CHARACTERIZING BRAIN AND BODY CONNECTIONS THROUGH DATA-EFFICIENT MEDICAL IMAGE SEGMENTATION

By

Qi Yang

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

Computer Science

December 16 2023

Nashville, Tennessee

Approved:

Professor Bennett A. Landman, Ph.D.

Professor Yuankai Huo, Ph.D.

Professor Ipek Oguz, Ph.D.

Professor Kurt G Schilling, Ph.D.

Professor Ann Zenobia Moore, Ph.D.

To my parents,

To my girlfriend,

To my cats

ACKNOWLEDGMENTS

My journey as a graduate student pursuing a Ph.D. degree with the MASI Lab at Vanderbilt University has been both a personal and collaborative adventure. First and foremost, I would like to express my heartfelt gratitude to my advisor, Dr. Bennett A. Landman. Your timely responses to every email and your adept ability to swiftly identify and correct errors in slide content that I had overlooked have been invaluable. Despite your frequent disapproval, working under your guidance has been exceptionally rewarding. You have significantly enhanced my skills in proposing valuable questions, crafting figures, writing academic manuscripts, and delivering oral presentations. I am deeply appreciative of your unwavering support during my challenging job search journey. I extend my thanks to Dr. Kurt G. Schilling for patiently answering what I feared were foolish questions, especially when I was new to diffusion MRI. My gratitude also goes to Dr. Yuankai Huo, who has provided excellent feedback to make my research more impactful. I wish I could spend more time discussing my work with you. Dr. Ann Zenobia Moore possesses hands-on knowledge about BLSA and has offered a plethora of suggestions for my research projects, for which I am grateful. I also thank Dr. Ipek Oguz, who introduced me to open-source medical image computing tools and gave me valuable suggestions to shape my research vision.

I wish to extend heartfelt acknowledgments to my lab mates during my Ph.D. journey. To the individuals in the MASI lab: Dr. Shunxing Bao, your patience and willingness to discuss both life and academic questions have been a source of solace and insight. Dr. Yucheng Tang, thank you for introducing me to the MASI lab and for your care and assistance when I first arrived in the United States. Dr. Riqiang Gao, your teachings have imparted many life lessons, making my journey smoother. Dr. Ho Hin Lee, your patience in repeatedly answering my queries and providing steadfast support is greatly appreciated. Dr. Leon Y. Cai, your detailed suggestions on manuscript readings have always been a boon. And Lucas, your conversations have brought me ease in stressful times. I sincerely thank all my colleagues in MASI for creating a friendly atmosphere in the lab. A warm thank you to all my friends who have supported me through hard times and brought joy into my life.

I would like to extend my deepest thanks to my girlfriend, Xin Yu, for her unwavering support throughout this journey. Together, we discussed academic questions, faced deadlines, and explored the beauty of life in our spare time with three lovely ragdolls: Disney(sleeping for a long time), August, and Yoyo. You have opened the door to another world for me, and together, starting from Tennessee, we will leave our footprints in every state of the United States!

Finally, I want to thank my parents, Zhimao Yang and Faping Chen. They always provided me with unconditional love and support throughout my life. They are the ones who always try their best to solve my

problems. I know that I can always call them first whenever the time zone I am. I don't think any language can accurately express my gratitude towards them at this time.

TABLE OF CONTENTS

Page

LIST OF TABLES . . . . . **ix**

LIST OF FIGURES . . . . . **x**

1 Introduction **1**

1.1 Overview 1
1.2 Structural MRI and diffusion MRI for brain 2
1.2.1 Structural MRI for brain 2
1.2.2 Diffusion MRI for brain 2
1.3 Single slice CT for thigh 3
1.4 Intra-modality medical image segmentation 5
1.4.1 Atlas-based segmentation 5
1.4.2 Learning-based segmentation 6
1.4.3 Transfer learning 7
1.5 Inter-modality medical image segmentation 8
1.6 Contributed Works 9
1.6.1 Contribution 1: The white matter pathway atlas for structure MRI 9
1.6.2 Contribution 2: Subject-specific segmentation of white matter pathway based on deep learning 10
1.6.3 Contribution 3: Label-efficient thigh segmentation based on transfer learning . . . 10
1.6.4 Contribution 4: Single slice CT muscle group segmentation with domain adaptation and self-training 10
1.6.5 Contribution 5: Characterize brain and body connection through linear and non-linear model 11

2 The white matter pathway atlas for structural T1w MRI image **12**

2.1 Introduction 12
2.2 Related works 12
2.3 Method 14
2.3.1 Data 15
2.3.2 Subject-level processing: tractography 16
2.3.3 Volumetric atlas creation 18
2.3.4 Suface-intersection atlas creation 18
2.3.5 Data visualization and validation 18
2.4 Technique validation 19
2.5 Discussion 20

3 Subject-specific segmentation of white matter based on deep learning **23**

3.1 Introduction 23
3.2 Related work 23
3.3 Materials and Methods 24
3.3.1 Data 26
3.3.2 Tractography 26
3.3.3 Registration and intensity normalization 27
3.3.4 Patch-wise network 27

3.3.5 Implementation details . . . 28
3.3.6 Baseline methods . . . 28
3.3.7 Metrics . . . 29
3.4 Results . . . 30
3.4.1 Fine-tune binary threshold . . . 30
3.4.2 Qualitative results . . . 31
3.4.3 Quantitative results . . . 32
3.5 Discussion . . . 35
3.5.1 Generalize to external dataset . . . 35
3.5.2 Limitations . . . 36
3.6 Conclusion . . . 36

4 Label-efficient thigh segmentation based on transfer learning . . . 37

4.1 Introduction . . . 37
4.2 Related work . . . 37
4.3 Methods . . . 38
4.3.1 Preprocessing . . . 39
4.3.2 Create a pseudo label for thigh . . . 39
4.3.3 Two stage training . . . 39
4.3.4 Data distribution . . . 40
4.3.5 Implementation details . . . 41
4.3.6 Baseline methods and metrics . . . 41
4.4 Experimental Results . . . 42
4.5 Conclusion and Discussion . . . 47

5 Single Slice Thigh CT Muscle Group Segmentation with Domain Adaptation and Self-Training 49

5.1 Introduction . . . 49
5.2 Material and method . . . 50
5.2.1 Data and preprocessing . . . 50
5.2.2 Train segmentation module from scratch . . . 52
5.2.3 Fine tune segmentation module in self training . . . 54
5.3 Experimental Results . . . 55
5.3.1 Implementation details and evaluation metrics . . . 55
5.3.2 Qualitative and quantitative results . . . 55
5.3.3 Ablation Study . . . 56
5.3.4 Sensitivity analysis . . . 57
5.4 Discussion . . . 59
5.5 Conclusion . . . 60

6 Characterize brain and body connection through linear and nonlinear model . . . 61

6.1 Introduction . . . 61
6.2 Materials and Method . . . 62
6.2.1 Brain and body feature extraction . . . 62
6.2.2 Selection layer . . . 62
6.2.3 Model Architecture . . . 64
6.3 Experiment and result . . . 65
6.3.1 Implementation details . . . 65
6.3.2 Characterize linear relationship . . . 66
6.3.3 Metrics . . . 66
6.3.4 Validation of the Gumbel softmax . . . 67
6.3.5 Choice of number of selecting features . . . 68
6.3.6 Real dataset . . . 69

6.3.7 Predict body composition area using brain region volumes . . . 69
6.3.8 Predict hippocampus volume using body composition area . . . 71
6.4 Discussion and conclusion . . . 73

7 Conclusion and Future Works . . . 75

7.1 Impact of the Dissertation . . . 75
7.2 Future works . . . 76
7.2.1 Slice to volume generation through conditional score-based diffusion model . . . 76
7.2.2 Leveraging large language model for text generation for characterization brain and body . . . 77

References . . . 78

A Copyright from Publishers . . . 91

A.1 Copyright from Elsvier . . . 91
A.2 Copyright from Wiley . . . 92
A.3 Copyright from SPIE . . . 93

LIST OF TABLES

Table		Page
1.1	The approximate HU values for common tissues.	5
2.1	Meta data information.	16
3.1	Dataset descriptions. * represents one typical case selected from the VU dataset.	26
3.2	The tractography algorithms and corresponding generated pathways.	26
3.3	number of scans for the training, validation, testing cohorts and external dataset.	28
3.4	Corner coordinates of pre-trained nice models out of 125 models, indexed starting at one.	28
4.1	The number of slices, thighs, lower legs, labeled thighs and lower legs for the cohort.	41
4.2	The mean DSC for each tissue of each method for thigh CT image. The highest result is bolded. The * means the method is significantly different from U-Net++ in the second stage (p-value < 0.05, Wilcoxon signed-rank test).	43
4.3	The mean DSC for each tissue of each method for lower leg CT image. The highest result is bolded. The * means the method is significantly different from U-Net++ in the second stage (p-value < 0.05, Wilcoxon signed-rank test).	45
5.1	Data distribution and image information for the whole pipeline.	52
5.2	The mean DSC and standard deviation for each muscle group and average performance	55
6.1	The explain variance score for linear and non-linear model. * represents p-value < 0.05 and indicates transformed brain features are related to the target variable significantly. The bold explain variance score means better prediction ability.	70
6.2	The EVS for linear and non-linear model based on male and female. * represents p-value < 0.05 and indicates transformed brain features are related to the target variable significantly. The bold explain variance score means better prediction ability.	71
6.3	The explain variance score for linear and non-linear model for left and right hippocampus volume prediction based on body composition areas. * represents p-value < 0.05 and indicates transformed body features are related to the target variable significantly. The bold explain variance score means better prediction ability.	73

LIST OF FIGURES

Figure Page

1.1 The triplane view of a T1w brain. There are sagittal, coronal, and axonal views of brain images. . . . 3
1.2 This is one of the classic pipelines of getting white matter region from dMRI. Diffusion tensor is derived from dMRI. Then the random seeds are used to track whole brain tractography. Starting from each voxel, the next point is like a continuation of the current one. Then, parcellation of brain tractography is performed to create different specific clusters. For each cluster, the density map is calculated for each voxel and converted into the binary mask. . . . 4
1.3 The left part is the low-dose CT slice for the thigh image. The middle part is the targeted label map. The right one is the legend for each label. . . . 4
1.4 The U-Net architecture for 3D image. The whole U-Net architecture includes encoding and decoding parts. . . . 6
1.5 With the introduction of a vision transformer (Vit), volumetric images are divided into patches. The left part is the overall ViT architecture. The right part is the encoder block includes the normalization and multi-head self-attention mechanisms. . . . 7
1.6 The roadmap of the dissertation . . . 9

2.1 Comparison of types of human brain atlases and regions present in each. Visualizations were made using FSLview tri-planar view for volumetric atlases and using MI-brain 3D-view for streamline atlases. Note that because atlases are in different spaces, visualized slices, anatomy, and orientation is not guaranteed to be the same across atlases. This figure is not exhaustive, and is only representative of the types of atlases and the information they contain. In general, from top-to bottom, left-to-right, atlases focus on cortical and sub-cortical gray matter, to regional white matter labels, to tractography-derived white matter pathways, to streamline-based atlases. Figure inspired by work on standardizing gray matter parcellations[89]. . . . 13
2.2 Experimental workflow and generation of atlases. Data from three repositories (HCP, BLSA, and VU) were curated. Subject-level processing includes tractography and registration to MNI space. Volumetric atlases for each set of bundle definitions are created by population-averaging in standard space. Point clouds are displayed which allow qualitative visualization of probability densities of a number of fiber pathways. Finally, surface atlases are created by assigning indices to the vertices of the MNI template white matter/gray matter boundary. . . . 15
2.3 Visualization of data contained in example volumetric and surface atlases. Example visualization for 10 pathways in the TractSeg nonlinear atlas are shown as both overlays and surfaces. . . . 19
2.4 Data validation. (a) Matrix of correlation coefficient of pathways plotted against all others indicates similarities within and across methodologies for bundle dissection. Solid white lines are used to visually separate bundle segmentation methods. (b) UMAP dimensionality reduction projected onto un-scaled 2D plane shows that many WM pathways are similar, but not the same, across methods. Object colors represent specific atlas bundles, with shape indicating segmentation methods. (c) Correlation coefficient of atlases separated by dataset indicates small, but significant, differences between datasets. Together, these justify the inclusion of all tractography methods, as well as separation of atlases by datasets. . . . 22

3.1 (a) WM is largely homogenous when imaged using most sources of MRI contrast, for example, T1w (left).(b) Traditional WM atlas (center) represents each voxel with one tissue class. (c) Modern approaches to bundle segmentation identify multiple overlapping structures (right). Diffusion tractography offers the ability to capture a multi-label description of WM voxels. 24
3.2 The pipeline of proposed WM bundle learning is presented, which integrates data processing and registration as well as bundle learning. We extract WM bundles from six different tractography methods. Structural images and corresponding tractograms are reoriented to the MNI template. Patch-wise, spatial-localized neural networks are utilized to learn WM bundle regions from a T1w MRI image. The output of each U-net is merged as the final step before segmentation. Representative samples of WM bundles acquired from six automatic tractography methods and the final learning result is visualized. 25
3.3 Each curve represents the average DSC of all WM bundles of all validation dataset scans per diffusion tractography algorithms for MAS, atlas- and learning-based methods at different threshold values. The 95 percent confidence interval is within the printed notch due to the large sample population size. The legend above each plot includes the optimal threshold for each tractography algorithm. 31
3.4 3D visualization of MAS, atlas- and learning-based results across six diffusion tractography algorithms by reconstruction of the left corticospinal tract (CST) surface on an affine reoriented coronal T1w MRI slice. The text below each image is quantitative DSC for each case. 32
3.5 Quantitative results of MAS,atlas-based method, and proposed learning methods on test cohorts from HCP, BLSA, and VU and external cohort from HCPLS, IXI and UG. The outlier percentage (top row) of all six algorithms is shown in the bar plot. Two measures are used to assess the overlap between algorithms deriving fiber mask from T1w and truth from dMRI: Dice (middle row) and surface distance (lower row). Each column presents the result of a different bundle segmentation algorithm and shows the proposed method, MAS, and single atlas-based method. Each boxplot includes each pathway of the bundle segmentation algorithm per every scan. The 95 percent confidence interval is within the printed notch due to the large sample size. The difference between methods was significant ($p < 0.005$, Wilcoxon signed-rank test, indicated by $*$). 33
3.6 Plots of overlap versus overreach for the left CST across all bundle segmentation algorithms for MAS, atlas- and learning-based methods are shown. The markers on each curve represent the overlap and overreach values at specific threshold values. The range of overreach for MAS is [0,6]. The range of overreach for atlas-based methods is [0,9]. The range of overreach for the learning-based method is [0,6]. 34
3.7 Each curve represents the average DSC of all WM bundles of all external dataset scans per diffusion tractography algorithm for atlas- and learning-based methods. The 95 percent confidence interval is within the printed line width due to the large sample size. The legend above each plot includes the optimal threshold for each tractography algorithm. . . 35

4.1 The first row and second row of (a) represent the middle thigh and lower leg from the same subject respectively. The left column is the original CT image and the right column is the target tissue label. Each tissue has a different area, and the imbalance of area makes segmentation of sparse tissue (intermuscular fat) challenging. The area of each tissue is shown in (b). 38
4.2 The proposed hierarchical coarse-to-fine thigh segmentation includes three parts: 1) The threshold and morphology are used to generate coarsely segmented pseudo labels. 2) Feeding pseudo labels into the deep learning model and training the model from scratch. 3) Using the optimized model from the previous stage as initialization, and fine-tuning the model with limited expert labels. The model from the first- and second-stage is optimized separately. 40

4.3 The fig shows the DSC comparison of thigh image using U-Net trained only with human labels, U-Net++ trained only with human labels, U-Net in stage 1, U-Net++ in stage 1, U-Net in stage 2 and U-Net++ in stage 2 in boxplots of five target tissues. 42
4.4 The plot shows the qualitative representation of the thigh slice segmentation. (a) represents three randomly selected source CT images after applying window [-150,100]. (b) represents the segmentation from U-Net only trained with human labels. (c) represents the segmentation from U-Net++ only trained with human labels. (d) and (e) represent the segmentation by using network U-Net and U-Net++ in stage 1 respectively. (f) and (g) is the segmentation by network U-Net and U-Net++ in stage 2, respectively. (h) is the ground truth. The yellow arrow points to the large difference between those methods and ground truth. The DSC values only show intermuscular fat segmentation performance for reference.. 43
4.5 The fig shows the DSC comparison of lower leg image using U-Net trained only with human labels, U-Net++ trained only with human labels, U-Net in stage 1, U-Net++ in stage 1, U-Net in stage 2 and U-Net++ in stage 2 in boxplots of five target tissues. 44
4.6 The plot shows the qualitative representation of the lower leg slice segmentation. (a) represent the source CT image after applying window [-150,100]. (b) represents the segmentation from U-Net only trained with human labels. (c) represents the segmentation from U-Net++ only trained with human labels. (d) and (e) represent the segmentation by using network U-Net and U-Net++ in stage 1 respectively. (f) and (g) is the segmentation by network U-Net and U-Net++ in stage 2 respectively. (h) is the ground truth. The text below each image is internal bone DSC. 44
4.7 Shows the relationship between mean DSC and added data for each fine-tuning. The violin plot includes 10 data points. Each data represents mean DSC across all tissues of the test cohort in one fine-tuning process. 46
4.8 The outliers from thigh and lower leg. The first row and third row are segmentation results on the thigh and lower leg. The second and fourth rows are the CT images after applying the window. Each column represents an outlier from the thigh and lower leg respectively. 47

5.1 A selective sample that highlights the inter-modality heterogeneity between MRI and CT and low-intensity difference among different muscle groups in CT. (a) The MR image is normalized by min-max. The original CT scale is clipped to [-200,500] and then normalized to [0,1]. (b) is the intensity distribution for four muscle groups. The overlap intensity among four muscle groups is observed from the second row. 50
5.2 Overview of proposed pipeline. In part (a), we adopt a CycleGAN design including two generators and two discriminators for MR and CT respectively. The segmentation module is trained by feeding synthetic CT images and corresponding MR ground truth. In part (b), the segmentation module from (a) is used to infer pseudo labels divided into hard and easy cohorts based on entropy maps. Then, the easy cohort pseudo-labels are refined based on anatomy processing (muscle and bone bask). In part (c), easy cohort pseudo-labels of CT images are used to fine-tune the segmentation module, and adversarial learning between easy and hard cohorts forces the segmentation module to adapt to hard cohort simultaneously to increase segmentation module robustness. 51
5.3 The preprocessing steps for dilating the ground truth of the MRI dataset. The blue contours in (a),(b) represent the muscle and bone boundaries extracted by level sets, and (c) represent the original ground truth. The quadriceps femoris muscle group is dilated in 6 iterations and the hamstring muscle group is dilated in 2 iterations. (d) represents the final truth after preprocessing. 52
5.4 Representation results of the proposed methods and baseline methods. Each row represents one subject. The proposed method reduces prediction errors on bones and around muscle group boundaries. The yellow arrows point to differences between the proposed method and AccSeg-Net, DISE, and SynSeg-net. The Input column images are rescaled for visualization purposes. 56

5.5 Quantitative results of DSC of baseline methods and the proposed method. * indicates ($p < 0.05$) significant difference between by Wilcoxon signed-rank test and ** indicates ($p < 0.02$ corrected by Bonferroni method[69]). The yellow arrows indicate outliers that are located at a far distance from the distribution, spanning from the 25th percentile to the 75th percentile, among the four methods. When calculating the standard deviation, these outliers are included in the calculation and can potentially influence the resulting standard deviation. Therefore, the box represents the data distribution from 25th percentile to 75th percentile rather than the standard deviation of the entire test dataset. 57
5.6 Graphic visualization for the four pipelines designed for the ablation study. (1) represents segmentation maps influenced by the segmentation module trained from scratch. (2) The pseudo-labels of the training data are inferenced by the segmentation module from scratch and then divided into two cohorts for fine-tuning. (3) The prediction map inferenced by the segmentation module from scratch is masked by a muscle mask. (4) Proposed pipeline. The pseudo-labels of the training data are inferenced by the segmentation module from scratch and then masked by a muscle mask for fine-tuning. 58
5.7 The quantitative results for four pipelines used in the ablation study. * indicates ($p < 0.05$) significant difference between by Wilcoxon signed-rank test and ** indicates ($p < 0.02$ corrected by Bonferroni method. 58
5.8 The sensitivity plot of proposed pipeline result. The x-axis represents the ratio between the eroded area and the muscle ground truth. The positive prediction value is calculated based on Eq. 5.10 59

6.1 For each visit, we compute a segmentation map of brain image and mid-thigh image of the same subject in BLSA during one visit. In the brain segmentation map (a), we use the BrainColor protocol to visualize each label. For the thigh segmentation map (b), we follow the protocol from Chapter 4. 63
6.2 This figure shows the concrete random variables at the beginning and end of the training. Each row represents one random variable. To better visualize random variables, k is set as 20, and d is set as 40. (a) represents the concrete variable at the beginning of training. (b) represents the concrete variable at the end of training. From a to b, the Concrete random variable becomes sparse and similar to a one-hot vector 64
6.3 The model used to characterize brain and body relationship (a) shows the legend of each block used in this figure. (b) shows the plain deep neural network including 6 blocks. Each block includes a fully connected layer with 256, 128, 64, 32, 16, 8 input features respectively. (c) shows the proposed model architecture. Except for the same blocks, the select layer is inserted between input features and the first block to select associated features. d represents the number of input features per each observation and k represents the number of selected features decided by the user. 65
6.4 The resulting plot for the toy example and success rate of selecting features in each run with different percentages of toy data. (a) shows the mean explain variance score across 5 output variables. Compared with the MLP model, the only difference is that Gumbel has a selected layer to subset real signals. (b) shows the success rate of real signals chosen by Gumbel Softmax. 68
6.5 The bar plot of expected selecting features and exact features Gumbel-softmax select. The five bar plots represent using brain features to predict body features including muscle, cortical bone, internal bone, subcutaneous fat, and intermuscular respectively. When we increase the number of selected features, the exact number of selected features is less than expected since there are duplicate selected features 68
6.6 The histogram for demographic information of BLSA dataset. (a) shows the age distribution of BLSA subjects. The BLSA dataset is used to investigate the aging effect. Most people are elder. (b) shows the sex distribution of BLSA subjects. (c) shows the distribution of visit numbers per subject. The maximal number of visits of one subject is 10. . . . 69

6.7 The bland-Altman plot of linear regression and proposed nonlinear methods. Each plot has the difference between prediction and truth as the y-axis and ground truth as the x-axis. The gray color represents the linear regression and the red color represents the proposed methods. We can observe the proposed methods have a smaller limit of agreement in muscle area prediction compared with linear regression as pointed out by the yellow arrow. We also find that there are two obvious clusters controlled by sex in cortical bone and subcutaneous fat prediction. 70
6.8 The bland-Altman plot of linear regression and proposed nonlinear methods. Each plot has the difference between prediction and truth as the y-axis and ground truth as the x-axis. The gray color represents the linear regression and the red color represents the proposed methods. As pointed out by the arrow, we can observe the proposed methods have a smaller limit of agreement in left hippocampus prediction compared with linear regression when we include whole brain volume into the estimation process. 72

A.1 Copyright from Elsvier 91
A.2 Copyright from Wiley 92
A.3 Copyright from SPIE 93

CHAPTER 1

Introduction

1.1 Overview

Over the past two decades, medical imaging techniques such as computed tomography (CT), magnetic resonance imaging (MRI), X-ray and diffusion MRI have been invaluable for the early detection, diagnosis, and treatment of diseases [18]. Given the growing size of medical images, varied pathology, and fatigue of physicians, researchers, and experts, computer-assisted algorithms for medical images have become increasingly necessary [143].

Computer algorithms for the delineation of anatomical structures and other regions of interest from the medical image are regarded as medical image segmentation algorithms[122]. The segmentation algorithm has an essential role in the numerous biomedical imaging applications such as quantification of tissues[90], localization of tumor [11], analysis of anatomical structure[66] and computer-aided surgery[68]. Initially, low-level edge and line detector filters[184] and deformation model[32] from 1970 to the 1990s were used to perform low-level pixel image level processing. At the end of the 1990s, active shape model[172], atlas[74] and hand-crafted features[193] and statistical classifiers[161] were used to extract the boundary of the region of interest in high-dimensional feature space. With the advent of deep learning technology, hand-crafted features are increasingly being replaced with features representing the data learned by deep neural networks to achieve medical image segmentation tasks with impressive performance[161].

Despite the newly improved performance, large, representative, and high-quality annotated datasets are the prerequisite for the advanced deep learning model[147]. However, we rarely have a perfect training dataset in the medical imaging field. We usually have a limited size of training data or low-quality annotation[146]. Even more, the region of interest (ROI) only can be derived from certain imaging modalities instead of images that we have to delineate. In this situation, knowledge transfer can be one useful way to solve the issue. Knowledge transfer refers to the use of knowledge from other image modalities[97] or similar task[186] to infer the bio-structure segmentation from images we have. This way can help researchers solve imperfect training dataset problems by leveraging existing publicly annotated datasets or other imaging modalities with annotation.

Deep learning brings us opportunities to segment bio-structures from the brain to the whole human body. Whole-brain segmentation plays a crucial role in both scientific and clinical research, facilitating quantitative comprehension of the human brain. This non-invasive method allows for the quantification of brain structures

using structural MRI. Except for the quantification of brain structures, a thorough comprehension of the human brain requires an understanding of its anatomical connectivity. White matter forms these long-range connections, organized into distinct tracts[148]. It is valuable to segment the white matter pathway from the brain. As for body composition, an increasing body of evidence supports an intimate brain-body connection in aging[12], with cardiovascular disease (CVD)[95], and neuro-degenerative disease[88]. Currently, the study begins to focus on the connection among different systems of the human body like brain-gut connection[79]. Investigating relationships between each other can bring new understanding to each one of them. Inspired by this, we focus on investigating knowledge transferring to segment the bio-structure of the brain and body to characterize relationships among them in this thesis.

1.2 Structural MRI and diffusion MRI for brain

Neuroimaging is a branch of medical imaging that focuses on the brain. In addition to diagnosing disease, neuroimaging also studies brain anatomy, the function of the brain parts, and the connection between brain parts. MRI is the popular imaging modality in the neuroimaging study of safety and high contrast among different tissues of the brain. In this thesis, structural MRI and diffusion MRI are two image modalities to discuss.

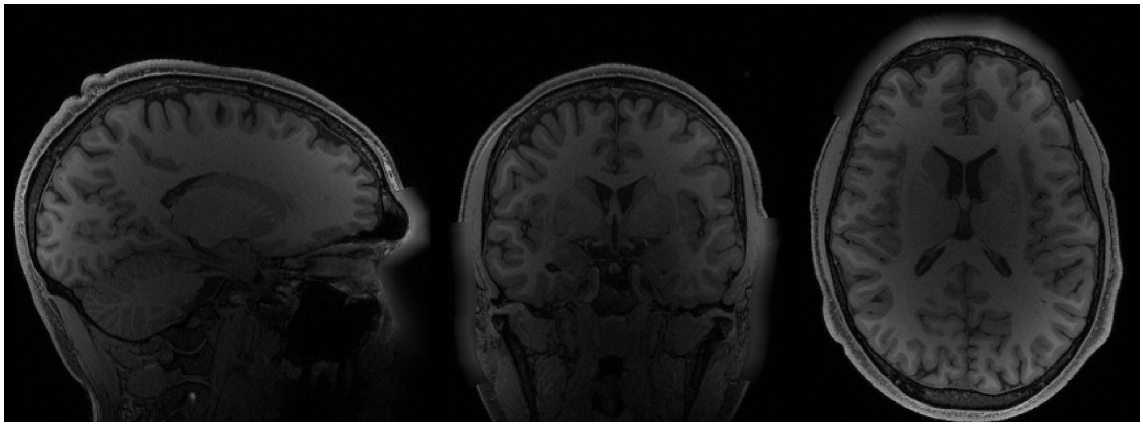
1.2.1 Structural MRI for brain

Structural MRI for the brain produces high contrast between gray and white matter, allowing for the quantification of gray matter, white matter, and cerebrospinal fluid (CSF)[55]. One popular image type for structural MRI is T1 weight MRI (T1w). T1w is one of the sequences of MRI to capture differences in the T1 relaxation time of tissues. T1w can help physicians capture the lesion, and tumor and observe the longitude development of the brain. The following Figure 1.1 is the brain T1w image in triplane view.

From Figure 1.1, the CSF of T1w is dark. The gray matter and white matter are gray and white respectively. However, the whole white matter is homogeneous and shows little contrast within it. It is hard to subdivide the white matter into other structures by only relying on contrast within T1w.

1.2.2 Diffusion MRI for brain

Diffusion MRI (dMRI) is a form of MRI imaging based on measuring the random motion of water within a voxel under a diffusion gradient pulse. The contrast of dMRI is the attenuation of the signal based on how easily the water molecules can diffuse in that region. The dMRI signal is related to the apparent diffusion coefficient ADC and the b-value by:



Sagittal view Coronal view Axonal view

Figure 1.1: The triplane view of a T1w brain. There are sagittal, coronal, and axonal views of brain images.

$$S = S_0 e^{-bADC} \tag{1.1}$$

where S_0 is the signal in the absence of diffusion gradient pulse and ADC is the apparent diffusion coefficient. b is the b-value used to represent the amplitude of the diffusion gradient pulse. DMRI works as a non-invasive imaging method to examine the white matter pathway in vivo. DMRI measures the random microscopic motion (or diffusion) of water molecules in the brain tissue as the contrast parameter[78]. The motion is anisotropic because of the arrangement of fibers within WM. Based on this property, tractography is performed to estimate the anatomical trajectories of WM. One popular workflow is to estimate local fiber orientation based on diffusion signal models like tensor[10] or other advanced models[2, 156] and reference long-range pathways from local orientation[9]. The subsequent dissection[23] of streamlines across from whole-brain fractograms, allows for the segmentation or mapping of WM pathways. Figure 1.2 presents one of the pipelines to parcellate white matter and convert WM pathways into binary segmentation masks by density map.

1.3 Single slice CT for thigh

CT represents a computerized X-ray imaging procedure in which an X-ray is aimed at a patient and quickly rotated around the body, producing signals that are processed by computers to generate a cross-sectional image of the body. However, CT is a radiation-intensive procedure[17]. The accumulated CT radiation dose could be dangerous to the human body. The low-dose CT scan is the preferred choice when it is suitable for patients' conditions such as low-dose CT used to screen the lung cancer module[124]. CT is the promising reference method for quantifying whole-body composition and skeletal muscle mass[109]. However, if we

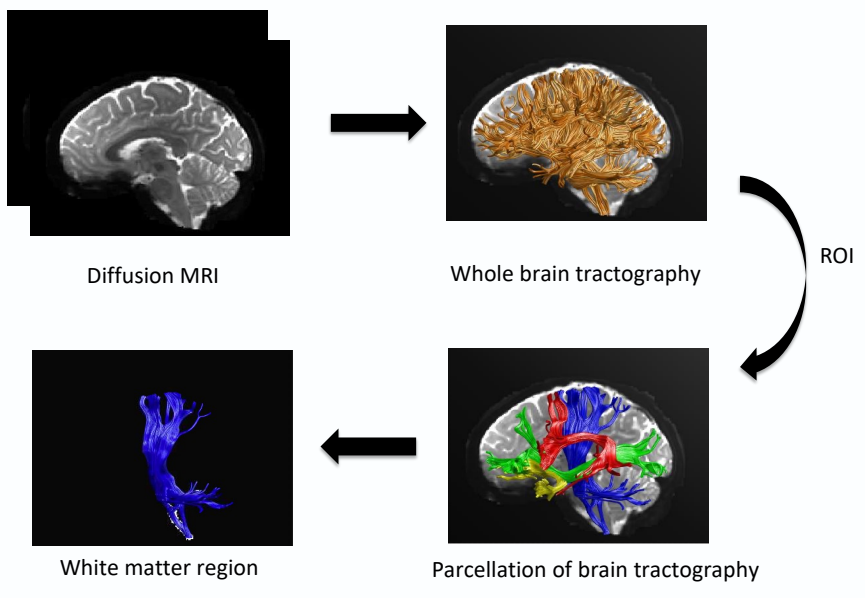


Figure 1.2: This is one of the classic pipelines of getting white matter region from dMRI. Diffusion tensor is derived from dMRI. Then the random seeds are used to track whole brain tractography. Starting from each voxel, the next point is like a continuation of the current one. Then, parcellation of brain tractography is performed to create different specific clusters. For each cluster, the density map is calculated for each voxel and converted into the binary mask.

perform a study on the whole CT image, the subject will suffer from extra radiation which is not what we expect. To address this concern, the researchers proposed to use one single slice CT to replace whole CT volume to measure body composition and demonstrated results derived from one single are related to 3D whole-body volume[109]. As for single slice, thigh composition is of great interest since it includes muscle, fat, and bone tissues. They are all the import health biomarkers since they can change along with disease[36]. One typical example is shown in Figure 1.3.

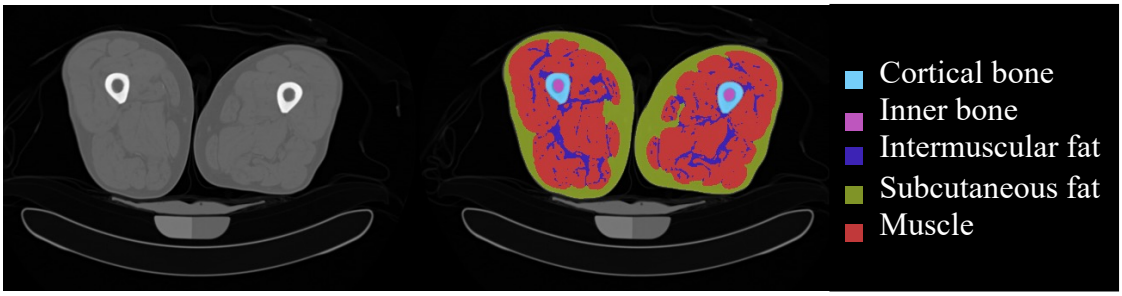


Figure 1.3: The left part is the low-dose CT slice for the thigh image. The middle part is the targeted label map. The right one is the legend for each label.

From Figure 1.3, the outmost fat is the subcutaneous fat. The subcutaneous fat contains the anatomical muscle. There are the femur and intermuscular fat within the muscle. The bone is divided into the cortical

Tissues	HU values
Air	-1000
Fat	[-190,-30]
Water	0
Muscle	[30,80]
Bones	[1000,+∞]

Table 1.1: The approximate HU values for common tissues.

bone and internal bone in our case. Segmentation is an important approach to quantify the properties including area, mass, etc. In thigh tissue segmentation, intensity-based segmentation is widely applied since the CT image has standard units as "Hounsfield units". The unit of the CT image has a corresponding physical meaning. In the Hounsfield scale, water is arbitrarily assigned a value of 0 HU[154]. Other values can be computed according to:

$$HU = 1000 \times (\mu_{tissue} - \mu_{water})/\mu_{water} \tag{1.2}$$

where μ is the CT linear attenuation coefficient. The Table 1.1 represents the approximate HU values for common tissues

1.4 Intra-modality medical image segmentation

The intra-modality medical image segmentation means delineating the same target on the dataset of the different patients within the same image modality such as CT, MRI, etc.

1.4.1 Atlas-based segmentation

In the earlier stage of medical image segmentation, researchers used low-level segmentation technology including but not limited to the threshold, region, edge, and clustering-based methods to extract regions of interest[94]. Those methods can achieve acceptable performance. However, those methods are limited to a small dataset and cannot combine human expert annotation knowledge into the segmentation process. The major difference between high-level and low-level segmentation methods is whether the method incorporates prior anatomical knowledge[104]. Atlas-based segmentation is one of the most popular high-level methods in the segmentation task. The atlas is defined as the combination of an intensity image (template) and its segmented image (the atlas labels)[20]. Organ segmentation measurements vary significantly across populations due to diverse demographic differences. Creating an atlas for each organ within a population aids not only in segmenting out-of-sample individuals but also in providing comparative statistical measures across populations[92, 93, 195]. After registering the atlas template and the target image, the atlas labels are propagated to the target image as the final segmentation label map[20]. At this point, the segmentation turns into a registration problem. The registration errors decide the accuracy of the segmentation performance. To

better deal with registration errors caused by a single atlas, multi-atlas segmentation (MAS)[166] is applied to perform label propagation to minimize outliers by discarding low agreement among many atlases bringing improvement in the accuracy for the well-defined shape of the objects.

1.4.2 Learning-based segmentation

Different from atlas-based segmentation, deep learning[91] is currently a popular method in medical image segmentation. Deep learning uses computational models that are composed of multiple processing layers to learn a representation of medical images. One of the popular models for medical image segmentation is U-Net[132]. The network architecture is shown in Figure 1.4. The left part of the network used to extract features is called the encoder, and the right part of the network that is restored from the features to the original image size is defined as a decoder. The encoder and decoders form the encoder-decoder structure. The skip connection brought by U-Net integrates the low-level features with high-level features to improve the performance of segmentation. The U-Net and its variant[196] have been applied extensively in varied modalities and different body parts of medical images[98, 151].

