left and right kidneys in target CT Urogram images were automatically localized and cropped using a random forest based localization method [9]. The previously cropped atlases were then registered to the cropped target CT Urogram images using affine and non-rigid registrations by NiftyReg [10]. A SIMPLE based context learning procedure was performed to select the best 10 registered atlases for each target kidney [11]. Finally, the left and right kidney segmentations were derived by performing the joint label fusion (JLF) [12] on the selected atlases.

2.3. Automated Segmentation of Pyelocalyceal Anatomy and Validation

Once the kidneys were cropped and segmented from the original excretory phase image, a Gaussian mixture model (GMM) was used to segment the pyelocalyceal anatomy within the kidneys. Empirically, a threshold (above 100 Hounsfield Unit (HU)) was applied to exclude tissues surrounding the kidney. The GMM with three components was then employed on the histogram of remaining intensities. The two components (from three total) with higher mean HU score were clustered and identified to be the pyelocalyceal anatomy segmentation. The component with smallest mean HU score represented residual kidney organ tissue not completed removed in the initial thresholding step. Finally, a 3D tree structure (center line) was derived from the pyelocalyceal anatomy segmentation using the method described in [13]. Briefly, the method calculated the 3D axis skeleton of 3D binary volume using a parallel thinning algorithm based on Eular table.

All pyelocalyceal segmentations were qualitatively evaluated by a radiologist and rated as having excellent, acceptable, or poor accuracy. A subset of the kidneys that resulted in excellent or acceptable segmentations were then manually segmented by a radiologist and the DICE coefficient was calculated.

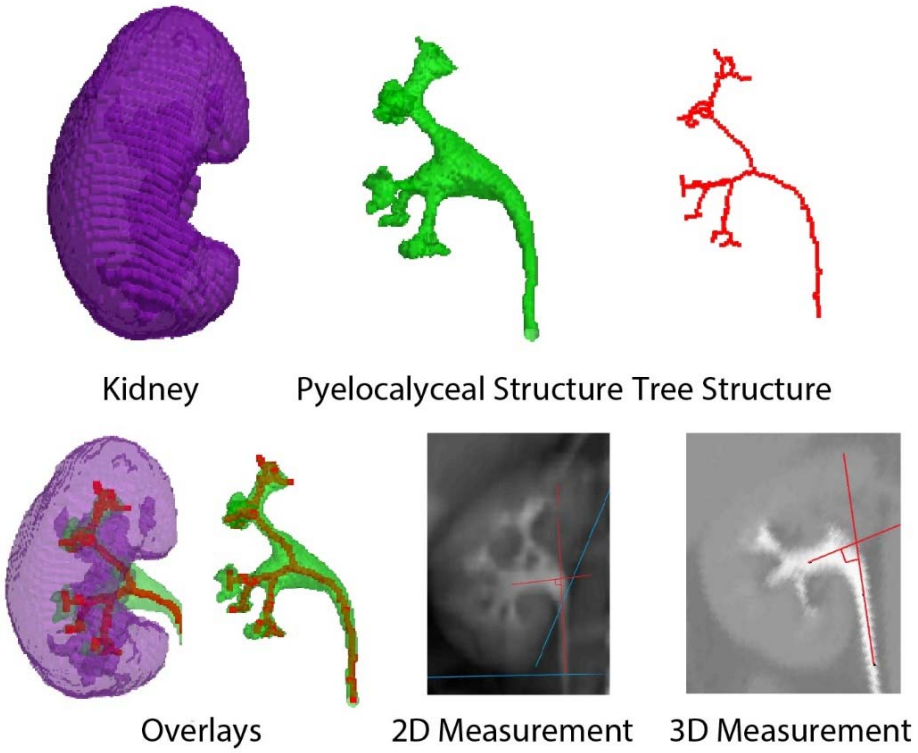


Figure XII.3 Quantitative results of the segmentation and angle measurements for a single kidney. Top row: 3D reconstruction of the kidney, 3D reconstruction of the pyelocalyceal structure, tree struc-ture. Bottom row: Overlays of reconstructions and tree structure, traditional 2D measurement [1] of IPA (red lines) using averaged 2D image (blue lines indicate key landmarks), and the 3D IPA measurement (red lines) using described method.

2.4. Measurement of Infundibulopelvic Angle in 2D and 3D images

The previously described Elbahnasy method for IPA measurement in 2D images was modified to allow for IPA measurement using 3D images and the above derived 3D tree structure [14]. Key landmarks corresponding to those in the Elbahnasy method were identified by a Urologist in 3D slicer software (https://www.slicer.org) using the kidney segmentation, pyelocalyceal anatomy and tree structure derived from above automated algorithm. The landmarks were as follows: (1) the center point of the proximal ureter at the lowest plane of the kidney, (2) the center point of the renal pelvis along medial margin of kidney, (3) a point in the inferior branch of the kidney drainage system. The three points were used to create a unique 4mm thick slice from the 3D volume, and the IPA was measured as the angle between the lines connecting points (1) and (2), and the center line through the lowest branch of the kidney drainage system. As a

comparison, traditional 2D measurements of the IPA were performed on the average cropped kidney image in the coronal direction (Figure XII.3).

3. Results

3.1. Patients

After exclusion of patients with imaging artifacts or inadequate collecting system distension, imaging of 8 renal units from 6 patients were identified to be appropriate for this feasibility study. Patients ranged in age from 42-80 years old and all had normal kidney function.

3.2. Pyelocalyceal Anatomy Segmentation

The pyelocalyceal anatomy was appropriately segmented in 8 of the 11 renal units with a rating of excellent or accep by a radiologist. Of these, 6 were manually segmented by the radiologist and DICE coefficients ranged from 0.62 to 0.88.

3.3. Infundibulopelvic Angle

Figure XII.3 demonstrates the segmentation results, tree structure, as well as 2D and 3D IPA measurements from a single example kidney. The IPA based on the 3D segmentations and tree structures ranged from 14.6 degrees to 81.5 degrees while IPA based on 2D reformatted images ranged from 9.4 degrees to 88.3 degrees (Table XII-1). Comparisons between the angles based on the 2D and 3D methods demonstrated a difference up to 35.6 %

Table XII-1 The angles (degree) obtained from 2D and 3D measurements

Kidney #	2D Measurement	3D Measurement	Absolute difference	Percent difference
1	19.2	23.7	4.5	18.99%
2	16.5	21.9	5.4	24.66%
3	66.9	70.1	3.2	4.57%
4	34.2	48.6	14.4	29.63%
5	57.4	60	2.6	4.33%
6	9.4	14.6	5.2	35.63%
7	23.1	19.7	3.4	17.26%
8	88.3	81.5	6.8	8.34%

4. Discussion

Kidney stone disease is a chronic condition that often requires many surgeries over a patient's lifetime. Each surgery is associated with risks and residual stones [346] can have severe consequences so appropriate initial surgical intervention is critical. In addition, anatomic variation may play a role in stone formation or burden of disease [347]. Thus, accurate characterization of patient anatomy can have both immediate and long-term effects with respect to surgical planning as well as lifelong management, such as the interval between imaging studies. This is the first method known to the authors for automated characterization the 3D pyelocalyceal tree. Results demonstrate that this algorithm is technically feasible and DICE coefficient indicate good segmentation results. The relative difference in the measured IPA between the 2D and 3D techniques was up to 35%. In fact, while prior studies have indicated that anatomic variation may be critical to predicting surgical success, the data are inconsistent and, as this preliminary data suggests, part of the discrepancy may be due to inaccuracies from utilization of 2D images. An inherent limitation of such automated algorithms is that the result will only be as good as the initial imaging, and

imaging quality of CT urograms can be dependent on multiple factors such as kidney function and level of hydration. We aim to further automate our algorithm, assess additional anatomic variables, both novel and previously described, and then correlate these with stone free rates after stone surgery. Outcomes from such studies may provide valuable tools for patient-specific stone management.

Chapter XIII. Conclusions and Future Work

1. Summary

The large-scale medical image processing and analyses are challenging for both brain and abdomen. For the brain, we have established an end-to-end large-scale medical image analysis framework in investigating lifespan aging by conducting robust and consistent whole brain volume and surface metrics (Chapters II, III, IV), controlling inter-subject variations (Chapters V, VI), and conducting robust statistical analyses (Chapter VII). We have generalized the multi-atlas label fusion theory from 3D to 4D for longitudinal whole brain segmentation (Chapter VIII). For the abdomen, we have proposed splenomegaly segmentation methods using multi-atlas approach, deep convolutional neural networks, and synthesis learning (Chapter IX, X, XI). Then, we applied abdomen segmentation methods to achieve a tree structure of the urinary collecting system, allowing for 3-dimensional characterization of the pyelocalyceal anatomy (Chapter XII).

2. Consistent Whole Brain Segmentation and Cortical Reconstruction

2.1. Summary

Whole brain multi-atlas segmentation and cortical surface reconstruction have long been regarded as two unrelated techniques. We proposed the first work, MaCRUISE, to combine multi-atlas segmentation with cortical surface reconstruction (Chapter III). This method was extended to achieve detailed surface parcellation (Chapter IV). Using such technique, 132 volume labels and 98 surface labels were achieved from a clinical acquired single T1w MRI scan.

2.2. Main Contributions

- MaCRUISE combined the previous independent volume segmentation and surface reconstruction into a uniformed and consistent framework.
- It achieved more robust surface reconstruction and more accurate volume segmentation

compared with state-of-the-art methods.

- Detailed annotations (132 volume labels and 98 surface labels) were achieved from a single T1w MRI scan.

2.3. Future Work

The processing speed is a major limitation in the MaCRUISE as the multi-atlas segmentation and the surface reconstruction are computational expensive. In recent years, the deep learning segmentation methods have been shown their advantages, especially on the computational time. Therefore, it is appealing if further efforts can be made to integrate the deep learning techniques with whole brain segmentation and surface reconstruction.

3. Large-scale Multi-Site Image Data Analysis

3.1. Summary

Recent developments on data sharing and computational power offer us an opportunity to explore large-scale medical image data. In this work, we have collected more than 5000 normal MRI scans from night projects and most of them are public available. We proposed the MLF algorithm to perform fast whole brain segmentation from machine learning perspective (Chapter II). With such large cohort, we presented the novel data-driven probabilistic atlas to achieve personalized prior in less than ten minutes (Chapter V). To deal with the large variations on imaging sequences and subjects, we proposed the multi-atlas-based TICV estimation method (Chapter VI). The TICV measurements were used as linear confounds in life-span brain volumetry analysis using C-RCS method (Chapter VII)

3.2. Main Contributions

- We revisited the whole brain segmentation problem from machine learning perspective. The AdaBoost learning method as well as PCA representation were used in the whole brain segmentation.

