Deep Learning-Based Automatic Segmentation of Skeletal Muscles and Generalisation Study

Zhicheng Lin

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

School of Electrical and Electronic Engineering
Faculty of Engineering
The University of Sheffield

2025

Abstract

Musculoskeletal disorders affect a significant portion of the population, with one in four people in the UK currently suffering from such conditions, which impact both individuals' work and social lives. While the mechanisms underpinning skeletal disorders are well-understood, a major challenge in advancing the understanding of muscle-related conditions lies in the difficulty of accurately measuring muscle tissue's physiological status.

Muscle disorders vary significantly in terms of their causes, affected muscles, progression rates, and treatment strategies. Furthermore, individuals with the same muscle disorder often exhibit different responses to the condition, highlighting the need for subject-specific, quantitative characterisation of muscle tissue in vivo. This could significantly enhance current diagnosis and treatment strategies and provide a more informed approach to evaluating the efficacy of new treatments in clinical trials. However, despite its potential, quantitative muscle tissue analysis has not yet been fully integrated into clinical practice. As an indispensable part of this quantitative analysis, manual segmentation of muscle images is labour-intensive, time-consuming, prone to inter- and intra-operator variability, highlighting the need for automatic segmentation methods.

The aim of this thesis was to develop, test, and analyse methods for deep learning based automatic muscle segmentation from medical imaging data. This work presents three distinct methods designed to address the limitations of the current method for muscle automatic segmentation in MR images. The outcome provides a comprehensive overview of both existing and novel methods for muscle segmentation pipeline and analysis from medical images.

The methods discussed in this thesis offer valuable insights for future research, providing a foundation for the quantitative study of muscle segmentation. By adopting the best-suited deep learning models or pre-/post-processing pipeline from this work, future studies can improve the understanding and treatment of muscle conditions. Ultimately, this research aims to promote the clinical adoption of computational tools for muscle disorder characterisation, enhancing diagnosis, treatment planning, and patient monitoring.

Acknowledgements

The journey of completing my PhD studies and thesis has been one of growth, learning, and immense gratitude. This accomplishment would not have been possible without the support and encouragement of many people, to whom I would like to express my heartfelt thanks.

First and foremost, I wish to acknowledge my supervisory team: Dr. Lingzhong Guo and Prof. Enrico Dall'Ara. I am deeply grateful to Lingzhong for guiding me through these years of study. Your profound academic expertise and unwavering kindness have been invaluable. Whenever I faced setbacks or felt lost, you were always there to support and encourage me, making this PhD journey not only rewarding but also incredibly enjoyable. To Enrico, I am sincerely thankful for your rigorous academic attitude and meticulous guidance. Your high standards and constructive feedback have driven me to continuously improve, helping me achieve significant results in my research.

I would also like to extend my gratitude to Prof. Zi-Qiang Lang for the guidance and recognition you have provided in the field of artificial intelligence. Collaborating with you has broadened my academic perspective and deepened my understanding of the subject, significantly expanding my horizons.

My heartfelt thanks go to my friends, without whose support I might never have embarked on this PhD journey. To those friends (you know who you are), thank you for instilling in me the confidence to persevere and for helping me stay optimistic during challenging times. On a brighter note, I am equally grateful for your unwavering enthusiasm in celebrating the successes I have achieved along the way.

I am deeply indebted to my family: my parents, Lan Ma and Qiang Lin; my grandparents, Guangshu Chen and Qi Lin; my grandmother, Suqin Liao; and my uncle, Ning Ma. You have always been my warmest harbor, providing the love, understanding, and support that enabled me to fully commit to my PhD journey. Your encouragement gave me the strength to face challenges with resilience and optimism. I also wish to express my profound gratitude to my late grandfather, Prof. Baohong Ma. Although it has been nine years since you left us, not a day goes by without me thinking of you. Your dedication to scientific research and your unwavering spirit continues to inspire me. I hope you are watching over me, and I strive to make you proud. I miss you deeply.

I love you all!

Publications & contributions

Journal entries

In print

1) **Lin Z**, Henson WH, Dowling L, Walsh J, Dall'Ara E, Guo L. Automatic segmentation of skeletal muscles from MR images using modified U-Net and a novel data augmentation approach. Front Bioeng Biotechnol. 2024 Feb. doi: 10.3389/fbioe.2024.1355735.
2) Henson WH, Li X, **Lin Z**, Guo L, Mazzá C, Dall'Ara E. Automatic segmentation of lower limb muscles from MR images of post-menopausal women based on deep learning and data augmentation. PLoS One. 2024 Apr. doi: 10.1371/journal.pone.0299099.
3) **Lin Z**, Dall'Ara E, Guo L. A novel mean shape based post-processing method for enhancing deep learning lower-limb muscle segmentation accuracy. PLoS One. 2024 Oct. doi: 10.1371/journal.pone.0308664.

In preparation

1) **Lin Z**, Lang Z, Guo L, Walker D, Wang M., Matella M, Murdoch C. (2025) Deep Learning-Based Electrical Impedance Spectroscopy Analysis for Malignant and Potentially Malignant Oral Disorder Detection. Scientific Report. Under review.
2) **Lin Z**, Kanz H, Dall'Ara E, Guo L. A Generalisation Study in Deep Learning-based Segmentation of Lower-limb Muscles Across Different Populations. Scientific Report. Under review.

Conference abstracts

Oral presentation

1) **Zhicheng L.**, W. H. Henson, C. Mazzà, E. Dall'Ara, Guo L. "Deep learning-based automatic segmentation of skeletal muscle", ESB 2023, Maastricht, the Netherlands.

Poster presentation

1) M Matella, K Hunter, Zi-Qiang Lang, **Zhicheng Lin**, D C Walker. "Virtual Tissue Constructs to Assess the Potential of Electrical Impedance Spectroscopy as a Method for Tissue Identification and Pathology Diagnosis", VPH 2024, Stuttgart, Germany.

Declaration

I, Zhicheng Lin, confirm that the work presented in this thesis is a product of my studies. Where information is portrayed to the reader that was derived from other sources, I confirm that this is indicated within the thesis.

Content

Chapter 1 General introduction and thesis overview

1.1 Thesis overview

Muscle image segmentation in medical imaging plays a crucial role in diagnosing and monitoring musculoskeletal diseases, such as sarcopenia [1], cerebral palsy [2], and muscle atrophy [3], [4]. These diseases often lead to abnormal muscle growth, loss of muscle mass, or changes in muscle function, making accurate segmentation of muscle tissues from biomedical images essential for assessing disease progression and developing appropriate treatments[5]. While manual segmentation remains the gold standard for muscle tissue delineation, it is labour-intensive, time-consuming, prone to inter- and intra-operator variability [6], [7], and not practical for large-scale clinical application. This has led to an increasing interest in deep learning (DL)-based automatic segmentation methods, which offer potential advantages in terms of segmentation accuracy, reproducibility, and efficiency.

However, one major challenge with DL models is their limited generalization capability across different populations due to factors like muscle morphology difference of difference population or limited dataset [8], which hampers their widespread clinical applicability. This thesis addresses several key issues in muscle segmentation steps, specifically focusing on data augmentation (pre-processing), advanced model structure modification, model generalization performance evaluation, and post-processing methods to improve the robustness, accuracy, and clinical utility of DL-based muscle segmentation models.

A significant challenge in training DL models for medical image segmentation is the lack of sufficient labelled data [8], especially in specialized cohorts such as children or elderly populations [9] due to the difficulties in data collection and the required labour intensive labelling operations. To address this issue, this thesis introduces a novel data augmentation pipeline based on the Statistical Shape Model (SSM) [10], [11] technique. The SSM-based augmentation uses prior anatomical knowledge to generate diverse, anatomically realistic training samples. This method allows the model to learn a broader range of muscle shapes and variations, particularly important when training on smaller datasets. By enhancing the model's feature learning ability to realistic anatomical variations, this approach significantly improves the robustness of DL models.

While data augmentation can improve the diversity of the training data, the model architecture itself also affects the final performance in segmentation accuracy. This thesis proposed a modified U-Net architecture, termed Attention-Feature-Fusion U-Net (AFFU) to further enhance the segmentation accuracy of muscle structures from Magnetic Resonance

Imaging (MRI) data. This model modification integrates attention mechanisms and feature fusion strategies to better capture and preserve the relevant features of muscle tissue. The introduction of attention gates helps the model focus on important anatomical regions [12], while feature fusion [13], [14] improves the integration of multi-scale information, leading to more precise muscle delineation. This structural modification addresses some of the limitations of traditional U-Net models, particularly in cases where muscle boundaries are less clearly defined, such as in populations with smaller or less well-defined boundaries among muscles.

Models' ability to generalize across diverse cohorts is vital in real world application [15]. The effectiveness of a segmentation model trained on a single cohort (e.g., elderly women) may not translate to other cohorts (e.g., children or young adults) due to physiological and anatomical differences in muscle morphology [16]. This thesis focuses on evaluating the generalization potential of the proposed AFFU model and compares its performance to that of other baseline models, such as U-Net, UNet++, and Attention-UNet. This evaluation uses datasets including three cohorts (children, young adults, and elderly women), to assess the model's robustness in handling variability in muscle size, structure, and tissue contrast. The results reveal that models trained on a single cohort often perform sub-optimally when applied to other cohorts, highlighting the need for improved generalization methods in muscle image segmentation and underscores the importance of evaluating model performance across multiple populations.

For an entire segmentation pipeline, post-processing operation is a final vital step which cannot be ignored. Despite some methods like hole-filling [17], binary closing [2] and Conditional Random Field (CRF) [18] have been used in applications, the medical image specific post-processing method is still not mature. In the final objective of this thesis, a post-processing method based on mean shape and SSM is introduced to further enhance the accuracy of muscle segmentation results. This method integrates anatomical knowledge during the post-processing phase to refine the segmentation boundaries and reduce false positives. The incorporation of mean shapes, which represent typical muscle structures, guides the model to produce more anatomically accurate results, even in cases where the initial segmentation might be noisy or incomplete. The performance of this novel post-processing technique is compared to traditional methods, such as Conditional Random Fields (CRF) and 2D post-processing methods, which focus on refining the edges and removing small segmentation errors. It demonstrates superior performance in accurately delineating

muscle boundaries, especially in challenging cases with low contrast or unclear muscle boundaries, thus improving the overall segmentation accuracy.

This thesis presents a comprehensive approach to improving the accuracy and generalization of DL models for muscle automatic segmentation in medical imaging. The findings highlight the importance of incorporating anatomical knowledge at various stages of the segmentation pipeline—from data augmentation to model design and post-processing—to improve the clinical relevance and applicability of muscle segmentation models. These advancements offer a step forward in creating more robust, accurate, and generalizable models for muscle segmentation, with potential applications in diverse clinical settings and patient populations.

1.2 Thesis structure

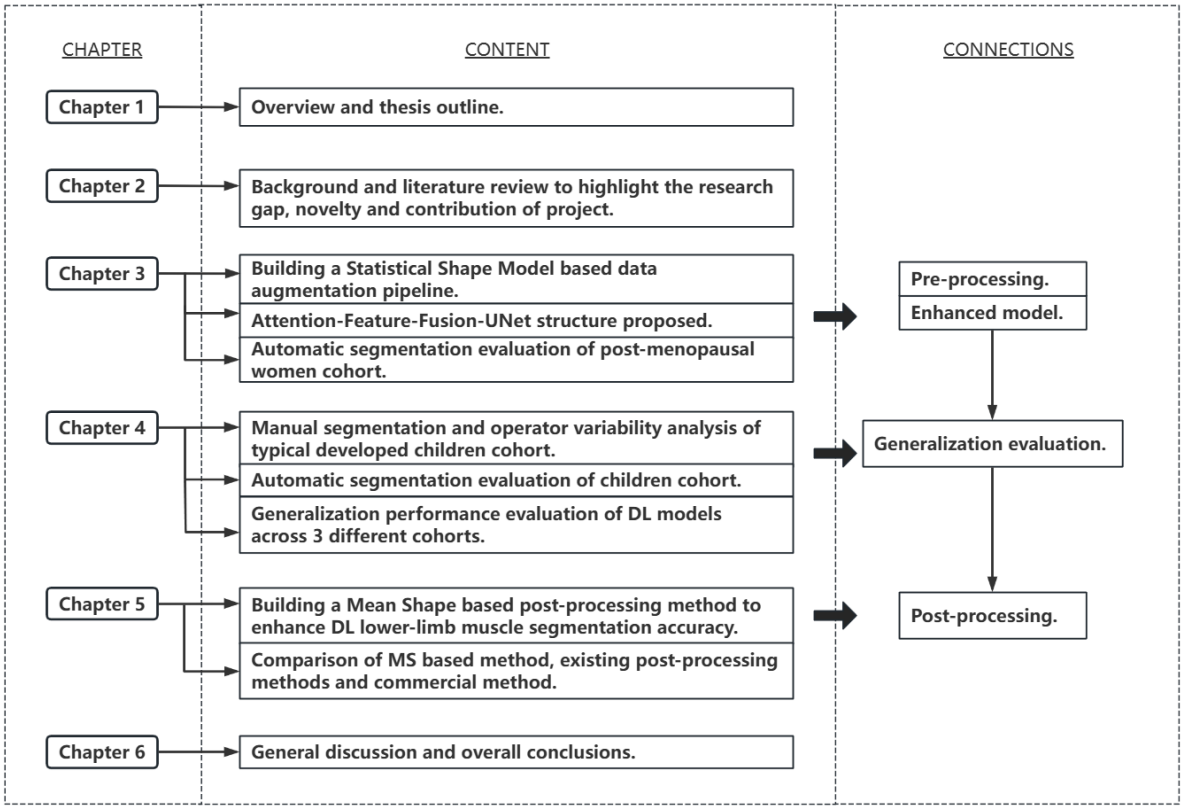


Figure 1.1 Schematic representation of thesis structure highlighting the main contents of each chapter and main connections.

The thesis consists of 6 chapters, as shown in Figure 1.1. Chapter 1 gives an overview of the thesis and motivation of each main chapter. Chapter 2 presents the literature review, technical background knowledge. Chapters 3,4,5 contain the methods, results and critical appraisal of proposed model, model's generalization comparison and novel pipeline in pre-/post-processing steps. The final chapter, Chapter 6, concludes the findings, novelty, contributions of this thesis, giving an overall conclusion and plan for future work.

Chapter 2 Background and literature review

2.1 The lower-limb musculoskeletal system

2.1.1 Anatomy of skeletal muscle

The skeletal muscles of the human body can be divided into three major parts: head and neck muscles, trunk muscles and limb muscles. The trunk muscles include the pectoralis, diaphragm, abdominal and back muscles, and the limb muscles are composed of the upper and lower limb muscles. There are more than 600 muscles in the human body, and skeletal muscle accounts for a large proportion of the body weight, whose total mass is about 40% of the total body mass. Each skeletal muscle consists of a tendon and a muscle belly containing sensory nerve endings. Under normal circumstances, the two ends of each muscle will be attached to two different bones. After being stimulated by nerve signals, they will contract. The two attached bones will be displaced around the joint to achieve relative displacement of the human body, realising mobility function.

Skeletal muscle is organized hierarchically (Figure 2.1). Individual muscle fibers, which are multinucleated and elongated cells, are surrounded by a connective tissue layer known as the endomysium. The fibres are collected into larger bunches known as fascicles, and these fascicles are in turn grouped into muscles as shown below. The fibres themselves are made up of smaller threads – known as myofibrils, which contain the proteins that cause contraction. Groups of muscle fibers form bundles called fascicles, which are enclosed by the perimysium. Multiple fascicles are further grouped together to form the entire muscle, which is wrapped in a dense connective tissue layer called the epimysium. This layered architecture supports both the mechanical function and the innervation/vascularization of skeletal muscle, making it essential for force transmission and physiological regulation. Skeletal muscles are activated by nerve impulses transmitted from motor neurons, which initiate contraction at specialized sites called neuromuscular junctions.

The primary functions of skeletal muscle are: 1) Generate energy to adjust the body temperature. 2) Convert fatty acids, glucose, and amino acids to provide energy for exercise and other cellular processes. 3) In daily life, the most crucial role of skeletal muscles is maintaining posture and body movement by connecting bones and joints. The skeletal muscles of the lower limbs as shown in Figure 2.2 play a vital role in the body's mobility, which is also the research object of this project.

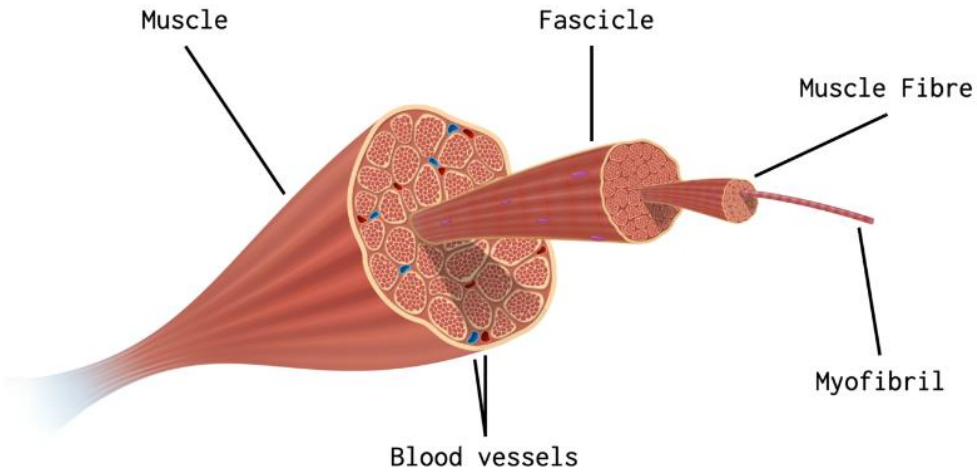


Figure 2.1 Muscle structure.

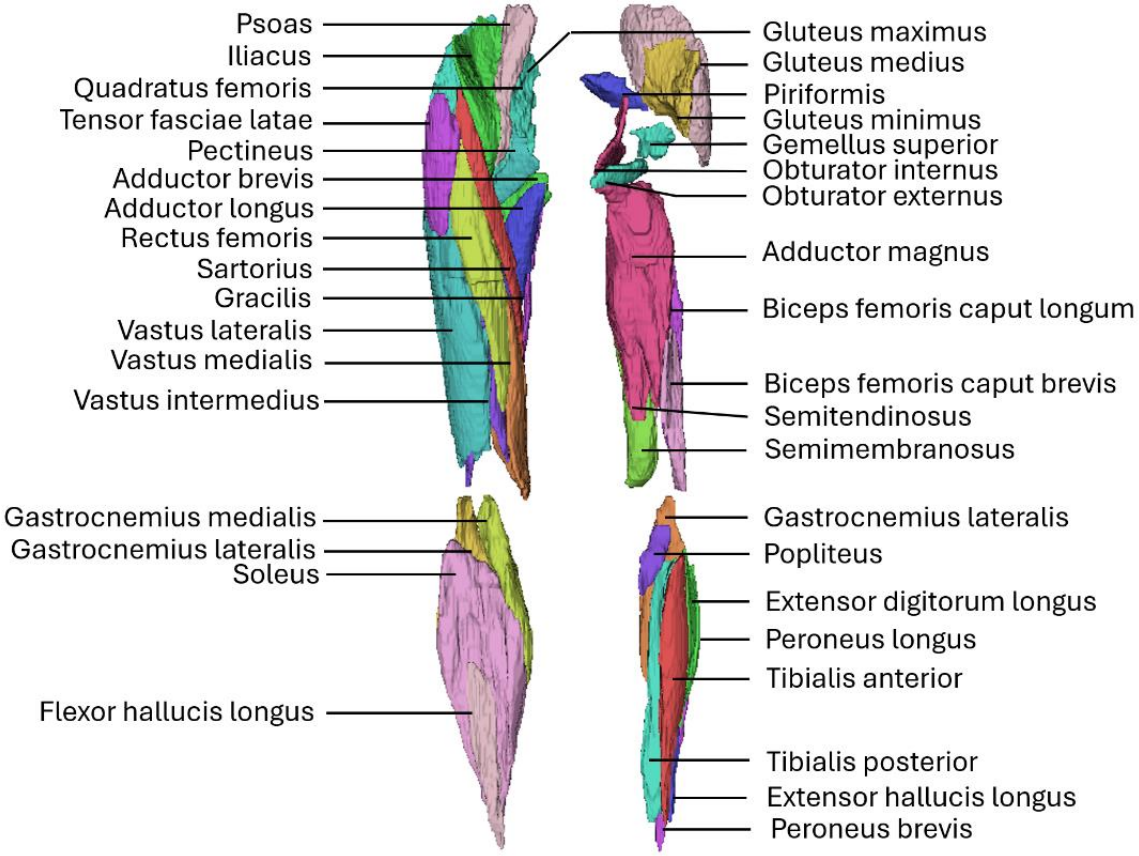


Figure 2.2 Anatomical model of lower limb muscles shown in anterior view.

The lower limb skeletal muscles are combined with pelvic girdle, thigh, calf and foot muscles. Lower limb muscles are stronger than upper ones, which are related to maintaining an upright posture, supporting weight, walking, running, and jumping.

The pelvic girdle muscles are mainly distributed around the hip joint, including the iliacus, gluteus maximus, gluteus medius and gluteus minimus. Their functions are primarily to bend and rotate the thighs and maintain the gait of the human body and the upright balance.

The thigh muscle shown in Figure 2.3 is the largest in the human body, distributed around the femur and divided into anterolateral, posterior, and medial groups. The anterolateral group includes the quadriceps, sartorius, and tensor fascia latae. The quadriceps comprises four parts: rectus femoris, vastus medialis, vastus lateralis and vastus medialis. Its function is to maintain the bending and straightening of the calf and thigh. The sartorius and tensor fascia latae is responsible for the abduction and internal rotation of the thigh and calf. The posterior group consists of the biceps femoris, semitendinosus, and semimembranosus. The biceps femoris is a fusiform muscle, and the semitendinosus and semimembranosus are pinnate muscles stretching the thigh and externally rotating the calf. The medial group is mainly the adductor Magnus, which undertakes the work of adducting and externally rotating the thigh.

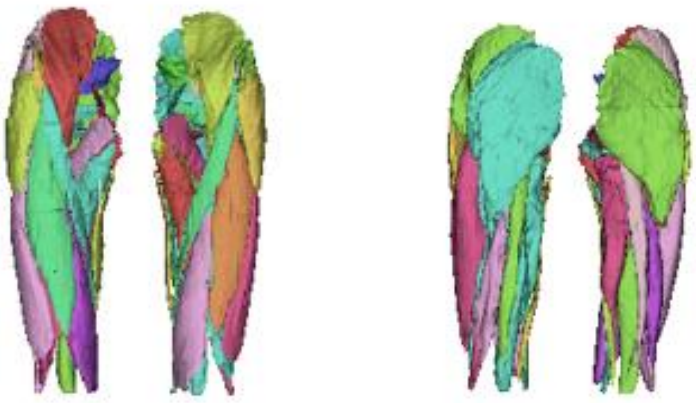


Figure 2.3 Anterior and posterior view of thigh muscles.

The calf muscles shown in Figure 2.4 are also divided into anterior, lateral and posterior groups. The anterior group is primarily the anterior tibialis muscle on the lower leg's anterolateral side. The posterior group consists of the gastrocnemius and soleus muscles, which maintain the essential upright function of the human body. Foot muscles are mainly located on the soles of the feet and are divided into dorsal and plantar muscles. They are used primarily to move the toes, maintain the foot arch, and affect posture in standing, walking and other human movements.

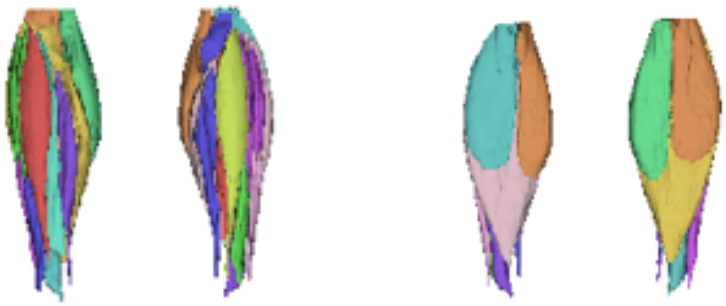


Figure 2.4 Anterior (left) and posterior (right) view of calf muscles.

2.1.2 Maintenance and treatment

Skeletal muscles account for about 40% of human body mass, and with ageing, muscle declines gradually. After middle age, the human body will lose about 1% of muscle every year, when it reaches a 30% reduction, it will affect normal muscle function, and when it comes to 40%, it will affect health [19]. Human ageing is divided into primary and secondary ageing. Natural ageing causes the body's structural function to decline, which is primary ageing. A further decline in function due to factors such as acquired diseases, or an unwholesome lifestyle is considered secondary ageing. Studies have shown that the degeneration of skeletal muscle structure and function in old age is one of the significant causes of secondary ageing [20]. According to a survey of 280,000 people in Europe [21], less than 17% of the people who regularly perform resistance training and most older adults have neglected muscle strength training for a long time, which also leads to the faster rate of muscle loss in the elderly than the young. In addition to providing average power, skeletal

muscle also undertakes functions such as maintaining human metabolism and bone health [22].

To maintain muscle health, besides regular muscle strength resistance training, researchers can also use CT, MRI, and other medical imaging methods to measure and diagnose muscle shape, quality, and related disease markers. Among them, skeletal muscle MRI is the most accurate and precise imaging technology for detecting muscle lesions. MRI technology has the advantages of the high soft-tissue resolution, repeatability, and no radiation. Recently, it has been increasingly used to diagnose and evaluate muscle diseases [23]. Skeletal muscle MRI can also accurately assess skeletal muscle morphological, volumetric, and structural changes, revealing muscle adipose, oedema, atrophy, and metabolic changes.

According to aetiology, muscle diseases are divided into two categories: acquired and hereditary. The former mainly includes inflammatory and endocrine myopathy, and the latter especially has muscular dystrophy, distal myopathy, congenital myopathy (CMs), metabolic myopathy, etc. The treatment and prognosis of these myopathies are very different; therefore, an early and accurate diagnosis is necessary. Myopathy cause changes in the size or shape of muscles (atrophy, hypertrophy), often seen on MRI slices of the lower limbs shown in Figure 2.5 [24]. The lower limb MRI scan sequence included T1WI (T1Weightedimage) and short-term inversion recovery (STIR). The presence of fat or a fat component in the lesion shows a high signal in T1 on MRI, as most lesions have a low signal in T1, so the nature of the lesion can be initially determined by the high signal in T1. At this point a STIR sequence can be added, where the presence of a high signal on this sequence indicates a fat or fat-containing lesion. Cerebral palsy (CP) is a neurological disorder that usually affects infants; children with cerebral palsy have smaller muscle volume than typically developing children [2]. Accurate muscle size, mass, and other parameters are needed if a comprehensive muscle growth survey is conducted in children with and without CP. The slices provided by MRI scans, as shown in Figure 2.6, can bring researchers a quantitative analysis of changes in children with CP. After slices' manual segmentation, researchers can do a quantitative analysis of ROI.

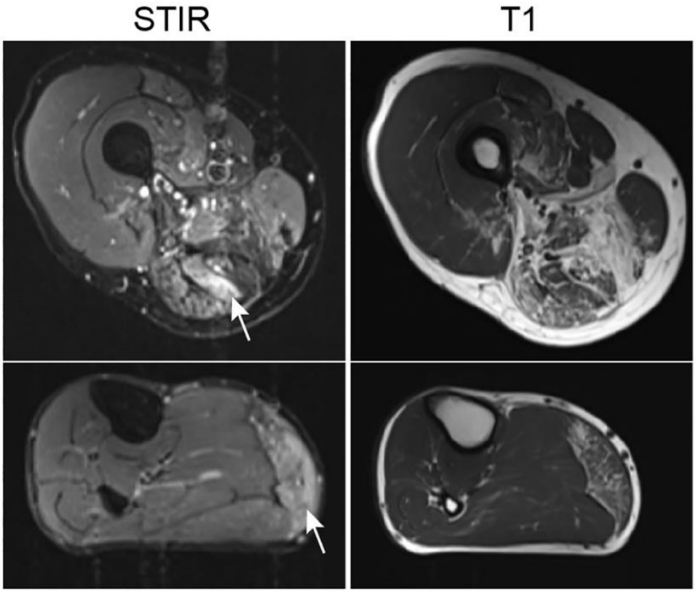


Figure 2.5 MRI reveals STIR lesions (arrows) of the semitendinosus and medial gastrocnemius muscles. T1 images reveal fat replacement.

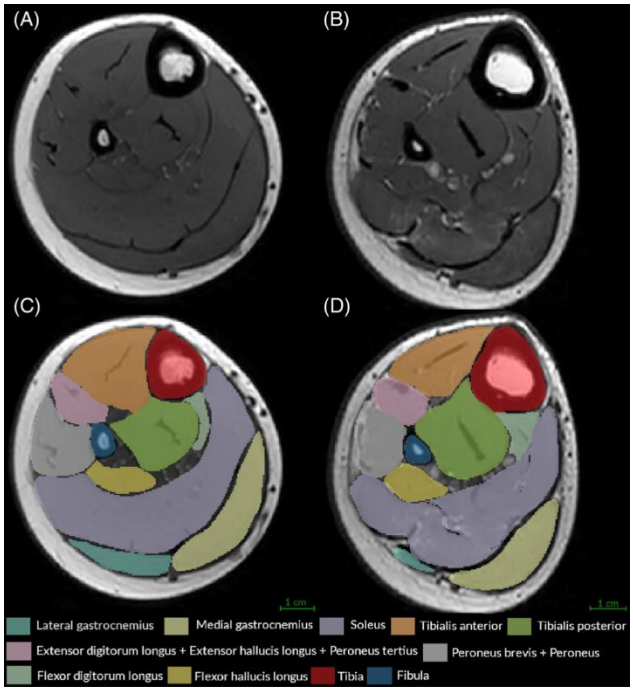


Figure 2.6 MRI scans of normal child (A) and a child with CP(B), and their corresponding manual segmentation labels (C and D, respectively). The MRI scan of normal child differs from that of a child with CP in the shapes, positions, and sizes/volumes of muscles.

2.2 Research on medical image processing

Medical imaging technology enables the non-invasive extraction [25] of tissue or organ shape, structure, and some physiological functions using specialized physical instruments. This technology is fundamental for clinical diagnosis, offering crucial visual information that supports accurate patient assessment. Recent developments in X-ray technology, computed tomography (CT), and magnetic resonance imaging (MRI) have made it possible to capture medical images with clearer textures information, providing clinicians with enhanced detail for more accurate diagnoses and contributing to a reduction in misdiagnosis rates. The discovery of X-rays by German researcher Wilhelm Roentgen in 1895 marked the official start of radiology [26]. Over the past century, the integration of CT scanning and nuclear magnetic resonance has driven substantial progress in medical imaging, advancing diagnostic capabilities and the overall field of medical imaging technology.

In 1973, magnetic resonance imaging (MRI) was introduced by Lauterbur as a bio-magnetic nuclear spin imaging technique [27]. The principle of MRI relies on generating a radio frequency signal within human tissues by applying a magnetic field and radio frequency pulse. By adjusting the intensity and duration of this pulse, the body is made to resonate within the magnetic field, allowing the receiving system to capture and analyze the resonance signals. Unlike X-ray imaging, MRI does not involve ionizing radiation or the use of potentially toxic contrast agents, making it a safer, non-invasive technique. Among the primary imaging methods—MRI, X-ray imaging, CT scanning, and ultrasound—MRI is especially favoured due to its superior soft tissue contrast and absence of radiation, making it particularly useful for studies in infant development, musculoskeletal imaging, and soft tissue differentiation.

Traditionally, medical diagnosis required doctors and researchers to manually identify and annotate regions of interest (ROIs) on MRI or CT slices. However, due to challenges like equipment variation, environmental factors, and imaging noise, manual processing results of these images can be heterogeneous. Additionally, the annual increase in medical imaging volume [28] has made it even more challenging to rely solely on experts like manual ROI segmentation. In response, computer-aided diagnosis (CAD) technology emerged, leveraging computer algorithms to enhance the visualization and analysis of medical images. CAD is now commonly applied for fully or semi-automatic ROI analysis and has become a prominent field in medical engineering research. Medical image automatic segmentation

primarily involves extracting regions of interest (ROI) from medical image slices. As the initial step in computer-aided diagnosis, segmentation is commonly defined as the identification of contours or sets of voxels that constitute the object of interest [29]. Automatic segmentation primarily focuses on organ structures and lesion areas. Segmenting organs and substructures enable the quantitative analysis of clinically relevant parameters, such as volume and shape, which are essential for medical assessment and diagnosis, e.g. liver automatic segmentation for anatomical volumetry analysis [30] and heart segmentation to assist the congenital heart disease treatment [31].

This section provides an overview of medical image automatic segmentation algorithms, including both traditional and deep learning-based segmentation approaches. Given space limitations and the focus of our study, only a few representative traditional methods are highlighted, while the main discussion will centre on deep learning-based automatic segmentation techniques.

2.2.1 Traditional segmentation algorithm

Traditional methods primarily use low-level features which refer to basic visual characteristics that can be extracted directly from pixel data [13] in images to extract ROI. They mainly employ classical mathematical theories, making themselves easy to understand and cost-effective to implement. At the same time, traditional algorithms do not require large amounts of data to train models, making them well-suited for small datasets. The computational complexity of traditional algorithms is relatively low, and running speed is faster, giving them an advantage in applications that require real-time processing. While traditional methods can effectively segment simple images to some extent, their segmentation results are often suboptimal for complex medical images [32] and natural scene images due to the irregular, complicated shapes and rich textures present in these images. In the following sub-chapters, some representative traditional segmentation methods are presented that utilize low-level features, including the pixel grey value segmentation method, edge detection method, and region segmentation method, which will be discussed briefly below.

Pixel grey value segmentation method

This method extracts ROI by using features like pixel intensity and edge pixel variations. The most representative method is the Threshold Segmentation Method (Thresholding) [33]. By setting different thresholds manually, pixels with varying grayscale values in a medical image

can be classified into distinct target regions. It is advantageous for simplicity and speed, making it widely used. However, it also has clear drawbacks. In practice, multiple thresholds often need to be tested to achieve optimal results. Additionally, in real images, like muscle MRI slices in this study, many muscle boundaries are closely connected and unclear. For example, the borders between the semitendinosus and semimembranosus muscles often have similar textures, making it difficult for simple thresholding to accurately segment these muscles. The Otsu method, introduced by Japanese scholar Otsu in 1979 [34], is a dynamic threshold segmentation algorithm that addresses the limitations of basic thresholding by using intra-class variance between the background and target regions. This approach is used in many applications about medical image processing field, such as in the work by Helen et al. [35], where an enhanced Otsu threshold method was used to segment lung tissue in CT images. Despite these improvements, the thresholding-based method still produces relatively rough results with unclear edge textures and is sensitive in noise. Consequently, thresholding techniques like Otsu's are now mainly applied in preprocessing medical images.

Edge detection segmentation method

As an object's most essential and direct feature, the contour or edge is the kernel of the edge detection segmentation method. The edge detection method focuses on identifying changes in intensity using first-order or second-order derivation along object boundaries to outline complete contours. Common edge detectors used to extract edges include the Roberts, Prewitt, Sobel, and Laplacian of Gaussian (LoG) detectors [36]. Each of these methods helps highlight boundaries within an image by detecting intensity changes, making them useful for edge-based segmentation tasks. However, the drawback of edge detection is that the ability to suppress noise is limited. If the target's contour edge pixels are much closer to the background pixels, the results will be poor.

Region segmentation method

Image region segmentation aims to label and identify object regions within an image by finding the set of pixels corresponding to a target area [37]. A key technique is the region-growing method [38], where segmentation begins from a point (seed) that meets certain criteria and expands outward. If adjacent regions share similar characteristics—such as grayscale, colour, or texture—they are merged into the growing segment. The growth stops when no further regions meet the merging conditions. This straightforward approach is used in some medical segmentation tasks; for instance, Adams et al. [38] applied a 3D region-

growing method for semi-automatic segmentation of the brain and heart in MR images. However, this method requires manually selecting starting points, which means segmentation accuracy heavily depends on operator precision, limiting its broader applicability.

In addition to the three popular traditional methods, contour-based segmentation techniques like the Snake Model [39], [40] are also widely used. The Snake Model uses a closed curve to outline the target region by minimizing an energy function, effectively adjusting itself to fit the object's boundaries. However, it also has limitations: it is very sensitive to the initial contour’s position and shape, as well as to the selection of hyper-parameters and can be easily disrupted by noise or edge pixels’ information, leading to low accuracy and reliability of segmentation results.

In general, traditional image segmentation methods can effectively segment simple natural images like shown in Figure 2.7; however, for complex medical images from MRI or CT scan with complicated structures and rich textures, their performance is often unsatisfied.

In Figure 2.7, traditional methods can achieve good results for segmentation of images with simple and few target objects. For example, by using region growing methods, the contour of the animal has been basically outlined. However, for the MRI slice, the traditional method can only differentiate the region of muscle, fat tissue and background, the specific individual muscles cannot be recognized in detail.

Moreover, in complex scenarios, traditional methods require substantial manual modification on specific data cohort and material resources to manually design features, resulting in weak generalization and poor adaptability to new contexts. Given these limitations, deep learning (DL) algorithms may be more suitable for complex images and higher accuracy demands.

The algorithms and some representative deep learning based automatic segmentation methods will be described in next section.

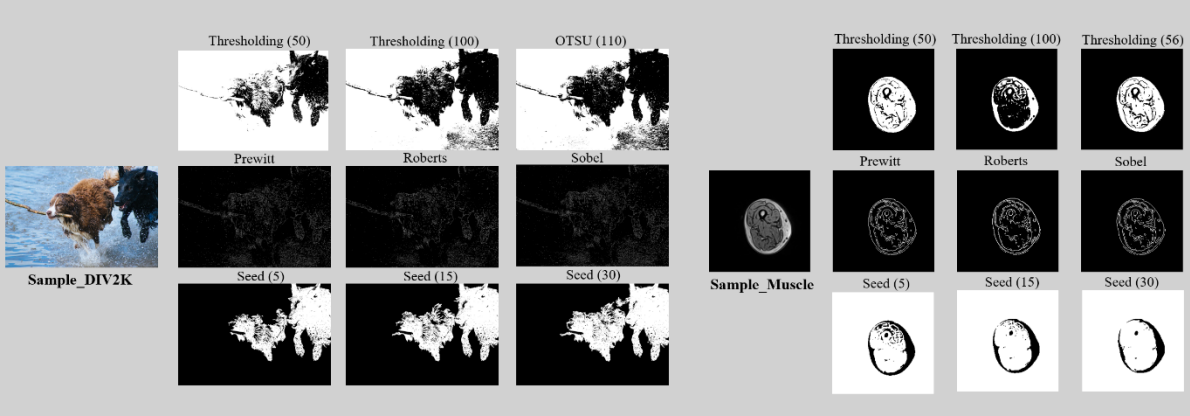


Figure 2.7 Visualization of segmentation based samples of a natural image from public dataset (DIV2K) and a human lower-limb MRI slice. For each case, the threshold method (threshold set to be 50/100 and automatically set using OTSU method), edge detection (three operators were used) and region-growing method (seed set to be 5/15/30) were used respectively.

2.2.2 Automatic medical image segmentation algorithms

The rapid development in medical imaging technology has led to a massive increase volume of medical image data generated daily worldwide, highlighting the need for computer-assisted image analysis in clinical diagnosis and treatment [41]. Medical images often have features like low contrast, complex textures, and unclear boundaries, which bring challenges for traditional segmentation methods. These conventional algorithms are limited by high computational demands and variability in results, making them less competitive as high-performance computing, machine learning (ML), and neural networks (NN) evolve.

With the rise of DL, particularly convolutional neural networks (CNNs)[42], [43], neural network algorithms are increasingly replacing traditional methods. Unlike traditional algorithms that extract only low-level features (e.g., colour, texture, shape), CNN-based deep learning models can capture higher level semantic information which may be invisible, improving segmentation accuracy. As shown in Figure 2.8 in VGG16 [42], as the size of the images continues to decrease, the semantic information extracted by the model becomes more advanced and abstract.

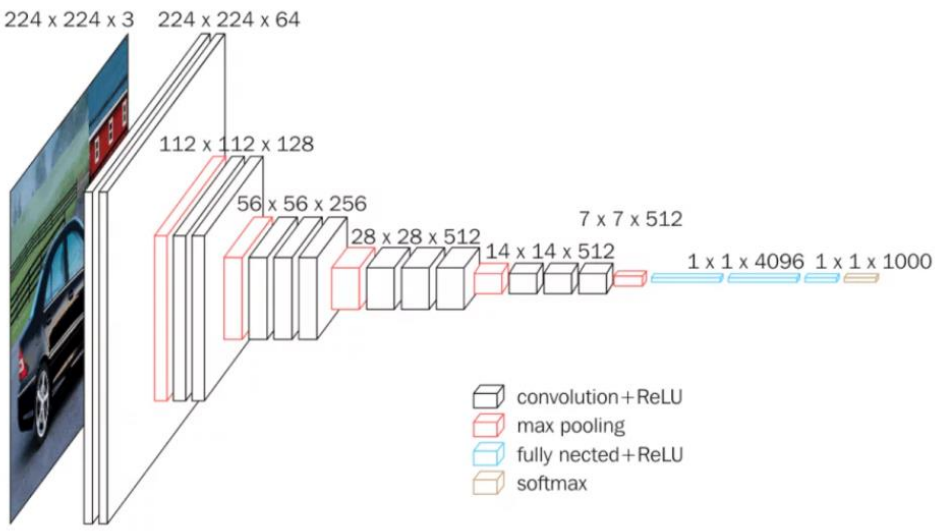


Figure 2.8 Structure of VGG16 which was proposed by Simonyan et al. [42] from the University of Oxford in the paper "Very Deep Convolutional Networks for Large-Scale Image Recognition".

CNNs enhance segmentation accuracy by extracting detailed image features through learned convolution kernels. Due to computation cost constraints, these networks typically consist of multiple small-sized convolutional layers stacked together, with down-sampling used to progressively reduce the spatial dimensions of each input slice. This process expands the receptive field (RF) of the convolution kernels, enabling a hierarchical extraction of features from shallow to deep layers and from local details to broader, global patterns. Through this approach, convolutional networks achieve multi-level feature extraction, essential for accurate and comprehensive image segmentation.

Supervised learning is a foundational and widely used approach for medical image segmentation, where pixel-level annotations perform as a critical training reference also known as ground truth (GT). This approach demands a large volume of high-quality labelled data, which is often challenging to collect due to patient privacy concerns and the cost of database construction. To enlarge the database, numerous public medical image datasets have emerged, e.g. LIDC-IDRI (Lung Image Database Consortium and Image Database Resource Initiative) [44], ISIC (International Skin Imaging Collaboration) [45], ChestX-ray14 [46], facilitating extensive research in supervised segmentation. Meanwhile, many notable neural networks like FCN [47], U-Net [48], DeepLab [49], and GANs [50] have accelerated advancements in medical image segmentation, pushing the developments of supervised learning in this field.

Minaee et al. [51] surveyed and summarized the latest literature on image segmentation, discussing over 100 DL-based segmentation methods proposed up until 2019. Among these, semantic segmentation networks can be broadly categorized into the following groups: 1. Fully convolutional networks (FCN), 2. Encoder-Decoder based models, 3. DeepLab family, 4. Models with attention mechanisms. In what follows, some representative models in each of these categories and corresponding medical applications will be discussed.

Fully Convolutional Network

The FCN [47] is the best paper of CVPR 2015 and the milestone of semantic segmentation technology in Deep Learning. Unlike classification models such as AlexNet and VGGNet [42], FCN replaces the fully connected layers in the neural network with convolutional layers, realising an end-to-end, pixel-to-pixel semantic segmentation network. The segmentation task is a particular case of the classification task, where the model processes the target from the entire image to every pixel in the image.

The production after the final convolutional layer is a multi-channel three-dimensional matrix. In the channel direction, each pixel in the different channel has information, representing the probability of belonging to each class. The channel index with the highest probability can become the category number of the pixel. There are two advantages of this change (fully convolutional operation):

a. The fully connected layer requires the input size of the network structure to be fixed because the weight matrix size of the linear layer is fixed, which is determined by the inside characteristics of the neural network, while the input image of FCN can be of any size, making it more flexible. The well-trained model can be used on any input size without any shape process, e.g., cropping and resizing.

b. Most of the model's parameters are concentrated in the fully connected layer. After being replaced by the convolutional layer, the total number of parameters of the FCN is significantly reduced.