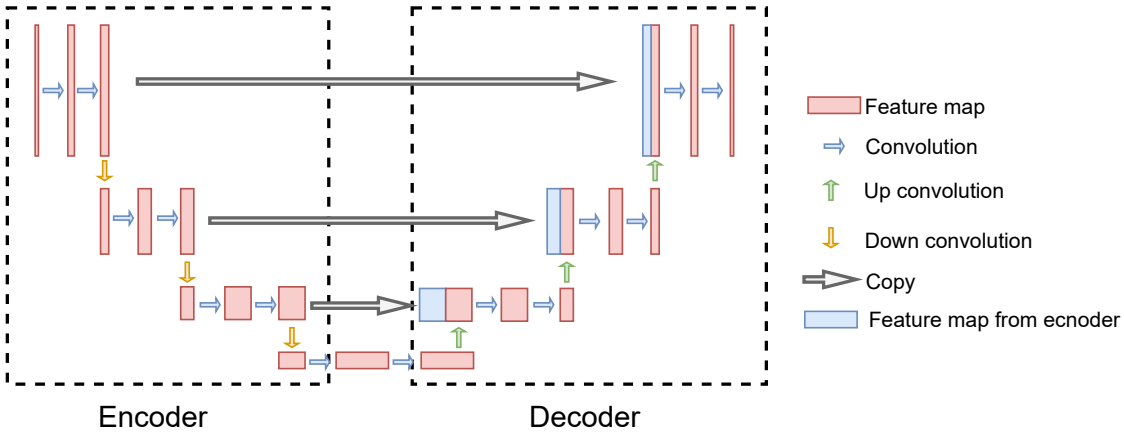


Figure 1.4: The U-Net architecture for 3D image. The whole U-Net architecture includes encoding and decoding parts.

Transformers are increasingly recognized as the alternative to the established Convolutional Neural Networks (CNNs). Dosovitskiy et al. introduced the Vision Transformer (ViT), a sophisticated extension of the conventional NLP transformer[39]. The transformer converts sequences of image patches into linear embeddings, leveraging an attention mechanism to find the connection between these embeddings, a departure from the traditional convolutional operations inherent in CNNs. As shown in Figure 1.5, each input image is systematically segmented into patches of dimensions 16×16, referred to as 'visual tokens.' Subsequently, these tokens undergo a projection into encoded vectors of predefined dimensions through a Multi-Layer Per-

ceptron (MLP). To retain the spatial context amidst the attentional process, a position encoding vector is integrated with the encoded vectors. Attention maps, derived from a multi-head attention network, culminate in an output prediction via a dual-layer MLP classification mechanism. Intrinsically, the ViT is adept at capturing intricate long-range relationships and dependencies between visual constituents, exhibiting minimal inductive biases in the visual domain. In the medical domain, the growth of interest in adapting transformer networks is also demonstrated[187, 192].

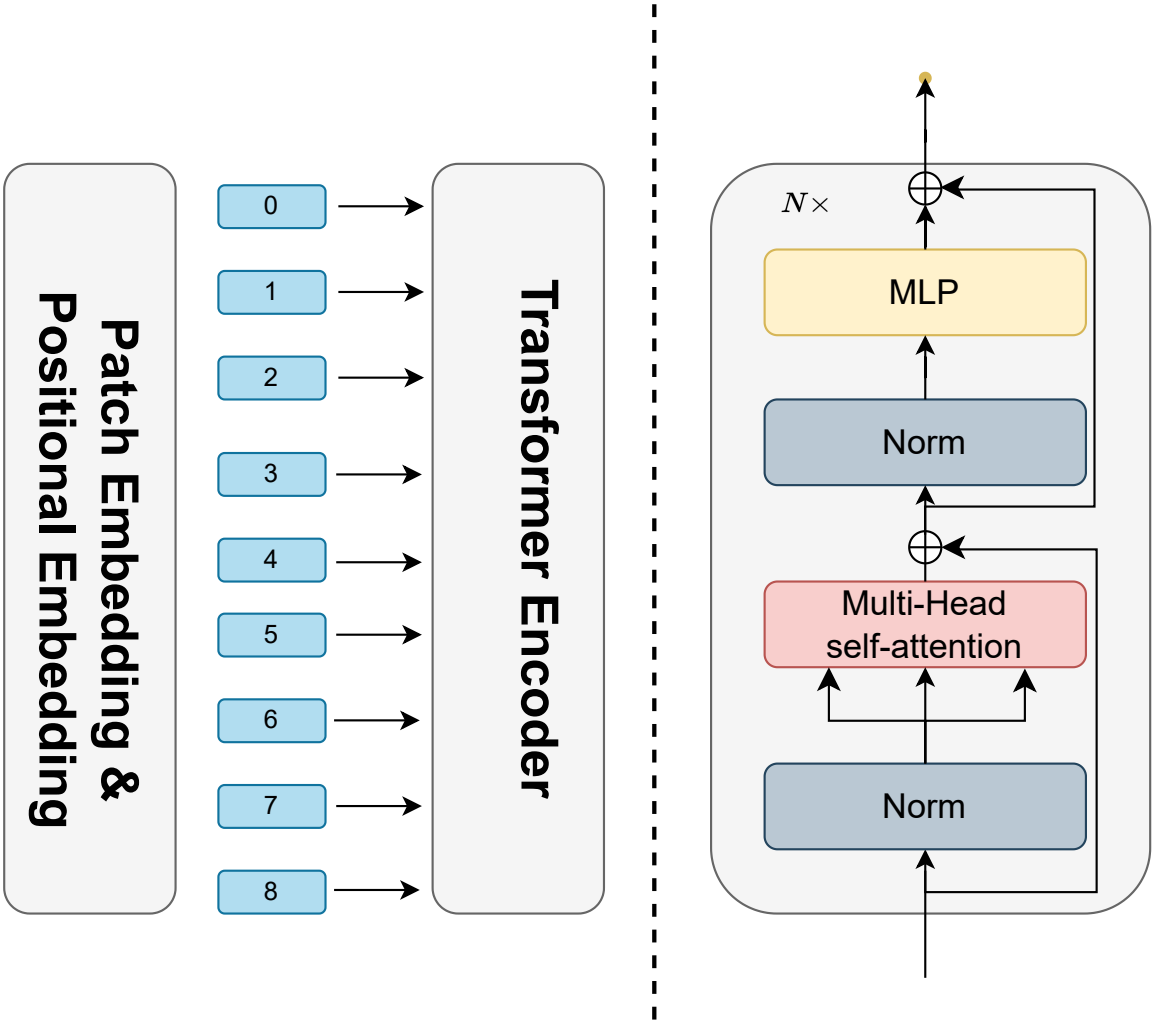


Figure 1.5: With the introduction of a vision transformer (Vit), volumetric images are divided into patches. The left part is the overall ViT architecture. The right part is the encoder block includes the normalization and multi-head self-attention mechanisms.

1.4.3 Transfer learning

The prerequisite for the impressive performance of deep learning is sufficient annotation data. Manual annotation is label-intensive and needs professional knowledge. It is hard to collect high-quality data for training

in the clinic. Transfer learning is an effective way to address the above challenges[104]. Transfer learning means that deep learning models can be trained on a large-scale labeled database. After the convergence of the pre-train model, the limited original annotated data is utilized to fine-tune the pre-train model to converge. Transferring learning from natural images to medical images is also de facto for medical image analysis[147].

1.5 Inter-modality medical image segmentation

Medical imaging has many kinds of modalities. Researchers cannot only segment on one image modality to help doctors since each modality has its own strengths and drawbacks. For instance, while it is difficult to annotate white matter pathways from T1w, the white matter can be easily parcellated into clusters based on dMRI. Thus, how to leverage the labels from the source domain to perform segmentation on the new domain is a popular topic in the research area.

One way to solve this problem is by the atlas-based segmentation method mentioned in Chapter 1.4. Transferring statistical average labels derived from the source domain to the target image by registering the template to the target image. The deep learning method has been the de facto standard for medical image segmentation. However, the learning method showed degraded performance when the model was applied to a domain different from what it trained on. Synthesis of a target image based on a generative adversarial network (GAN)[58] is one category to address this issue. GAN includes two models: generative models (G) that capture the data distribution and discriminative model (D) that estimates the probability that a sample came from training data rather than G. The training procedure for G is to maximize the probability of D making a mistake. To learn the generator's distribution, the prior input noise is defined as variable $p_z(z)$, and mapping from noise to data space is regarded as $G(z)$. The target function can be represented as:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{data(x)}}[logD(x)] + \mathbb{E}_{z\sim p(z)}[log(1-D(G(z)))] \tag{1.3}$$

Based on GAN, CycleGAN[43] is created to perform image-to-image translation for unpaired images and provides a promising tool for cross-modality synthesis (MRI-CT). CycleGAN introduces cycle loss by utilizing two generators $(G_{(A-B)}, G_{(B-A)})$ to force image content consistent when translating from domain A to domain B.

$$\begin{aligned} L_{cyc}(G_{(A-B)}, G_{(B-A)}) = {} & \mathbb{E}_{x\sim p_{data(x)}}||G_{B-A}(G_{A-b}(x)) - x||_1 \\ & + \mathbb{E}_{y\sim p_{data(y)}}||G_{A-B}(G_{B-A}(y)) - y||_1 \end{aligned} \tag{1.4}$$

where L_{cyc} represents the cycle loss. x is input from domain A and y is input from domain B. The segmentation

neural network can be trained on synthesizing images to learn segmentation maps. This pipeline provides a solid foundation for domain adaptation. Its variants have been applied to the whole medical image field.

1.6 Contributed Works

To characterize the relationship between the brain and the body, we proposed several methodologies and tools. We constructed white matter atlases from a large cohort of subjects (**contribution 1**). Building on this, we introduced a localized patch-based convolutional neural network designed specifically to predict white matter pathways from T1 structure MRI data personally (**contribution 2**). Turning our attention to body composition, we proposed label-efficient approaches to segment the muscle, bone, and fat tissues from a single CT slice (**contribution 3**). To further classify muscle into distinct groups, we utilized domain adaptation to transfer labels from publicly available MRI data to individual CT slices (**contribution 4**). Concluding our previous contributions, we introduced the gumble-softmax structure into deep neural networks, empowering them to predict brain features based on body features and reciprocally (**contribution 5**). The roadmap of the dissertation is shown in Figure 1.6

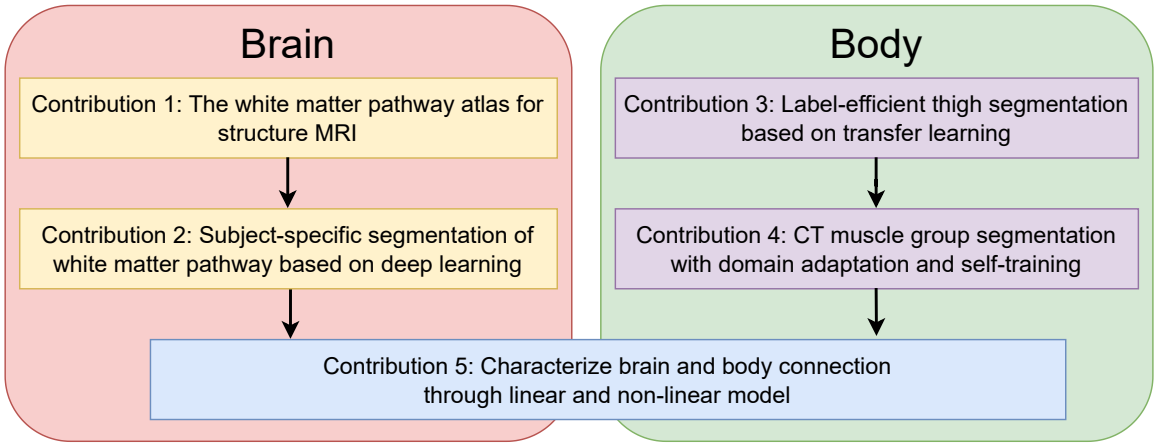


Figure 1.6: The roadmap of the dissertation

1.6.1 Contribution 1: The white matter pathway atlas for structure MRI

The atlases in neuroscience research are valuable, serving not only as indispensable tools for individual subject studies but also for drawing inferences and comparing diverse populations. Notwithstanding the extensive array of human brain structure atlases accessible to researchers, there remains a conspicuous scarcity of resources specifically dedicated to the white matter or atlases derived from limited populations.

To bridge this gap, we proposed to create white matter pathway atlases by collecting data from a large-scale, multi-site population. We selected six state-of-the-art white matter pathway segmentation methods to

curate the atlas segmentation masks. To ensure the reliability and relevance of our atlas, we validated the atlas by comparing atlas from different populations.

This part will be covered in Chapter 2.

1.6.2 Contribution 2: Subject-specific segmentation of white matter pathway based on deep learning

DMRI can be used to probe the connectivity and microstructure of human brain tissue in vivo. However, many legacy or time-constrained studies only have T1w. Annotating white matter pathways from T1w is challenging due to its limited contrast for white matter. Though the white matter atlas existed, it serves as an averaged representation of a cohort population. Convolutional neural networks (CNN) exhibit potential in capturing subject-specific variations compared with atlas. However, no prior work deriving white matter pathway segmentation from T1w utilizing CNN.

To bridge the gap, we introduced localized patch-based CNNs to predict white matter pathways directly from T1w. We validated our methodology on an extensive cohort population, employing six state-of-the-art techniques to derive white matter pathway data as our benchmark.

This part will be covered in Chapter 3.

1.6.3 Contribution 3: Label-efficient thigh segmentation based on transfer learning

Medical image segmentation plays a pivotal role in quantifying the volumes of muscle, bone, and fat, thereby providing insights into body composition. Deep neural networks, which have shown promise in this domain, typically require extensive annotated data for training from scratch. However, acquiring human annotations for medical images, even for a single CT slice, is a process that is both time-consuming and labor-intensive.

To handle the human annotation problem, we introduced a transfer learning-based approach. Initially, we train the model using pseudo labels and subsequently fine-tune it with a limited set of human expert annotations to achieve accurate body composition.

This chapter will be covered in Chapter 4.

1.6.4 Contribution 4: Single slice CT muscle group segmentation with domain adaptation and self-training

Segmenting muscles into distinct groups provides fine-grained features that can be utilized to understand the relationship between the brain and the body. However, annotating muscles on an individual slice proves difficult due to the similar intensity shared among different muscle groups. Further complicating this task is the absence of a 3D context and the inherent challenges of annotation.

To address the above challenges, we utilize muscle group annotations from publicly available MRI volume

datasets. Through domain adaptation, we effectively transfer muscle labels from MRI volumes to single CT slices.

This part will be covered in Chapter 5.

1.6.5 Contribution 5: Characterize brain and body connection through linear and non-linear model

Characterizing brain and body connection is crucial for understanding how body composition influences brain diseases, and vice versa. Yet, most prior research has primarily relied on conventional metrics like BMI or hip-waist ratio, rather than detailed fat, muscle, and bone distribution.

To close the gap, our study uses 133 regions of interest and body composition as quantitative metrics. When we design the model architecture, we integrate the Gumbel-softmax into a deep neural network to extract relevant input features related to output features.

This part will be covered in Chapter 6.

CHAPTER 2

The white matter pathway atlas for structural T1w MRI image

This work was previously published[64]. Permission to include the work as part of the dissertation has been obtained, see Appendix A.

2.1 Introduction

The creation and application of medical image-based brain atlases is widespread in neuroanatomy and neuroscience research. Atlases have proven to be a valuable tool to enable studies on individual subjects and facilitate inferences and comparisons of different populations, leading to insights into development, cognition, and disease[20], [89, 153, 54]Through the process of spatial normalization, images can be aligned with atlases to facilitate comparisons of brains across subjects, time, or experimental conditions. Additionally, atlases can be used for label propagation, where anatomical labels are propagated from the atlas to new data to identify a priori regions of interest. With these applications in mind, several human brain atlases have been created Figure 2.1), with variations in the number of labels, the regions of the brain that are delineated, the methods used to generate labels, and the population or individuals used to create the atlas (for a review of the existing atlases and their standardization, see recent work by [89].

2.2 Related works

Despite the wide variety of human brain atlases available to the research community, there is a distinct lack of resources available to describe the white matter of the brain. For example, most atlases emphasize cortical or sub-cortical gray matter, and do not contain a label for white matter[183, 136, 56, 130, 158, 102, 103, 135, 112, 86, 83, 6, 107, 13, 44, 75, 82, 157, 106], or only label white matter as a single homogenous structure, or simply separate into the “cerebral white matter” of the left and right hemispheres[105, 101, 37].

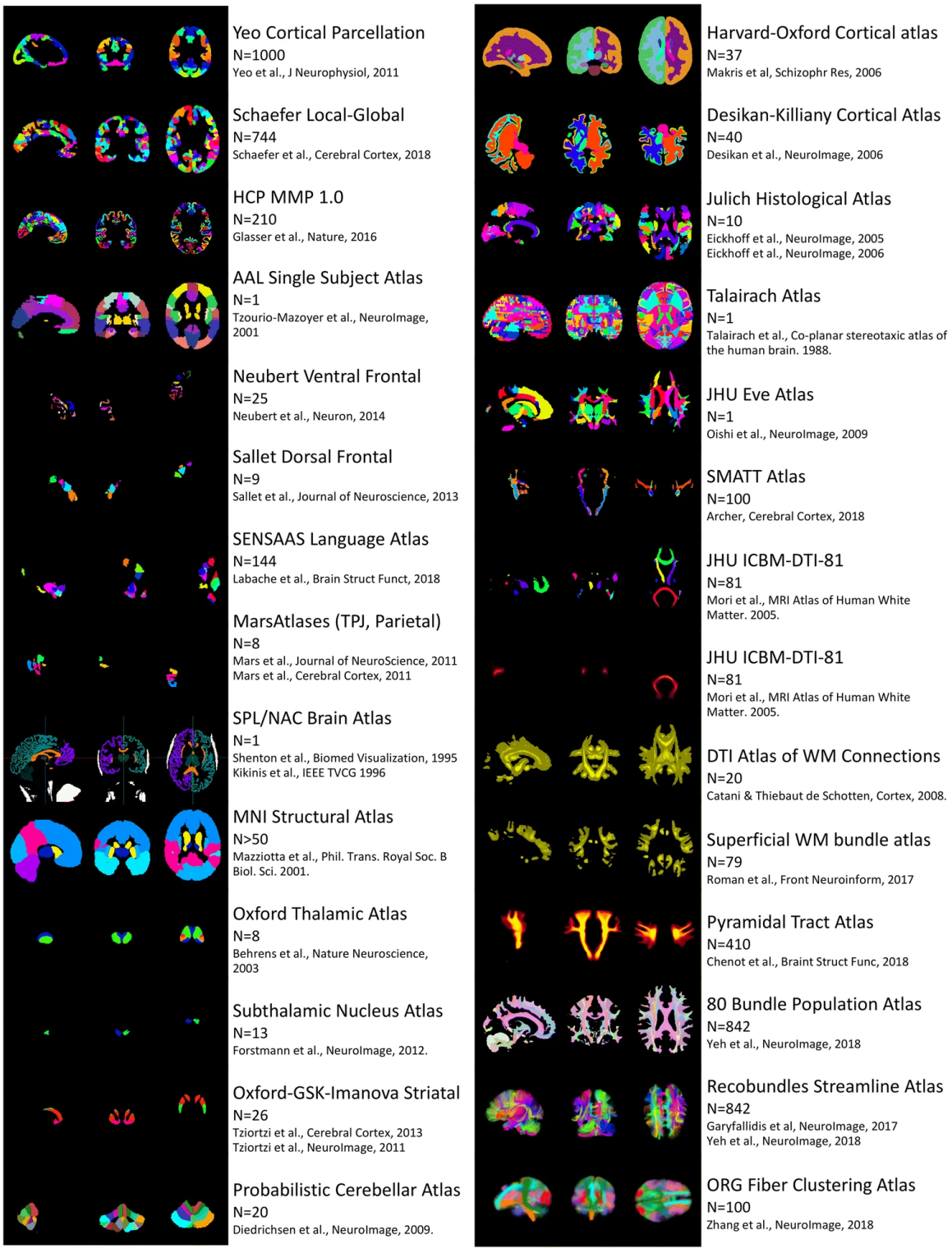


Figure 2.1: Comparison of types of human brain atlases and regions present in each. Visualizations were made using FSLview tri-planar view for volumetric atlases and using MI-brain 3D-view for streamline atlases. Note that because atlases are in different spaces, visualized slices, anatomy, and orientation is not guaranteed to be the same across atlases. This figure is not exhaustive, and is only representative of the types of atlases and the information they contain. In general, from top-to bottom, left-to-right, atlases focus on cortical and sub-cortical gray matter, to regional white matter labels, to tractography-derived white matter pathways, to streamline-based atlases. Figure inspired by work on standardizing gray matter parcellations[89].

Some atlases do indeed include labels for white matter. However, in many cases these labels are for "regions" of the white matter rather than labels for specific white matter bundles[41, 40, 42, 149, 87, 117] For example, an atlas may contain a label for the "anterior limb of the internal capsule" or "corona radiata" which are descriptions of regions through which several white matter bundles are known to pass. While these regions are certainly scientifically useful, the white matter pathways themselves would be more informative for network neuroscience investigations or applications where white matter structure, connectivity, and location are paramount. Additionally, regional labels do not overlap, whereas the fiber bundles of the brain are known to be organized as a complex mixture of structures, overlapping to various degrees. To overcome these limitations, several atlases have been created using diffusion MRI fiber tractography, a technique which allows the investigator to perform a "virtual dissection" of various white matter bundles of the brain. Examples include population-based templates [194, 162]or atlases of association and projection pathways[182, 110, 118, 24, 22] atlases of the superficial U-fibers connecting adjacent gyri [116, 60] and atlases created from tractography on diffusion data averaged over large population cohorts[116, 180, 181]. In particular, several atlases have been made with a focus on a single pathway or a set of pathways with functional relevance[46] for example the pyramidal tract[28] the sensorimotor tracts[4] or lobular-specific connections[24, 129, 152],Existing tractography-based atlases, however, typically suffer from one or more limitations: (1) small population sample sizes, (2) restriction to very few white matter pathways, and (3) the use of out-dated modelling for tractography (specifically the use of diffusion tensor imaging which is associated with a number of biases and pitfalls). Further, it is not clear whether the same pathway defined using one atlas results in the same structure when compared to another atlas due to differences in the procedures utilized to define and dissect the bundle under investigation. A final type of atlas, streamline-based atlases[180, 53, 194, 131, 61], have become popular in recent years. These are composed of millions of streamlines and can be used as a resource to cluster sets of streamlines on new datasets, thus they nicely complement the use and application of volumetric atlases when diffusion MRI is available. In this work, we introduce the Pandora white matter bundle atlas. The Pandora atlas is actually a collection of 4-dimensional population-based atlases represented in both volumetric and surface coordinates in a standard space. Importantly, the atlases are based on a large number of subjects, and are created from multiple state-of-the-art tractography and dissection techniques, resulting in a sizable number of (possibly overlapping) white matter labels. In the following, we describe the creation of these atlases, validate the use of multiple subject populations and multiple tractography methodologies.

2.3 Method

Figure 2.2 presents an overview of the pipeline and methodology used to create these atlases. Briefly, we retrieved and organized data from 3 large repositories (Figure 2.2, Data). For each subject, we performed

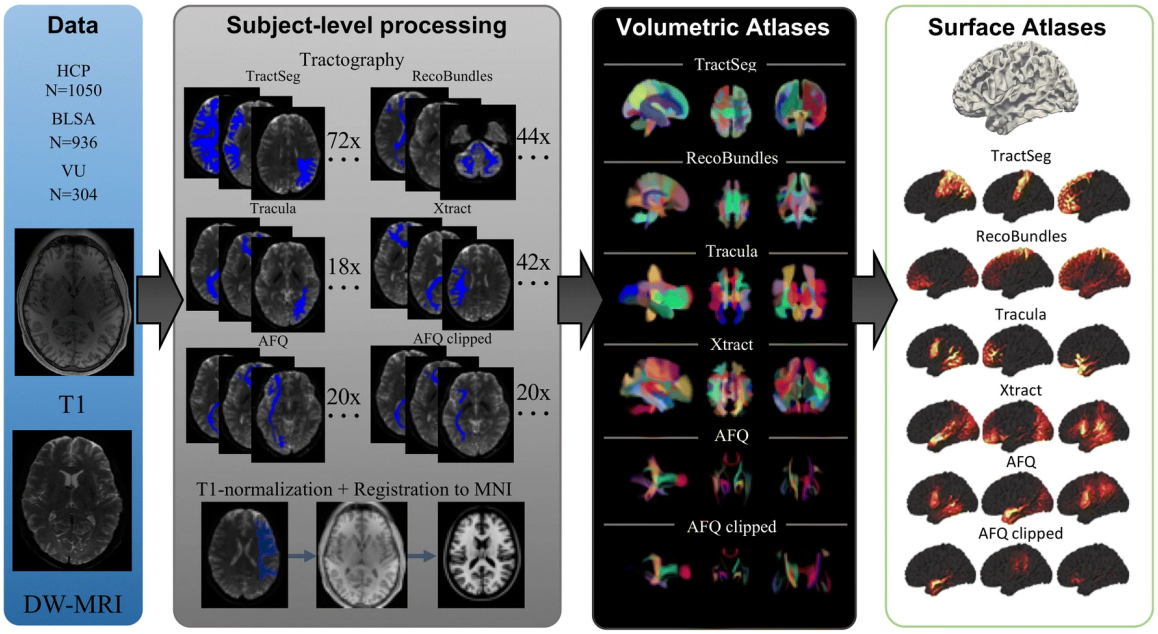


Figure 2.2: Experimental workflow and generation of atlases. Data from three repositories (HCP, BLSA, and VU) were curated. Subject-level processing includes tractography and registration to MNI space. Volumetric atlases for each set of bundle definitions are created by population-averaging in standard space. Point clouds are displayed which allow qualitative visualization of probability densities of a number of fiber pathways. Finally, surface atlases are created by assigning indices to the vertices of the MNI template white matter/gray matter boundary.

six different automated methods of tractography and subsequent white matter dissection (Figure 2.2, Subject-level processing: tractography), and registered all data to a standard volumetric space (Figure 2.2, Subject-level processing: registration). Next, a probabilistic map was created separately for each white matter bundle in standard space in order to create the volumetric atlases (Figure 2.2, Volumetric atlas creation). Finally, a surface mesh of the boundary between white and gray matter was created, and the volumetric maps were used to assign probabilities along this surface to create the surface-intersection atlases (Figure 2.2, Surface Atlas creation).

2.3.1 Data

We used de-identified images from the Baltimore Longitudinal Study of Aging (BLSA), Human Connectome Project (HCP) S1200 release, and Vanderbilt University (Figure 2.2, Data). The BLSA is a long-running study of human aging in community-dwelling volunteers and is conducted by the Intramural Research Program of the National Institute on Aging, NIH. Cognitively normal BLSA participants with diffusion MRI data were included in the present study, using only one scan per participant, even if multiple follow-ups were available. HCP data are freely available and unrestricted for non-commercial research purposes and

	HCP	BLSA	VU
Subjects	1060	963	303
Age	28.8±3.5	66.2±14.82	29.7±11.5
Age Range	[22,35]	[22.4,95.1]	18,75
Handedness	N/A	86L,843R,35N/A	30L,270R,3N/A
Sex	488M,573F	431M,532F	134M,169F

Table 2.1: Meta data information.

are composed of healthy young adults. This study accessed only de-identified participant information. All datasets from Vanderbilt University were acquired as part of a shared database for MRI data gathered from healthy volunteers. A summary of the data is given in Table 2.1, including number of subjects, age, sex, and handedness. All human datasets were acquired under research protocols approved by the local Institutional Review Boards.

All datasets included a T1-weighted image, as well as a set of diffusion-weighted images (DWIs). Briefly, the BLSA acquisition (Philips 3T Achieva) included T1-weighted images acquired using an MPRAGE sequence. Diffusion-weighted images were acquired using a single-shot EPI sequence, and consisted of a single b-value(700=s/mm^2), with 33 volumes. HCP acquisition (custom 3T Siemens Skyra) included T1-weighted images acquired using a 3D MPRAGE sequence. Diffusion images were acquired using a single-shot EPI sequence, and consisted of three b-values (b=1000, 2000, and 3000 s/mm^2), with 90 directions (and 6 b=0 s/mm^2) per shell. The scans collected at Vanderbilt included healthy controls from several projects. A typical acquisition is below, although some variations exist across projects. T1-weighted images acquired using an MPRAGE sequence. Diffusion images were acquired using a single-shot EPI sequence and consisted of a single b-value (1000 s/mm^2), with 65 volumes.

Data pre-processing included correction for susceptibility distortions, subject motion, eddy current correction[3] and b-table correction[137].

2.3.2 Subject-level processing: tractography

Six methods for tractography and virtual bundle dissection were employed on all diffusion datasets in native space (Figure 2, Subject-level processing). These included (1) TractSeg[169] (2) Recobundles[53] (3) Tracula[182] (4) XTract[168] (5) Automatic Fiber-tract Quantification (AFQ)[179] and (6) post-processing of AFQ where only the stem of the bundle was retained, which we call AFQ-clipped. Algorithms were chosen because they are fully automated, validated, and represent a selection of the state-of-the art methods in the field. In all cases, algorithms were run using default parameters or parameters recommended by original authors. Briefly, TractSeg is based on convolutional neural networks and performs bundle-specific tractography based on a field of estimated fiber orientations[169, 170], and delineates 72 bundles. We implemented the dockerized version at (https://github.com/MIC-DKFZ/TractSeg) which generates fiber orientations using

constrained spherical deconvolution using MRtrix software[155] Recobundles segments streamlines based on their shape-similarity to a dictionary of expertly delineated model bundles. Recobundles was run using DIPY[52] software (https://dipy.org) after performing whole-brain tractography using spherical deconvolution and DIPY LocalTracking algorithm. The bundle-dictionary contains 80 bundles, but only 44 were selected to be included in the Pandora atlas after consulting with the algorithm developers based on internal quality assurance (for example removing cranial nerves which are often not used in brain imaging). Of note, Recobundles is a method to automatically extract and recognize bundles of streamlines using prior bundle models, and the implementation we chose uses the DIPY bundle dictionary for extraction, although others can be used. Tracula (https://surfer.nmr.mgh.harvard.edu/fswiki/Tracula) uses probabilistic tractography with anatomical priors based on an atlas and Freesurfer[47, 48, 31] (https://surfer.nmr.mgh.harvard.edu) cortical parcellations to constrain the tractography reconstructions. Tracula used the ball-and-stick model of diffusion from FSL's[77] bedpostx algorithm to reconstruct white matter pathways, and resulted in 18 bundles segmented per subject. Xtract (https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/XTRACT) is a recent automated method for probabilistic tractography based on carefully selected inclusion, exclusion, and seed regions, selected for 42 tracts in the human brain. Xtract also utilized the ball-and-stick model (bedpostx) of diffusion for local reconstruction. AFQ (https://github.com/yeatmanlab/AFQ) is a technique that identifies the core of the major fiber tracts with the aim of quantifying tissue properties within and along the tract, although we only extracted the bundle profile itself. The default in AFQ is to use tensor based deterministic tractography, followed by fiber segmentation utilizing methodology defined by[165], and removal of outlier streamlines. In our case, we extracted the full profile of the bundle, as well as the core of the bundle which was performed in the AFQ software by a clipping operation. For this reason, we called these AFQ and AFQ-clipped, respectively. Both of these methods resulted in 20 bundles. In total, we present 216 bundles in the atlas. Output from all algorithms were in the form of streamlines, tract-density maps, or probability maps. In all cases, pathways were binarized at the subject level, indicating the voxel-wise existence or non-existence of the bundle in that subject, for that pathway. These binary maps were used to create the population atlases after deformation to standard space. Exhaustive manual quality assurance (QA) was performed on tractography results. QA included displaying overlays of binarized pathways over select slices for all subjects, inspecting and verifying appropriate shape and location of all bundles on all subjects. We note that not all methods were able to successfully reconstruct all pathways on all subjects, for this reason, some atlases contain information from fewer than all 2443 subjects.

2.3.3 Volumetric atlas creation

Once all data were in MNI space, population-based atlases were created by following methods previously used to create tractography atlases[28, 19, 152]. For each pathway, the binarized maps were summed and set to a probabilistic map between 0 and 100 population overlap (Figure 2.2, Volumetric Atlas). Thus, each pathway was represented as a 3D volume, and concatenation of all volumes results in the 4D volumetric atlas. Atlases were additionally separated based on the method used to create the atlas, as well as separated by dataset (BLSA, HCP, VU) if population-specific or method-specific analysis is required.

2.3.4 Suface-intersection atlas creation

To overlay each pathway onto the MNI template surfaces, a standard FreeSurfer pipeline[47] was used to reconstruct the white/gray matter cortical surfaces directly from the MNI ICBM template image. Each of the probability maps overlaid over the volumetric atlas was then transferred to the reconstructed surfaces to create the surface atlas. However, the reconstructed cortical surfaces do not necessarily guarantee unique voxel-to-vertex matching (normally, more than one vertex belongs to a single voxel) even if they perfectly trace the white- and gray-matter boundary. This potentially degenerates vertex-to-voxel mapping without a voxel-wise resampling scheme. Therefore, the probability to a given vertex was obtained by tri-linear resampling of the associated voxel for sub-voxel accuracy.

2.3.5 Data visualization and validation

Qualitative validation of the atlases included pathway visualization as an overlay of the population probability on the MNI ICBM template image, or visualization of population-probability on the white matter/gray matter surface. These displays were used in QA during atlas creation, ensuring acceptable probability values, as well as agreement with expected anatomy, shape, and location. To quantify similarities and differences across pathways and methods, a pathway-correlation measure was used. The pathway-correlation was calculated between two pathways by taking the correlation coefficient of all voxels where either pathway has a probability > 0. This correlation coefficient ranged from -1 to 1, where a value of 1 indicates a perfect correlation of population densities. Thus, this metric measures the coherence between population maps obtained from the bundles and was used to assess if the distribution of population probabilities in space is similar. We used this measure to test similarities/differences between the pathways from different bundle dissection methods (to justify the use of different tractography methods) as well as between pathways generated from the different datasets (to justify making available atlases separated by dataset, as well as understand differences in results based on populations). Finally, a uniform manifold approximation and projection (UMAP)[108] was used for dimensionality reduction in order to further assess similarities and differences in pathways across methodolo-

gies. The UMAP is a general non-linear dimension reduction that is particularly well suited for visualizing high-dimensional datasets.

2.4 Technique validation

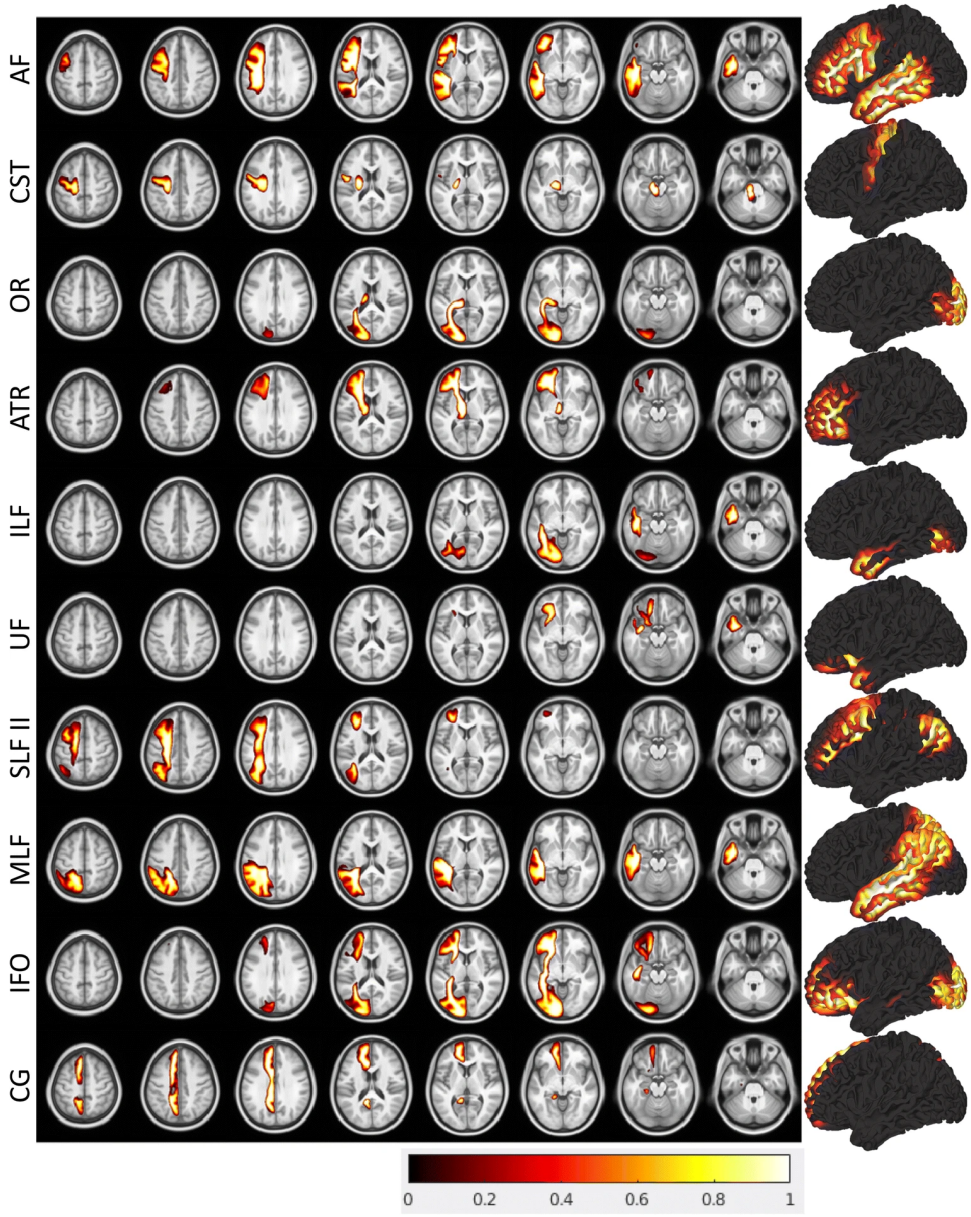


Figure 2.3: Visualization of data contained in example volumetric and surface atlases. Example visualization for 10 pathways in the TractSeg nonlinear atlas are shown as both overlays and surfaces.