- We reduced the computational time for achieving whole brain segmentation (with 132 labels) from more than 30 hours to less than 10 minutes using more than 3000 training volumes.
- Data-driven probabilistic atlases were established from a dictionary learned from large-scale training cohort.
- Multi-atlas based simultaneous TICV and PFV estimation method was proposed to achieve more accurate performance than state-of-the-art methods.
- We proposed C-RCS regression method to model the non-linear developmental trajectories of life-span brain volumetry.
- We showed the changes of structural connectives and volumetric trajectories from global, network, and regional levels.

3.3. Future Work

Large-scale medical image analysis has been regarded as one of the major future directions of medical image analysis. However, the computational efficiency, the robustness on multi-site even multi-sequence data, and the large-scale data mining algorithms are among the key barriers in large-scale image analyses. The deep learning techniques as well as Big Data mining techniques developed in computer vision, machine learning, and bioinformatics are promising solutions for the next generation large-scale medical image analyses.

4. Longitudinal Whole Brain Segmentation

4.1. Summary

To improve reproducibility, longitudinal segmentation (4D) approaches have been investigated to reconcile temporal variations with traditional 3D approaches. We propose the longitudinal label fusion algorithm, 4DJLF, to incorporate the temporal consistency modeling via non-local patch-intensity covariance models (Chapter XIII).

4.2. Main Contributions

- We generalized the multi-atlas label fusion theory from 3D to 4D for longitudinal scenarios. 4DJLF is under the general label fusion framework by simultaneously incorporating the spatial and temporal covariance on all longitudinal time points.
- The proposed algorithm is a longitudinal generalization of a leading joint label fusion method (JLF) that has proven adaptable to a wide variety of applications.
- The spatial temporal consistency of atlases is modeled in a probabilistic model inspired from both voting based and statistical fusion.

4.3. Future Work

It is challenging to reconcile temporal inconsistency while keep sensitivity. To develop spatial temporal consistent whole brain MRI segmentation method is essential, yet challenging task. One major limitation is that we did not have a longitudinal MRI cohort with detailed manual segmentations. Therefore, it would be valuable if we can provide such validation dataset as a publicly available dataset for the community.

5. Multi-atlas Based Abdomen Image Processing

5.1. Summary

Non-invasive splenomegaly segmentation from 3D MRI or CT is challenging given the diverse structural variations of human abdomens as well as the wide variety of clinical acquisition schemes. We proposed the multi-atlas based fully automated and semi-automated splenomegaly segmentation methods (Chapter IX). Then, the multi-atlas segmentation technique was applied to the kidney to get tree structures and 3D measurements for renal collecting system (Chapter XII).

5.2. Main Contributions

- The automated segmentation method using the selective and iterative method for

performance level estimation (SIMPLE) atlas selection was used to address the concerns of inhomogeneity for clinical splenomegaly MRI.

- The semi-automated craniocaudal spleen length-based SIMPLE atlas selection (L-SIMPLE) was proposed to integrate a spatial prior in a Bayesian fashion and guide iterative atlas selection.
- The graph cuts refinement was employed to achieve the final splenomegaly segmentation from the probability maps from multi-atlas segmentation.
- For the kidney, we propose a novel non-invasive framework that automatically achieves a tree structure of the urinary collecting system using CT urograms, allowing for 3-dimensional characterization of the pyelocalyceal anatomy.

5.3. Future Work

For splenomegaly segmentation, the computational time would be further reduced when performing atlas selection (e.g., using the information from spleen length). Another direction worth pursing is to use the spleen length L and its spatial information to initialize or leverage the image registration. For kidney structure analyses, the landmark annotation step can be fully automated in the future.

6. Deep Learning Based Abdomen Image Processing

6.1. Summary

In recent years, deep convolutional neural networks segmentation methods have demonstrated advantages for abdominal organ segmentation. First, we proposed the SSNet to address spatial variations when segmenting extraordinarily large spleens (Chapter X). The SSNet was designed based on the framework of image-to-image conditional generative adversarial networks. Second, we proposed a novel end-to-end synthesis and segmentation network (EssNet) to achieve the unpaired MRI to CT image synthesis and CT splenomegaly segmentation simultaneously without using manual labels on CT (Chapter XI).

6.2. Main Contributions

- For splenomegaly segmentation, we proposed SSNet for the fast splenomegaly segmentation. Global convolutional network (GCN) was used as the generator to reduce false negatives, while the Markovian discriminator (PatchGAN) was used to alleviate false positives.
- We proposed the EssNet that enabled the end-to-end simultaneous synthesis learning and segmentation. Using EssNet, we achieved accurate spleen segmentation without having ground truth labels in the target modality.

6.3. Future Work

The SSNet and EssNet were designed using 2D frameworks rather than 3D due to the limitation that we did not have large enough 3D training dataset. A promising direction is to extend the SSNet and EssNet from 2D to 3D to have better spatial consistency. Another appealing direction is to combine traditional medial image techniques (e.g. registration, preprocessing, postprocessing) with deep learning techniques to further leverage the segmentation performance.

7. Concluding Remarks

The application of medical image analysis to the large-scale images is a challenging field. In this dissertation, we address these challenges by proposing new algorithms and improving already developed tools for automated large-scale medical image processing and data. We have addressed many key obstacles for performing large-scale medical image processing and analyses . Medical image analysis on large-scale imgae data is improving and new techniques are constantly emerging. By its nature, medical image research is a data intensive and collaborative discipline, which requires the new infrastructures and techniques to systematically extract, examine and result in new knowledge. Yet both image processing and data analysis for Big Data medical images are still maturing, which still leaves room for either adapting the existing techniques for Big Data scenario or even proposing new approaches. The is also room for applying the

large-scale medical image analysis on understanding the fundamental problems and diseases in human brain and abdomen .

Appendix A: Publications

1. Journal Articles

1. **Yuankai Huo**, Justin Blaber, Stephen M. Damon, Brian D. Boyd, Shunxing Bao, Prasanna Parvathaneni, Camilo Bermudez Noguera, Shikha Chaganti, Vishwesh Nath, Jasmine M. Greer, Ilwoo Lyu, William R. French, Allen T. Newton, Baxter P. Rogers, and Bennett A. Landman. "Towards Portable Large-scale Image Processing with High-Performance Computing". Journal of Digital Image. (accepted with minor revision)
2. **Yuankai Huo**, Jiaqi Liu, Zhoubing Xu, Robert L. Harrigan, Albert Assad, Richard G. Abramson, and Bennett A. Landman. "Robust Multicontrast MRI Spleen Segmentation for Splenomegaly Using Multi-Atlas Segmentation." IEEE Transactions on Biomedical Engineering 65, no. 2 (2018): 336-343.
3. **Yuankai Huo**, Andrew J. Asman, Andrew J. Plassard, and Bennett A. Landman. "Simultaneous total intracranial volume and posterior fossa volume estimation using multi‐atlas label fusion." Human brain mapping 38, no. 2 (2017): 599-616
4. **Yuankai Huo**, Aaron Carass, Susan M. Resnick, Dzung L. Pham, Jerry L. Prince, Bennett A. Landman. “Consistent Cortical Reconstruction and Multi-atlas Brain Segmentation”. NeuroImage. Volume 138, September 2016, Pages 197–210 PMC4927397
5. Qu Tian, Susan M. Resnick, Bennett A. Landman, **Yuankai Huo**, Vijay K. Venkatraman, Christopher E. Gonzalez, Eleanor M. Simonsick, Michelle D. Shardell, Luigi Ferrucci, Stephanie A. Studenski “Lower gray matter integrity is associated with greater lap time variation in high-functioning older adults.” Experimental Gerontology. 2016. 2016 May;77:46-51
6. Andrew J. Asman, **Yuankai Huo***, Andrew J. Plassard, and Bennett A. Landman, “Multi-atlas Learner Fusion: An efficient segmentation approach for large-scale data”, Medical Image Analysis (MedIA), 2015 Dec;26(1):82-91 (***Corresponding Author**)

2. Highly Selective Conference Publications

1. **Yuankai Huo**, Katherine Aboud, Hakmook Kang, Laurie E. Cutting, Bennett A. Landman. “Mapping Lifetime Brain Volumetry with Covariate-Adjusted Restricted Cubic Spline Regression from Cross-sectional Multi-site MRI”. In International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), Athens, Greece, October 2016. (Oral Presentation).
2. Prasanna Parvathaneni, Baxter P. Rogers, **Yuankai Huo**, Kurt G. Schilling, Allison E. Hainline, Adam W. Anderson, Neil D. Woodward, and Bennett A. Landman. "Gray Matter Surface based Spatial Statistics (GS-BSS) in Diffusion Microstructure." In International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 638-646. Springer, Cham, 2017.

3. Conference Publications

1. **Yuankai Huo**, Zhoubing Xu, Shunxing Bao, Albert Assad, Richard G. Abramson, and Bennett A. Landman. "Adversarial Synthesis Learning Enables Segmentation Without Target Modality Ground Truth." Accpted in ISBI 2018.
2. **Yuankai Huo**, Shunxing Bao, Prasanna Parvathaneni, and Bennett A. Landman. "Improved Stability of Whole Brain Surface Parcellation with Multi-Atlas Segmentation." In Proceedings of the SPIE Medical Imaging Conference. Houston, Texas, February 2018
3. **Yuankai Huo**, Zhoubing Xu, Shunxing Bao, Camilo Bermudez, Andrew J. Plassard, Jiaqi Liu, Yuang Yao, Albert Assad, Richard G. Abramson, and Bennett A. Landman. "Splenomegaly Segmentation using Global Convolutional Kernels and Conditional Generative Adversarial Networks." In Proceedings of the SPIE Medical Imaging Conference. Houston, Texas, February 2018
4. Shunxing Bao, **Yuankai Huo**, Prasanna Parvathaneni, Andrew J. Plassard, Camilo Bermudez, Yuang Yao, Ilwoo Llyu, Aniruddha Gokhale, and Bennett A. Landman. "A Data Colocation Grid

Framework for Big Data Medical Image Processing-Backend Design." In Proceedings of the SPIE Medical Imaging Conference. Houston, Texas, February 2018.