The FCN architecture operates by progressively down-sampling the input image through repeated convolution and pooling layers. This down-sampling compresses the spatial dimensions of the image, allowing the network to capture deeper semantic information by shrinking the input size by factors of 32, 16, or 8. As the feature map spatial size decreases, channel depth increases, more abstract features could be learned. In the final stages of FCN

processing, the feature map is up-sampled back to the original image size. The earliest FCN version, FCN-32s [47], utilized this approach, where it directly up-sampled the feature map by 32 times to match the original dimensions. However, this method produced relatively coarse results, as significant detail was lost during the interpolation process. To improve on this, FCN-16s and FCN-8s were introduced, which achieve finer segmentation by incorporating information from both the deep and shallow layers. This approach fuses the coarse, high-level information from deeper layers, with the finer, spatially detailed features from shallower layers, before up-sampling, thereby restoring the original image size with improved segmentation accuracy. This fusion strategy effectively enhances the resolution and precision of the final segmentation output, capturing more detail and improving alignment with GT.

As a milestone, FCNs have three significant contributions:

a. Fully convolution

The fully convolutional model can accept inputs of different sizes and is more flexible.

b. Up-sampling with deconvolution

In the semantic segmentation task, after the convolution operation, the image size will be significantly reduced, but the semantic segmentation requires semantic annotation of each pixel of the image; the image needs to be restored to the original image size, which means being up-sampled to enlarge the shape. Standard up-sampling methods include bilinear interpolation algorithm and so on. The original text proposes that the interpolation algorithm will cause the feature map to lose some semantic information. Therefore, in training, the author uses deconvolution for up-sampling. Compared with other sampling methods, deconvolution has the characteristics of parameter self-learning.

c. Skip architecture

In FCN, the skip-level structure method is adopted, combining fine layers (more advanced convolutional layers) and coarse layers (up-sampled layers). Using this method, the image segmentation can be as fine as possible on the premise of preserving the global features. The author compared the performance of the three structures (FCN32s, FCN16s, and FCN8s) through experiments. From 32s to 16s, the model's performance improved significantly, reaching an mIOU (mean intersection over union) of 62.4; from 16s to 8s, the performance increased by only 0.3 percentage points, but smoothness and detail were slightly improved.

Although FCN(8s) achieves better results, the annotations obtained are still not fine enough and insensitive to details in the image, e.g. unsmooth points along the contour. Since FCN ignores global information when extracting semantic information, some improved methods are proposed to improve this shortcoming. In the ParseNet proposed by Liu et al. [52], the global average pooling is first performed on each channel direction of the feature map. Then the size is restored through the convolution operation, and the original feature map is calculated by weighted summation. The advantage is that the feature maps can obtain different global information on each channel. The average pooling operation proposed in ParseNet is similar in concept to the attention mechanism, which will be clarified below. In 2017, PSPNET [53] proposed the spatial pyramid module to better extract image context and multi-scale information and reduced the probability of image category mis-segmentation in the FCN network. The core of the spatial pyramid module is to get the features of four different scales through pooling operation, which means getting more receptive field information.

Encoder-Decoder based model

In the same year that FCN was proposed, an encoder-decoder-based model, U-Net, was also introduced in the segmentation field. The U-Net architecture, proposed by Ronneberger et al. in 2015 [48], was presented at MICCAI. It was built based on FCN's framework, incorporating a U-shaped encoder-decoder structure with skip connections to merge encoder and decoder information at each layer. The encoder-decoder design was first introduced by Hinton et al. in Nature (2006) [54]. At that time, the primary function was not for segmentation but for compressing images, saving cost and removing noise. It is only necessary to keep the images' features and a trained decoder when storing images.

The kernel part of the U-Net (Figure 2.9) includes two objects. The encoder is the feature extraction part, with four down-sampling operations, so that the spatial size of the feature map at the bottom is 1/16 of the original size, but the channel depth is increased by 16 times, which can increase the receptive field (RF). With the progress of down-sampling, the RF will expand step by step, equivalent to the compressed image, and the perceptible area per unit area becomes wilder. As the down-sampling proceeds, the low-frequency information of the image is more perceived. The decoder is the up-sampling part. After each up-sampling operation, the output of the encoder in the same layer will be fused by the skip connection. The fusion method of U-Net here is to concatenate and fuse features along the channel

dimension, while the fusion in FCN-16s and FCN-8s is point-by-point addition. Up-sampling can be regarded as combining information at various levels while restoring the size of the image. Still, this combining process is connected layer by layer, which will cause certain features to be lost.

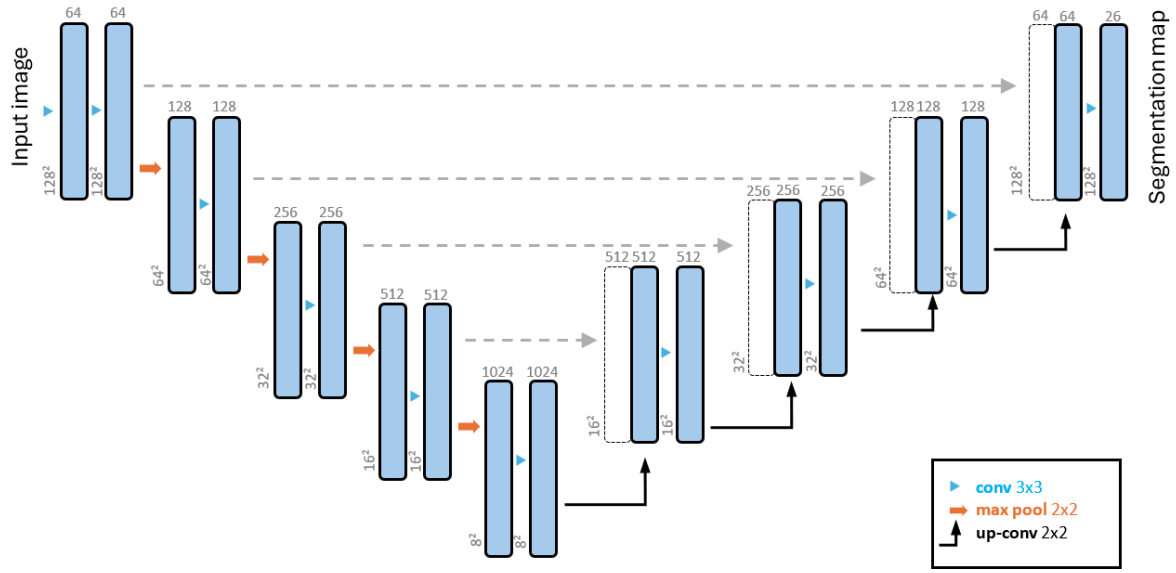


Figure 2.9 Structure of U-Net.

As the typical and basic benchmark model, U-Net has been improved by many scholars from different application fields in recent years. For example, UNet++ [55] and UNet3+ [56] are based on structurally modified U-Net models. UNet++ uses more short connections to strengthen the fusion of different RF in each layer. By comparing U-Nets of different layer depths, Zhou et al. [55] found that UNet++ can be regarded as a combination of overlap U-Nets of different layer depths through short connections. As for how many layers of U-Net works best, model can decide itself during back-propagation process.

Features or RF from each model layer have different sensitivities to target objects of different sizes. For instance, it is easy to identify large size objects by capturing the wild features with clear contours. In practice, the edge information represented by a few pixels of large or small things is easily affected by the down-sampling and up-sampling processes, repeatedly.

UNet++ has RF of different sizes, so in experiments, the segmentation accuracy of the UNet++ model in lung nodules, colon polyps and liver are on average higher than U-Net with 0.6 IOU points. In experimental comparisons in later chapters, UNet++ is used as a benchmark in the control group to compare with other modified models. Although the improvement of UNet++ is not obvious in the experiments after, the design idea explained by

Zhou et al. provides a good way for the modified U-Net model in the following chapters in this thesis. Although UNet++ utilises multi-scale information through nesting and dense short skip connections, essentially, the framework mainly relies on short connections, making some primary features are hardly used well. UNet3+ proposed in 2020 [56] is based on UNet++, which abandons frequent short connections but combines high-level semantic information and low-level semantic information in feature maps of different layers. In the liver and spleen datasets used in this paper, Unet3+ achieves better results than UNet and UNet++.

Another classic representative model is SegNet [57], using a similar structure as U-Net. The innovation is that the decoder uses the pooling index function when up-sampling. It is pointed out that the deconvolution used for up-sampling in FCN allows the model to learn parameters by itself. However, too many parameters may make it challenged for model to find the optimal global point. For SegNet, when using pooling down-sampling in the encoder, the position of each pixel in the result affined from the original map is recorded. Then when the decoder uses up-sampling to restore, combined with the pooling index, the pixels are first restored to the original position, and other places are filled with 0, then the convolution operation is performed. Badrinarayanan et al. conclude that such improvements can reduce the number of training parameters and improve boundary division. The primary purpose of SegNet is to design an efficient architecture for road and indoor scene understanding. Compared to SegNet, U-Net and its variants are still more suitable for segmentation in medical images.

DeepLab family

As mentioned above, the image semantic segmentation model is adapted from the neural network classification model. However, due to repeated pooling and down-sampling operations, certain detailed textures are lost in the feature map. Additionally, high-dimensional features' translation invariance reduces the resolution, making it challenging to accurately define the semantics of individual pixels.

DeepLab V1 [49] proposed in 2018 uses atrous convolution (also known as dilated convolution), which typically reduces resolution but increases the RF in CNN models like U-Net. The RF in a CNNs refers to the region in the original image that influences a given pixel in the feature map of each layer. Wilder RF allows the model to capture more contextual information from a wilder region in the original image. In U-Net, the RF is increased by down-sampling the input image multiple times, typically reducing the image to 1/16 of

original size. While down-sampling deepens the network and effectively increases the RF, it also loses some critical information, especially after multiple pooling layers. By using atrous convolution, DeepLab V1 could preserve spatial resolution and focus on essential details without losing context. In the ordinary convolution operation, the convolution kernel of 3×3 size is used (rate=1), the RF obtained in the feature map of the next layer is 3×3, and in DeepLab V1 (rate=2), the RF of the first layer is 5×5. Atrous convolution brings a larger field of view.

Although translation invariance enhances the model's ability to abstract data hierarchically, it also hinders low-level visual tasks, ignoring some low-level semantic information. DeepLab V1 adopts the condition random filed (CRF) post-processing method to solve this problem. In CRF, each pixel is used as a node, and the relationship between pixels is used as an edge, constituting a conditional random field. A binary potential function describes the relationship between pixels, and similar pixels are encouraged to be assigned the same label, while pixels with a significant difference are assigned different labels. The Deeplab V2 [58] proposed later adds the ASPP (atrous spatial pyramid pooling) structure to the base of DeeplabV1. A parallel sampling of given inputs with atrous convolutions of different sampling rates is equivalent to capturing the context of the image at multiple scales to achieve fine labels.

The DeepLab series of models are mainly constructed from the aspect of increasing the receptive field. While retaining the input image features as much as possible, the receptive field inside the model is enlarged, thereby offsetting the loss of information due to the need to obtain higher-dimensional semantic information.

Attention mechanism

Attention mechanism refers to a technique which can simulate human perception, such as vision or hearing, in deep learning, making models to focus more on specific features of the input data that are most relevant to the target task, enhancing prediction performance [59].

RNNSearch, an attention model [60], was the first model to apply attention mechanisms to machine translation tasks. In computer vision field, models primarily utilize self-attention mechanisms to calculate attention weights by analyzing the statistical relationships between pixels, this method that has been applied to various task scenarios [61], [62], [63]. Based on the dimensions of features considered, attention mechanisms can be categorized into spatial attention, channel attention and scale attention. Spatial attention is one of the most widely used module. It selectively focuses on different spatial regions of the input by assigning

varying attention weights to each pixel and is extensively applied to tasks such as image classification [61], and image segmentation [64].

In terms of image segmentation, representative models include Attention UNet [65], SE module [64], and the Convolutional Block Attention Module (CBAM) [66].

In [65], a novel Attention Gate (AG) structure was proposed, which automatically enables the model to learn target structures of different sizes during self-study. The Attention UNet integrates feature maps from two layers into an Attention Gate (AG) to generate a single-channel weight grid. This grid is multiplied with the feature map, assigning varying weights to different feature pixels. The AG can update weight grid through back-propagation. Its low computational cost, and easy integration into standard U-Net series make it effective for challenging tasks like CT pancreas segmentation.

Squeeze-and-Excitation module (SE) [64], developed by Hu et al. at Momenta, won the 2017 ImageNet image classification task by a large margin. Its kernel components are the squeeze and excitation modules. The squeeze module reduces the spatial dimensions of the feature map to 1×1 using global average pooling, then uses a fully connected layer to predict channel-wise weights, resulting in a 1×1×C weight vector. This vector is multiplied with the original feature map, enabling information exchange between channels. By integrating this process between adjacent layers, SE significantly enhances network accuracy.

The AG and SE enhance accuracy by focusing on spatial and channel attention in feature maps. CBAM [66] combines both, applying channel and spatial attention to independently weight feature maps along these dimensions. CBAM calculates weights using both average and maximum pooling operation, with experiments showing superior performance compared to single pooling strategy. Its flexibility and simplicity make CBAM a more adaptable structure than the AG, allowing integration into any CNN architecture.

2.3 Research gaps and motivation

Muscle manual segmentation is very laboured, 10 hours per subject in average for the pilot study using the same older women dataset [6], so that it is not a viable option in most cases for muscle segmentation which also highlights the importance of muscle automatic segmentation study. Accurate muscle segmentation is crucial for various applications such as tracking muscle degeneration in neuromuscular diseases, planning of orthopedic surgery, and monitoring the effectiveness of rehabilitation programs. In such contexts, even small segmentation errors may lead to significant misinterpretation of muscle volume or morphology, which underlines the need for both automation and precision. In general, automatic segmentation of muscles in medical imaging compared to natural objects segmentation is a more complex task and can be divided into three key steps: data pre-processing, model based automatic segmentation and post-processing. Each step plays a vital role in obtaining more accurate results. However, despite there are many significant progress and development in medical image segmentation, there remain critical gaps specific to muscles segmentation application, particularly in the context of considering the anatomical geometrical features carried by the muscles themselves. In addition, the generalization of the model in muscle automatic segmentation is also a direction that few studies have covered. This section will focus on research based on the three key steps and model generalization analysis, analyse the gaps and introduce our study's motivation. This work is particularly motivated by the clinical need for accurate and efficient muscle segmentation, as manual methods are prohibitively time-consuming and impractical in routine clinical workflows.

Pre-processing step

Pre-processing is a fundamental stage for preparing medical images in segmentation task, such as ensuring consistent image quality, removing noise, improving imaging condition and data augmentation.

DL based models are deeply data-driven, the data quality and quantity could decide the level of model's performance [67]. During medical image models training, one common problem is the size limitation of available training database, due to many factors like data privacy, time consuming scanning operations, which could affect the segmentation performance [8]. To solve the problem, many data augmentation methods are proposed such as random scaling, cropping, flipping, noise addition and spatial transformations to expand the database size [43], [68]. However, due to the strict requirements to maintain the anatomical features of

medical tissue, MRI-specific image augmentation is more challenging compared to normal natural objects augmentation.

To date, there has been insufficient research on this subject. Thyreau et al. [69] segmented the hippocampus by enhancing the contrast at image borders. Ronneberger et al. and Dong et al. [48], [70] used deformable registration techniques to segment brain tumours and cells. Shin et al. [71] used GAN networks to generate synthetic MRI images of abnormal brain tumours, improving model performance. In [9], the deformable registration techniques was applied to augment muscle data and just achieved a slightly better accuracy (DSC, 0.73).

The methods mentioned above appear to focus on using nonlinear deformations or simple image modification to modify target objects and generate new data, which is particularly suitable for targets like cells and tumours which don't have a fixed shape. However, for structures of tissues like muscles, which may exhibit volume differences across cohorts but maintain consistent topological structures, these methods have limitations. They cannot ensure that the newly augmented data meets the anatomical constraints in practice.

To address this gap, out study proposes a Statistical Shape Model based data augmentation method. The kernel idea of it is to compute the mean shape of the same individual lower limb muscle across different subjects under same cohort, and then, by incorporating principal component analysis (PCA) techniques, introduce deformations within a certain range. This method ensures that the augmented data is not only distinct from the original data but also retains its authenticity.

Model based automatic segmentation of muscle

With the popularity of U-Net series models applied in image segmentation, many modified structures are proposed to improve the prediction accuracy. Only a few studies have used deep learning modified models to perform muscle segmentation from MRI slices. Zhu et al. [2] segmented 13 muscles from the shank of 4 children using T1-weighted MRI images. The dataset included 15 subjects for training and 1 for validation, with ages ranging from 5.4 to 14.8 years (6 females, 14 males, and 6 subjects with cerebral palsy). In their study, both the original U-Net and its best-performing variant yielded similar accuracy, the proposed hybrid model, H-DenseUNet achieved the best score (DSC 0.90, ASSD 0.5 mm). Ni et al. [17] proposed a cascaded three-dimensional deep convolutional neural network to segment all lower limb muscles of a testing cohort includes 13 healthy adults, achieving DSC around 0.90 in average. The accuracy is comparable to the manual segmentation. Other modifications like

UNet++ [55], MultiResUNet [72] achieved success in other tissue's segmentation. They employ multiple short connections to enhance the integration of different receptive fields within each layer or focus solely on integration within the same layer. However, in the expansive path of the U-Net architecture, deep features are gradually combined with shallow features from the bottom to the top. This process allows semantic elements from adjacent layers to fuse, but as it progresses upward, semantic gaps persist between non-adjacent layers, leading to information loss.

To address this and improve the integration of semantic features from different layers, this study proposed a modified U-Net networks, by adopting feature fusion and attention mechanisms [12], [66] to weaken the side effects from information loss.

Post-processing step

For muscle segmentation, post-processing stands as a final and critical step to refine the prediction results from the initial segmentation algorithm, CNN in this study. During literature review, some post-processing methods have been used to enhance the final prediction results. For instance, in a liver segmentation task [73], where the assumption is that an organ should be represented as a single volume with the largest labelled region, a 2D post-processing technique was employed. This method involved erosion, identification and selection of the largest region, dilation, and edge smoothing, resulting in a 4% improvement in Intersection over Union score (IoU). Or like Conditional Random Field (CRF) post-processing methods presented in [18] and [74], it was attached to the end of DeepLab V1 and resulting in 4% improvement of IoU.

Although the deep leaning models can make segmentations with DSC around 0.90 for research in muscle segmentation [2], [17], there is still some error existed in the predicted image, like miss-segmentation areas and local error close to the main cohort, which could be improved. In previous studies, operations like hole-filling [17] and binary closing [2] methods were used to fill the empty holes or remove outliers in muscles.

Although these methods achieved good segmentation results, previous studies did not consider the 2D or 3D shape of the segmentation target. This is particularly relevant for muscles, which, despite variations in volume across different subjects, exhibit similar spatial morphology and topological structure [75], [76]. Moreover, these methods have a significant limitation: if the segmentation of a continuous structure results in the target being split into

two disconnected parts, the traditional approach may easily remove one of the segments which belong to the right areas. This would further degrade the segmentation results.

Considering the existing gap, our study proposes a mean shape based post-processing method to increase the segmentation accuracy of the output from convolutional neural networks, with a focus on accurately delineating muscle boundaries and reducing false positives.

Generalization performance

For muscle segmentation to be adopted in real-world clinical settings, models must demonstrate strong generalization across different populations and imaging conditions. Generalization of DL-based models refers to a model's ability to make accurate predictions on unseen data that comes from the similar distribution as training dataset. The goal of generalization is to enable the model to capture the main representative features in the training data, allowing it to perform well on previously unseen subjects. This capability is critical for the practical application of deep learning algorithms, as it ensures models can reliably operate in real-world scenarios beyond their training datasets.

In this study, the model’s generalization performance means how the model which trained on limited kinds of cohorts or subjects performs segmentation on unseen or different cohorts’ muscles. Despite some research achieved good results (DSC around 0.90) on lower-limb muscle [2], [7], [17], their target category is single, just for athlete, children or old women. In the previous study, there is no experiment to train and test the model on two different cohorts, like children and the old people, athletes and the unhealthy people.

To address this gap, our study aims to investigate the generalization performance of deep learning models in segmenting lower-limb muscles across diverse cohorts. Specifically, we will explore how a model trained on one cohort (e.g., typical developed children) performs when applied to a significantly different cohort (e.g., young people or old women). It will provide insights into the robustness and adaptability of segmentation models.

Chapter 3 Automatic segmentation of skeletal muscles on PMW from MR images using FFU/AFFU and SSM based data augmentation approach

This chapter is based on a paper published in Frontiers in Bioengineering and Biotechnology (2024): 'Automatic Segmentation of Skeletal Muscles from MR Images using Modified U-Net and a Novel Data Augmentation Approach' by **Lin Z**, Henson WH, Dowling L, Walsh J, Dall'Ara E, Guo L. Doi: https://doi.org/10.3389/fbioe.2024.1355735

3.1 Introduction

The skeletal muscles of the lower limbs play a fundamental role in generating strength and facilitating movement in daily activities and are particularly important when assessing conditions related to mobility impairments. For instance, diseases and chronic conditions such as sarcopenia—age-related loss of muscle mass and strength—can significantly reduce mobility in older adults [1]. Similarly, stroke patients often experience sarcopenia due to muscle denervation, inflammation, and inadequate nutrient intake, leading to muscle mass loss and impaired function [3], [4]. In children, conditions like cerebral palsy, caused by non-progressive brain injury at birth, can lead to musculoskeletal deformities and further hinder movement [77]. Monitoring changes in muscle properties over time can help provide a better understanding of these musculoskeletal diseases and lead to more effective diagnostic approaches and interventions [5].

However, the effective assessment of muscle properties remains challenging when using clinical imaging techniques, such as Computed Tomography (CT) and Magnetic Resonance Imaging (MRI). Among these, MRI is preferred for evaluating muscle volume and shape as it avoids the ionizing radiation associated with CT [6], [23]. Muscle segmentation from biomedical images is a crucial step for measuring the geometrical properties of skeletal muscles, which are linked to functional muscle properties [2], [6], [78]. Traditionally, manual segmentation of MRI 2D slices has been used to delineate muscle structures. However, this process is very time-consuming — requiring approximately ten hours to manually segment 35 lower limb muscles of one subject—and prone to variability, with inter-operator reproducibility errors exceeding 10% for certain muscles [6], [79].

To address these challenges, numerous image segmentation methods have been proposed, traditionally based on digital image processing and optimization algorithms. These conventional methods [34], [35], [38] normally extract low-level semantics of images, such as colour, texture, or shape information, without considering high-level semantic information, and could lead to imprecise labelling results. With the emergence of Deep Learning (DL), Artificial Neural Networks (ANNs), and Convolutional Neural Networks (CNNs) [42], [43], traditional segmentation methods are gradually being replaced since semantic segmentation algorithms based on DL and CNNs can extract mid-to-high-level semantic information from images, improving the segmentation results [80].

In recent years, deep learning (DL)-based segmentation techniques have shown promise in replacing manual methods. Zhu et al. [2] developed a hybrid U-Net model that achieved a Dice Similarity Coefficient (DSC) of 0.88 for segmenting 11 lower limb muscles (lateral gastrocnemius, medial gastrocnemius, soleus, plantaris, tibialis anterior, tibialis posterior, extensor digitorum longus, peroneus brevis, popliteus, flexor digitorum longus and flexor hallucis longus) in both healthy children and those with cerebral palsy. Similarly, Ni et al. [17] created a 3D CNN that performed comparably to manual segmentation in young athletes, reaching a DSC of about 0.90. However, while these models have been successful within specific groups, they haven't been widely tested in older or more diverse populations, like post-menopausal or obese women. Compared to children and young men, post-menopausal women experience a decline in muscle content with aging, resulting in less distinct boundaries between individual muscles and an increase in fat content, which poses a greater challenge for automatic individual muscle segmentation. Moreover, the segmentation of skeletal muscles in post-menopausal obese women is potentially more challenging due to the thick layer of fat around the muscles. While automatic image segmentation approaches based on CNN algorithms are usually calibrated and tested on similar cohorts of subjects, to improve the applicability of the models it would be beneficial to assess the model's accuracy for different testing datasets.

Currently, one of the most popular DL networks used for medical image segmentation is U-Net [2] and has been shown to be powerful in medical image segmentation, e.g., cell segmentation [81], liver and tumour segmentation [82]. Despite the success of existing U-Net type networks, they suffer from some limitations including the hard coding of the receptive field size, as well as that they do not account for inherent noise in the data. Several improved U-Net networks have been proposed, where UNet++ [55], an improved version of U-Net, employs multiple short connections to enhance the integration of different receptive fields at each layer, compromising increasing computational cost. In [72], the MultiResUNet was proposed to enhance the semantic integration between the encoder and decoder; however, it only focuses the integration within the same layer. In the expansion path of the U-Net architecture, deep features are gradually combined with shallow features from bottom to top. This process allows the fusion of semantic elements from adjacent layers, but as moving towards the upper layers, there remains a semantic gap between non-adjacent layers, resulting in information loss. To address this and improve the integration of semantic features from

different layers, in this study modified U-Net networks have been developed, by adopting feature fusion and attention mechanisms [12], [66].

One common issue with supervised learning for image segmentation is the limited size of available training datasets, which can significantly affect the performance of automatic segmentation methods [8]. To address this, data augmentation techniques such as random scaling, cropping, rotation, flipping, noise addition, and spatial and grayscale transformations are often used to expand the dataset size [43], [68]. However, augmenting medical images is more challenging compared to natural images, due to the strict need to maintain medical and anatomical accuracy.

There has been relatively little research on MRI-specific image augmentation. For instance, Thyreau et al. [69] segmented the hippocampus by adjusting the contrast of image borders, while Ronneberger et al. [81] and Dong et al. [70] applied deformable registration techniques for brain tumour and cell segmentation. Shin et al. [71] used GAN networks to synthesize abnormal brain tumour MRI images, optimizing model performance through data augmentation. Despite their advancements, these methods face limitations when applied to muscle segmentation. Deformable registration can excessively distort muscle shapes in new subjects, leading to a loss of anatomical consistency, such as the spatial location, size, and shape of the muscles [9]. Moreover, randomly applying deformations to target images does not effectively enhance the variety of image features for model learning, especially in the case of muscle slices.

The first aim of Chapter 3 was to design modified U-Net structures (AFFU/FFU) by introducing features from different layer, fusing them by attention modules and evaluate them with other popular benchmark models (U-Net and UNet++) on post-menopausal women cohort. The second aim of this chapter was to propose a novel data augmentation method based on Statistical Shape Model (SSM) for extending the size of the MRI datasets. By comparing the performance of the same model before and after adding augmented data, the potential of the new data augmentation method can be investigated.

3.2 Materials and methods

3.2.1 Participants

The data from three different cohorts of post-menopausal women (PMW) were utilized in this study. The first cohort consisted of T1-weighted magnetic resonance images (MR images) of 11 PMW (PMW-1) with no muscle disease who were recruited by the Metabolic Bone Centre (Sheffield, UK) as part of larger studies (approved by the East of England–Cambridgeshire and Hertfordshire Research Ethics Committee and the Health Research Authority, Reference 16/EE/0049) [6]. Two other cohorts recruited from a previous observational (approved by the Leeds West Research Ethics Committee, Reference 20/YH/0274) study were included in this work: 8 PMW (PMW-2) with no muscle diseases and 10 obese PMW (PMW-OB, detailed statistical information seen in Table 3.1) [83]. PMW-1 was funded by Engineering and Physical Sciences Research Council (EPSRC, Frontier Multisim Grant, EP/K03877X/1, and EP/S032940/1), and PMW-2/OB was funded by the Medical Research Council (MRC) and Versus Arthritis (Medical Research Council Versus Arthritis Centre for Integrated Research into Musculoskeletal Ageing, CIMA, (MR/R502182/1).

Table 3.1 Statistical information.

Cohort	Age (year)	Weight (kg)	Height (cm)	BMI
PMW-1	69.0±6.7	66.9±7.7	159±3	26.5±3.4
PMW-2	65.8±3.9	57.1±5.8	161±3	22.0±2.1
PMW-OB	64.7±4.5	84.1±15.6	159±9	33.0±3.2

All the three cohorts' full lower-limb MRI scans were performed on the 1.5T Siemens Magnetom Avanto or 1.5T Siemens Magnetom Aera (Siemens AG, Erlangen, Germany) under same scanning protocol. Four sequences were taken to capture the hip, thigh, knee, and shank [9]. To minimize scanning time without losing detailed joint geometries the joints were scanned with a higher resolution (pixel size of 1.05 mm x 1.05 mm and slice thickness of 3.00 mm) than the central portions of the long bones (pixel size of 1.15 mm x 1.15 mm and slice thickness of 5.00 mm). The sequences were then combined in MATLAB (R2006a) to create a continuous 3D image from the hip to the ankle [84]. This was done by standardizing the resolution of each imaging sequence from the different sections through tri-linear interpolation (interp3, MATLAB R2006a) to be 1.00 mm x 1.00 mm x 1.00 mm. In this chapter, a total of 25 lower limb muscles of PMW-1 cohort spanning from the knee to the hip

were manually segmented using Mimics (version 26.0; Materialise) and Amira (version 2022.2; Thermofisher) by three experienced operators for each subject [6], which is considered the gold standard for muscle segmentation. The same muscles from PMW-2/OB were manually segmented by the same author using Amira (version 2022.2; Thermofisher).

The inter- and intra-operator reproducibility of the manual segmentation on PMW-1 was found to differ from muscles to muscles (range CoV: 4.2%-22.8%) [6] and only some of them where the intra-operator CoV is less than 10% were reproducible enough to be used for calibrating the DL models. Ultimately, during the model evaluation phase, 16 muscles were selected as the objective muscles for analysis: rectus femoris (RF), vastus intermedius (VI), vastus lateralis (VL), vastus medialis (VM), sartorius (SAT), semimembranosus (SMB), semitendinosus (SMT), gracilis (GRA), biceps femoris caput brevis (BCB), biceps femoris caput longum (BCL), adductor magnus (AM), adductor brevis (AB), adductor longus (AL), gluteus maximus (GM), iliacus (IL), and tensor fasciae latae (TFL).

The setting of the computer used to train the model is configured as: GPU: NVIDIA GeForce RTX 3060 Ti; RAM: 16.0 GB.

3.2.2 Modified U-Net: Feature-Fusion-UNet/Attention-Feature-Fusion-UNet

In this section, we used U-Net [81] and UNet++ as benchmarks for our segmentation task [70], [81]. U-Net [2], [81] is a 2D multi-layer Encoder-Decoder U-shaped neural network. The Encoder consists of convolutions and down sampling operations, enabling it to learn the features of input images and transmit them to the lower layers, generating a feature map, which is referred to as feature extraction. In the down sampling process, the receptive field gradually expands, resembling a compressed image, allowing for a larger perceived area per unit. Additionally, down sampling captures more low-frequency information from the image. The Decoder utilizes features to restore the original resolution of the feature map and perform pixel-level prediction. After each up-sampling operation, the output of the encoder at the corresponding layer is merged using skip connections. In U-Net, feature fusion is achieved by concatenating and merging features along the channel dimensions. UNet++ [55] is a nested version of the U-Net architecture used for semantic segmentation, particularly in medical image analysis. It improves upon the U-Net by enhancing feature extraction and has shown superior performance in medical image segmentation tasks. It incorporates convolution layers

on skip pathways to bridge the semantic gap and employs dense skip connections to enhance gradient flow, distinguishing it from the original U-Net.

Two modified U-Net structures were developed in this chapter. A Feature-Fusion-UNet (FFU) incorporated a fusion part into the basic U-Net model to improve the integration of semantic features from different layers. In FFU, the fourth, third, and second decoder layers' feature mappings were using up-sampling operations to zoom in features of 8/4/2 times to maintain consistent sizes and then extracted by using the Atrous Spatial Pyramid Pooling (ASPP) [49] blocks to fuse the semantic information, which was subsequently concatenated with the output. Furthermore, to address potential issues such as gradient loss, a connection structure resembling the residual network was integrated into the top layer, facilitating improved gradient transfer and making the model converging faster [85].

Furthermore, the Attention-Feature-Fusion-UNet (AFFU) was proposed based on FFU to explore the employment of multiple attention mechanisms to achieve the desired segmentation accuracy. The AFFU incorporated attention gates into each individual layer [12] and used CBAM block [66] to help compensate for any deficiencies in a single attention gate, and mitigated losses during resizing, enhancing the neural network's receptive field using a pyramid pooling technique. The structure of FFU and AFFU is shown in Figure 3.1.

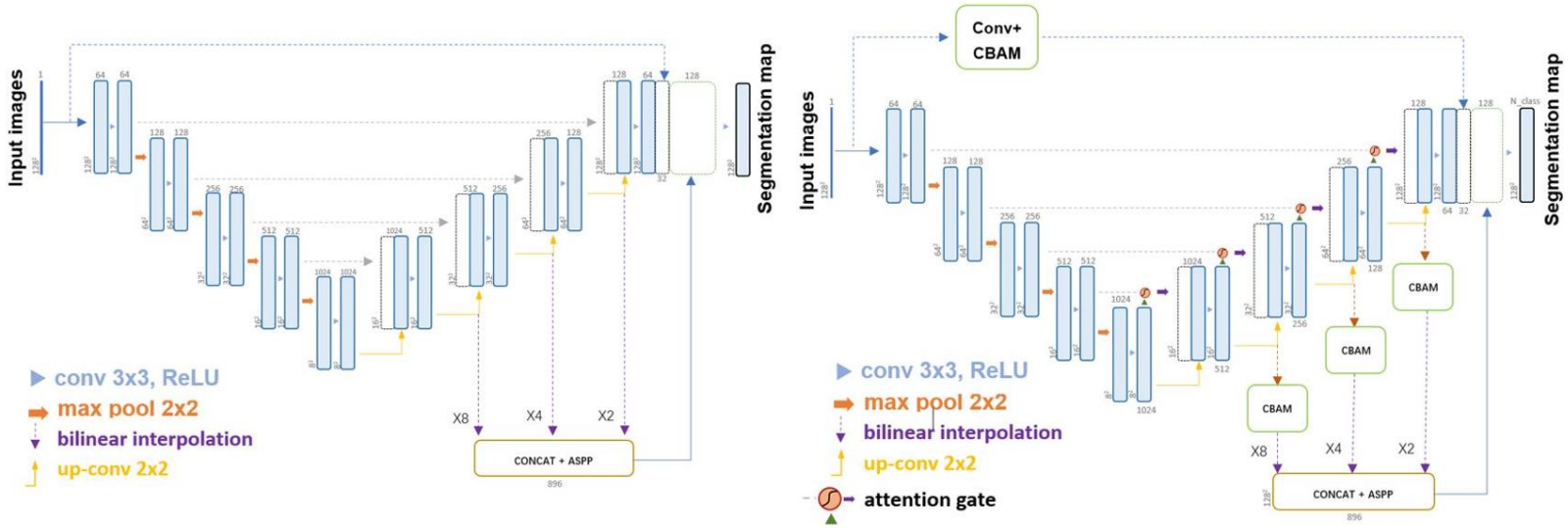


Figure 3.1 Structure of FFU (left) and AFFU (right).

3.2.3 SSM based data augmentation

The proposed data augmentation method was based on statistical shape modelling (SSM) [10], [86], [87], [88] which is a recently developed method that involves modelling medical images to obtain an average shape model and deformation models of a certain object class in the training dataset.

To define homologous surface points in the overall dataset, correspondence points are represented by sampling each shape surface. All n three-dimensional correspondence on a shape sample can be encoded as a $3n$ vector set to represent the entire shape, which can be subjected to statistical analysis. The accuracy of the analysis is determined by the quality of point selection. Generally, correspondence points are selected manually as landmarks on the target surface, but the increasing quantity of data samples makes manual selection more cumbersome and unsatisfactory [10]. Parametric methods have been used to address this issue, such as [88] which uses a method based on fixed spherical parameterization mapping to select points. However, most medical organs' shapes cannot be accurately parameterized, so these methods still have limitations. Cates et al. [10] proposed a non-parametric analysis system —— Shape Modeling with Entropy-Based Particle Systems (ShapeWorks correspondence optimization algorithm), which can optimize the distribution of a set of shape-similar correspondences and seeks a compact distribution of corresponding points in the shape space through an optimization algorithm to simplify the model. The basic idea of ShapeWorks is to construct a set of dynamic particles, each representing a set of correspondence points constrained on the surface of one shape. By balancing the negative entropy of particle distribution on each shape surface and the entropy of each shape's distribution in shape space, and using gradient descent, the positions of the correspondence can be optimized.

Assuming that there are N samples with m correspondences on each contour, which can be seen as a 3×1 vector in shape space, x_k represents a set of particles from kth shape, $x_k \in R^{3m}$, $x_k = \{x_k^1, x_k^2, x_k^3, \dots, x_k^{3m},\}$, $k = \{1,2, \dots, N\}$ and let $Z = \{x_1, x_2, \dots, x_N\}$, Z is a random variable of a shape in the shape space after aligning all shapes to the same coordinate system. Then the method consists in minimizing the energy function where H is an estimation of differential entropy

$$Q = H(Z) - \sum_{N} H(x_k). \tag{3.1}$$

where the first term in Q minimization in *Eq 3.1* targets a compact distribution of each individual's sample in the shape space and the second term aims to increase the entropy of the correspondence distribution by achieving a more uniform distribution of points on each shape. The algorithm balances individual shape variation and overall shape similarity by minimizing both terms, allowing it to accurately model shape variations in the dataset.

The method was implemented using the publicly available software tool ShapeWorks [88]. n different subject shape models of the same muscle were provided as input. The software then generated a mean shape model, from which $N-1$ modes were extracted using principal component analysis (PCA). Multiple deformations were produced based on the average shape model with different modes.

Moreover, in order to generate new MR images associated to the newly generated muscle labels, a correspondence mapping between the original and new labels was calculated using the deformable registration algorithm Sheffield Image Registration Toolkit (ShIRT)[9], [89]. ShIRT is a non-linear image registration algorithm that effectively tackles the significant anatomical variability commonly found in medical images [9], [89]. Its unique approach involves calculating a displacement function that maps every pixel or voxel in the reference image to its corresponding point in the target image. Through an iterative optimization process, the algorithm further refines the mapping matrix to minimize the cost function, which is determined by summing the squared differences between the intensity values of the two images. To generate a 3D displacement field, the algorithm uses trilinear interpolation between grid nodes. This field is then applied to transform the reference image, resulting in a registered image produced through trilinear interpolation.

By setting two parameters, the Nodal Spacing (NS) and the smoothing coefficient λ, ShIRT performed a non-linear deformable registration with high dimensional variability between input images or shapes. The two parameters of the deformable registration (NS and λ) were chosen from a previous sensitivity analysis equal to 5 voxels (5mm) and 50, respectively [9]. The obtained mapping was applied to the original MR images using the same transformation matrix, resulting in a new set of segmented subjects without the need for manual segmentation.

The pipelines of the proposed data-augmentation method are shown in Figure 3.2.

Figure 3.3 reveals that the application of the ±1 standard deviation mode to modify the mean shape yields noticeable alterations in volume and introduces heightened diversity in terms of texture within the generated new data.

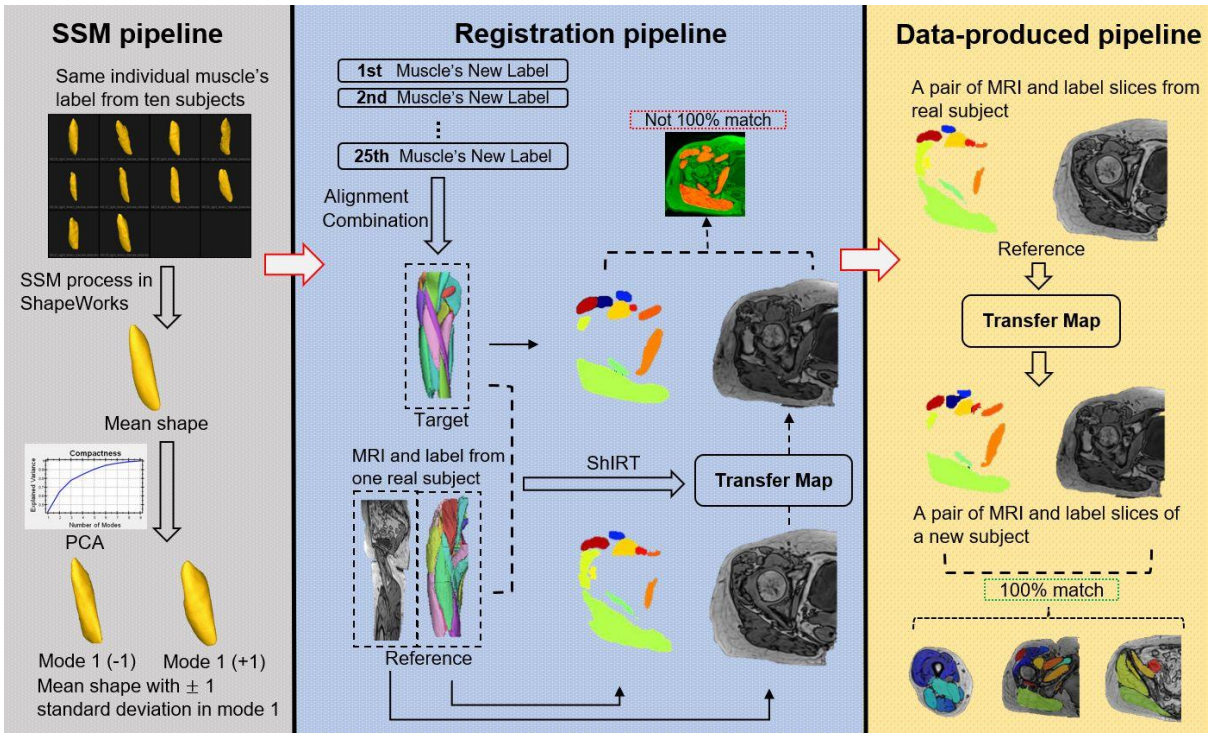


Figure 3.2 Data augmentation combined with Statistical Shape Model (SSM) process, registration, and data-produced pipelines**.** This figure illustrates the process for generating new subjects with matched labels without the need for manual image segmentation. The process is accomplished by using SSM and deformable image registration on ten original subjects' MRI. The SSM pipeline (based on ShapeWorks software) inputs ten shapes of the same muscle from different subjects, which were all from PMW-1 and applies an iterative optimization formula to generate corresponding points (default 1024 points) on all surfaces, calculates the mean shape, and uses PCA to extract nine principal features. New shapes are generated by adding or subtracting modes at different magnitudes from the mean shape. The registration pipeline involves obtaining new samples for all 25 muscles using SSM. A new muscle label (target) is formed by aligning the new samples with the original samples, and an original subject (reference) is selected. The MRI, label, and target's label (mean shape) of the reference are inputted into ShIRT, which generates a mapping matrix that aligns the reference with the target labels. Using this map, the MR image of the original sample is transformed, resulting in an MR image that matches the target. However, due to image noise and the diversity of the target during registration, the transformed MR image may not perfectly overlap with the annotation generated by SSM [90], and therefore, cannot be directly used for subsequent model training. However, the transformation map generated by this pipeline can be preserved. The data-produced pipeline applies the mapping matrix generated by SSM to the label and MR image of the reference, resulting in fully matched image and label data (bottom right).