We begin with a qualitative validation of the data, thoroughly inspecting and visualizing all volumes and surfaces from each atlas. An example visualization for 10 selected pathways from the TractSeg sets of atlases is shown in Figure 2.3. All pathways overlay in the correct location, with the correct shape and trajectory, as

expected. Population agreement is generally high in the core of the bundle (values 1) with larger variability along the periphery of pathways. Through this qualitative validation process, differences in the methodologies were noted including some possessing high sensitivity (larger volumes, greater agreement across subjects) and those with higher specificity (smaller, well-defined pathways with lower population agreement).

Next, to assess differences within and between tractography techniques, we show pathway-correlations against all other pathways as a large 216x216 matrix of correlations (Figure 2.4, a) and also plotting the UMAP projection of each pathway on a 2D plane (Figure 2.4, b). As expected, most pathways are quite different from others (for example we do not expect the optic radiations to share any overlap whatsoever with the uncinate fasciculus, regardless of methodology), however there are clearly clusters of pathways sharing some similarity, due to both spatial overlap of pathways with comparable anatomies (for example inferior longitudinal fasciculus and inferior frontal occipital fasciculus), as well as methods representing the same pathway. We identified a core group of 20 pathways that are commonly dissected in all methods, and clusters of these pathways are apparent in the UMAP projection (for example, the corticospinal tracts, forceps major and minor, optic radiations, and inferior longitudinal fasciculi are quite similar across algorithms). Thus, certain pathways are similar, but not exactly the same, across methodologies, justifying the use of all six state-of-the art methods for bundle dissection. Finally, we quantify differences across datasets by showing boxplots of the pathway-correlations after separating by source of data (Figure 2.4, c). While all methods show quite high correlations, it is clear that BLSA and VU datasets and bundles are more similar to each other than to HCP datasets. This is expected as HCP data quality, SNR, resolution, and acquisitions are quite different from the more clinically feasible BLSA and VU sets. Thus, bundles are also different based on dataset source. Because of this, in addition to combining results from all subjects, we also supply atlases separated by dataset.

2.5 Discussion

Here, we have created and made available the white matter bundle atlas, that addresses a number of limitations of current human brain atlases by providing a set of population-based volumetric and surface atlases in standard space, based on a large number of subjects, including many pathways from multiple diffusion MRI tractography bundle segmentation methods. We envision the use of these atlases for spatial normalization and label propagation in ways similar to standard usage of volumetric brain atlases. These labels can be used not only for statistical analysis across population and individuals, but also for priors for tractography, relating neuroimaging findings to structural pathways or to inform future methodologies for parcellating and segmenting white matter based on functional, molecular, or alternative contrasts. Similarly, although much less frequently used in the field, the surface-based atlas can also be used to relate functional MRI findings (which

are largely applied to cortex, with some evidence for signal contrast in white matter), as priors for cortico-cortical tractography and future bundle segmentations, as a tool for gray matter based spatial statistics, and again for relating alternative neuroimaging findings to structure.

As a simple example workflow. An investigator may be interested in relating tumour localization on a structural image to specific white matter pathways hypothesized to be involved in some functional network. The investigator may choose to register their image to the MNI template, and can either warp their data to template space or apply the inverse transform to get white matter labels into the subject native space. The investigator could then relate tumour location to the probability of given pathways, or could simply threshold the probabilistic maps at a given threshold (for example 0.5) and relate these to the existence/non-existence of the bundle being displaced by the tumour. We currently recommend the use of the concatenation of all datasets for standard investigative studies unless a population-specific template is required. While differences between datasets are clear and expected, the increased population variability that results from including data from all sources is likely an advantage when investigators are using their own data with possible differences in acquisition, resolution, and subjects. However, future work will investigate creation and dissemination of age-specific white matter analysis, as well as including an age-adjusted surface mesh instead of using the MNI template to generate the surface. We have chosen to include a large number of algorithms for streamline generation and bundle dissection. Our results (Figure 2.4) show that even if the same white matter structure is segmented using different techniques, the results are not guaranteed to be the same. This is because different algorithms or workflows may define bundles in different ways, with different approaches taken to segment the structure of interest. Thus, an investigator could use our atlas with the set of protocols that they agree with most, or alternatively, could relate findings to all white matter pathways across all methodologies in our atlas. We note that we have chosen six standard algorithms to create this atlas, although others exist and new ones are continually developed based on improvements in both our understanding of anatomical connections and our ability to reconstruct these connections with tractography. These methods were chosen because they are fully automated, and robust, bundle segmentation techniques that can be easily run on several thousand diffusion datasets. Inclusion of other tractography and/or segmentation methods are likely additions in future iterations of the atlas, and are easily integrated with existing deformation fields and data organization. The addition of tract orientation maps[169] or orientation-density maps[125, 126]may facilitate the development of bundle segmentation algorithms or act as priors for bundle specific tractography. Finally, future iterations can include variations and concatenations of gray matter and/or regional atlases in the same space, continually adding to the number of features to be investigated with a single dataset in standard space.

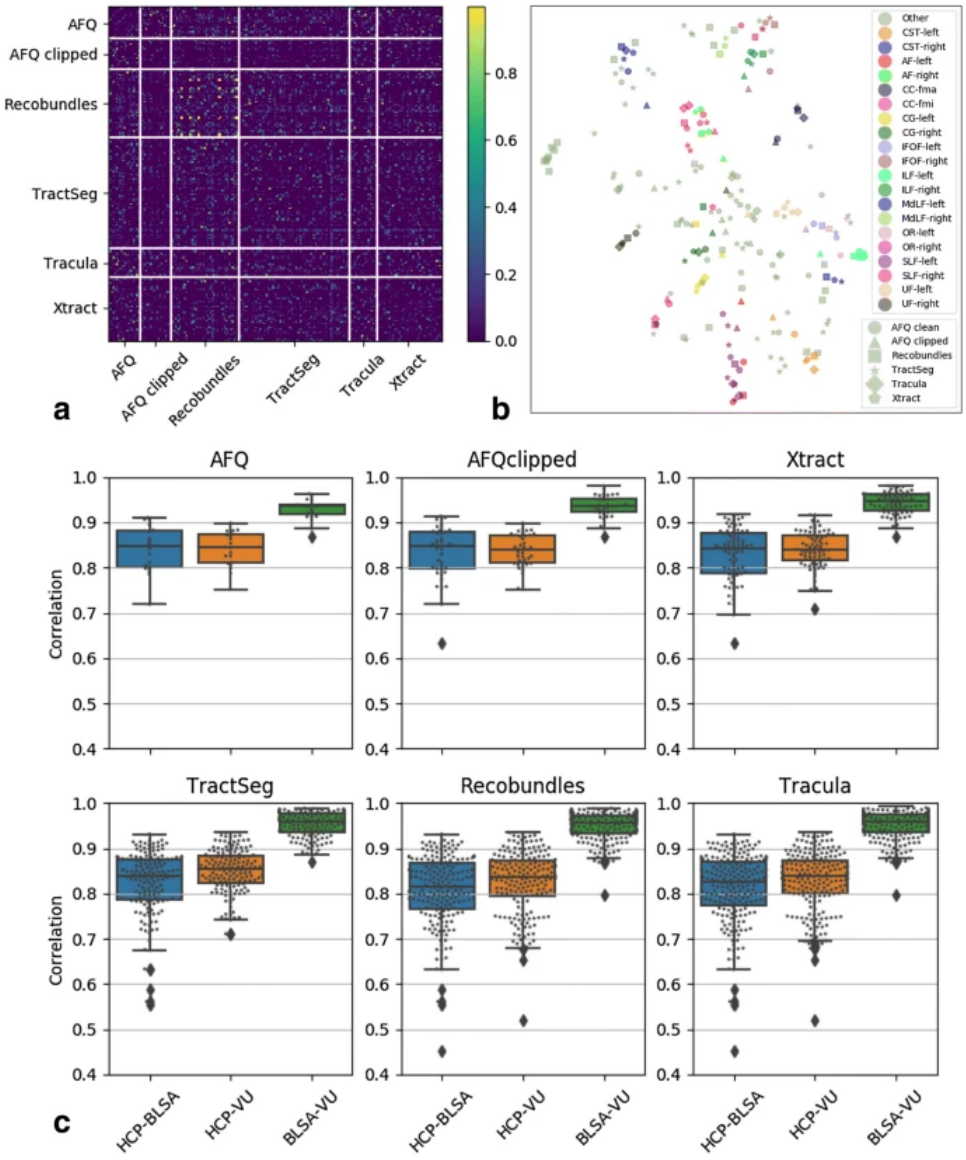


Figure 2.4: Data validation. (a) Matrix of correlation coefficient of pathways plotted against all others indicates similarities within and across methodologies for bundle dissection. Solid white lines are used to visually separate bundle segmentation methods. (b) UMAP dimensionality reduction projected onto un-scaled 2D plane shows that many WM pathways are similar, but not the same, across methods. Object colors represent specific atlas bundles, with shape indicating segmentation methods. (c) Correlation coefficient of atlases separated by dataset indicates small, but significant, differences between datasets. Together, these justify the inclusion of all tractography methods, as well as separation of atlases by datasets.

CHAPTER 3

Subject-specific segmentation of white matter based on deep learning

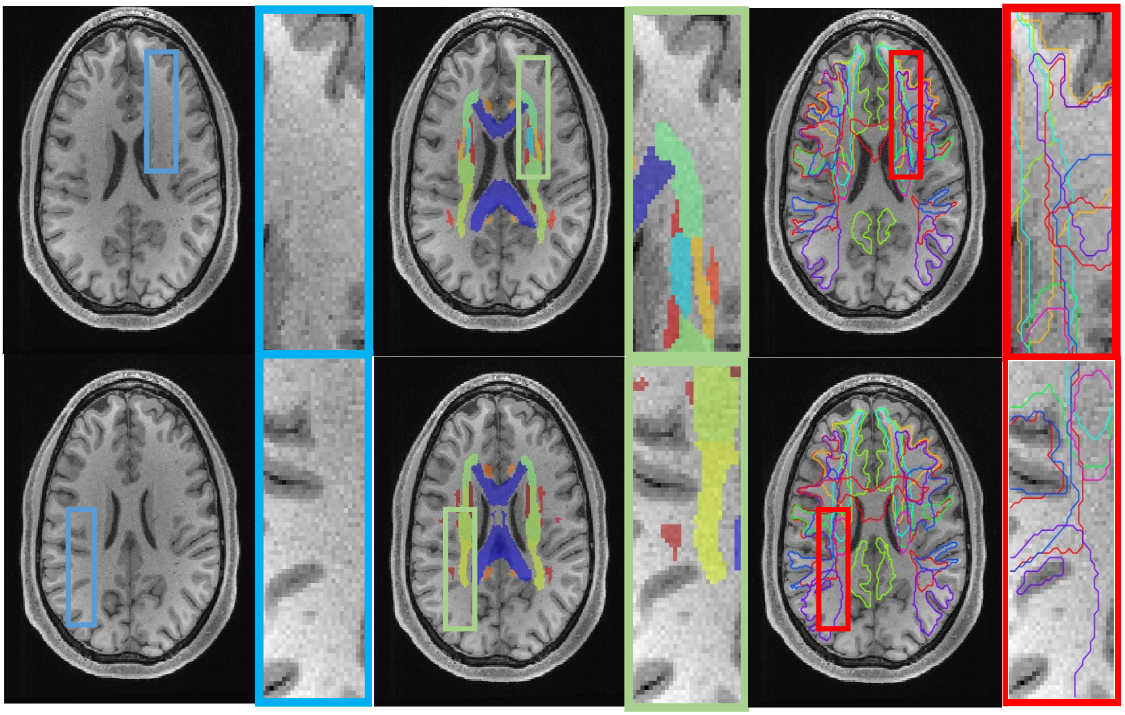
This work was previously published [174]. Permission to include the work as part of the dissertation has been obtained, see Appendix A.

3.1 Introduction

As mentioned in chapter 1, dMRI can be used to non-invasively probe the connectivity and microstructure of human brain tissue in vivo[127]. It allows for creation of tractograms of whole brain and parcellation of specific white matter pathway. However, legacy or time-constrained study only has T1w instead of dMRI. It is challenging to annotate white matter pathways from T1w since it has little contrast for white matter. Deriving label only from T1w is our discussed topic in this chapter.

3.2 Related work

Image registration is an established way of transferring different WM labels from population-based atlases to T1w MRI and can isolate different WM regions in T1w MRI. In general, WM atlases from the dMRI community can be divided into two categories: streamline-based atlases[28, 61, 180, 115] and volumetric atlases[65, 110, 116]. Streamline-based WM atlases contain streamlines assigned to various WM pathways, while volumetric WM atlases contain labels indicating the pathway assignment(s) of a given voxel. One such widely used volumetric atlas was proposed by Mori et al. and recognizes 48 different WM labels.[110] WM atlas are very popular in neuroimaging analysis but have key limitations. They require that different WM regions are not overlapping and often contain a limited amount of information outside deep WM (Figure 3.1). To navigate these limitations, Hansen et al. recently proposed the Pandora WM bundle atlases, which are volumetric atlases that present 216 overlapping WM pathways from 2300 healthy subjects. This approach has subsequently allowed for both the identification of overlapping pathways and improved WM labeling outside the deep structures on T1w MRI without dMRI[65].



(a) T1w brain MRI. (b) Atlas overlay (c) Tractography overlay

Figure 3.1: (a) WM is largely homogenous when imaged using most sources of MRI contrast, for example, T1w (left).(b) Traditional WM atlas (center) represents each voxel with one tissue class. (c) Modern approaches to bundle segmentation identify multiple overlapping structures (right). Diffusion tractography offers the ability to capture a multi-label description of WM voxels.

Compared with a large cohort population atlas, deep convolutional neural networks have the potential to capture subject-specific variations. Among convolutional neural networks, the U-Net[29] has obtained impressive results for performing 3D medical image segmentation including brain[72] and abdomen[151]. Brebisson et al. proposed a deep neural network learning 2D and 3D patches from structural brain MRI to predict the anatomical class of each voxel[33]. DeepNat leverages a hierarchical multi-task network to achieve brain segmentation with 3D patches[164] SLANT[72] learns spatially localized 3D patches from structural MRI to achieve brain structure segmentation. Additionally, current deep learning approaches[33, 7] have demonstrated superior performance compared with atlas-based methods on healthy brain segmentation from structural images.

3.3 Materials and Methods

The pipeline of predicting WM labels directly from T1w MRI with deep learning includes four steps: tractography, registration, normalization and patch-wise networks as shown in Figure 3.2.

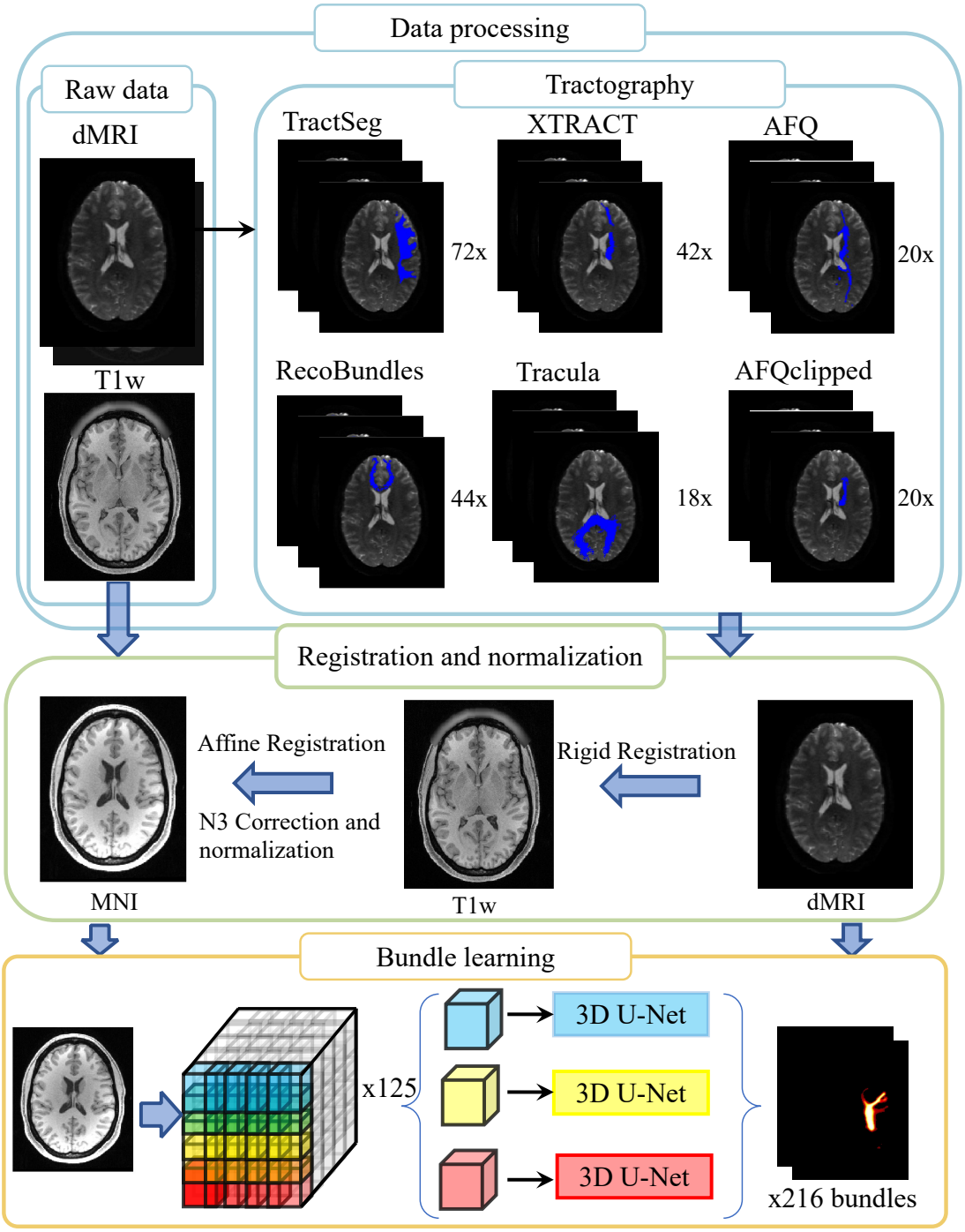


Figure 3.2: The pipeline of proposed WM bundle learning is presented, which integrates data processing and registration as well as bundle learning. We extract WM bundles from six different tractography methods. Structural images and corresponding tractograms are reoriented to the MNI template. Patch-wise, spatial-localized neural networks are utilized to learn WM bundle regions from a T1w MRI image. The output of each U-net is merged as the final step before segmentation. Representative samples of WM bundles acquired from six automatic tractography methods and the final learning result is visualized.

Dataset Name	T1w voxel size(mm)	Diffusion voxel size (mm)	B-value	Diffusion volume
BLSA	1.0×1.0×1.0	0.81×0.81×2.2	700	1B0 + 32DWIs
HCP	0.7×0.7×0.7	1.25×1.25×1.25	1000,2000,3000	(6B0 + 90DWIs)×3
VU*	1×1×1	2.5×2.5×2.5	1000	1B0 + 64DWIs
HCP_LS	0.8×0.8×0.8	1.5×1.5×1.5	1000 2500	5B0 + 76DWIs
IXI	0.93×0.93×1.2	1.75×1.75×2.35	1000	1B0 + 15DWIs
UG	1×1×1	2×2×2	2000	6B0 + 48DWIs

Table 3.1: Dataset descriptions. * represents one typical case selected from the VU dataset.

The algorithm name	Number of generated bundles for each algorithm
TractSeg[169]	72
RecoBundles[53]	44
XTRACT[168]	42
Tracula[182]	18
AFQ[179]	20
AFQclipped[179]	20

Table 3.2: The tractography algorithms and corresponding generated pathways.

3.3.1 Data

We use 2,416 de-identified images from the Baltimore Longitudinal Study of Aging (BLSA)[45], 1,105 images from Human Connectome Project (HCP) S1200 release[160], and 349 images from Vanderbilt University (VU) to train all deep neural networks. We also select three open-source datasets to perform external validation to test the generalizability of the proposed learning method. We study 26 images from HCP lifeSpan (HCPLS)[160], 394 images from IXI (IXI, http://brain-development.org/ixi-dataset), and 12 images from the Unilateral Glaucoma dataset (UG, https://openneuro.org/datasets/ds001743/versions/1.0.1). All above images include paired T1w MRI and dMRI. The voxel resolution of T1w MRI, voxel resolution of dMRI, the B0-value and diffusion volumes are shown in Table 4.1.

3.3.2 Tractography

dMRI is often subject to artifacts, which can deteriorate the accuracy of extracting WM bundles. To correct these artifacts, we perform correction for susceptibility distortions, subject motion, eddy currents, and b-tables prior to analysis[21]. We perform tractography on preprocessed dMRI. We select six popular tractography algorithms to recognize pathways and annotate WM bundles. All six algorithms were run using default parameters. The number of generated pathways of all six tractography algorithms is shown in Table 4.2.

Note that AFQclipped clips the center of each of the 20 AFQ bundles with ROI based exclusion criteria. According to anatomic name for each bundle, we notice that there are ten common bundles across from all six tractography algorithms. Thus, we do not regard 216 bundles from TractSeg, RecoBundles, XTRACT, Tracula, AFQ and AFQclipped as unique bundles.

3.3.3 Registration and intensity normalization

To ensure that all inputs have the same voxel size and dimension, we register all pathways derived from dMRI through all six bundle segmentation algorithms and T1w MRI, to the Montreal Neurological Institute (MNI) ICBM 152 asymmetric template[49]. First, we rigidly register the b = 0 s/mm2 volume of each dMRI to the T1w MRI of the same subject using FSL[145]. Then, after performing N3 correction of bias field and normalization of white matter intensity by FreeSurfer[47] on raw T1w MRI, corrected T1w MRI is registered to the MNI template with antsRegistrationSyn in ANTs[5]. By linking these registration steps, all pathways are rigidly registered to T1w MRI of the same subject. All pathways are affine reoriented to the MNI template and serve as ground truth. The affine transformation is also applied to the raw T1w MRI. After registration, we skull strip all structural images on the MNI template with the bet tool in FSL, clip, and normalize the background and the 98th percentile of within-brain intensity to intensity units 0 and 1. The pipeline of registration and normalization is visualized in the Figure 3.2

3.3.4 Patch-wise network

After transforming all image and ground-truth pathways to MNI template, the high-resolution image volume could not fit into the 12G GPU (GTX 1080Ti) memory using current popular network architectures. Inspired by SLANT[72], we designed 125 overlapped 3D U-Nets to cover the entire MNI volume and subdivide each image into corresponding 125 patches. Each patch ψ_n represented by one coordinate (x_n, y_n, z_n) and patch size (d_x, d_y, d_z),$n \in 1, 2, ...125$

$$\psi_n = [x_n : (x_n + d_x), y_n : (y_n + d_y), z_n : (z_n + d_z)] \tag{3.1}$$

Where ψ_n represnets the nth patch,x_n, y_n, z_n represent the corner coordinates of the nth patch. $x_n$and $z_n \in [1, 25, 50, 74, 98]$ and $y_n \in [1, 34, 67, 101, 134]$. The length d_x, the width d_y and depth d_z is 96.

To merge the outputs of the U-Nets after training, the pixel-wise output represents an activation value of the neural network rather than specific WM pathways. Thus, the average way is adopted to get the final value instead of majority vote:

$$p_{whole}(i) = \frac{1}{n_k} \sum_{k=1}^{n_k} p_k(i) \tag{3.2}$$

Where p_{whole} represents all pixels within the structural image and p_whole (i) means the ith pixel. k indexes the U-Nets that covers ith pixel. p_k (i) represents the final value of ith pixel of kth U-net. Networks not covering a particular voxel are excluded in the final merge process.

	Scans of train	Scans of validation	Scans of test	Scans of external validation
TractSeg	2803	213	754	431
RecoBundles	2789	211	754	430
XTRACT	2786	211	751	427
Tracula	2538	189	693	428
AFQ	2730	201	726	367
AFQclipped	2730	201	726	367

Table 3.3: number of scans for the training, validation, testing cohorts and external dataset.

Corner coordinate index	Corner coordinate (x,y,z)
2,2,2	25,34,25
2,4,2	25,101,25
4,2,2	101,25,25
4,4,2	101,101,25
3,3,3	50,67,50
2,2,4	25,34,74
2,4,4	25,101,74
4,2,4	101,25,74
4,4,4	101,101,74

Table 3.4: Corner coordinates of pre-trained nice models out of 125 models, indexed starting at one.

3.3.5 Implementation details

We divided the HCP, BLSA, and VU data into training, validation, and test cohorts evenly based on subjects and used HCPLS, IXI and UG as the external dataset. We kept the splitting strategy consistent across learning all six diffusion tractography algorithms. To remove data corrupted by registration or failed diffusion tractography algorithms, exhaustive human review was performed on verifying acceptable image registration and inspecting appropriate shape and location of all bundles[65]. The resultant number of scans for the training, validation, testing cohorts and external dataset is shown in Table 4.3

Inspired by the AssemblyNet[30], we adopt a transfer learning technology to utlize the weights from trained U-Net to initialize the nearest U-Nets. In the beginning, we trained nine U-Nets and their corner coordinates are shown in Table 3.4. Then, the trained nine models in the Table 3.4 order are used to initialize the nearest 116 models (every model is trained only once).

We used pytorch[121] to implement baseline U-Net[29] as the convolution neural network to learn patches from anatomical images and set the batch size 1. The output channel depends on the number of WM bundles recognized by the bundle segmentation algorithm. We set a learning rate of 0.0001 and do not perform learning rate decay during the training process. We adopted the sum binary cross-entropy for each effective WM bundle as a loss function and train all models using the Adam optimizer. When we inferred the WM regions based on deep neural networks, we appended a sigmoid function to the output of each patch-wise neural network to map the final merged output to [0,1].

3.3.6 Baseline methods

We compare the quantitative performance of transferring labels with the traditional atlas-based approach as the baseline method. Here, we use the Pandora atlas[65], which is a 4D collection population-based

atlases. The Pandora atlas used the same cohorts and diffusion tractography algorithms to generate each corresponding WM bundle same as we learn in this study. All volumes of the Pandora atlas are on the same MNI template as we use here. Each volume of the 4D atlas is in the form of a probability map indicating the probability of a pixel being in a specific WM bundle. To make a fair comparison, we transfer the label from the atlas to the affine reoriented MNI template as the final probability map for each target scan.

In addition to the atlas-based method, we use multi-atlas segmentation (MAS) as another baseline method. We selected 20 subjects whose tracts all passed human review as the single atlas. There are 9 subjects from HCP, 9 subjects from BLSA, and 2 subjects from VU among 20 subjects. To make sure the output is in the same template as other comparison methods, all 20 atlas and target images are affine reoriented to the MNI template as well as the corresponding white matter tracts. All 20 atlases are registered to target images through non-rigid registration[137]. Then, the new labels are obtained through joint label fusion[166] methods. The whole pipeline of MAS is implemented by ANTs[5] python package and performed for all 216 bundles for each target scan.

3.3.7 Metrics

To evaluate the accuracy of our proposed method, we compare the segmentation results against the ground truth provided by diffusion tractography. Additionally, we compare the accuracy of the proposed method against the accuracy achieved by Pandora atlas and MAS. To quantify the agreement between segmentation and truth, we use four measures: Dice coefficient (DSC), average symmetry surface distance, bundle overlap, and bundle overreach. We use DSC as the main evaluation measurement for different bundle segmentation algorithms by comparing binary WM bundle prediction against the ground truth voxel-by-voxel:

$$DSC = \frac{2|R \cap T|}{|R| + |T|} = \frac{2|TP|}{2|TP| + |FP| + |FN|} \tag{3.3}$$

where TP is true positive, FP is false positive, FN is false negative, R represents the segmentation result generated by the proposed method or atlas-based method and T represents the corresponding ground truth. Average symmetry surface distance[67] is given in millimeters and based on surface vertices between the proposed or atlas-based segmentation, R, and the ground-truth segmentation, T. For each vertex on the surface of R, $(S(R))$, the Euclidean distance to closest surface vertices of truth $(S(T))$ can be defined in $d(S_R, S(T))$:

$$d(S_R, S(T)) = min_{S_R \in S(R)} ||S_R - S(T)||$$
$$ASSD = \frac{1}{|S(R)| + S|T|} \left(\sum_{S_R \in S(R)} d(S_R, S(T)) + \sum_{S_T \in S(T)} d(S_T, S(R)) \right) \tag{3.4}$$

where $|S(R)|$ represents the number of vertices of the resulting surface and $|S(T)|$represents the number of vertices on the ground-truth surface. S_R represents a vertex from the atlas-based or proposed segmentation. S_T represents a vertex from the ground truth. Bundle overlap[138] is the proportion of voxels that contain the ground truth region that is also overlapped by the results of the learning- or atlas-based methods.

$$OL = \frac{R \cap T}{T} = \frac{|TP|}{|TP| + |FN|} \tag{3.5}$$

Bundle overreach[138] is the number of voxels containing results from proposed or atlas-based methods that are outside of the ground truth volume divided by the total number of voxels within the ground truth.

$$OR = \frac{R \setminus T}{T} = \frac{|FP|}{|TP| + |FN|} \tag{3.6}$$

where operator $\setminus$ denotes the relative complement operation The non-parametric Wilcoxon signed-rank test[171] for paired distributions was used to calculate test significance when comparing learning-based results with corresponding atlas-based results.

3.4 Results

3.4.1 Fine-tune binary threshold

The outputs of the MAS, atlas-based and proposed methods have been mapped to [0,1] and represent a probability that a given voxel is included in the WM pathway. The binary threshold to convert the probability to a yes or no is important and influences the performance of both the atlas-based and proposed methods. Starting from 0, we sweep thresholds until 1 with a step size of 0.01, using the validation datasets to calculate mean DSC across all WM pathways of all scans. The optimal threshold and the curve of relationships between mean DSC and binary threshold for the MAS, atlas- and learning-based methods are shown in Figure 3.3. The optimal thresholds are the values where the mean DSC across all pathways from all scans are highest for MAS, atlas- and learning-based methods.

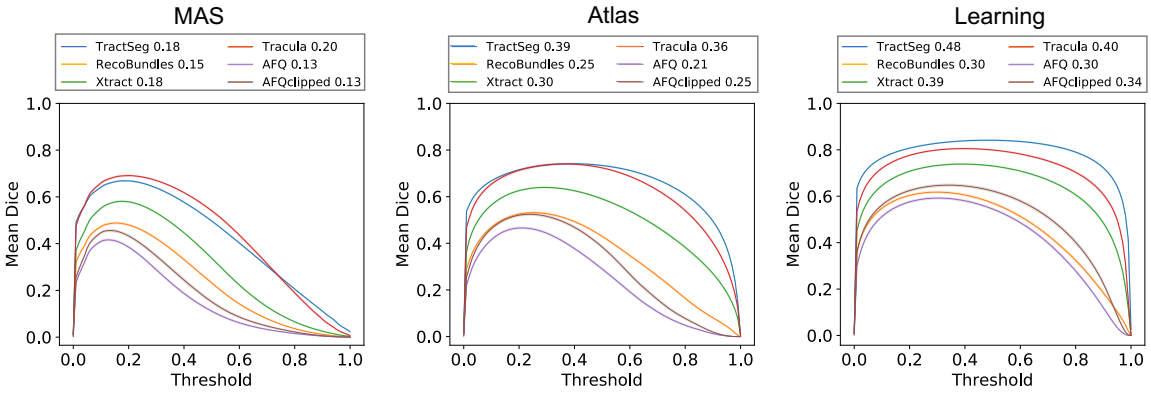


Figure 3.3: Each curve represents the average DSC of all WM bundles of all validation dataset scans per diffusion tractography algorithms for MAS, atlas- and learning-based methods at different threshold values. The 95 percent confidence interval is within the printed notch due to the large sample population size. The legend above each plot includes the optimal threshold for each tractography algorithm.

3.4.2 Qualitative results

We select one scan from the HCP test cohort to visualize the left corticospinal tract (CST) across all six bundle segmentation algorithms to see an intra-subject variance of bundle segmentation algorithms and visualize the difference between results derived from T1w images and ground truths from dMRI. We use the optimal threshold values calculated in Table 5 to binarize each output, using a marching cube[142] to extract and render the CST surface. 3D visualization is shown in Figure 3.4 .

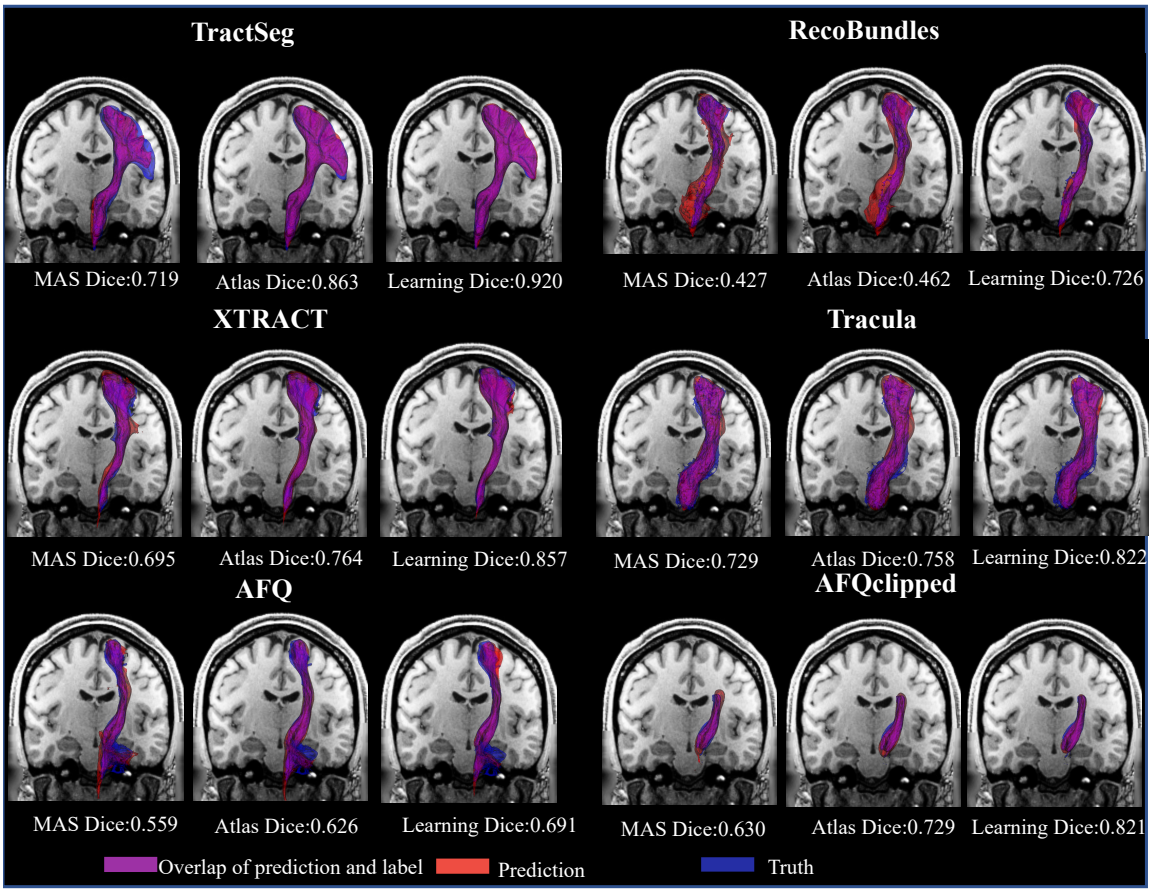
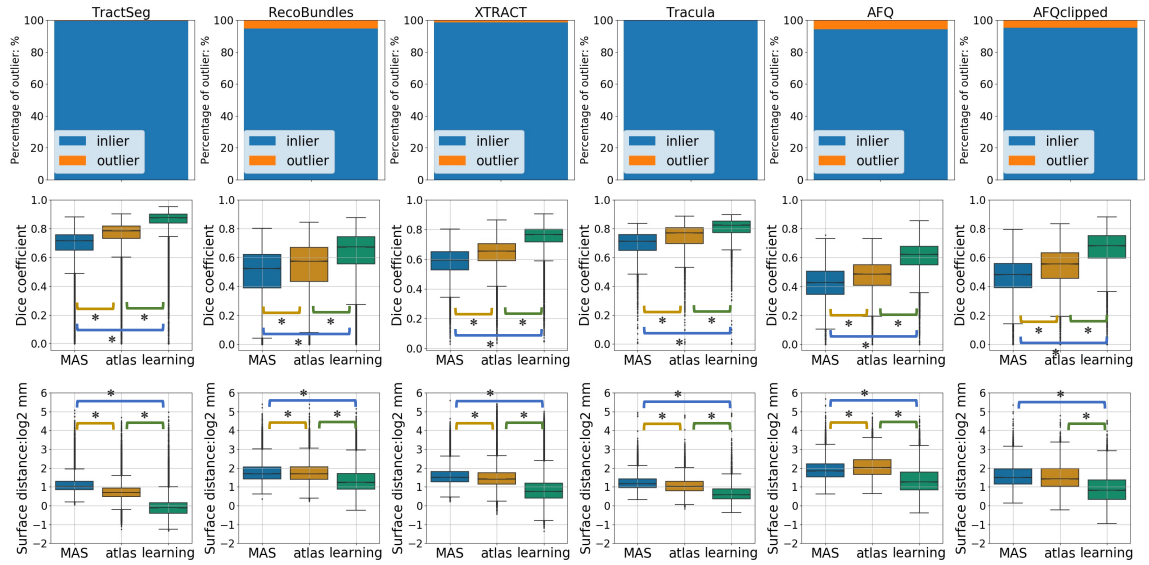


Figure 3.4: 3D visualization of MAS, atlas- and learning-based results across six diffusion tractography algorithms by reconstruction of the left corticospinal tract (CST) surface on an affine reoriented coronal T1w MRI slice. The text below each image is quantitative DSC for each case.