5. **Yuankai Huo**, Vaughn Braxton, S. Duke Herrell, Bennett Landman, and Smita De. "Automated Characterization of Pyelocalyceal Anatomy Using CT Urograms to Aid in Management of Kidney Stones." In Computer Assisted and Robotic Endoscopy and Clinical Image-Based Procedures, pp. 99-107. Springer, Cham, 2017.
6. **Yuankai Huo**, Susan M. Resnick, and Bennett A. Landman. "4D Multi-atlas Label Fusion using Longitudinal Images." In International Workshop on Patch-based Techniques in Medical Imaging, pp. 3-11. Springer, Cham, 2017.
7. Peijun Hu, **Yuankai Huo**, Dexing Kong, J. Jeffrey Carr, Richard G. Abramson, Katherine G. Hartley, and Bennett A. Landman. "Automated Characterization of Body Composition and Frailty with Clinically Acquired CT." In International Workshop and Challenge on Computational Methods and Clinical Applications in Musculoskeletal Imaging, pp. 25-35. Springer, Cham, 2017
8. **Yuankai Huo**, Jiaqi Liu, Zhoubing Xu, Robert L. Harrigan, Albert Assad, Richard G. Abramson, Bennett A. Landman. “Multi-atlas Segmentation Enables Robust Multi-contrast MRI Spleen Segmentation for Splenomegaly” In Proceedings of the SPIE Medical Imaging Conference. Orlando, Florida, February 2017. (Oral presentation)
9. Shunxing Bao, Frederick D. Weitendorf, Andrew J. Plassard, **Yuankai Huo**, Aniruddha Gokhale, Bennett A. Landman. “Theoretical and Empirical Comparison of Big Data Image Processing with Apache Hadoop and Sun Grid Engine”. Orlando, Florida, February 2017.
10. Jiaqi Liu, **Yuankai Huo**, Zhoubing Xu, Albert Assad, Richard G. Abramson, Bennett A. Landman. “Multi-Atlas Spleen Segmentation on CT Using Adaptive Context Learning” In Proceedings of the SPIE Medical Imaging Conference. Orlando, Florida, February 2017.
11. **Yuankai Huo**, Aaron Carass, Susan M. Resnick, Dzung L. Pham, Jerry L. Prince, Bennett A. Landman.2016 “Combining Multi-atlas Segmentation with Brain Surface Estimation,” In Proceedings of the SPIE Medical Imaging Conference 2016. (Oral presentation)

12. **Yuankai Huo**, Katherine Swett, Susan M. Resnick, Laurie E. Cutting, Bennett A. Landman. “Data-driven Probabilistic Atlases Capture Whole-brain Individual Variation”, MICCAI MAPPING Workshop, Munich, Germany, October 2015.

4. Conference Abstracts

1. Katherine Swett, **Yuankai Huo**, Elyce Williams, Susan M. Resnick, Bennett A. Landman and Laurie E. Cutting. Socioeconomic status predicts prefrontal cortex volume across the lifespan: a big data, cross-sectional MRI study. In Cognitive Neuroscience Society. San Francisco, California, March 28th, 2015.
2. Camilo Bermudez, **Yuankai Huo**, Andrew Plassard, Katherine Elizabeth Aboud, Laurie Cutting, and Bennett Landman. "Prediction of chronological age from hierarchical brain volumes using a random forest regression can provide a personalized lifetime metric of aging.(P6. 078)." (2017): P6-078.

Appendix B: Biography

Yuankai Huo was born in Suzhou, China in 1985. He received his B.E. degree in telecommunication engineering from Nanjing University of Posts and Telecommunications in Nanjing, China, in 2008, his M.E. degree in information and telecommunication engineering from Southeast University in Nanjing, China, in 2011. Then he moved to U.S. and received his M.S. degree in computer science from Columbia University in the city of New York, New York, in 2014.

Before arriving at Vanderbilt University, he worked at Department of Psychiatry, Columbia University Medical Center, as staff research associate from 2011 to 2014. He also worked for Siemen Healthier at, Princeton, New Jersey, in the summer of 2017. As a graduate student, his work focused on developing robust medical image segmentation algorithms for large-scale and multi-site imaging data. Through his work, he had research interest in structural and functional image computing, machine learning in medical imaging, and large-scale medical image data analyses.

REFERENCES

1. Hedman, A.M., et al., *Human brain changes across the life span: a review of 56 longitudinal magnetic resonance imaging studies.* Hum Brain Mapp, 2012. **33**(8): p. 1987-2002.

2. Bushberg, J.T. and J.M. Boone, *The essential physics of medical imaging*. 2011: Lippincott Williams & Wilkins.

3. Glasser, O., *Wilhelm Conrad Röntgen and the early history of the Roentgen rays*. 1993: Norman Publishing.

4. Dhawan, A.P., *Medical image analysis*. Vol. 31. 2011: John Wiley & Sons.

5. Pham, D.L., C. Xu, and J.L. Prince, *Current methods in medical image segmentation 1.* Annual review of biomedical engineering, 2000. **2**(1): p. 315-337.

6. Sharma, N. and L.M. Aggarwal, *Automated medical image segmentation techniques.* Journal of medical physics, 2010. **35**(1): p. 3.

7. Bankman, I., *Handbook of medical image processing and analysis*. 2008: academic press.

8. Zitova, B. and J. Flusser, *Image registration methods: a survey.* Image and Vision Computing, 2003. **21**(11): p. 977-1000.

9. MacDonald, D., et al., *Automated 3-D extraction of inner and outer surfaces of cerebral cortex from MRI.* NeuroImage, 2000. **12**(3): p. 340-356.

10. Leventon, M.E., *Statistical models in medical image analysis*. 2000, Massachusetts Institute of Technology.

11. McAuliffe, M.J., et al. *Medical image processing, analysis and visualization in clinical research.* in *Computer-Based Medical Systems, 2001. CBMS 2001. Proceedings. 14th IEEE Symposium on*. 2001. IEEE.

12. Duncan, J.S. and N. Ayache, *Medical image analysis: Progress over two decades and the challenges ahead.* Ieee Transactions on Pattern Analysis and Machine Intelligence, 2000. **22**(1): p. 85-106.

13. Ayache, N. and J. Duncan, *20th anniversary of the Medical Image Analysis journal (MedIA)*. 2016.

14. Paragios, N., J. Duncan, and N. Ayache, *Handbook of Biomedical Imaging: Methodologies and Clinical Research*. Vol. 779. 2015: Springer.

15. Rueckert, D., B. Glocker, and B. Kainz, *Learning clinically useful information from images: Past, present and future.* Med Image Anal, 2016. **33**: p. 13-18.

16. Comaniciu, D., et al., *Shaping the Future through Innovations: From Medical Imaging to Precision Medicine.* arXiv preprint arXiv:1605.02029, 2016.

17. Frangi, A.F., Z.A. Taylor, and A. Gooya, *Precision Imaging: more descriptive, predictive and*

integrative imaging. Med Image Anal, 2016. **33**: p. 27-32.

18. Noble, J.A., *Reflections on ultrasound image analysis.* Med Image Anal, 2016. **33**: p. 33-37.

19. Suinesiaputra, A., et al., *Cardiac image modelling: Breadth and depth in heart disease.* Med Image Anal, 2016. **33**: p. 38-43.

20. Weese, J. and C. Lorenz, *Four challenges in medical image analysis from an industrial perspective.* Med Image Anal, 2016. **33**: p. 44-49.

21. Criminisi, A., *Machine learning for medical images analysis.* Med Image Anal, 2016. **33**: p. 91-93.

22. de Bruijne, M., *Machine learning approaches in medical image analysis: From detection to diagnosis.* Med Image Anal, 2016. **33**: p. 94-97.

23. Zhang, S. and D. Metaxas, *Large-Scale medical image analytics: Recent methodologies, applications and Future directions.* Med Image Anal, 2016. **33**: p. 98-101.

24. Van Horn, J.D. and A.W. Toga, *Human neuroimaging as a "Big Data" science.* Brain imaging and behavior, 2014. **8**(2): p. 323-331.

25. Wells, W.M., *Medical Image Analysis–past, present, and future.* Med Image Anal, 2016. **33**: p. 4-6.

26. Poline, J.-B., et al., *Data sharing in neuroimaging research.* Frontiers in neuroinformatics, 2012. **6**: p. 9.

27. Phan, K.L., et al., *Functional neuroanatomy of emotion: a meta-analysis of emotion activation studies in PET and fMRI.* Neuroimage, 2002. **16**(2): p. 331-348.

28. Arnone, D., et al., *Magnetic resonance imaging studies in bipolar disorder and schizophrenia: meta-analysis.* The British Journal of Psychiatry, 2009. **195**(3): p. 194-201.

29. Frazier, T.W. and A.Y. Hardan, *A meta-analysis of the corpus callosum in autism.* Biological psychiatry, 2009. **66**(10): p. 935-941.

30. Ma, B., J. Tromp, and M. Li, *PatternHunter: faster and more sensitive homology search.* Bioinformatics, 2002. **18**(3): p. 440-445.

31. Gorgolewski, K., et al., *Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python.* Frontiers in neuroinformatics, 2011. **5**: p. 13.

32. Evangelou, E., et al., *Non‐replication of association for six polymorphisms from meta‐analysis of genome‐wide association studies of Parkinson's disease: Large‐scale collaborative study.* American Journal of Medical Genetics Part B: Neuropsychiatric Genetics, 2010. **153**(1): p. 220-228.

33. Toga, A.W., et al., *Mapping the human connectome.* Neurosurgery, 2012. **71**(1): p. 1.

34. Schumann, G., et al., *The IMAGEN study: reinforcement-related behaviour in normal brain function and psychopathology.* Molecular psychiatry, 2010. **15**(12): p. 1128-1139.

35. Paley, M.R. and P.R. Ros, *Imaging of spleen disorders*, in *The Complete Spleen*. 2002, Springer. p.

259-280.

36. De Odorico, I., et al., *Normal splenic volumes estimated using three-dimensional ultrasonography.* Journal of ultrasound in medicine, 1999. **18**(3): p. 231-236.

37. Rodrigues, A.J., et al., *Sonographic assessment of normal spleen volume.* Clinical Anatomy, 1995. **8**(4): p. 252-255.

38. Spielmann, A.L., D.M. DeLong, and M.A. Kliewer, *Sonographic evaluation of spleen size in tall healthy athletes.* American Journal of Roentgenology, 2005. **184**(1): p. 45-49.

39. Prassopoulos, P., et al., *Determination of normal splenic volume on computed tomography in relation to age, gender and body habitus.* Eur Radiol, 1997. **7**(2): p. 246-8.

40. Bezerra, A.S., et al., *Determination of splenomegaly by CT: is there a place for a single measurement?* AJR Am J Roentgenol, 2005. **184**(5): p. 1510-3.

41. Linguraru, M.G., et al., *Assessing splenomegaly: automated volumetric analysis of the spleen.* Acad Radiol, 2013. **20**(6): p. 675-84.

42. Behrad, A. and H. Masoumi. *Automatic spleen segmentation in MRI images using a combined neural network and recursive watershed transform*. in *Neural Network Applications in Electrical Engineering (NEUREL), 2010 10th Symposium on*. 2010. IEEE.

43. Farraher, S.W., et al., *Liver and Spleen Volumetry with Quantitative MR Imaging and Dual-Space Clustering Segmentation 1.* Radiology, 2005. **237**(1): p. 322-328.