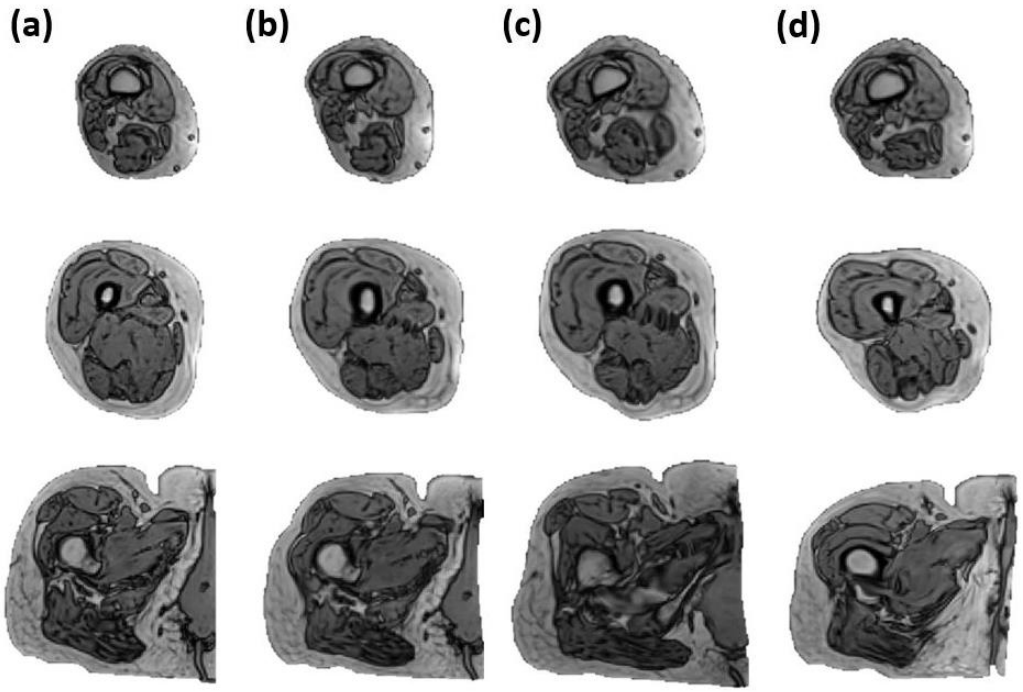


Figure 3.3 Augmentation results of one subject. The figure displays the outcomes derived from subjecting an individual to the augmentation process. Column (a) illustrates three slices of the original MRI (distal to proximal from top to bottom) from the right lower limb of one subject. (b), (c), and (d) illustrate the results obtained by mapping the subject to the mean shape and mean shape ± 1 standard deviation mode in PCA, respectively.

3.2.4 Data pre-processing and training procedure

The dataset used in model training and testing consisted of slices obtained from the right lower limb's main thigh part (knee to hip) of each subject. All images were cropped to a width of 256 pixels for input, keeping the original resolution fixed. During training process, random clipping was applied to further crop the input images to a width of 125 pixels to speed up training time and enhance the model's robustness against noise.

To reduce the computational requirements and considering the memory size of the equipment, a batch size of 16 was chosen for the U-Net and UNet++, while a batch size of 10 was used for our proposed models. The initial learning rate was set to 0.01 and decreased by a factor of 0.9 for each epoch. The model was trained for a total of 100 epochs, and the best weights on the validation set were retained for making predictions on the test set. The hyper-parameters were empirically tuned for best performance [84].

3.2.5 Experimental design

The experiment was split into three parts.

3.2.5.1 Leave-one-out cross-validation on PMW-1

In the first part, four CNNs (U-Net, UNet++, FFU and AFFU) were trained and tested on PMW-1 data (11 subjects) using the leave-one-out approach. In each trial, a single subject is used as a test sample and the rest is used as training, ensuring that each subject is used in the test set once. This approach was used to verify whether FFU and AFFU could improve the segmentation performance compared with the benchmark U-Net and UNet++ models.

3.2.5.2 Cross-dataset generalization testing

The second part involved training on 10 subjects from PMW-1 and then utilizing the trained network to test on: the remaining 1 subject from PMW-1, 8 subjects from the PMW-2 dataset, and 10 subjects from the PMW-OB dataset. The objective of this part of the study was to evaluate the model's generalization ability across different cohort (healthy and obese post-menopausal women).

3.2.5.3 Augmentation-based training for enhanced performance

For the third part, 10 subjects from the PMW-1 dataset used in the previous part were used as input to the augmentation pipeline, expanding the dataset to 37 subjects in total, based on the

mean shape of the muscles and including a variability of ±1 standard deviation associated to the PCA mode that describes best the variability in the dataset (mode 1). The augmented dataset was then used to re-train the model weights obtained from the second part of the study. This part of the study aimed to investigate whether training with the augmented dataset improved the segmentation performance of each model on the testing data (1*PMW-1, 8*PMW-2, 10*PMW-OB).

3.2.6 Evaluation metrics

For each part of the study three metrics were used: Dice Similarity Coefficient (DSC), Relative Volume Error (RVE), and Hausdorff distance (HD) [2], [9], [70]. These metrics provide complementary insights into the performance of the segmentation models and help assess their accuracy, volume similarity, and the degree of local errors.
The Dice Similarity Coefficient (DSC) was calculated by using equation Eq 3.2, which shows the similarity between reference shapes (R) and predicted shapes (P).

$$DSC = \frac{2|R \cap P|}{|R| + |P|} \tag{3.2}$$

where R is the set of voxels in the reference labels from manual work and P is from the model's prediction results. A DSC equal to 1 indicates a perfect match between the segmentation result and the ground truth, while a value of 0 represents no overlap.

Relative Volume Error (RVE) was utilized to assess the similarity between the segmentation results and the ground truth in terms of total volume of the individual muscle. RVE was defined as

$$RVE = \frac{V_{pred} - V_{ref}}{V_{ref}} \times 100\% \tag{3.3}$$

where V_{ref} and V_{pred} are the volume of reference and prediction cohort respectively. For the convenience of display, 1-RVE was used as measurement of accuracy.
The Hausdorff distance (HD) is a metric that measures the maximum distance between the reference and predicted external surfaces. It provides information about the extent of local errors in the model's prediction. HD was calculated as

$$HD(R, P) = max\{d(R, p), d(r, P)\} \tag{3.4}$$

where r is the voxel on surface of R and p is the voxel on surface of P, d is a function to find the minimum distance between r or p and the nearest point within the cohort P or R.

3.2.7 Statistics

When comparing two different models the difference between same individual muscle metrics were tested for normality using a Kolmogorov-Smirnov test, concluding that the difference between the paired values were not normally distributed. Therefore, a Wilcoxon signed-rank test was utilized. The paired values of each model's DSC, 1-RVE and HD metrics were compared across the four models. A significance level α=0.05 was considered.

3 Results

3.3.1 Performance of the models on the PMW-1 cohort

In total, each model predicted 176 labels representing individual lower limb muscle (11 subjects with 16 muscle labels each) through cross-validation. Table 3.2 presents the score of each model to predict the three considered metrics, DSC,1-RVE and HD. Significant differences between the accuracy of the models and the accuracy of the U-Net or the UNet++ are reported in Figure 3.4.

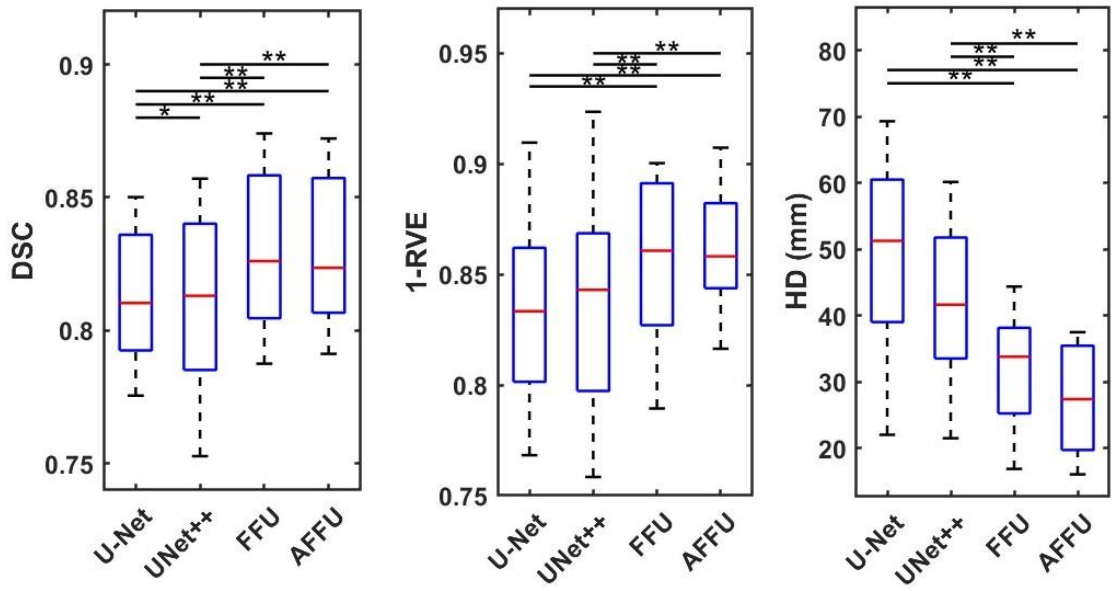


Figure 3.4 Performance of each model established using cross-validation. Box plots for the DSC (left), 1-RVE (middle) and HD (right, lower scores indicate better performance) for each model tested on the PMW-1 cohort (16 mean values of each individual muscle calculated from 176 predicted labels) are shown (* indicates $p<0.05$, ** indicates $p<0.01$).

Table 3.2 Model performance comparison under three metrics. The *p*-values in the table represent the significance results of the Wilcoxon signed-rank test conducted on muscles in all validation subjects between the models and U-Net. The *p** values represent the significance results of the Wilcoxon signed-rank test conducted on muscles in all validation subjects across the cross-validation process between the models and UNet++. Significant differences are reported in bold ($p < 0.05$ *or* $p^* < 0.05$).

	DSC				
	Mean ±SD	p	Gain vs U-Net [%]	p^*	Gain vs UNet++ [%]
U-Net	0.810 ±0.108	N/A	N/A	N/A	N/A
UNet++	0.811 ±0.098	**0.026**	NS	N/A	N/A
AFFU	0.828 ±0.079	$\boldsymbol{p < 0.001}$	2.22	$\boldsymbol{p < 0.001}$	2.10
FFU	0.826 ±0.087	$\boldsymbol{p < 0.001}$	1.98	$\boldsymbol{p < 0.001}$	1.85
	1-RVE				
	Mean ±SD	p	Gain vs U-Net [%]	p^*	Gain vs UNet++ [%]
U-Net	0.832 ±0.151	N/A	N/A	N/A	N/A
UNet++	0.837 ±0.141	0.578	NS	N/A	N/A
AFFU	0.859 ±0.122	**0.001**	3.25	$\boldsymbol{p < 0.001}$	2.63
FFU	0.853 ±0.123	**0.002**	2.52	**0.006**	1.91
	HD				
	Mean ±SD	p	Gain vs U-Net [%]	p^*	Gain vs UNet++ [%]
U-Net	49.9 ±48.6	N/A	N/A	N/A	N/A
UNet++	44.7 ±43.1	0.294	NS	N/A	N/A
AFFU	29.9 ±26.5	$\boldsymbol{p < 0.001}$	40.1	$\boldsymbol{p < 0.001}$	33.1
FFU	32.4 ±35.7	$\boldsymbol{p < 0.001}$	35.1	$\boldsymbol{p < 0.001}$	27.5

Table 3.3 Average per-muscle performance of each model calculated from the cross-validation results. Each data recorded in the table represents the mean segmentation performance of each model for the same muscle across different validation samples. The values for the Hausdorff distance are rounded to one decimal place for improved readability.

	U-Net			UNet++			AFFU			FFU		
	DSC	1-RVE	HD [mm]	DSC	1-RVE	HD [mm]	DSC	1-RVE	HD [mm]	DSC	1-RVE	HD [mm]
RF	0.817	0.846	26.6	0.822	0.850	21.5	0.830	0.864	17.3	0.844	0.889	25.1
VI	0.809	0.910	47.5	0.817	0.924	31.5	0.825	0.907	29.1	0.825	0.900	34.0
VL	0.829	0.818	50.8	0.843	0.839	37.3	0.856	0.861	37.5	0.858	0.865	34.0
VM	0.848	0.833	55.2	0.851	0.829	35.6	0.872	0.858	30.6	0.874	0.862	43.4
SAT	0.788	0.782	43.6	0.789	0.788	43.1	0.822	0.840	16.0	0.817	0.833	16.9
SMB	0.808	0.814	51.7	0.807	0.853	43.4	0.819	0.866	48.2	0.811	0.843	38.8
SMT	0.776	0.789	47.4	0.785	0.785	37.1	0.804	0.812	35.7	0.798	0.806	33.6
GRA	0.718	0.768	22.0	0.719	0.766	23.3	0.752	0.853	17.8	0.728	0.792	17.4
BCB	0.800	0.775	83.2	0.785	0.758	74.5	0.809	0.791	23.9	0.815	0.789	35.8
BCL	0.811	0.832	62.1	0.818	0.829	42.0	0.847	0.859	25.7	0.845	0.859	52.0
AM	0.850	0.877	58.9	0.857	0.899	43.3	0.869	0.901	34.7	0.866	0.893	37.5
AB	0.753	0.833	34.4	0.753	0.858	41.3	0.761	0.848	21.4	0.759	0.838	18.1
AL	0.812	0.847	26.2	0.809	0.847	27.1	0.823	0.855	18.1	0.828	0.867	30.2
GM	0.901	0.882	69.3	0.900	0.884	60.1	0.907	0.906	61.1	0.908	0.900	44.4
IL	0.843	0.879	67.6	0.837	0.879	81.8	0.859	0.899	35.2	0.859	0.894	31.7
TFL	0.798	0.834	52.7	0.783	0.807	72.2	0.791	0.816	25.3	0.788	0.821	25.4

Overall, the AFFU was found to stand on the first place among all four models, producing the best average DSC (0.828 ± 0.079), RVE (0.859 ± 0.122) and HD (29.9 ± 26.5 mm) across 16 muscles. FFU was the second-best model with DSC (0.826 ± 0.087), RVE (0.853 ± 0.123) and HD (32.4 ±35.7 mm). On the test set, both models outperformed U-Net (and UNet++) with the following average improvements: 2.2% (2.1%) in DSC, 3.3% (2.6%) in 1-RVE, 40.1% (33.1%) in HD for AFFU; 2.0% (1.8%) in DSC, 2.5% (1.9%) in 1-RVE, 35.1% (27.5%) in HD for FFU. The mean metrics value of different muscles for each model are shown in the supplementary material (Table 3.3). Figure 3.5 shows the segmentation visualization results of some muscles, where it could be observed that the results obtained from AFFU and FFU, exhibit generally superior segmentation compared to U-Net and UNet++.

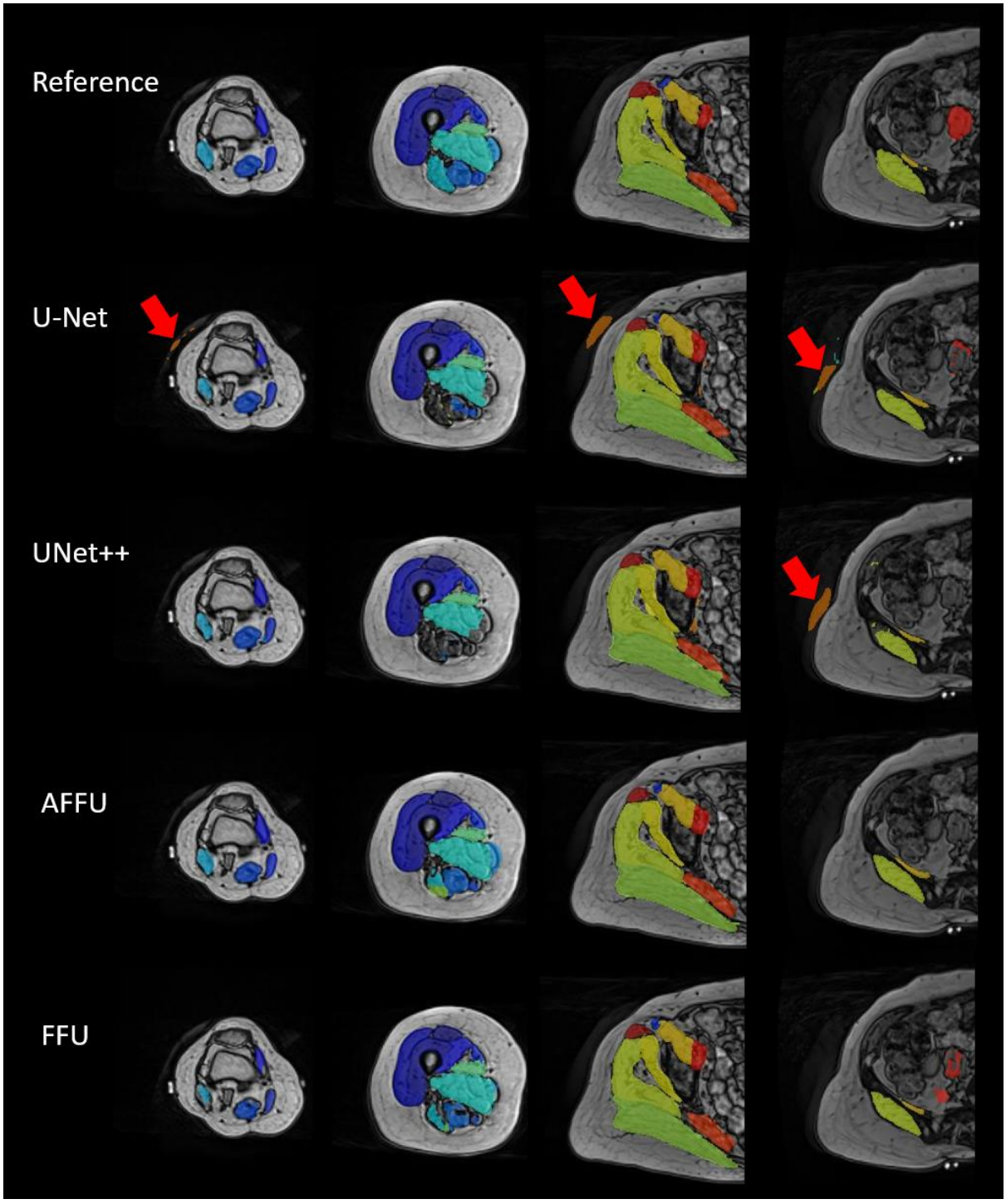


Figure 3.5 Segmentation visualization of muscles of one subject from prediction of each model and manual reference. The first row represents the gold standard generated by manual muscle segmentation by experienced operators. In the second and third rows, the performance of U-Net and UNet++ is depicted, wherein some local errors (red arrows) around the contour compared to the reference are noticeable, along with a few mis-segmented points. The fourth and fifth rows present the results obtained from AFFU and FFU, which exhibit generally superior segmentation compared to U-Net and UNet++.

3.3.2 Performance of the models on different cohorts after training on the PMW-1 cohort

In this section, three datasets (i.e., PMW-1, PMW-2, and PMW-OB cohorts), were used to evaluate the performance of the models after having trained the models on the data from the PMW-1 cohort. Table 3.4 and Figure 3.6 provide an overview of the average performance of each model, pooling together the results from different subsets (1*PMW-1, 8*PMW-2, 10*PMW-OB).

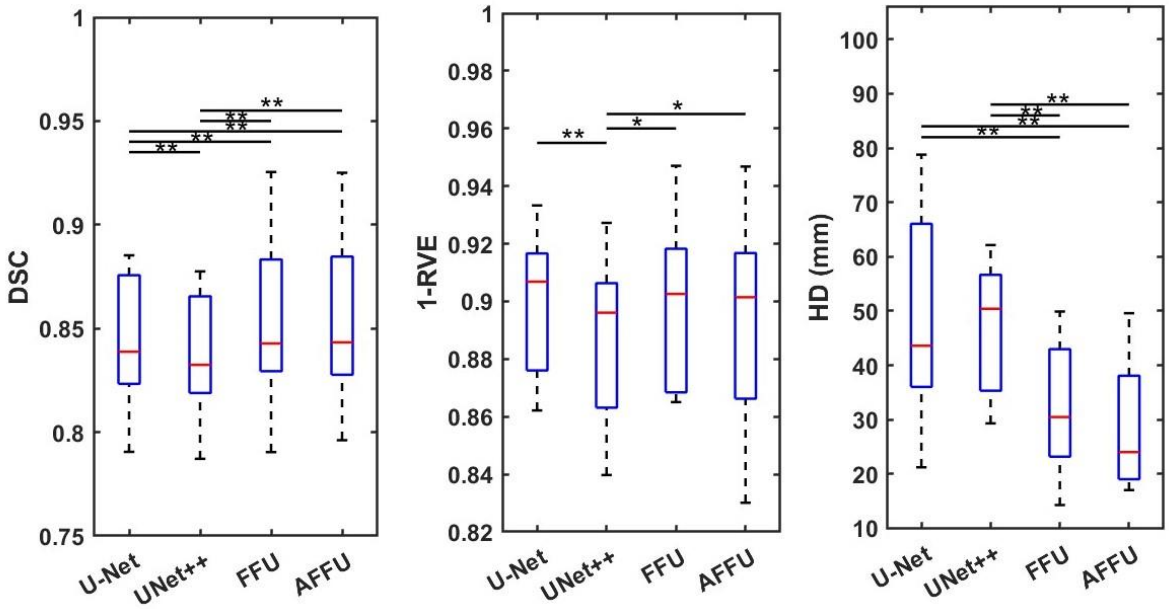


Figure 3.6. Performance of each model on all 19 test subjects. For each model and each metric, 16 mean values calculated from 304 predicted labels (16 labels for each of the 19 subjects) are shown (* indicates $p<0.05$, ** indicates $p<0.01$).

Table 3.4 Average model performance comparison including all test subjects. Significant differences are reported in bold.

	DSC				
	Mean ±SD	p	Gain vs U-Net [%]	p^*	Gain vs UNet++ [%]
U-Net	0.843 ±0.062	N/A	N/A	N/A	N/A
UNet++	0.837 ±0.067	$\mathbf{p < 0.001}$	-0.70	N/A	N/A
AFFU	0.848 ±0.058	$\mathbf{p < 0.001}$	0.59	$\mathbf{p < 0.001}$	1.31
FFU	0.848 ±0.058	$\mathbf{p < 0.001}$	0.59	$\mathbf{p < 0.001}$	1.31
	1-RVE				
	Mean ±SD	p	Gain vs U-Net [%]	p^*	Gain vs UNet++ [%]
U-Net	0.893 ±0.097	N/A	N/A	N/A	N/A
UNet++	0.884 ±0.103	$\mathbf{p < 0.001}$	-1.01	N/A	N/A
AFFU	0.894 ±0.096	0.674	NS	**0.014**	1.13
FFU	0.894 ±0.095	0.686	NS	**0.011**	1.13
	HD				
	Mean ±SD [mm]	p	Gain vs U-Net [%]	p^*	Gain vs UNet++ [%]
U-Net	49.1 ±52.8	N/A	N/A	N/A	N/A
UNet++	51.7 ±49.5	0.242	NS	N/A	N/A
AFFU	30.2 ±38.7	$\mathbf{p < 0.001}$	38.5	$\mathbf{p < 0.001}$	41.6
FFU	35.9 ±52.5	$\mathbf{p < 0.001}$	26.9	$\mathbf{p < 0.001}$	30.6

The AFFU and FFU models were more accurate in segmenting the muscles compared to the U-Net (0.59% and 0.59% higher DSC, respectively; 39% and 27% lower HD, respectively) and the UNet++ (1.31% and 1.31% higher DSC, respectively; 1.13% and 1.13% higher 1-RVE, respectively; 42% and 31% lower HD, respectively). The AFFU also showed a significant improvement compared to FFU on HD (15.9%, p=0.010). The full results for individual muscles and each model are shown in Table 3.5.

Table 3.5 Average per-muscle performance of each model calculated from all test subjects. The data in the table represent the average segmentation performance of the same muscle for each model across all test scans for part2.

	U-Net			UNet++			AFFU			FFU		
	DSC	1-RVE	HD [mm]	DSC	1-RVE	HD [mm]	DSC	1-RVE	HD [mm]	DSC	1-RVE	HD [mm]
RF	0.869	0.933	21.2	0.855	0.921	30.7	0.883	0.947	18.1	0.881	0.947	22.2
VI	0.791	0.813	45.7	0.787	0.816	32.6	0.796	0.805	38.4	0.790	0.799	29.5
VL	0.842	0.916	72.2	0.838	0.912	51.4	0.846	0.903	49.6	0.837	0.893	66.1
VM	0.882	0.910	57.2	0.877	0.895	32.9	0.887	0.915	37.7	0.887	0.907	38.2
SAT	0.822	0.908	41.5	0.821	0.901	55.5	0.840	0.890	21.9	0.840	0.903	25.0
SMB	0.844	0.886	62.3	0.832	0.869	46.1	0.844	0.866	33.0	0.843	0.865	29.3
SMT	0.836	0.906	37.1	0.833	0.902	37.7	0.840	0.918	31.6	0.841	0.918	45.3
GRA	0.832	0.895	30.0	0.831	0.871	29.3	0.840	0.900	21.4	0.842	0.900	24.2
BCB	0.812	0.862	35.0	0.804	0.857	90.2	0.816	0.867	19.9	0.821	0.866	33.5
BCL	0.885	0.911	37.5	0.876	0.910	57.8	0.886	0.906	26.1	0.887	0.909	49.9
AM	0.884	0.918	69.8	0.877	0.903	53.7	0.886	0.916	48.8	0.886	0.919	40.7
AB	0.775	0.827	32.0	0.767	0.840	44.8	0.770	0.830	17.0	0.770	0.816	17.9
AL	0.835	0.920	40.1	0.821	0.897	49.5	0.843	0.921	17.1	0.843	0.919	17.0
GM	0.919	0.928	72.9	0.920	0.927	62.1	0.925	0.957	65.3	0.925	0.964	90.4
IL	0.842	0.874	78.8	0.839	0.866	101.7	0.859	0.900	20.6	0.857	0.902	31.5
TFL	0.824	0.878	52.3	0.817	0.860	51.3	0.810	0.858	17.2	0.812	0.870	14.3

Table 3.6 Average performance of each model for PMW-2 and PMW-OB. The p-values in the table represent the significance results of the Wilcoxon signed-rank test conducted on muscles in all validation subjects between the models and U-Net or UNet++ about two different cohorts PMW-2 and PMW-OB. Significant differences are reported in bold ($p < 0.05$ *or* $p * < 0.05$).

	DSC (PMW-2)				
	Mean ±SD	p	Gain vs U-Net [%]	p^*	Gain vs UNet++ [%]
U-Net	0.828 ±0.073	N/A	N/A	N/A	N/A
UNet++	0.820 ±0.080	$\mathbf{p < 0.001}$	-0.97	N/A	N/A
AFFU	0.833 ±0.065	**0.008**	0.60	$\mathbf{p < 0.001}$	1.58
FFU	0.834 ±0.065	**0.006**	0.72	0.743	NS
	1-RVE (PMW-2)				
	Mean ±SD	p	Gain vs U-Net [%]	p^*	Gain vs UNet++ [%]
U-Net	0.880 ±0.105	N/A	N/A	N/A	N/A
UNet++	0.868 ±0.116	$\mathbf{p < 0.001}$	-1.36	N/A	N/A
AFFU	0.873 ±0.105	0.470	NS	0.264	NS
FFU	0.876 ±0.105	0.737	NS	0.166	NS
	HD (PMW-2)				
	Mean ±SD [mm]	p	Gain vs U-Net [%]	p^*	Gain vs UNet++ [%]
U-Net	45.0 ±47.6	N/A	N/A	N/A	N/A
UNet++	47.3 ±44.1	0.606	NS	N/A	N/A
AFFU	25.9 ±27.9	$\mathbf{p < 0.001}$	42.4	$\mathbf{p < 0.001}$	45.2
FFU	29.8 ±32.2	$\mathbf{p < 0.001}$	33.8	$\mathbf{p < 0.001}$	40.0
	DSC (PMW-OB)				
	Mean ±SD	p	Gain vs U-Net [%]	p^*	Gain vs UNet++ [%]
U-Net	0.857 ±0.049	N/A	N/A	N/A	N/A
UNet++	0.853 ±0.051	$\mathbf{p < 0.001}$	-0.47	N/A	N/A
AFFU	0.862 ±0.048	$\mathbf{p < 0.001}$	0.58	$\mathbf{p < 0.001}$	1.06
FFU	0.861 ±0.049	**0.006**	0.47	$\mathbf{p < 0.001}$	0.94
	1-RVE (PMW-OB)				
	Mean ±SD	p	Gain vs U-Net [%]	p^*	Gain vs UNet++ [%]
U-Net	0.914 ±0.072	N/A	N/A	N/A	N/A
UNet++	0.908 ±0.073	**0.003**	-0.66	N/A	N/A
AFFU	0.919 ±0.076	0.242	NS	**0.021**	1.21
FFU	0.917 ±0.075	0.484	NS	**0.046**	0.99
	HD (PMW-OB)				
	Mean ±SD [mm]	p	Gain vs U-Net [%]	p^*	Gain vs UNet++ [%]
U-Net	54.5 ±58.4	N/A	N/A	N/A	N/A
UNet++	57.4 ±54.8	0.446	NS	N/A	N/A
AFFU	34.8 ±46.8	$\mathbf{p < 0.001}$	36.1	$\mathbf{p < 0.001}$	39.4
FFU	42.5 ±65.7	$\mathbf{p < 0.001}$	22.0	$\mathbf{p < 0.001}$	26.0

The performance of each model for segmenting the muscles of the 8 subjects from the PMW-2 cohort and the 10 subjects from the PMW-OB cohort are reported in Table 3.6. In terms of

(1-RVE), there were no significant differences in accuracy of the models. Each model achieved high (1-RVE) values in the range 0.87-0.92, indicating accurate volume estimation. For DSC, AFFU demonstrated a gain of approximately 0.6% over U-Net in both groups ($p=0.008$ for PMW-2, $p<0.001$ for PMW-OB). FFU showed a significant improvement compared to U-Net for the PMW-2 (+0.72%, $p=0.006$) and PMW-OB (0.47%, $p = 0.006$) cohorts. For HD, both AFFU and FFU consistently exhibited smaller local errors compared to U-Net (22.0%-42.4%, $p<0.001$). On average, they achieved a reduction of approximately 20 mm in local distance, indicating improved localization accuracy.

All muscle segmentation results are reported in Table 3.7 and Table 3.8.

Table 3.7 Average per-muscle performance of each model calculated from eight healthy old women.

PMW-2												
	U-Net			UNet++			AFFU			FFU		
	DSC	1-RVE	HD [mm]	DSC	1-RVE	HD [mm]	DSC	1-RVE	HD [mm]	DSC	1-RVE	HD [mm]
RF	0.825	0.942	23.6	0.800	0.921	31.7	0.852	0.939	18.8	0.846	0.949	19.1
VI	0.780	0.845	34.3	0.775	0.844	32.8	0.780	0.856	31.7	0.773	0.845	32.4
VL	0.826	0.893	61.5	0.819	0.890	41.0	0.827	0.891	32.6	0.817	0.879	50.2
VM	0.861	0.858	34.6	0.856	0.841	35.5	0.865	0.862	21.2	0.867	0.851	21.9
SAT	0.796	0.899	33.5	0.801	0.888	41.8	0.813	0.854	19.2	0.817	0.870	21.8
SMB	0.817	0.834	71.3	0.793	0.792	36.5	0.830	0.854	22.1	0.827	0.846	26.9
SMT	0.829	0.903	38.4	0.822	0.894	27.9	0.823	0.919	29.5	0.828	0.925	29.8
GRA	0.830	0.935	24.9	0.831	0.917	24.4	0.828	0.874	20.5	0.836	0.889	17.9
BCB	0.788	0.819	34.3	0.775	0.795	93.4	0.788	0.806	22.0	0.798	0.810	22.0
BCL	0.874	0.909	45.3	0.863	0.900	60.4	0.870	0.885	22.8	0.874	0.897	29.7
AM	0.872	0.860	63.3	0.860	0.829	51.0	0.873	0.870	48.2	0.873	0.871	46.7
AB	0.774	0.771	35.5	0.770	0.781	49.5	0.767	0.783	16.1	0.772	0.773	16.6
AL	0.831	0.920	34.6	0.823	0.901	32.6	0.840	0.888	19.4	0.842	0.901	18.4
GM	0.920	0.943	68.8	0.920	0.945	52.7	0.925	0.968	54.4	0.924	0.971	72.2
IL	0.823	0.880	87.5	0.822	0.890	112.9	0.849	0.910	18.6	0.845	0.911	37.3
TFL	0.806	0.865	29.2	0.797	0.859	31.9	0.796	0.812	16.9	0.799	0.837	14.4

Table 3.8 Average per-muscle performance of each model calculated from ten old obese women. The data in Tables 3.7/3.8 represent the average performance evaluation results of each muscle for each model across different cohorts, PMW-2, and PMW-OB.

PMW-OB												
	U-Net			UNet++			AFFU			FFU		
	DSC	1-RVE	HD [mm]	DSC	1-RVE	HD [mm]	DSC	1-RVE	HD [mm]	DSC	1-RVE	HD [mm]
RF	0.903	0.924	20.3	0.897	0.917	31.9	0.911	0.959	18.0	0.911	0.949	25.7
VI	0.796	0.778	55.8	0.794	0.783	29.9	0.803	0.758	44.9	0.799	0.757	26.8
VL	0.855	0.957	84.2	0.853	0.953	61.1	0.861	0.926	65.3	0.853	0.919	83.1
VM	0.899	0.954	77.5	0.893	0.941	31.3	0.903	0.960	52.7	0.903	0.953	53.2
SAT	0.847	0.939	49.7	0.842	0.933	69.4	0.863	0.935	24.1	0.862	0.947	28.2
SMB	0.866	0.930	57.5	0.866	0.934	53.6	0.859	0.872	42.3	0.859	0.875	31.5
SMT	0.844	0.941	38.1	0.845	0.944	47.5	0.854	0.946	35.0	0.854	0.943	60.6
GRA	0.845	0.904	34.5	0.843	0.880	33.5	0.854	0.928	22.0	0.854	0.923	30.1
BCB	0.832	0.924	37.8	0.828	0.936	95.5	0.839	0.945	19.2	0.842	0.939	45.0
BCL	0.895	0.925	33.6	0.886	0.931	60.3	0.898	0.935	30.8	0.897	0.931	70.3
AM	0.899	0.965	78.8	0.896	0.963	58.0	0.900	0.958	50.8	0.899	0.962	37.1
AB	0.777	0.877	31.3	0.766	0.882	44.2	0.774	0.878	18.3	0.770	0.858	19.8
AL	0.835	0.916	46.8	0.817	0.887	66.3	0.842	0.945	15.2	0.841	0.932	15.6
GM	0.923	0.929	78.8	0.923	0.928	69.8	0.928	0.962	78.8	0.929	0.971	112.7
IL	0.854	0.873	75.2	0.850	0.852	97.4	0.865	0.894	22.4	0.865	0.897	27.6
TFL	0.848	0.884	72.7	0.842	0.858	68.7	0.834	0.906	16.9	0.835	0.909	13.1

3.3.3 Effect of augmentation data

The muscle volume distribution after the addition of the augmented data is reported in Figure 3.7, showing that the augmented data fills some gaps in the original data distribution. If tested on the all subjects pooled together, small significant improvements in DSC were observed by using the augmented training dataset ($p<0.001$, +0.6% for AFFU; $p<0.001$, +1.5% for U-Net, Table 3.9). However, when considering 1-RVE and HD, the performance remained almost unchanged (very small 0.9% improvement in RVE was observed for U-Net with the augmented dataset). The full results can be found in Table 3.10.

Table 3.9 Model comparison under three metrics with/without augmentation data on all test subjects pooled together. Significant differences are reported in bold. Column A/B represent the scores achieved by the models trained without/with augmented data.

		A	B	p	Gain [%]
DSC Mean ±SD [%]	U-Net	0.843 ±0.062	0.856 ±0.054	$\boldsymbol{p < 0.001}$	1.54
	AFFU	0.848 ±0.058	0.853 ±0.055	$\boldsymbol{p < 0.001}$	0.59
RVE Mean ±SD [%]	U-Net	0.893 ±0.097	0.901 ±0.091	$\boldsymbol{p < 0.001}$	0.90
	AFFU	0.894 ±0.096	0.897 ±0.094	0.286	NS
HD Mean ±SD [mm]	U-Net	49.1 ±52.8	43.5 ±51.3	$\boldsymbol{p < 0.001}$	NS
	AFFU	30.2 ±38.7	31.3 ±43.4	**0.002**	-3.64

Table 3.10 Average per-muscle performance of U-Net and AFFU calculated with and without augmentation data on all test scans. The results for each muscle represent the average values obtained from predictions on all test scans using the same model. Column A/B represent the scores achieved by the models trained without/with augmented data.

	U-Net						AFFU					
	DSC		1-RVE		HD (mm)		DSC		1-RVE		HD (mm)	
	A	B	A	B	A	B	A	B	A	B	A	B
RF	0.869	0.888	0.933	0.930	21.2	18.3	0.883	0.892	0.947	0.936	18.1	16.7
VI	0.791	0.800	0.813	0.814	45.7	52.8	0.796	0.797	0.805	0.792	38.4	27.3
VL	0.842	0.851	0.916	0.916	72.2	60.8	0.846	0.844	0.903	0.909	49.6	48.0
VM	0.882	0.893	0.910	0.909	57.2	39.1	0.887	0.894	0.915	0.915	37.7	31.1
SAT	0.822	0.843	0.908	0.897	41.5	38.2	0.840	0.846	0.890	0.903	21.9	20.9
SMB	0.844	0.864	0.886	0.919	62.3	35.4	0.844	0.857	0.866	0.897	33.0	27.4
SMT	0.836	0.855	0.906	0.921	37.1	42.1	0.840	0.853	0.918	0.913	31.6	31.2
GRA	0.832	0.845	0.895	0.920	30.0	23.2	0.840	0.843	0.900	0.914	21.4	20.1
BCB	0.812	0.824	0.862	0.855	35.0	35.4	0.816	0.824	0.867	0.865	19.9	19.4
BCL	0.885	0.892	0.911	0.909	37.5	44.1	0.886	0.894	0.906	0.919	26.1	24.0
AM	0.884	0.890	0.918	0.913	69.8	75.3	0.886	0.886	0.916	0.916	48.8	42.6
AB	0.775	0.787	0.827	0.855	32.0	22.4	0.770	0.770	0.830	0.778	17.0	16.4
AL	0.835	0.851	0.920	0.939	40.1	18.2	0.843	0.843	0.921	0.937	17.1	17.4
GM	0.919	0.925	0.928	0.936	72.9	92.8	0.925	0.925	0.957	0.954	65.3	96.6
IL	0.842	0.852	0.874	0.898	78.8	62.9	0.859	0.858	0.900	0.907	20.6	39.4
TFL	0.824	0.836	0.878	0.877	52.3	35.5	0.810	0.821	0.858	0.892	17.2	22.6

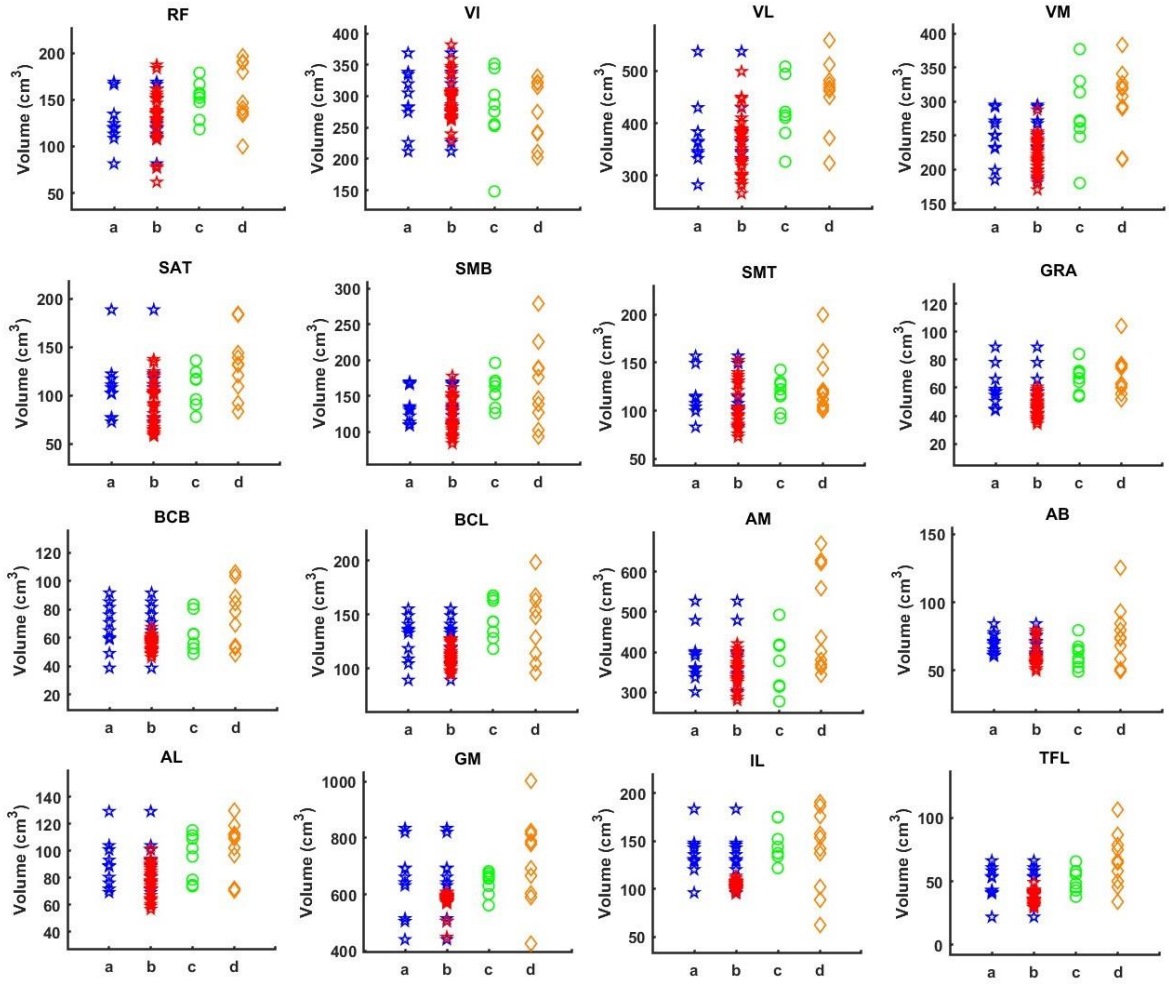


Figure 3.7. Muscle volume distribution. The figure illustrates the volume distribution of 16 muscles in four sets: 10 PMW-1 training samples (column a, blue), training samples with added augmentation data (column b, blue and red), PMW-2 (column c, green), and PMW-OB (column d, orange) cohorts. In each column (b), the red dots represent the distribution of the additional segmented subjects, while the blue dots represent the values from the original scans.

U-Net demonstrated a noticeable increase in the DSC for all three cohorts (PMW-1: 1.2%, $p=0.006$; PMW-2:2.2%, $p<0.001$; PMW-OB:1.1%, $p<0.001$, Table 3.9). On the other hand, AFFU shows a small significant increase in the DSC only for PMW-2 (1.1%, $p<0.001$, Table 3.9). Significant improvements in RVE, and HD for the PMW-2 cohort were observed (+1.93% for U-Net and +1.6% for AFFU in RVE, and +19.8% for U-Net in HD, Table 3.14). In contrast, in PMW-OB, only the DSC metric showed significant improvement, while RVE and HD remained relatively unchanged. The results obtained for single muscles are reported in Table 3.11, 3.12 and 3.13.

Table 3.11 Average per-muscle performance of U-Net and AFFU with/without augmentation data on one post-menopausal woman. Since only one test scan was included, each data point in above table represents the mean value of three repetitions from the same test subject. Column A/B represent the scores achieved by the models trained without/with augmented data.