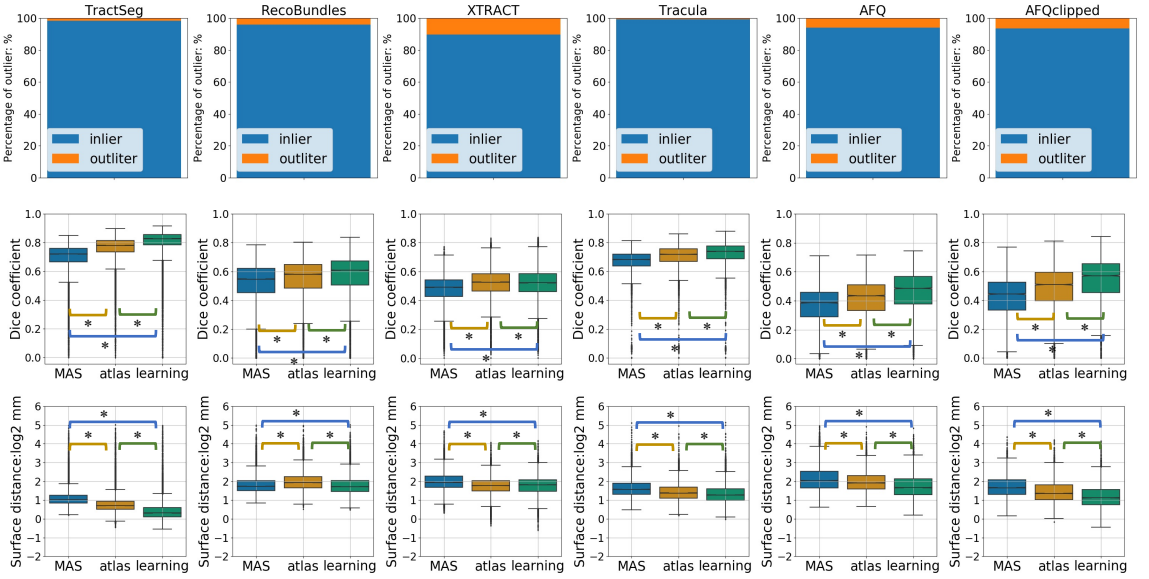
From Figure 3.4, we find the learning-based method per bundle segmentation algorithm has a higher overlap compared with the atlas-based method and MAS according to the areas of magenta overlap for this subject. Although the pathway of all six tractography algorithms has varied shapes, the learning method still can make good predictions for the largest pathway of TractSeg and the smallest pathway of AFQclipped.

3.4.3 Quantitative results

We used the optimal threshold values fine-tuned from the validation datasets to binarize the output on the testing and external datasets. To examine their overall performance, we evaluated all 216 bundles using the DSC and average symmetry surface distance (Figure 3.5).



(a) Evaluation on test cohort



(b) Evaluation on external cohort

Figure 3.5: Quantitative results of MAS,atlas-based method, and proposed learning methods on test cohorts from HCP, BLSA, and VU and external cohort from HCPLS, IXI and UG. The outlier percentage (top row) of all six algorithms is shown in the bar plot. Two measures are used to assess the overlap between algorithms deriving fiber mask from T1w and truth from dMRI: Dice (middle row) and surface distance (lower row). Each column presents the result of a different bundle segmentation algorithm and shows the proposed method, MAS, and single atlas-based method. Each boxplot includes each pathway of the bundle segmentation algorithm per every scan. The 95 percent confidence interval is within the printed notch due to the large sample size. The difference between methods was significant ($p < 0.005$, Wilcoxon signed-rank test, indicated by $*$).

From Figure 3.5(a), the blue bar plot represents the percentage of pathways that successfully passed the human reviewing process across the whole test cohort. All learning-based methods perform statistically better than the atlas-based methods and MAS. When using ground truths derived from TractSeg, the MAS, atlas-

and learning-based methods achieve the highest median DSC of 0.72, 0.78, and 0.87 and the smallest average symmetry surface distance of 2.04mm, 1.62 mm, and 0.92 mm respectively. Compared with the atlas method, the learning method shows the largest improvement in median DSC for AFQ from 0.48 to 0.62 and reduces the median average symmetry surface distance from 4.08 mm to 2.40 mm. The same pathway for different subjects may have varied shapes and localization. The proposed method is able to adapt to these differences more robustly than atlas-based methods.

In Figure 3.5(b), the blue bar plot represents the percentage of pathways that successfully passed the human reviewing process across the external dataset. All learning-based methods perform statistically better compared to atlas-based methods except for XTRACT. However, the difference between the atlas- and learning-based methods is less pronounced. The median DSC of the learning-based method on XTRACT is 0.522, lower than 0.527 of the atlas-based method. Compare 3.5(a) to 3.5(b), variations in MRI contrast across scanners decrease the performance.

We perform bundle overlap and bundle overreach on the left CST pathway of all six bundle segmentation algorithms to analyze the relationship between the spatial overlap of the proposed method and threshold (Figure 3.6).

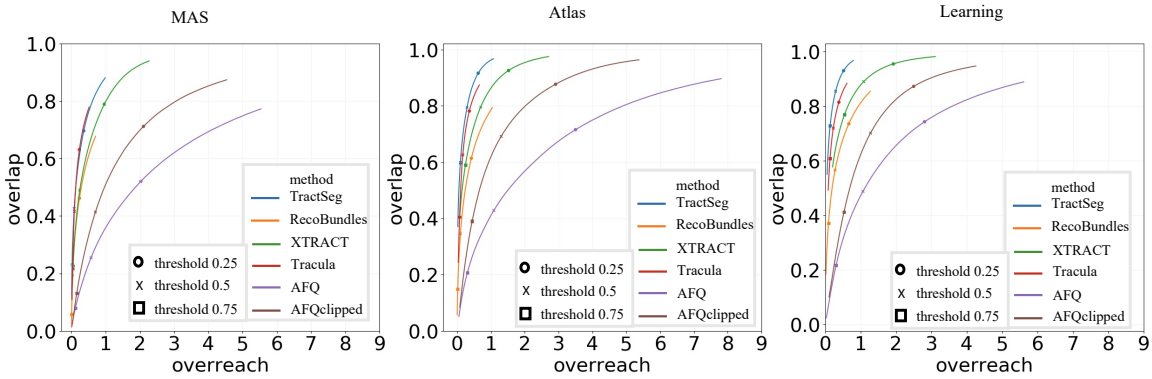


Figure 3.6: Plots of overlap versus overreach for the left CST across all bundle segmentation algorithms for MAS, atlas- and learning-based methods are shown. The markers on each curve represent the overlap and overreach values at specific threshold values. The range of overreach for MAS is [0,6]. The range of overreach for atlas-based methods is [0,9]. The range of overreach for the learning-based method is [0,6].

From Figure 3.6, all MAS, atlas- and learning-based methods for all six diffusion tractography algorithms identified WM bundles with a high overlap but suffer from high overreach except for Tracula and TractSeg. As for AFQ, when the overreach is about 5 times the actual ground truth volume, the MAS method has overlap values of 0.75 and 0.8 with the atlas-based method. The proposed method has an overlap value of 0.9 higher compared with the MAS and atlas-based methods.

We calculated the overall binary thresholds from the validation cohort. Thus, we want to investigate

whether the optimal thresholds calculated from validation datasets can be generalized to external datasets. We show the curve of the relationship between DSC and binary threshold on the external datasets in Figure 3.7 Comparing Figure3.7 to Figure 3.7, the biggest difference between thresholds estimated from the validation dataset and the external datasets is in XTRACT. The binary threshold for MAS is shifted from 0.18 to 0.34. The binary threshold for the atlas-based method in XTRACT is shifted from 0.30 to 0.56. The binary threshold for the proposed method in XTRACT is shifted from 0.39 to 0.77. We can learn that the optimal threshold shifts when fed different MRI contrasts.

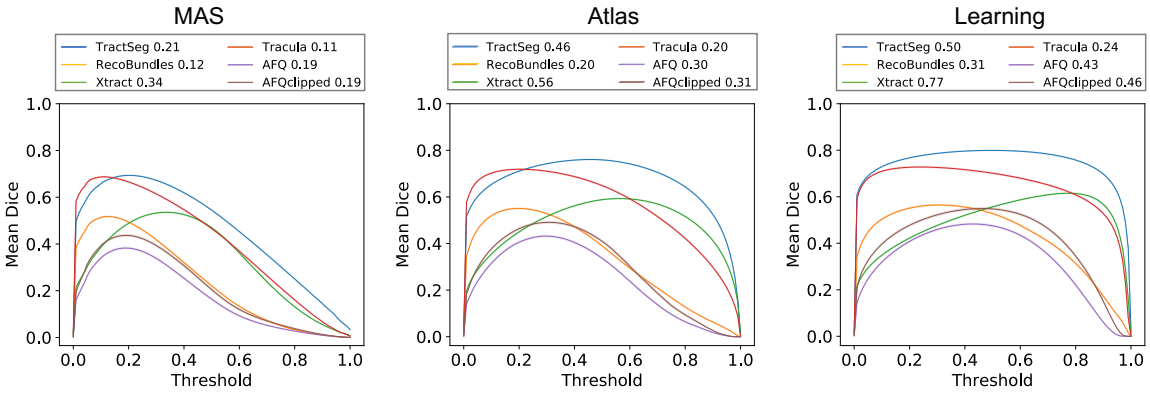


Figure 3.7: Each curve represents the average DSC of all WM bundles of all external dataset scans per diffusion tractography algorithm for atlas- and learning-based methods. The 95 percent confidence interval is within the printed line width due to the large sample size. The legend above each plot includes the optimal threshold for each tractography algorithm.

3.5 Discussion

In this study, we aim to propose a spatial localized patch-wise framework to segment white matter structure with six different definition schemes only from anatomical images. We envision this framework as a tool to estimate a coarse WM region of interest rather than segmentation with more details derived from dMRI. We provide a probability map of six different tractography algorithms for users to adjust binary threshold and choose their preference scheme.

3.5.1 Generalize to external dataset

The deep learning model is sensitive to intensity distributions of T1w MRI which are not seen in the training cohort. The different scale settings of the raw T1w images bring the shift of the optimal threshold which is shown in the external dataset. Apart from varied intensity distribution, the external dataset itself also deteriorates the performance of deep neural network. 394 IXI images domain the whole dataset. The acquisition of IXI requires 15 DWIs which is smaller compared to the original training cohort. Thus, the generated ground truth by six algorithms does not have high quality (they might have less streamlines or streamlines stop in

advance during the tracking process).

3.5.2 Limitations

Currently, the variance in performance of the learning-based method is obvious. Inherent definitions and ways of extracting WM tracts by bundle segmentation methods bring challenges to the proposed learning framework. One possible area for important is developing robust labels by merging common labels belonging to the same white matter bundle across from all six tractography algorithms. By doing this way, each white matter bundle label will include six bundle definition schemes, having common parts and specific parts for each tractography algorithm. The robust label may make full use of complementary information among six tractography algorithms. White matter pathways usually are in the form of streamlines, which can provide a connection at the sub-pixel level. However, information is lost when converting streamlines into masks, even with their own built-in function, bringing the noise to the label.

3.6 Conclusion

We propose a spatial localized patch-wise framework to delineate WM structure based on structural T1w images. We use this framework to learn WM regions under six bundle segmentation algorithms and compare the result of the framework to atlas-based methods and MAS. When optimal threshold is utilized to evaluate scans that have the same acquisition as the training datasets, the learning-based methods are statistically superior to the atlas-based methods and MAS.

CHAPTER 4

Label-efficient thigh segmentation based on transfer learning

This work was previously published [178, 176]. Permission to include the work as part of the dissertation has been obtained, see Appendix A

4.1 Introduction

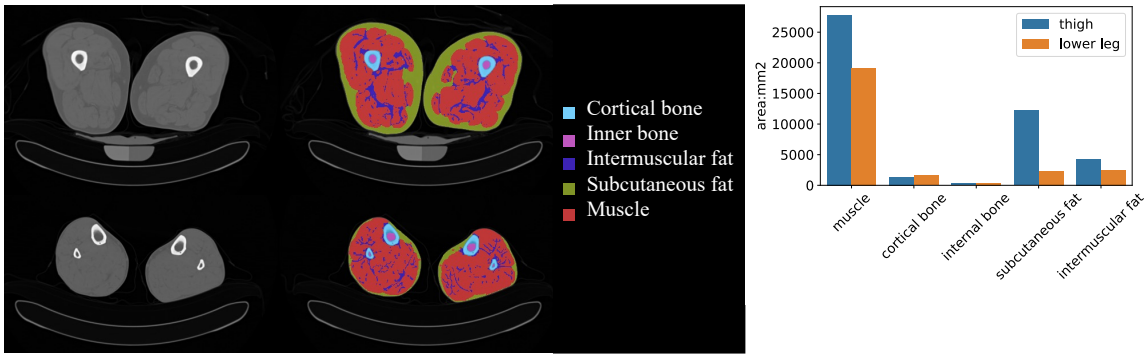
Estimating volumes and masses of total body components is important for cancer, joint replacement, and exercise physiology[76]. Full-body CT scans can be used to calculate whole-body composition directly. However, it is hard to acquire typical full-body CT in the usual medical context due to the intense radiology dose. Mourtzakis et al. proposed that body components measured on abdomen or thigh slices are highly correlated with the mass of whole-body tissues[111]. Thus, accurate segmentation of thigh slices can quantify tissue area properties to estimate body composition without requiring additional irradiation or examinations. So, this paper aims to segment muscle, fat, and bones from 2D thigh and lower leg CT slices.

4.2 Related work

Several recent techniques have been proposed to address thigh and lower leg segmentation on CT images. Senseney et al. proposed an automatic region growing method using morphology operation and threshold to extract bone muscle and fat in CT thigh and abdomen images[141]. Tan et al. proposed to use a variational Bayesian Gaussian mixture model to cluster fat, marrow, muscle bone, and air on 3D CT scans[150]. Felinto et al. proposed to use the Gaussian mixture model and relative position to cluster similar tissues for inter-muscular fat and muscle segmentation[34]. With impressive performance of the deep neural network-based segmentation, Zhu et al. applied the H-DenseU-Net on MRI lower leg data of children with and without cerebral palsy[198]. Rohm et al. created a 3D heterogeneous MRI lower leg dataset and trained a convolution network to segment muscle[128].

Deep learning methods show impressive performance in segmentation tasks. However, this performance depends on sufficient human annotation[123, 167]. In the medical imaging field, human annotation requires professional knowledge, which is very time-consuming and thus expensive. To avoid annotating new data, many researchers used common data augmentation methods such as rotation, intensity shift, and scaling, to artificially enhance the diversity and quality of the training data[144]. Image synthesis is another way for data augmentation. Generative adversarial networks (GAN)[57] have been utilized to synthesize new labeled data for segmentation. However, GAN is notorious for training and is hard to implement in practical tasks[113].

The main limitation of data augmentation is data bias generated during the data augmentation process. To preserve original data distribution, leveraging the power of unannotated data is another solution to train a model with limited annotation data. Liu et al. proposed one framework of unsupervised segmentation for medical images[100]. Chen et al. proposed to use self-supervised learning with image context restoration to achieve brain tumor segmentation with a limited dataset[27]. Instead of self-supervised learning and unsupervised learning, transfer learning is another way to train with limited label data. First, a model is trained from scratch on a large-scale dataset with a similar task. Then the model is fine-tuned with human annotated data. Tajbakhsh et al. showed that a fine-tuned network could outperform networks that were trained from scratch with better robustness[147]. To better segment muscle, cortical bone, internal bone, subcutaneous fat, and intermuscular fat with limited annotated data, we propose a novel two-stage transfer learning-based framework. We use an approximate hand-crafted method to generate pseudo labels for 1883 thighs to train the model in the first stage and fine-tune with 125 human label thighs in the second stage to achieve segmentation. We test the model on the thigh slice and use the lower leg slice as external data to demonstrate the generalizability of the proposed framework. The target tissue and corresponding legend can be found in FigIV-1. This paper is a significant extension of our accepted work[178] of SPIE 2022.



(a) (b)

Figure 4.1: The first row and second row of (a) represent the middle thigh and lower leg from the same subject respectively. The left column is the original CT image and the right column is the target tissue label. Each tissue has a different area, and the imbalance of area makes segmentation of sparse tissue (intermuscular fat) challenging. The area of each tissue is shown in (b).

4.3 Methods

We designed a two-stage coarse-to-fine deep learning method to achieve thigh and lower leg segmentation on low-dose CT slices with deep learning. We first split a CT thigh slice into single left and right thigh images. In the first stage, we train the deep neural network with approximate hand-crafted labels. In the second stage, we fine-tune the model from the first stage with human expert labels to recover more details.

4.3.1 Preprocessing

The preprocessing pipeline works for both thigh and lower leg images with minor differences. For each thigh image, we first set the field of view of the CT thigh slice including the left thigh, right thigh, table, and phantom. Next, we use the threshold of -500 Hounsfield Unit (HU) to binarize the input thigh slice. We use a square kernel 25 $\times$ 25 to erode binary images and create three independent eroded masks. Then, we choose the left thigh and right thigh according to area size (the area of the table mask should be smaller compared with the thigh mask) and center position (the center of the left thigh mask and right thigh mask should be at approximate horizontal axis). After picking the eroded mask of two thighs, we dilate the chosen mask with the same kernel size. Based on those two masks, we find the maximal bounding box for each thigh and crop the original CT slice from 512 $\times$ 512 to 256 $\times$ 256 without changing the pixel resolution and intensity range of the whole CT slice. Different from preprocessing on the thigh, we use kernel size 10 $\times$10 to erode and dilate the mask of the lower leg. Finally, we manually review all the thighs and lower legs and exclude cropped images including other tissue (e.g. the table).

4.3.2 Create a pseudo label for thigh

Each CT slice has specific intensity units for each tissue. We use a CT window of [-190, -30] HU for fat, [30,80] HU for muscle, and [1000,inf] HU for bones[43]. We proposed the following pipeline with seven steps to extract five target tissues coarsely by using CT intensity and morphology. (1) create a cortical bone binary mask image with a threshold of 1000. (2) inverse the cortical bone mask from step (1) and find the surrounding internal bone. (3) use a threshold of 0 HU to binarize the thigh image and create a muscle mask. (4) fill the holes and remove bones from steps (1) and (2). (5) subtract the muscle mask from step (4) to create an intermuscular fat mask based on the assumption that intermuscular fat is within the muscle. (6) binarize the input image with a threshold of -500 HU. (7) subtract the result of step (4) from step (6) to create a subcutaneous fat mask. Five coarse approximate segmentation masks are shown in Figure 4.2. They are fused into one mask before being fed into the deep neural network.

4.3.3 Two stage training

U-Net++[154] is an encoder-decoder network where the encoder and decoder are connected through a series of nested, dense skip connections. The nested skip connections can help bridge the semantic gap between the feature maps of the encoder and decoder, which is helpful in segmenting fine-grain details of target tissues like intermuscular fat in our case. Thus, we use U-Net++ as our backbone to infer segmentation results. Transfer learning refers to reusing a model developed for a task as the starting point for a model on a different or same task, which alleviates the challenge of limited training data. Thus, we design a two-stage transfer

learning strategy. In the first stage, we use approximate pseudo labels to train U-Net++ from scratch and choose the best model according to performance on the validation dataset. Then, the best model is loaded as initialization. Human expert-labeled data are used to fine-tune the model until converges. The whole pipeline is shown in Figure 4.2.

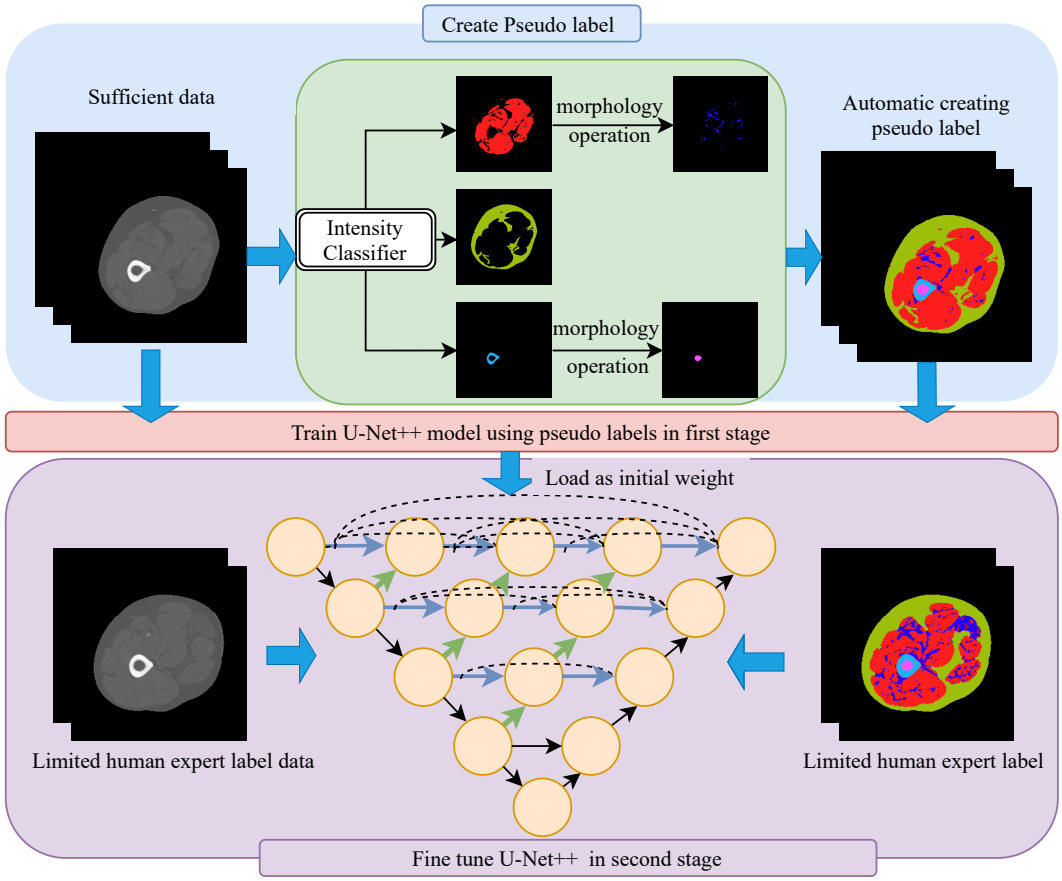


Figure 4.2: The proposed hierarchical coarse-to-fine thigh segmentation includes three parts: 1) The threshold and morphology are used to generate coarsely segmented pseudo labels. 2) Feeding pseudo labels into the deep learning model and training the model from scratch. 3) Using the optimized model from the previous stage as initialization, and fine-tuning the model with limited expert labels. The model from the first- and second-stage is optimized separately.

4.3.4 Data distribution

We use 3022 de-identified CT thigh slices from the Baltimore Longitudinal Study of Aging (BLSA)[45] and 121 de-identified thigh slices from the Genetic and Epigenetic Signatures of Translational Aging Laboratory Testing (GESTALT) study as well as 9141 de-identified lower leg slices from the BLSA. All data are under Internal Review Board approval. The image size is 512 $\times$ 512. In the preprocessing and quality assurance stages, 5 thigh images are discarded since they include other structures (e.g. the table). Note that for some thigh slices, only the left thigh is manually labeled instead of both thighs. As for labels of lower leg images,

Study Name	Cohort	Slices	All thighs or lower legs	Labelled thighs or lower legs
BLSA	First stage thigh training	944	942 L and 941 R	0
BLSA & GESTALT	Second stage thigh training	117	117 L and 8 R	125
BLSA & GESTALT	Second stage thigh validation	26	26 L and 5 R	31
BLSA & GESTALT	Second stage thigh testing	65	65 L and 8 R	73
BLSA	External lower leg testing	9141	9141 L and 9141 R	39
BLSA	External thigh testing	1991	1987 L and 1991 R	0

Table 4.1: The number of slices, thighs, lower legs, labeled thighs and lower legs for the cohort.

we manually refine the result from the proposed method as ground truth. We divide labelled thigh slices into training, validation, and testing cohorts for stage 2 in ratio of 60 percent,10 percent, and 30 percent, respectively. No subject had images in both the training and validation or testing cohorts.

4.3.5 Implementation details

Our experiments are implemented in Python 3.7 with PyTorch 1.7. We apply a window of [-150, 100] HU to normalize each input image. In the first stage, the initial learning rate for U-Net and U-Net++ is 0.002 and 0.0002, respectively. In the fine-tuning stage, the initial learning rate for both U-Net and U-Net ++ is 0.0001. We conducted the experiment to train only with human expert labels by using U-Net and U-Net++. The learning rate for U-Net is 0.01, and the learning rate for U-Net++ is 0.001. The learning rate decayed to 0 linearly until the end of the training epoch in both stages. Resize and crop are used as online data augmentation. The max-training-epoch is set to 200 with a batch size of 8. The optimizer used in training is stochastic gradient descent (SGD).

4.3.6 Baseline methods and metrics

The U-Net[123][29] is considered an alternate architecture because of its impressive performance on medical image segmentation. To validate the effectiveness of the transfer learning strategy, both U-Net and U-Net++ training with human labels only are also regarded as baseline methods. To evaluate the accuracy of our proposed method, we compare the segmentation results against the ground truth provided by expert labels. To quantify the agreement between segmentation and truth, we use the Dice Similarity Coefficient (DSC) as the main evaluation measurement for inference results by comparing each binary tissue against the ground truth voxel-by-voxel:

$$DSC = \frac{2|R \cap T|}{|R| + |T|} \tag{4.1}$$

where R represents the segmentation result generated by the deep learning model and T represents the corresponding ground truth.

4.4 Experimental Results

Figure 4.3 compares the DSC of the muscle, cortical bone, inner bone, subcutaneous fat, and intermuscular fat between U-Net++ and U-Net using only human labels, in stage 1 and stage 2. The boxplot presented is evaluated across 73 single thighs. Table 4.2 shows the mean DSC of each tissue of all six methods. Overall, the average DSC across all five tissues of U-Net++ in the second stage is significantly better than all the other five methods. Except for subcutaneous fat, the proposed method has the highest mean DSC of the rest tissues. The proposed method makes the largest improvement from 0.681 to 0.782 on mean DSC for sparse and small intermuscular fat compared with U-Net trained only with human labels. 4.4 compares the qualitative result produced by all six methods. Compared with U-Net in stage 2, the proposed method can yield superior performance and segments more details of intermuscular fat.

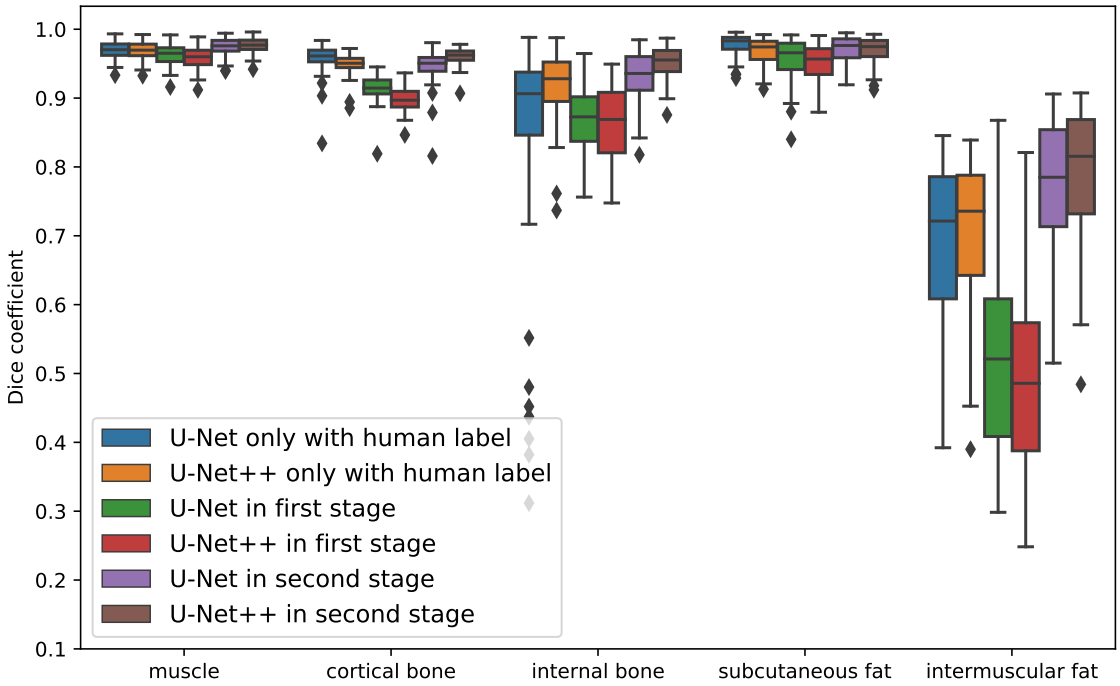


Figure 4.3: The fig shows the DSC comparison of thigh image using U-Net trained only with human labels, U-Net++ trained only with human labels, U-Net in stage 1, U-Net++ in stage 1, U-Net in stage 2 and U-Net++ in stage 2 in boxplots of five target tissues.

Method	Muscle	Cortical bone	Internal bone	Subcutaneous fat	Intermuscular fat	Average
U-Net only with human labels	0.967*	0.958	0.852*	**0.977***	0.681*	0.887*
U-Net++ only with human label	0.966*	0.947*	0.919*	0.967*	0.695*	0.899*
U-Net in the first stage	0.960*	0.915*	0.868*	0.955*	0.609*	0.841*
U-Net++ in the first stage	0.957*	0.898*	0.865*	0.949*	0.481*	0.830*
U-Net in the second stage	0.971	0.946*	0.932*	0.900	0.762*	0.916*
U-Net++ in the second stage	**0.973**	**0.960**	**0.951**	0.969	**0.782**	**0.927**

Table 4.2: The mean DSC for each tissue of each method for thigh CT image. The highest result is bolded. The * means the method is significantly different from U-Net++ in the second stage (p-value $<$ 0.05, Wilcoxon signed-rank test).

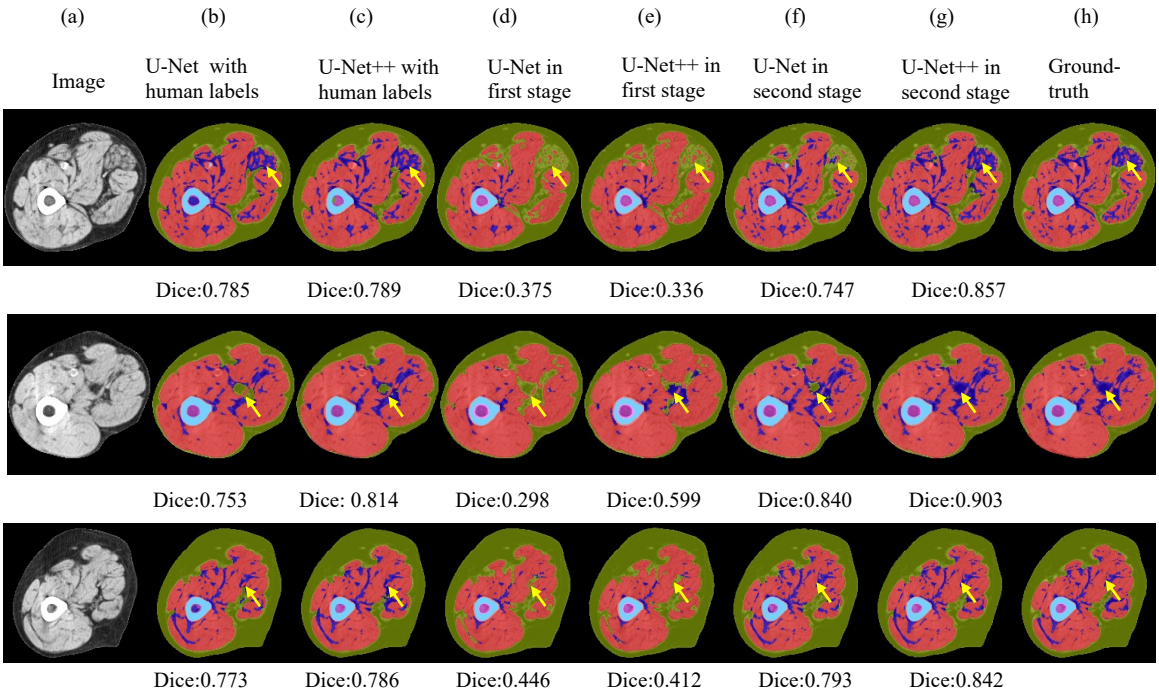


Figure 4.4: The plot shows the qualitative representation of the thigh slice segmentation. (a) represents three randomly selected source CT images after applying window [-150,100]. (b) represents the segmentation from U-Net only trained with human labels. (c) represents the segmentation from U-Net++ only trained with human labels. (d) and (e) represent the segmentation by using network U-Net and U-Net++ in stage 1 respectively. (f) and (g) is the segmentation by network U-Net and U-Net++ in stage 2, respectively. (h) is the ground truth. The yellow arrow points to the large difference between those methods and ground truth. The DSC values only show intermuscular fat segmentation performance for reference..

To test the generalizability of the proposed methods, we apply the model to preprocessed lower-leg images. The experimental setting is the same as in thigh's experiment. Figure 4.5 compares the performance of all six methods on lower leg images. Table 4.3 shows the mean DSC of each tissue of all six methods on lower leg images. The average DSC of all tissues of the proposed method decreased from 0.927 to 0.855 when compared with the thigh experiment, but still the significantly best among all the other five methods. However, the U-Net trained only with human labels has the highest mean DSC 0.923 in cortical bones and the U-Net++ trained only with human labels has the highest mean DSC 0.893 in internal bones. Figure **??**

demonstrated the confidence level of the bone results with qualitative representations.

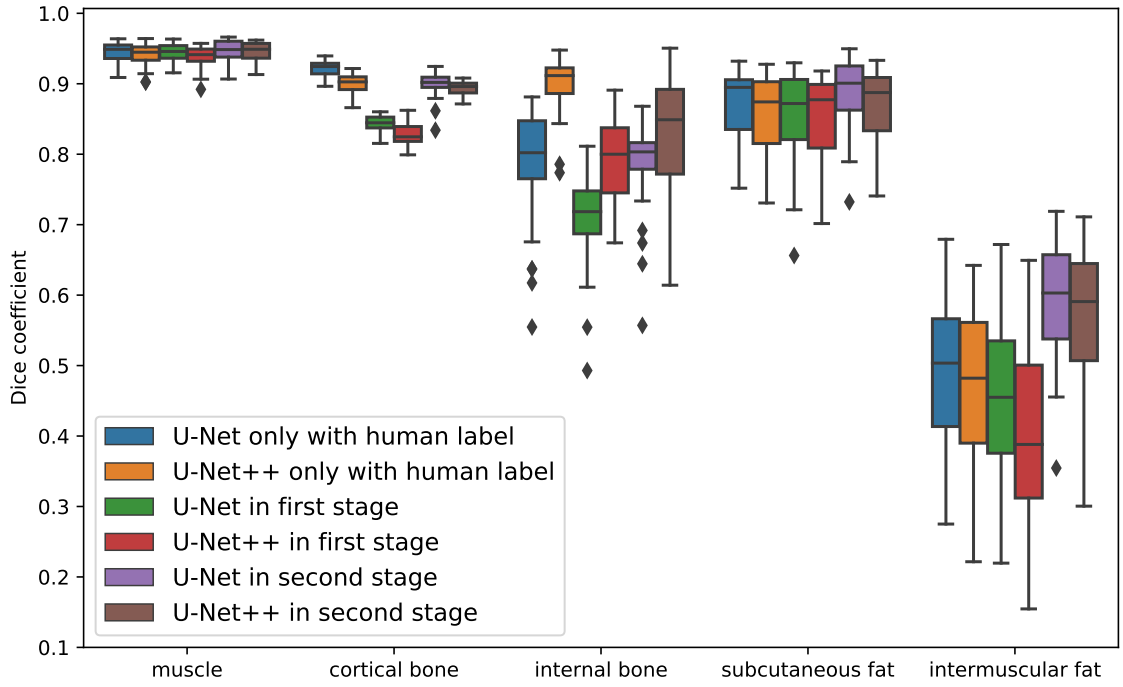


Figure 4.5: The fig shows the DSC comparison of lower leg image using U-Net trained only with human labels, U-Net++ trained only with human labels, U-Net in stage 1, U-Net++ in stage 1, U-Net in stage 2 and U-Net++ in stage 2 in boxplots of five target tissues.