44. Wu, J., *An automated human organ segmentation technique for abdominal magnetic resonance images*. 2010: McMaster University.

45. Parker, J.R., *Algorithms for image processing and computer vision*. 2010: John Wiley & Sons.

46. Pham, D.L. and J.L. Prince, *Adaptive fuzzy segmentation of magnetic resonance images.* IEEE transactions on medical imaging, 1999. **18**(9): p. 737-752.

47. McInerney, T. and D. Terzopoulos, *Deformable models in medical image analysis: a survey.* Medical image analysis, 1996. **1**(2): p. 91-108.

48. Boykov, Y. and G. Funka-Lea, *Graph cuts and efficient ND image segmentation.* International journal of computer vision, 2006. **70**(2): p. 109-131.

49. Cootes, T.F., et al., *Active shape models-their training and application.* Computer vision and image understanding, 1995. **61**(1): p. 38-59.

50. Cootes, T.F., G.J. Edwards, and C.J. Taylor, *Active appearance models.* Ieee Transactions on Pattern Analysis and Machine Intelligence, 2001. **23**(6): p. 681-685.

51. Bezdek, J., L. Hall, and L. Clarke, *Review of MR image segmentation techniques using pattern recognition.* Medical Physics, 1993. **20**(4): p. 4.

52. Cocosco, C.A., A.P. Zijdenbos, and A.C. Evans, *A fully automatic and robust brain MRI tissue classification method.* Med Image Anal, 2003. **7**(4): p. 513-27.

53. Van Leemput, K., et al., *Automated model-based tissue classification of MR images of the brain.* IEEE Trans Med Imaging, 1999. **18**(10): p. 897-908.

54. Wells, W.M., et al., *Adaptive segmentation of MRI data.* IEEE Trans Med Imaging, 1996. **15**(4): p. 429-42.

55. Dale, A.M., B. Fischl, and M.I. Sereno, *Cortical surface-based analysis: I. Segmentation and surface reconstruction.* Neuroimage, 1999. **9**(2): p. 179-194.

56. Han, X., et al. *Graph-based topology correction for brain cortex segmentation*. in *Information Processing in Medical Imaging*. 2001. Springer.

57. Mangin, J.-F., et al., *From 3D magnetic resonance images to structural representations of the cortex topography using topology preserving deformations.* Journal of Mathematical Imaging and Vision, 1995. **5**(4): p. 297-318.

58. Han, X., et al., *CRUISE: cortical reconstruction using implicit surface evolution.* Neuroimage, 2004. **23**(3): p. 997-1012.

59. Fischl, B., *FreeSurfer.* Neuroimage, 2012. **62**(2): p. 774-81.

60. Stewart, J.C., et al., *Depressive symptom clusters and 5-year incidence of coronary artery calcification: the coronary artery risk development in young adults study.* Circulation, 2012. **126**(4): p. 410-7.

61. Barnes, J., et al., *Head size, age and gender adjustment in MRI studies: a necessary nuisance?* Neuroimage, 2010. **53**(4): p. 1244-55.

62. Perlaki, G., et al., *Are there any gender differences in the hippocampus volume after head-size correction? A volumetric and voxel-based morphometric study.* Neuroscience Letters, 2014. **570**: p. 119-123.

63. Whitwell, J.L., et al., *Normalization of cerebral volumes by use of intracranial volume: implications for longitudinal quantitative MR imaging.* American Journal of Neuroradiology, 2001. **22**(8): p. 1483-1489.

64. Farias, S.T., et al., *Maximal brain size remains an important predictor of cognition in old age, independent of current brain pathology.* Neurobiology of Aging, 2012. **33**(8): p. 1758-1768.

65. Peelle, J.E., R. Cusack, and R.N.A. Henson, *Adjusting for global effects in voxel-based morphometry: Gray matter decline in normal aging.* Neuroimage, 2012. **60**(2): p. 1503-1516.

66. Nordenskjold, R., et al., *Intracranial volume estimated with commonly used methods could introduce bias in studies including brain volume measurements.* Neuroimage, 2013. **83**: p. 355-360.

67. Westman, E., et al., *Regional Magnetic Resonance Imaging Measures for Multivariate Analysis in Alzheimer's Disease and Mild Cognitive Impairment.* Brain Topography, 2013. **26**(1): p. 9-23.

68. Smith, S.M., *Fast robust automated brain extraction.* Hum Brain Mapp, 2002. **17**(3): p. 143-55.

69. Davis, P. and E. Wright, *A new method for measuring cranial cavity volume and its application to the assessment of cerebral atrophy at autopsy.* Neuropathology and applied neurobiology, 1977. **3**(5): p.

341-358.

70. Perneczky, R., et al., *Head circumference, atrophy, and cognition: implications for brain reserve in Alzheimer disease.* Neurology, 2010. **75**(2): p. 137-42.

71. Commowick, O., S.K. Warfield, and G. Malandain, *Using Frankenstein's creature paradigm to build a patient specific atlas.* Med Image Comput Comput Assist Interv, 2009. **12**(Pt 2): p. 993-1000.

72. Liu, H. and H. Motoda, *Feature selection for knowledge discovery and data mining*. Vol. 454. 2012: Springer Science & Business Media.

73. Wu, X., et al., *Data mining with big data.* IEEE transactions on knowledge and data engineering, 2014. **26**(1): p. 97-107.

74. Powell, S., et al., *Registration and machine learning-based automated segmentation of subcortical and cerebellar brain structures.* Neuroimage, 2008. **39**(1): p. 238-47.

75. Morra, J.H., et al., *Comparison of AdaBoost and support vector machines for detecting Alzheimer's disease through automated hippocampal segmentation.* IEEE Trans Med Imaging, 2010. **29**(1): p. 30-43.

76. Hao, Y., et al., *Local label learning (LLL) for subcortical structure segmentation: application to hippocampus segmentation.* Hum Brain Mapp, 2014. **35**(6): p. 2674-97.

77. Han, X., *Learning-boosted label fusion for multi-atlas auto-segmentation*, in *Machine Learning in Medical Imaging*. 2013, Springer. p. 17-24.

78. Zikic, D., B. Glocker, and A. Criminisi, *Encoding atlases by randomized classification forests for efficient multi-atlas label propagation.* Med Image Anal, 2014. **18**(8): p. 1262-73.

79. Magnotta, V.A., et al., *Measurement of brain structures with artificial neural networks: two- and three-dimensional applications.* Radiology, 1999. **211**(3): p. 781-90.

80. Liao, S., Y. Gao, and D. Shen, *Sparse patch based prostate segmentation in CT images.* Med Image Comput Comput Assist Interv, 2012. **15**(Pt 3): p. 385-92.

81. Xu, Z., *Automatic Segmentation of the Human Abdomen on Clinically Acquired CT*. 2016, Vanderbilt University.

82. Huo, Y., et al. *Multi-atlas Segmentation Enables Robust Multi-contrast MRI Spleen Segmentation for Splenomegaly*. in *SPIE Medical Imaging*. 2017. International Society for Optics and Photonics.

83. Wu, M., et al., *Optimum template selection for atlas-based segmentation.* Neuroimage, 2007. **34**(4): p. 1612-8.

84. Guimond, A., J. Meunier, and J.-P. Thirion, *Average brain models: A convergence study.* Computer vision and image understanding, 2000. **77**(2): p. 192-210.

85. Gass, T., G. Székely, and O. Goksel, *Semi-supervised segmentation using multiple segmentation hypotheses from a single atlas*, in *Medical Computer Vision. Recognition Techniques and Applications in Medical Imaging*. 2013, Springer. p. 29-37.

86. Rohlfing, T., D.B. Russakoff, and C.R. Maurer, Jr., *Performance-based classifier combination in atlas-based image segmentation using expectation-maximization parameter estimation.* IEEE Trans Med Imaging, 2004. **23**(8): p. 983-94.

87. Heckemann, R.A., et al., *Automatic anatomical brain MRI segmentation combining label propagation and decision fusion.* Neuroimage, 2006. **33**(1): p. 115-26.

88. Avants, B.B., et al., *Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain.* Med Image Anal, 2008. **12**(1): p. 26-41.

89. Ourselin, S., et al., *Reconstructing a 3D structure from serial histological sections.* Image and Vision Computing, 2001. **19**(1-2): p. 25-31.

90. Klein, A., et al., *Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration.* Neuroimage, 2009. **46**(3): p. 786-802.

91. Wang, H.Z., et al., *Multi-Atlas Segmentation with Joint Label Fusion.* Ieee Transactions on Pattern Analysis and Machine Intelligence, 2013. **35**(3): p. 611-623.

92. Asman, A.J. and B.A. Landman, *Non-local statistical label fusion for multi-atlas segmentation.* Med Image Anal, 2013. **17**(2): p. 194-208.

93. Coupé, P., et al., *Patch-based segmentation using expert priors: Application to hippocampus and ventricle segmentation.* Neuroimage, 2011. **54**(2): p. 940-954.

94. Artaechevarria, X., A. Munoz-Barrutia, and C. Ortiz-de-Solorzano, *Combination strategies in multi-atlas image segmentation: application to brain MR data.* IEEE Trans Med Imaging, 2009. **28**(8): p. 1266-77.

95. Sabuncu, M.R., et al., *A generative model for image segmentation based on label fusion.* IEEE Trans Med Imaging, 2010. **29**(10): p. 1714-29.

96. Warfield, S.K., K.H. Zou, and W.M. Wells, *Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation.* IEEE Trans Med Imaging, 2004. **23**(7): p. 903-921.

97. Aljabar, P., et al., *Multi-atlas based segmentation of brain images: atlas selection and its effect on accuracy.* Neuroimage, 2009. **46**(3): p. 726-38.

98. Asman, A.J., A.S. Dagley, and B.A. Landman, *Statistical label fusion with hierarchical performance models.* Proc Soc Photo Opt Instrum Eng, 2014. **9034**: p. 90341E.

99. Isgum, I., et al., *Multi-atlas-based segmentation with local decision fusion--application to cardiac and aortic segmentation in CT scans.* IEEE Trans Med Imaging, 2009. **28**(7): p. 1000-10.

100. Rohlfing, T., et al., *Evaluation of atlas selection strategies for atlas-based image segmentation with application to confocal microscopy images of bee brains.* Neuroimage, 2004. **21**(4): p. 1428-42.

101. Langerak, T.R., et al., *Multiatlas-based segmentation with preregistration atlas selection.* Med Phys, 2013. **40**(9): p. 091701.

102. Langerak, T.R., et al., *Label fusion in atlas-based segmentation using a selective and iterative method*

for performance level estimation (SIMPLE). IEEE Trans Med Imaging, 2010. **29**(12): p. 2000-8.