	PMW-1											
	U-Net						AFFU					
	DSC		1-RVE		HD (mm)		DSC		1-RVE		HD (mm)	
	A	B	A	B	A	B	A	B	A	B	A	B
RF	0.874	0.891	0.955	0.937	11.9	10.7	0.854	0.887	0.882	0.985	13.8	12.0
VI	0.817	0.820	0.900	0.922	35.3	39.8	0.853	0.853	0.861	0.866	27.9	21.5
VL	0.834	0.843	0.678	0.690	37.6	25.7	0.856	0.863	0.759	0.755	28.4	25.8
VM	0.880	0.899	0.880	0.865	35.1	25.9	0.898	0.899	0.893	0.884	19.3	15.9
SAT	0.786	0.790	0.660	0.556	24.7	21.2	0.821	0.798	0.731	0.621	21.6	16.0
SMB	0.837	0.839	0.869	0.895	38.3	33.4	0.810	0.817	0.903	0.938	27.8	24.0
SMT	0.811	0.804	0.585	0.572	16.7	16.3	0.820	0.813	0.624	0.632	14.5	14.0
GRA	0.725	0.746	0.495	0.468	26.1	25.0	0.791	0.781	0.834	0.727	21.6	21.3
BCB	0.799	0.804	0.585	0.613	12.6	22.6	0.802	0.806	0.563	0.602	10.9	11.5
BCL	0.881	0.884	0.782	0.760	13.4	11.6	0.892	0.884	0.794	0.767	6.3	6.2
AM	0.836	0.853	0.901	0.863	31.3	32.2	0.853	0.856	0.861	0.863	34.1	32.7
AB	0.764	0.777	0.774	0.820	11.6	9.6	0.758	0.750	0.727	0.697	11.4	10.1
AL	0.858	0.867	0.963	0.948	17.4	13.9	0.870	0.859	0.941	0.990	17.4	13.2
GM	0.881	0.888	0.787	0.806	47.3	16.0	0.893	0.893	0.822	0.829	17.5	14.4
IL	0.866	0.857	0.834	0.813	43.9	32.6	0.879	0.874	0.870	0.853	19.5	25.9
TFL	0.738	0.782	0.927	0.960	32.6	21.1	0.686	0.743	0.747	0.890	23.4	19.7

Table 3.12 Average per-muscle performance of U-Net and AFFU with/without augmentation data on ten old obese women. Each data point represents the mean performance of each muscle for a particular model across all individuals within the same cohort.

	PMW-OB											
	U-Net						AFFU					
	DSC		1-RVE		HD (mm)		DSC		1-RVE		HD (mm)	
	A	B	A	B	A	B	A	B	A	B	A	B
RF	0.903	0.911	0.924	0.925	20.3	18.0	0.911	0.912	0.959	0.941	18.0	16.4
VI	0.796	0.806	0.778	0.775	55.8	69.1	0.803	0.804	0.758	0.754	44.9	27.3
VL	0.855	0.862	0.957	0.945	84.2	69.9	0.861	0.856	0.926	0.926	65.3	60.3
VM	0.899	0.905	0.954	0.947	77.5	51.1	0.903	0.910	0.960	0.954	52.7	38.5
SAT	0.847	0.864	0.939	0.928	49.7	50.2	0.863	0.867	0.935	0.950	24.1	24.9
SMB	0.866	0.875	0.930	0.940	57.5	39.9	0.859	0.870	0.872	0.905	42.3	32.7
SMT	0.844	0.863	0.941	0.951	38.1	44.9	0.854	0.863	0.946	0.935	35.0	33.7
GRA	0.845	0.858	0.904	0.942	34.5	28.0	0.854	0.850	0.928	0.930	22.0	19.0
BCB	0.832	0.838	0.924	0.901	37.8	35.3	0.839	0.847	0.945	0.919	19.2	17.1
BCL	0.895	0.897	0.925	0.917	33.6	63.6	0.898	0.901	0.935	0.938	30.8	27.5
AM	0.899	0.903	0.965	0.960	78.8	77.3	0.900	0.896	0.958	0.951	50.8	40.2
AB	0.777	0.787	0.877	0.888	31.3	21.5	0.774	0.771	0.878	0.791	18.3	17.3
AL	0.835	0.847	0.916	0.937	46.8	20.7	0.842	0.840	0.945	0.951	15.2	18.7
GM	0.923	0.928	0.929	0.936	78.8	114.8	0.928	0.929	0.962	0.958	78.8	137.7
IL	0.854	0.861	0.873	0.893	75.2	67.3	0.865	0.864	0.894	0.896	22.4	38.6
TFL	0.848	0.855	0.884	0.868	72.7	53.3	0.834	0.838	0.906	0.911	16.9	28.6

Table 3.13 Average per-muscle performance of U-Net and AFFU with/without augmentation data on eight old healthy women. Each data point represents the mean performance of each muscle for a particular model across all individuals within the same cohort.

	PMW-2											
	U-Net						AFFU					
	DSC		1-RVE		HD (mm)		DSC		1-RVE		HD (mm)	
	A	B	A	B	A	B	A	B	A	B	A	B
RF	0.825	0.858	0.942	0.935	23.6	19.7	0.852	0.868	0.939	0.925	18.8	17.6
VI	0.780	0.790	0.845	0.849	34.3	34.1	0.780	0.782	0.856	0.831	31.7	27.9
VL	0.826	0.838	0.893	0.908	61.5	53.8	0.827	0.826	0.891	0.907	32.6	35.3
VM	0.861	0.876	0.858	0.868	34.6	25.7	0.865	0.874	0.862	0.870	21.2	23.7
SAT	0.796	0.822	0.899	0.902	33.5	25.5	0.813	0.824	0.854	0.879	19.2	16.6
SMB	0.817	0.852	0.834	0.896	71.3	30.0	0.830	0.846	0.854	0.881	22.1	21.2
SMT	0.829	0.852	0.903	0.926	38.4	41.8	0.823	0.846	0.919	0.920	29.5	30.2
GRA	0.830	0.842	0.935	0.950	24.9	16.9	0.828	0.841	0.874	0.916	20.5	21.3
BCB	0.788	0.810	0.819	0.828	34.3	37.2	0.788	0.797	0.806	0.830	22.0	23.3
BCL	0.874	0.886	0.909	0.919	45.3	23.7	0.870	0.888	0.885	0.914	22.8	21.9
AM	0.872	0.879	0.860	0.861	63.3	78.2	0.873	0.878	0.870	0.877	48.2	46.9
AB	0.774	0.789	0.771	0.818	35.5	25.0	0.767	0.770	0.783	0.771	16.1	16.1
AL	0.831	0.852	0.920	0.941	34.6	15.6	0.840	0.844	0.888	0.912	19.4	16.3
GM	0.920	0.926	0.943	0.952	68.8	74.8	0.925	0.924	0.968	0.965	54.4	55.4
IL	0.823	0.840	0.880	0.915	87.5	61.2	0.849	0.849	0.910	0.926	18.6	42.1
TFL	0.806	0.820	0.865	0.878	29.2	15.0	0.796	0.810	0.812	0.869	16.9	15.5

Augmenting the training data has enabled U-Net to learn additional sample features, thereby enhancing its performance and decreasing the likelihood of segmentation errors, for example by effectively filling the muscle volume that were previously not well identified (Figure 3.8).

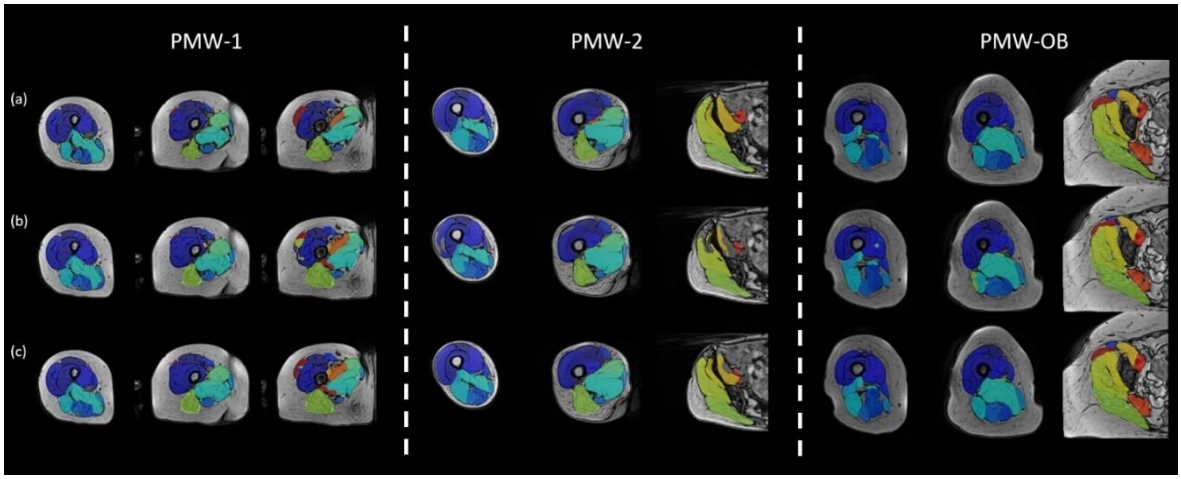


Figure 3.8 Segmentation performance visualization of U-Net trained without/with augmentation data. The figure exhibits the segmentation performance of U-Net on a subject chosen from each of the three cohorts (PMW-1/2/OB). (a) displays the gold standard manual segmentation, while (b) and (c) present the results for the U-Net model trained without or with augmented data, respectively.

Table 3.14 Comparison of models across three metrics for different cohorts trained with augmentation data. Significant differences are reported in bold.

		PMW-1			
		Without aug.	With aug.	p	Gain [%]
DSC (Mean ±SD [%])	U-Net	0.824 ±0.051	0.834	**0.006**	1.21
	AFFU	0.833 ±0.056	0.836	0.796	NS
RVE (Mean ±SD [%])	U-Net	0.786 ±0.145	0.780	0.918	NS
	AFFU	0.801 ±0.104	0.806	0.717	NS
HD (Mean ±SD [mm])	U-Net	27.2 ±12.1	22.4 ±9.0	**0.020**	17.6
	AFFU	19.7 ±7.5	17.8 ±7.0	**0.011**	9.6
		PMW-2			
		Without aug.	With aug.	p	Gain [%]
DSC (Mean ±SD [%])	U-Net	0.828 ±0.073	0.846	$\boldsymbol{p < 0.001}$	2.17
	AFFU	0.833 ±0.065	0.842	$\boldsymbol{p < 0.001}$	1.08
RVE (Mean ±SD [%])	U-Net	0.880 ±0.105	0.897	$\boldsymbol{p < 0.001}$	1.93
	AFFU	0.873 ±0.105	0.887	$\boldsymbol{p < 0.001}$	1.6
HD (Mean ±SD [mm])	U-Net	45.0 ±47.6	36.1	$\boldsymbol{p < 0.001}$	19.8
	AFFU	25.9 ±27.9	27.0	0.065	NS
		PMW-OB			
		Without aug.	With aug.	p	Gain [%]
DSC (Mean ±SD [%])	U-Net	0.857 ±0.049	0.866	$\boldsymbol{p < 0.001}$	1.05
	AFFU	0.862 ±0.048	0.864	0.097	NS
RVE (Mean ±SD [%])	U-Net	0.914 ±0.072	0.916	0.472	NS
	AFFU	0.919 ±0.076	0.913	0.067	NS
HD (Mean ±SD [mm])	U-Net	54.5 ±58.4	51.6	0.380	NS
	AFFU	34.8 ±46.8	36.2	0.071	NS

3.3.4 Analysis for individual muscles

The PMW-OB cohort had generally higher mean muscle volumes for the muscles with volume in the high range (VM, VL, AM, GM) compared to the other two cohorts. The muscles were ranked in order of volume for the PMW-1 cohort (mean volume: 189.9 ±162.5 cm^3). Compared with PMW-1 and PMW-2 (197.2 ±162.8 cm^3) cohorts, some muscles such as VL, AM and GM in the PMW-OB (217.1 ±185.9 cm^3) cohort showed higher values (Figure 3.9).

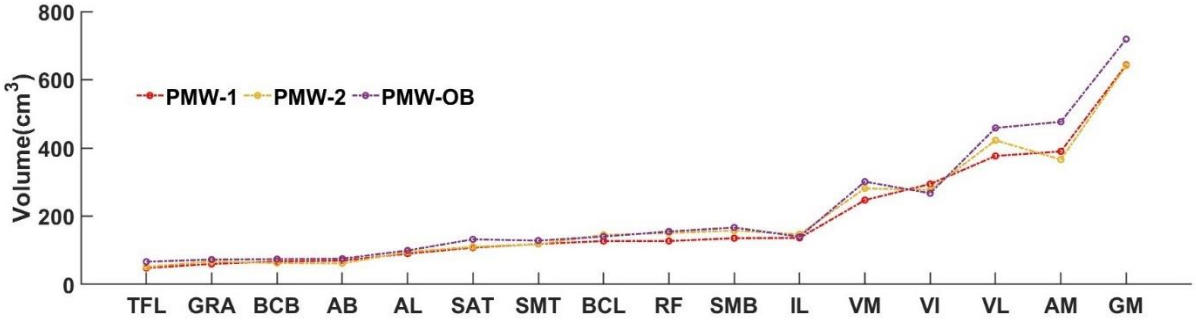


Figure 3.9 Individual muscle volume comparison among PMW-1/2/OB cohorts.

Overall, a trend of increasing DSC for all models was observed as muscle volume increased (Figure 3.10). In particular, for the PMW-1 and PMW-2 cohorts, muscles with volume in the middle range were better segmented by AFFU and FFU compared to U-Net (improvement of 0.72%-4.71%(PMW-1), 0.17%-3.23%(PMW-2) for AFFU and 0.78%-4.19%(PMW-1),0.18%-2.66%(PMW-2) for FFU, for most muscles with volume in the middle or middle-high range (arrows in Figure 3.10). Similar trends were found for the PMW-OB cohort, with AFFU that led to better predictions than U-Net for most muscles with volume in the middle-low range (improvement of 0.33%-1.87%). For the PMW-1 cohort a not-significant trend of increased accuracy in DSC for larger muscles was found (Figure 3.10). Data augmentation affected similarly the segmentation of the individual muscles, with main gains for the U-Net and AFFU model applied to the PMW-2 cohort (Figure 3.11).

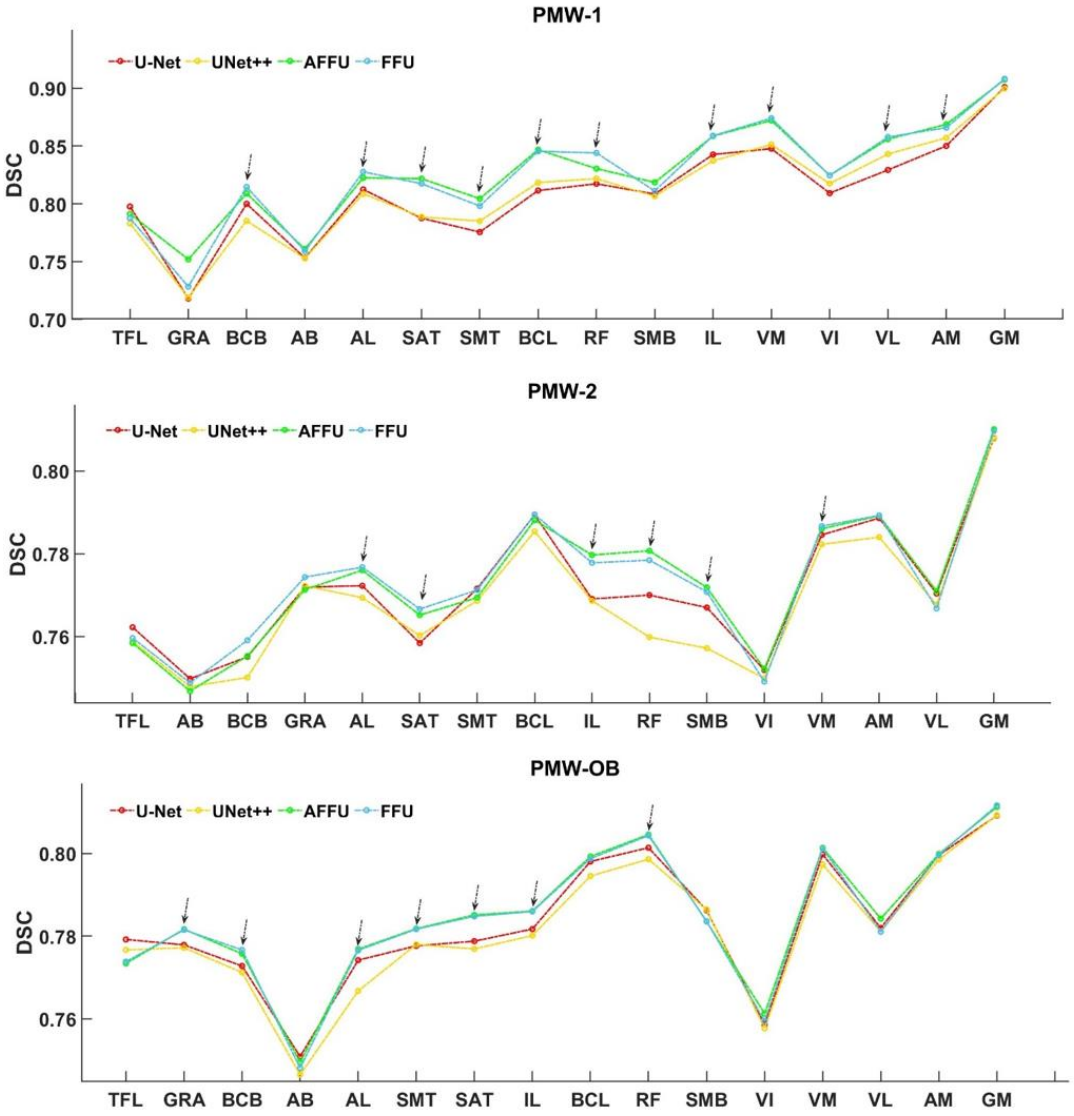


Figure 3.10 Relationship between muscle volume and DSC index in women cohorts. Muscle volume and DSC for the three different cohorts of subjects. The muscles were arranged in ascending order according to the mean muscle volume of each specific cohort, and each point in the figure represents the average DSC value of the muscle across the subjects.

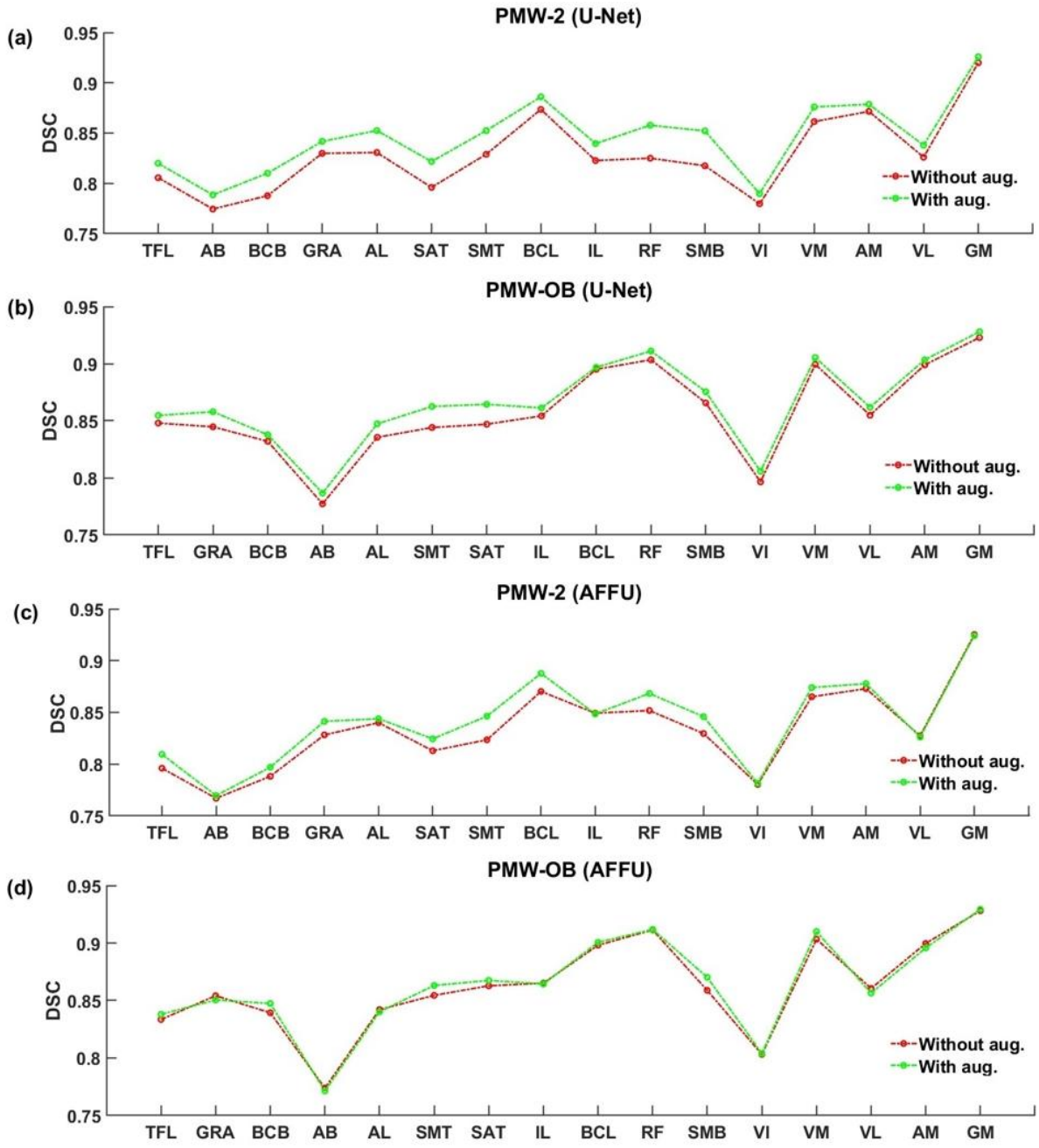


Figure 3.11 Per-muscle evaluation trained with/without augmentation of U-Net (a-b) and AFFU (c-d). The plots display the DSCs before and after data augmentation.

3.4 Discussion

This chapter aimed to evaluate the performance of different DL models to automatically segment individual muscles of the lower limbs from the MRI images of different cohorts of post-menopausal women. The new models proposed in this section, that include the fusion part (FFU) and multi-added attention mechanism (AFFU), were found to be significantly more accurate than the standard U-Net and UNet++ models. The proposed SSM based data augmentation method also enhanced the segmentation accuracy of the models trained on small datasets.

AFFU achieved average DSC scores of 0.83-0.85 depending on the training dataset, providing a reasonable automatic segmentation of most muscles even if trained on different cohorts of women. In this chapter, models were trained on same dataset (PMW-1) and tested on different cohorts (PMW-1/-2/OB). Higher accuracy was found for previous automatic segmentation tools trained and tested on young cohorts (DSC of 0.85 and 0.91 in [2] and [17], respectively). The difference may be due to differences between this study and previous studies: fewer training datasets were used in this study (11 subjects versus 20-60 subjects [2], [17]); more complex image texture was processed for the tested cohorts including post-menopausal women and obese post-menopausal women compared to the younger healthy subjects in those studies; with muscle loss with aging, structures like tendons, ligaments and blood vessels become more prominent which typically shows dark area on T1-weighted images [91], [92]; training was performed on the PMW-1 cohort and then tested on the other two cohorts, evaluating the possible application of the models if not trained on the same cohort.

Model AFFU and FFU showed better performance than U-Net and UNet++ when tested on different cohorts. This effect indicates that feature fusion part and attention mechanisms could learn more features than U-Net in a small sample size (n~10) and the fusion of features from layers 2/3/4 into U-Net could enhance its learning capability. Moreover, the addition of attention mechanisms led to more robust performance across different samples and in some cases, attention mechanisms have led to accuracy improvements of up to 20% compared to state-of-the-art non attention models [93].

This improvement may be because the attention mechanism allows the model to notice more general features, rather than information that is specific to post-menopausal women. The results also indicate that the proposed models are more resistant to noise interference

compared to the standard models. Although the differences in DSC may not be significant for some results, the HD, that is associated with local errors, showed a significant improvement of approximately 15-25mm compared to the benchmark. As the muscle MR scans approaches the hip and its image texture becomes more complex, due to the factor that multiple muscle group, e.g. hip flexors, extensors and adductors, converge in this area, their varying signal intensities of different tissues introduce more complicated overall image texture [94].

There is a discernible rise in the occurrence of mis-segmentation areas within the U-Net predictions (Figure 3.5). These areas which indicated by red arrows in Figure 3.5 can be attributed to noise interference, underlining the need for further investigation into the factors influencing the segmentation accuracy. As can be observed in Figures 3.8 and 3.9, there were a few fat infiltrations in the muscles of the PMW-OB cohort, which were generally larger than those of the other two cohorts for the larger muscles. Assuming that the neural network has the same ability to identify the target region during segmentation, the impact of noise on parameters such as DSC and RVE is less obvious in numerical values when working on larger muscles. This shows that on some fat or edge treatments, the model still does not learn relevant features from the PMW-1 cohort to the PMW-OB cohort. Even so, the results of the model on PMW-OB (U-Net: DSC=0.857 ±0.049; AFFU: DSC=0.862 ±0.048) cohort are comparable compared to segmentation results on younger subjects in the literature [2], [17]. U-Net and AFFU exhibited generally lower scores (Figure 3.10) for the same muscles in the PMW-OB cohort compared to PMW-1. This indicates that a higher proportion of fat might have a larger impact on noise in segmentation, leading to larger local errors (higher HD). However, this should be confirmed with further analysis to determine if the segmentation accuracy is influenced by the training with data from different cohorts.

When training data augmentation based on the PMW-1 cohort was used, improvements were found mainly for U-Net tested on the PMW-1 cohort, as expected. The improvement of the AFFU and FFU models was not particularly obvious when tested on the PMW-2 and PMW-OB cohorts. This effect could be due to the simpler structure of U-Net, which lacks attention mechanisms or other powerful features learning structures, leading to its higher sensitivity to the quantity of training data. In [95], U-Net performance was improved by 30% compared to baseline after training on larger datasets. In contrast, the model with attention mechanisms can learn more features and information with relatively fewer subjects. Deep learning is a supervised learning approach, with the amount of data often determining the depth of the model's learning. The U-Net and AFFU models achieved a similar level of accuracy in

segmenting the muscles after training with augmentation data, indicating that the models learned as much relevant information as possible from the dataset. However, the relationship between training dataset size and model accuracy is non-linear [96]. Moreover, considering that the training dataset was based only on the data from the PMW-1 cohort, the improvement in accuracy for the PMW-2 and PMW-OB cohorts is probably due to the inclusion in the enlarged dataset of muscle features also beneficial for the other cohorts. Nevertheless, it remains to be investigated what would happen if the augmentation would be based also on the PMW-2 and PMW-OB cohorts, which will be tested in future work.

The DSC performance was positively correlated with muscle volume for most models and muscles (except AB and VI). This trend shows that the automatic segmentation performance of the model is generally better for larger muscles. However, more variability was found for the testing cohorts PMW-2 and PMW-OB. This result is important, especially if the segmentation of the models is used to assess the function of large muscles due to musculoskeletal diseases [97], or for creating biomechanical applications of personalized multi-body dynamics models[98], applications for which it is more important to assess the properties of large muscles. Nevertheless, for improving the applicability of the models, they should be better optimized for automatically segmenting smaller muscles, such as TFL and GRA, where there is more room for improvement. This trend is not reflected in the literature [17], due to the fact that there is less difference between the muscles of the young athlete sample compared to the older women sample, and with higher image definition. However, the variability is also found generally higher in small volume muscles like quadratus femoris (DSC = 0.81 ±0.057) or muscles with irregular shapes in [17].

The technique developed in this section lies in the feature fusion and attention fusion, which has been demonstrated to be highly beneficial for muscle segmentation of post-menopausal women compared to the classical U-Net [81]. Moreover, it is noteworthy that the proposed models are generic and modular as such it can be easily applied to natural/medical image segmentation, classification, as well as regression problems.

AFFU/FFU spent around 10 hours in the training period, but less than 0.5 seconds per MR image on the test period. When using a well-trained model to make a new subject prediction, it takes less to segment a subject compared to the manual segmentation [6], [9].

The study in this chapter is affected by some limitations. Firstly, it is evident that despite using data augmentation techniques, the limited quantity of data from real subjects still significantly affects model training. Although no overfitting was observed during the training phase by monitoring the loss function's decline, training with a larger number of independent real subjects should be further considered in future studies also included the cross-cohort generalization analysis. A second limitation relates to the proposed novel data augmentation approach, which uses only the feature (mode 1) with the highest variation from the PCA analysis. In future studies, incorporating more features and extending the variability of those may lead to the creation of realistic subjects across a wide range of volumes and shapes, with the potential of improving the segmentation accuracy of the models also for different cohorts of subjects. Lastly, after visualizing the model's segmentation results, it became apparent that some muscles exhibited noticeable segmentation errors when observed in 3D. These segmentation errors could potentially be mitigated through post-processing approaches. Since muscles generally exhibit specific geometric shapes, leveraging geometric models in post-processing might alleviate some limitations inherent in the model itself.

3.5 Conclusion

This chapter has shown that using multi attention mechanisms and fusion structure improve the automatic segmentation of muscles from MR images from different cohorts of post-menopausal women, compared to standard U-Net models. Moreover, it has been shown that these models are less affected by the number of used training datasets, compared to the U-Net model, which benefited from the augmentation of the training dataset.

To optimizing this approach for further clinical research, the next step of the analysis should focus on investigating the potential of models' cross-cohort generalizations performance on different cohort, e.g. young people, children, which will be done and shown in the next chapter.

Chapter 4 Generalization study of AFFU on different cohorts

4.1 Introduction

In previous studies (Chapter 3), the automatic muscle segmentation is proved to be crucial for enhancing the precision and efficiency of muscle analysis in medical imaging, enabling more accurate related disease diagnosis, and the modified U-Net based DL model, AFFU (Attention-Feature-Fusion-U-Net) has been found to more accurately segment human muscles from a post-menopausal women cohort than popular benchmark DL models, U-Net and UNet++. In particular, it has showed stronger robustness against the noise which affected by HD score. However, the generalization of AFFU and other baseline models was not evaluated and analysed in different subjects from different cohorts (e.g. different ages, gender, genetic background). Robust generalization allows models to be applied across different age cohorts, imaging devices and patients under various condition[16], making them more clinically viable. Moreover, models with good generalization ability could reduce the cost for retraining of fine-tuning on new dataset [15] when the new data comes from different or other unseen populations, the well-trained models can be directly used without retraining or fine-tuning of parameters.

Muscles in different age groups, e.g. between young children and old women, exhibit obvious physiological differences, which poses a challenge for the generalization performance of a data driven model as DL models. A model that performs well on data from old women cohorts does not necessarily exhibit the same performance on children.

Accurate lower-limb muscle segmentation in children is crucial, some diseases like cerebral palsy, sarcopenia and neurological disorders can cause abnormal muscle growth with muscle loss or motion function abnormity where accurate segmented label could aid clinicians in monitoring the progression of the disease during early diagnosis [2], [14], [77]. However, their smaller muscle size [7] makes precise segmentation more challenging. Additionally, the rapid growth in children introduces greater variability between subjects, and their muscle boundaries are often less defined compared to adults [14], further complicating the segmentation process. The less defined boundaries in children can be attributed to several developmental and physiological factors [99], [100], [101]: Children generally have lower levels of intramuscular fat and less developed connective tissue, which contribute to reduced contrast between muscle and surrounding tissues; additionally, children's lower muscle echo intensity (EI) and smaller muscle size can limit the clarity of muscle boundaries in imaging; physiological factors, such as different muscle activation patterns and ongoing development

of muscle quality, further contribute to this effect. As children grow, these factors evolve, leading to more distinct muscle boundaries in adulthood. Some DL models have been applied on these cohort, e.g. Zhu et al. applied a deep learning (DL) model to segment the lower limb muscles in children (includes 14 male, 6 female with mean age: 9.0 ± 3.2 years old, height: 133.2 ± 22.1 cm, weight: 32.3 ± 16.5 kg.), including 6 of those with cerebral palsy, achieving an average DSC of 0.88.

However, few studies have integrated data from different age groups for model training and testing or performed reproducibility analysis on muscle segmentation [102] during the manual segmentation process used to establish the gold standard. Reproducibility analysis can offer more reliable reference points for assessing the performance of DL models across various muscle groups.

The first aim of Chapter 4, therefore, is to investigate the inter- and intra- operator's reproducibility of manual segmentation on typical developed children cohort (TDC). The training of DL model relies on the ground truth from manual segmentation. After analysis, the segmented data is used as ground truth to identify and calibrate the accuracy of the models and to select which muscles to select in assessment process [6], [103].

The second aim of this chapter is to explore the cross-cohort generalization potential of AFFU, as proposed in the previous section, alongside other popular benchmark models (U-Net, UNet++, and Attention-UNet), different models could be more or less accurate depending on the considered cohort. The dataset spans a wide range of cohorts, from children to young adults and older adults. During the generalization evaluation, we will examine the feature learning capabilities of each model and assess whether a model performs significantly better for one cohort over another. This analysis will help determine if the model's performance varies depending on the specific training dataset used, potentially indicating issues such as overfitting or underfitting.

4.2 Materials and methods

4.2.1 Participants

The subjects of the study included 9 typical developed children (TDC), 8 healthy young adults (HY), and 11 healthy post-menopausal women (PMW).

The T1-weighted TDC images were collected using a Siemens Magnetom Vida Machine (Siemens, Erlangen Germany), with an echo time of 2.46 ms, repetition time of 5.44 ms, flip angle of 9 degrees. The data was collected with an ethics approval from ethics committee of University of Vienna (reference number 00587). All scans are recorded with the Dixon method and are a composition (done by the MRI machine automatically, pixel size 1.00x1.00x1.00 mm^3, slice spacing 1.00 mm) of 3 single scans from coronal plane. First scan includes hip and proximal femur, second scan includes most of the femur, knee and proximal tibia and third scan includes tibia and feet.

The PMW subjects selected for model training and analysis were the same as those used in Chapter 3.

The HY was collected and anonymized as part of the MultiSim project (project title: 'Clinical data to inform the MultiSim Project: Development of a modeling framework focused on the human musculoskeletal system', STH Project Number: STH19009, REC Number: 16/EE/0049). The scanning protocol was the same that operators used in the cohort PMW (details in section 3.2.1).

4.2.1.1 Cohort statistics

The detailed parameters of the three cohorts are given below (Table 4.1).

Table 4.1 Statistical information

Cohort	Age (year)	Weight (kg)	Height (cm)	BMI
TDC (N=9) (4F/5M)	12.1±0.7 (10.8 to 13.0)	42.9±9.6 (34.2 to 62.3)	158±9 (142 to 168)	17.1±2.9 (14.9 to 23.2)
HY (N=8) (4F/4M)	26.4±3.5 (22.0 to 31.0)	63.1±8.4 (53.0 to 76.0)	171±10 (160 to 188)	21.4±1.4 (19.0 to 23.9)
PMW (N=11) (11F)	69.0±6.7 (59.1 to 83.0)	66.9±7.7 (56.8 to 78.6)	159±3 (154 to 164)	26.5±3.4 (21.2 to 31.1)

In the table, four parameters for each cohort are presented as mean ± SD (range: min to max). The number of subjects considered in each cohort, along with gender distribution, is indicated as N=x and xF/yM (females/males).

4.2.1.2 Manual segmentation and operator variability analysis of TDC cohort

The operation process of manual segmentation and inter(intra)-operator variability analysis on PMW [6] was described in Section 3.2.1.

The overview of the methods used in this section was presented in Figure 4.1. The MR images of TDC were segmented by three operators: one PhD student experienced in manual muscle segmentation, and two Master students who had no previous experience of muscle segmentation. The two master students were trained on at least 15 subjects using Armira (version 2022.2; Thermofisher) to get used to the image segmentation process and the muscle anatomy. In each right limb, 25 muscles of the thigh, from knee to hip, were segmented. Three subjects were randomly selected for the reproducibility study. Each operator performed three independent segmentations on right thigh part of each selected subject. To avoid the influence of previous operation, the three segmentation operations on the same sample were not performed consecutively. The muscle segmentation on both sides of the thigh part of all 9 subjects was performed by the same operator (one PhD student) and used as the ground truth in further model training.

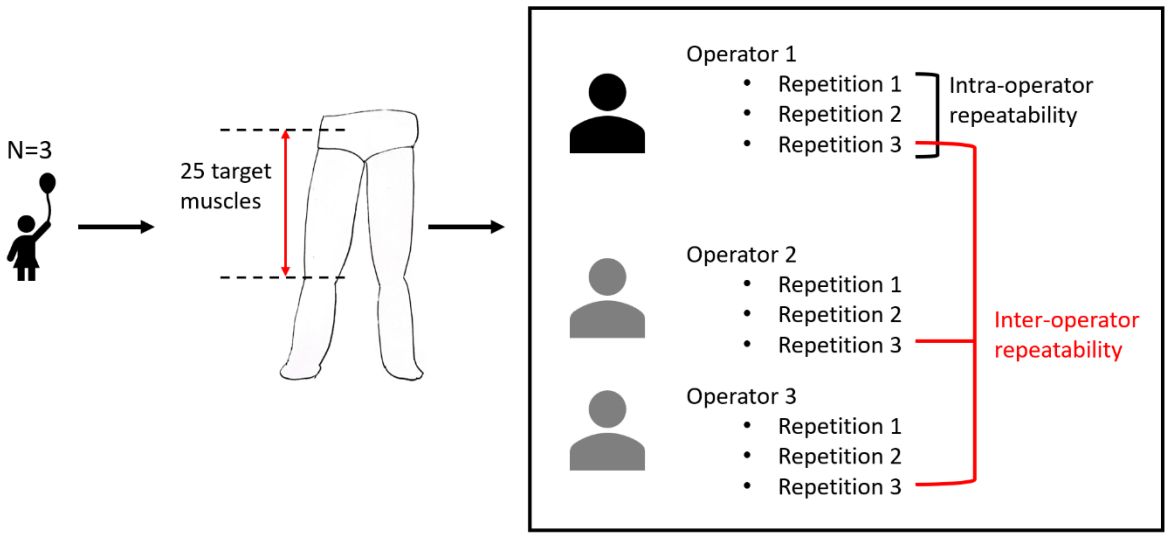


Figure 4.1 Overview of the manual segmentation operation.

Inter-operator variability of the TDC muscle segmentation procedure was accessed by calculating the precision error (PE)[103] of the three operators in terms of the muscle volume (V_m) and length (L_m). Considering that operators may become more familiar with the same case with repeated process, this analysis was based on the third trial from each operator on each subject. Intra-operator variability was determined by calculating the average PE (Precision Error) of the three segmentation results performed by the same operator (one PhD student) on the three subjects. According to literature[103], values of PE in intra-operator variability can be acceptable when lower than 10%. Moreover, a specific label was considered to be segmented reproducibly, and thus suitable for subsequent deep learning model training, only when both the PE of V_m and L_m were lower than 10%.

The Psoas muscle was removed from the analysis since in some subjects it was not fully scanned and partially cut off from the view [6], [7].

The pixel resolution of the MR images was 1.00x1.00x1.00 mm^3, so the V_m for each muscle was calculated by summing the number of pixels in the labeled segmentation image (MATLAB, R2022a). L_m was calculated as the length of the centreline from each muscles' 3D segmentation results. The centreline was generated by linking the points representing the topological skeleton of each muscle cross-section in the 2D MR images slices [6]. The measurement of L_m was automatically performed using the module 'Vascular Modelling Toolkit - Extract Centreline' in 3D slicer (version: 5.7.0) [104].

4.2.2 Deep learning models

Four Convolutional Neural Networks (CNNs) were used in experimental parts: U-Net [81], UNet++ [55], AFFU (Attention-Feature-Fusion-UNet)[14], and Attention UNet [12]. The first two network architectures are well-known, state-of-the-art designs that have been widely used in the field of medical image segmentation task. The AFFU is the modified U-Net structure proposed in Chapter 3 and showed better performance than other benchmark models (U-Net and UNet++) on PMW cohorts [14]. Although Attention-UNet has not been widely tested in the context of muscle segmentation task, it was included in this study as a benchmark model due to its relatively simple attention mechanism compared to AFFU and with less training cost. All the model structure details were included in the previous sections (Section 2.2.2 and Section 3.2.2).

4.2.3 Data pre-processing

All MR image slices were converted from DICOM (Digital Imaging and Communications in Medicine) to PNG (Portable Network Graphic) format in MATLAB (R2022a). The slices were adjusted into left and right parts according to the body's symmetry and cropped to a size of 256x256 pixels. For the left side of TDC cohort which needed to be used, the ‘Flip’ function in MATLAB was applied to create a mirrored image. The MR slices had a resolution of 1.00x1.00 mm, with 1.00 mm between slices. The corresponding segmentation labels for each MR image slice were processed using the same protocol.

4.2.4 Experimental design

4.2.4.1 Task 1: Testing on TDC with model trained on PMW cohort

Task 1 consisted in testing different CNNs (U-Net and AFFU), which were pre-trained based on the PMW cohort as described in Chapter 3, on the TDC cohort. The objective is to investigate the generalization performance of the models when trained on a single cohort (PMW in this study) and then directly tested on another cohort (TDC). Additionally, this task aims to compare the performance of AFFU, which integrates additional attention mechanisms, with U-Net to assess whether these attention modules enhance the predictive accuracy of the models, even when trained on a different cohort. The test dataset included both sides of the thigh muscles from the TDC subjects to investigate whether CNNs trained on one side of the muscle exhibit any obvious differences in performance when tested on bilateral muscle segmentation. To reduce the uncertainty in the performance estimates[105], each model’s training procedure was repeated three times, and the average segmentation results for each target muscle were used to represent the final performance.

4.2.4.2 Task 2: Leave-one-out cross-validation on TDC cohort

Task 2 focused on evaluating the performance of different CNNs (U-Net, UNet++, AFFU and Attention-UNet) on TDC images. It consists of two parts: In the first part, models were trained exclusively on MR images from right-side muscles and then tested on both sides. In the second part, models were trained on images of muscles from both sides and subsequently tested on both sides. By controlling the training set's inclusion of left-side muscles and evaluating the test results, the generalization performance of AFFU and other models in

extracting muscle features and their robustness ability against the difference from symmetry between both side of the body were accessed. During training, a leave-one-out (LOO) cross-validation strategy was employed. In each repetition, one subject was designated as the test set while the remaining subjects were used for training. The final model result was the average of all repetitions (9 in total).

4.2.4.3 Task 3: Comprehensive testing across different age groups

In Task 3, AFFU and other CNNs were trained and tested on subjects from different age groups, including typical developed children (TDC), young healthy adults (HY), and older post-menopausal women (PMW). The objective was to determine whether the models' overall feature learning capability remains consistent when trained on different datasets, including multiple age groups, or on a single cohort. In training, each model's training procedure was repeated three times to reduce the effect of model's uncertainty, with the average performance on the test dataset presented as the result. For each repetition, two different subjects were randomly selected as the test dataset from each cohort, 6 subjects as the test set in each time, while the remaining 22 subjects were used for training.

The summary of how the data used in each task could be seen in Table 4.2 below.

Table 4.2 Data utilization.

	Training data	Testing data	Number of evaluated muscles
Task 1	PMW	TDC	18
Task 2	TDC	TDC	18
Task 3	TDC+HY+PMW	TDC+HY+PMW	15

4.2.5 Evaluation metrics and statistics

For operator variability analysis, the precision error (PE) [103] was used to assess the reproducibility in segmenting each lower-limb muscle. PE was defined as:

$$PE = \sqrt{\sum_{j=1}^{m} \frac{CV_j^2}{m}}, CV = \frac{SD}{\mu} \tag{4.1}$$

here CV is coefficient of variance and is defined as the ratio between the standard deviation (SD) and the mean (μ) of the properties (V_m or L_m) of each muscle, m is the number of subjects (equal to 3 in this section).

For segmentation task of this section, four metrics were used: Dice Similarity Coefficient (DSC), Relative Volume Error (RVE), Hausdorff distance (HD) and Average Symmetric Surface Distance (ASSD) [2], [9], [17], [70]. DSC, RVE and HD were defined in detail in Section 4.x. The ASSD measures the mean distance of surface voxels between the predicted label boundaries and the reference label boundaries. The ASSD is defined as

$$ASSD = \frac{\sum_{p \in B_P} d(p, B_R) + \sum_{r \in B_R} d(r, B_P)}{|B_R| + |B_P|} \tag{4.2}$$

where B_R and B_P are the boundaries of reference labels and predicted or post-processed labels, respectively. $d(x, Y)$ is the Euclidean distance between voxel x and surface Y. A smaller ASSD score, closer to 0, indicates that the segmentation results are closer to the gold standard.

When comparing two different models or same model's performance on both side muscles, the differences were tested for normality using the Kolmogorov-Smirnov test, which indicated that the differences between the paired values were not normally distributed. Consequently, a Wilcoxon signed-rank test was employed [7], [14]. The paired values of each model's DSC, RVE, HD and ASSD metrics were compared across the four models, with a significance level of $a = 0.05$.