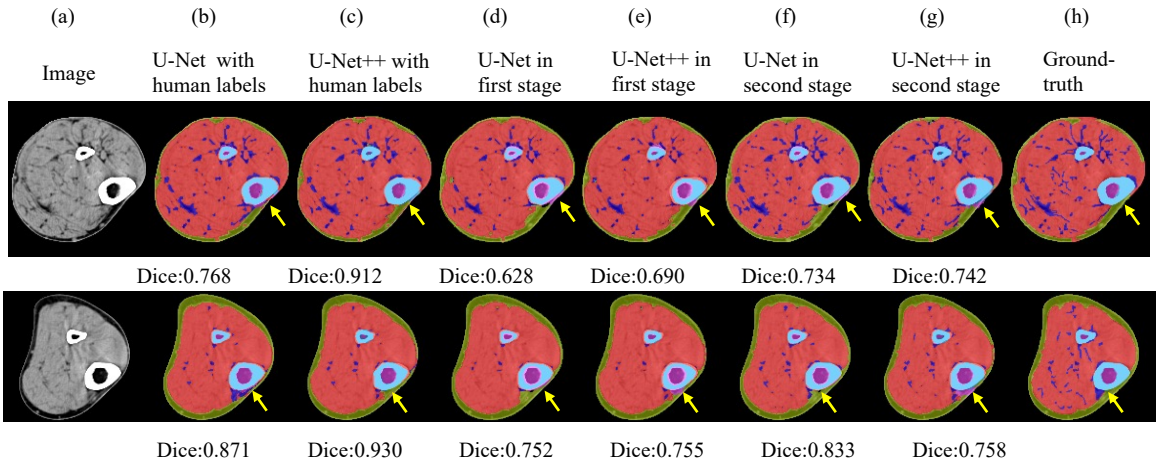


Figure 4.6: The plot shows the qualitative representation of the lower leg slice segmentation. (a) represent the source CT image after applying window [-150,100]. (b) represents the segmentation from U-Net only trained with human labels. (c) represents the segmentation from U-Net++ only trained with human labels. (d) and (e) represent the segmentation by using network U-Net and U-Net++ in stage 1 respectively. (f) and (g) is the segmentation by network U-Net and U-Net++ in stage 2 respectively. (h) is the ground truth. The text below each image is internal bone DSC.

Method	Muscle	Cortical bone	Internal bone	Subcutaneous fat	Intermuscular fat	Average
U-Net only with human labels	0.944*	**0.922***	0.786*	0.876*	0.494*	0.805*
U-Net++ only with human label	0.941*	0.900*	**0.901***	0.857*	0.469*	0.814*
U-Net in the first stage	0.945*	0.843*	0.708*	0.850*	0.443*	0.758*
U-Net++ in the first stage	0.939*	0.838*	0.791*	0.852*	0.393*	0.761*
U-Net in the second stage	**0.946***	0.901	0.787*	0.891*	**0.590**	0.823
U-Net++ in the second stage	**0.945**	0.893	0.836	0.870	0.573	**0.823**

Table 4.3: The mean DSC for each tissue of each method for lower leg CT image. The highest result is bolded. The * means the method is significantly different from U-Net++ in the second stage (p-value < 0.05, Wilcoxon signed-rank test).

After training on pseudo labels, we want to investigate the relationship between the number of human expert data and the performance of the model in the second (fine-tuned) stage. We fed 1, 5, 10, 20, 30, 60, 90, and all expert human label thighs to fine-tune the model from the first stage respectively. The fine-tuning process is repeated 10 times. Each time the data is randomly picked from the training cohort (125 thighs) and inference from the test cohort (73 thighs). The distribution of mean dice of each retrain is shown in Figure 4.7. When only feeding one thigh to model, the variance of mean DSC is largest compared with others. Also, the variance becomes smaller when increasing feeding data. When feeding 30 thighs, the mean DSC of 10 re-training is 0.924 almost equivalent to the mean DSC 0.931 of for feeding all training data.

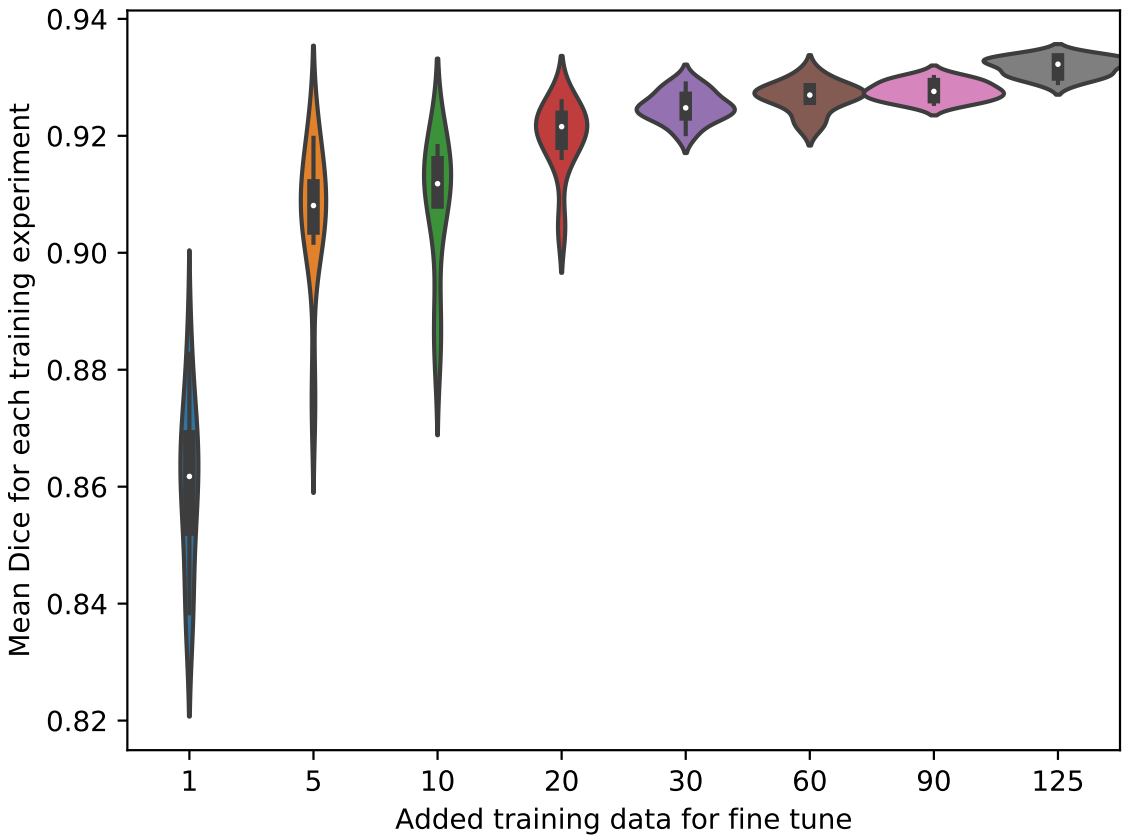


Figure 4.7: Shows the relationship between mean DSC and added data for each fine-tuning. The violin plot includes 10 data points. Each data represents mean DSC across all tissues of the test cohort in one fine-tuning process.

We apply the proposed models on additional CT scans of thighs and lower legs, which do not have human-generated label maps for comparison. We overlay the segmentation results on CT images with colormap and undergo human review. each segmentation result to find outliers. 10 out of 3982 thigh images and 136 out of 18282 lower leg images fail human review and are regarded as outliers. Figure 4.8 shows four outliers from thigh and lower leg images respectively.

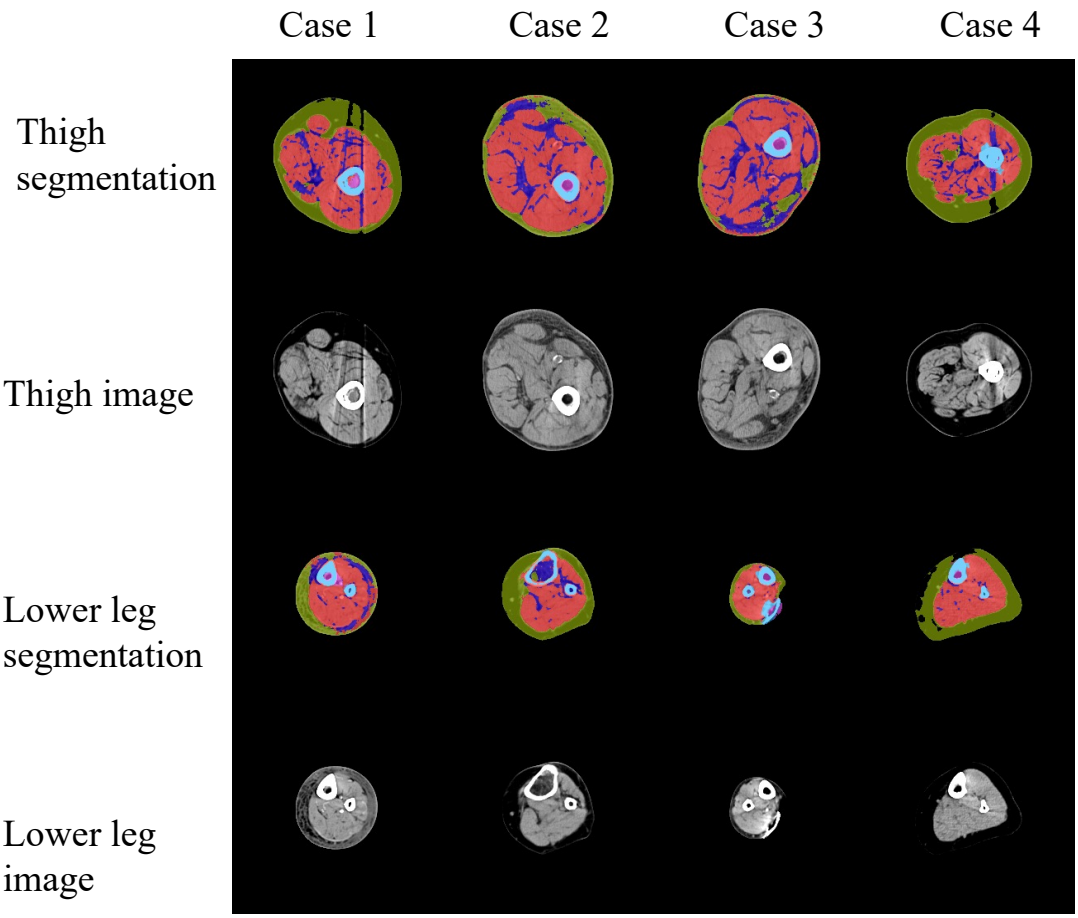


Figure 4.8: The outliers from thigh and lower leg. The first row and third row are segmentation results on the thigh and lower leg. The second and fourth rows are the CT images after applying the window. Each column represents an outlier from the thigh and lower leg respectively.

4.5 Conclusion and Discussion

Herein, we proposed a transfer learning-based method to achieve accurate and robust thigh tissue segmentation, focusing on muscle, cortical bone, internal bone, subcutaneous fat, and intermuscular fat. The proposed framework can achieve accurate segmentation on thigh CT slices with limited human labels. Compared with methods only trained with human expert labels, the superior performance of the proposed framework demonstrates the effectiveness of two-stage training. We applied the model only trained on thigh slices to the lower leg image. The results show that our model still recognizes muscle and fat with high agreement, which demonstrates the proposed framework has the generalizability to the similar anatomical structure of CT image. Then, we analyze the relationship between the size of the fine-tuning dataset and performance. The result shows that the proposed framework can use a smaller size of training cohort to keep performance as all training data, which indicates the proposed framework has the potential to fully exploit the data and

increase data efficiency.

One major limitation of the proposed method is that the performance of bone structure is inferior to models only trained with human labels. As shown in FigIV-7, the proposed method might regard subcutaneous fat around cortical bone as internal bone, which leads to inferior performance. The reason could be that the boundary of the pseudo label between cortical bone and internal bone is not clearly defined, which made the model misclassify subcutaneous fat around cortical bones. Thus, how to improve the quality of pseudo-labels is one of the important future directions. In summary, the proposed pipeline has the potential to be applied in other medical scenes with low human effort, which makes better use of human expert labels.

CHAPTER 5

Single Slice Thigh CT Muscle Group Segmentation with Domain Adaptation and Self-Training

This work was previously published [175]. Permission to include the work as part of the dissertation has been obtained, see Appendix A.

5.1 Introduction

Thigh muscle group segmentation is essential for assessing muscle anatomy, computing the muscle size/volume, and estimating muscle strength[70]. Quantitative thigh muscle assessment from segmentation can be a potential indicator of metabolic syndrome[99]. The loss of the thigh muscle and associated functional capabilities are closely related to aging[50]. Accurate measurement of thigh muscle cross-sectional area, volumes, and mass can help researchers understand and study the effect of aging on the composition of the human body. Thus, extracting subject-specific muscle groups is an essential step.

MR imaging is the most common imaging technique in previous muscle analyses given its high contrast for soft tissue[185]. Many human efforts have been put into MR imaging for muscle analysis. Barnouin et al. optimize reproducible manual muscle segmentation[8]. Schlaeger et al. construct a reference database (MyoSegmentTum) including the satorious, hamstring, quadriceps femoris, and gracilis muscle groups for 3D MR volume[139]. Compared with MR imaging, however, the short acquisition time of CT is better suited for routine clinical use[185]. In a longitudinal body composition study, single slice CT for each subject also reduces unnecessary radiation [190, 178, 189, 191, 188]. Accurate segmentation of muscle groups on a single slice can aid in understanding thigh components and the effects of aging on muscle[119].

Direct human manual annotation on single-slice CT is labor-intensive and challenging due to similar intensity among different muscle groups in CT. Leveraging publicly available annotation from existing MR resources (source domain) like MyoSegmentTum for CT (target domain) is a promising direction to overcome the problem of muscle group segmentation. Methods handling domain shift or heterogeneity among modalities are called domain adaptation (DA)[59]. DA aims to minimize differences among domains. DA has two challenging tasks that need to be addressed in our case: 1) homogeneous intensity of different muscle groups of CT images as mentioned before, and 2) inter-modality heterogeneity including contrast and anatomic appearance, The above two challenges can be found in Fig.1. With the above challenges for thigh muscle segmentation problems, we propose a new DA pipeline to achieve CT thigh muscle segmentation. We build a segmenter trained with synthetic CT images in CycleGAN[199]. We infer segmentation maps on real CT images by the segmenter and divide the segmentation maps into two cohorts based on entropy.

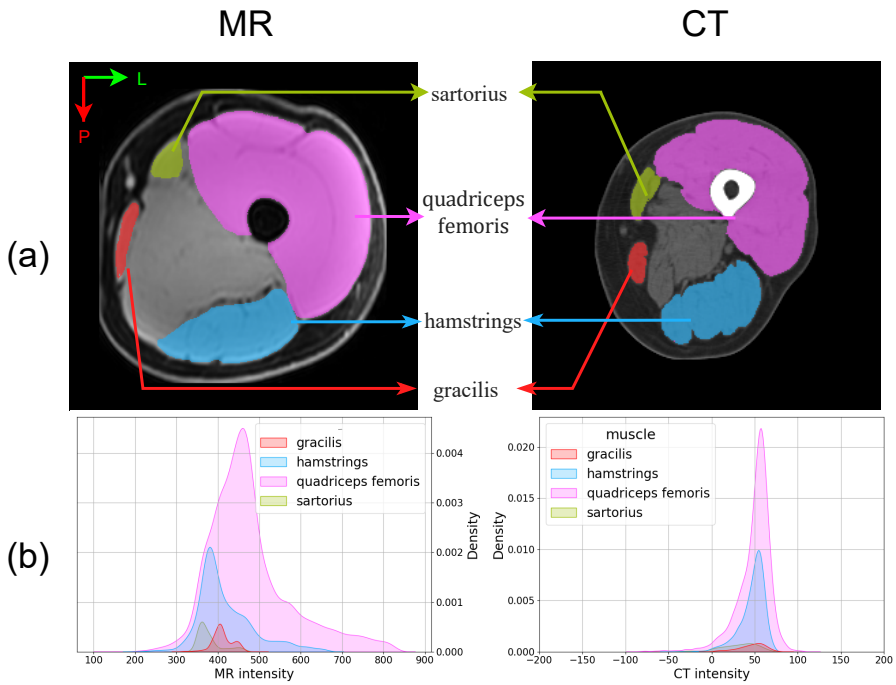


Figure 5.1: A selective sample that highlights the inter-modality heterogeneity between MRI and CT and low-intensity difference among different muscle groups in CT. (a) The MR image is normalized by min-max. The original CT scale is clipped to [-200,500] and then normalized to [0,1]. (b) is the intensity distribution for four muscle groups. The overlap intensity among four muscle groups is observed from the second row.

The entropy can work as an indicator for prediction map quality[163]. Based on the anatomic context, the whole muscle and bone masks of CT images are utilized to correct the wrong prediction brought by domain shift. Self-training is applied on two cohorts to make the segmenter adapt to high entropy cohorts to enhance robustness and preserve the segmentation performance on low entropy cohorts.

5.2 Material and method

To solve challenges 1) and 2), we proposed a pipeline that includes three key parts as described in (Fig 5.2): (1) preprocessing on 2D single thigh slice and 3D public MRI volume, and (2) training segmentation module by feeding synthesized CT images, and (3) fine-tuning segmentation module by applying self-training on the CT training datasets.

5.2.1 Data and preprocessing

We use two datasets in our study. One is the Baltimore Longitudinal Study of Aging (BLSA)[45], and the other one is MyoSegmenTUM[139]. The BLSA is a longitudinal dataset and collects 2D mid-thigh CT slices for each subject during the visit. BLSA study protocols are approved by the National Institutes of Health Intramural Institutional Review Board and all participants provided written informed consent. MyoSeg-

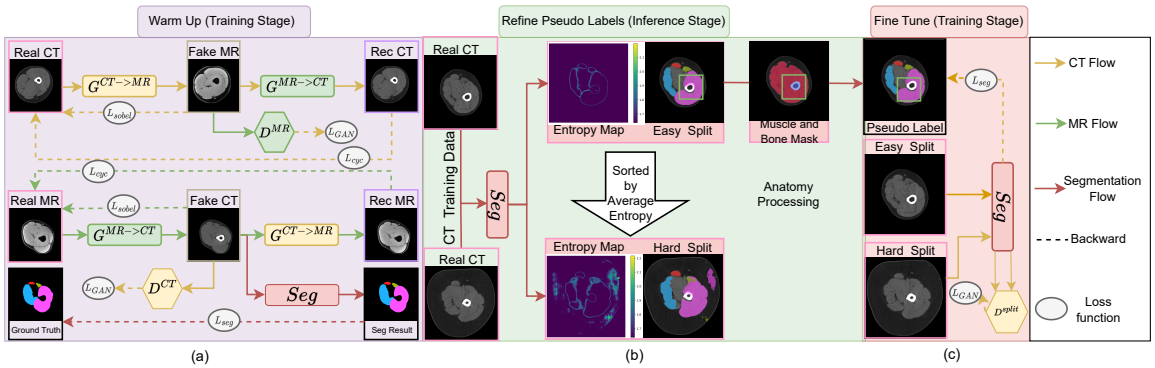


Figure 5.2: Overview of proposed pipeline. In part (a), we adopt a CycleGAN design including two generators and two discriminators for MR and CT respectively. The segmentation module is trained by feeding synthetic CT images and corresponding MR ground truth. In part (b), the segmentation module from (a) is used to infer pseudo labels divided into hard and easy cohorts based on entropy maps. Then, the easy cohort pseudo-labels are refined based on anatomy processing (muscle and bone bask). In part (c), easy cohort pseudo-labels of CT images are used to fine-tune the segmentation module, and adversarial learning between easy and hard cohorts forces the segmentation module to adapt to hard cohort simultaneously to increase segmentation module robustness.

menTUM is a 3D MRI thigh dataset providing annotations for four muscle groups including the sartorius, hamstring, quadriceps femoris, and gracilis muscle groups.

We used 1123 de-identified 2D low-dose single CT thigh slices of 763 participants from the BLSA. All data are de-identified under Institute Review Board approval. The slice has a size of 512×512 pixels. We split one single CT slice into left thigh and right thigh images with size 256 ×256 pixels by following the pipeline in [177]. During the preprocessing steps, 11 images were discarded due to low quality or abnormal anatomic appearance. The CT images are the target domain in our case.

MyoSegmentTUM consists of water-fat MR images of 20 sessions of 15 healthy volunteers. The water protocol MR is selected as the source image. We select 1980 mid-thigh slices from MR volumes to reduce the anatomical gap between MR and CT slices at the mid-thigh position. The MR slices are divided into left and right thigh images based on image morphology operation. Each image has 300 × 300 pixels.

The original label of the MR slices is placed at each group with a margin of 2 mm to the outer boundary, as shown in Fig 5.3(c). The incomplete ground truth makes the whole domain adaptation pipeline more challenging. To address this concern, we extract whole cross-sectional muscle and bone contour by using level set[96]. We use a binary 3×3 kernel to dilate the quadriceps femoris and hamstring muscle with six and two iterations, respectively. The complete muscle mask is obtained after performing the level set and dilation operation, as shown in Fig 5.3(d).

We feed random pairs of CT and MR images to the proposed method. All 1980 MR images are fed into the training cohort. For CT, we divide all CT images into training, validation, and test cohorts based on

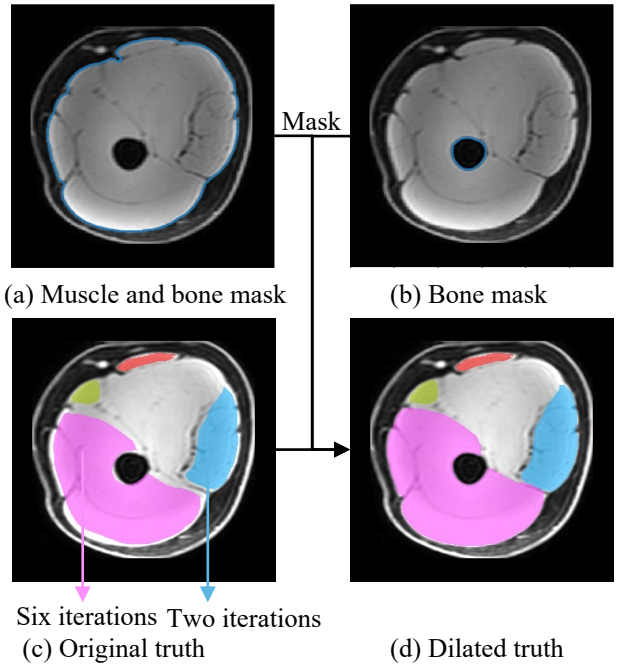


Figure 5.3: The preprocessing steps for dilating the ground truth of the MRI dataset. The blue contours in (a),(b) represent the muscle and bone boundaries extracted by level sets, and (c) represent the original ground truth. The quadriceps femoris muscle group is dilated in 6 iterations and the hamstring muscle group is dilated in 2 iterations. (d) represents the final truth after preprocessing.

	Participants	Images including left and right thigh	Image resolution	Pixel dimension (mm × mm)
CT training cohort	669	2044	256 × 256	0.97 × 0.97
CT validation cohort	19	38	256 × 256	0.97 × 0.97
CT test cohort	75	152	256 × 256	0.97 × 0.97
MR training cohort	15	1980	300 × 300	1 × 1

Table 5.1: Data distribution and image information for the whole pipeline.

participants. The training cohort includes 2044 CT thigh images from 669 participants. The validation cohort consists of 38 CT thigh images from 19 participants. The test cohort consists of 152 CT thigh images from 75 participants. Each CT image in the validation and test cohort has ground truth manually annotated from scratch to work for evaluation. The data distribution can be found in Table 5.1.

5.2.2 Train segmentation module from scratch

Inspired by SynSeg-net[71], we design a U-Net[133] segmentation module (*Seg*). We train the *Seg* with CycleGAN[199] in an end-to-end fashion as shown in Fig 2(a). CycleGAN aims to solve the image-to-image translation problem, in an unsupervised manner without requiring paired images. CycleGAN uses the idea of cycle consistency that we translate one image from one domain to the other and back again we should arrive at where we started [199]. Thus, we have two generators and discriminators in our framework.

Generator $G^{X\to Y}$ represents the mapping function X:→Y. Two generators $G^{MR\to CT}$ and $G^{CT\to MR}$ are utilized to synthesis fake CT ($G^{MR\to CT}(x_{MR})$)and fake MR ($G^{CT\to MR}(x_{CT})$) images respectively. The discriminator D^{CT} and D^{MR} determine whether the input image (CT or MR) is synthetic or real. The adversarial loss is applied to generators and discriminators and is defined as:

$$L_{GAN}^{CT}(G^{MR\to CT},D^{CT},X_{MR},Y_{CT}) = \mathbb{E}_{y\sim Y_{CT}}[logD^{CT}(y)] + \mathbb{E}_{x\sim X_{MR}}[1-logD^{CT}(G^{MR\to CT}(x))]$$

$$L_{GAN}^{MR}(G^{CT\to MR},D^{MR},X_{CT},Y_{MR}) = \mathbb{E}_{y\sim Y_{MR}}[logD^{MR}(y)] + \mathbb{E}_{x\sim X_{CT}}[1-logD^{MR}(G^{CT\to MR}(x))] \tag{5.1}$$

The above adversarial loss cannot guarantee that individual images are anatomically aligned to the desired output since there are no constraints for the mapping function. Cycle loss[199] is introduced to reduce possible space for the mapping function by minimizing the difference between images and cycle-reconstructed images. The loss function is:

$$L_{cyc}^{CT} = ||G^{MR\to CT}(G^{CT\to MR}(x_{CT})) - x_{CT}||_1 \tag{5.2}$$

$$L_{cyc}^{MR} = ||G^{CT\to MR}(G^{MR\to CT}(x_{MR})) - x_{MR}||_1 \tag{5.3}$$

To regularize the generator, we applied identity loss[199] to regularize generators. The identity loss is expressed as:

$$L_{Identity} = \mathbb{E}[||G^{MR\to CT}(x_{MR}) - x_{MR}||_1] + \mathbb{E}[||G^{CT\to MR}(x_{CT}) - x_{CT}||_1] \tag{5.4}$$

We further added an edge loss to preserve boundary information. Modified Sobel operator[81] is utilized to extract edge magnitude. The edge loss is calculated based on the difference in edge magnitude of two images. The edge loss is expressed as Eq.5.5

$$v = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \quad h = \begin{bmatrix} 0 & 0 & 0 \\ -1 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

$$sobel(x,y) = \| \sqrt{\|v*x\|_2 + \|h*x\|_2} - \sqrt{\|v*y\|_2 + \|h*y\|_2} \|_1$$

$$L_{edge} = sobel(G^{MR\to CT}(x_{MR}),x_{MR}) + sobel(G^{CT\to MR}(x_{CT}),x_{CT}) \tag{5.5}$$

where v and h are vertical and horizontal kernels, * represents the convolution between kernel and image. As for segmentation, weighted cross entropy loss is applied to supervise segmentation module L_{seg}

After defining all loss functions, we combine them together by assigning different weights $\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5$ for the loss function L. The L_{CT}^{GAN} is similar to L_{MR}^{GAN} and we set the same weight λ_1 for them. L_{cyc}^{CT} is symmetrical to L_{cyc}^{MR}, and the same weight λ_2 is assigned for those two losses. The final loss function is defined as

$$L = \lambda_1(L_{GAN}^{CT} + L_{GAN}^{MR}) + \lambda_2(L_{cyc}^{CT} + L_{cyc}^{MR}) + \lambda_3 L_{Identity} + \lambda_4 L_{edge} + \lambda_5 L_{seg} \tag{5.6}$$

5.2.3 Fine tune segmentation module in self training

Even though we train the segmenter from scratch by feeding synthesized CT images, the segmentation module is not robust to all CT cases as shown in Fig 5.2(b)(the segmentation map of hard split has incorrect prediction). The segmentation performance is still limited since synthetic data cannot transfer all information from real CT images. Inspired by[120], we adopted a self-training framework to handle this challenge. We infer all pseudo labels and probability maps for real CT images in the training cohort. The entropy calculated based on probability for each class works as a measurement to evaluate the confidence of the segmentation map in unsupervised domain adaptation[163]. All segmentation maps are ranked by average entropy map I^{CT} from low to high. The larger the entropy, the more potential error the segmentation map includes. All segmentation maps are divided into easy and hard splits based on ranking order. The first λ of training samples are easy split and the rest are hard split.

$$\begin{aligned} p^{CT} &= softmax(Seg(x^{CT})) \\ I^{CT} &= -\sum_{i=1}^{class} (p_i^{CT} log_2(p_i^{CT})) \end{aligned} \tag{5.7}$$

Where p^{CT} is the probability map for each muscle class and I^{CT} is the entropy map calculated based on p^{CT}.

Anatomical context such as spatial distribution is an important prior for medical image segmentation. To reduce incorrect prediction induced by noise and appearance shift in synthetic images, we leverage muscle and bone masks from[177] to mask out erroneous predictions for easy split as shown in Fig 5.2(b).

As shown in Figure 5.2(c), we construct the discriminator D^{split} from scratch. Different from [120], the segmentation module is further trained by aligning the entropy map of easy splits to ones of hard splits. At the same time, the segmentation module is fine-tuned by feeding rectified pseudo labels of the easy split after anatomical processing and supervised by weighted cross entropy lossL_{seg}^{easy}. The loss function $L^{finetune}$ can be expressed as:

$$\begin{aligned} L_{GAN}^{split} &= \mathbb{E}_{x \sim X_{CT}^{easy}}[logD^{split}(x)] + \mathbb{E}_{y \sim Y_{CT}^{hard}}[1 - logD^{split}(y)] \\ L^{finetune} &= \lambda_6 L_{GAN}^{split} + L_{seg}^{easy} \end{aligned} \tag{5.8}$$

Method	Gracilis muscle	Hamstring muscle	Quadriceps femoris	Sartorius muscle	Average of four muscles
AccSeg-Net	0.753(0.128)	0.882(0.075)	0.91(0.028)	0.708(0.176)	0.813(0.08)
DISE	0.786(0.159)	0.895(0.078)	0.928(**0.023**)	0.76(0.201)	0.843(0.09)
SynSeg-net	0.838(0.110)	0.869(0.072)	0.936(0.028)	0.802(0.164)	0.861(0.063)
Proposed	**0.876(0.085)**	**0.898(0.055)**	**0.941**(0.024)	**0.837(0.099)**	**0.888(0.041)**

Table 5.2: The mean DSC and standard deviation for each muscle group and average performance

where X_{CT}^{easy} is the easy split of the CT training cohort and X_{CT}^{hard} is the hard split.L_{seg}^{easy} is a weighted cross-entropy loss for the segmentation module only trained on easy cohort.

5.3 Experimental Results

We compare the proposed pipeline with three state-of-the-art domain adaptation methods including SynSeg-net[71], AccSeg-Net[197] and DISE[25]. Then we perform an ablation study to demonstrate the effectiveness of the fine-tuning stage and sensitivity analysis for the proposed method.

5.3.1 Implementation details and evaluation metrics

We used Python 3.7.8 and Pytorch 1.10 to implement the whole framework. The baseline and proposed methods are run on Nvidia RTX 5000 16GB GPU. For training from scratch, we set λ_1=1.0, λ_2=30.0, λ_3=0.5, λ_4=1.0, λ_5=1.0. In the segmentation module, the weights for background, gracilis muscle, hamstring muscle, quadriceps femoris, and sartorius muscle are set as [1,10,1,1,10] in the weighted cross-entropy loss, respectively. For the training data divided into easy and hard cohorts, we set the first $\lambda = \frac{2}{3}$ as the easy cohort and the rest as the hard cohort. For the fine-tuning stage, we set λ_6=0.001. The initial learning rate for the training model from scratch is 0.0002. We set the maximal training epochs as 100. Before the first 50 epochs, the learning rate is constant at 0.0002, and then it decreases to 0 linearly. We clip the original CT intensity to [-200,500]. For the MR images, we perform min-max normalization. All CT images and MR images are normalized to [-1,1].

Dice similarity coefficient (DSC)[38] is used to evaluate the overlap between segmentation and ground truth. Briefly, we consider S as the segmentation, G as the ground truth, and $||$ as the L^1 norm operation.

$$DSC(S,G) = \frac{2|S \cap G|}{|S| + |G|} \tag{5.9}$$

5.3.2 Qualitative and quantitative results

A detailed comparison of quantitative performance is shown in Table **??** and Fig 5.5. All methods are trained with the same training dataset and inference is performed on the same testing dataset. From Table **??**, the proposed method achieves the highest mean DSC of 0.888 with the lowest standard deviation of 0.041. The

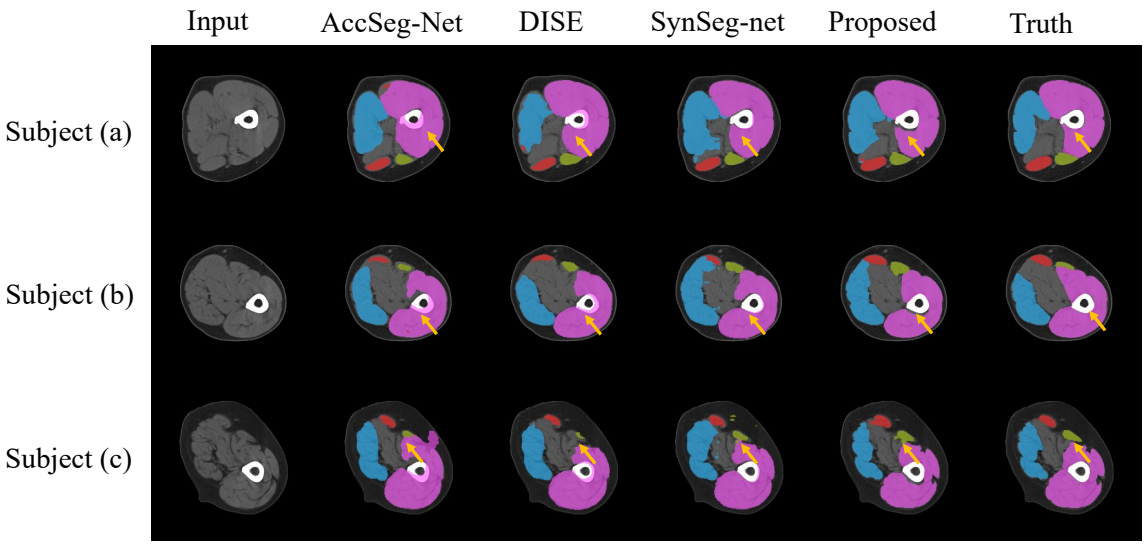


Figure 5.4: Representation results of the proposed methods and baseline methods. Each row represents one subject. The proposed method reduces prediction errors on bones and around muscle group boundaries. The yellow arrows point to differences between the proposed method and AccSeg-Net, DISE, and SynSeg-net. The Input column images are rescaled for visualization purposes.

proposed method significantly differed from all baseline methods with $p < 0.05$ under Wilcoxon signed-rank test. The proposed method achieves the best DSC for each muscle group and the lowest standard deviation except for the quadriceps femoris. Compared with AccSeg-Net, the proposed method makes the largest improvement in sartorius muscle increasing mean DSC from 0.708 to 0.837, and decreasing standard deviation from 0.176 to 0.099. In Fig.5.5, compared with the second best-performing method SynSeg-net, our method further reduces outliers and has a tighter and better DSC distribution. In Fig.5.4, while the baseline methods makes incorrect predictions on bone, our method is more robust and has fewer incorrect predictions as shown in Fig.5.4.

5.3.3 Ablation Study

To investigate the effectiveness of the anatomical processing step and adversarial learning in fine tune stage, we designed 1)“From scratch”, 2)“From scratch + Fine tune”, 3)“From scratch + muscle mask” and 4)“From scratch + muscle mask + Fine tune” pipelines by modifying the procedures of the proposed pipeline. “From scratch” represents the result directly from method section B. “From scratch + Fine tune” means splitting pseudo-labels from scratch into easy and hard cohorts and performing adversarial learning between two cohorts. “From scratch + muscle mask” represents that the muscle masked derived from [177] is used to mask out noise for the final prediction map. “From scratch + muscle mask + Fine tune” represents the proposed pipeline. The graphic description for each pipeline is shown in Fig 5.6.

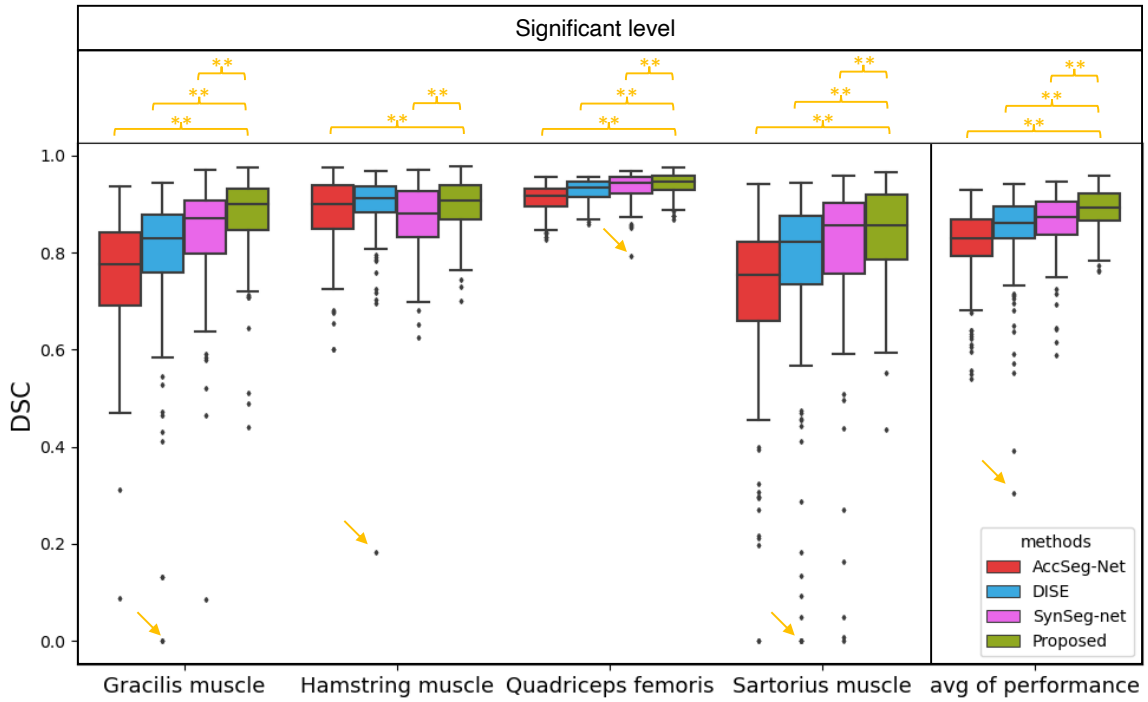


Figure 5.5: Quantitative results of DSC of baseline methods and the proposed method. * indicates ($p < 0.05$) significant difference between by Wilcoxon signed-rank test and ** indicates ($p < 0.02$ corrected by Bonferroni method[69]). The yellow arrows indicate outliers that are located at a far distance from the distribution, spanning from the 25th percentile to the 75th percentile, among the four methods. When calculating the standard deviation, these outliers are included in the calculation and can potentially influence the resulting standard deviation. Therefore, the box represents the data distribution from 25th percentile to 75th percentile rather than the standard deviation of the entire test dataset.