103. Rousseau, F., P.A. Habas, and C. Studholme, *A supervised patch-based approach for human brain labeling.* IEEE Trans Med Imaging, 2011. **30**(10): p. 1852-62.

104. Wang, H. and P.A. Yushkevich, *Multi-atlas segmentation without registration: A supervoxel-based approach*, in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013*. 2013, Springer. p. 535-542.

105. MacDonald, D., et al., *Automated 3-D extraction of inner and outer surfaces of cerebral cortex from MRI.* Neuroimage, 2000. **12**(3): p. 340-56.

106. Dale, A.M., B. Fischl, and M.I. Sereno, *Cortical surface-based analysis. I. Segmentation and surface reconstruction.* Neuroimage, 1999. **9**(2): p. 179-94.

107. Xu, C., et al., *Reconstruction of the human cerebral cortex from magnetic resonance images.* IEEE Trans Med Imaging, 1999. **18**(6): p. 467-80.

108. Shattuck, D.W. and R.M. Leahy, *BrainSuite: an automated cortical surface identification tool.* Med Image Anal, 2002. **6**(2): p. 129-42.

109. Liu, T., et al., *Reconstruction of central cortical surface from brain MRI images: method and application.* Neuroimage, 2008. **40**(3): p. 991-1002.

110. Kim, J.S., et al., *Automated 3-D extraction and evaluation of the inner and outer cortical surfaces using a Laplacian map and partial volume effect classification.* Neuroimage, 2005. **27**(1): p. 210-21.

111. Fischl, B., M.I. Sereno, and A.M. Dale, *Cortical surface-based analysis. II: Inflation, flattening, and a surface-based coordinate system.* Neuroimage, 1999. **9**(2): p. 195-207.

112. Mathalon, D.H., et al., *Correction for head size in brain-imaging measurements.* Psychiatry Res, 1993. **50**(2): p. 121-39.

113. Ashburner, J. and K.J. Friston, *Unified segmentation.* Neuroimage, 2005. **26**(3): p. 839-851.

114. Ananth, H., et al., *Cortical and subcortical gray matter abnormalities in schizophrenia determined through structural magnetic resonance imaging with optimized volumetric voxel-based morphometry.* American Journal of Psychiatry, 2014.

115. Pengas, G., et al., *Comparative reliability of total intracranial volume estimation methods and the influence of atrophy in a longitudinal semantic dementia cohort.* J Neuroimaging, 2009. **19**(1): p. 37-46.

116. Buckner, R.L., et al., *A unified approach for morphometric and functional data analysis in young, old, and demented adults using automated atlas-based head size normalization: reliability and validation against manual measurement of total intracranial volume.* Neuroimage, 2004. **23**(2): p. 724-38.

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

118. Lemieux, L., et al., *Automatic segmentation of the brain and intracranial cerebrospinal fluid in T1-*

weighted volume MRI scans of the head, and its application to serial cerebral and intracranial volumetry. Magn Reson Med, 2003. **49**(5): p. 872-84.

119. Driscoll, I., et al., *Longitudinal pattern of regional brain volume change differentiates normal aging from MCI.* Neurology, 2009. **72**(22): p. 1906-13.

120. Keihaninejad, S., et al., *A robust method to estimate the intracranial volume across MRI field strengths (1.5T and 3T).* Neuroimage, 2010. **50**(4): p. 1427-37.

121. Aguilar, C., et al., *Automated CT-based segmentation and quantification of total intracranial volume.* Eur Radiol, 2015. **25**(11): p. 3151-60.

122. Hansen, T., et al., *How Does the Accuracy of Intracranial Volume Measurements Affect Normalized Brain Volumes? Sample Size Estimates Based on 966 Subjects from the HUNT MRI Cohort.* American Journal of Neuroradiology, 2015.

123. Safran, C., et al., *Toward a national framework for the secondary use of health data: an American Medical Informatics Association White Paper.* Journal of the American Medical Informatics Association, 2007. **14**(1): p. 1-9.

124. Murdoch, T.B. and A.S. Detsky, *The inevitable application of big data to health care.* JAMA, 2013. **309**(13): p. 1351-2.

125. Jiang, T., et al., *Multimodal magnetic resonance imaging for brain disorders: advances and perspectives.* Brain Imaging and Behavior, 2008. **2**(4): p. 249-257.

126. Van Horn, J.D. and A.W. Toga, *Multi-site neuroimaging trials.* Current opinion in neurology, 2009. **22**(4): p. 370.

127. Hall, D., et al., *Sharing heterogeneous data: the national database for autism research.* Neuroinformatics, 2012. **10**(4): p. 331-339.

128. Resnick, S.M., et al., *Longitudinal magnetic resonance imaging studies of older adults: a shrinking brain.* J Neurosci, 2003. **23**(8): p. 3295-301.

129. Thambisetty, M., et al., *Longitudinal changes in cortical thickness associated with normal aging.* Neuroimage, 2010. **52**(4): p. 1215-23.

130. Poldrack, R.A., et al., *Toward open sharing of task-based fMRI data: the OpenfMRI project.* Frontiers in neuroinformatics, 2013. **7**: p. 12.

131. Biswal, B.B., et al., *Toward discovery science of human brain function.* Proceedings of the National Academy of Sciences, 2010. **107**(10): p. 4734-4739.

132. Marcus, D.S., et al., *Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults.* J Cogn Neurosci, 2007. **19**(9): p. 1498-507.

133. Landman, B.A., et al., *Multi-parametric neuroimaging reproducibility: a 3-T resource study.* Neuroimage, 2011. **54**(4): p. 2854-66.

134. Eichner, E.R., *Splenic function: normal, too much and too little.* Am J Med, 1979. **66**(2): p. 311-20.

135. Yetter, E.M., et al., *Estimating splenic volume: sonographic measurements correlated with helical CT determination.* American Journal of Roentgenology, 2003. **181**(6): p. 1615-1620.

136. Lamb, P., et al., *Spleen size: how well do linear ultrasound measurements correlate with three-dimensional CT volume assessments?* The British journal of radiology, 2002. **75**(895): p. 573-577.

137. Hu, P., et al., *Automatic 3D liver segmentation based on deep learning and globally optimized surface evolution.* Physics in medicine and biology, 2016. **61**(24): p. 8676.

138. Huo, Y., et al. *Multi-atlas Segmentation Enables Robust Multi-contrast MRI Spleen Segmentation for Splenomegaly*. in *SPIE Medical Imaging*. 2017. International Society for Optics and Photonics.

139. Peng, C., et al., *Large Kernel Matters--Improve Semantic Segmentation by Global Convolutional Network.* arXiv preprint arXiv:1703.02719, 2017.

140. Luc, P., et al., *Semantic segmentation using adversarial networks.* arXiv preprint arXiv:1611.08408, 2016.

141. Kamnitsas, K., et al. *Unsupervised domain adaptation in brain lesion segmentation with adversarial networks*. in *International Conference on Information Processing in Medical Imaging*. 2017. Springer.

142. Roy, S., A. Carass, and J.L. Prince, *Magnetic resonance image example-based contrast synthesis.* IEEE transactions on medical imaging, 2013. **32**(12): p. 2348-2363.

143. Burgos, N., et al., *Attenuation correction synthesis for hybrid PET-MR scanners: application to brain studies.* IEEE transactions on medical imaging, 2014. **33**(12): p. 2332-2341.

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

145. Zhu, J.-Y., et al., *Unpaired image-to-image translation using cycle-consistent adversarial networks.* arXiv preprint arXiv:1703.10593, 2017.

146. Wolterink, J.M., et al. *Deep MR to CT Synthesis Using Unpaired Data*. in *International Workshop on Simulation and Synthesis in Medical Imaging*. 2017. Springer.

147. Zhao, C., et al. *A Supervoxel Based Random Forest Synthesis Framework for Bidirectional MR/CT Synthesis*. in *International Workshop on Simulation and Synthesis in Medical Imaging*. 2017. Springer.

148. Chartsias, A., et al. *Adversarial Image Synthesis for Unpaired Multi-modal Cardiac Data*. in *International Workshop on Simulation and Synthesis in Medical Imaging*. 2017. Springer.

149. Asman, A.J., et al., *Multi-atlas learner fusion: An efficient segmentation approach for large-scale data.* Med Image Anal, 2015. **26**(1): p. 82-91.

150. Huo, Y., et al. *Mapping Lifetime Brain Volumetry with Covariate-Adjusted Restricted Cubic Spline Regression from Cross-Sectional Multi-site MRI*. in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2016. Springer.

151. Huo, Y., et al., *Consistent cortical reconstruction and multi-atlas brain segmentation.* NeuroImage, 2016.

152. Huo, Y., et al. *Data-driven Probabilistic Atlases Capture Whole-brain Individual Variation*. in *1 st MICCAI Workshop on*. 2015.

153. Huo, Y., et al., *Simultaneous total intracranial volume and posterior fossa volume estimation using multi‐atlas label fusion.* Human Brain Mapping, 2016.

154. Huo, Y., S.M. Resnick, and B.A. Landman. *4D Multi-atlas Label Fusion using Longitudinal Images*. in *International Workshop on Patch-based Techniques in Medical Imaging*. 2017. Springer.

155. Huo, Y., et al., *Splenomegaly Segmentation using Global Convolutional Kernels and Conditional Generative Adversarial Networks.* arXiv preprint arXiv:1712.00542, 2017.

156. Huo, Y., et al., *Adversarial Synthesis Learning Enables Segmentation Without Target Modality Ground Truth.* arXiv preprint arXiv:1712.07695, 2017.

157. Huo, Y., et al., *Automated Characterization of Pyelocalyceal Anatomy Using CT Urograms to Aid in Management of Kidney Stones*, in *Computer Assisted and Robotic Endoscopy and Clinical Image-Based Procedures*. 2017, Springer. p. 99-107.

158. Crespo-Facorro, B., et al., *Human frontal cortex: an MRI-based parcellation method.* Neuroimage, 1999. **10**(5): p. 500-19.

159. Tsang, O., et al., *Comparison of tissue segmentation algorithms in neuroimage analysis software tools.* Conf Proc IEEE Eng Med Biol Soc, 2008. **2008**: p. 3924-8.

160. Wolz, R., et al., *LEAP: learning embeddings for atlas propagation.* Neuroimage, 2010. **49**(2): p. 1316-25.

161. Chakravarty, M.M., et al., *Performing label‐fusion‐based segmentation using multiple automatically generated templates.* Human brain mapping, 2013. **34**(10): p. 2635-2654.

162. Jia, H., P.T. Yap, and D. Shen, *Iterative multi-atlas-based multi-image segmentation with tree-based registration.* Neuroimage, 2012. **59**(1): p. 422-30.