4.2.6 Training procedures

All models were trained using identical protocols on an NVIDIA GeForce RTX™ 3060 Ti GPU with 8GB of memory. During the training, the input images were randomly cropped to a size of 128x128 pixels before being fed into the models which not only reduces computational cost but also helps reduce the impact of image noise. For the model: U-Net, UNet++, and Attention-UNet, the batch size was set to 16, while for AFFU, the batch size was set to 8 (due to memory size limitation). The epoch was set to 100 in task 2 and 200 in task 3. The initial learning rate was 0.001 and multiplied by 0.9 after every 5 epochs. The other hyper-parameters were empirically tuned for best performance [7], [14].

4.3 Results

4.3.1 Operator reproducibility analysis of TDC cohort

Table 4.3 Reproducibility of muscle segmentation.

	Intra-op PE [%]		Inter-op PE [%]	
Muscles	V_m	L_m	V_m	L_m
Rectus femoris †‡	2.3	3.5	7.8	2.2
Vastus intermedius †‡	3.8	1.4	10.9	15.7
Vastus lateralis †‡	2.7	2.3	7.6	7.4
Vastus medialis †‡	3.1	1.5	7.3	5.3
Sartorius †	2.5	1.6	18.6	1.1
Semimembranosus †‡	2.8	2.2	13.6	5.0
Semitendinosus †‡	1.9	1.2	9.7	4.2
Gracilis †	4.0	0.6	19.4	2.2
Biceps femoris caput brevis †‡	4.5	3.1	24.1	8.3
Biceps femoris caput longum †‡	3.2	1.7	11.1	3.8
Adductor magnus †‡	2.7	3.9	6.9	6.8
Adductor brevis †	24.4	18.3	32.9	37.9
Adductor longus †	5.5	3.9	12.7	7.0
Gemellus superior	21.6	28.0	36.6	44.6
Gluteus maximus †‡	1.2	1.5	6.9	1.6
Gluteus medius	3.0	3.3	4.0	6.7
Gluteus minimus	16.4	7.6	29.6	8.7
Iliacus †‡	5.2	2.8	9.8	3.5
Obturator externus	9.4	35.1	32.6	38.1
Obturator internus	4.8	9.3	17.5	28.7
Pectineus	5.1	4.7	23.4	8.9
Piriformis	7.0	26.5	51.9	27.5
Psoas	4.1	3.4	15.1	12.5
Quadratus femoris	17.5	27.9	32.3	44.6
Tensor fasciae latae †	5.0	2.2	17.8	5.4
	Intra-op PE < 10%		Inter-op PE < 10%	

Inter- and intra-operator PE for V_m and L_m was calculated by three operators (inter-op) and by one operator over three repetitions (intra-op) on three random chosen subjects from the TDC cohort. The values of inter-op and intra-op PE lower than 10% were marked in light blue and in orange, respectively. Muscles marked in the left column with a dagger symbol (†) or double dagger (‡) denote those muscles that exhibited high reproducibility in the intra- or inter- operator analysis as reported in reference [6] on post-menopausal women. That study excluded Gemellus superior, Obturator externus, Obturator internus, Pectineus, Piriformis and Psoas.

The intra-operator PE was lower than the inter-operator PE for most of the tested muscles (Table 4.3), excluding L_m of Rectus femoris and Sartorius. This result indicates that a single operator tends to be more consistent in muscle segmentation than multiple operators. This consistency is more pronounced in V_m measurements than in L_m.

Muscles like the Adductor brevis, Piriformis, Gemellus superior, Obturator externus and Quadratus femoris exhibit high PE in both inter- and intra-op analyses, indicating challenges in consistently segmenting these muscles regardless of the operator.

From the manual segmentation visualization (Figure 4.2), obvious perceptual differences can be observed in the thigh region (upper section), particularly in muscles like the Vastus intermedius (VI) and Vastus lateralis (VL), especially between operator 3 and operators 1/2. Similarly, in the area close to hips, such as the Adductor brevis (AB), significant differences in boundary delineation between operators are also evident. These differences align with the results shown in Table 4.2 (the inter-operator PE for the AB exceeds 30%, while the inter-operator PE for the VI is greater than 10%).

18 muscles such as the Rectus femoris, Vastus lateralis, and Gluteus maximus exhibit relatively low PE in intra-op analysis for both V_m and L_m, indicating high reproducibility from the same operator. These 18 muscles were included in the next DL models' evaluation stage of Task 1 and Task 2.

Except for L_m of Obturator internus of subject 1 (CV of 14%), the rest of the muscles marked in blue in Table 5.3 had a CV lower than 10%.

Among the 16 highly reproducible muscles reported in reference [6] and the 18 highly reproducible muscles identified in this section, the manual segmentations of 15 muscles are

consistently reproducible in both studies and were included in the evaluation stage in Task 3 (where both cohorts were used for training of the models).

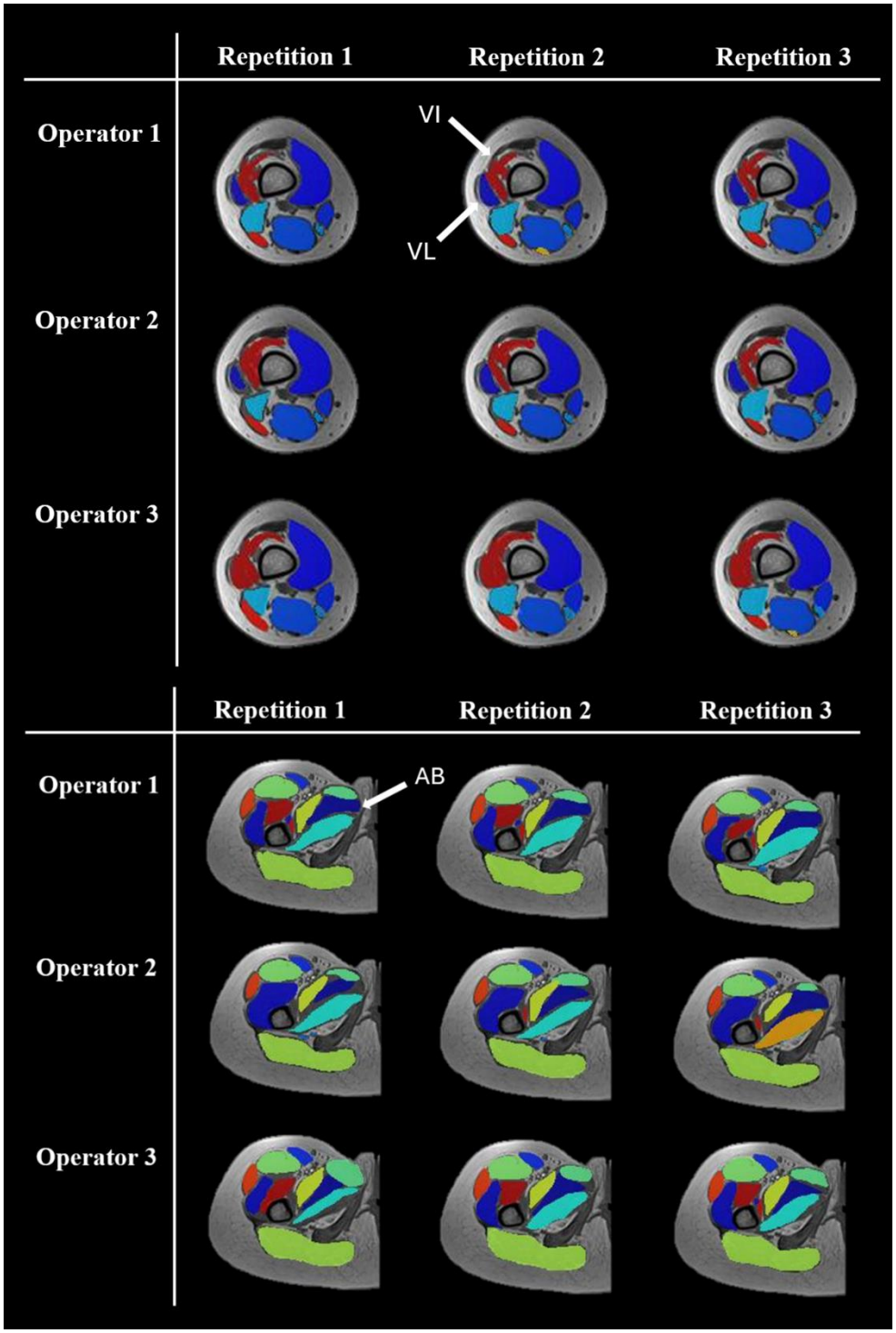


Figure 4.2 Visualization of manual segmentation cross 3 repetitions among 3 operators. The label combined MR images in the upper and lower sections are from the same subject selected for the repeatability analysis. Repetition 1-3 corresponds to the first through the last trial segments performed by the same operator.

Table 4.4 Intra-operator coefficient of variance.

	CV of V_m (%)			CV of L_m (%)		
Muscles	Sub 1	Sub 2	Sub 3	Sub 1	Sub 2	Sub 3
Rectus femoris	0.7	3.3	2.1	2.8	5.2	0.8
Vastus intermedius	2.2	4.7	4.2	0.8	1.8	1.3
Vastus lateralis	3.4	2.4	2.0	3.3	0.5	2.2
Vastus medialis	2.3	3.3	3.4	2.0	0.7	1.5
Sartorius	0.5	3.8	1.9	0.5	0.7	2.6
Semimembranosus	0.9	4.7	0.5	3.4	1.2	0.9
Semitendinosus	3.1	1.0	0.4	1.4	1.4	0.7
Gracilis	2.8	4.8	4.0	0.9	0.2	0.5
Biceps femoris caput brevis	5.6	5.3	1.3	3.5	3.2	2.7
Biceps femoris caput longum	2.3	5.0	0.8	0.6	2.1	1.9
Adductor magnus	4.5	0.9	1.0	4.3	5.0	1.5
Adductor brevis	39.6	10.1	10.4	30.0	2.2	10.0
Adductor longus	1.8	7.9	5.1	5.7	3.1	1.8
Gemellus superior	9.8	11.3	34.3	25.5	31.0	27.3
Gluteus maximus	1.3	1.3	0.9	0.2	2.5	0.5
Gluteus medius	3.9	3.3	0.7	0.8	3.1	4.6
Gluteus minimus	27.4	6.0	4.5	11.2	4.7	4.9
Iliacus	2.3	7.9	4.0	3.5	2.6	2.2
Obturator externus	14.0	7.5	3.4	5.8	37.6	62.1
Obturator internus	6.9	3.8	2.8	14.1	4.4	6.3
Pectineus	5.6	6.0	3.2	6.1	1.5	5.0
Piriformis	6.9	9.0	4.0	15.5	29.0	32.2
Psoas	4.6	3.8	3.7	2.0	5.2	1.9
Quadratus femoris	20.0	8.1	21.2	47.5	6.5	6.9
Tensor fasciae latae	2.6	8.2	0.5	2.2	2.5	1.7

Intra-operator coefficient of variance (CV) for V_m and L_m calculated by same operator (operator 1) over three repetitions of three subjects (subject 1-3). Muscles marked in blue indicate high reproducibility.

Table 4.5 Muscle volume and length variability analysis.

Muscles	Volume (cm3)			Length (cm)		
	Left	Right	Difference	Left	Right	Difference
Rectus femoris	135 (84, 207) **	159 (98, 240)	18 %	29.3 (27.8, 38.8)	30.1 (27.1, 32.9)	3 %
Vastus intermedius	255 (191, 349)	246 (188, 331)	-3 %	33.5 (30.2, 39.5)	32.3 (29.4, 39.8)	-3 %
Vastus lateralis	358 (265, 403) *	410 (294, 453)	14 %	32.1 (26.6, 34.5)	30.7 (27.2, 34.3)	-4 %
Vastus medialis	218 (150, 295)	227 (164, 296)	4 %	30.2 (27.5, 33.5)	29.9 (25.1, 35.5)	-1 %
Sartorius	62 (54, 82)	69 (49, 110)	12 %	49.0 (42.6, 51.1)	48.2 (42.6, 50.4)	-2 %
Semimembranosus	100 (90, 151)	110 (85, 166)	10 %	26.7 (24.6, 29.9)	25.3 (21.2, 30.4)	-5 %
Semitendinosus	94 (61, 120) *	107 (68, 130)	14 %	31.0 (26.1, 33.8)	31.4 (24.7, 34.4)	1 %
Gracilis	45 (31, 54)	48 (32, 63)	8 %	32.5 (24.6, 35.2) *	30.6 (25.6, 32.2)	-6 %
Biceps femoris caput brevis	50 (32, 63)	49 (29, 85)	-2 %	22.7 (21.4, 26.0)	23.1 (18.8, 26.5)	2 %
Biceps femoris caput longum	93 (58, 138)	101 (52, 146)	9 %	27.7 (23.4, 32.8) **	25.6 (22.6, 30.7)	-7 %
Adductor magnus	279 (206, 363) **	304 (214, 405)	9 %	28.5 (22.6, 31.6)	28.4 (20.7, 29.4)	0 %
Adductor brevis	61 (37, 70)	58 (34, 79)	-6 %	13.7 (12.3, 16.2)	12.5 (9.6, 14.8)	-8 %
Adductor longus	64 (52, 80)	74 (47, 97)	15 %	19.8 (17.1, 22.7) **	18.5 (14.8, 21.8)	-6 %
Gemellus superior	5 (4, 9)	5 (3, 8)	3 %	3.8 (3.1, 6.2)	4.5 (3.6, 7.9)	17 %
Gluteus maximus	407 (281, 553) **	465 (296, 602)	14 %	25.9 (22.8, 28.5)	26.0 (24.1, 28.4)	0 %
Gluteus medius	179 (108, 205)	181 (118, 214)	2 %	15.7 (14.7, 18.2)	15.7 (13.5, 18.5)	0 %
Gluteus minimus	48 (38, 61)	51 (38, 70)	7 %	12.4 (9.3, 13.7)	10.2 (4.8, 12.8)	-18 %
Iliacus	97 (68, 113)	99 (67, 123)	3 %	22.7 (19.3, 24.6)	21.5 (19.5, 23.3)	-5 %
Obturator externus	26 (20, 35)	24 (17, 30)	-9 %	6.1 (3.5, 10.4)	7.8 (6.7, 9.4)	28 %
Obturator internus	28 (21, 35) **	24 (15, 30)	-16 %	8.3 (7.0, 9.2) **	6.6 (5.1, 8.5)	-21 %
Pectineus	31 (23, 36)	33 (23, 45)	7 %	10.6 (9.4, 12.6)	10.5 (9.5, 12.4)	0 %
Piriformis	18 (10, 22)	16 (10, 23)	-11 %	5.2 (4.2, 8.0)	6.2 (4.6,7.5)	20 %
Quadratus femoris	10 (2,14)	13 (2, 17)	31 %	4.4 (3.2, 5.6)	5.6 (2.8, 6.4)	27 %
Tensor fasciae latae	32 (19, 39)	30 (23, 55)	-7 %	14.2 (12.2, 16.5)	13.8 (12.2, 15.4)	-2 %

Median (minimum, maximum) values of muscle (excluding Psoas) volume and length for the right and left lower-limb muscles. Muscles marked in the column Left with * or ** denote significant difference ($p < 0.01$or $p < 0.05$) between limbs in V_m and L_m, respectively. The difference between is calculated as (Median_R – Median_L)/Median_L. All the segmentation results used were performed by the same operator.

4.3.2 DL model testing results

4.3.2.1 Task 1 results: TDC cohort testing results from model trained on PMW

Table 4.6 Model comparison for the four used metrics of U-Net/AFFU trained on PMW cohort.

	Right part			
Model	DSC	RVE	HD (mm)	ASSD (mm)
U-Net	0.531 ± 0.226	**-0.278 ± 0.394**	116.0 ± 87.9	10.9 ± 8.8
AFFU	**0.559 ± 0.250**	-0.337 ± 0.363	**76.37 ± 66.6**	**9.2 ± 14.2**
	Left part			
U-Net	0.495 ± 0.219	**-0.287 ± 0.433**	126.3 ± 86.4	11.4 ± 6.6
AFFU	**0.537 ± 0.241**	-0.343 ± 0.387	**92.8 ± 74.3**	**10.2 ± 14.3**

DSC, RVE, HD and ASSD found for 18 muscles with high repeatability included in the automatic segmentation results. Values in the table are the mean value (± standard deviation, SD) averaged across 9 TDC subjects of 3 repetitions of each model, 162 (9x18) labels were calculated for each value. The best value under each metrics is highlighted in bold.

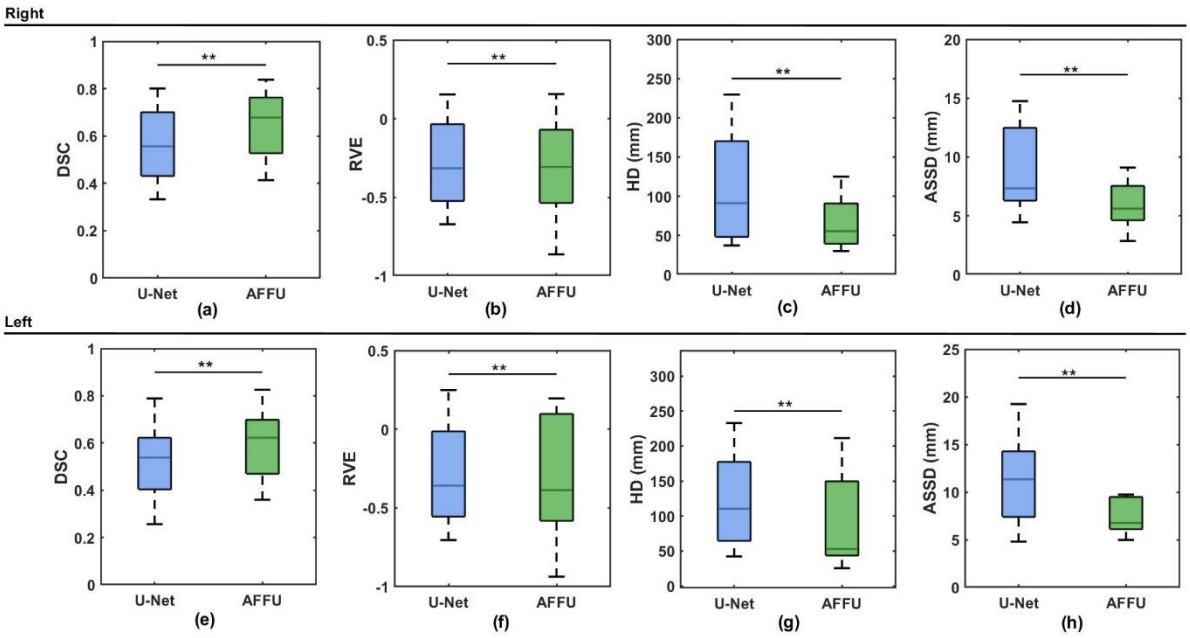


Figure 4.3 Model comparison of both sides of U-Net/AFFU trained on PMW cohort. Panels a-d represent the segmentation results for the right-side muscles, while panels e-h represent the left-side muscles. The box plots for each model are constructed from the medians of each target muscles in all subjects (18 points in total) averaged over the three repetitions.

Statistically significant differences between results of both models are reported with * ($p < 0.05$) or ** ($p < 0.01$).

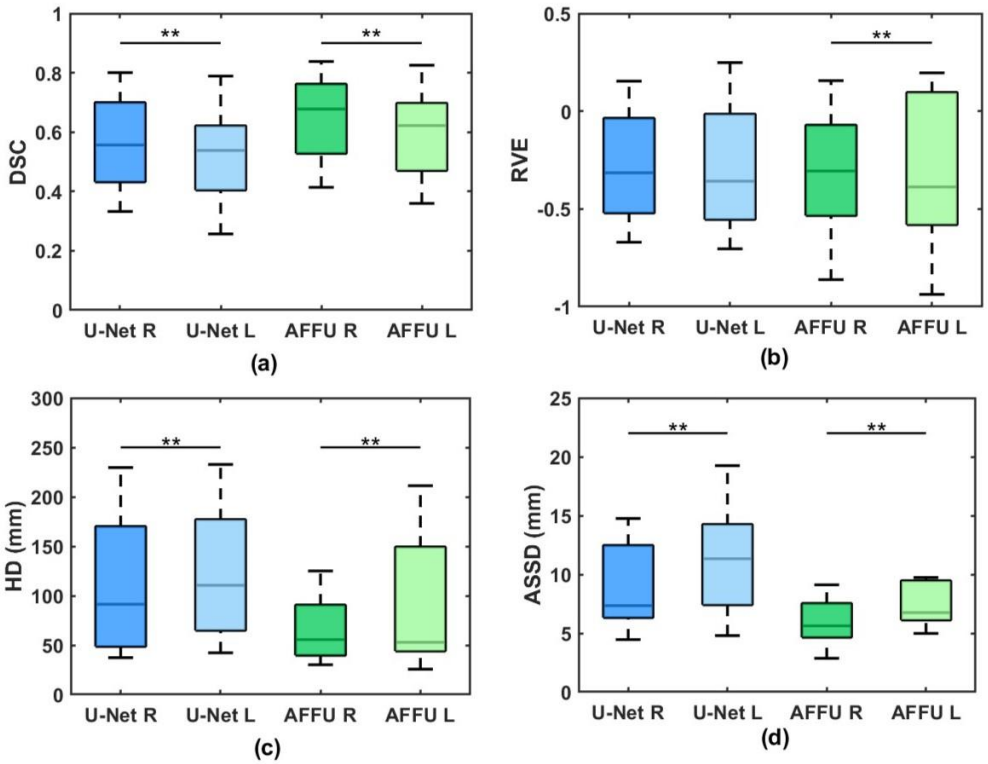


Figure 4.4 Comparison of muscle segmentation performance of the same model on different lower-limb sides. Each boxplots includes 18 medians value from 9 subjects across 3 repetitions. In each subplot, dark and light colour represents right and left side respectively.

When trained on the PMW cohort and tested on TDC cohort, all models didn't perform well. However, under low scores, the AFFU still performed better than U-Net among DSC (0.559/0.537, right/left, $p < 0.05$), HD (76.4/92.8 mm, $p < 0.05$) and ASSD (9.2/10.2 mm, $p < 0.05$) with following average improvement: 5.3% of right (8.5% of left) in DSC, 34.5% (26.5%) in HD and 15.6% (10.5%) in ASSD. Regarding RVE, U-Net outperformed AFFU with improvement around 16.5% of both sides.

Compared to the left side, AFFU and U-Net demonstrated significantly better results on the right-side muscles, with the following improvements: a 4.1% (AFFU) and 7.3% (U-Net) increase in DSC, a 1.7% (AFFU) improvement in RVE, a 17.7% (AFFU) and 7.9% (U-Net) reduction in HD, and a 9.8% (AFFU) and 4.4% (U-Net) reduction in ASSD.

4.3.2.2 Task 2 results: TDC test results from model trained on TDC cohort

Table 4.7 Model comparison for the four metrics used in this study for models trained on right leg muscles from TDC cohort.

	Right part			
Model	DSC	RVE	HD (mm)	ASSD (mm)
U-Net	0.861 ± 0.054	0.111 ± 0.169	31.5 ± 46.0	2.0 ± 1.4
UNet++	**0.868 ± 0.045**	0.117 ± 0.133	32.5 ± 57.7	1.9 ± 1.3
Attention UNet	0.866 ± 0.049	0.119 ± 0.155	32.5 ± 52.1	1.9 ± 1.6
AFFU	0.864 ± 0.048	**0.096 ± 0.161**	**20.8 ± 32.8**	**1.6 ± 0.9**
	Left part			
U-Net	0.829 ± 0.091	0.084 ± 0.191	51.3 ± 62.9	3.1 ± 3.0
UNet++	**0.836 ± 0.063**	0.109 ± 0.158	38.9 ± 56.4	2.4 ± 1.6
Attention UNet	0.831 ± 0.083	0.097 ± 0.193	42.5 ± 51.2	2.7 ± 2.1
AFFU	0.834 ± 0.064	**0.076 ± 0.145**	**29.2 ± 45.1**	**2.1 ± 1.3**

The values in the table represent the mean (± standard deviation, SD) averaged across 9 TDC subjects using leave-one-out method per model. A total of 162 (9x18) labels were calculated for each value. The best value under each metric is highlighted in bold font.

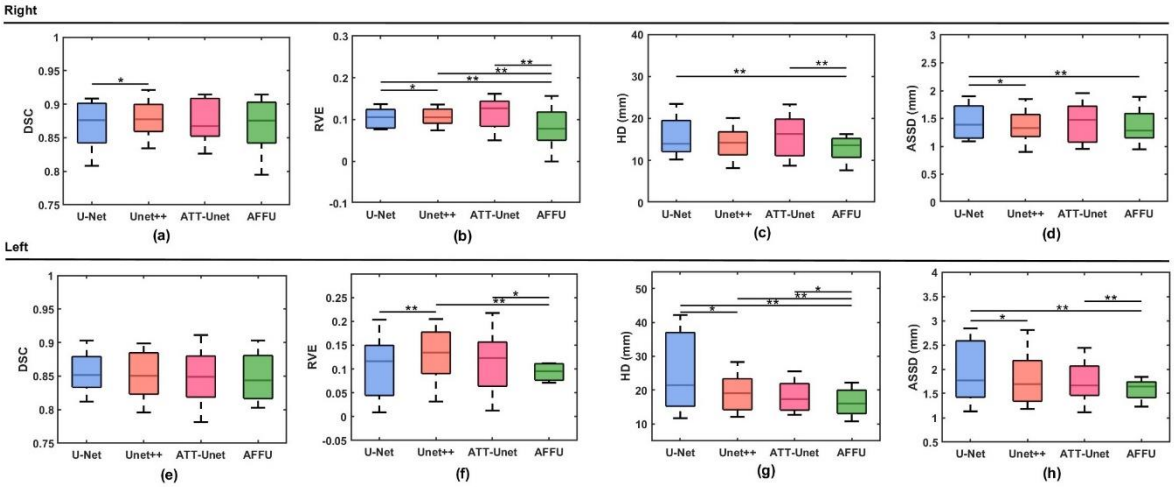


Figure 4.5 Performance comparison for each model. Box plots of four metrics for U-Net (blue), UNet++ (orange), Attention-UNet (red) and AFFU (green) are shown (* indicates $p < 0.05$, ** indicates $p < 0.01$). Panels a-d represent the results for the right-side muscles, while panels e-h represent the score for left-side muscles.

For the right-side, UNet++ achieved the highest DSC score (0.868 ± 0.045), which is 1% higher than U-Net (0.861 ± 0.054, $p < 0.01$). AFFU achieved the highest score in RVE (0.096 ± 0.161), HD (20.8 ± 32.8 mm) and ASSD (1.6 ± 0.9 mm), outperforming other

models with following improvement: 13.5% (U-Net, $p < 0.01$), 17.9% (UNet++, $p < 0.01$) and 19.3% (Attention-UNet, $p < 0.01$) in DSC, 34.0% (U-Net, $p < 0.01$), 36.0%(Attention-UNet, $p < 0.01$) in HD, 20% (U-Net, $p < 0.01$) in ASSD.

For the left-side muscles, the models did not exhibit significant differences in DSC scores, with an average around 0.833. Similar to the right side, AFFU also achieved better score in RVE (0.076 ± 0.145), HD (29.2 ± 45.1 mm) and ASSD (2.1 ± 1.3 mm), with following improvement compared to others: 30.3% (UNet++, $p < 0.01$), 21.6% (Attention-UNet, $p < 0.01$) in DSC, 43.1% (U-Net, $p < 0.01$), 24.9% (UNet++, $p < 0.01$) and 31.3% (Attention-UNet, $p < 0.05$) in HD, 32.3% (U-Net, $p < 0.05$), 22.2% (Attention-UNet, $p < 0.05$) in ASSD.

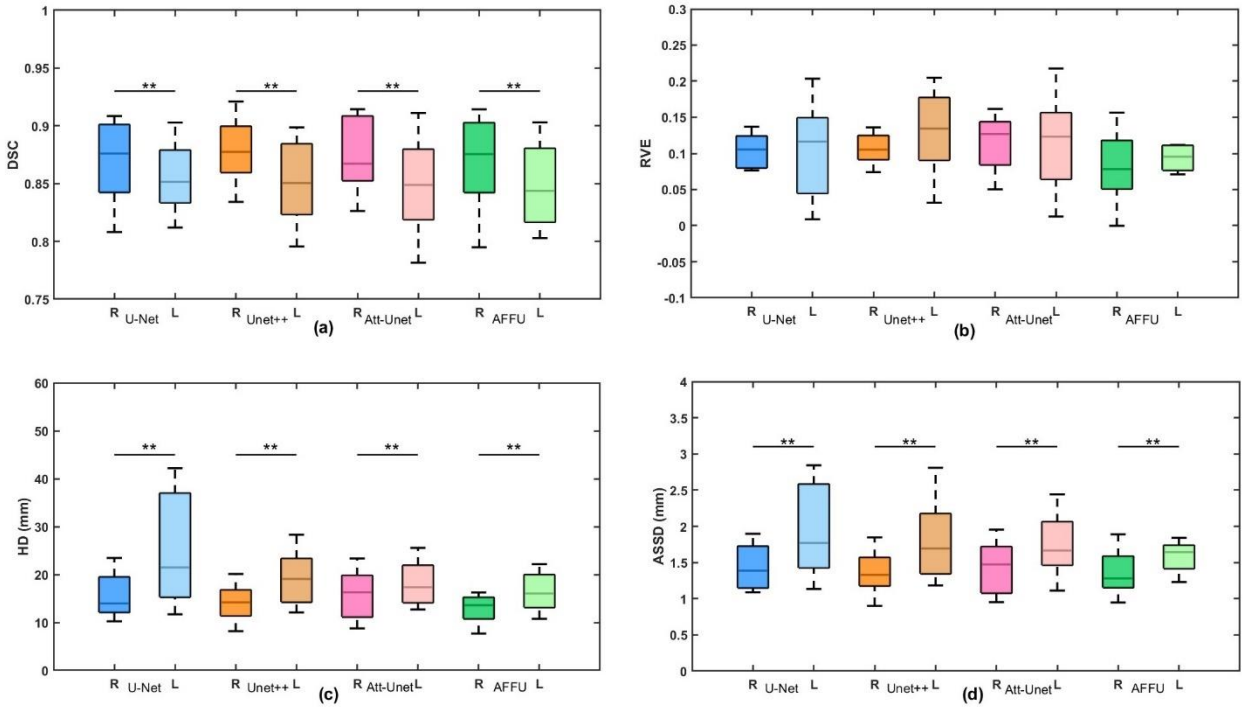


Figure 4.6 Comparison of muscle segmentation performance of the same model on different lower-limb sides.

When trained using only the right-side muscles, the models showed differences in predicting muscles on both sides (except for RVE), with the overall results being superior for the right-side muscles, as expected.

The average improvement for the right-side muscles compared to the left-side was 3.6% to 4.2% in DSC, 16.5% to 38.6% in HD, and 20.8% to 35.5% in ASSD.

Table 4.8 p value from statistical analysis of each model.

	p value			
Muscles	U-Net	UNet++	Attention UNet	AFFU
Rectus femoris	0.301	0.301	0.164	0.203
Vastus intermedius	0.098	**0.004**	**0.020**	**0.004**
Vastus lateralis	0.570	**0.020**	0.250	0.129
Vastus medialis	**0.020**	**0.012**	**0.004**	0.074
Sartorius	**0.020**	**0.039**	**0.020**	**0.004**
Semimembranosus	**0.020**	**0.020**	**0.027**	**0.004**
Semitendinosus	**0.008**	**0.008**	**0.004**	**0.020**
Gracilis	0.164	**0.012**	**0.020**	**0.012**
Biceps femoris caput brevis	**0.004**	**0.004**	**0.020**	0.055
Biceps femoris caput longum	0.203	**0.012**	**0.012**	**0.020**
Adductor magnus	0.098	0.055	**0.012**	**0.004**
Adductor longus	**0.039**	0.074	**0.027**	0.164
Gluteus maximus	**0.004**	**0.008**	**0.008**	**0.012**
Gluteus medius	0.652	0.129	0.164	0.301
Iliacus	0.570	0.129	0.164	0.098
Obturator internus	0.910	0.129	1.000	0.129
Pectineus	0.652	0.074	0.820	0.734
Tensor fasciae latae	0.074	**0.039**	**0.027**	**0.008**

The p values in the table were calculated by comparing the DSC values for the bilateral side of the same muscle across the 9 test samples for each model. Bold font indicates p values less than 0.05.

Three (or more) models showed significant differences (Figure 4.6, Table 4.8) in segmentation prediction results between the bilateral muscles for Vastus intermedius, Vastus medialis, Sartorius, Semimembranosus, Semitendinosus, Gracilis, Biceps femoris caput brevis, Biceps femoris caput longum, Gluteus maximus, and Tensor fasciae latae.

Table 4.9 Model comparison for the four metrics used in this study, for models trained on bilateral muscles from TDC cohort.

	Right part			
Model	DSC	RVE	HD (mm)	ASSD (mm)
U-Net	0.860 ± 0.051	0.149 ± 0.172	34.7 ± 57.5	2.1 ± 1.9
UNet++	0.864 ± 0.049	0.172 ± 0.164	28.5 ± 50.5	1.8 ± 1.3
Attention UNet	0.864 ± 0.049	0.140 ± 0.163	31.9 ± 54.5	1.8 ± 1.4
AFFU	**0.865 ± 0.049**	**0.119 ± 0.170**	**23.2 ± 40.7**	**1.7 ± 1.1**
	Left part			
U-Net	0.867 ± 0.051	0.141 ± 0.122	39.7 ± 54.2	2.2 ± 2.2
UNet++	0.869 ± 0.043	0.164 ± 0.110	36.9 ± 56.9	2.1 ± 2.3
Attention UNet	0.868 ± 0.063	0.137 ± 0.135	37.1 ± 58.4	2.2 ± 2.2
AFFU	**0.871 ± 0.055**	**0.117 ± 0.129**	**24.4 ± 41.1**	**1.7 ± 1.3**

Data was obtained in the same manner as described in Table 4.7.

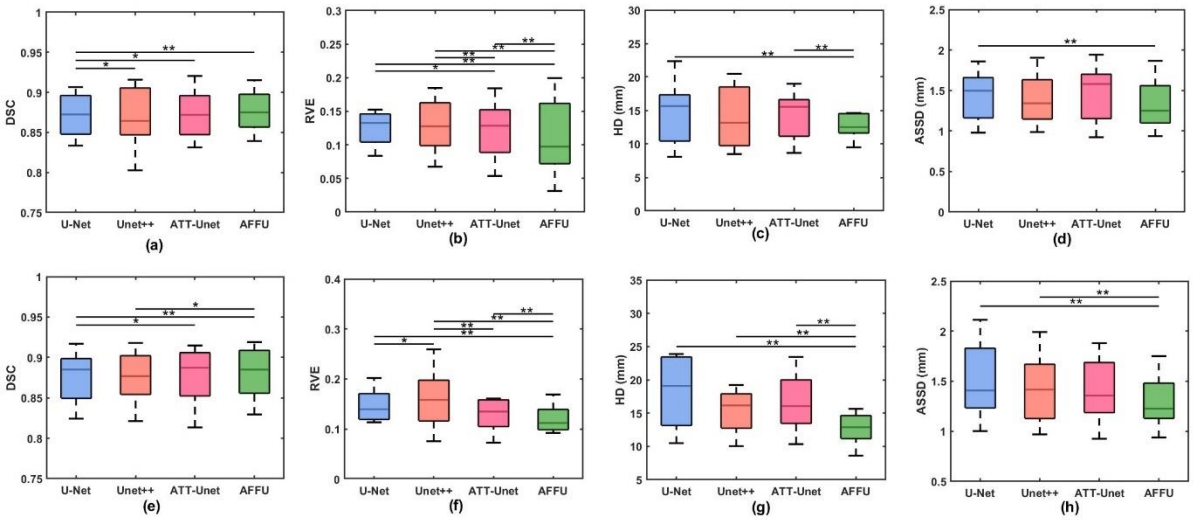


Figure 4.7 Performance comparison of each model. Box plots of four metrics for U-Net (blue), UNet++ (orange), Attention-UNet (red) and AFFU (green) are shown (* indicates $p < 0.05$, ** indicates $p < 0.01$). Panels (a)-(d) represent the results for the right-side muscles, while panels (e)-(h) represent the score for left-side muscles.

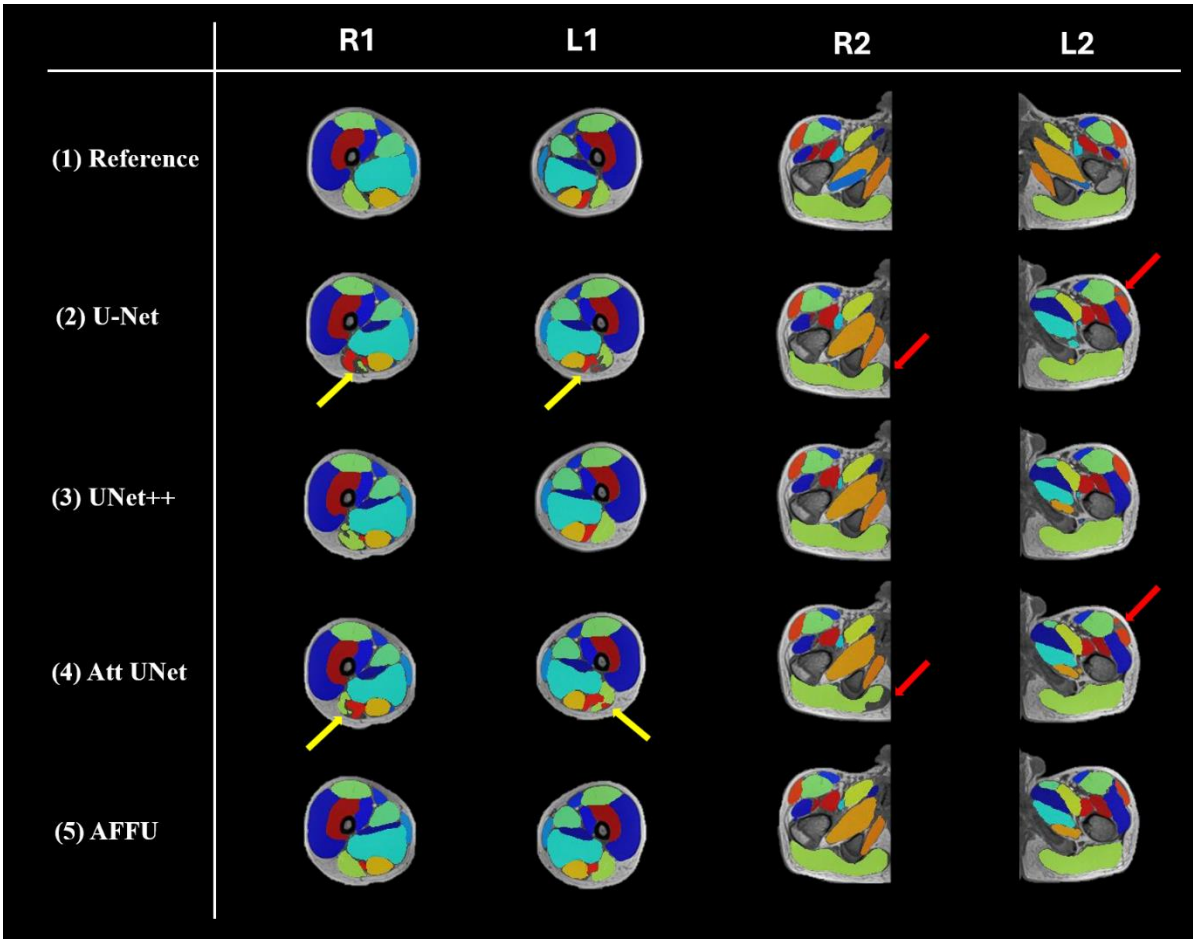


Figure 4.8 Visualization of model performance. The images displayed are from the same test subject. Rows (1) to (5) represent the segmentation results from the reference (ground truth) and each of the models. Columns R1/2 and L1/2 correspond to the images from the right and left sides, respectively. The yellow arrows indicate blank regions[106], while the red arrows highlight mis-segmentation errors.

AFFU achieved best scores in RVE (0.119 ± 0.170), HD (23.2 ± 40.7 mm) and ASSD (1.7 ± 1.1 mm) with following improvement compared to others: 20.1% (U-Net, $p < 0.01$), 30.8% (UNet++, $p < 0.01$), 15% (Attention UNet, $p < 0.01$) in RVE, 33.1% (U-Net, $p < 0.01$), 27.3% (Attention UNet, $p < 0.01$) in HD, 19.0% (U-Net, $p < 0.01$) in ASSD. From the visualization demo shown in Figure 4.8, it can be seen that compared to ground truth, AFFU and UNet++ achieved more closer segmented results than U-Net and Attention UNet with less blank regions or mis-segmentation [106].

For the left-side, AFFU was superior than the other models, with DSC (0.871 ± 0.055), HD (24.4 ± 41.1 mm) and ASSD (1.7 ± 1.3 mm), with following improvement compared to others: 0.5% (U-Net, $p < 0.01$), 0.3% (UNet++, $p < 0.05$) in DSC, 17% (U-Net, $p < 0.01$), 28.7% (UNet++, $p < 0.01$), 14.6% (Attention UNet, $p < 0.01$) in RVE, 38.5% (U-Net, $p < 0.01$), 33.9% (UNet++, $p < 0.01$), 34.2% (Attention UNet, $p < 0.01$) in HD, 22.7% (U-Net, $p < 0.01$), 19.0% (UNet++, $p < 0.01$) in ASSD.

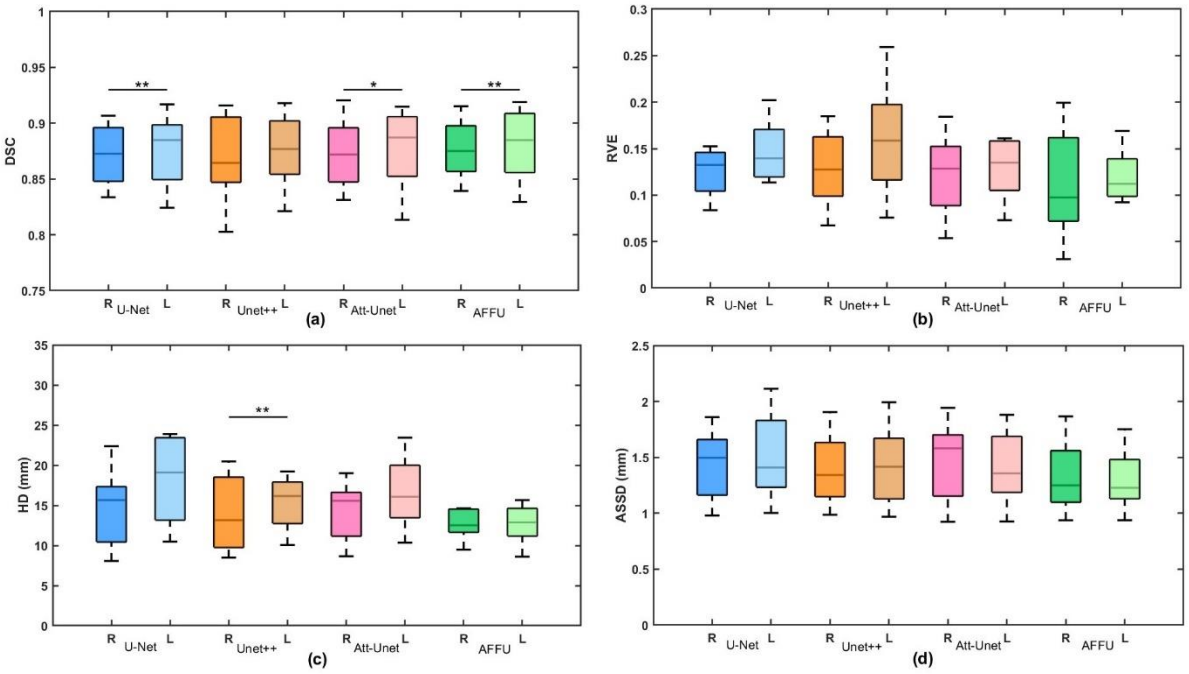


Figure 4.9 Comparison of muscle segmentation performance of the same model on different lower-limb sides.