.

As shown in Fig 5.7, compared with "From scratch", the proposed pipeline significantly increases mean DSC from 0.870 to 0.888 and demonstrates that the anatomical processing step plus fine-tuning stage can improve segmentation performance. Compared with "From scratch + Fine tune", the proposed pipeline significantly increased mean DSC from 0.878 to 0.888, which shows that the muscle mask can help the segmentation module discriminate noise outside the muscle mask. Compared with "From scratch + muscle mask", the pipeline shows that adversarial learning can make the segmentation module adapt to the hard split improving DSC from 0.878 to 0.888 on the test dataset instead of only relying on the muscle mask.

5.3.4 Sensitivity analysis

As shown in Fig. 5.1, the thigh muscle is homogeneous, and hard to discriminate the muscle group based on intensity alone. Furthermore, it is difficult to delineate the boundary of muscle groups by visual assessment on CT images. To check whether prediction maps cover central areas of muscle groups, we perform a sensitivity analysis to the proposed method. For each muscle group, we apply a binary 3×3 kernel to erase every muscle

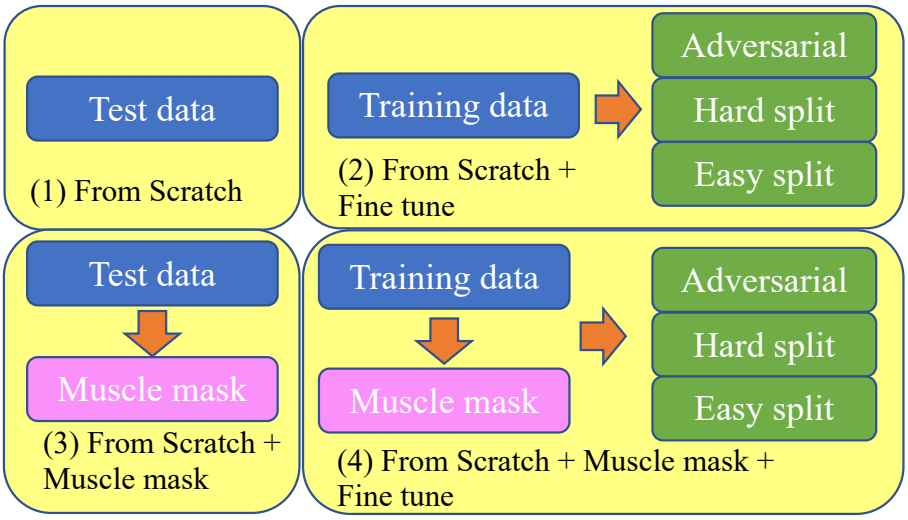


Figure 5.6: Graphic visualization for the four pipelines designed for the ablation study. (1) represents segmentation maps influenced by the segmentation module trained from scratch. (2) The pseudo-labels of the training data are inferenced by the segmentation module from scratch and then divided into two cohorts for fine-tuning. (3) The prediction map inferenced by the segmentation module from scratch is masked by a muscle mask. (4) Proposed pipeline. The pseudo-labels of the training data are inferenced by the segmentation module from scratch and then masked by a muscle mask for fine-tuning.

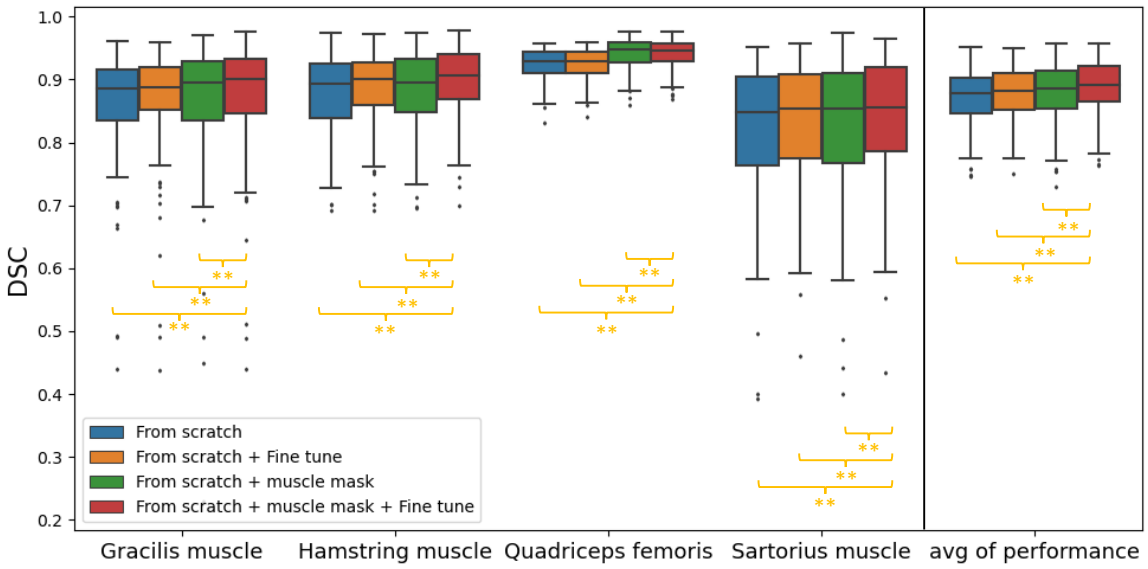


Figure 5.7: The quantitative results for four pipelines used in the ablation study. * indicates ($p < 0.05$) significant difference between by Wilcoxon signed-rank test and ** indicates ($p < 0.02$ corrected by Bonferroni method.

group iteratively until the predicted muscle group is empty. The area ratio is defined as the rate between

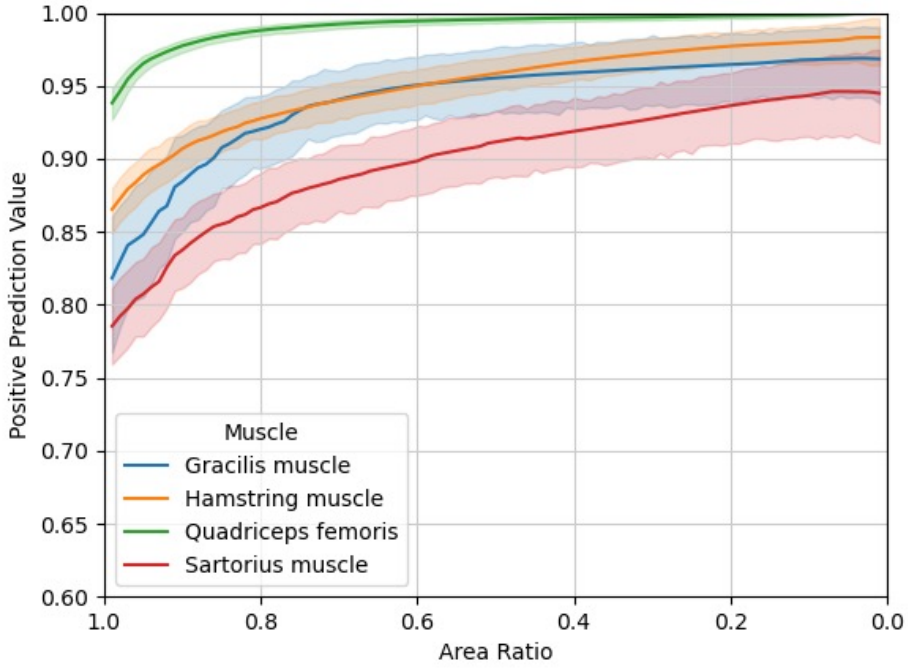


Figure 5.8: The sensitivity plot of proposed pipeline result. The x-axis represents the ratio between the eroded area and the muscle ground truth. The positive prediction value is calculated based on Eq. 5.10

eroded muscle mask and manual ground truth. The positive predictive value (PPV) is defined as

$$PPV = \frac{|S \cap G|}{|S|} \tag{5.10}$$

where S represents the segmentation and G represents the ground truth. $||$ represents the L^1 norm operator. From Fig 5.8, the quadriceps femoris has the highest initial PPV of 0.94 and the sartorius has lowest PPV of 0.78. The PPV of all four groups muscle are more than 0.85 when the area ratio is 0.8. The quadriceps femoris, hamstring, gracilis, and sartorius muscle have a final PPV of 1.0, 0.97, 0.96, and 0.95 respectively.

5.4 Discussion

In this work, we study thigh CT and achieve single-slice muscle group segmentation by proposing a two stages pipeline to leverage manual label from MR 3D volume. In the first stage, we selected single thigh CT slices from 3D volumes and split slices into left and right thigh images. The real MR and CT images were fed into a CycleGAN framework to generate synthetic CT images. The generated synthesized images were input into the segmentation module. We use the original annotation from the MR volumes to supervise the segmentation module. In the second stage, the pseudo-labels of CT images in the training cohort are inferenced by the segmentation module (*Seg*). Based on the assumption that uncertainty is related to wrong

predictions, we divided the training cohort into easy and hard splits based on inference entropy. We observe that the bone in MR is dark. However, bone in CT is bright. The significant contrast between MR and CT causes domain shift during CycleGAN incurring wrong predictions on the bone area. To address the domain shift problem, the muscle mask from [177] is used to correct the noise map. Finally, adversarial learning is utilized to align the prediction map between easy and hard split to make the segmentation module robust to real CT images. To our best knowledge, this is the first pipeline to perform domain adaptation on thigh CT images. We collect all modules into one container to let the public and more researchers take advantage of our contribution. The segmentation module can be directly used for single-slice CT muscle group segmentation.

Although the proposed pipeline can handle current challenges in domain adaption, limitations still exist in the process of the proposed pipeline. One limitation is the dependence on pseudo labels when training from scratch. need researchers to empirically tune the hyperparameters need to tuned empirically to make the generative model synthesis anatomy consistent images since we need to balance the generator and discriminator simultaneously. Another limitation is that even though the entropy map is closely related to noise, prediction errors cannot be found only based on entropy maps. It means that the segmentation module might learn incorrect patterns in the fine-tuning stage and needs further study, which is beyond the scope of this manuscript.

5.5 Conclusion

In summary, we present a novel pipeline to leverage muscle group annotations from MR 3D volumes in segmenting single thigh CT slices. In this study, we (1) proposed a pipeline to solve the domain adaptation problem for CT thigh images, (2) applied the proposed pipeline to CT thigh images and demonstrated the effectiveness and robustness of the pipeline, and (3) packed all modules into a container for researchers to extract muscle groups conveniently and directly without manual annotation. As our current pipeline includes multi-stages, the way to improve the whole pipeline is to bundle it into one end-to-end framework.

CHAPTER 6

Characterize brain and body connection through linear and nonlinear model

6.1 Introduction

An increasing body of evidence supports an intimate brain-body connection in aging[12], with cardiovascular disease (CVD)[95], cognitive decline[88], and dementia[88]. Recently, Bobb et.al[15] performed a longitudinal and cross-sectional study and found that higher BMI was associated with lower gray matter volume in several ROIs and with declines in volume in temporal and occipital gray matter over time. Beck et.al [12] conducted a study using a mixed cross-sectional and longitudinal design to examine the relationship between cross-sectional body magnetic resonance imaging (MRI) measurements of adipose tissue distribution and longitudinal changes in brain structure by estimating the brain age gap. Deng et al.[35] included traditional body metrics such as BMI of 322,336 participants to investigate the longitudinal association between life course adiposity and risk of all-cause incident dementia and to explore the underlying mechanisms driven by metabolites.

Nevertheless, the majority of these studies have primarily focused on establishing a connection between baseline brain measurements and subsequent outcomes. In accordance with [51], these studies can be categorized as within-sample correlations since they establish a relationship between two variables measured concurrently or analyze the correlation between a variable in a group at an initial time-point and another variable in the same group at a future time-point. However, the ultimate goal is to perform individualized predictions of output measurements/features given input features. Consequently, it is essential to develop a reliable relationship that can effectively predict outcomes for new samples, subjects, or cases based on relationships developed by previous individuals. Vakli and colleagues [159] used a computer model, CNN, to estimate BMI by analyzing brain MRI scans along with age and sex data. But, BMI doesn't give a complete picture of one's body composition, like fat and muscle percentages. Thus, we take the form of a tabular to record brain and body fine-grained features as measurements.

Tabular data refers to structured data organized in a table format consisting of rows and columns. In our case, each row corresponds to a particular subject's visit and represents the features associated with that subject. Conversely, each column represents a specific feature across all the datasets. According to [16], tabular data is inherently distinct from homogeneous data types, such as images, speech, and audio. It exhibits unique characteristics, including the presence of dense numerical features and sparse categorical features. Dense numerical features encompass continuous or discrete numerical values, whereas sparse categorical

features represent discrete values from a limited set of categories. This combination of diverse data types contributes to the heterogeneous nature of tabular data. Furthermore, the correlation among the features in tabular data is generally weaker compared to the correlations introduced by spatial or semantic relationships in image or speech data, which poses a challenge to deep neural networks. One helpful way to mitigate this challenge is to make full use of regularization. Different from homogenous data forms, feature importances of tabular data are different. Based on this observation, [80, 142] proposed the regularization to select sensitive intermediate features and remain inactive to insensitive intermediate features to explore the potential of a deep neural network. Abid et al.[1] proposed to apply Gumbel-softmax in concrete autoencoder, an end-to-end differentiable method for global feature selection, which efficiently identifies a subset of the most informative features and simultaneously learns a neural network to reconstruct the input data from the selected features. Inspired by [1], we extend Gumbel-softmax to characterize brain and body connection instead of self-supervised learning. To be specific, we use muscle, fat, and bone areas derived from single thigh slices as body metrics/features instead of BMI metrics. We use volumes of 133 brain regions as brain metrics/features. To characterize the brain and body connection, we develop a generalizable model that can explain the relationship between brain and body to predict brain measurements given by the body and vice versa.

6.2 Materials and Method

6.2.1 Brain and body feature extraction

We use the BLSA dataset[45] to characterize connections between the brain and body. BLSA collects a structure MRI of the brain and a mid-thigh single CT slice per each subject during their visit. We apply SLANT[72] on brain images to achieve whole-brain segmentation with 133 labels. SLANT divides the standard brain structure MRI image into 27 patches and trains corresponding 27 UNet models to perform segmentation on each patch. Finally, the prediction result is achieved through a majority vote of 27 models. As for the mid-thigh slice, we use the proposed segmentation method[178] to perform body composition by segmenting muscle, cortical bone, internal bone, subcutaneous fat, and intermuscular fat on a single slice. The volume of the brain region and the average area of the left and right thigh region are quantified from a segmentation map (shown in Figure 6.1) as brain and body features respectively.

6.2.2 Selection layer

Feature selection is an important concept in machine learning or deep learning. Briefly, we select a subset of useful features to build a good predictor for a specified response variable[63]. It can help to avoid model overfitting on noise and increase the interpretation of the model. To achieve feature selection, inspired by

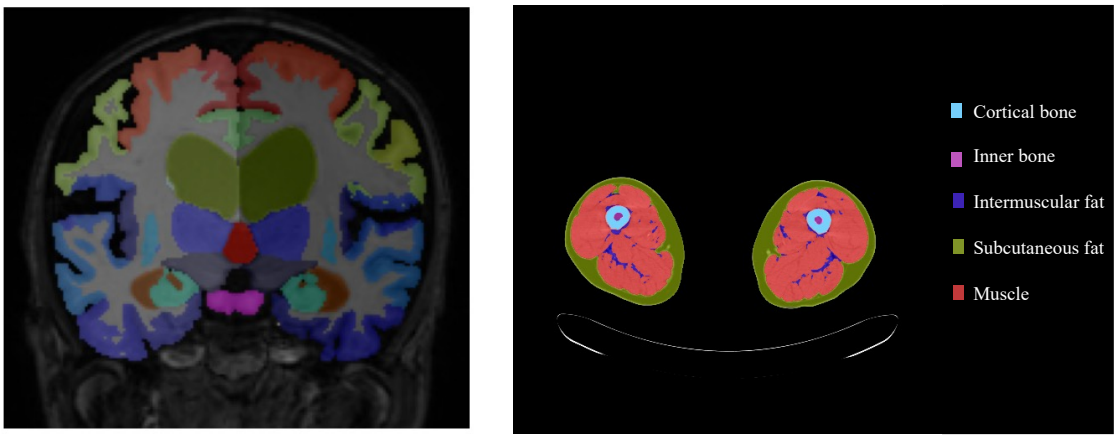


Figure 6.1: For each visit, we compute a segmentation map of brain image and mid-thigh image of the same subject in BLSA during one visit. In the brain segmentation map (a), we use the BrainColor protocol to visualize each label. For the thigh segmentation map (b), we follow the protocol from Chapter 4.

[26], Gumbel-softmax is adopted to construct a select layer by choosing features associated with the response variable. The Gumbel-softmax uses a continuous relaxation distribution to approximate a categorical random variable represented as a one-hot vector in $\mathbb{R}^d$ with category probability $p_1, p_2, ...p_d$. The Gumbel-softmax distribution is in the form of Eq 6.2.

$$g_j = -log(-log(u_j)), u_j \in Uniform(0,1) \tag{6.1}$$

$$m_j = \frac{exp((log(p_j)+g_j)/\tau)}{\sum_{k=1}^{d} exp((log(p_k)+g_k)/\tau)} \tag{6.2}$$

where m_j refers to the j^{th} element in a concrete random variable τ refers to temperature, which controls the sharpness of the distribution. In the limit $\tau \to 0$, the concrete random variable smoothly approaches the discrete distribution, outputting a one-hot vector with probability $\frac{p_j}{\sum_p p_j}$. When $\tau \to \infty$, the distribution becomes more uniform. To allow the Gumbel-softmax to explore different possibilities and avoid getting stuck in local optima, g_j as random noise sampled from Gumbel distribution is introduced to Gumbel Softmax distribution to explore different possibilities. Through re-parameterization, [85], Gumbel softmax distribution becomes more similar to the Gumbel distribution.

The concrete random variable is deployed from Eq.6.2 and used to choose features. To select k features from original d input features, k dimensional concrete random variables $\mathbf{m}^i, i \in 1...k$ are generated and $\mathbf{m}^i \in \mathbb{R}^d$. The selector layer outputs $\mathbf{x} \cdot \mathbf{m}^i$ for i^{th} selected feature, which is a linear combination of the input features weighted by $\mathbf{m}^i$ during training. However, as τ tends to 0, each random variable in the selector layer outputs

only one of the input features. During the validation and inference stage, "argmax" operator is used to choose a subset of features.

We follow Abid et al. [1] to set up an annealing schedule. The temperature is set for all of the concrete variables, initially beginning with a high-temperature T_0 and gradually decaying the temperature until a final temperature T_B according to a first-order exponential decay: $T(b) = T_0(T_B/T_0)^{b/B}$ where $T(b)$ is the temperature at bth epoch and B is the total number of epochs. Decreasing temperature smoothly is helpful for the Gumbel softmax distribution to better describe the real categorical distribution and avoid local minima during training.

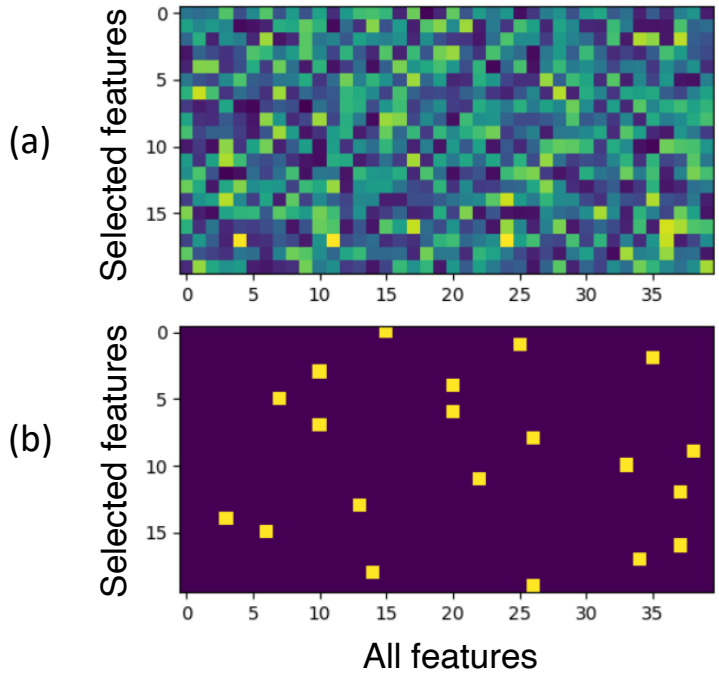


Figure 6.2: This figure shows the concrete random variables at the beginning and end of the training. Each row represents one random variable. To better visualize random variables, k is set as 20, and d is set as 40. (a) represents the concrete variable at the beginning of training. (b) represents the concrete variable at the end of training. From a to b, the Concrete random variable becomes sparse and similar to a one-hot vector

6.2.3 Model Architecture

Kadra et al.[80] demonstrated that well-regularized simple deep multilayer perceptron (MLP) outperforms specialized neural architectures and traditional machine learning methods. The plain deep neural network is designed to characterize the relationship between body features and brain features in a non-linear way. The designed model architecture includes 6 fully connected layer blocks. Each block contains one fully connected layer, followed by LeakyRelu to add the non-linear transformation to features. To subset associated original input features with predicted variables, the select layer is inserted after input features as shown in Figure 6.2.

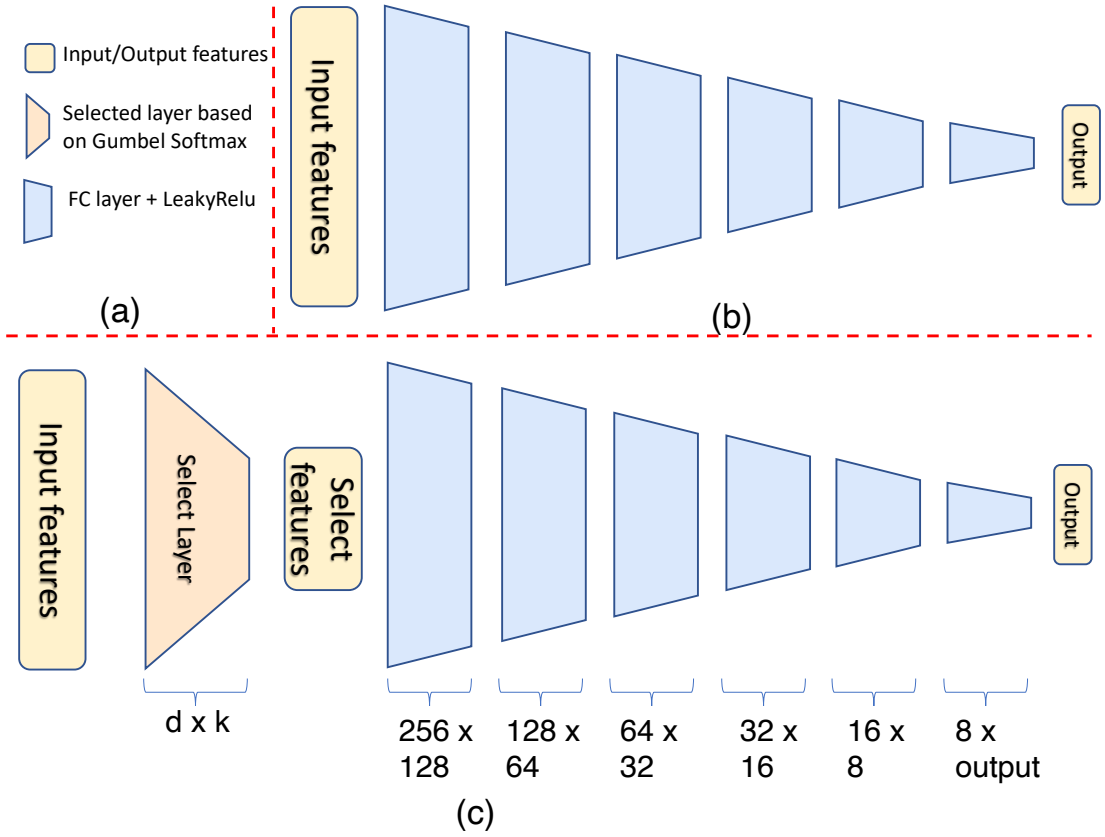


Figure 6.3: The model used to characterize brain and body relationship (a) shows the legend of each block used in this figure. (b) shows the plain deep neural network including 6 blocks. Each block includes a fully connected layer with 256, 128, 64, 32, 16, 8 input features respectively. (c) shows the proposed model architecture. Except for the same blocks, the select layer is inserted between input features and the first block to select associated features. d represents the number of input features per each observation and k represents the number of selected features decided by the user.

6.3 Experiment and result

The experiment is performed on 15,000 samples within synthetic and 2167 pairs within real BLSA datasets including brain and thigh features. The experiment of the synthetic dataset is to validate the efficacy of Gumbel softmax to check whether Gumbel softmax can select the associated features. The experiment of the real dataset is to predict body areas such as muscle, internal bone, cortical bone, subcutaneous fat, and intermuscular fat by using brain region volumes and predicting hippocampus volume with body region areas.

6.3.1 Implementation details

The model architecture is implemented in Keras [62]. We use mean square error as the loss function for all experiments. The deep neural network is optimized by the Adam[84] optimizer with a learning rate of 0.001. 5000 is set for whole training epochs and each batch has 16 samples. The synthetic dataset and real dataset

are divided into training, validation, and test cohorts. The tested epoch is selected based on the performance of the validation cohort. To reduce the effect of feature scaling, each feature and predicted response within the training cohort is normalized by using Eq. 6.3.

$$\boldsymbol{F}_{norm} = \frac{\boldsymbol{F} - \bar{f}}{\sigma} \tag{6.3}$$

where $\bar{f}$ is the average value of $\boldsymbol{F}$ and σ is the standard deviation of $\boldsymbol{F}$. By performing normalization, the feature is normalized to a distribution with a mean of 0 and a standard deviation of 1. The mean $\bar{f}$ and standard deviation σ are applied to validation and test cohort to normalize the corresponding feature to alleviate data distribution shift between training and other cohorts.

6.3.2 Characterize linear relationship

Regarding linear characterization, the linear regression model from the statsmodel package [140] is adopted to characterize linear relationships. The control variables are demographic information, such as sex, age, and mild cognitive impairment (MCI) status. Additionally, when estimating the volume of the hippocampus, whole brain volume is included as a control variable. The input features are potentially high-dimensional, with up to 133 dimensions. To address this, Principal Component Analysis (PCA) is applied to perform feature reduction by extracting the component with the largest variance. PCA is performed on the training dataset, and the resulting transformation matrix is applied to the validation and test cohorts to ensure consistent features are obtained. The linear regression formula for predicting body features from brain features takes the form:

$$output feature = Intercept + \beta_0 age + \beta_1 sex + \beta_2 MCI + \beta_3 input feature + \varepsilon \tag{6.4}$$

6.3.3 Metrics

Explained variance score (EVS) is used to explain the dispersion of errors of prediction for a given dataset prediction, which is suitable for regression tasks. EVS has the upper bound of 1. When EVS equals 1, it means perfect prediction. If EVS is smaller than 0, it means the prediction is worse than using the average of ground truth as the prediction. $\hat{\boldsymbol{y}}$ represents the prediction, and $\boldsymbol{y}$ represents the ground-truth. The formula for explaining variance is written as:

$$\begin{aligned} EVS(\boldsymbol{y}, \hat{\boldsymbol{y}}) &= 1 - \frac{Var(\boldsymbol{y} - \hat{\boldsymbol{y}})}{Var(\boldsymbol{y})} \\ Var(\boldsymbol{y}) &= \sum_i (y_i - \bar{y})^2 \end{aligned} \tag{6.5}$$

where $\bar{y}$ is the average of $\boldsymbol{y}$ and y_i is the ith element of $\boldsymbol{y}$.

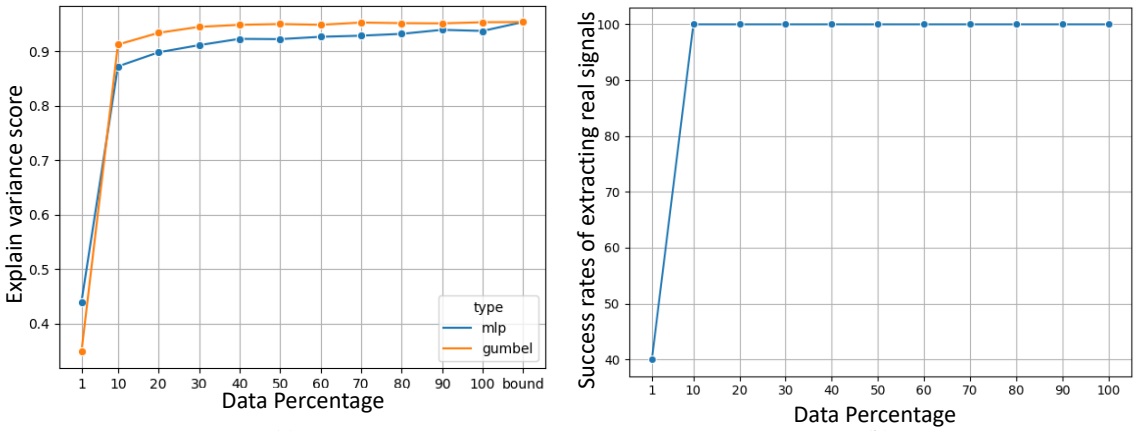
6.3.4 Validation of the Gumbel softmax

The purpose of this section is to investigate the ability of Gumbel softmax to separate meaningful signals from noise, as well as the potential to enhance the regularization of neural networks and improve the understanding of the input-output relationship. To test the hypothesis, we followed the approach proposed in [26] and created synthetic data with an explicit non-linear mapping from input to output. Specifically, we randomly generated 133 input features from a standard Gaussian distribution and selected 5 out of the 133 features that are related to 5 output responses. We then used the following formula to express the non-linear relationship between the input and output.

$$
\begin{aligned}
y_1 &= \sin(x_1) + 3\exp(x_4) + \eta \\
y_2 &= 4\cos(x_1) + 3|x_2| + \eta \\
y_3 &= \exp(x_1) + 5\sin(x_2) + 6|x_3| + \eta \\
y_4 &= \exp(x_1) + 10\sin(x_4) + 6\exp(x_5) + \eta \\
y_5 &= 4|x_5| + 4\exp(x_3) + \eta
\end{aligned} \tag{6.6}
$$

where $\eta \sim \mathcal{N}(0,1)$ is regarded as noise added into the non-linear relationship.

To train our model, we generated 15,000 samples for the training cohort, 1,000 samples for the validation cohort, and 9,000 samples for the test cohort. The Eq 6.3 is used to normalize features and responses. We varied the percentage of training data used to train the model and analyzed its performance. Figure 6.4 displays the results. As depicted in Figure 6.4 (a), our proposed model consistently outperformed the plain MLP model in terms of EVS, except when using only 1 percent of the training data. Gumbel Softmax is employed to select 10 relevant features and selected features are always the real signals after removing duplicates, except when using only 1 percent of the training data. In conclusion, the Gumbel Softmax method improved our model's prediction ability, by accurately selecting real signals, as evidenced by the results in Figure 6.4(b).



(a) (b)

Figure 6.4: The resulting plot for the toy example and success rate of selecting features in each run with different percentages of toy data. (a) shows the mean explain variance score across 5 output variables. Compared with the MLP model, the only difference is that Gumbel has a selected layer to subset real signals. (b) shows the success rate of real signals chosen by Gumbel Softmax.

6.3.5 Choice of number of selecting features

The number of subset features is an important hyperparameter that might influence the performance of neural networks. To choose suitable k, we performed the experiment with k values starting from 10 to 130 with step 10 to predict body region features(muscle area, cortical bone area, internal bone area, subcutaneous fat area, and intermuscular fat area) from brain volumes. The result can be found in Fig. 6.5. It can be observed that even though we plan to subset 130 brain features, only 12, 8, 15, 20, and 30 unique features are selected for predicting muscle, cortical bone, internal bone, subcutaneous fat, and intermuscular fat respectively. Most of the selected features are repeated. In such a scenario, the input features are over-trained and cannot reflect the effect of real signals. Based on this observation, we choose the upper bound of k as 10 when we use brain features to predict body features.

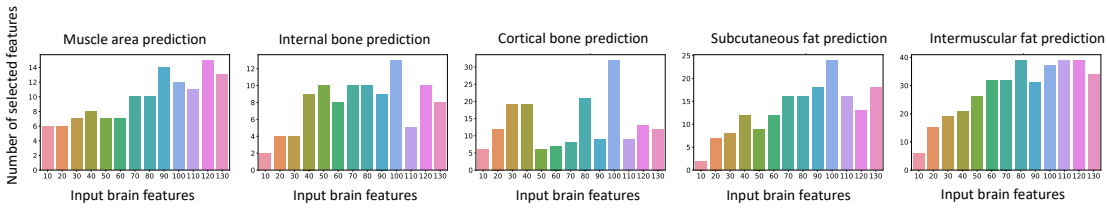


Figure 6.5: The bar plot of expected selecting features and exact features Gumbel-softmax select. The five bar plots represent using brain features to predict body features including muscle, cortical bone, internal bone, subcutaneous fat, and intermuscular respectively. When we increase the number of selected features, the exact number of selected features is less than expected since there are duplicate selected features

6.3.6 Real dataset

The BLSA dataset (Philips 3T Achieva) includes T1-weighted images acquired using an MPRAGE sequence (TE = 3.1 ms, TR = 6.8 ms, slice thickness = 1.2 mm, number of Slices = 170, flip angle = 8 deg, FOV = 256×240mm, acquisition matrix = 256×240, reconstruction matrix = 256×256, reconstructed voxel size = 1×1mm). The thigh single slice has resolution 512 $\times$ 512 and pixel size = 1x1mm. We follow the pipeline in Chapter 4 to divide the thigh slice into left and right images. Then the segmentation model is deployed to perform body composition. The demographic information for the BLSA dataset can be found in Figure 6.6.

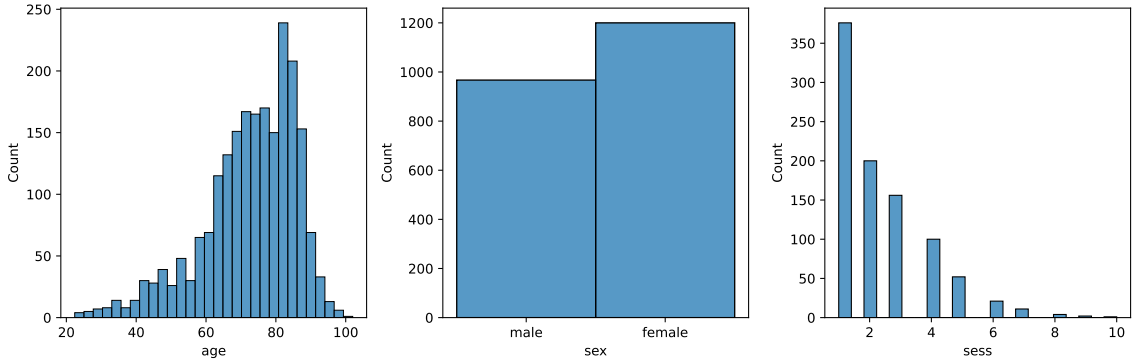


Figure 6.6: The histogram for demographic information of BLSA dataset. (a) shows the age distribution of BLSA subjects. The BLSA dataset is used to investigate the aging effect. Most people are elder. (b) shows the sex distribution of BLSA subjects. (c) shows the distribution of visit numbers per subject. The maximal number of visits of one subject is 10.