163. Weisenfeld, N.I. and S.K. Warfield, *Learning likelihoods for labeling (L3): a general multi-classifier segmentation algorithm.* Med Image Comput Comput Assist Interv, 2011. **14**(Pt 3): p. 322-9.

164. Breiman, L., *Arcing classifier (with discussion and a rejoinder by the author).* The annals of statistics, 1998. **26**(3): p. 801-849.

165. Freund, Y. and R.E. Schapire. *A desicion-theoretic generalization of on-line learning and an application to boosting*. in *Computational learning theory*. 1995. Springer.

166. Biswal, B.B., et al., *Toward discovery science of human brain function.* Proc Natl Acad Sci U S A, 2010. **107**(10): p. 4734-9.

167. Kawas, C., et al., *A prospective study of estrogen replacement therapy and the risk of developing Alzheimer's disease: the Baltimore Longitudinal Study of Aging.* Neurology, 1997. **48**(6): p. 1517-21.

168. D'Haese, P.F., et al., *Cranial Vault and its CRAVE tools: A clinical computer assistance system for deep brain stimulation (DBS) therapy.* Med Image Anal, 2012. **16**(3): p. 744-753.

169. Tackett, J.L., et al., *Common genetic influences on negative emotionality and a general psychopathology factor in childhood and adolescence.* J Abnorm Psychol, 2013. **122**(4): p. 1142-53.

170. Klein, A., et al. *Open labels: online feedback for a public resource of manually labeled brain images.* in *16th Annual Meeting for the Organization of Human Brain Mapping*. 2010.

171. Evans, A.C., et al. *3D statistical neuroanatomical models from 305 MRI volumes*. in *Nuclear Science Symposium and Medical Imaging Conference, 1993., 1993 IEEE Conference Record.* 1993. IEEE.

172. Asman, A.J. and B.A. Landman, *Formulating spatially varying performance in the statistical fusion framework.* IEEE Trans Med Imaging, 2012. **31**(6): p. 1326-36.

173. Wang, H., et al., *A learning-based wrapper method to correct systematic errors in automatic image segmentation: consistently improved performance in hippocampus, cortex and brain segmentation.* Neuroimage, 2011. **55**(3): p. 968-85.

174. Dice, L.R., *Measures of the amount of ecologic association between species.* Ecology, 1945. **26**(3): p. 297-302.

175. Wilcoxon, F., *Individual comparisons by ranking methods.* Biometrics bulletin, 1945: p. 80-83.

176. Pearson, K., *LIII. On lines and planes of closest fit to systems of points in space.* The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 1901. **2**(11): p. 559-572.

177. Torgerson, W.S., *Theory and methods of scaling.* 1958.

178. Tenenbaum, J.B., V. de Silva, and J.C. Langford, *A global geometric framework for nonlinear dimensionality reduction.* Science, 2000. **290**(5500): p. 2319-+.

179. Belkin, M. and P. Niyogi, *Laplacian eigenmaps for dimensionality reduction and data representation.* Neural Computation, 2003. **15**(6): p. 1373-1396.

180. Roweis, S.T. and L.K. Saul, *Nonlinear dimensionality reduction by locally linear embedding.* Science, 2000. **290**(5500): p. 2323-+.

181. Wolz, R., et al., *Multi-method analysis of MRI images in early diagnostics of Alzheimer's disease.* PLoS One, 2011. **6**(10): p. e25446.

182. Aljabar, P., D. Rueckert, and W.R. Crum, *Automated morphological analysis of magnetic resonance brain imaging using spectral analysis.* Neuroimage, 2008. **43**(2): p. 225-35.

183. Gerber, S., et al., *Manifold modeling for brain population analysis.* Med Image Anal, 2010. **14**(5): p. 643-53.

184. Breiman, L., et al., *Classification and regression trees*. 1984: CRC press.

185. Quinlan, J.R., *C4. 5: Programs for Machine Learning.* 1993.

186. Hastie, T., et al., *The elements of statistical learning*. Vol. 2. 2009: Springer.

187. Criminisi, A., et al., *Regression forests for efficient anatomy detection and localization in computed tomography scans.* Med Image Anal, 2013. **17**(8): p. 1293-303.

188. Balafar, M.A., et al., *Review of brain MRI image segmentation methods.* Artificial Intelligence Review, 2010. **33**(3): p. 261-274.

189. Lim, S.P. and H. Haron, *Surface reconstruction techniques: a review.* Artificial Intelligence Review, 2014. **42**(1): p. 59-78.

190. Pham, D.L. and J.L. Prince, *Adaptive fuzzy segmentation of magnetic resonance images.* IEEE Trans Med Imaging, 1999. **18**(9): p. 737-52.

191. Keshavan, M.S., et al., *A comparison of stereology and segmentation techniques for volumetric measurements of lateral ventricles in magnetic resonance imaging.* Psychiatry Res, 1995. **61**(1): p. 53-60.

192. Brewer, J.B., et al., *Fully-automated quantification of regional brain volumes for improved detection of focal atrophy in Alzheimer disease.* AJNR Am J Neuroradiol, 2009. **30**(3): p. 578-80.

193. Fischl, B., et al., *Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain.* Neuron, 2002. **33**(3): p. 341-55.

194. Brewer, J.B., *Fully-automated volumetric MRI with normative ranges: translation to clinical practice.* Behav Neurol, 2009. **21**(1): p. 21-8.

195. Fischl, B. and A.M. Dale, *Measuring the thickness of the human cerebral cortex from magnetic resonance images.* Proc Natl Acad Sci U S A, 2000. **97**(20): p. 11050-5.

196. Han, X., et al., *Reliability of MRI-derived measurements of human cerebral cortical thickness: the effects of field strength, scanner upgrade and manufacturer.* Neuroimage, 2006. **32**(1): p. 180-94.

197. Winkler, A.M., et al., *Measuring and comparing brain cortical surface area and other areal quantities.* Neuroimage, 2012. **61**(4): p. 1428-43.

198. Fan, S.W., et al., *Quantitative MRI analysis of the surface area, signal intensity and MRI index of the central bright area for the evaluation of early adjacent disc degeneration after lumbar fusion.* Eur Spine J, 2012. **21**(9): p. 1709-15.

199. Symms, M., et al., *A review of structural magnetic resonance neuroimaging.* J Neurol Neurosurg Psychiatry, 2004. **75**(9): p. 1235-44.

200. Feczko, E., et al., *An MRI-based method for measuring volume, thickness and surface area of entorhinal, perirhinal, and posterior parahippocampal cortex.* Neurobiology of Aging, 2009. **30**(3): p. 420-31.

201. Cabezas, M., et al., *A review of atlas-based segmentation for magnetic resonance brain images.* Comput Methods Programs Biomed, 2011. **104**(3): p. e158-77.

202. Doan, N.T., J.O. de Xivry, and B. Macq. *Effect of inter-subject variation on the accuracy of atlas-based segmentation applied to human brain structures.* in *SPIE Medical Imaging*. 2010. International Society for Optics and Photonics.

203. Iglesias, J.E. and M.R. Sabuncu, *Multi-atlas segmentation of biomedical images: A survey.* Med Image Anal, 2015. **24**(1): p. 205-19.

204. Fischl, B., et al., *High-resolution intersubject averaging and a coordinate system for the cortical surface.* Human brain mapping, 1999. **8**(4): p. 272-284.

205. Lyttelton, O., et al., *An unbiased iterative group registration template for cortical surface analysis.* Neuroimage, 2007. **34**(4): p. 1535-1544.

206. Tosun, D. and J.L. Prince, *A Geometry-Driven Optical Flow Warping for Spatial Normalization of Cortical Surfaces.* IEEE Trans Med Imaging, 2008. **27**(12): p. 1739-1753.

207. Tosun, D., M.E. Rettmann, and J.L. Prince, *Mapping techniques for aligning sulci across multiple brains.* Med Image Anal, 2004. **8**(3): p. 295-309.

208. Yeo, B.T.T., et al., *Spherical Demons: Fast Diffeomorphic Landmark-Free Surface Registration.* IEEE Trans Med Imaging, 2010. **29**(3): p. 650-668.

209. Desikan, R.S., et al., *An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest.* Neuroimage, 2006. **31**(3): p. 968-80.

210. Destrieux, C., et al., *Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature.* Neuroimage, 2010. **53**(1): p. 1-15.

211. Fischl, B., et al., *Automatically parcellating the human cerebral cortex.* Cerebral Cortex, 2004. **14**(1): p. 11-22.

212. Thompson, P.M., et al., *Cortical change in Alzheimer's disease detected with a disease-specific population-based brain atlas.* Cerebral Cortex, 2001. **11**(1): p. 1-16.

213. Chung, M.K., et al., *Deformation-based surface morphometry applied to gray matter deformation.* Neuroimage, 2003. **18**(2): p. 198-213.

214. Fornito, A., et al., *Surface-based morphometry of the anterior cingulate cortex in first episode schizophrenia.* Human brain mapping, 2008. **29**(4): p. 478-489.

215. Fischl, B., et al., *Sequence-independent segmentation of magnetic resonance images.* Neuroimage, 2004. **23 Suppl 1**: p. S69-84.

216. Han, X. and B. Fischl, *Atlas renormalization for improved brain MR image segmentation across scanner platforms.* IEEE Trans Med Imaging, 2007. **26**(4): p. 479-86.

217. Fischl, B., et al., *Automatically parcellating the human cerebral cortex.* Cereb Cortex, 2004. **14**(1): p. 11-22.

218. Lehmann, M., et al., *Atrophy patterns in Alzheimer's disease and semantic dementia: a comparison of FreeSurfer and manual volumetric measurements.* Neuroimage, 2010. **49**(3): p. 2264-2274.

219. Cardinale, F., et al., *Validation of FreeSurfer-estimated brain cortical thickness: comparison with histologic measurements.* Neuroinformatics, 2014. **12**(4): p. 535-542.

220. Shiee, N., et al., *Reconstruction of the human cerebral cortex robust to white matter lesions: method*

and validation. Hum Brain Mapp, 2014. **35**(7): p. 3385-401.

221. Landman, B.A., et al., *System for integrated neuroimaging analysis and processing of structure.* Neuroinformatics, 2013. **11**(1): p. 91-103.

222. Han, X., et al., *Topology correction in brain cortex segmentation using a multiscale, graph-based algorithm.* IEEE Trans Med Imaging, 2002. **21**(2): p. 109-21.

223. Bazin, P.L. and D.L. Pham, *Topology-preserving tissue classification of magnetic resonance brain images.* IEEE Trans Med Imaging, 2007. **26**(4): p. 487-96.

224. Bazin, P.L. and D.L. Pham, *Homeomorphic brain image segmentation with topological and statistical atlases.* Med Image Anal, 2008. **12**(5): p. 616-25.

225. Huo, Y., et al. *Combining multi-atlas segmentation with brain surface estimation*. in *SPIE Medical Imaging*. 2016. International Society for Optics and Photonics.