After incorporating bilateral muscles MR images in the training dataset, the differences in model performance between the left and right muscles became less pronounced (Figure 4.9). Three models showed better DSC scores in left side compared with right side, with the following gains: 0.8% of U-Net ($p < 0.01$), 0.5% of Attention UNet ($p < 0.05$) and 0.7% of AFFU ($p < 0.01$).

Table 4.10 Model generalization evaluation comparison for models trained on TDC cohort and tested on PMW and HY cohorts.

	PMW			
Model	DSC	RVE	HD (mm)	ASSD (mm)
U-Net	0.060 ± 0.091	-0.951 ± 0.073	85.2 ± 59.5	28.3 ± 20.5
AFFU	0.056 ± 0.079	-0.956 ± 0.069	90.0 ± 57.5	29.6 ± 21.5
	HY			
U-Net	0.585 ± 0.158	-0.362 ± 0.245	87.0 ± 57.4	7.5 ± 3.5
AFFU	0.649 ± 0.133	-0.289 ± 0.213	74.7 ± 54.7	6.0 ± 2.1

Following the generalization evaluation in Task 1, as a supplementary experiment, we also selected three repetitions for each model from the leave-one-out training strategy described in first part in this subsection and then tested them on PMW and HY datasets. Both training and testing were based on the right lower limb muscles, with the evaluation covering 16 muscles[6], [14].

Both models failed to segment the muscles from PMW cohort which can be observed from DSC and RVE score (Table 4.10) and showed the similar generalization performance as in Task 1 on PMW cohort, for DSC from 0.585 to 0.649 and RVE from -0.362 to -0.289.

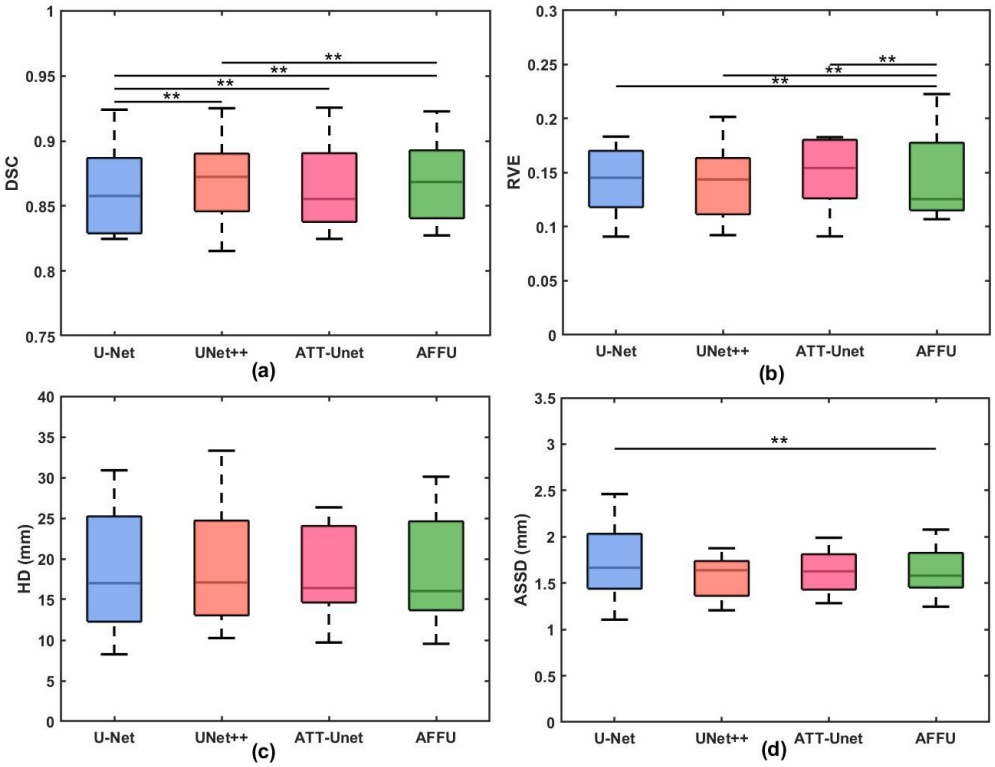
4.3.2.3 Task 3 results: Results from model trained on different age groups

Table 4.11 Model comparison for the four metrics used in this study, model trained on combined dataset including different age groups.

	Right part			
Model	DSC	RVE	HD (mm)	ASSD (mm)
U-Net	0.858 ± 0.046	0.145 ± 0.135	18.4 ± 24.9	1.7 ± 0.9
UNet++	0.870 ± 0.042	0.143 ± 0.119	17.5 ± 31.0	1.7 ± 1.1
Attention UNet	0.869 ± 0.042	0.149 ± 0.117	17.8 ± 45.4	1.7 ± 1.1
AFFU	**0.872 ± 0.041**	**0.127 ± 0.117**	**16.7 ± 33.9**	**1.6 ± 1.0**

The values in the table represent the mean (± standard deviation, SD) averaged across 9 test subjects from different age groups in 3 repetitions where each repetition includes 1 subject from each cohort. A total of 135 (9x15) labels were calculated for each value.

Figure 4.10 Performance comparison of each model. Box plots of four metrics for U-Net



(blue), UNet++ (orange), Attention-UNet (red) and AFFU (green) are shown (* indicates $p < 0.05$, ** indicates $p < 0.01$). Each boxplot includes 15 medians of each muscle from 9 testing subjects.

AFFU provided the best predictions (DSC: 0.872 ± 0.041, RVE: 0.127 ± 0.117, HD: 16.7 ± 33.9 mm, ASSD: 1.6 ± 1.0 mm) among all models, with small improvement: 1.6% (U-Net, $p < 0.01$), 0.2% (UNet++, $p < 0.01$) in DSC, 12.4% (U-Net, $p < 0.01$), 11.2% (UNet++, $p < 0.01$) and 14.8% (Attention UNet, $p < 0.01$) in RVE, 5.9% (U-Net, $p < 0.01$) in ASSD. All models show no significant differences in HD (16.7-18.4 mm).

Table 4.12 Model comparison for different age groups.

	TDC			
Model	DSC	RVE	HD (mm)	ASSD (mm)
U-Net	0.860 ± 0.040	0.189 ± 0.118	22.3 ± 33.3	1.7 ± 0.8
UNet++	0.869 ± 0.035	0.190 ± 0.097	22.2 ± 30.2	1.8 ± 1.3
Attention UNet	0.867 ± 0.037	0.203 ± 0.010	41.2 ± 72.7	1.9 ± 1.7
AFFU	0.866 ± 0.037	0.187 ± 0.083	29.1 ± 50.0	1.6 ± 1.1
	HY			
Model	DSC	RVE	HD (mm)	ASSD (mm)
U-Net	0.835 ± 0.050	0.122 ± 0.167	27.8 ± 19.4	2.2 ± 0.9
UNet++	0.851 ± 0.045	0.109 ± 0.137	34.3 ± 41.3	2.3 ± 1.1
Attention UNet	0.845 ± 0.047	0.104 ± 0.143	28.5 ± 22.0	2.1 ± 0.8
AFFU	0.850 ± 0.043	0.084 ± 0.136	31.3 ± 24.5	2.1 ± 0.9
	PMW			
Model	DSC	RVE	HD (mm)	ASSD (mm)
U-Net	0.869 ± 0.042	0.137 ± 0.106	25.6 ± 19.6	2.0 ± 0.9
UNet++	0.877 ± 0.043	0.131 ± 0.108	25.0 ± 15.2	1.9 ± 0.8
Attention UNet	0.881 ± 0.034	0.131 ± 0.080	26.3 ± 19.8	1.9 ± 0.7
AFFU	0.874 ± 0.041	0.117 ± 0.101	24.3 ± 19.3	2.0 ± 0.9

For the TDC cohort, AFFU (RVE: 0.187 ± 0.083, ASSD: 1.6 ± 1.1 mm) outperformed Attention UNet (RVE: 0.203 ± 0.010, $p < 0.05$) and U-Net (1.7 ± 0.8 mm, $p < 0.05$) only in terms of RVE and ASSD, respectively, with improvements of 7.9% and 6.7%.

For the HY cohort, AFFU showed better performance on DSC (0.850 ± 0.043), RVE (0.084 ± 0.136) and ASSD (2.1 ± 0.9 mm) compared to others, with following gains: 20.0% (U-Net, $p < 0.01$) in DSC, 31.1% (U-Net, $p < 0.01$), 22.9% (UNet++, $p < 0.01$) and 19.2% (Attention UNet, $p < 0.01$) in RVE, 4.5% (U-Net, $p < 0.05$) and 8.7% (UNet++, $p < 0.05$) in ASSD. There is no obvious difference in HD between AFFU and other models.

For the PMW cohort, AFFU didn't show better score in HD and ASSD, just outperformed U-Net in DSC and RVE with gain 0.6% ($p < 0.05$) in DSC, 14.6% ($p < 0.01$) in RVE. UNet++

(DSC: 0.877 ± 0.043) and Attention UNet (0.881 ± 0.034) achieved higher scores in DSC than AFFU (0.874 ± 0.041) with gains 0.3% ($p < 0.05$) and 0.8% ($p < 0.01$), respectively.

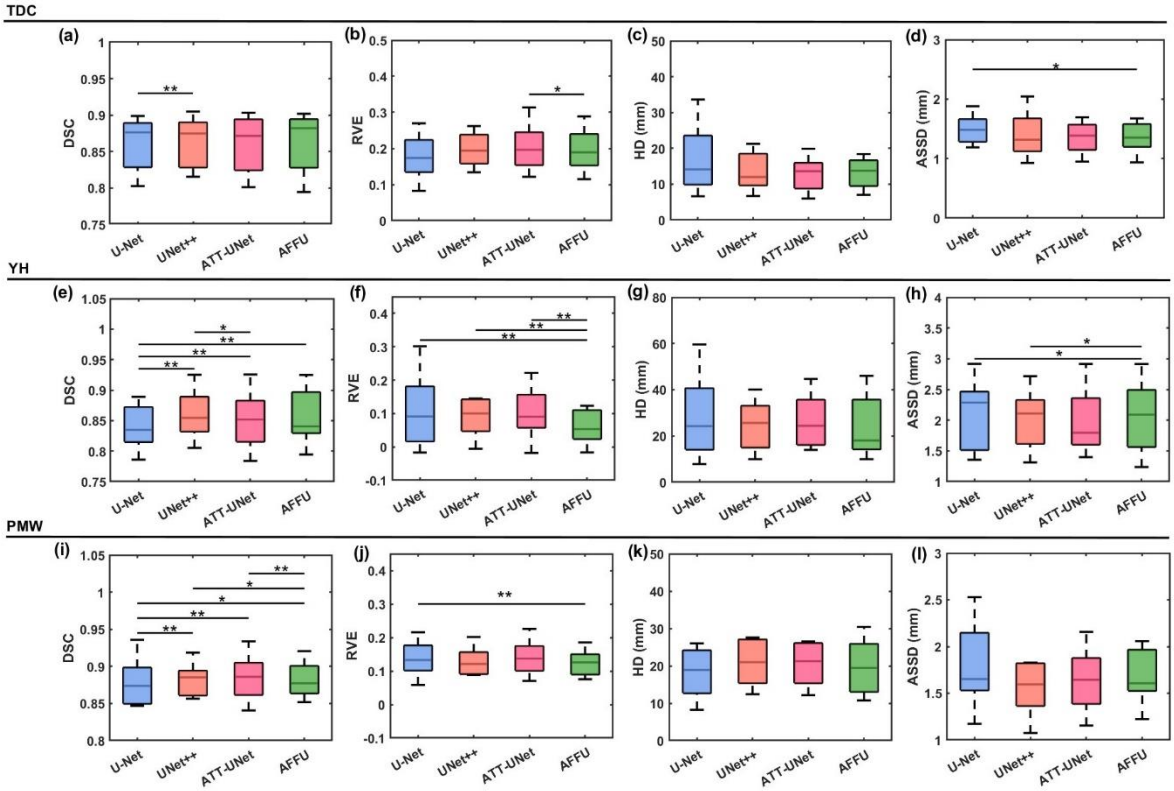


Figure 4.11 Performance comparison of each model under different cohort. Box plots of four metrics for U-Net (blue), UNet++ (orange), Attention-UNet (red) and AFFU (green) are shown (* indicates $p < 0.05$, ** indicates $p < 0.01$). Each boxplot includes 15 medians of each muscle from 3 testing subjects. Panels (a)-(d), (e)-(h) and (i)-(l) represent the results for TDC, YH and PMW, respectively.

Table 4.13 The number of trainable parameters (rounded to millions) of each model and training time (per epoch).

	U-Net	UNet++	Attention UNet	AFFU
Size	31M	10M	32M	38M
Time	1.2 min	1.5 min	1.2 min	7.8 min

AFFU has the highest number of model parameters, with 38 million, and is the most time-consuming, requiring 8 minutes per epoch. In contrast, UNet++ has the smallest model size, with only 10 million parameters, and takes less than 2 minutes per epoch on average.

4.4 Discussion

Discussion about reproducibility on TDC cohort

In operator reproducibility analysis (Table 4.3), 18 muscles achieved high reproducibility scores in the intra-operator assessments, with 15 of these consistent with the results in reference [6] where the same cohort PMW includes 11 post-menopausal women was used. Additionally, 8 muscles demonstrated high reproducibility in the inter-operator assessments, with 7 consistent with the results in reference [6]. For these 18 muscles, the coefficients of variation (CV) for volume and length, obtained from the same operator's segmentations across three different subjects (Table 4.4), indicated that for most muscles, the operator's error across the three separate segmentation instances for each subject was kept below 5%. This consistency suggests that the operator's ability to distinguish most muscles remained reliable, with low variation due to cohort differences, whether in typically developing children or healthy post-menopausal women.

For the Adductor brevis, Gemellus superior, Obturator externus, Piriformis, and Quadratus femoris muscles, the precision errors (PE) for both volume and length exceeded 10% in both the inter- and intra-operator analyses. This indicates that within the TDC cohort, these muscles were challenging to segment consistently, both between different operators and across multiple instances by the same operator. As reported in Figure 4.13, Adductor brevis and Obturator externus are surrounded by several other muscles, making their boundaries difficult to define clearly and increasing the likelihood of confusion with adjacent muscles. Meanwhile, Gemellus superior, Piriformis, and Quadratus femoris, located close to the hip part and relatively small in volume (10-20 cm^3, Table 4.5), were more susceptible to regional noise, leading to lower reproducibility.

In the bilateral muscle parameter comparison, the Gluteus maximus, Vastus lateralis, and Adductor magnus exhibited significant volume differences, consistent with the findings reported in reference [6] using same PWM cohort, with the average volume of the right-side muscles being greater than that of the left-side muscles. Other muscles, such as the Rectus femoris and Gracilis, also showed significant differences in either volume or length. This clearly indicates that caution should be exercised when assuming limb symmetry in the assignment of children's musculotendon parameters. which is also mentioned in reference [6] for healthy populations.

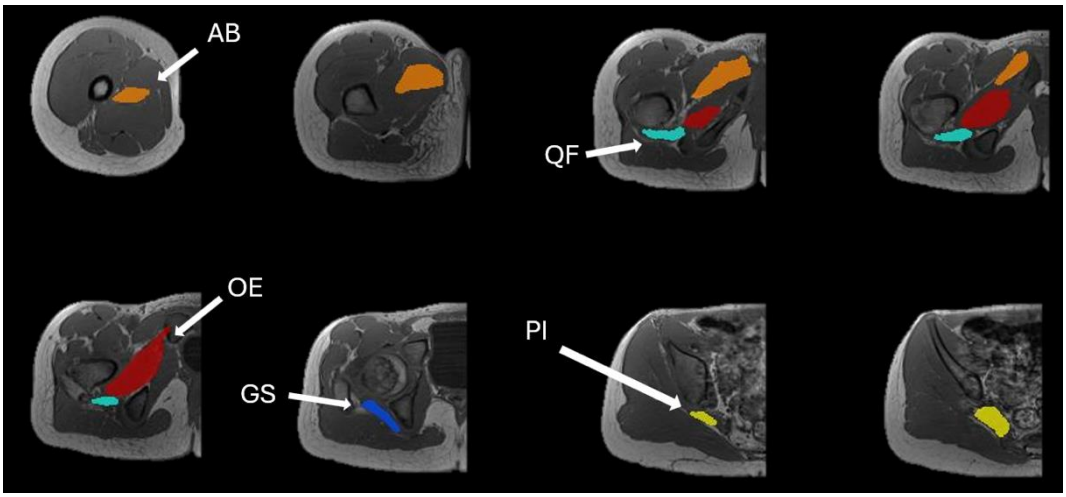


Figure 4.13 Visualization of manual segmentation on muscles with low reproducibility. 5 muscles were included and in different colour. (AB: Adductor brevis (orange), GS: Gemellus superior (blue), OE: Obturator externus (red), PI: Piriformis (yellow), QF: Quadratus femoris (tiffany blue))

Discussion about task 1

The results from testing models trained on PMW and evaluated on TDC cohort reveal insights into the generalization capabilities of DL models across distinct population groups. All models didn't show good performance on any of the calculated metrics. The poorer performance of both AFFU and U-Net on the TDC cohort compared to the PMW cohort underscores the challenges of applying DL models across different age groups and body compositions. Compared to performance on the PMW dataset [14], both AFFU and U-Net showed worse results on the TDC dataset, with DSC scores decreasing by 32.5% and 34.4%, respectively. Conversely, when trained on TDC and tested on PMW and HY cohorts, the models also demonstrated unsatisfactory performance (Table 4.10).

Specifically, the decrease highlights how anatomical differences between old women and children may significantly hinder the models' ability to accurately segment muscle structures. Factors such as variations in muscle size, fat tissue distribution, and tissue characteristics between these groups likely contribute to the decline in segmentation accuracy, emphasizes the importance of training or fine-tuning models on datasets that are representative of the target population [107], [108].

Discussion about task 2

The results demonstrate that training on a more diverse dataset, incorporating images from both sides of the lower limbs, leads to improved performance. This improvement highlights the importance of a balanced dataset, particularly in medical imaging tasks, where anatomical differences between sides could influence the segmentation outcome [16], [17].

In the first stage of task 2, when only the right-side images were used for training, all models exhibited a performance bias toward the right muscles, with significant drops in accuracy of left side, especially in metrics such as HD and ASSD. These differences were most evident in models like UNet++ and Attention UNet, underscoring the difficulty of generalizing from a single side of the body to both. This phenomenon may be caused by differences in dominant feet (left or right) [109] or other imbalance during muscle growth.

However, once bilateral data were incorporated, the gap between the two sides narrowed, indicating better model generalization. The AFFU model consistently showed strong performance across all metrics, especially in handling the variability between the left and right muscles. The model's performance in RVE, HD, and ASSD suggests that it could better handle volume differences and boundary delineations, making it a robust option for muscle segmentation tasks (Figure 4.7). The consistent performance of the AFFU model across both stages indicates that advanced architectures can capitalize on more comprehensive datasets to enhance segmentation accuracy.

From Table 4.8, it is evident that when training is performed only on the right-side muscles, the prediction results for both sides of 10 muscles show significant differences between both sides. In contrast, the results for the remaining muscles are relatively similar between both sides. Therefore, in further study, if dataset is small and unilateral training just on one side muscles is performed, it might be advisable to exclude the other side’s limb prediction results for these 10 muscles from the analysis after model training.

Discussion about task 3

When comparing the performance of the models across all age groups combined (Table 4.11), AFFU consistently demonstrates superior performance compared to U-Net, UNet++, and Attention-UNet.

AFFU marginally outperforms all other models in DSC (0.872 ± 0.041), demonstrating slightly better segmentation accuracy. The improvement, although modest, signifies that AFFU is highly effective in matching predicted segmentations with ground-truth labels,

suggesting a superior overall capability in capturing detailed features. Even small amounts of noise from mis-segmented pixels can have a significant impact on texture analysis based on medical image segmentation[106], [110]. It also shows a significant reduction in RVE (0.127 ± 0.117) compared to the other models, especially over the baseline U-Net. This improvement indicates that AFFU is particularly adept at handling variations in object size, which is often challenging in medical image segmentation tasks[14]. The reduced error reflects its ability to maintain consistency across multiple instances. In terms of HD and ASSD, there is not significant difference among all models.

When analyzing performance across different cohorts—children (TDC), healthy young people (HY), and post-menopausal women (PMW)—the models exhibit varying degrees of success, and AFFU's generalization ability shows robustness.

In TDC, AFFU did well in reducing volume error rates and improving surface distance accuracy. Although the overall segmentation accuracy is relatively comparable across models, AFFU's ability to maintain low error margins indicates better generalization for this younger age group.

For healthy young people, AFFU demonstrates its most significant improvements in this cohort. It consistently outperforms other models in all key metrics, indicating that it generalizes exceptionally well to this population. The substantial improvement in volume error reduction, along with high boundary accuracy, shows AFFU's strength in segmenting the often less predictable anatomical features of younger individuals.

For PMW, AFFU's performance remains competitive, but its advantage is less pronounced compared to other models, particularly in DSC where UNet++ and Attention-UNet slightly outperform it. However, AFFU still shows better scores in reducing volume errors, reinforcing its general ability to provide accurate predictions in more complex cases. This indicates that, while it may not always achieve the highest accuracy for certain age groups, giving different performance in the same cohort reported in other study [14] which only trained and tested in PMW, it still maintains consistency and reliability.

In Task 3, although the batch size was adjusted to be 8 for AFFU with more complex architecture to better accommodate GPU performance, the training time for AFFU was still 8 times longer than that of U-Net, UNet++, and Attention-UNet. Therefore, while AFFU can be recommended for use with smaller datasets, it is insufficient to solely pursue modifications to model structures without considering the anatomical features inherent in the data itself when

dealing with larger datasets. Future research should explore faster post-processing methods that incorporate muscle shape information to refine the predicted labels generated by the model.

4.5 Conclusion

In this chapter, we explored the impact of various models on different cohort's muscle segmentation, with a particular focus on how different training datasets influenced the outcomes. The findings clearly demonstrate that the quality and relevance of the training dataset play a more crucial role in achieving high performance than the choice of the model itself. While different models can yield varying results, it is the richness and representativeness of the data from certain cohort used to train these models that most significantly determines their ability to generalize and perform well across diverse cohort, from children, young people to old women.

We also evaluated the performance and generalization capabilities of different popular used DL models, with a particular focus on AFFU, in segmenting muscle structures across three distinct population cohorts: typical developed children, young adults, and post-menopausal women. Through comprehensive comparisons with other models (U-Net, UNet++, and Attention-UNet), AFFU consistently demonstrated advantages, particularly in reducing volume errors and maintaining spatial accuracy. AFFU's architecture proved highly adaptable, excelling in younger cohorts where anatomical variability is more pronounced [111], [112]. While its performance remained competitive in older populations, the gap narrowed slightly compared to UNet++ and Attention-UNet. However, AFFU's ability to generalize effectively across diverse datasets highlights its robustness, even when trained on different cohorts.

A notable limitation of AFFU is its longer training time (e.g. 5 times higher in Task 3, Table 4.13) compared to other models. However, its superior performance on critical metrics such as boundary delineation and error reduction make it well-suited for applications, particularly in cases where training efficiency is less critical, such as with smaller datasets. Future research should focus on leveraging external information to further optimize the prediction results of DL models, aiming to achieve higher accuracy while maintaining fast training speeds.

Chapter 5 Mean shape based post-processing method for enhancing deep learning lower-limb muscle segmentation accuracy

This chapter is based on a paper published in PLOS ONE (2024): 'A novel mean shape based post-processing method for enhancing deep learning lower-limb muscle segmentation accuracy' by **Lin Z**, Dall'Ara E, Guo L. Doi: https://doi.org/10.1371/journal.pone.0308664

5.1 Introduction

In the previous chapters, we proposed an enhanced muscle segmentation model based on U-Net, AFFU, and designed a novel data augmentation based on SSM technique to address the issue of limited training dataset, demonstrating improvements in segmentation accuracy. We also conducted a generalization evaluation of various deep learning (DL) models, including the proposed AFFU, across different age groups—elderly women, young adults, and typical developed children. These experiments highlighted that training models on a single cohort significantly limits their generalization ability to another cohort.

While continuous modification and improvements to DL model architectures and data augmentation techniques are valuable, they do not always provide the directional corrections necessary to enhance model performance meaningfully. One of the drawbacks of deep learning is its lack of interpretability. Blindly modifying the structure and trying different combinations of various mechanism modules like attention gate or transformer is cumbersome and insufficient. Therefore, Chapter 5 will focus on the crucial aspect of post-processing for muscle segmentation.

In muscle segmentation post-processing, it is crucial to consider muscle anatomical area, volume, and age-related regional muscle loss. Innovative techniques should focus on accurately measuring muscle average area essential for studying how muscle mass affects muscle force changes with age and training [113]. Our study is centred on this pivotal aspect, aiming to enhance the accuracy of lower limb muscle segmentation by developing innovative post-processing techniques.

In the field of medical imaging, particularly in muscle segmentation, post-processing stands as a crucial step to refine and validate the results obtained from initial segmentation algorithms (e.g., convolutional neural networks (CNNs)). In a study [114] on the automatic segmentation of lumbar paraspinal muscles, the impact of CNN architecture and training was evaluated. The study emphasized the use of post-processing transformations, such as spatial connection and closing analysis, to reduce false positives and further improve CNN accuracy. Another study [115] focused on multi-muscle deep learning segmentation for the quantification of muscle fat infiltration in cervical spine conditions. The study reported high accuracy and reliability for the CNN model and highlighted the use of post-processing techniques to minimize bias and improve segmentation accuracy. Furthermore, research on transfer learning for data-efficient abdominal muscle segmentation with CNNs emphasized

the potential of transfer learning to produce data-efficient skeletal muscle segmentation models, highlighting the importance of leveraging advanced techniques to improve the efficiency of muscle segmentation [116]. For lower-limb muscle segmentation task, some morphological operations like hole-filling [17] and binary closing [2] methods were utilized to fill the empty holes or remove the outliers. In summary, post-processing techniques play a crucial role in refining and validating muscle segmentation results obtained from CNNs, ultimately improving the accuracy and efficiency of medical imaging analysis. This process is not merely a technical afterthought but a pivotal component that significantly impacts the overall accuracy and clinical utility of segmentation. The traditional post-processing techniques based on pixel grey values like thresholding and region growing, which are straightforward and effective in certain natural images, often fall short in dealing with the complexity of muscle tissues, struggling to differentiate between muscle groups and adjacent structures with similar densities or intensities.

In recent years, some new post-processing methods for enhancing image segmentation have been developed. For example, in a task of liver segmentation [73], assuming an organ should be a single volume with the maximum sized label region, a 2D post-processing method based on erosion operation, identification, and selection of the largest region, dilation operation, and edge smoothing was utilized to achieve a 4% percentage gain of Intersection over Union (IoU). The first three steps were based on the process of the normal images and operated in MATLAB (function: imerode, bwlabel, and imdilate). Edge smoothing was achieved through a 2D convolution with a uniform square filter applied to the binary labelled mask. This 2D post-processing algorithm was significant for its ability to refine liver segmentation outputs, particularly in isolating the main liver region from extraneous segments and smoothing the edges for a more accurate representation of the liver's shape and boundaries. Another example is the Conditional Random Field (CRF) post-processing methods, as presented in [18] and [74]. It represents a significant advancement in the field of image segmentation, particularly in refining the segmentation results of Deep Convolutional Neural Networks (DCNNs). CRF was originally attached to the end of DeepLab V1[18], and achieved 4% improvement of IoU. The CRF in DeepLab V1 focused on recovering detailed local structure, instead of further smoothing the already smooth outputs from modern DCNNs. The DenseCRF proposed in [74] presents an efficient inference algorithm for fully connected CRFs, increasing the performance of the unary classifiers on the VOC 2010 from 13% to 22% accuracy.

Although they have achieved commendable segmentation benefits, previous studies did not consider the 2D or 3D shape of the segmentation targets. This is particularly relevant for muscles, which, despite varying in volume across different subjects, share similarities in spatial morphology and topological structure [14, 15]. The effectiveness of CRF and 2D post-processing methods depends considerably on the performance of the underlying CNN models; if the models do not achieve accurate results or good feature maps, the post-processing techniques will not be likely to yield significant improvements. Incorporating mean shapes or representative shapes for optimizing the segmentation results can enable the inclusion of the anatomical significance of the segmentation targets or in biomedical modelling, leading to potentially more accurate results, such as in bone segmentation [11] and in human lower-limb anatomy modelling [117]. In this study, we propose a new post-processing approach based on the mean shape (MS) and a Statistical Shape Model (SSM) [10] to further refine deep learning-based muscle segmentation results, with a focus on accurately delineating muscle boundaries and reducing false positives.

The goal of this chapter was to propose a new mean shape based post-processing method and to assess the effect of it with existing methods (CRF, and 2D post-processing method) on the accuracy of a previously developed CNN approach to segment individual muscles of the lower limb from MRI scans [4]. The secondary goal of the study was to compare the accuracy of different segmentation tools, including that of a commercial semi-automatic muscle segmentation toolbox (Mimics 26.0, Materialise).

5.2 Materials and methods

5.2.1 Participants

Two groups of different cohorts from healthy post-menopausal women (PMW) were used in the chapter. The first cohort (PMW-1) includes 10 PMW T1-weighted magnetic resonance images with no muscle disease (also used in Chapter 3 and 4) [6] and the second (PMW-2) was recruited from a previous observational study and consisted of 8 healthy PMW without muscle diseases (also used in Chapter 3) [83]. The detailed parameters of both cohorts are given below (Table 5.1). Although the two cohorts were imaged in separate studies, the scans were conducted at the same hospital using two different 1.5T MRI scanners, following the same scanning protocol. Detailed scanning and post-processing operations using MATLAB (R2006a) and manual segmentation processes using Mimics (version 26.0; Materialise) and Amira (version 2022.2; Thermofisher) have been already described in Chapter 3 (section 3.2.1).

Table 5.1 Statistical information.

Cohort	Age (year)	Weight (kg)	Height (cm)	BMI
PMW-1	69.0±6.7	66.9±7.7	159±3	26.5±3.4
PMW-2	65.8±3.9	57.1±5.8	161±3	22.0±2.1

Considering the inter- and intra-operator reproducibility, manual segmented results was found to vary significantly between different muscles, with a coefficient of variation (CoV) ranging from 4.2% to 22.8% [6], [7]. In this section, 25 muscles from PMW-1, covering the area from knee to hip same as Chapter 3, were used during model training. Of these, 16 muscles with higher reproducibility were chosen for further analysis, including rectus femoris (RF), vastus intermedius (VI), vastus lateralis (VL), vastus medialis (VM), sartorius (SAT), semimembranosus (SMB), semitendinosus (SMT), gracilis (GRA), biceps femoris caput brevis (BCB), biceps femoris caput longum (BCL), adductor magnus (AM), adductor brevis (AB), adductor longus (AL), gluteus maximus (GM), iliacus (IL), and tensor fasciae latae (TFL).

5.2.2 Experimental design and U-Net

U-Net [81] was used as the deep learning (DL) network during the experimental phase, the model layers setting detail is given in Table 5.2 and the flowchart of structure is shown in Figure 2.8 in Chapter 2. The primary objective of this chapter is to investigate whether the proposed mean shape based post-processing method improves segmentation results. To achieve this, as a widely used segmentation model, U-Net was chosen for training. Compared to other advanced models, U-Net has fewer parameters, ensuring it does not excessively consume computational resources. Additionally, if the post-processing method proposed in this study prove to be effective, it can be easily and directly applied to the output results of other studies that use U-Net for training. This will facilitate direct comparisons between the original results and those obtained using our method.

Table 5.2 U-Net layers setting.

1	Im input 256x256x1	18	Depth concat of 2 inputs
2	64 3x3 convs, stride [1,1], BN + ReLU	19	512 3x3 convs, stride [1,1], BN + ReLU
3	64 3x3 convs, stride [1,1], BN + ReLU	20	512 3x3 convs, stride [1,1], BN + ReLU
4	2x2 max pool stride [2,2]	21	256 2x2 transp convs stride [2,2], BN +ReLU
5	128 3x3 convs, stride [1,1], BN + ReLU	22	Depth concat of 2 inputs
6	128 3x3 convs, stride [1,1], BN + ReLU	23	256 3x3 convs, stride [1,1], BN + ReLU
7	2x2 max pool stride [2,2]	24	256 3x3 convs, stride [1,1], BN + ReLU
8	256 3x3 convs, stride [1,1], BN + ReLU	25	128 2x2 transp convs stride [2,2], BN +ReLU
9	256 3x3 convs, stride [1,1], BN + ReLU	26	Depth concat of 2 inputs
10	2x2 max pool stride [2,2]	27	128 3x3 convs, stride [1,1], BN + ReLU
11	512 3x3 convs, stride [1,1], BN + ReLU	28	128 3x3 convs, stride [1,1], BN + ReLU
12	512 3x3 convs, stride [1,1], BN + ReLU	29	64 2x2 transp convs stride [2,2], BN +ReLU
13	2x2 max pool stride [2,2]	30	Depth concat of 2 inputs
14	1024 3x3 convs, stride [1,1], BN + ReLU	31	64 3x3 convs, stride [1,1], BN + ReLU
15	1024 3x3 convs, stride [1,1], BN + ReLU	32	64 3x3 convs, stride [1,1], BN + ReLU
16	2x2 max pool stride [2,2]	33	26 1x1 convs stride [1,1]
17	512 2x2 transp convs stride [2,2], BN +ReLU	34	Softmax

During the training period, the dataset, consisting of 10 samples where arounds 400 slices included in each, was randomly split into training and validation sets with a 9:1 ratio. This split ensures that the model is adequately trained on a larger dataset while allowing validation on a separate subset. The initial learning rate for the model was set to 0.001, which decayed multiplicatively by a factor of 0.9 every 5 epochs, over a total of 100 epochs. The model was trained with the multi-class cross entropy loss function. Other hyper-parameters, including batch size, dropout rate, and regularization strength, were tuned empirically using grid search and cross-validation for optimal network performance, ensuring that GPU memory and capacity were not exceeded. After the loss converged on the training set, the best-performing set of weights from each epoch on the validation set was retained. The PMW-1 cohort was employed as the training set for the U-Net, while the PMW-2 cohort served as the test set to evaluate model performance and obtain preliminary results (Figure 5.1). To decrease the effect caused by model uncertainty, the process above was repeated three times. The model's preliminary prediction results were retained as a control group and used as input for next post-processing steps. After obtaining the U-Net output results, rough 3D morphological operations, including erosion and dilation, were applied in MATLAB to decrease local segmentation errors and improve the overall quality of the segmentation. Subsequently, 2D post-processing methods described in section 2.5.2 [73] or the mean shape-based post-processing method outlined in section 2.3 were applied to refine the segmentation results, yielding 'Result—U-Net+2D' and 'Result—U-Net+MS' (Figure 5.1), respectively. As an alternative approach, the label map outputted from the last SoftMax layer of U-Net was used as input into the Conditional Random Field (CRF) post-processing method described in section 2.5.1, resulting in 'Result—U-Net+CRF'. The score for each segmentation metric, including DSC, RVE, HD and ASSD, was defined as the average of the three repetitions for each subject. Initially, the corresponding muscle areas in each lower limb MR image of PMW-2 were manually selected using a thresholding technique, which involved setting intensity value thresholds to delineate muscle regions, followed by simulation to obtain the results labeled as 'Result—Mimics' (Figure 5.1).

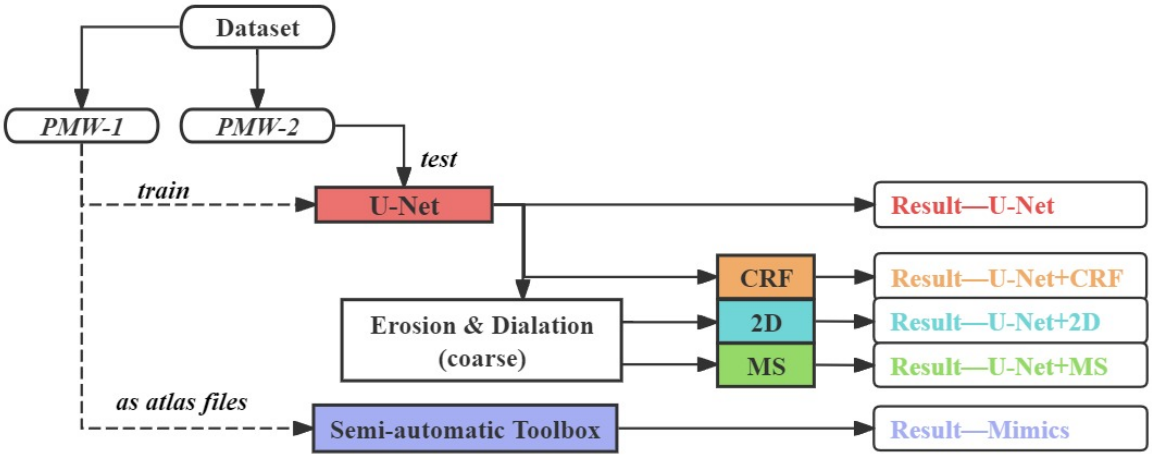


Figure 5.1 Experimental design pipeline.

5.2.3 Proposed mean shape (MS) based post-processing

Considering the background that the same muscles in different individuals exhibit similar three-dimensional (3D) shapes and topological structures [75], [76], this chapter proposed a post-processing method based on Statistical Shape Models (SSM). This similarity is crucial for enhancing the accuracy of segmentation results. This method aimed to further optimize the predicted segmentation results of the DL model by using the geometrical mean shape information obtained from the Statistical Shape Models (SSM) in combination with deformable registration techniques. The non-linear deformable image registration algorithm, ShIRT [9], [89], was employed to align the predicted segmentation with the anatomical structure, improving the accuracy.

To establish homologous surface points within a dataset, three-dimensional surfaces are sampled to identify corresponding points, which serve as references for shape analysis. These points are represented as a set of $3n$-dimensional vectors, where n is the number of points selected for shape representation and statistical analysis. The accuracy of this representation is heavily reliant on the quality of point selection. Typically, corresponding points are manually chosen as landmarks on the target surface. However, as the volume of data increases, manual selection becomes cumbersome and often unsatisfactory, leading to potential inaccuracies in the analysis [10]. To tackle this challenge, parametric methods have been proposed. One such method, detailed in [88], employs a fixed spherical parameterization mapping to facilitate point selection, providing a systematic approach to establishing correspondence points. However, most organs' shapes cannot be accurately parameterized, so these methods still have limitations. Cates [10] introduced a non-parametric analysis system called the Shape Modelling with Entropy-Based Particle Systems, known as the ShapeWorks correspondence optimization algorithm, which can optimize the distribution of a set of shape-similar correspondences and seeks a compact distribution of corresponding points in the shape space through an optimization algorithm to simplify the model. The fundamental concept of ShapeWorks involves creating a dynamic set of particles, where each particle represents a collection of correspondence points constrained to the surface of a shape. This allows for flexible adaptation to varying shapes and structures. The detail of the algorithm has been given in Section 3.2.3 in Chapter 3.

In this chapter, the non-parametric analysis system proposed by Cates [10], [88], designed for shape correspondence optimization, was utilized through the publicly available software tool

ShapeWorks. The same in Chapter 3, to use the software, ***n*** shape models representing the same muscle from different subjects were provided as input. The software then generated a mean shape model, which serves as a reference point for further analysis. From this mean shape model, n-1 modes of variation are extracted using principal component analysis (PCA).

The detail operation is given in **Figure 5.2**.

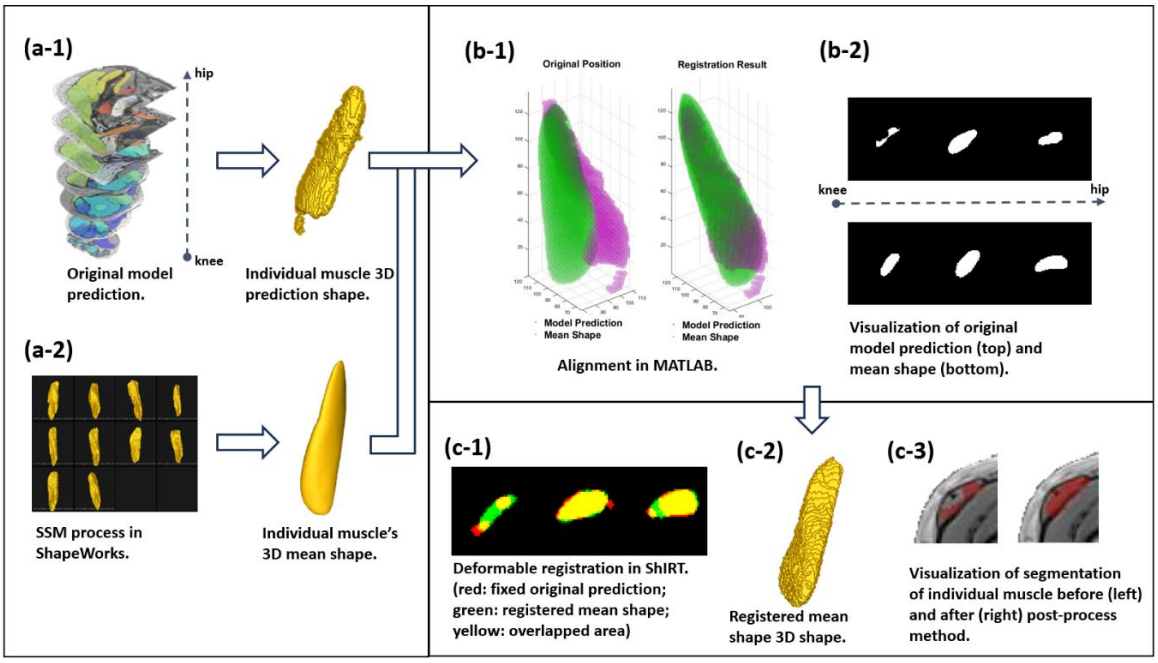


Figure 5.2 Pipeline for mean shape based post-processing method. Initially, the muscle segmentation results (original predictions) from the neural network were stacked, and subsequently, each muscle label was extracted to generate distinct 3D shapes (a-1). Using the ShapeWorks software, the mean shape for each muscle was generated (a-2) using gold standard manual segmentation from PMW-1. To reduce substantial spatial disparities between the original prediction and the mean shape, a rigid registration based on an iterative closest point algorithm (ICP) was performed in MATLAB (fixed shape: original prediction; moving shape: mean shape; function: pcregistericp (moving, fixed)) (b-1). Variations in muscle (e.g., TFL) slices at corresponding positions of both, from the knee to the hip, are depicted (b-2). In the final step, the two shapes underwent deformable registration in ShIRT, and the registered mean shape was retained as the post-processing segmentation results (c-1/2/3).

After generating the mean shape (MS) of each individual muscle and performing alignment operations in MATLAB, the ShIRT tool was employed to carry out deformable image registration between the model-predicted shape and the corresponding MS. ShIRT is a non-linear deformable image registration algorithm (Chapter 3), specifically designed to address the significant anatomical variability commonly encountered in medical images [9], [89].

This method enables the flexible alignment of images with complex anatomical differences. It calculates a displacement function that maps each voxel (a 3D pixel representing a volume element) in the reference image to its corresponding point in the target image. Using an iterative optimization process, the algorithm refines the mapping matrix to minimize the cost function. This function is calculated by summing the squared differences between the intensity values of corresponding voxels in the two images. ShIRT calculates node displacements using an isotropic hexahedral grid, which is overlaid onto both the fixed and moving images. The distance between nodes, known as Nodal Spacing (NS), ensures uniform spacing across all dimensions for accurate registration. During registration, optimal nodal displacements are continuously smoothed using a coefficient λ, which helps control the smoothness of the displacement field and ensures efficient convergence of the registration process. The algorithm uses trilinear interpolation between grid nodes to generate a continuous 3D displacement field, ensuring smooth transitions between voxel displacements. This displacement field is then applied to transform the reference image into alignment with the target image. Previous sensitivity analyses have shown that the optimal Nodal Spacing (NS) for muscle segmentation tasks is 5 voxels (equivalent to 5 mm) [9], which was adopted in this study. During hyper-parameter tuning, the Nodal Spacing (NS) was fixed at 5 mm, while the smoothing coefficient λ was automatically optimized by the software to achieve the best registration performance.

5.2.4 Semi-automatic muscle segmentation in Mimics for comparison

The Muscle Segmentation toolbox in Mimics (Mimics 26.0) allows for the semi-automatic segmentation of multiple lower limb muscles simultaneously, as opposed to segmenting them individually.