6.3.7 Predict body composition area using brain region volumes

The body composition refers to muscle, cortical bone, internal bone, subcutaneous fat, and intermuscular fat. In this section, brain region volumes are used to predict those areas by using the BLSA dataset. The whole BLSA dataset is divided into training, validation, and test cohorts randomly based on subjects. The training cohort has 1295 sessions from 553 subjects. The validation cohort includes 412 sessions from 184 subjects and the test cohort includes 460 sessions from 186 subjects. The input features and predicted variables are normalized using Eq 6.3. To have the same experimental setting as linear regression, we concatenate age, sex, and MCI to the feature map after the select layer. We feed 132 brain region features to the select layer and the following architecture. We iterate the number of select features from 1 to 10 with step 1 and choose the best performance of the validation dataset as the optimal model. The prediction results are evaluated by EVS. The Atlman-Blank plot is shown in Fig 6.7 to visualize the relationship between prediction and target ground truth in linear and non-linear. The explained variance score is shown in table 6.1

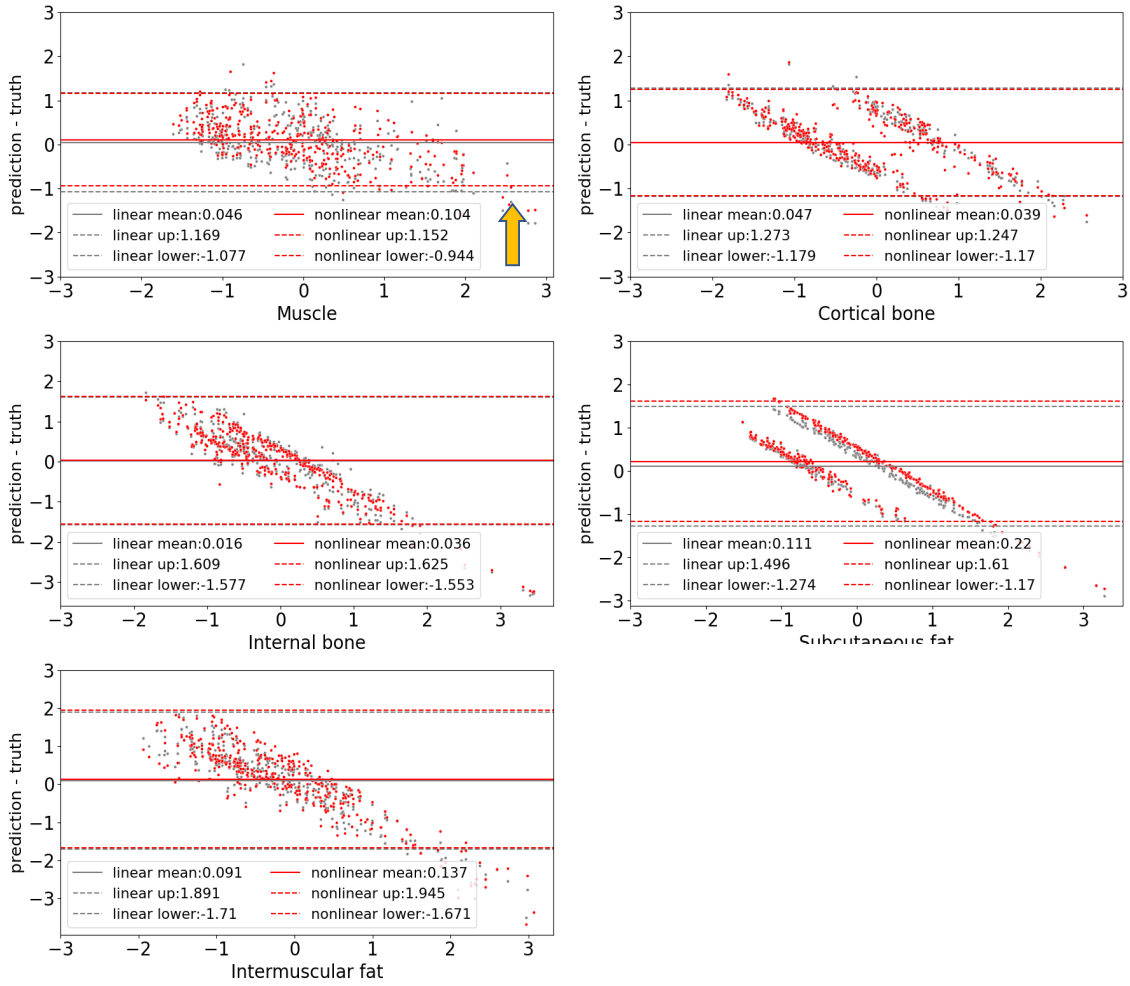


Figure 6.7: The bland-Altman plot of linear regression and proposed nonlinear methods. Each plot has the difference between prediction and truth as the y-axis and ground truth as the x-axis. The gray color represents the linear regression and the red color represents the proposed methods. We can observe the proposed methods have a smaller limit of agreement in muscle area prediction compared with linear regression as pointed out by the yellow arrow. We also find that there are two obvious clusters controlled by sex in cortical bone and subcutaneous fat prediction.

Region	Linear regression (Linear relationship)	Proposed model (Non-linear relationship)
Muscle area	0.598*	**0.639**
Cortical bone area	0.527*	**0.542**
Internal bone area	0.213	**0.215**
Subcutaneous fat area	**0.282**	0.226
Intermuscular fat area	**0.069**	0.050

Table 6.1: The explain variance score for linear and non-linear model. * represents p-value < 0.05 and indicates transformed brain features are related to the target variable significantly. The bold explain variance score means better prediction ability.

From Table 6.1, we observe that from linear regression transformed brain features are significantly asso-

Region	Female Linear Regression (Linear relationship)	Female Proposed Model (Non-linear relationship)	Male Linear Regression (Linear relationship)	Male Proposed Model (Non-linear relationship)
Muscle area	**0.413**	0.395	**0.395***	0.373
Cortical bone area	**-0.031**	-0.092	**0.044***	-0.122
Internal bone area	0.009	**0.074**	0.121*	**0.188**
Subcutaneous fat area	0.009	**0.020**	-0.031	**0.083**
Intermuscular fat area	-0.082	**-0.051**	0.003	**0.007**

Table 6.2: The EVS for linear and non-linear model based on male and female. * represents p-value $<$ 0.05 and indicates transformed brain features are related to the target variable significantly. The bold explain variance score means better prediction ability.

ciated with muscle and cortical bone area. The proposed model has a better EVS of 0.639, 0.542, and 0.215 than the linear model in muscle, cortical bone, and internal bone area. From Fig 6.7, the linear and non-linear plot of cortical bone and subcutaneous fat has two obvious clusters, which are separated by sex.

To remove the sex effect, we divided the whole BLSA database based on sex and followed previous data split rules. The linear regression formula is shown as:

$$output\,feature = intercept + \beta_0 age + \beta_1 MCI + \beta_2 input\,feature + \varepsilon \tag{6.7}$$

Table 6.2 displays the EVS for males and females in both linear and non-linear relationships for four different body composition areas. With the exception of muscle, the EVS for all other body compositions is either close to zero or negative, regardless of whether a linear or non-linear model is used. For the muscle area, the linear model has a slightly better explained variance score of 0.413 and 0.395 compared with the non-linear model.

6.3.8 Predict hippocampus volume using body composition area

As mentioned in [134, 173, 114], body composition is an important indicator of cognitive function. The hippocampus is highly associated with cognitive function[14]. We hypothesize that body metrics are associated with hippocampus structure. Thus, the body features are used to predict hippocampus volume to test this hypothesis. Similar to section 6.3.5, the whole brain volume (aggregating all region volumes of SLANT), age, sex, and MCI are concatenated to the feature map after the select layer in the non-linear model. We iterate the number of select features from 1 to 5 with step 1 and choose an optimal model based on performance on the validation cohort. As for the linear model, PCA is applied to perform feature reduction on body composition features and takes the form of Eq. 6.8

$$output\,feature = intercept + \beta_0 age + \beta_1 sex + \beta_2 MCI + \beta_3 Wholebrain + \beta_4 input\,feature + \varepsilon \tag{6.8}$$

Two experiments are conducted to examine the impact of whole brain volume on the relationship between transformed body features and left and right hippocampus volume. The first experiment does not include whole brain volume, while the second experiment does. The resulting EVS is presented in Table 6.3 and bland-Altman plot is shown in Figure 6.8

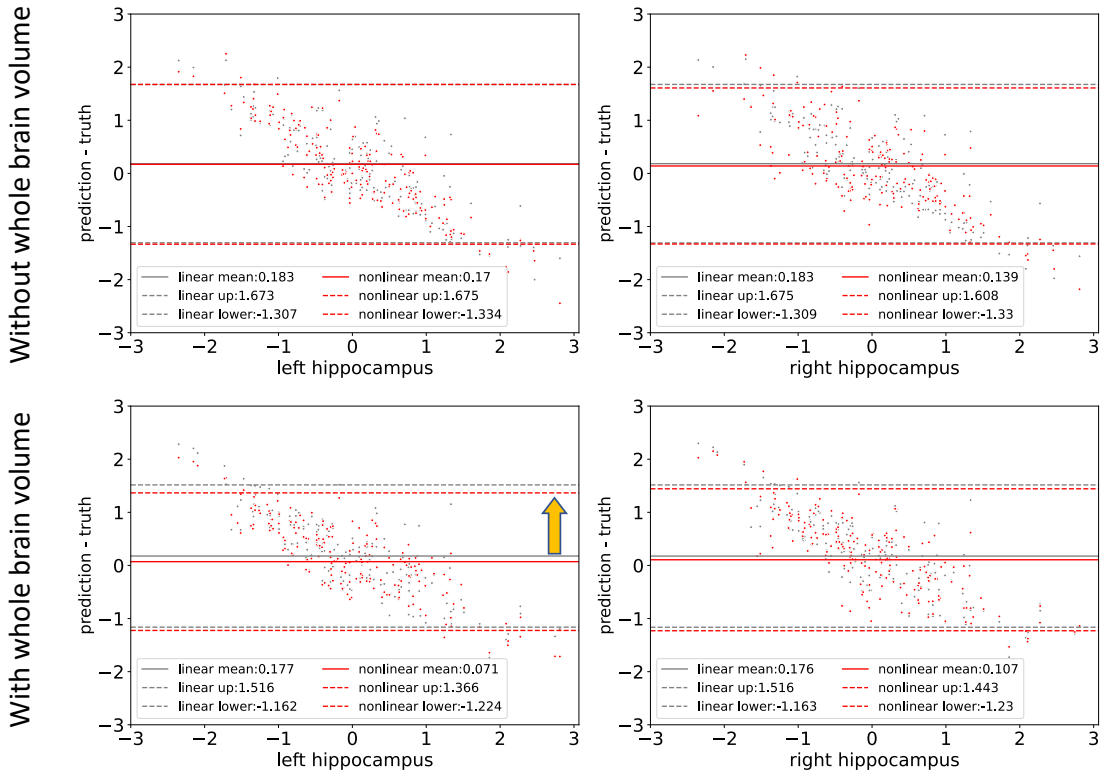


Figure 6.8: The bland-Altman plot of linear regression and proposed nonlinear methods. Each plot has the difference between prediction and truth as the y-axis and ground truth as the x-axis. The gray color represents the linear regression and the red color represents the proposed methods. As pointed out by the arrow, we can observe the proposed methods have a smaller limit of agreement in left hippocampus prediction compared with linear regression when we include whole brain volume into the estimation process.

From table 6.3, it is apparent that without whole brain volume, transformed body features are significantly associated with both left and right hippocampus volume in the linear model. After introducing whole brain volume, both the linear and non-linear models produce better EVS. The non-linear model outperforms the linear model in terms of EVS, except for right hippocampus volume.

	Linear regression with whole brain volume	Proposed model with whole brain volume	Linear regression without whole brain volume	Proposed model without whole brain volume
Left hippocampus volume	0.415	**0.443**	0.215*	**0.221**
Right hippocampus volume	0.372	**0.446**	**0.227***	0.194

Table 6.3: The explain variance score for linear and non-linear model for left and right hippocampus volume prediction based on body composition areas. * represents p-value < 0.05 and indicates transformed body features are related to the target variable significantly. The bold explain variance score means better prediction ability.

6.4 Discussion and conclusion

In this study, we have proposed to apply Gumbel-softmax to regularize MLP to analyze the relationship between brain volumes and body regions. The Gumbel-softmax is capable of subsetting input features and regularizing the deep neural network to reflect the relationship between selected features and their associated output variables. Our results, presented in Fig 6.4, demonstrate that the Gumbel-softmax model can accurately recognize the real signals associated with synthetic output variables. Additionally, our proposed model has been shown to have better EVS when compared to the plain MLP neural network. Furthermore, when we increased the amount of data fed into the networks, both networks exhibited an increase in EVS.

In our real data experiment, we use brain volume features along with demographic information such as age, sex, and MCI diagnosis to predict body regions for both male and female subjects. Our results, as presented in Table 6.1, demonstrate that the transformed brain features better predict muscle and cortical bone area and that these transformed features are significantly related to muscle and cortical bone metrics in linear regression analysis. When we used the proposed model to investigate the relationship between input brain features and body composition metrics, the introduction of a large-scale model and non-linear transformation led to better EVS compared to the linear model. The proposed model has higher EVS in muscle prediction. We find that selecting features of Gumbel-Softmax are left gyrus rectus volume, left inferior lateral ventricle volume, and right ventral diencephalon volume for muscle prediction. Based on [73], ventral diencephalon volume is associated with obesity and BMI, which supports the effectiveness of selecting real signals of Gumbel-softmax indirectly.

As shown in Figure 6.7, the difference between the predicted values and the actual values gradually shifted from positive to negative as the actual values increased. This indicates that the predictions were centered around the mean values with limited variance and that the output variable could not be fully predicted. Additionally, we observe two distinct clusters on the Bland-Altman plot for predicting cortical bone and subcutaneous fat, which is attributed to sex. To investigate this further, we separated the entire dataset into two cohorts based on gender and repeated the same experiment setting for male and female body composition metric prediction. Combining Table 6.2 with Table 6.1, we find that the EVS drops significantly for all body

composition metrics prediction except for muscle. This indicates that sex explains a significant portion of the EVS when fitting the whole dataset and predicting body composition metrics except for muscle.

In another real data experiment, we attempt to predict hippocampus volume using body metrics. We investigated the effect of whole brain volume in the linear regression model during the experiment. The results presented in our analysis show that transformed body metrics were significantly associated with hippocampus volume when whole brain volume was not included. However, when we include whole brain volume, the body metrics are not significantly associated with hippocampus volume. We believe that the reason for this finding is that transformed body composition metrics contain similar information to whole brain volume. Therefore, body composition metrics become insignificant when including whole brain volume in the model. Moreover, our analysis reveals that the proposed model had better EVS compared to the linear regression model, except for predicting left hippocampus volume without whole brain volume.

In this study, we proposed to apply Gumbel-softmax to plain MLP to characterize the relationship between brain features and body features in a non-linear approach. We demonstrate the effectiveness of Gumbel-softmax using a synthetic dataset. We next use brain region volumes to predict body regions and find the muscle that is most predictable with the highest EVS compared with cortical bone, internal bone, subcutaneous fat, and intermuscular fat. Combined with another experiment using body regions to predict hippocampus volume, we noticed that the non-linear model has better EVS compared with linear regression when the output variable is predictable.

In the literature, there exists a significant amount of heterogeneity regarding characterizing the relationship between the brain and body. Most of them focused on correlation within samples instead of predicting new individuals. Different from previous literature, this study aims to predict body features using brain features and vice versa directly and can potentially help resolve some of the discrepancies regarding which brain signals are associated with body variables. With the proposed approach's ability to link input features with output variables directly in a non-linear way, it could potentially enable researchers to better understand the connection between the brain and body.

CHAPTER 7

Conclusion and Future Works

7.1 Impact of the Dissertation

In this dissertation, we leverage statistical, machine learning, and deep learning approaches within healthcare, focusing primarily on advanced medical image analysis. Our exploration encompasses foundational techniques including data assurance, segmentation, and synthesis. Specifically, we address the challenges associated with structure MRI and single CT thigh slices from the BLSA dataset.

BLSA commenced in 1958 and has consistently enrolled volunteers across a broad age range, conducting follow-ups every 1 to 4 years. At each visit, certified technicians collect clinical and functional data, including brain structure MRI, diffusion MRI, and single thigh CT slices, to study the aging process. This dissertation primarily utilizes the BLSA dataset for investigations.

The brain atlas is an essential way to understand the structure of the white matter pathway. However, we found that many current atlases overlook or oversimplify white matter structures. While some atlases label white matter using diffusion MRI, they often have limitations in coverage, methodology, and sample size. We resent a new population-based collection of white matter atlases represented in both volumetric and surface coordinates in a T1w standard space. Technique validation shows that certain pathways are not exactly the same even with the same name. It also demonstrated that certain pathways are different across different data sources. The details can be found in Chapter 2.

Even though the white matter atlas can propagate white matter labels to T1w without dMRI, the average representation of atlas impedes the subject-specific segmentation of the white matter pathway. We proposed spatial localized deep neural networks to derive personalized white matter pathways from T1w. We validated that deep neural networks have better generalizability and specificity for personalized T1w. The details can be found in Chapter 3.

After investigating brain white matter tractography, we focus on the body composition of mid-thigh CT slices. Body composition refers to quantifying the distribution of muscle, fat, and bone. Tissue segmentation is the foundation for body composition. While the deep neural network has impressive performance in medical image segmentation, it requires tons of annotated data. Annotating biomedical images is not only tedious and time-consuming but also demands costly, specialty-oriented knowledge and skills, which are not easily accessible. To handle this challenge, we proposed a two-stage transfer learning strategy. We derive pseudo labels from CT slices based on anatomic appearance and warm-up segmentation model from scratch. Then,

expert labels are fed to optimize the segmentation model. Our proposed methods are evaluated on the thigh and lower leg slices. We validated that transfer learning based on pseudo labels can achieve data-efficient segmentation compared with only relying on limited human expert labels. The details can be found in Chapter 4.

Except for body composition, muscle group segmentation is also important for sarcopenia diagnosis and muscle strength estimate. However, we still face annotation problems. We proposed the domain adaptation and self-training method to transfer labels from public MRI volume to single CT slices. The experiment and ablation study demonstrate the effectiveness and robustness of the proposed methods. We validated that self-training can improve robustness when handling the domain gap between MR and CT. The details can be found in Chapter 5.

Previous contributions focused on the investigation of the brain and body separately. We have taken strides towards understanding anatomical priors of white matter pathways derived from structure T1w and quantitative body composition within a single CT slice through data-efficient learning. Our final contribution pulled together these innovations to characterize the relationship between MRI-derived structural brain measures and quantified body composition. We proposed to use Gumbel-softmax to extract a subset of input features as real signals to predict outputs through the deep neural network. We validated the effectiveness of Gumbel-softmax by synthetic dataset and found that a deep neural network has better explainability compared with a linear model. The details can be found in Chapter 6.

7.2 Future works

In this section, we propose to explore future directions based on the current contributed works.

7.2.1 Slice to volume generation through conditional score-based diffusion model

Generative modeling has undergone a remarkable evolution in recent years, with the advent of diverse frameworks specifically designed to estimate intricate and high-dimensional data distributions. Notably, the score-based diffusion model has risen to prominence, distinguished by its innovative approach to data distribution modeling. Unlike traditional generative models, which typically parameterize the data distribution or its transformation directly, score-based diffusion models emphasize the estimation of the gradient of the log-likelihood of data. This score elucidates strategies to adjust a random sample to enhance its probability within the model's perceived data distribution.

Leveraging these scores, the model employs a methodical diffusion process to incrementally transform samples from a Gaussian distribution into those that represent the target data distribution closely. By avoiding an explicit definition of the likelihood function directly, the model sidesteps pitfalls such as mode collapse, a

challenge frequently encountered in other generative frameworks. Instead, it harnesses noise-corrupted data iterations during training, capitalizing on the interplay between noise intensity and the gradient of the log data density.

Within the context of the BLSA dataset, body composition analysis is often conducted using a single slice of the abdomen or mid-thigh, obtained during imaging acquisition phases. However, the positional variance inherent in using a single slice can introduce noise, potentially affecting subsequent analyses. To address this challenge, we can transform the existing 3D volumes of the thigh or abdomen into sequential slices along pointed directions akin to video frames. During training, we innovatively employ random masking of preceding frames either to anticipate current frames or to utilize future frames for current frame prediction, thereby emulating volume generation in an auto-regressive manner.

7.2.2 Leveraging large language model for text generation for characterization brain and body

Large Language Models (LLMs) are important foundation models in natural language processing. Trained on vast amounts of textual data, these models demonstrate an impressive ability to understand, generate, and interact using human language. Their architecture and scale enable them to capture contextual meanings from diverse linguistic sources.

LLMs can be applied in multi-modality domains such as the medical imaging domain. For instance, the model is expected to predict brain volume, brain ages, or other brain-related metrics given single thigh slices. This designed paradigm offers unparalleled advantages. The inherent explainability ensures that generated records are both detailed and contextually relevant, translating intricate visual data into interpretable content. Their precision ensures no nuanced feature is overlooked, while their versatility enables concurrent processing of diverse data forms, such as textual medical histories, for a holistic analysis. LLMs guarantee consistent and rapid record generation. Their adaptability also allows for outputs tailored to varied audiences, and the inherent feedback loop ensures ongoing refinement and accuracy improvement in response to expert input.

References

[1] Abubakar Abid, Muhammad Fatih Balin, and James Zou. Concrete autoencoders for differentiable feature selection and reconstruction. *arXiv preprint arXiv:1901.09346*, 2019.

[2] Daniel C Alexander. Multiple-fiber reconstruction algorithms for diffusion mri. *Annals of the New York Academy of Sciences*, 1064(1):113–133, 2005.

[3] J L Andersson, S Skare, and J Ashburner. How to correct susceptibility distortions in spin-echo echo-planar images: Application to diffusion tensor imaging. *Neuroimage*, 20, 2003.

[4] D B Archer, D E Vaillancourt, and S A Coombes. A template and probabilistic atlas of the human sensorimotor tracts using diffusion MRI. *Cereb Cortex*, 28, 2018.

[5] Brian B Avants, Nick Tustison, Gang Song, and Others. Advanced normalization tools (ANTS). *Insight j*, 2(365):1–35, 2009.

[6] K M Baarsen, M Kleinnijenhuis, S Jbabdi, S N Sotiropoulos, J A Grotenhuis, and A M Walsum. A probabilistic atlas of the cerebellar white matter. *Neuroimage*, 124, 2015.

[7] Siqi Bao and Albert C.S. Chung. Multi-scale structured CNN with label consistency for brain MR image segmentation. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging and Visualization*, 6(1):113–117, jan 2018.

[8] Yoann Barnouin, Gillian Butler-Browne, Thomas Voit, David Reversat, Noura Azzabou, Gaëlle Leroux, Anthony Behin, Jamie S McPhee, Pierre G Carlier, and Jean-Yves Hogrel. Manual segmentation of individual muscles of the quadriceps femoris using MRI: a reappraisal. *Journal of Magnetic Resonance Imaging*, 40(1):239–247, 2014.

[9] Peter J Basser. Fiber-tractography via diffusion tensor mri (dt-mri). In *Proceedings of the 6th Annual Meeting ISMRM, Sydney, Australia*, volume 1226, 1998.

[10] Peter J Basser, James Mattiello, and Denis LeBihan. Estimation of the effective self-diffusion tensor from the nmr spin echo. *Journal of Magnetic Resonance, Series B*, 103(3):247–254, 1994.

[11] Stefan Bauer, Roland Wiest, Lutz-P Nolte, and Mauricio Reyes. A survey of MRI-based medical image analysis for brain tumor studies. *Physics in Medicine & Biology*, 58(13):R97, 2013.

[12] Dani Beck, Ann-Marie G De Lange, Dag Alnæs, Ivan I Maximov, Mads L Pedersen, Olof Dahlqvist Leinhard, Jennifer Linge, Rozalyn Simon, Geneviève Richard, Kristine M Ulrichsen, et al. Adipose tissue distribution from body mri is associated with cross-sectional and longitudinal brain age in adults. *NeuroImage: Clinical*, 33:102949, 2022.

[13] T E Behrens, H Johansen-Berg, M W Woolrich, S M Smith, C A Wheeler-Kingshott, and P A Boulby. Non-invasive mapping of connections between human thalamus and cortex using diffusion imaging. *Nat Neurosci*, 6, 2003.

[14] Luis EB Bettio, Luckshi Rajendran, and Joana Gil-Mohapel. The effects of aging in the hippocampus and cognitive decline. *Neuroscience & Biobehavioral Reviews*, 79:66–86, 2017.

[15] Jennifer F Bobb, Brian S Schwartz, Christos Davatzikos, and Brian Caffo. Cross-sectional and longitudinal association of body mass index and brain volume. *Human brain mapping*, 35(1):75–88, 2014.

[16] Vadim Borisov, Tobias Leemann, Kathrin Seßler, Johannes Haug, Martin Pawelczyk, and Gjergji Kasneci. Deep neural networks and tabular data: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[17] David J Brenner and Eric J Hall. Computed tomography—an increasing source of radiation exposure. *New England journal of medicine*, 357(22):2277–2284, 2007.

[18] Herb Brody. Medical imaging. *Nature*, 502(7473):S81–S81, 2013.

[19] U Bürgel, K Amunts, L Hoemke, H Mohlberg, J M Gilsbach, and K Zilles. White matter fiber tracts of the human brain: Three-dimensional mapping at microscopic resolution, topography and intersubject variability. *Neuroimage*, 29, 2006.

[20] Mariano Cabezas, Arnau Oliver, Xavier Lladó, Jordi Freixenet, and Meritxell Bach Cuadra. A review of atlas-based segmentation for magnetic resonance brain images. *Computer methods and programs in biomedicine*, 104(3):e158—-e177, 2011.

[21] Leon Y Cai, Qi Yang, Colin B Hansen, Vishwesh Nath, Karthik Ramadass, Graham W Johnson, Benjamin N Conrad, Brian D Boyd, John P Begnoche, Lori L Beason-Held, and Others. PreQual: An automated pipeline for integrated preprocessing and quality assurance of diffusion weighted MRI images. *Magnetic Resonance in Medicine*, 86(1):456–470, 2021.

[22] Marco Catani and Michel Thiebaut De Schotten. *Atlas of human brain connections*. Oxford University Press, 2012.

[23] Marco Catani, Robert J Howard, Sinisa Pajevic, and Derek K Jones. Virtual in vivo interactive dissection of white matter fasciculi in the human brain. *Neuroimage*, 17(1):77–94, 2002.

[24] Marco Catani and Michel Thiebaut de Schotten. A diffusion tensor imaging tractography atlas for virtual in vivo dissections. *Cortex*, 44(8):1105–1132, 2008.

[25] Wei-Lun Chang, Hui-Po Wang, Wen-Hsiao Peng, and Wei-Chen Chiu. All about structure: Adapting structural information across domains for boosting semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1900–1909, 2019.

[26] Jianbo Chen, Le Song, Martin Wainwright, and Michael Jordan. Learning to explain: An information-theoretic perspective on model interpretation. In *International Conference on Machine Learning*, pages 883–892. PMLR, 2018.

[27] Liang Chen, Paul Bentley, Kensaku Mori, Kazunari Misawa, Michitaka Fujiwara, and Daniel Rueckert. Self-supervised learning for medical image analysis using image context restoration. *Medical image analysis*, 58:101539, 2019.

[28] Q Chenot, N Tzourio-Mazoyer, F Rheault, M Descoteaux, F Crivello, L Zago, E Mellet, G Jobard, M Joliot, B Mazoyer, and L Petit. A population-based atlas of the human pyramidal tract in 410 healthy participants. *Brain Struct Funct*, 224, 2019.

[29] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*, pages 424–432. Springer, 2016.

[30] Pierrick Coupé, Boris Mansencal, Michaël Clément, Rémi Giraud, Baudouin Denis de Senneville, Vinh-Thong Ta, Vincent Lepetit, and José V Manjon. AssemblyNet: A novel deep decision-making process for whole brain MRI segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 466–474, 2019.

[31] A M Dale, B Fischl, and M I Sereno. Cortical surface-based analysis. I Segmentation and surface reconstruction. *Neuroimage*, 9, 1999.

[32] Christos Davatzikos and N Bryan. Using a deformable surface model to obtain a shape representation of the cortex. *IEEE transactions on medical imaging*, 15(6):785–795, 1996.

[33] Alexandre De Brébisson and Giovanni Montana. Deep neural networks for anatomical brain segmentation. Technical report, 2015.

[34] Jonas de Carvalho Felinto, Katia Maria Poloni, Paulo Guilherme de Lima Freire, Jessica Bianca Aily, Aline Castilho de Almeida, Maria Gabriela Pedroso, Stela Márcia Mattiello, and Ricardo José Ferrari. Automatic segmentation and quantification of thigh tissues in CT images. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 10960 LNCS, pages 261–276. Springer Verlag, 2018.

[35] Yue-Ting Deng, Yu-Zhu Li, Shu-Yi Huang, Ya-Nan Ou, Wei Zhang, Shi-Dong Chen, Ya-Ru Zhang, Liu Yang, Qiang Dong, Jian-Feng Feng, et al. Association of life course adiposity with risk of incident dementia: a prospective cohort study of 322,336 participants. *Molecular Psychiatry*, pages 1–11, 2022.

[36] Richard A Dennis, Douglas E Long, Reid D Landes, Kalpana P Padala, Prasad R Padala, Kimberly K Garner, James N Wise, Charlotte A Peterson, and Dennis H Sullivan. Tutorial for using SliceOmatic to calculate thigh area and composition from computed tomography images from older adults. *PloS one*, 13(10):e0204529, 2018.

[37] R S Desikan, F Segonne, B Fischl, B T Quinn, B C Dickerson, and D Blacker. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage*, 31, 2006.

[38] Lee R Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945.

[39] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[40] S B Eickhoff, S Heim, K Zilles, and K Amunts. Testing anatomically specified hypotheses in functional imaging using cytoarchitectonic maps. *Neuroimage*, 32, 2006.

[41] S B Eickhoff, T Paus, S Caspers, M H Grosbras, A C Evans, K Zilles, and K Amunts. Assignment of functional activations to probabilistic cytoarchitectonic areas revisited. *Neuroimage*, 36, 2007.

[42] S B Eickhoff, K E Stephan, H Mohlberg, C Grefkes, G R Fink, K Amunts, and K Zilles. A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. *Neuroimage*, 25, 2005.

[43] Klaus Engelke, Oleg Museyko, Ling Wang, and Jean Denis Laredo. Quantitative analysis of skeletal muscle by computed tomography imaging—State of the art, oct 2018.

[44] S Ewert, P Plettig, N Li, M M Chakravarty, D L Collins, T M Herrington, A A Kühn, and A Horn. Toward defining deep brain stimulation targets in MNI space: A subcortical atlas based on multimodal MRI, histology and structural connectivity. *Neuroimage*, 170, 2018.

[45] Luigi Ferrucci. The Baltimore Longitudinal Study of Aging (BLSA): A 50-Year-Long Journey and Plans for the Future. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, 63(12):1416–1419, dec 2008.

[46] Teresa D Figley, Behnoush Mortazavi Moghadam, Navdeep Bhullar, Jennifer Kornelsen, Susan M Courtney, and Chase R Figley. Probabilistic white matter atlases of human auditory, basal ganglia, language, precuneus, sensorimotor, visual and visuospatial networks. *Frontiers in human neuroscience*, 11:306, 2017.

[47] B Fischl. FreeSurfer. *Neuroimage*, 62, 2012.

[48] B Fischl, M I Sereno, and A M Dale. Cortical surface-based analysis. II: Inflation, flattening, and a surface-based coordinate system. *Neuroimage*, 9, 1999.

[49] Vladimir Fonov, Alan C Evans, Kelly Botteron, C Robert Almli, Robert C McKinstry, D Louis Collins, Brain Development Cooperative Group, and Others. Unbiased average age-appropriate atlases for pediatric studies. *Neuroimage*, 54(1):313–327, 2011.

[50] Peter Francis, Mark Lyons, Mathew Piasecki, Jamie Mc Phee, Karen Hind, and Philip Jakeman. Measurement of muscle health in aging. *Biogerontology*, 18(6):901–911, 2017.

[51] John DE Gabrieli, Satrajit S Ghosh, and Susan Whitfield-Gabrieli. Prediction as a humanitarian and pragmatic contribution from human cognitive neuroscience. *Neuron*, 85(1):11–26, 2015.

[52] Eleftherios Garyfallidis, Matthew Brett, Bagrat Amirbekian, Ariel Rokem, Stefan Van Der Walt, Maxime Descoteaux, and Ian Nimmo-Smith. Dipy, a library for the analysis of diffusion MRI data. *Frontiers in neuroinformatics*, 8:8, 2014.

[53] Eleftherios Garyfallidis, Marc Alexandre Côté, Francois Rheault, Jasmeen Sidhu, Janice Hau, Laurent Petit, David Fortin, Stephen Cunanne, and Maxime Descoteaux. Recognition of white matter bundles using local and global streamline-based registration and clustering, apr 2018.

[54] J C Gee, M Reivich, and R Bajcsy. Elastically deforming 3D atlas to match anatomical brain images. *J Comput Assist Tomogr*, 17, 1993.

[55] George Gifford, Robert McCutcheon, and Philip McGuire. Neuroimaging studies in people at clinical high risk for psychosis. In *Risk Factors for Psychosis*, pages 167–182. Elsevier, 2020.

[56] Matthew F Glasser, Timothy S Coalson, Emma C Robinson, Carl D Hacker, John Harwell, Essa Yacoub, Kamil Ugurbil, Jesper Andersson, Christian F Beckmann, Mark Jenkinson, and Others. A multi-modal parcellation of human cerebral cortex. *Nature*, 536(7615):171–178, 2016.

[57] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[58] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[59] Hao Guan and Mingxia Liu. Domain adaptation for medical image analysis: a survey. *IEEE Transactions on Biomedical Engineering*, 69(3):1173–1185, 2021.

[60] Miguel Guevara, Claudio Román, Josselin Houenou, Delphine Duclap, Cyril Poupon, Jean François Mangin, and Pamela Guevara. Reproducibility of superficial white matter tracts using diffusion-weighted imaging tractography. *Neuroimage*, 147:703–725, 2017.

[61] P. Guevara, D. Duclap, C. Poupon, L. Marrakchi-Kacem, P. Fillard, D. Le Bihan, M. Leboyer, J. Houenou, and J. F. Mangin. Automatic fiber bundle segmentation in massive tractography datasets using a multi-subject bundle atlas. *NeuroImage*, 61(4):1083–1099, jul 2012.

[62] Antonio Gulli and Sujit Pal. *Deep learning with Keras*. Packt Publishing Ltd, 2017.

[63] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

[64] Colin B Hansen, Qi Yang, Ilwoo Lyu, Francois Rheault, Cailey Kerley, Bramsh Qamar Chandio, Shreyas Fadnavis, Owen Williams, Andrea T Shafer, Susan M Resnick, et al. Pandora: 4-d white matter bundle population-based atlases derived from diffusion mri fiber tractography. *Neuroinformatics*, 19(3):447–460, 2021.

[65] Colin B Hansen, Qi Yang, Ilwoo Lyu, Francois Rheault, Cailey Kerley, Bramsh Qamar Chandio, Shreyas Fadnavis, Owen Williams, Andrea T Shafer, Susan M Resnick, and Others. Pandora: 4-D white matter bundle population-based atlases derived from diffusion MRI fiber tractography. *Neuroinformatics*, pages 1–14, 2020.

[66] Rolf A Heckemann, Joseph V Hajnal, Paul Aljabar, Daniel Rueckert, and Alexander Hammers. Automatic anatomical brain MRI segmentation combining label propagation and decision fusion. *NeuroImage*, 33(1):115–126, 2006.

[67] Tobias Heimann, Brain Van Ginneken, Martin A. Styner, Yulia Arzhaeva, Volker Aurich, Christian Bauer, Andreas Beck, Christoph Becker, Reinhard Beichel, György Bekes, Fernando Bello, Gerd Binnig, Horst Bischof, Alexander Bornik, Peter M.M. Cashman, Ying Chi, Andrés Córdova, Benoit M. Dawant, Márta Fidrich, Jacob D. Furst, Daisuke Furukawa, Lars Grenacher, Joachim Hornegger, Dagmar Kainmüller, Richard I. Kitney, Hidefumi Kobatake, Hans Lamecker, Thomas Lange, Jeongjin Lee, Brian Lennon, Rui Li, Senhu Li, Hans Peter Meinzer, Gábor Németh, Daniela S. Raicu, Anne Mareike Rau, Eva M. Van Rikxoort, Mikaël Rousson, Lászlo Ruskó, Kinda A. Saddi, Günter Schmidt, Dieter Seghers, Akinobu Shimizu, Pieter Slagmolen, Erich Sorantin, Grzegorz Soza, Ruchaneewan Susomboon, Jonathan M. Waite, Andreas Wimmer, and Ivo Wolf. Comparison and evaluation of methods for liver segmentation from CT datasets. *IEEE Transactions on Medical Imaging*, 28(8):1251–1265, aug 2009.

[68] C L Hoad and A L Martel. Segmentation of MR images for computer-assisted surgery of the lumbar spine. *Physics in Medicine & Biology*, 47(19):3503, 2002.

[69] Myles Hollander, Douglas A Wolfe, and Eric Chicken. *Nonparametric statistical methods*. John Wiley & Sons, 2013.

[70] Martin Hudelmaier, Wolfgang Wirth, Maria Himmer, Susanne Ring-Dimitriou, Alexandra Sänger, and Felix Eckstein. Effect of exercise intervention on thigh muscle volume and anatomical cross-sectional areas—Quantitative assessment using MRI. *Magnetic resonance in medicine*, 64(6):1713–1720, 2010.

[71] Yuankai Huo, Zhoubing Xu, Hyeonsoo Moon, Shunxing Bao, Albert Assad, Tamara K Moyo, Michael R Savona, Richard G Abramson, and Bennett A Landman. Synseg-net: Synthetic segmentation without target modality ground truth. *IEEE transactions on medical imaging*, 38(4):1016–1025, 2018.

[72] Yuankai Huo, Zhoubing Xu, Yunxi Xiong, Katherine Aboud, Prasanna Parvathaneni, Shunxing Bao, Camilo Bermudez, Susan M. Resnick, Laurie E. Cutting, and Bennett A. Landman. 3D whole brain segmentation using spatially localized atlas network tiles. *NeuroImage*, 194:105–119, jul 2019.

[73] In-seong Hwang and Soon-Beom Hong. Association between body mass index and subcortical volume in pre-adolescent children with autism spectrum disorder: An exploratory study. *Autism Research*, 15(12):2238–2249, 2022.

[74] Juan Eugenio Iglesias and Mert R Sabuncu. Multi-atlas segmentation of biomedical images: a survey. *Medical image analysis*, 24(1):205–219, 2015.