226. Tustison, N.J., et al., *N4ITK: improved N3 bias correction.* IEEE Trans Med Imaging, 2010. **29**(6): p. 1310-20.

227. Carass, A., et al., *Simple paradigm for extra-cerebral tissue removal: algorithm and analysis.* Neuroimage, 2011. **56**(4): p. 1982-92.

228. Landman, B. and S. Warfield. *MICCAI 2012 workshop on multi-atlas labeling*. in *Medical Image Computing and Computer Assisted Intervention Conference 2012: MICCAI 2012 Grand Challenge and Workshop on Multi-Atlas Labeling Challenge Results*. 2012.

229. Zeng, X., et al., *Segmentation and measurement of the cortex from 3-D MR images using coupled-surfaces propagation.* IEEE Trans Med Imaging, 1999. **18**(10): p. 927-37.

230. Sethian, J.A., *Level set methods and fast marching methods: evolving interfaces in computational geometry, fluid mechanics, computer vision, and materials science*. Vol. 3. 1999: Cambridge university press.

231. Xu, C. and J.L. Prince, *Snakes, shapes, and gradient vector flow.* IEEE Trans Image Process, 1998. **7**(3): p. 359-69.

232. Osher, S. and J.A. Sethian, *Fronts propagating with curvature-dependent speed: algorithms based on Hamilton-Jacobi formulations.* Journal of computational physics, 1988. **79**(1): p. 12-49.

233. Osher, S. and R. Fedkiw, *Level set methods and dynamic implicit surfaces*. Vol. 153. 2006: Springer Science & Business Media.

234. Plassard, A.J., et al., *Evaluation of Atlas-Based White Matter Segmentation with Eve.* Proc SPIE Int Soc Opt Eng, 2015. **9413**.

235. Shiee, N., et al., *A topology-preserving approach to the segmentation of brain images with multiple sclerosis lesions.* Neuroimage, 2010. **49**(2): p. 1524-35.

236. Cohen, J., *Statistical power analysis for the behavioral sciences*. 1977: Lawrence Erlbaum Associates, Inc.

237. Shock, N.W., et al., *Normal human aging: The Baltimore longitudinal study of aging.* NIH Publication 1984: p. No. 84-2450.

238. Li, B., F. Bryan, and B.A. Landman, *Next Generation of the Java Image Science Toolkit (JIST): Visualization and Validation.* Insight J, 2012. **2012**: p. 1-16.

239. Lucas, B.C., et al., *The Java Image Science Toolkit (JIST) for rapid prototyping and publishing of neuroimaging software.* Neuroinformatics, 2010. **8**(1): p. 5-17.

240. Tustison, N.J., et al., *Large-scale evaluation of ANTs and FreeSurfer cortical thickness measurements.* Neuroimage, 2014. **99**: p. 166-79.

241. Plassard, A.J., et al. *On the fallacy of quantitative segmentation for T1-weighted MRI*. in *SPIE Medical Imaging*. 2016. International Society for Optics and Photonics.

242. Jog, A., et al., *MR image synthesis by contrast learning on neighborhood ensembles.* Medical image analysis, 2015. **24**(1): p. 63-76.

243. Huo, Y., et al., *Consistent cortical reconstruction and multi-atlas brain segmentation.* NeuroImage, 2016. **138**: p. 197-210.

244. Thambisetty, M., et al., *Longitudinal changes in cortical thickness associated with normal aging.* Neuroimage, 2010. **52**(4): p. 1215-1223.

245. Asman, A.J., et al., *Multi-atlas learner fusion: An efficient segmentation approach for large-scale data.* Medical image analysis, 2015. **26**(1): p. 82-91.

246. Cachia, A., et al., *A generic framework for the parcellation of the cortical surface into gyri using geodesic Voronoı diagrams.* Medical Image Analysis, 2003. **7**(4): p. 403-416.

247. Fischl, B., et al., *Automatically parcellating the human cerebral cortex.* Cerebral cortex, 2004. **14**(1): p. 11-22.

248. Van Essen, D.C., et al., *Parcellations and hemispheric asymmetries of human cerebral cortex analyzed on surface-based atlases.* Cerebral cortex, 2011. **22**(10): p. 2241-2262.

249. Parvathaneni, P., et al. *Gray Matter Surface Based Spatial Statistics (GS-BSS) in Diffusion Microstructure*. in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2017. Springer.

250. Landman, B.A., et al., *Multi-parametric neuroimaging reproducibility: a 3-T resource study.* Neuroimage, 2011. **54**(4): p. 2854-2866.

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

252. Heimann, T. and H.P. Meinzer, *Statistical shape models for 3D medical image segmentation: a review.* Med Image Anal, 2009. **13**(4): p. 543-63.

253. Frey, B.J. and D. Dueck, *Clustering by passing messages between data points.* Science, 2007. **315**(5814): p. 972-976.

254. Gouttard, S., et al., *Assessment of reliability of multi-site neuroimaging via traveling phantom study.* Med Image Comput Comput Assist Interv, 2008. **11**(Pt 2): p. 263-70.

255. Hartley, S.W., et al., *Analysis and validation of automated skull stripping tools: a validation study based on 296 MR images from the Honolulu Asia aging study.* Neuroimage, 2006. **30**(4): p. 1179-86.

256. Fein, G., et al., *Controlling for premorbid brain size in imaging studies: T1-derived cranium scaling factor vs. T2-derived intracranial vault volume.* Psychiatry Res, 2004. **131**(2): p. 169-76.

257. Smith, S.M., et al., *Accurate, robust, and automated longitudinal and cross-sectional brain change analysis.* Neuroimage, 2002. **17**(1): p. 479-89.

258. Boyes, R.G., et al., *Cerebral atrophy measurements using Jacobian integration: comparison with the boundary shift integral.* Neuroimage, 2006. **32**(1): p. 159-169.

259. Weiskopf, N., et al., *Unified segmentation based correction of R1 brain maps for RF transmit field inhomogeneities (UNICORT).* Neuroimage, 2011. **54**(3): p. 2116-24.

260. Ridgway, G., et al., *Estimation of total intracranial volume; a comparison of methods.* Alzheimer's & Dementia, 2011. **7**(4): p. S62-S63.

261. Malone, I.B., et al., *Accurate automatic estimation of total intracranial volume: a nuisance variable with less nuisance.* Neuroimage, 2015. **104**: p. 366-72.

262. Badie, B., D. Mendoza, and U. Batzdorf, *Posterior fossa volume and response to suboccipital decompression in patients with Chiari I malformation.* Neurosurgery, 1995. **37**(2): p. 214-8.

263. Nyland, H. and K.G. Krogness, *Size of posterior fossa in Chiari type 1 malformation in adults.* Acta Neurochir (Wien), 1978. **40**(3-4): p. 233-42.

264. Sgouros, S., M. Kountouri, and K. Natarajan, *Posterior fossa volume in children with Chiari malformation Type I.* J Neurosurg, 2006. **105**(2 Suppl): p. 101-6.

265. Schaerer, J., et al., *Accurate intracranial cavity volume estimation using multiatlas segmentation.* Alzheimer's & Dementia, 2012. **8**(4): p. P272.

266. Van Leemput, K. and M.R. Sabuncu. *A cautionary analysis of staple using direct inference of segmentation truth*. in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2014. Springer.

267. Rohlfing, T., D.B. Russakoff, and C.R. Maurer, *Expectation maximization strategies for multi-atlas multi-label segmentation.* Inf Process Med Imaging, 2003. **18**: p. 210-21.

268. Rohlfing, T., D.B. Russakoff, and C.R. Maurer Jr, *Extraction and application of expert priors to combine multiple segmentations of human brain tissue*, in *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2003*. 2003, Springer. p. 578-585.

269. Commowick, O. and S.K. Warfield, *Incorporating priors on expert performance parameters for segmentation validation and label fusion: a maximum a posteriori STAPLE*, in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2010*. 2010, Springer. p. 25-32.

270. Landman, B.A., et al., *Robust statistical fusion of image labels.* IEEE Trans Med Imaging, 2012.

31(2): p. 512-22.

271. Asman, A.J. and B. Landman, *Robust statistical label fusion through consensus level, labeler accuracy, and truth estimation (COLLATE).* Medical Imaging, IEEE Transactions on, 2011. **30**(10): p. 1779-1794.

272. Asman, A.J. and B.A. Landman, *Hierarchical performance estimation in the statistical label fusion framework.* Med Image Anal, 2014. **18**(7): p. 1070-81.

273. Akhondi-Asl, A., et al., *A logarithmic opinion pool based STAPLE algorithm for the fusion of segmentations with associated reliability weights.* IEEE Trans Med Imaging, 2014. **33**(10): p. 1997-2009.

274. Shen, D., et al., *EditorialMachine Learning and Data Mining in Medical Imaging.* Biomedical and Health Informatics, IEEE Journal of, 2015. **19**(5): p. 1587-1588.

275. Huo, Y., et al., *Data-driven probabilistic atlases capture whole-brain individual variation*, in *In: Proceedings of the 1st Miccai 2015 Workshop on Management and Processing of images for Population Imaging-MICCAI-MAPPING2015, C. Barillot, M. Dojat, D. Kennedy and W. Niessen.* 2015. p. 7.

276. Huo, Y., et al., *Combining Multi-atlas Segmentation with Brain Surface Estimation.* Proc SPIE Int Soc Opt Eng, 2016. **9784**.

277. Panda, S., et al., *Robust Optic Nerve Segmentation on Clinically Acquired CT.* Proc SPIE Int Soc Opt Eng, 2014. **9034**: p. 90341G.

278. Harrigana, R.L., et al., *Robust optic nerve segmentation on clinically acquired CT.*

279. Harrigan, R.L., et al. *Constructing a statistical atlas of the radii of the optic nerve and cerebrospinal fluid sheath in young healthy adults*. in *SPIE Medical Imaging*. 2015. International Society for Optics and Photonics.

280. Harrigan, R.L., et al. *Short term reproducibility of a high contrast 3-D isotropic optic nerve imaging sequence in healthy controls*. in *SPIE Medical Imaging*. 2016. International Society for Optics and Photonics.

281. Asman, A.J., et al., *Groupwise multi-atlas segmentation of the spinal cord's internal structure.* Med Image Anal, 2014. **18**(3): p. 460-71.

282. Feeman, T.G., *The mathematics of medical imaging: a beginner's guide*. 2010: Springer Science & Business Media.

283. Dempster, A.P., N.M. Laird, and D.B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm.* Journal of the royal statistical society. Series B (methodological), 1977: p. 1-38.