In this section, semi-automatic means that manual adjustment of parameters is required to define the target muscle area for segmentation before running the software.

Before using the toolbox, atlas files must be created from previously segmented samples, which act as reference templates for the software during the segmentation process. The Muscle Segmentation tool contains a voting mechanism between the atlases with the highest correspondence that determines whether to assign a certain voxel to a certain muscle. The

Mimics includes an atlas file package containing 10 subjects (5 males and 5 females) aged between 23 and 60 years. Given the significant difference between the demographic range of these samples and the testing cohort used in this chapter, PMW-1 was selected as the atlas file for the software, ensuring good anatomical relevance and alignment with the post-menopausal cohort being studied.

The atlas files consist of MR scan images overlaid with masks (Figure 5.3 (a), in green) that define the areas to be segmented, along with individual masks for each segmented muscle (Figure 5.3 (b), in different color), which are used to guide the segmentation process. Prior to running the program, a manual mask (Figure 5.3 (c), in green) must be added on the top of the original image. Gradient thresholding technique is then applied to ensure the mask closely fits the muscle area, while minimizing background noise. Afterward, the MR images of new samples are input into the toolbox. Upon running the embedded algorithms, segmentation results for each muscle are generated.

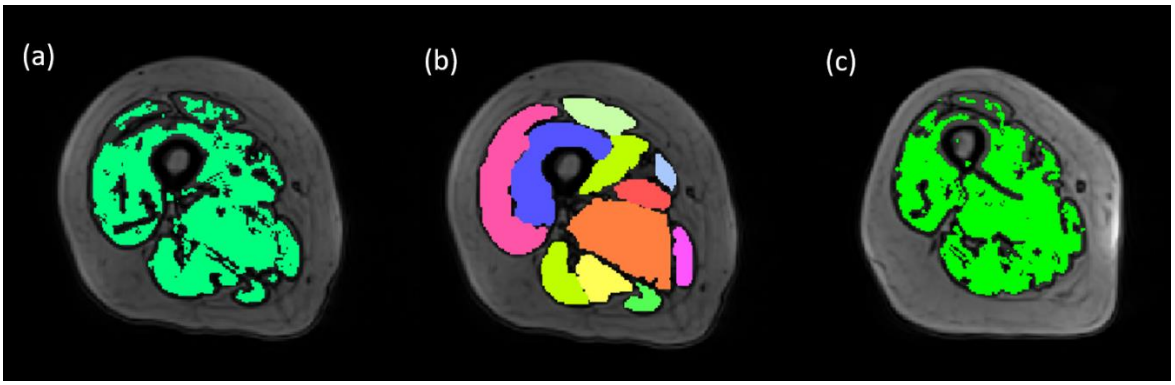


Figure 5.3 The mask and segmentation visualization of atlas and new subjects in Mimics.

5.2.5 Existing post-processing methods for comparison

5.2.5.1 Fully connected conditional random fields (CRFs)

In [74], a fully connected CRF method that utilized pairwise potentials among all pixels in the image was proposed to compensate for the restriction in complexity of inference and the limitation from unsupervised segmentation [118], [119]. The pairwise edge potentials were defined using a set of Gaussian kernels in an arbitrary feature space. This method was based on a mean field approximation to the CRF distribution which was optimized through message passing steps [74] to avoid the sudden increase in computation caused by image size. Fully

connected conditional random fields conform to the Gibbs distribution [120], as shown in the following formula:

$$P(x = X|I) = \frac{1}{Z(I)} e^{-E(X|I)} \tag{5.1}$$

where x is the observed value, I is the colour information of pixel, $E(X|I)$ is the energy function which consists of a unary potential function and a binary potential function, as shown in the following formula:

$$E(x|I) = \sum_i \psi_u(x_i) + \sum_{i,j} \psi_p(x_i, y_j) \tag{5.2}$$

where x_i, y_j are the labels assigned to pixel i and j, respectively.

The unary potential function $\psi_u(x_i)$ is used to measure the probability of the observed pixel to different classes which is from the output feature maps of CNNs and defined as:

$$\psi_u(x_i) = -\log P(x_i|I) \tag{5.3}$$

If $P(x_i|I)$ is very high (close to 1), the model has a high degree of confidence of pixel i in the prediction belongs to class x_i. $\psi_u(x_i)$ will tend to 0. If $P(x_i|I)$ is very low (close to 0), $\psi_u(x_i)$ will rise, making $E(X|I)$ larger. The unary term should be as small as possible to show that the model is confident in the current prediction and meets prediction goals.

A binary potential function measures the probability $\psi_p(x_i, y_j)$ of two events occurring at the same time where if the pixel values are very close, then the probability that the two pixels belong to the same category should be relatively high; on the contrary, if the colour difference is relatively large, then the probability of these two pixels belong to different category should be relatively high. This energy term is designed to split the segmentation results as far as possible from the edges of the image or between different targets.

$$\psi_p(x_i, y_j) = u(x_i, y_j) \sum \omega^m K_G^m(f_i, f_j) \tag{5.4}$$

$u(x_i, y_j)$ is the label compatibility item which restricts the conditions of conduction between pixels, and only under the different label conditions can energy be transmitted to each other, $u(x_i, y_j) = 1, if\ x_i \neq y_j,\ u(x_i, y_j) = 0, if\ x_i = y_j$.

ω^m is the weight parameter. $K_G^m(f_i, f_j)$ is Gaussian kernel including appearance kernel and smoothness kernel in terms of colour vector I_i and I_j and positions p_i and p_j, and defined as below:

$$K_G(f_i, f_j) = W^{(1)} e^{-\frac{|p_i - p_j|^2}{2\theta_\alpha^2} - \frac{|I_i - I_j|^2}{2\theta_\beta^2}} + W^{(2)} e^{-\frac{|p_i - p_j|^2}{2\theta_\gamma^2}} \quad (5.5)$$

$W^{(1)}$ and $W^{(2)}$ are the weight of appearance kernel and smoothness kernel, respectively. The concept behind the appearance kernel stems from the observation that adjacent pixels sharing similar colours are probable candidates for belonging to the same class. The weights of nearness and similarity are controlled by the parameters θ_α and θ_β. The smoothness kernel removes small, isolated regions and is controlled by θ_γ. This is the desired result, because if the pixels are physically close and the colors are similar, their labels should also be the same.

When adjacent pixel *i* and *j* share the same label ($x_i = y_j$), pairwise term $\psi_p(x_i, y_j)$ will be 0, and contributes nothing to energy function. This is the desired result, because if the pixels are physically close and the colors are similar, their labels should also be the same.

When $x_i \neq y_j$, $u(x_i, y_j) = 1$, in this case, the value of the Gaussian kernel determines the size of the pairwise term $\psi_p(x_i, y_j)$. If the difference of pixel *i* and *j* on position and color is obvious, $K_G(f_i, f_j)$ will close to 0, and contribute less to energy function which the model allows. If the position and color information of pixel *i* and *j* is very similar, but with different predicted label, the value of pairwise term will increase, with energy term increases also. In this case, CRF will modify the prediction results by minimize the energy term to make their label same.

The detailed derivation of CRFs can be found in [74]. In Python, the CRF method can be implemented using the 'pydensecrf' library provided where $W^{(1)}$, θ_α and θ_β was set to 10, $W^{(2)}$and θ_γ was set to 1 during learning in [74].

5.2.5.2 2D post-processing method

In this algorithm, an erosion operator (using the MATLAB function *imerode*) was initially applied to remove small noise from the main cohort. Next, 2D connected components were identified using the *bwlabel* function. Assuming that the target region is a single volume with

a smooth surface, the largest connected area was retained, and any empty holes within it were filled using function *imfill* and *dilate*. A 2D convolution with a uniform square filter (*conv2D*) was then applied to smooth the result. These operations were performed on each predicted 2D slice, with tuning details provided in Table 5.3 [73]. Unlike single-class segmentation, such as liver segmentation, the multi-class segmentation of muscles in this study required separate extraction and post-processing of each muscle, which were then combined for the final output.

Table 5.3 Parameters used in the 2D post-processing method.

Operation	MATLAB function	Parameters
Erode	imerode(label, SE) SE = strel('disk',r)	$r = 3$
Bwlabel	bwlabel(label,conn)	conn = 8
Dilate	Imdilate(label,SE) SE = strel('disk',r)	$r = 3$
Fill holes	imfill(label,'holes')	
Smoothness	kernel = ones(w)/w^2 image = conv2(single(label),kernel,'same') image = image>threshold	$w = 5$ threshold = 0.5

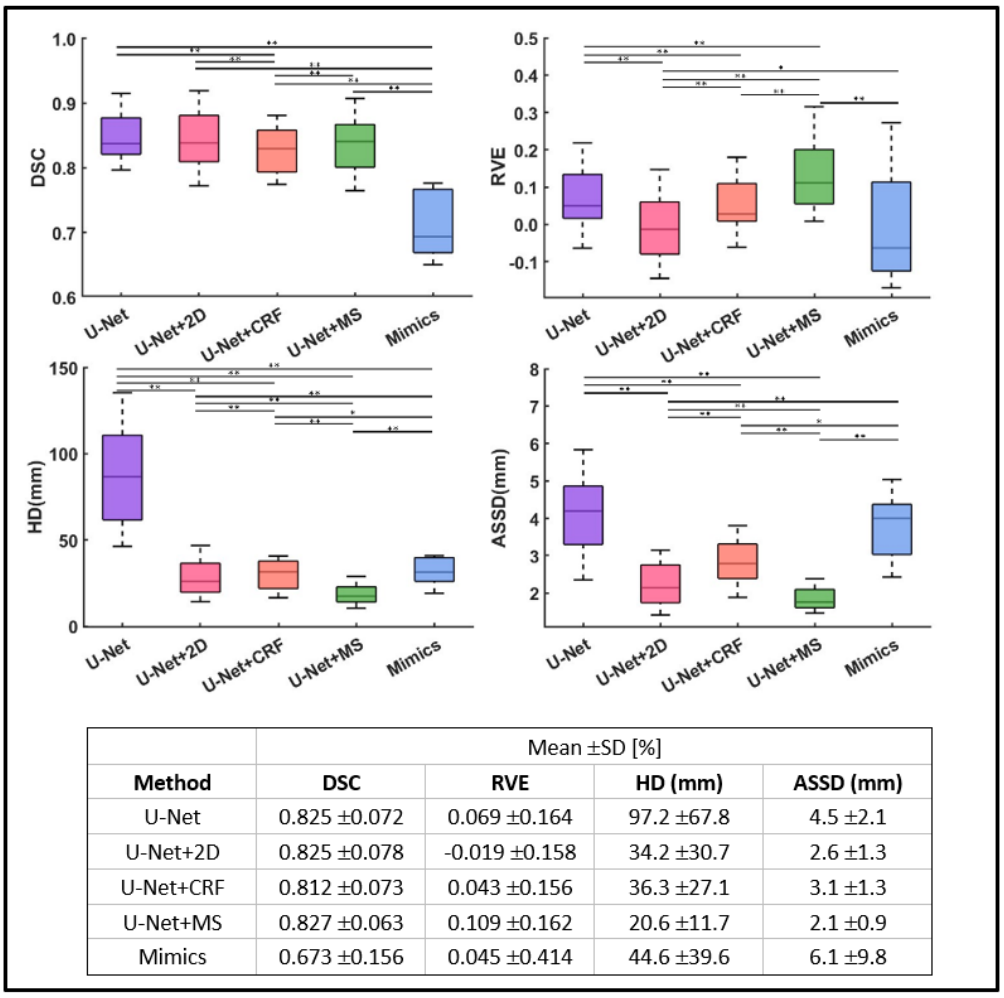
5.2.6 Evaluation metrics and statistics

Four metrics were used in this study: Dice Similarity Coefficient (DSC), Relative Volume Error (RVE), Hausdorff Distance (HD), and Average Symmetric Surface Distance (ASSD) [2, 21, 27]. These metrics can be used to comprehensively evaluate the accuracy of segmentation results (DSC and RVE) and assess the impact of incorrect segmentation region and local error (HD and ASSD). The definition of metrics can be found in Chapter 3 and 4. The 8 testing subjects were also evaluated separately using these 4 metrics to observe the bias effect caused by individual cohort to the main performance. Wilcoxon signed rank test [14] was used to statistically test differences for each metrics between the different methods because when compare the difference between the paired values, they were not normally distributed. A significance level $a = 0.05$ was considered.

5.3 Results

5.3.1 Performance comparison under different methods.

A total of 128 labels, representing individual lower-limb muscles (16 muscle labels per subject across 8 subjects from the PMW-2 cohort), were evaluated. These labels were used to compare the performance of various post-processing methods and the semi-automatic segmentation tool against the original prediction results produced by the trained U-Net model.



	Mean ±SD [%]			
Method	**DSC**	**RVE**	**HD (mm)**	**ASSD (mm)**
U-Net	0.825 ±0.072	0.069 ±0.164	97.2 ±67.8	4.5 ±2.1
U-Net+2D	0.825 ±0.078	-0.019 ±0.158	34.2 ±30.7	2.6 ±1.3
U-Net+CRF	0.812 ±0.073	0.043 ±0.156	36.3 ±27.1	3.1 ±1.3
U-Net+MS	0.827 ±0.063	0.109 ±0.162	20.6 ±11.7	2.1 ±0.9
Mimics	0.673 ±0.156	0.045 ±0.414	44.6 ±39.6	6.1 ±9.8

Fig 5.4 Performance comparison of each method. Box plots for the DSC (top left), RVE (top right), HD (bottom left), and ASSD (bottom right) for each method (16 medium values of each muscle calculated from 128 labels) are shown (* indicates $p<0.05$, ** indicates $p<0.01$). Lower scores indicate better performance in HD and ASSD. U-Net denotes original prediction from U-Net; U-Net+2D denotes 2D post-processing; U-Net+CRF denotes fully connected conditional random fields; U-Net+MS denotes mean shape based post-process method; Mimics denotes the results from Mimics semi-automatic muscle segmentation toolbox.

It can be observed that original results from U-Net, and results from U-Net+2D, and U-Net+MS exhibit jointly optimal results in terms of DSC (Fig 4.), averaging around 0.826 (0.825-0.827), representing an improvement of 1.7% ($p<0.01$) and 22.7% ($p<0.01$) over U-Net+CRF (0.812) and Mimics (0.673), respectively.

The U-Net+2D showed the best performance in RVE, with its absolute values being closest to zero, indicating a high degree of alignment with the gold standard in terms of volume prediction. It outperformed the other methods with a percentage difference between the absolute values of the RVE for one method and U-Net+2D from 2.4% to 9.0% ($p<0.05$). U-Net+MS shows the worst RVE performance (0.109 ±0.162).

U-Net+MS exhibited the best performance in terms of HD (mean value: 20.6mm ±11.7mm), better than U-Net (78% improvement, $p<0.01$), Mimics (53.8% improvement, $p<0.01$), U-Net+2D (39.8% improvement, $p<0.01$) and U-Net+CRF (43.3% improvement, $p<0.01$). Similarly, the U-Net+MS was associated with the lowest ASSD (mean value: 2.1mm ±0.9mm), better than U-Net (53.3% improvement, $p<0.01$), Mimics (65.6% improvement, $p<0.01$), U-Net+2D (19.2% improvement, $p<0.01$) and U-Net+CRF (47.6% improvement, $p<0.01$).

The results from the different methods, applied to the 8 individual test subjects, demonstrated consistent variability and similar trends across all subjects (Figure 5.5). When analyzing results for individual subjects, similar trends were observed as those seen when results were averaged across the entire cohort for each method (Figure 5.4). The analysis showed that minimal individual bias existed, and it had negligible impact on the overall results of this experiment.

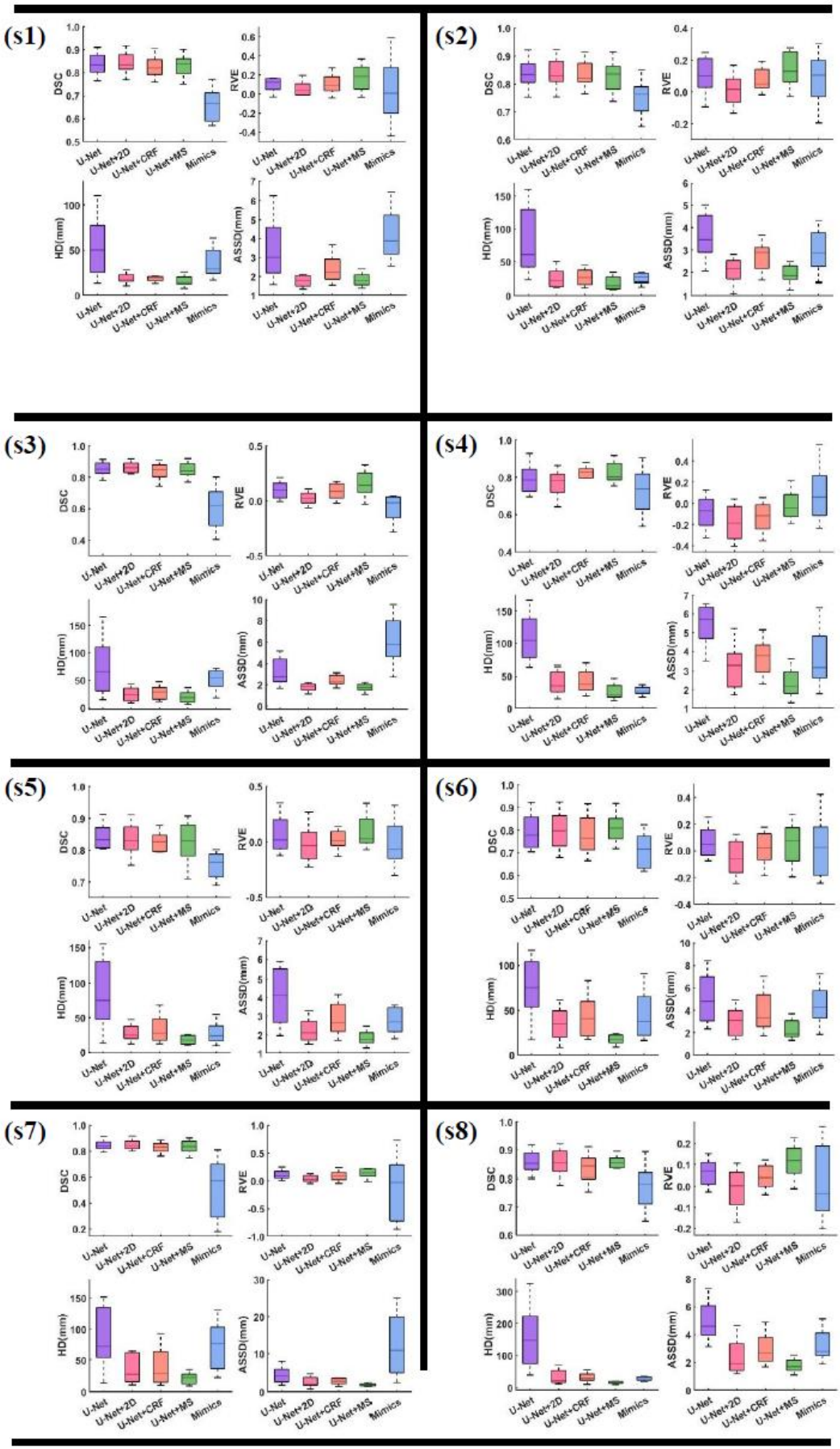
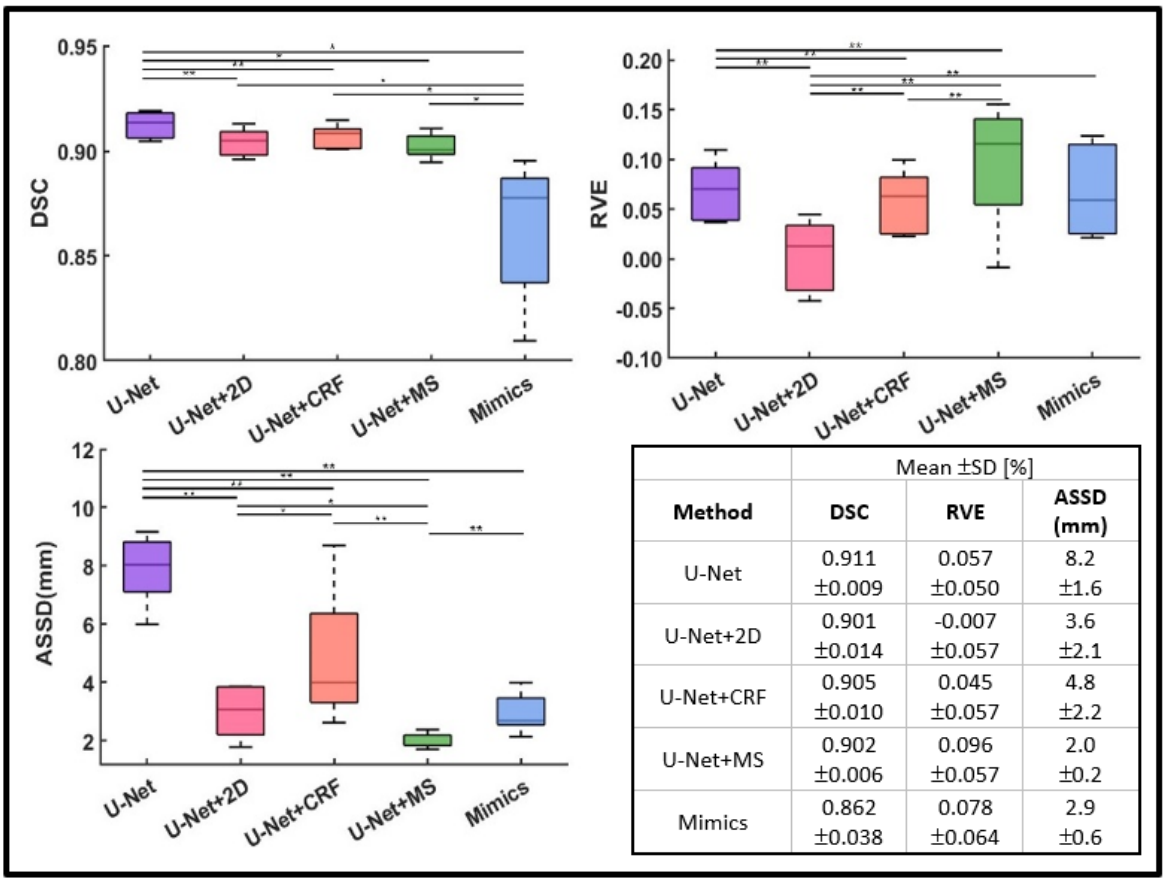

(s1)
(s2)
(s3)
(s4)
(s5)
(s6)
(s7)
(s8)
DSC
RVE
HD(mm)
ASSD(mm)
U-Net
U-Net+2D
U-Net+CRF
U-Net+MS
Mimics

Figure 5.5 Performance comparison of each method on individual testing subjects from PMW-2. Subplots (s1-s8) represent the performance of four metrics (DSC/RVE/HD/ASSD) for each subject (8 in total) under different methods (each boxplot includes 16 individual muscles' score). The median number of each boxplot is connected by dotted red lines to see if the overall trend is the same.



Method	Mean ±SD [%] DSC	RVE	ASSD (mm)
U-Net	0.911 ±0.009	0.057 ±0.050	8.2 ±1.6
U-Net+2D	0.901 ±0.014	-0.007 ±0.057	3.6 ±2.1
U-Net+CRF	0.905 ±0.010	0.045 ±0.057	4.8 ±2.2
U-Net+MS	0.902 ±0.006	0.096 ±0.057	2.0 ±0.2
Mimics	0.862 ±0.038	0.078 ±0.064	2.9 ±0.6

Figure 5.6 Performance comparison of each method after pooling all labels into one. Box plots for the DSC (top left), RVE (top right), and ASSD (bottom left) for each method (8 values from 8 subject under each metric) are shown (* indicates $p<0.05$, ** indicates $p<0.01$).

When the labels of the 16 target muscles were combined into a single label, the DSC, RVE, HD and ASSD were used to assess the recognition performance of each method on the overall target muscle region (Figure 5.6). Since the entire muscle cohort is significantly larger and longer than individual muscles, comparing the Hausdorff Distance (HD) across methods becomes irrelevant and was therefore excluded from the analysis.

The U-Net achieved the highest DSC value of 0.911 ± 0.009, indicating superior segmentation accuracy compared to the ground truth. Mimics had the lowest DSC value at 0.862 ± 0.038, reflecting a significant 5.4% ($p<0.01$) reduction in segmentation accuracy

compared to U-Net, indicating that this difference is statistically significant. The other methods (U-Net+2D, U-Net+CRF, U-Net+MS) show slightly lower DSC values than U-Net, indicating minor deviations in segmentation accuracy with U-Net+2D at 0.901 ± 0.014 ($p<0.01$), U-Net+CRF at 0.905 ± 0.010 ($p<0.01$), and U-Net+MS at 0.902 ± 0.006 ($p<0.05$).

The 2D post-processing method showed the best performance in RVE, with a value closest to zero (-0.007 ± 0.057), indicating the highest accuracy in volume prediction compared to the gold standard. The U-Net+MS method demonstrated the worst performance in RVE (0.096 ± 0.057), indicating a greater deviation from the gold standard in volume prediction compared to the other methods.

The U-Net+MS approach demonstrated the best performance in terms of ASSD, with the lowest mean value (2.0 mm ± 0.2 mm), indicating the highest precision in boundary segmentation compared to the gold standard. In contrast, the U-Net approach had the highest ASSD value (8.2 mm ± 1.6 mm). The U-Net+2D, U-Net+CRF, and Mimics methods demonstrated intermediate performances, with ASSD values of 3.6 mm ± 2.1 mm, 4.8 mm ± 2.2 mm, and 2.9 mm ± 0.6 mm, respectively, indicating varying levels of boundary precision.

5.3.2 Relationship between muscle volume and segmentation accuracy.

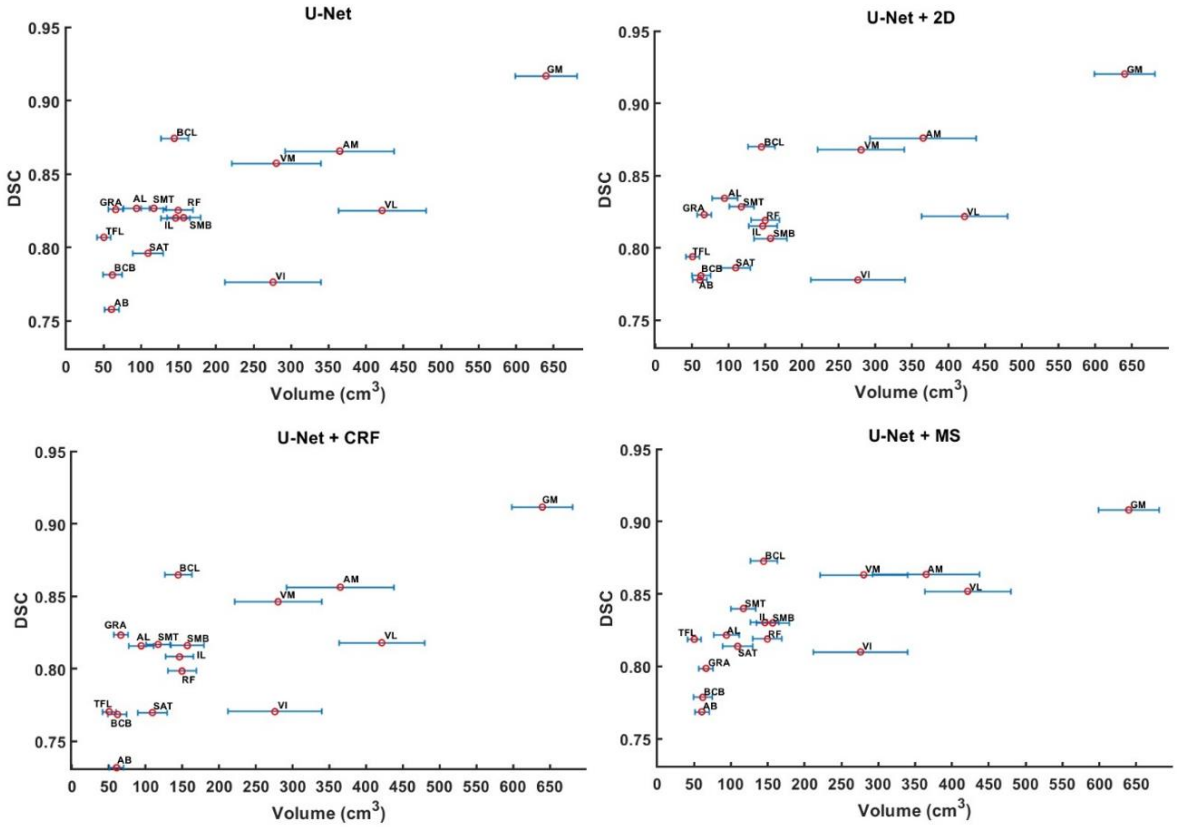


Figure 5.7 Relationship between muscle volume and DSC score under 4 methods (U-Net (top left), U-Net+2D (top right), U-Net+CRF (bottom left) and U-Net+MS (bottom right)) of test subjects**.** The muscles were arranged in ascending order according to the mean muscle volume of PMW-2 cohort. The error bar of each point denotes the standard deviation of muscle volume among 8 test subjects.

As muscle volume increases, the accuracy of segmentation under each method also shows an increase. For muscles with large volume differences between samples, such as VI, or for smaller and irregularly shaped muscles like AB and BCB, the segmentation precision is low.

5.3.3 Visualization.

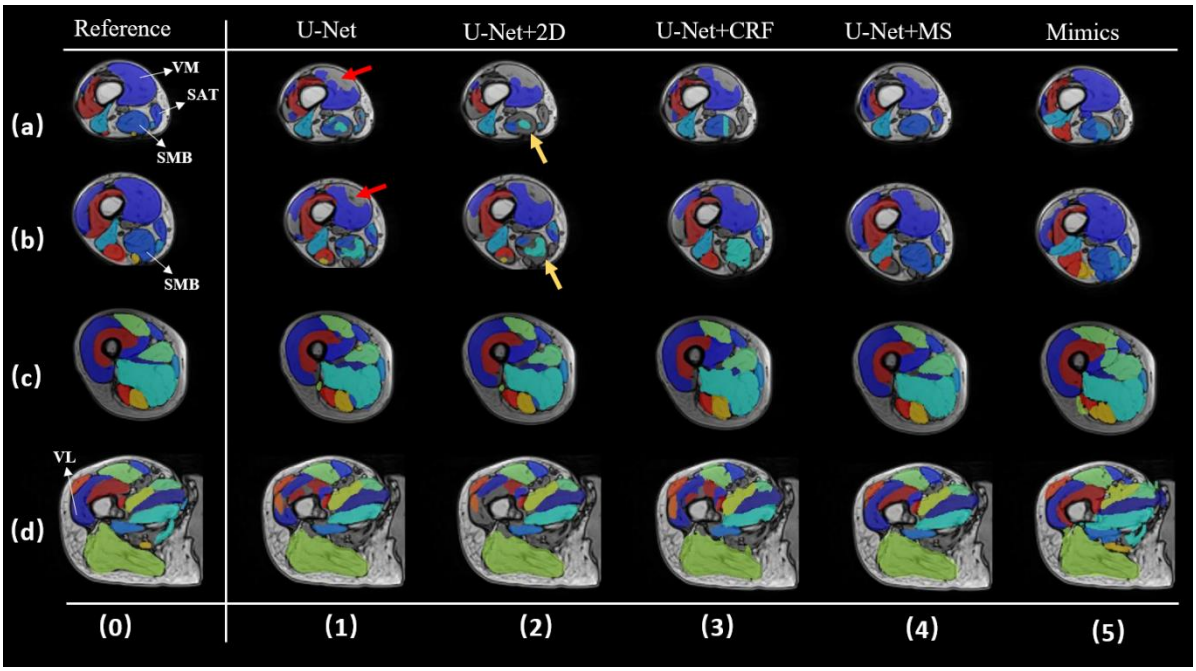


Figure 5.8 Segmentation performance visualization of each method. The figure exhibits the segmentation performance of each method on one subject from a location close to knee to hip, rows a-d. Column 0 denotes the golden standard, and columns 1-5 denote different segmentation results.

Compared to the gold standard (column 0, Figure 5.8), the U-Net segmentation results from each method primarily exhibited two issues: blank regions and mis-segmentation errors. Blank regions refer to areas that should be assigned to a specific muscle class but remain unsegmented, leaving empty spaces (as indicated by the red arrows in a-1, b-1). Mis-segmentation occurs when areas that should be classified as muscle A are incorrectly identified as muscle B or C (as indicated by the yellow arrows in a-2, b-2).

In the original predictions of the U-Net, as illustrated in a-1, there were substantial blank areas in the results related to VM. After 2D post-processing method (a-2) the issue persisted, and, conversely, there were less regions well identified. However, results obtained through U-Net+CRF (a-3), and U-Net+MS (a-4) showed improvement, with some small areas correctly filled.

The original results also exhibit problems of mis-segmentation in muscles like SAT and SMB. In a-1, the SAT prediction from U-Net was entirely mis-segmented, and the SMB predicted area included a substantial region of incorrectly segmented muscles. In U-Net+2D (a-2) and U-Net+CRF (a-3) results, errors in the SAT segment were removed, leaving behind

blank areas. However, in the SMB area, incorrect prediction areas were not fully removed, and correct predictions were mistakenly removed in some cases. MS effectively addressed these issues (a-4), where both SMB and SAT muscle regions were relatively well segmented. The same issue was well-reflected in the SMB area (b-4). After using approach U-Net+2D or U-Net+CRF, the SMB area worsened (b-1/2), whereas the U-Net+MS method reduced errors.

In the VL area (d-1), the U-Net prediction introduced errors by incorrectly splitting the properly segmented VL muscle into two parts (indicated by yellow arrows). As a result, in U-Net+2D, one of the correctly predicted VL regions was erroneously removed. In contrast, the U-Net+MS method (d-4) demonstrated superior performance, resolving these issues more effectively than the U-Net+2D (d-2) and U-Net+CRF (d-3) approaches.

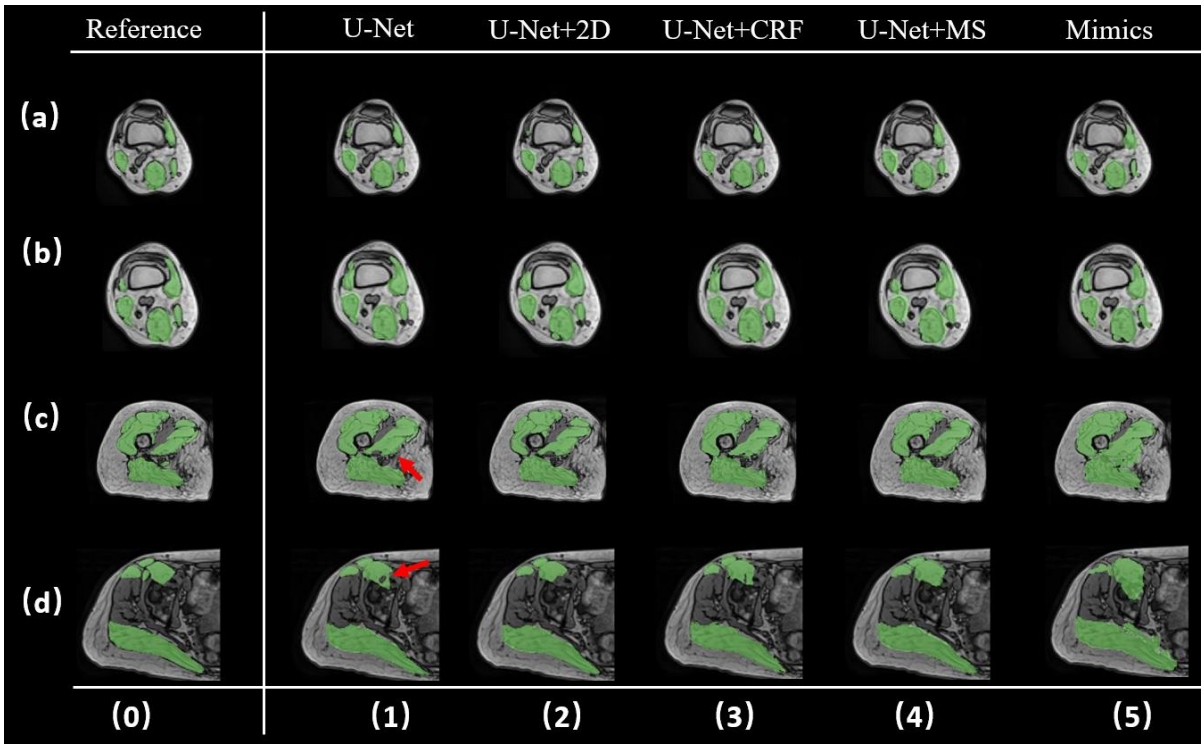


Figure 5.9 Segmentation performance visualization of each method after pooling all labels into one. The figure exhibits the segmentation performance of each method on one testing subject from a location close to knee to hip, rows a-d. Column 0 denotes the golden standard, and columns 1-5 denote different segmentation results.

When the labels of 16 target muscles are treated as one label, the U-Net+MS method can correct the empty spaces more or less in the U-Net results (as indicated by the red arrow in Figure 5.9-c-1 and d-1). Except for Mimics, the other methods do not present a large difference in this visualization case.

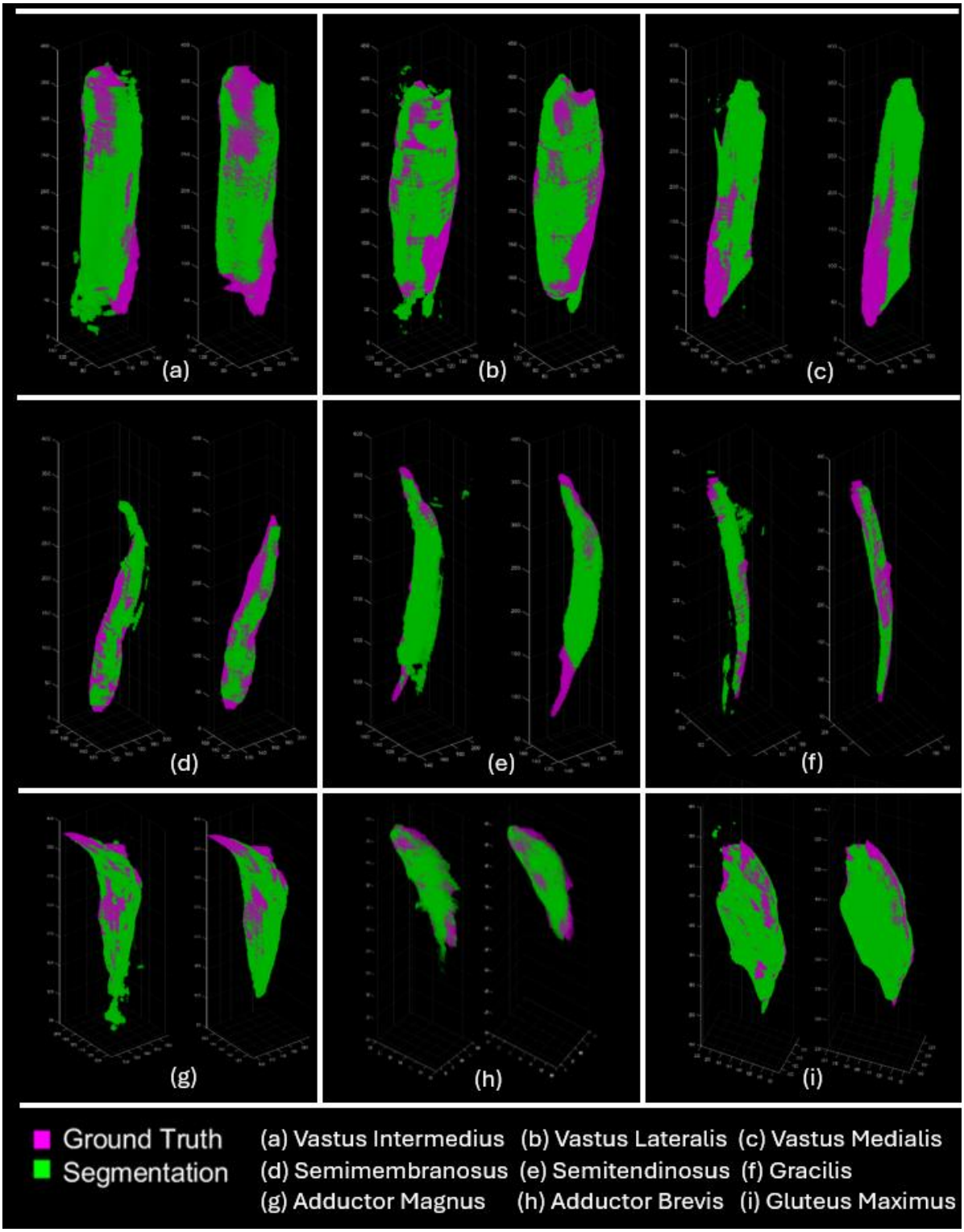


Figure 5.10 Comparison between ground truth and segmentation results with/without post-processing step. GT (in purple) and segmented results (in green) are overlapped and visualized by converting them into point clouds in MATLAB. Parts a-i are the comparison of VI, VL, VM, SMB, SMT, GRA, AM, AB and GM from one testing subject, respectively. In each part, the left side is the comparison of U-Net results with GT, and the right side is the comparison of U-Net+MS results with GT.

Figure 5.10 provides a direct intuition of how post-processing step reduces HD and weakens the side effects from local errors. The segmentation after post-processing step is more similar to GT in geometry, especially at both ends of each muscle, and post-processing step can improve obvious unreasonable areas (e.g. in Figure 5.10-g) leading to lower HD score with better performance.

5.4 Discussion

In this chapter, we aim to propose a novel muscle segmentation post-processing method based on mean shape from a statistical shape model to improve the accuracy of the deep learning segmentation results and compare it with other existing post-process methods and commercial automatic segmentation software. Overall, our method, which uses the mean shape and deformable registration, has been found to be significantly more accurate than the original prediction from the U-Net, the post-processed results from selected 2D post-processing and Conditional Random Fields (CRFs) methods, and the direct segmented results from the Mimics semi-automatic muscle segmentation tool.

U-Net segmentation followed by MS post-processing demonstrated the best performance in terms of HD (20.6mm) and ASSD (2.1mm) and had similar DSC score (0.827) compared to the original (U-Net, no postprocessing) results (0.825) and U-Net segmentation followed by 2D post-processing method (0.825). However, U-Net+MS exhibited the highest RVE at 0.109, indicating a systematic overestimation of muscle volumes, which is further confirmed by the high RVE for pooled labels (Figure 5.6). This indicates that, while using the MS method on U-Net prediction results can provide better corrections for some local mis-segmentation points and large outlying points, this approach is less accurate than others when the correct estimation of the volume of the muscles is required.

Compared to U-Net+MS results, the results from Mimics have lower RVE but on average lower DSC scores. Figure 5.11 shows the Mimics segmentation results (green) of two muscles (Vastus Intermedius and Sartorius) compared with the ground truths (purple), where the RVE scores of both muscles were good, 0.1% and 0.15%, but the DSC scores were lower than 0.70, 0.571 and 0.684. The lower RVE and the lower DSC normally indicates that there are large mis-identified volumes (non-overlap volumes) but that the combination of overestimated volume in one region and underestimated volume in another region lead to similar overall volume, as shown in Figure 5.11. This issue confirms that reporting only the RVE in image segmentation studies is not enough and could lead to misinterpretation of the results.