[75] I Ilinsky, A Horn, P Paul-Gilloteaux, P Gressens, C Verney, and K Kultas-Ilinsky. Human motor thalamus reconstructed in 3D from continuous sagittal sections with identified subcortical afferent territories. *eNeuro*, 5, 2018.

[76] Aliz Lacoste Jeanson, Jn Dupej, Chiara Villa, and Jaroslav Brek. Body composition estimation from selected slices: equations computed from a new semi-automatic thresholding method developed on whole-body ct scans. *PeerJ*, 5:e3302, 2017.

[77] M Jenkinson, C F Beckmann, T E Behrens, M W Woolrich, and S M Smith. Fsl. *Neuroimage*, 62, 2012.

[78] Ben Jeurissen, Maxime Descoteaux, Susumu Mori, and Alexander Leemans. Diffusion mri fiber tractography of the brain. *NMR in Biomedicine*, 32(4):e3785, 2019.

[79] MP Jones, JB Dilley, Douglas Drossman, and MD Crowell. Brain–gut connections in functional gi disorders: anatomic and physiologic relationships. *Neurogastroenterology & Motility*, 18(2):91–103, 2006.

[80] Arlind Kadra, Marius Lindauer, Frank Hutter, and Josif Grabocka. Well-tuned simple nets excel on tabular datasets. *Advances in neural information processing systems*, 34:23928–23941, 2021.

[81] Nick Kanopoulos, Nagesh Vasanthavada, and Robert L Baker. Design of an image edge detection filter using the sobel operator. *IEEE Journal of solid-state circuits*, 23(2):358–367, 1988.

[82] M C Keuken, P L Bazin, A Schafer, J Neumann, R Turner, and B U Forstmann. Ultra-high 7T MRI of structural age-related changes of the subthalamic nucleus. *J Neurosci*, 33, 2013.

[83] R Kikinis, M E Shenton, D V Iosifescu, R W McCarley, P Saiviroonporn, H H Hokama, A Robatino, D Metcalf, C G Wible, C M Portas, R M Donnino, and F A Jolesz. A digital brain atlas for surgical planning, model-driven segmentation, and teaching. *IEEE Trans Vis Comput Graph*, 2, 1996.

[84] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[85] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[86] L Labache, M Joliot, J Saracco, G Jobard, I Hesling, L Zago, E Mellet, L Petit, F Crivello, B Mazoyer, and N Tzourio-Mazoyer. A SENtence Supramodal areas AtlaS (SENSAAS) based on multiple task-induced activation mapping and graph analysis of intrinsic connectivity in 144 healthy right-handers. *Brain Struct Funct*, 224, 2019.

[87] J L Lancaster, M G Woldorff, L M Parsons, M Liotti, C S Freitas, L Rainey, P V Kochunov, D Nickerson, S A Mikiten, and P T Fox. Automated Talairach atlas labels for functional brain mapping. *Hum Brain Mapp*, 10, 2000.

[88] Eric B Larson. The body-mind connection in aging and dementia. *Journal of the American Geriatrics Society*, 61(7):1210, 2013.

[89] Ross M Lawrence, Eric W Bridgeford, Patrick E Myers, Ganesh C Arvapalli, Sandhya C Ramachandran, Derek A Pisner, Paige F Frank, Allison D Lemmer, Aki Nikolaidis, and Joshua T Vogelstein. Standardizing human brain parcellations. *Scientific data*, 8(1):1–9, 2021.

[90] Stephen M Lawrie and Suheib S Abukmeil. Brain abnormality in schizophrenia: a systematic and quantitative review of volumetric magnetic resonance imaging studies. *The British Journal of Psychiatry*, 172(2):110–120, 1998.

[91] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[92] Ho Hin Lee, Yucheng Tang, Shunxing Bao, Yan Xu, Qi Yang, Xin Yu, Agnes B Fogo, Raymond Harris, Mark P de Caestecker, Jeffery M Spraggins, et al. Supervised deep generation of high-resolution arterial phase computed tomography kidney substructure atlas. In *Medical Imaging 2022: Image Processing*, volume 12032, pages 736–743. SPIE, 2022.

[93] Ho Hin Lee, Yucheng Tang, Shunxing Bao, Qi Yang, Xin Yu, Kevin L Schey, Jeffery M Spraggins, Yuankai Huo, and Bennett A Landman. Unsupervised registration refinement for generating unbiased eye atlas. In *Medical Imaging 2023: Image Processing*, volume 12464, pages 470–476. SPIE, 2023.

[94] Lay Khoon Lee, Siau Chuin Liew, and Weng Jie Thong. A review of image segmentation methodologies in medical image. *Advanced computer and communication engineering technology*, pages 1069–1080, 2015.

[95] Glenn N Levine. The mind-heart-body connection. *Circulation*, 140(17):1363–1365, 2019.

[96] Chunming Li, Chenyang Xu, Changfeng Gui, and Martin D Fox. Distance regularized level set evolution and its application to image segmentation. *IEEE transactions on image processing*, 19(12):3243–3254, 2010.

[97] Kang Li, Shujun Wang, Lequan Yu, and Pheng-Ann Heng. Dual-teacher++: Exploiting intra-domain and inter-domain knowledge with reliable transfer for cardiac segmentation. *IEEE Transactions on Medical Imaging*, 2020.

[98] Xiaomeng Li, Hao Chen, Xiaojuan Qi, Qi Dou, Chi-Wing Fu, and Pheng-Ann Heng. H-DenseUNet: hybrid densely connected UNet for liver and tumor segmentation from CT volumes. *IEEE transactions on medical imaging*, 37(12):2663–2674, 2018.

[99] K I Lim, S J Yang, T N Kim, H J Yoo, H J Kang, W Song, Sei-Hyun Baik, D S Choi, and Kyung Mook Choi. The association between the ratio of visceral fat to thigh muscle area and metabolic syndrome: the Korean Sarcopenic Obesity Study (KSOS). *Clinical endocrinology*, 73(5):588–594, 2010.

[100] Chen Liu, Matthew Amodio, Liangbo L Shen, Feng Gao, Arman Avesta, Sanjay Aneja, Jay C Wang, Lucian V Del Priore, and Smita Krishnaswamy. Cuts: A fully unsupervised framework for medical image segmentation. *arXiv preprint arXiv:2209.11359*, 2022.

[101] N Makris, J M Goldstein, D Kennedy, S M Hodge, V S Caviness, S V Faraone, M T Tsuang, and L J Seidman. Decreased volume of left and total anterior insular lobule in schizophrenia. *Schizophr Res*, 83, 2006.

[102] R B Mars, S Jbabdi, J Sallet, J X O'Reilly, P L Croxson, E Olivier, M P Noonan, C Bergmann, A S Mitchell, M G Baxter, T E J Behrens, H Johansen-Berg, V Tomassini, K L Miller, and M F S Rushworth. Diffusion-weighted imaging tractography-based parcellation of the human parietal cortex and comparison with human and macaque resting-state functional connectivity. *J Neurosci*, 31, 2011.

[103] R B Mars, J Sallet, U Schuffelgen, S Jbabdi, I Toni, and M F Rushworth. Connectivity-based subdivisions of the human right "temporoparietal junction area": Evidence for different areas participating in different cortical networks. *Cereb Cortex*, 22, 2012.

[104] Sébastien Martin, Jocelyne Troccaz, and Vincent Daanen. Automated segmentation of the prostate in 3D MR images using a probabilistic atlas and a spatially constrained deformable model. *Medical physics*, 37(4):1579–1590, 2010.

[105] J Mazziotta, A Toga, A Evans, P Fox, J Lancaster, K Zilles, R Woods, T Paus, G Simpson, B Pike, C Holmes, L Collins, P Thompson, D MacDonald, M Iacoboni, T Schormann, K Amunts, N Palomero-Gallagher, S Geyer, L Parsons, K Narr, N Kabani, G L Goualher, D Boomsma, T Cannon, R Kawashima, and B Mazoyer. A probabilistic atlas and reference system for the human brain: International consortium for brain mapping (ICBM). *Philos Trans R Soc Lond Ser B Biol Sci*, 356, 2001.

[106] J Mazziotta, A Toga, A Evans, P Fox, J Lancaster, K Zilles, R Woods, T Paus, G Simpson, B Pike, C Holmes, L Collins, P Thompson, D MacDonald, M Iacoboni, T Schormann, K Amunts, N Palomero-Gallagher, S Geyer, L Parsons, K Narr, N Kabani, G le Goualher, J Feidler, K Smith, D Boomsma, H H Pol, T Cannon, R Kawashima, and B Mazoyer. A four-dimensional probabilistic atlas of the human brain. *Journal of the American Medical Informatics Association : JAMIA*, 8, 2001.

[107] J C Mazziotta, A W Toga, A Evans, P Fox, and J Lancaster. A probabilistic atlas of the human brain: Theory and rationale for its development. The international consortium for brain mapping (ICBM). *Neuroimage*, 2, 1995.

[108] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[109] N Mitsiopoulos, R N Baumgartner, S B Heymsfield, W Lyons, D Gallagher, and R Ross. Cadaver validation of skeletal muscle measurement by magnetic resonance imaging and computerized tomography. *Journal of applied physiology*, 85(1):115–122, 1998.

[110] Susumu Mori, Kenichi Oishi, Hangyi Jiang, Li Jiang, Xin Li, Kazi Akhter, Kegang Hua, Andreia V. Faria, Asif Mahmood, Roger Woods, Arthur W. Toga, G. Bruce Pike, Pedro Rosa Neto, Alan Evans, Jiangyang Zhang, Hao Huang, Michael I. Miller, Peter van Zijl, and John Mazziotta. Stereotaxic white matter atlas based on diffusion tensor imaging in an ICBM template. Technical Report 2, 2008.

[111] Marina Mourtzakis, Carla MM Prado, Jessica R Lieffers, Tony Reiman, Linda J McCargar, and Vickie E Baracos. A practical and precise approach to quantification of body composition in cancer patients using computed tomography images acquired during routine care. *Applied Physiology, Nutrition, and Metabolism*, 33(5):997–1006, 2008.

[112] F X Neubert, R B Mars, A G Thomas, J Sallet, and M F Rushworth. Comparison of human ventral frontal cortex areas for cognitive control and language with areas in monkey frontal cortex. *Neuron*, 81, 2014.

[113] Weili Nie and Ankit B Patel. Towards a better understanding and regularization of GAN training dynamics. In *Uncertainty in Artificial Intelligence*, pages 281–291. PMLR, 2020.

[114] Hye-Mi Noh, Sohee Oh, Hong Ji Song, Eun Young Lee, Jin-Young Jeong, Ohk-Hyun Ryu, Kyung-Soon Hong, and Dong-Hyun Kim. Relationships between cognitive function and body composition among community-dwelling older adults: a cross-sectional study. *BMC geriatrics*, 17:1–9, 2017.

[115] Lauren J. O'Donnell and Carl Fredrik Westin. Automatic tractography segmentation using a high-dimensional white matter atlas. *IEEE Transactions on Medical Imaging*, 26(11):1562–1575, nov 2007.

[116] K Oishi, K Zilles, K Amunts, A Faria, H Jiang, X Li, K Akhter, K Hua, R Woods, A W Toga, G B Pike, P Rosa-Neto, A Evans, J Zhang, H Huang, M I Miller, P C M van Zijl, J Mazziotta, and S Mori. Human brain white matter atlas: Identification and assignment of common anatomical structures in superficial white matter. *Neuroimage*, 43, 2008.

[117] Kenichi Oishi, Andreia Faria, Hangyi Jiang, Xin Li, Kazi Akhter, Jiangyang Zhang, John T. Hsu, Michael I. Miller, Peter C.M. van Zijl, Marilyn Albert, Constantine G. Lyketsos, Roger Woods, Arthur W. Toga, G. Bruce Pike, Pedro Rosa-Neto, Alan Evans, John Mazziotta, and Susumu Mori. Atlas-based whole brain white matter analysis using large deformation diffeomorphic metric mapping: Application to normal elderly and Alzheimer's disease participants. *NeuroImage*, 46(2):486–499, jun 2009.

[118] Kenichi Oishi, Andreia V Faria, Peter CM Van Zijl, and Susumu Mori. *MRI atlas of human white matter*. Academic Press, 2010.

[119] T J Overend, D A Cunningham, D H Paterson, and M S Lefcoe. Thigh composition in young and elderly men determined by computed tomography. *Clinical Physiology*, 12(6):629–640, 1992.

[120] Fei Pan, Inkyu Shin, Francois Rameau, Seokju Lee, and In So Kweon. Unsupervised intra-domain adaptation for semantic segmentation through self-supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3764–3773, 2020.

[121] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In H Wallach, H Larochelle, A Beygelzimer, F d Alché-Buc, E Fox, and R Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[122] Dzung L Pham, Chenyang Xu, and Jerry L Prince. Current methods in medical image segmentation. *Annual review of biomedical engineering*, 2(1):315–337, 2000.

[123] Kenneth A Philbrick, Alexander D Weston, Zeynettin Akkus, Timothy L Kline, Panagiotis Korfiatis, Tomas Sakinis, Petro Kostandy, Arunnit Boonrod, Atefeh Zeinoddini, Naoki Takahashi, and Others. RIL-contour: a medical imaging dataset annotation tool for and with deep learning. *Journal of digital imaging*, 32(4):571–581, 2019.

[124] Cristiano Rampinelli, Daniela Origgi, and Massimo Bellomi. Low-dose CT: technique, reading methods and image interpretation. *Cancer Imaging*, 12(3):548, 2012.

[125] F Rheault, E St-Onge, J Sidhu, K Maier-Hein, N Tzourio-Mazoyer, L Petit, and M Descoteaux. Bundle-specific tractography with incorporated anatomical and orientational priors. *Neuroimage*, 186, 2019.

[126] Francois Rheault, Etienne St-Onge, Jasmeen Sidhu, Quentin Chenot, Laurent Petit, and Maxime Descoteaux. Bundle-specific tractography. In *Computational Diffusion MRI*, pages 129–139. Springer, 2018.

[127] Alard Roebroeck, Karla L Miller, and Manisha Aggarwal. Ex vivo diffusion mri of the human brain: Technical challenges and recent advances. *NMR in Biomedicine*, 32(4):e3941, 2019.

[128] Marlena Rohm, Marius Markmann, Johannes Forsting, Robert Rehmann, Martijn Froeling, and Lara Schlaffke. 3D Automated Segmentation of Lower Leg Muscles Using Machine Learning on a Heterogeneous Dataset. *Diagnostics*, 11(10):1747, sep 2021.

[129] K Rojkova, E Volle, M Urbanski, F Humbert, F Dell'Acqua, and M Thiebaut de Schotten. Atlasing the frontal lobe connections and their variability due to age and education: A spherical deconvolution tractography study. *Brain Struct Funct*, 221, 2016.

[130] E T Rolls, C C Huang, C P Lin, J Feng, and M Joliot. Automated anatomical labelling atlas 3. *Neuroimage*, 206, 2020.

[131] Claudio Román, Miguel Guevara, Ronald Valenzuela, Miguel Figueroa, Josselin Houenou, Delphine Duclap, Cyril Poupon, Jean François Mangin, and Pamela Guevara. Clustering of whole-brain white matter short association bundles using HARDI data. *Frontiers in Neuroinformatics*, 11, dec 2017.

[132] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. Technical report.

[133] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241, 2015.

[134] Caterina Rosano, Eleanor M Simonsick, Tamara B Harris, Steven B Kritchevsky, Jennifer Brach, Marjolein Visser, Kristine Yaffe, and Anne B Newman. Association between physical and cognitive function in healthy elderly: the health, aging and body composition study. *Neuroepidemiology*, 24(1-2):8–14, 2005.

[135] J Sallet, R B Mars, M P Noonan, F X Neubert, S Jbabdi, J X O'Reilly, N Filippini, A G Thomas, and M F Rushworth. The organization of dorsal frontal cortex in humans and macaques. *J Neurosci*, 33, 2013.

[136] A Schaefer, R Kong, E M Gordon, T O Laumann, X N Zuo, A J Holmes, S B Eickhoff, and B T T Yeo. Local-global Parcellation of the human cerebral cortex from intrinsic functional connectivity MRI. *Cereb Cortex*, 28, 2018.

[137] K G Schilling, F C Yeh, V Nath, C Hansen, O Williams, S Resnick, A W Anderson, and B A Landman. A fiber coherence index for quality control of B-table orientation in diffusion MRI scans. *Magn Reson Imaging*, 58, 2019.

[138] Kurt G Schilling, Vishwesh Nath, Colin Hansen, Prasanna Parvathaneni, Justin Blaber, Yurui Gao, Peter Neher, Dogu Baran Aydogan, Yonggang Shi, Mario Ocampo-Pineda, and Others. Limits to anatomical accuracy of diffusion tractography using modern approaches. *NeuroImage*, 185:1–11, 2019.

[139] Sarah Schlaeger, Friedemann Freitag, Elisabeth Klupp, Michael Dieckmeyer, Dominik Weidlich, Stephanie Inhuber, Marcus Deschauer, Benedikt Schoser, Sarah Bublitz, Federica Montagnese, and Others. Thigh muscle segmentation of chemical shift encoding-based water-fat magnetic resonance images: the reference database MyoSegmenTUM. *PLoS One*, 13(6):e0198200, 2018.

[140] Skipper Seabold and Josef Perktold. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*, 2010.

[141] Justin Senseney, Paul F Hemler, and Matthew J McAuliffe. Automated segmentation of computed tomography images. In *2009 22nd IEEE International Symposium on Computer-Based Medical Systems*, pages 1–7, 2009.

[142] Ira Shavitt and Eran Segal. Regularization learning networks: deep learning for tabular datasets. *Advances in Neural Information Processing Systems*, 31, 2018.

[143] Dinggang Shen, Guorong Wu, and Heung-Il Suk. Deep learning in medical image analysis. *Annual review of biomedical engineering*, 19:221–248, 2017.

[144] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):1–48, 2019.

[145] Stephen M Smith, Mark Jenkinson, Mark W Woolrich, Christian F Beckmann, Timothy E J Behrens, Heidi Johansen-Berg, Peter R Bannister, Marilena De Luca, Ivana Drobnjak, David E Flitney, and Others. Advances in functional and structural MR image analysis and implementation as FSL. *Neuroimage*, 23:S208–S219, 2004.

[146] Nima Tajbakhsh, Holger Roth, Demetri Terzopoulos, and Jianming Liang. Guest Editorial Annotation-Efficient Deep Learning: The Holy Grail of Medical Imaging. *IEEE Transactions on Medical Imaging*, 40(10):2526–2533, 2021.

[147] Nima Tajbakhsh, Jae Y Shin, Suryakanth R Gurudu, R Todd Hurst, Christopher B Kendall, Michael B Gotway, and Jianming Liang. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE transactions on medical imaging*, 35(5):1299–1312, 2016.

[148] Hikaru Takeuchi, Yasuyuki Taki, Yuko Sassa, Hiroshi Hashizume, Atsushi Sekiguchi, Ai Fukushima, and Ryuta Kawashima. White matter structures associated with creativity: evidence from diffusion tensor imaging. *Neuroimage*, 51(1):11–18, 2010.

[149] J Talairach and P Tournoux. *Co-planar stereotaxic atlas of the human brain : 3-dimensional proportional system : An approach to cerebral imaging*. Georg Thieme, Stuttgart ; New York, 1988.

[150] Chaowei Tan, Zhennan Yan, Shaoting Zhang, Boubakeur Belaroussi, Hui Jing Yu, Colin Miller, and Dimitris N Metaxas. An automated and robust framework for quantification of muscle and fat in the thigh. In *2014 22nd International Conference on Pattern Recognition*, pages 3173–3178, 2014.

[151] Yucheng Tang, Riqiang Gao, Ho Hin Lee, Shizhong Han, Yunqiang Chen, Dashan Gao, Vishwesh Nath, Camilo Bermudez, Michael R Savona, Richard G Abramson, and Others. High-resolution 3D abdominal segmentation with random patch network fusion. *Medical Image Analysis*, 69:101894, 2021.

[152] Michel Thiebaut de Schotten, Dominic H. Ffytche, Alberto Bizzi, Flavio Dell'Acqua, Matthew Allin, Muriel Walshe, Robin Murray, Steven C. Williams, Declan G.M. Murphy, and Marco Catani. Atlasing location, asymmetry and inter-subject variability of white matter tracts in the human brain with MR diffusion tractography. *NeuroImage*, 54(1):49–59, jan 2011.

[153] A W Toga. *Brain warping*. Academic Press, San Diego, 1999.

[154] Arthur W Toga, Arthur W Toga, John C Mazziotta, and John C Mazziotta. *Brain mapping: the methods*, volume 1. Academic press, 2002.

[155] J D Tournier, R Smith, D Raffelt, R Tabbara, T Dhollander, M Pietsch, D Christiaens, B Jeurissen, C H Yeh, and A Connelly. MRtrix3: A fast, flexible and open software framework for medical image processing and visualisation. *Neuroimage*, 202, 2019.

[156] Jacques-Donald Tournier, Susumu Mori, and Alexander Leemans. Diffusion tensor imaging and beyond. *Magnetic resonance in medicine*, 65(6):1532, 2011.

[157] A C Tziortzi, G E Searle, S Tzimopoulou, C Salinas, J D Beaver, M Jenkinson, M Laruelle, E A Rabiner, and R N Gunn. Imaging dopamine receptors in humans with [11C]-(+)-PHNO: Dissection of D3 signal and anatomy. *Neuroimage*, 54, 2011.

[158] N Tzourio-Mazoyer, B Landeau, D Papathanassiou, F Crivello, O Etard, N Delcroix, B Mazoyer, and M Joliot. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage*, 15, 2002.

[159] Pál Vakli, Regina J Deák-Meszlényi, Tibor Auer, and Zoltán Vidnyánszky. Predicting body mass index from structural mri brain images using a deep convolutional neural network. *Frontiers in Neuroinformatics*, 14:10, 2020.

[160] David C Van Essen, Kamil Ugurbil, Edward Auerbach, Deanna Barch, T E J Behrens, Richard Bucholz, Acer Chang, Liyong Chen, Maurizio Corbetta, Sandra W Curtiss, and Others. The Human Connectome Project: a data acquisition perspective. *Neuroimage*, 62(4):2222–2231, 2012.

[161] Annegreet van Opbroek, Fedde van der Lijn, and Marleen de Bruijne. Automated brain-tissue segmentation by multi-feature SVM classification. *The MIDAS Journal*, 2013.

[162] A Varentsova, S Zhang, and K Arfanakis. Development of a high angular resolution diffusion imaging human brain template. *Neuroimage*, 91, 2014.

[163] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2517–2526, 2019.

[164] Christian Wachinger, Martin Reuter, and Tassilo Klein. DeepNAT: Deep convolutional neural network for segmenting neuroanatomy. *NeuroImage*, 170:434–445, apr 2018.

[165] S Wakana, A Caprihan, M M Panzenboeck, J H Fallon, M Perry, R L Gollub, K Hua, J Zhang, H Jiang, P Dubey, A Blitz, P van Zijl, and S Mori. Reproducibility of quantitative tractography methods applied to cerebral white matter. *Neuroimage*, 36, 2007.

[166] Hongzhi Wang, Jung W Suh, Sandhitsu R Das, John B Pluta, Caryne Craige, and Paul A Yushkevich. Multi-atlas segmentation with joint label fusion. *IEEE transactions on pattern analysis and machine intelligence*, 35(3):611–623, 2012.

[167] Yuli Wang, Peiyu Duan, Zhangxing Bian, Anqi Feng, and Yuan Xue. Efficient annotation for medical image analysis: A one-pass selective annotation approach. *arXiv preprint arXiv:2308.13649*, 2023.

[168] Shaun Warrington, Katherine L Bryant, Alexandr A Khrapitchev, Jerome Sallet, Marina Charquero-Ballester, Gwenaëlle Douaud, Saad Jbabdi, Rogier B Mars, and Stamatios N Sotiropoulos. Xtract-standardised protocols for automated tractography in the human and macaque brain. *Neuroimage*, 217:116923, 2020.

[169] Jakob Wasserthal, Peter Neher, and Klaus H. Maier-Hein. TractSeg - Fast and accurate white matter tract segmentation. *NeuroImage*, 2018.

[170] Jakob Wasserthal, Peter F Neher, Dusan Hirjak, and Klaus H Maier-Hein. Combined tract segmentation and orientation mapping for bundle-specific tractography. *Medical image analysis*, 58:101559, 2019.

[171] Frank Wilcoxon. Individual comparisons by ranking methods. In *Breakthroughs in statistics*, pages 196–202. Springer, 1992.

[172] Andreas Wimmer, Grzegorz Soza, and Joachim Hornegger. A generic probabilistic active shape model for organ segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 26–33. Springer, 2009.

[173] R Wirth, C Smoliner, CC Sieber, and D Volkert. Cognitive function is associated with body composition and nutritional risk cognitive function is associated with body composition and nutritional risk of geriatric patients. *The journal of nutrition, health & aging*, 15:706–710, 2011.

[174] Qi Yang, Colin B Hansen, Leon Y Cai, Francois Rheault, Ho Hin Lee, Shunxing Bao, Bramsh Qamar Chandio, Owen Williams, Susan M Resnick, Eleftherios Garyfallidis, et al. Learning white matter subject-specific segmentation from structural mri. *Medical Physics*, 49(4):2502–2513, 2022.

[175] Qi Yang, Xin Yu, Ho Hin Lee, Leon Y Cai, Kaiwen Xu, Shunxing Bao, Yuankai Huo, Ann Zenobia Moore, Sokratis Makrogiannis, Luigi Ferrucci, et al. Single slice thigh ct muscle group segmentation with domain adaptation and self-training. *Journal of Medical Imaging*, 10(4):044001–044001, 2023.

[176] Qi Yang, Xin Yu, Ho Hin Lee, Yucheng Tang, Shunxing Bao, Kristofer S Gravenstein, Ann Zenobia Moore, Sokratis Makrogiannis, Luigi Ferrucci, and Bennett A Landman. Label efficient segmentation of single slice thigh ct with two-stage pseudo labels. *Journal of Medical Imaging*, 9(5):052405–052405, 2022.

[177] Qi Yang, Xin Yu, Ho Hin Lee, Yucheng Tang, Shunxing Bao, Kristofer S Gravenstein, Ann Zenobia Moore, Sokratis Makrogiannis, Luigi Ferrucci, and Bennett A Landman. Label efficient segmentation of single slice thigh CT with two-stage pseudo labels. *Journal of Medical Imaging*, 9(5):52405, 2022.

[178] Qi Yang, Xin Yu, Ho Hin Lee, Yucheng Tang, Shunxing Bao, Kristofer S Gravenstein, Ann Zenobia Moore, Sokratis Makrogiannis, Luigi Ferrucci, and Bennett A Landman. Quantification of muscle, bones, and fat on single slice thigh ct. In *Medical Imaging 2022: Image Processing*, volume 12032, pages 422–429. SPIE, 2022.

[179] Jason D. Yeatman, Robert F. Dougherty, Nathaniel J. Myall, Brian A. Wandell, and Heidi M. Feldman. Tract Profiles of White Matter Properties: Automating Fiber-Tract Quantification. *PLoS ONE*, 7(11), nov 2012.

[180] F C Yeh, S Panesar, D Fernandes, A Meola, M Yoshino, J C Fernandez-Miranda, J M Vettel, and T Verstynen. Population-averaged atlas of the macroscale human structural connectome and its network topology. *Neuroimage*, 178, 2018.

[181] F C Yeh and W Y Tseng. NTU-90: A high angular resolution brain atlas constructed by q-space diffeomorphic reconstruction. *Neuroimage*, 58, 2011.

[182] Anastasia Yendiki, Patricia Panneck, Priti Srinivasan, Allison Stevens, Lilla Zöllei, Jean Augustinack, Ruopeng Wang, David Salat, Stefan Ehrlich, Tim Behrens, Saad Jbabdi, Randy Gollub, and Bruce Fischl. Automated probabilistic reconstruction of white-matter pathways in health and disease using an atlas of the underlying anatomy. *Frontiers in Neuroinformatics*, 5, oct 2011.

[183] B T Yeo, F M Krienen, J Sepulcre, M R Sabuncu, D Lashkari, and M Hollinshead. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *J Neurophysiol*, 106, 2011.

[184] Juha Yla-Jaaski and O Kubler. Segmentation and analysis of 3D volume images. In *9th International Conference on Pattern Recognition*, pages 951–952. IEEE Computer Society, 1988.

[185] Futoshi Yokota, Yoshito Otake, Masaki Takao, Takeshi Ogawa, Toshiyuki Okada, Nobuhiko Sugano, and Yoshinobu Sato. Automated muscle segmentation from CT images of the hip and thigh using a hierarchical multi-atlas method. *International Journal of Computer Assisted Radiology and Surgery*, 13(7):977–986, jul 2018.

[186] Fei Yu, Jie Zhao, Yanjun Gong, Zhi Wang, Yuxi Li, Fan Yang, Bin Dong, Quanzheng Li, and Li Zhang. Annotation-free cardiac vessel segmentation via knowledge transfer from retinal images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 714–722. Springer, 2019.

[187] Xin Yu, Yucheng Tang, Qi Yang, Ho Hin Lee, Shunxing Bao, Yuankai Huo, and Bennett A Landman. Enhancing hierarchical transformers for whole brain segmentation with intracranial measurements integration. *arXiv preprint arXiv:2309.04071*, 2023.

[188] Xin Yu, Yucheng Tang, Qi Yang, Ho Hin Lee, Shunxing Bao, Ann Zenobia Moore, Luigi Ferrucci, and Bennett A Landman. Accelerating 2d abdominal organ segmentation with active learning. In *Medical Imaging 2022: Image Processing*, volume 12032, pages 893–899. SPIE, 2022.

[189] Xin Yu, Yucheng Tang, Qi Yang, Ho Hin Lee, Riqiang Gao, Shunxing Bao, Ann Zenobia Moore, Luigi Ferrucci, and Bennett A Landman. Longitudinal variability analysis on low-dose abdominal ct with deep learning-based segmentation. *arXiv preprint arXiv:2209.14217*, 2022.

[190] Xin Yu, Qi Yang, Yucheng Tang, Riqiang Gao, Shunxing Bao, Leon Y Cai, Ho Hin Lee, Yuankai Huo, Ann Zenobia Moore, Luigi Ferrucci, et al. Reducing positional variance in cross-sectional abdominal ct slices with deep conditional generative models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 202–212. Springer, 2022.

[191] Xin Yu, Qi Yang, Yucheng Tang, Riqiang Gao, Shunxing Bao, Leon Y Cai, Ho Hin Lee, Yuankai Huo, Ann Zenobia Moore, Luigi Ferrucci, et al. Deep conditional generative models for longitudinal single-slice abdominal computed tomography harmonization. *arXiv preprint arXiv:2309.09392*, 2023.

[192] Xin Yu, Qi Yang, Yinchi Zhou, Leon Y Cai, Riqiang Gao, Ho Hin Lee, Thomas Li, Shunxing Bao, Zhoubing Xu, Thomas A Lasko, et al. Unest: local spatial representation learning with hierarchical transformer for efficient medical segmentation. *Medical Image Analysis*, page 102939, 2023.

[193] Nan Zhang, Su Ruan, Stéphane Lebonvallet, Qingmin Liao, and Yuemin Zhu. Kernel feature selection to fuse multi-spectral MRI images for brain tumor segmentation. *Computer Vision and Image Understanding*, 115(2):256–269, 2011.

[194] Shengwei Zhang and Konstantinos Arfanakis. Evaluation of standardized and study-specific diffusion tensor imaging templates of the adult human brain: Template characteristics, spatial normalization accuracy, and detection of small inter-group FA differences. *NeuroImage*, 172:40–50, may 2018.

[195] Yinchi Zhou, Ho Hin Lee, Yucheng Tang, Xin Yu, Qi Yang, Shunxing Bao, Jeffrey M Spraggins, Yuankai Huo, and Bennett A Landman. Multi-contrast computed tomography atlas of healthy pancreas. *arXiv preprint arXiv:2306.01853*, 2023.

[196] Zongwei Zhou, Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. UNet++: A Nested U-Net Architecture for Medical Image Segmentation. Technical report.

[197] Zongwei Zhou, Vatsal Sodha, Jiaxuan Pang, Michael B Gotway, and Jianming Liang. Models genesis. *Medical image analysis*, 67:101840, 2021.

[198] Jiayi Zhu, Bart Bolsterlee, Brian V.Y. Chow, Chengxue Cai, Robert D. Herbert, Yang Song, and Erik Meijering. Deep learning methods for automatic segmentation of lower leg muscles and bones from MRI scans of children with and without cerebral palsy. *NMR in Biomedicine*, 2021.

[199] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

Appendix A

Copyright from Publishers

A.1 Copyright from Elsvier

Author rights in Elsvier's proprietary journals include re-use portions, excerpts, and their own figures or tables in other works. Our Ch2 is under Elsvier publisher.

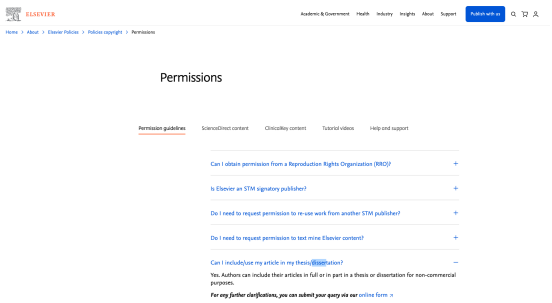
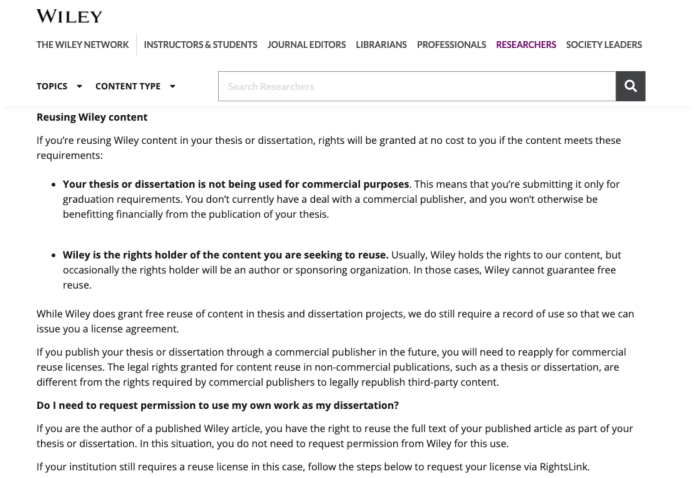


Figure A.1: Copyright from Elsvier

A.2 Copyright from Wiley

Author rights in Wiley's proprietary journals include re-use portions, excerpts, and their own figures or tables in other works. Our Medical Physics (Ch 3) is under Wiley's publisher.



WILEY

THE WILEY NETWORK | INSTRUCTORS & STUDENTS | JOURNAL EDITORS | LIBRARIANS | PROFESSIONALS | RESEARCHERS | SOCIETY LEADERS

TOPICS | CONTENT TYPE | Search Researchers

Reusing Wiley content

If you're reusing Wiley content in your thesis or dissertation, rights will be granted at no cost to you if the content meets these requirements:

- **Your thesis or dissertation is not being used for commercial purposes**. This means that you're submitting it only for graduation requirements. You don't currently have a deal with a commercial publisher, and you won't otherwise be benefitting financially from the publication of your thesis.
- **Wiley is the rights holder of the content you are seeking to reuse.** Usually, Wiley holds the rights to our content, but occasionally the rights holder will be an author or sponsoring organization. In those cases, Wiley cannot guarantee free reuse.

While Wiley does grant free reuse of content in thesis and dissertation projects, we do still require a record of use so that we can issue you a license agreement.

If you publish your thesis or dissertation through a commercial publisher in the future, you will need to reapply for commercial reuse licenses. The legal rights granted for content reuse in non-commercial publications, such as a thesis or dissertation, are different from the rights required by commercial publishers to legally republish third-party content.

Do I need to request permission to use my own work as my dissertation?

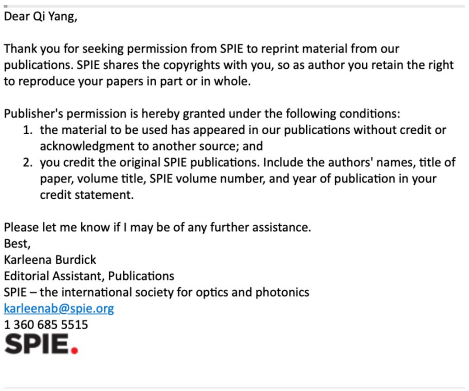
If you are the author of a published Wiley article, you have the right to reuse the full text of your published article as part of your thesis or dissertation. In this situation, you do not need to request permission from Wiley for this use.

If your institution still requires a reuse license in this case, follow the steps below to request your license via RightsLink.

Figure A.2: Copyright from Wiley

A.3 Copyright from SPIE

Author rights in SPIE's proprietary journals include re-use portions, excerpts, and their own figures or tables in other works. Our Ch 4 and Ch 5 are under SPIE.



Dear Qi Yang,

Thank you for seeking permission from SPIE to reprint material from our publications. SPIE shares the copyrights with you, so as author you retain the right to reproduce your papers in part or in whole.

Publisher's permission is hereby granted under the following conditions:

1. the material to be used has appeared in our publications without credit or acknowledgment to another source; and
2. you credit the original SPIE publications. Include the authors' names, title of paper, volume title, SPIE volume number, and year of publication in your credit statement.

Please let me know if I may be of any further assistance.
Best,
Karleena Burdick
Editorial Assistant, Publications
SPIE – the international society for optics and photonics
karleenab@spie.org
1 360 685 5515
SPIE.

Figure A.3: Copyright from SPIE