284. McLachlan, G. and T. Krishnan, *The EM algorithm and extensions*. Vol. 382. 2007: John Wiley & Sons.

285. Bellman, R., *Dynamic Programming and Lagrange Multipliers.* Proc Natl Acad Sci U S A, 1956. **42**(10): p. 767-9.

286. Sjolund, J., et al. *Skull Segmentation in MRI by a Support Vector Machine Combining Local and Global Features*. in *Pattern Recognition (ICPR), 2014 22nd International Conference on*. 2014. IEEE.

287. Segonne, F., et al., *A hybrid approach to the skull stripping problem in MRI.* Neuroimage, 2004. **22**(3): p. 1060-75.

288. Han, X., C. Xu, and J.L. Prince, *A topology preserving level set method for geometric deformable models.* Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2003. **25**(6): p. 755-768.

289. Shrout, P.E. and J.L. Fleiss, *Intraclass correlations: uses in assessing rater reliability.* Psychol Bull, 1979. **86**(2): p. 420-8.

290. Torrado-Carvajal, A., et al., *Multi-atlas and label fusion approach for patient-specific MRI based skull estimation.* Magn Reson Med, 2016. **75**(4): p. 1797-807.

291. Durrleman, S. and R. Simon, *Flexible regression models with cubic splines.* Stat Med, 1989. **8**(5): p. 551-61.

292. Harrell, F., *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. 2015: Springer.

293. Stone, C.J. and C.-Y. Koo. *Additive splines in statistics*. 1986.

294. Anderberg, M.R., *Cluster analysis for applications: probability and mathematical statistics: a series of monographs and textbooks*. Vol. 19. 2014: Academic press.

295. Efron, B. and R.J. Tibshirani, *An introduction to the bootstrap*. 1994: CRC press.

296. Roy, S., et al., *Temporal filtering of longitudinal brain magnetic resonance images for consistent segmentation.* NeuroImage: Clinical, 2016. **11**: p. 264-275.

297. Xue, Z., D. Shen, and C. Davatzikos, *CLASSIC: consistent longitudinal alignment and segmentation for serial image computing.* Neuroimage, 2006. **30**(2): p. 388-99.

298. Reuter, M., et al., *Within-subject template estimation for unbiased longitudinal image analysis.* Neuroimage, 2012. **61**(4): p. 1402-18.

299. Pham, D.L., *Spatial models for fuzzy clustering.* Computer Vision and Image Understanding, 2001. **84**(2): p. 285-297.

300. Wolz, R., et al., *Measurement of hippocampal atrophy using 4D graph-cut segmentation: application to ADNI.* Neuroimage, 2010. **52**(1): p. 109-118.

301. Li, G., et al. *Multi-atlas based simultaneous labeling of longitudinal dynamic cortical surfaces in infants*. in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2013. Springer.

302. Guo, Y., et al. *Segmentation of infant hippocampus using common feature representations learned for multimodal longitudinal data*. in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2015. Springer.

303. Wang, L., et al. *Consistent Multi-Atlas Hippocampus Segmentation for Longitudinal MR Brain Images with Temporal Sparse Representation*. in *International Workshop on Patch-based Techniques in Medical Imaging*. 2016. Springer.

304. Fair, D.A., et al., *Atypical default network connectivity in youth with attention-deficit/hyperactivity disorder.* Biological psychiatry, 2010. **68**(12): p. 1084-1091.

305. McCormick, P.A. and K.M. Murphy, *Splenomegaly, hypersplenism and coagulation abnormalities in liver disease.* Baillieres Best Pract Res Clin Gastroenterol, 2000. **14**(6): p. 1009-31.

306. Klein, B., et al., *Splenomegaly and solitary spleen metastasis in solid tumors.* Cancer, 1987. **60**(1): p. 100-2.

307. Woodruff, A.W., *Mechanisms involved in anaemia associated with infection and splenomegaly in the tropics.* Trans R Soc Trop Med Hyg, 1973. **67**(3): p. 313-28.

308. Hosey, R.G., et al., *Ultrasound assessment of spleen size in collegiate athletes.* Br J Sports Med, 2006. **40**(3): p. 251-4; discussion 251-4.

309. Bezerra, A.S., et al., *Determination of splenomegaly by CT: is there a place for a single measurement?* American Journal of Roentgenology, 2005. **184**(5): p. 1510-1513.

310. Thomsen, C., et al., *Determination of T1- and T2-relaxation times in the spleen of patients with splenomegaly.* Magn Reson Imaging, 1990. **8**(1): p. 39-42.

311. Mazonakis, M., et al., *Estimation of spleen volume using MR imaging and a random marking technique.* European Radiology, 2000. **10**(12): p. 1899-1903.

312. Mihaylova, A. and V. Georgieva, *A Brief Survey of Spleen Segmentation in MRI and CT Images.* International Journal, 2016. **5**(7).

313. Campadelli, P., E. Casiraghi, and S. Pratissoli, *A segmentation framework for abdominal organs from CT scans.* Artif Intell Med, 2010. **50**(1): p. 3-11.

314. Campadelli, P., et al., *Automatic abdominal organ segmentation from CT images.* ELCVIA: electronic letters on computer vision and image analysis, 2009. **8**(1): p. 001-14.

315. Chen, X., et al., *Medical image segmentation by combining graph cuts and oriented active appearance models.* IEEE Trans Image Process, 2012. **21**(4): p. 2035-46.

316. Linguraru, M.G., et al., *Automated segmentation and quantification of liver and spleen from CT images using normalized probabilistic atlases and enhancement estimation.* Med Phys, 2010. **37**(2): p. 771-783.

317. Lay, N., et al. *Rapid multi-organ segmentation using context integration and discriminative models*. in *International Conference on Information Processing in Medical Imaging*. 2013. Springer.

318. Heinrich, M.P. and M. Blendowski. *Multi-organ segmentation using vantage point forests and binary context features*. in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2016. Springer.

319. Lin, E.C. *Radiation risk from medical imaging*. in *Mayo Clinic Proceedings*. 2010. Elsevier.

320. Semelka, R.C., et al., *Imaging strategies to reduce the risk of radiation in CT studies, including selective substitution with MRI.* Journal of Magnetic Resonance Imaging, 2007. **25**(5): p. 900-909.

321. Pauly, O., et al. *Fast multiple organ detection and localization in whole-body MR Dixon sequences*. in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2011. Springer.

322. Xu, Z., et al., *Efficient multi-atlas abdominal segmentation on clinically acquired CT with SIMPLE context learning.* Med Image Anal, 2015. **24**(1): p. 18-27.

323. Xu, Z., et al., *Improving Spleen Volume Estimation Via Computer-assisted Segmentation on Clinically Acquired CT Scans.* Acad Radiol, 2016. **23**(10): p. 1214-20.

324. Huo, Y., et al. *Multi-atlas segmentation enables robust multi-contrast MRI spleen segmentation for splenomegaly*. in *SPIE Medical Imaging*. 2017. International Society for Optics and Photonics.

325. Liu, J., et al. *Multi-Atlas Spleen Segmentation on CT Using Adaptive Context Learning*. in *SPIE Medical Imaging*. 2017. International Society for Optics and Photonics.

326. Schreibmann, E., D.M. Marcus, and T. Fox, *Multiatlas segmentation of thoracic and abdominal anatomy with level set ‐based local search.* Journal of Applied Clinical Medical Physics, 2014. **15**(4): p. 22-38.

327. Tustison, N.J., et al., *N4ITK: improved N3 bias correction.* IEEE transactions on medical imaging, 2010. **29**(6): p. 1310-1320.

328. Jenkinson, M. and S. Smith, *A global optimisation method for robust affine registration of brain images.* Medical image analysis, 2001. **5**(2): p. 143-156.

329. Heinrich, M.P., et al., *MRF-based deformable registration and ventilation estimation of lung CT.* IEEE Trans Med Imaging, 2013. **32**(7): p. 1239-48.

330. Xu, Z., et al., *Evaluation of six registration methods for the human abdomen on clinically acquired CT.* IEEE Transactions on Biomedical Engineering, 2016. **63**(8): p. 1563-1572.

331. Song, Z., et al. *Integrated graph cuts for brain MRI segmentation*. in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2006. Springer.

332. Wolz, R., et al., *Automated abdominal multi-organ segmentation with subject-specific atlas generation.* IEEE transactions on medical imaging, 2013. **32**(9): p. 1723-1730.

333. Jimenez-del-Toro, O., et al., *Cloud-based evaluation of anatomical structure segmentation and landmark detection algorithms: VISCERAL anatomy benchmarks.* IEEE transactions on medical imaging, 2016. **35**(11): p. 2459-2475.

334. Xu, Z., et al., *Efficient multi-atlas abdominal segmentation on clinically acquired CT with SIMPLE context learning.* Medical image analysis, 2015. **24**(1): p. 18-27.

335. Huo, Y., et al., *Robust Multi-contrast MRI Spleen Segmentation for Splenomegaly using Multi-atlas Segmentation.* IEEE Transactions on Biomedical Engineering, 2017.

336. Isola, P., et al., *Image-to-image translation with conditional adversarial networks.* arXiv preprint arXiv:1611.07004, 2016.

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

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

339. Kingma, D. and J. Ba, *Adam: A method for stochastic optimization.* arXiv preprint arXiv:1412.6980, 2014.

340. Deng, J., et al. *Imagenet: A large-scale hierarchical image database*. in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. 2009. IEEE.

341. Zhang, Z., et al., *MDNet: A Semantically and Visually Interpretable Medical Image Diagnosis Network.* arXiv preprint arXiv:1707.02485, 2017.

342. Huo, Y., et al., *Simultaneous total intracranial volume and posterior fossa volume estimation using multi‐atlas label fusion.* Human brain mapping, 2017. **38**(2): p. 599-616.

343. Huo, Y., et al., *Robust Multi-contrast MRI Spleen Segmentation for Splenomegaly using Multi-atlas Segmentation.* IEEE Transactions on Biomedical Engineering, 2017. **(In Press)**.

344. Bobo, M.F., et al. *Fully Convolutional Neural Networks Improve Abdominal Organ Segmentation*. in *SPIE Medical Imaging*. 2018. International Society for Optics and Photonics.

345. Johnson, J., A. Alahi, and L. Fei-Fei. *Perceptual losses for real-time style transfer and super-resolution*. in *European Conference on Computer Vision*. 2016. Springer.

346. Chew, B.H., et al., *Natural history, complications and re-intervention rates of asymptomatic residual stone fragments after ureteroscopy: a report from the EDGE Research Consortium.* The Journal of urology, 2016. **195**(4): p. 982-986.

347. Zomorrodi, A., A. Buhluli, and S. Fathi, *Anatomy of the collecting system of lower pole of the kidney in patients with a single renal stone: a comparative study with individuals with normal kidneys.* Saudi Journal of Kidney Diseases and Transplantation, 2010. **21**(4): p. 666.