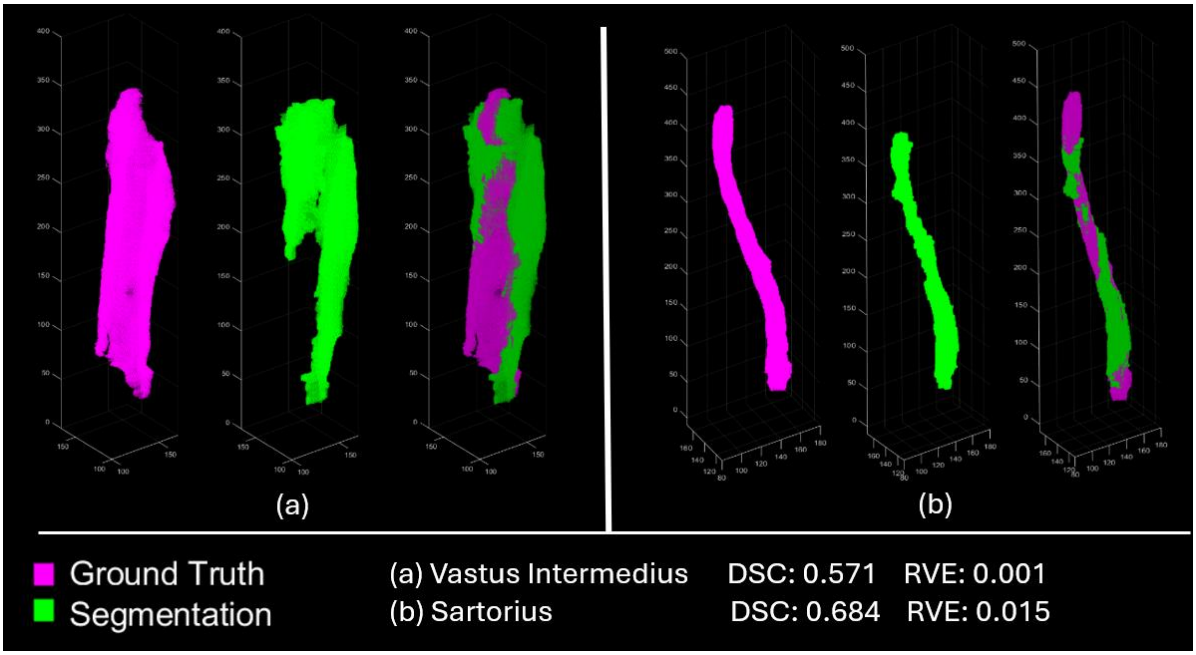


Figure 5.11 Comparison between ground truth and segmentation results from Mimics. In each part, the ground truth (left), segmentation result (middle) and the overlapping section of both (right) are shown. Parts a and b represent the results of muscle vastus intermedius and sartorius of same testing subject.

The proposed MS method retains the mean shape after deformable registration, making it superior in handling local errors compared to other methods in the control group, as evidenced by improvements in HD and ASSD. In 2D post-processing method, a 4% performance improvement was achieved in liver segmentation. However, compared to the liver, the individual muscle segmented in our study is more complex in size, shape, texture, and surface variations, leading the segmentation more challenging. In fact, when all muscles labels are pooled together, better metrics (DSC: 0.91) can be observed (Figure 5.6). It is worth noting that the volume-related accuracy (RVE: -0.019) is highest for the U-Net+2D approach. The comparison between the predicted muscles from the U-Net approach alone or followed by 2D post-processing approach (Figure 5.8 d-1 and d-2) suggests that for certain type of muscles involving more than two independent regions, the 2D post-processing method will result in poorer outcomes. This limitation of the 2D post-processing method is due because it only retains the larger region of the muscle. The CRF algorithm itself relies more on the predicted label map relationships between target pixels and surrounding pixels [74], [80], [120], sometimes overlooking anatomical significance and relying solely on mathematical inference. This could lead to obvious errors, as seen in the SMB area (Figure 5.8 A-3). In contrast to the literature [74], which used CRF based post-processing method for

the segmentation of natural objects such as birds and trees, where there is more contrast between the object to segment and the background, the differences between adjacent muscles, or between muscles and fat tissues, are sometimes not as clear in the MRIs and will affect the performance of CRF inference. In fact, CRF only achieved a DSC score of 0.812, which is worse than the untreated results of the DL model. As muscle volume increases (Fig 5.7), the DSC under each methods also increases, consistent with the results reported in [14]. For example, the DSC for the Gluteus Maximus (GM) reaches values above 90%. Nevertheless, for smaller muscles, with volume below 200 cm^3, a wide range of DSC scare was found, from 73% to 88%, highlighting the heterogeneous performance of the model. It is therefore very important to choose the right model for different applications. Based on the DSC scores obtained in this experiment, a recommendation table (Table 5.4) is given.

Table 5.4 Recommended methods based on different target muscles.

Muscle	**Chosen method / DSC**		**Chosen method / HD(mm)**		**Chosen method / ASSD(mm)**	
Rectus Femoris	U-Net	/ 0.825	U-Net + MS	/ 19.9	U-Net + MS	/ 2.3
Vastus Intermedius	U-Net + MS	/ 0.810	U-Net + MS	/ 22.9	U-Net + MS	/ 2.5
Vastus Lateralis	U-Net + MS	/ 0.852	U-Net + MS	/ 28.6	U-Net + MS	/ 2.2
Vastus Medialis	U-Net + 2D	/ 0.868	U-Net + MS	/ 18.6	U-Net + MS	/ 2.1
Sartorius	U-Net + MS	/ 0.814	U-Net + MS	/ 23.0	U-Net + MS	/ 1.7
Semimembranosus	U-Net + MS	/ 0.830	U-Net + MS	/ 25.7	U-Net + MS	/ 2.7
Semitendinosus	U-Net + MS	/ 0.840	U-Net + MS	/ 24.4	U-Net + MS	/ 2.0
Gracilis	U-Net	/ 0.826	U-Net + MS	/ 15.2	U-Net + MS	/ 1.5
Biceps Femoris Caput Brevis	U-Net	/ 0.781	U-Net + MS	/ 28.2	U-Net + MS	/ 2.5
Biceps Femoris Caput Longum	U-Net	/ 0.874	U-Net + MS	/ 25.7	U-Net + MS	/ 1.8
Adductor Magnus	U-Net + 2D	/ 0.876	U-Net + MS	/ 25.9	U-Net + MS	/ 2.4
Adductor Brevis	U-Net + 2D	/ 0.778	U-Net + MS	/ 14.1	U-Net + MS	/ 2.2
Adductor Longus	U-Net + 2D	/ 0.834	U-Net + MS	/ 15.3	U-Net + MS	/ 2.0
Gluteus Maximus	U-Net + 2D	/ 0.920	U-Net + MS	/ 15.8	U-Net + MS	/ 1.9
Iliacus	U-Net + MS	/ 0.830	U-Net + MS	/ 13.0	U-Net + MS	/ 1.8
Tensor Fasciae Latae	U-Net + MS	/ 0.819	U-Net + MS	/ 13.4	U-Net + MS	/ 1.7

In Table 5.4, the best method which shows the best performance (DSC/HD/ASSD) under individual muscle among 4 methods (U-Net, U-Net+2D, U-Net+CRF and U-Net+MS) is recommended as the chosen method with the corresponding value reported on the right.

Different post-processing approaches were found to be best suited to segment the 16 lower limb muscles considered in this chapter, depending on the muscle and the considered metric (Table 5.4). Therefore, depending on the different applications, the user can choose which post-processing method to use in order to obtain the best results. It can be shown that the post-processing method has improved the results of most muscle segmentation where U-Net+MS was more widely recommended, especially for the local analyses related to HD and ASSD.

Compared to the prediction results of U-Net+MS and U-Net, the results obtained from the automated part of the semi-automatic muscle segmentation toolbox provided by Mimics are

lower than average level. It only approaches the gold standard more closely in RVE (-0.019) and outperforms U-Net in HD (44.6mm). During the segmentation process, the software combines manually selected muscle regions using thresholding (as shown in Figure 5.3 (a) and (c)) with pre-prepared atlas files from PMW-1 cohort for segmentation. This operation could be influenced by the differences or bias between atlas files and the new input samples (i.e. increasing the number of files in the atlas is likely to improve the outcomes). The atlas files in this study were derived from PMW-1 cohort, while input samples were from PMW-2, with general consistency in age, volume ($p > 0.05$, Table 5.5), etc. Therefore, the operation of Mimics in this study could be affected by individual differences between the two cohorts more than for the other methods. However, in the process of operation, the results automatically generated by Mimics are generally expected to be corrected manually by the operator, leading to better yet less efficient results.

Table 5.5 Muscle average volume comparison between PMW-1 and PMW-2.

Muscle	**PMW-1 [cm3]**	**PMW-2 [cm3]**
Rectus femoris	127	151
Vastus intermedius	294	277
Vastus lateralis	376	422
Vastus medialis	247	281
Sartorius	108	110
Semimembranosus	136	158
Semitendinosus	119	117
Gracilis	60	67
Biceps femoris caput brevis	68	63
Biceps femoris caput longum	127	145
Adductor magnus	390	366
Adductor brevis	70	62
Adductor longus	90	95
Gluteus maximus	643	641
Iliacus	136	147
Tensor fasciae latae	48	51
		$\boldsymbol{p > 0.05}$

Muscle segmentation is a critical prerequisite for subsequent analysis and diagnosis, playing a key role in both texture analysis (TA) [121] and the quantitative assessment of related parameters [6]. In TA, the image features included in the segmentation results can significantly affect conclusions. Therefore, the segmentation process must prioritize removing local errors and minimizing noise-induced influences. Methods demonstrating superior performance in terms of HD and ASSD are better suited for TA, as they ensure greater boundary precision and error minimization. In quantitative analyses of muscles, such as mass and maximal isometric force [6], paying extra attention to accurate volume measurement is crucial throughout the segmentation process. Hence, the RVE metric assumes

greater importance in the evaluation of segmentation results. Therefore, considering the four considered evaluation metrics, MS post-processing is more suitable for TA, while the 2D post-processing method is more applicable for quantitative analysis.

The study in this chapter is also affected by some limitations. Firstly, it is limited by the number of the datasets used (10 subjects for training and 8 subjects for testing). The number of unique images from each individual collected during MRI acquisitions was from 435 to 550 slices, leading to a total number of training slices of approximately 5000. However, considering that only 10 subjects were used for training, the heterogeneity of features in the muscles to be segmented was probably limited. As a result, when training DL models, the features learned by convolutional operation and the shape features on the average shape geometry in the MS method generated based on the training set is probably incomplete. Although features including all samples in a specific range are not displayed, with the increase of data and the richness of subject types, the loss caused by incomplete feature learning can be reduced to a certain extent. In future work, more subjects from different cohorts will be collected to increase the range of muscle volumes and have larger distributions of muscle shapes like the data augmentation modification mentioned in chapter 3. Larger datasets are likely also to increase the DL model performance as previously found in Rohm et al [122], the DSC score increase around 2% after adding augmentation data in chapter 3. In addition, the performance of the post-processing methods still depends on the quality of the results produced by previous methods such as DL model and automatic segmentation tools. The performance of the model will affect the quality of post-processing results, the better model which proved achieved better segmentation than U-Net, e.g. Attention-Feature-Fusion-UNet (AFFU, chapter 3) achieved 2.0% gain in DSC compared to U-Net, and Spatial Channel UNet (SC-UNet) which considered the knowledge of where the image slice was acquired from within the lower limbs (along the axial direction)[7], can be used and linked with MS method in further study to observe is there any significant improvement.

5 Conclusion

In this chapter, we showed that using mean shape based post-processing method after an automatic segmentation with U-Net improves most metrics for segmenting individual muscles of the lower limb from MR images compared to the U-Net alone. This approach

reduced the local segmentation errors compared to segmentations with U-Net alone or followed by post-processing methods based on 2D analyses or Conditional Random Field.

The method we proposed also holds potential as a reference for other medical segmentation tasks which involve entire geometric organ shape like the liver and heart. By refining segmentation accuracy, this approach can be adapted to improve the results of various complex anatomical structures, making it broadly applicable to different areas of medical image analysis.

Chapter 6 General discussion and conclusions

6.1 Conclusions

The research conducted in this PhD thesis achieves the aims and objectives outlined in Chapters 1 and 2. Chapters 3-5 focus on three main research directions, addressing the difficulties in automatic medical image segmentation through pre-processing, DL model enhancement, data augmentation, post-processing, and model generalization evaluation, while proposing corresponding applications.

Chapter 3 focuses on DL model structure enhancement with feature fusion module and attention mechanism and data augmentation based on statistical shape modelling. The proposed model, AFFU, was found achieving better segmentation accuracy than benchmark models U-Net and UNet++ with 2.1% gain in DSC, 3.0% in RVE and 35% in HD when trained and tested on same cohort (PMW-1). When trained on PMW-1 and tested on PMW-2/-OB, the difference on DSC performance for all chosen models was not obvious, however, in local error which represented from HD, AFFU showed better results than U-Net with gain around 40% in PMW-2/-OB. This indicated that for small size dataset (n~10), the feature fusion mechanism could help model extract more representative features from training set and leading to more accurate segmentation detail in areas such as adjacent boundaries. When the novel data augmentation method proposed in this thesis was applied, simple models (U-Net) tend to benefit more compared to the complex model (AFFU). This could be explained by the fact that complex models already capture enough features from limited samples or small dataset, so, augmented data generated from the same cohort doesn't significantly improve their segmentation accuracy. In contrast, augmented data helps simple models learn a wider range of sample features more effectively. When computational power and budget are limited, optimizing a complex model or using data augmentation with a simpler model can serve as alternative strategies to consider.

Building on the results from Chapter 3, Chapter 4 focuses on exploring the cohort-related generalization performance of AFFU and other benchmark models across different cohorts (children, young people and older women). The intra- and inter- subject segmentation was first completed by three operators on children cohort and 18 muscles were found to be with high reproducibility (PE lower than 10%) and could be used to evaluate DL models prediction. The most muscles were also achieved high intra-operator’s reproducibility in PMW cohort [6].

Models trained on specific cohort groups, such as PMW (task 1) and TDC (task 2), exhibited limited generalization when applied to other cohorts. The significant decline in segmentation accuracy across age groups underscores the influence of anatomical differences, such as muscle size and fat distribution, and emphasizes the need for training datasets that are representative of the target population. Training with a balanced, bilateral dataset significantly improves model performance, reducing bias between left and right muscles. This demonstrates the importance of dataset diversity in overcoming challenges like muscle asymmetry and anatomical variability. Among the models, AFFU showed consistent accuracy and robustness handling of muscle volume differences, making it a reliable choice for segmentation tasks, particularly when using diverse datasets. Moreover, AFFU consistently outperformed other models in segmentation accuracy and robustness across combined age groups in training set, particularly good in volume error reduction and boundary delineation. However, its higher computational demands suggest that while it is effective for smaller datasets, more efficient post-processing and model optimizations are necessary for larger datasets.

For the final attached operation in automatic segmentation, Chapter 5 introduces a novel muscle segmentation post-processing method based on mean shapes (MS) from SSM, demonstrating significant improvements over existing methods such as 2D post-processing, CRF, and commercial tools like Mimics. The MS method excels in addressing local segmentation errors, achieving superior HD and ASSD metrics while maintaining comparable DSC scores, though it shows a tendency for volume overestimation (RVE). In comparison, MS is optimal for texture analysis (TA), prioritizing boundary precision, while 2D post-processing is better suited for volume-dependent quantitative analyses. This work highlights the need for application-specific post-processing approaches and sets the foundation for refining muscle segmentation in clinical and research contexts.

6.2 Limitations

While this thesis presents promising advancements in the automatic segmentation of lower-limb skeletal muscles, the evaluation of model's generalization and corresponding post-processing techniques, there are still several limitations must be acknowledged.

First, the availability and diversity of training dataset and different cohorts emerge as a constraint across all studies [123]. In Chapter 3, despite the proposed data augmentation technique helps increasing data variability, the limited number of real subject datasets significantly influenced feature extraction ability of models during model training. Although no overfitting was observed, the inclusion of a larger number of independent real subjects, along with cross-cohort generalization analyses, is essential to enhance robustness [124]. Similarly, in Chapter 5, the limited dataset size (10 subjects for training and 8 for testing) constrained the heterogeneity of features and, consequently, the model's ability to learn comprehensive shape and muscle characteristics.

Second, methodological limitations highlight areas for refinement. In Chapter 3, limited by computational cost, the proposed data augmentation method relied on only one main feature (mode 1) derived from PCA analysis, which restricted the variability of the generated artificial data. With the increase of data size (n>20 or more), the first main mode in PCA could not represent the most features among input subjects. Incorporating additional features and extending variability in future studies could improve segmentation accuracy for broader applications. Additionally, the segmentation errors observed in 3D visualizations underline the potential of geometric post-processing approaches to improve these issues. Chapter 5 further emphasizes that the performance of the proposed mean shape-based post-processing technique is inherently dependent on the quality of preceding segmentation results. Employing more advanced segmentation architectures, such as AFFU or SC-UNet, could potentially amplify the accuracy of post-processing methods.

Third, the computational efficiency of certain methods presents practical limitations. As identified in Chapter 4, AFFU demonstrated significantly longer training times compared to other models. While its performance metrics prove its use in critical applications, future research should aim to balance accuracy and computational efficiency, potentially by integrating external prior information or optimizing model architectures biased to lightweight models [16].

Finally, although Chapter 4 qualitatively analysed the cohort-related generalization performance of DL models and addressed a gap in this field, it lacked the quantitative analysis that is more critical in practical applications. If a quantitative study could determine that adding at least m subjects (e.g. children) to n subjects (e.g. older women) from the same cohort optimizes the model's prediction on the testing set, this finding could maximize process efficiency in real applications. It would provide clear guidance for balancing data collection costs with model performance.

6.3 Impact

The study conducted throughout this PhD has contributed to the publication of three papers (two as first author [14], [106] and one as a co-author [7]). Moreover, corresponding code have been written and shared on GitHub: 1) the traditional U-Net, UNet++, AFFU coded in Pytorch, 2) the mean shape based post processing, CRF [120] and 2D-processing [73] method written in MATLAB and Pytorch. One further paper is currently finished and has been submitted: “A Generalisation Study in Deep Learning-based Segmentation of Lower-limb Muscles Across Different Populations” (first author).

The data and code generated during this thesis writing have been made publicly accessible to encourage future researchers to prioritize using and adapting existing tools for data analysis. The thesis highlights the advancement and limitations of automatic segmentation tools based on DL and pre-/post-processing method which also play vital roles during entire segmentation operation, hoping that our exploration will support future researchers in applying these tools to address more clinical image segmented research questions.

The thesis aimed at improving the process of human lower-limb muscle’s automatic segmentation, not only pursuing more advanced DL models. Instead, it combined DL with various methods, incorporating factors such as SSM to enhance the interpretability of automatic segmentation. The potential of this study is, the first being in providing operators with an entire automatic segmentation pipeline with pre-processing (data augmentation), advanced DL model training and post-processing strategy which could be promoted to other medical organs’ segmentation or utilized under scenario with insufficient data. The second is for the first time it enables the qualitative cohort related generalization evaluation among different DL models.

6.4 Future work

Future work should be focused on two areas:

1) To address limitations in feature variability introduced by the current proposed data augmentation method, further experiments should focus on incorporating multiple modes (2^{nd}, 3^{rd} modes etc.) from PCA that can capture a wider range of anatomical variability. As dataset sizes increase, these augmented data should be evaluated carefully whether could represent real subjects.

2) The quantitative analysis of the model generalization is still necessary. Future studies should focus on assembling larger diverse datasets that encompass multiple cohorts, including different age groups, genders, or ethnic backgrounds. Cross-cohort generalization analysis should be conducted quantitatively to ensure model's robustness across different populations and find the middle point considering the computation cost and segmentation accuracy. A systematic study of the trade-offs between dataset size, cohort composition, and prediction accuracy could help establish guidelines for scalable model development.

Reference

[1] M. G. Pandy and T. P. Andriacchi, “Muscle and Joint Function in Human Locomotion,” *Annu. Rev. Biomed. Eng.*, vol. 12, no. 1, pp. 401–433, Jul. 2010, doi: 10.1146/annurev-bioeng-070909-105259.

[2] J. Zhu *et al.*, “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*, vol. 34, no. 12, p. e4609, Dec. 2021, doi: 10.1002/nbm.4609.

[3] A. J. Cruz-Jentoft *et al.*, “Sarcopenia: European consensus on definition and diagnosis,” *Age and Ageing*, vol. 39, no. 4, pp. 412–423, Jul. 2010, doi: 10.1093/ageing/afq034.

[4] S. Nishioka, A. Yamanouchi, T. Matsushita, E. Nishioka, N. Mori, and S. Taguchi, “Validity of calf circumference for estimating skeletal muscle mass for Asian patients after stroke,” *Nutrition*, vol. 82, p. 111028, Feb. 2021, doi: 10.1016/j.nut.2020.111028.

[5] Krzysztofik, Wilk, Wojdała, and Gołaś, “Maximizing Muscle Hypertrophy: A Systematic Review of Advanced Resistance Training Techniques and Methods,” *IJERPH*, vol. 16, no. 24, p. 4897, Dec. 2019, doi: 10.3390/ijerph16244897.

[6] E. Montefiori, B. M. Kalkman, W. H. Henson, M. A. Paggiosi, E. V. McCloskey, and C. Mazzà, “MRI-based anatomical characterisation of lower-limb muscles in older women,” *PLoS ONE*, vol. 15, no. 12, p. e0242973, Dec. 2020, doi: 10.1371/journal.pone.0242973.

[7] W. H. Henson, X. Li, Z. Lin, L. Guo, C. Mazzá, and E. Dall’Ara, “Automatic segmentation of lower limb muscles from MR images of post-menopausal women based on deep learning and data augmentation,” *PLoS ONE*, vol. 19, no. 4, p. e0299099, Apr. 2024, doi: 10.1371/journal.pone.0299099.

[8] M. H. Hesamian, W. Jia, X. He, and P. Kennedy, “Deep Learning Techniques for Medical Image Segmentation: Achievements and Challenges,” *J Digit Imaging*, vol. 32, no. 4, pp. 582–596, Aug. 2019, doi: 10.1007/s10278-019-00227-x.

[9] W. H. Henson, C. Mazzá, and E. Dall’Ara, “Deformable image registration based on single or multi-atlas methods for automatic muscle segmentation and the generation of augmented imaging datasets,” *PLoS ONE*, vol. 18, no. 3, p. e0273446, Mar. 2023, doi: 10.1371/journal.pone.0273446.

[10] J. Cates, P. T. Fletcher, M. Styner, M. Shenton, and R. Whitaker, “Shape Modeling and Analysis with Entropy-Based Particle Systems,” *Information Processing in Medical Imaging*, vol. 4584, pp. 333–345, 2007, doi: 10.1007/978-3-540-73273-0_28.

[11] F. Ambellan, A. Tack, M. Ehlke, and S. Zachow, “Automated segmentation of knee bone and cartilage combining statistical shape knowledge and convolutional neural networks: Data from the Osteoarthritis Initiative,” *Medical Image Analysis*, vol. 52, pp. 109–118, Feb. 2019, doi: 10.1016/j.media.2018.11.009.

[12] O. Oktay *et al.*, “Attention U-Net: Learning Where to Look for the Pancreas,” May 2018, Accessed: Dec. 05, 2023. [Online]. Available: http://arxiv.org/abs/1804.03999

[13] Y. Zhang, C. Guo, and P. Zhao, “Medical Image Fusion Based on Low-Level Features,” *Computational and Mathematical Methods in Medicine*, vol. 2021, pp. 1–13, Jun. 2021, doi: 10.1155/2021/8798003.

[14] Z. Lin, W. H. Henson, L. Dowling, J. Walsh, E. Dall’Ara, and L. Guo, “Automatic segmentation of skeletal muscles from MR images using modified U-Net and a novel data augmentation approach,” *Front. Bioeng. Biotechnol.*, vol. 12, p. 1355735, Feb. 2024, doi: 10.3389/fbioe.2024.1355735.

[15] L. Zhang *et al.*, “Generalizing Deep Learning for Medical Image Segmentation to Unseen Domains via Deep Stacked Transformation,” *IEEE Trans. Med. Imaging*, vol. 39, no. 7, pp. 2531–2540, Jul. 2020, doi: 10.1109/TMI.2020.2973595.
[16] P. Sharma *et al.*, “A lightweight deep learning model for automatic segmentation and analysis of ophthalmic images,” *Sci Rep*, vol. 12, no. 1, p. 8508, May 2022, doi: 10.1038/s41598-022-12486-w.
[17] R. Ni, C. H. Meyer, S. S. Blemker, J. M. Hart, and X. Feng, “Automatic segmentation of all lower limb muscles from high-resolution magnetic resonance imaging using a cascaded three-dimensional deep convolutional neural network,” *J. Med. Imag.*, vol. 6, no. 04, p. 1, Dec. 2019, doi: 10.1117/1.JMI.6.4.044009.
[18] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs,” presented at the Computer Vision and Pattern Recognition, Jun. 2016. Accessed: Dec. 17, 2023. [Online]. Available: http://arxiv.org/abs/1412.7062
[19] D. J. Wilkinson, M. Piasecki, and P. J. Atherton, “The age-related loss of skeletal muscle mass and function: Measurement and physiology of muscle fibre atrophy and muscle fibre loss in humans,” *Ageing Research Reviews*, vol. 47, pp. 123–132, Nov. 2018, doi: 10.1016/j.arr.2018.07.005.
[20] J. A. Bennie, K. De Cocker, and S. Tittlbach, “The epidemiology of muscle-strengthening and aerobic physical activity guideline adherence among 24,016 German adults,” *Scandinavian Med Sci Sports*, vol. 31, no. 5, pp. 1096–1104, May 2021, doi: 10.1111/sms.13922.
[21] J. A. Bennie, K. De Cocker, J. J. Smith, and G. H. Wiesner, “The epidemiology of muscle-strengthening exercise in Europe: A 28-country comparison including 280,605 adults,” *PLoS ONE*, vol. 15, no. 11, p. e0242220, Nov. 2020, doi: 10.1371/journal.pone.0242220.
[22] H. A. Parry, M. D. Roberts, and A. N. Kavazis, “Human Skeletal Muscle Mitochondrial Adaptations Following Resistance Exercise Training,” *Int J Sports Med*, vol. 41, no. 06, pp. 349–359, Jun. 2020, doi: 10.1055/a-1121-7851.
[23] J. Verdú-Díaz *et al.*, “Accuracy of a machine learning muscle MRI-based tool for the diagnosis of muscular dystrophies,” *Neurology*, vol. 94, no. 10, pp. e1094–e1102, Mar. 2020, doi: 10.1212/WNL.0000000000009068.
[24] S. Holm-Yildiz *et al.*, “Muscle biopsy and MRI findings in ANO5 -related myopathy,” *Muscle and Nerve*, vol. 64, no. 6, pp. 743–748, Dec. 2021, doi: 10.1002/mus.27419.
[25] L. Riehakainen, C. Cavallini, P. Armanetti, D. Panetta, D. Caramella, and L. Menichetti, “In Vivo Imaging of Biodegradable Implants and Related Tissue Biomarkers,” 2021.
[26] JAMA, “Wilhelm Konrad Roentgen—The Centennial of His Birth—Semicentennial of the X-Rays,” *JAMA Network*, vol. 323, no. 15, p. 1512, 2020, doi: 10.1001/jama.2019.13400.
[27] “Paul Lauterbur and Sir Peter Mansfield for MRI,” *European Heart Journal*, vol. 40, no. 24, pp. 1898–1899, Jun. 2019, doi: 10.1093/eurheartj/ehz397.
[28] ITN, “Global Medical Imaging Trends,” *itn imaging technology news*, 2022. [Online]. Available: https://www.itnonline.com/article/global-medical-imaging-trends
[29] G. Litjens *et al.*, “A survey on deep learning in medical image analysis,” *Medical Image Analysis*, vol. 42, pp. 60–88, Dec. 2017, doi: 10.1016/j.media.2017.07.005.
[30] M. Gross *et al.*, “Automated MRI liver segmentation for anatomical segmentation, liver volumetry, and the extraction of radiomics,” *Eur Radiol*, vol. 34, no. 8, pp. 5056–5065, Jan. 2024, doi: 10.1007/s00330-023-10495-5.

[31] S. Tayebi Arasteh *et al.*, “Automated segmentation of 3D cine cardiovascular magnetic resonance imaging,” *Front. Cardiovasc. Med.*, vol. 10, p. 1167500, Oct. 2023, doi: 10.3389/fcvm.2023.1167500.
[32] Nikolaj Buhl, “Guide to Image Segmentation in Computer Vision: Best Practices.” [Online]. Available: https://encord.com/blog/image-segmentation-for-computer-vision-best-practice-guide/
[33] S. Abdulateef and M. Salman, “A Comprehensive Review of Image Segmentation Techniques,” *IJEEE*, vol. 17, no. 2, pp. 166–175, Dec. 2021, doi: 10.37917/ijeee.17.2.18.
[34] N. Otsu, “A Threshold Selection Method from Gray-Level Histograms,” *IEEE Trans. Syst., Man, Cybern.*, vol. 9, no. 1, pp. 62–66, Jan. 1979, doi: 10.1109/TSMC.1979.4310076.
[35] R. Helen, N. Kamaraj, K. Selvi, and V. Raja Raman, “Segmentation of pulmonary parenchyma in CT lung images based on 2D Otsu optimized by PSO,” in *2011 International Conference on Emerging Trends in Electrical and Computer Technology*, Nagercoil, India: IEEE, Mar. 2011, pp. 536–541. doi: 10.1109/ICETECT.2011.5760176.
[36] M. Roushdy, “Comparative Study of Edge Detection Algorithms Applying on the Grayscale Noisy Image Using Morphological Filter,” vol. 6, no. 4, 2006.
[37] A. Norouzi *et al.*, “Medical Image Segmentation Methods, Algorithms, and Applications,” *IETE Technical Review*, vol. 31, no. 3, pp. 199–213, May 2014, doi: 10.1080/02564602.2014.906861.
[38] R. Adams and L. Bischof, “Seeded region growing,” *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 16, no. 6, pp. 641–647, Jun. 1994, doi: 10.1109/34.295913.
[39] M. Kass, A. Witkin, and D. Terzopoulos, “Snakes: Active contour models,” *Int J Comput Vision*, vol. 1, no. 4, pp. 321–331, Jan. 1988, doi: 10.1007/BF00133570.
[40] Chenyang Xu and J. L. Prince, “Snakes, shapes, and gradient vector flow,” *IEEE Trans. on Image Process.*, vol. 7, no. 3, pp. 359–369, Mar. 1998, doi: 10.1109/83.661186.
[41] K. Doi, “Computer-aided diagnosis in medical imaging: Historical review, current status and future potential,” *Computerized Medical Imaging and Graphics*, vol. 31, no. 4–5, pp. 198–211, Jun. 2007, doi: 10.1016/j.compmedimag.2007.02.002.
[42] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” Apr. 2015, Accessed: Dec. 05, 2023. [Online]. Available: http://arxiv.org/abs/1409.1556
[43] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017, doi: 10.1145/3065386.
[44] S. G. Armato *et al.*, “The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): A Completed Reference Database of Lung Nodules on CT Scans,” *Medical Physics*, vol. 38, no. 2, pp. 915–931, Feb. 2011, doi: 10.1118/1.3528204.
[45] N. Codella *et al.*, “Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC),” Mar. 29, 2019, *arXiv*: arXiv:1902.03368. Accessed: Nov. 09, 2024. [Online]. Available: http://arxiv.org/abs/1902.03368
[46] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, “ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, pp. 3462–3471. doi: 10.1109/CVPR.2017.369.

[47] J. Long, E. Shelhamer, and T. Darrell, “Fully Convolutional Networks for Semantic Segmentation,” Mar. 08, 2015, *arXiv*: arXiv:1411.4038. Accessed: Nov. 09, 2024. [Online]. Available: http://arxiv.org/abs/1411.4038
[48] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation,” May 18, 2015, *arXiv*: arXiv:1505.04597. Accessed: Nov. 09, 2024. [Online]. Available: http://arxiv.org/abs/1505.04597
[49] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018, doi: 10.1109/TPAMI.2017.2699184.
[50] I. J. Goodfellow *et al.*, “Generative Adversarial Networks,” Jun. 10, 2014, *arXiv*: arXiv:1406.2661. Accessed: Nov. 09, 2024. [Online]. Available: http://arxiv.org/abs/1406.2661
[51] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, “Image Segmentation Using Deep Learning: A Survey,” Nov. 15, 2020, *arXiv*: arXiv:2001.05566. Accessed: Nov. 09, 2024. [Online]. Available: http://arxiv.org/abs/2001.05566
[52] W. Liu, A. Rabinovich, and A. C. Berg, “ParseNet: Looking Wider to See Better,” Nov. 19, 2015, *arXiv*: arXiv:1506.04579. Accessed: Nov. 09, 2024. [Online]. Available: http://arxiv.org/abs/1506.04579
[53] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, “Pyramid Scene Parsing Network,” Apr. 27, 2017, *arXiv*: arXiv:1612.01105. Accessed: Nov. 09, 2024. [Online]. Available: http://arxiv.org/abs/1612.01105
[54] G. E. Hinton and R. R. Salakhutdinov, “Reducing the Dimensionality of Data with Neural Networks,” *Science*, vol. 313, no. 5786, pp. 504–7, Jul. 2006, doi: 10.1126/science.1127647.
[55] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, “UNet++: A Nested U-Net Architecture for Medical Image Segmentation,” Jul. 2018, Accessed: Dec. 05, 2023. [Online]. Available: http://arxiv.org/abs/1807.10165
[56] H. Huang *et al.*, “UNet 3+: A Full-Scale Connected UNet for Medical Image Segmentation,” in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain: IEEE, May 2020, pp. 1055–1059. doi: 10.1109/ICASSP40776.2020.9053405.
[57] V. Badrinarayanan, A. Kendall, and R. Cipolla, “SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017, doi: 10.1109/TPAMI.2016.2644615.
[58] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs,” May 12, 2017, *arXiv*: arXiv:1606.00915. Accessed: Nov. 12, 2024. [Online]. Available: http://arxiv.org/abs/1606.00915
[59] A. Vaswani *et al.*, “Attention Is All You Need,” Aug. 02, 2023, *arXiv*: arXiv:1706.03762. Accessed: Nov. 19, 2024. [Online]. Available: http://arxiv.org/abs/1706.03762
[60] D. Bahdanau, K. Cho, and Y. Bengio, “Neural Machine Translation by Jointly Learning to Align and Translate,” May 19, 2016, *arXiv*: arXiv:1409.0473. Accessed: Nov. 19, 2024. [Online]. Available: http://arxiv.org/abs/1409.0473
[61] F. Wang *et al.*, “Residual Attention Network for Image Classification,” Apr. 23, 2017, *arXiv*: arXiv:1704.06904. Accessed: Nov. 19, 2024. [Online]. Available: http://arxiv.org/abs/1704.06904

[62] J. Lu, C. Xiong, D. Parikh, and R. Socher, “Knowing When to Look: Adaptive Attention via A Visual Sentinel for Image Captioning,” Jun. 06, 2017, *arXiv*: arXiv:1612.01887. Accessed: Nov. 19, 2024. [Online]. Available: http://arxiv.org/abs/1612.01887
[63] J. Fu *et al.*, “Dual Attention Network for Scene Segmentation,” Apr. 21, 2019, *arXiv*: arXiv:1809.02983. Accessed: Nov. 19, 2024. [Online]. Available: http://arxiv.org/abs/1809.02983
[64] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, “Squeeze-and-Excitation Networks,” May 16, 2019, *arXiv*: arXiv:1709.01507. Accessed: Nov. 19, 2024. [Online]. Available: http://arxiv.org/abs/1709.01507
[65] O. Oktay *et al.*, “Attention U-Net: Learning Where to Look for the Pancreas,” May 20, 2018, *arXiv*: arXiv:1804.03999. Accessed: Nov. 19, 2024. [Online]. Available: http://arxiv.org/abs/1804.03999
[66] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, “CBAM: Convolutional Block Attention Module,” Jul. 2018, Accessed: Dec. 05, 2023. [Online]. Available: http://arxiv.org/abs/1807.06521
[67] N. Bhatt, N. Bhatt, P. Prajapati, V. Sorathiya, S. Alshathri, and W. El-Shafai, “A Data-Centric Approach to improve performance of deep learning models,” *Sci Rep*, vol. 14, no. 1, p. 22329, Sep. 2024, doi: 10.1038/s41598-024-73643-x.
[68] P. T. Jackson, A. Atapour-Abarghouei, S. Bonner, T. Breckon, and B. Obara, “Style Augmentation: Data Augmentation via Style Randomization,” Apr. 2019, Accessed: Dec. 05, 2023. [Online]. Available: http://arxiv.org/abs/1809.05375
[69] B. Thyreau, K. Sato, H. Fukuda, and Y. Taki, “Segmentation of the hippocampus by transferring algorithmic knowledge for large cohort processing,” *Medical Image Analysis*, vol. 43, pp. 214–228, Jan. 2018, doi: 10.1016/j.media.2017.11.004.
[70] H. Dong, G. Yang, F. Liu, Y. Mo, and Y. Guo, “Automatic Brain Tumor Detection and Segmentation Using U-Net Based Fully Convolutional Networks,” *Medical Image Understanding and Analysis*, vol. 723, pp. 506–517, 2017, doi: 10.1007/978-3-319-60964-5_44.
[71] H.-C. Shin *et al.*, “Medical Image Synthesis for Data Augmentation and Anonymization using Generative Adversarial Networks,” Sep. 2018, Accessed: Dec. 05, 2023. [Online]. Available: http://arxiv.org/abs/1807.10225
[72] N. Ibtehaz and M. S. Rahman, “MultiResUNet : Rethinking the U-Net architecture for multimodal biomedical image segmentation,” *Neural Networks*, vol. 121, pp. 74–87, Jan. 2020, doi: 10.1016/j.neunet.2019.08.025.
[73] P. Furtado, “Loss, post-processing and standard architecture improvements of liver deep learning segmentation from Computed Tomography and magnetic resonance,” *Informatics in Medicine Unlocked*, vol. 24, p. 100585, 2021, doi: 10.1016/j.imu.2021.100585.
[74] P. Krähenbühl and V. Koltun, “Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials,” presented at the NIPS 2011, arXiv, Oct. 2012. Accessed: Dec. 17, 2023. [Online]. Available: http://arxiv.org/abs/1210.5644
[75] N. Kamiya *et al.*, “Surface Muscle Segmentation Using 3D U-Net Based on Selective Voxel Patch Generation in Whole-Body CT Images,” *Applied Sciences*, vol. 10, no. 13, p. 4477, Jun. 2020, doi: 10.3390/app10134477.
[76] S. Biasotti, A. Cerri, A. Bronstein, and M. Bronstein, “Recent Trends, Applications, and Perspectives in 3D Shape Similarity Assessment,” *Computer Graphics Forum*, vol. 35, no. 6, pp. 87–119, Sep. 2016, doi: 10.1111/cgf.12734.
[77] A. D’Souza, B. Bolsterlee, A. Lancaster, and R. D. Herbert, “Muscle architecture in children with cerebral palsy and ankle contractures: an investigation using diffusion

tensor imaging," *Clinical Biomechanics*, vol. 68, pp. 205–211, Aug. 2019, doi: 10.1016/j.clinbiomech.2019.06.013.
[78] E. Lareau-Trudel *et al.*, "Muscle Quantitative MR Imaging and Clustering Analysis in Patients with Facioscapulohumeral Muscular Dystrophy Type 1," *PLoS ONE*, vol. 10, no. 7, p. e0132717, Jul. 2015, doi: 10.1371/journal.pone.0132717.
[79] G. Davico *et al.*, "Intra-operator Repeatability of Manual Segmentations of the Hip Muscles on Clinical Magnetic Resonance Images," *J Digit Imaging*, vol. 36, no. 1, pp. 143–152, Oct. 2022, doi: 10.1007/s10278-022-00700-0.
[80] H. Tao, W. Li, X. Qin, and D. Jia, "Image semantic segmentation based on convolutional neural network and conditional random field," in *2018 Tenth International Conference on Advanced Computational Intelligence (ICACI)*, Xiamen: IEEE, Mar. 2018, pp. 568–572. doi: 10.1109/ICACI.2018.8377522.
[81] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," arXiv, May 2015. Accessed: Dec. 05, 2023. [Online]. Available: http://arxiv.org/abs/1505.04597
[82] D. T. Kushnure and S. N. Talbar, "MS-UNet: A multi-scale UNet with feature recalibration approach for automatic liver and tumor segmentation in CT images," *Computerized Medical Imaging and Graphics*, vol. 89, p. 101885, Apr. 2021, doi: 10.1016/j.compmedimag.2021.101885.
[83] D. Lisa, "Increased muscle fat infiltration is associated with reduced muscle strength in older women with obesity and dynapenia," 2023.
[84] William H. Henson, Lin Xinshan, Lin Zhicheng, Guo Lingzhong, Claudia Mazzá, and Enrico Dall'Ara, "Automatic segmentation of lower limb muscles from MR images of post-menopausal women based on deep learning and data augmentation," 2023.
[85] T. Bachlechner, B. P. Majumder, H. H. Mao, G. W. Cottrell, and J. McAuley, "ReZero is All You Need: Fast Convergence at Large Depth," Jun. 24, 2020, *arXiv*: arXiv:2003.04887. Accessed: Sep. 20, 2024. [Online]. Available: http://arxiv.org/abs/2003.04887
[86] N. Ravikumar, A. Gooya, S. Çimen, A. F. Frangi, and Z. A. Taylor, "Group-wise similarity registration of point sets using Student's t-mixture model for statistical shape models," *Medical Image Analysis*, vol. 44, pp. 156–176, Feb. 2018, doi: 10.1016/j.media.2017.11.012.
[87] A. L. Clouthier, C. R. Smith, M. F. Vignos, D. G. Thelen, K. J. Deluzio, and M. J. Rainbow, "The effect of articular geometry features identified using statistical shape modelling on knee biomechanics," *Medical Engineering & Physics*, vol. 66, pp. 47–55, Apr. 2019, doi: 10.1016/j.medengphy.2019.02.009.
[88] Joshua Cates, Shireen Elhabian, and Ross Whitaker, "Shapeworks: particle-based shape correspondence and visualization software.," in *Statistical Shape and Deformation Analysis*, Academic Press, 2017.
[89] D. Barber and D. Hose, "Automatic segmentation of medical images using image registration: diagnostic and simulation applications," *Journal of Medical Engineering & Technology*, vol. 29, no. 2, pp. 53–63, Jan. 2005, doi: 10.1080/03091900412331289889.
[90] D. C. Barber, E. Oubel, A. F. Frangi, and D. R. Hose, "Efficient computational fluid dynamics mesh generation by image registration," *Medical Image Analysis*, vol. 11, no. 6, pp. 648–662, Dec. 2007, doi: 10.1016/j.media.2007.06.011.
[91] K. R. O'Brien *et al.*, "Robust T1-Weighted Structural Brain Imaging and Morphometry at 7T Using MP2RAGE," *PLoS ONE*, vol. 9, no. 6, p. e99676, Jun. 2014, doi: 10.1371/journal.pone.0099676.
[92] "MRI Signal-to-Noise Ratio (SNR) and Image Quality," mrimaster.com. [Online]. Available: https://mrimaster.com/snr/

[93] G. Ioannides, A. Chadha, and A. Elkins, “Gaussian Adaptive Attention is All You Need: Robust Contextual Representations Across Multiple Modalities,” Jan. 30, 2024, *arXiv*: arXiv:2401.11143. Accessed: Sep. 26, 2024. [Online]. Available: http://arxiv.org/abs/2401.11143
[94] J. D. De Certaines *et al.*, “Application of texture analysis to muscle MRI: 1-What kind of information should be expected from texture analysis?,” *EPJ Nonlinear Biomed Phys*, vol. 3, no. 1, p. 3, Mar. 2015, doi: 10.1140/epjnbp/s40366-015-0017-1.
[95] H. Dieckhaus *et al.*, “Logistic Regression–Based Model Is More Efficient Than U-Net Model for Reliable Whole Brain Magnetic Resonance Imaging Segmentation,” *Topics in Magnetic Resonance Imaging*, vol. 31, no. 3, pp. 31–39, Jun. 2022, doi: 10.1097/RMR.0000000000000296.
[96] H. Ning, Z. Li, C. Wang, and L. Yang, “Choosing an appropriate training set size when using existing data to train neural networks for land cover segmentation,” *Annals of GIS*, vol. 26, no. 4, pp. 329–342, Oct. 2020, doi: 10.1080/19475683.2020.1803402.
[97] F. Galbusera, A. Cina, M. Panico, D. Albano, and C. Messina, “Image-based biomechanical models of the musculoskeletal system,” *Eur Radiol Exp*, vol. 4, no. 1, p. 49, Dec. 2020, doi: 10.1186/s41747-020-00172-3.
[98] P. M. Graffy, J. Liu, P. J. Pickhardt, J. E. Burns, J. Yao, and R. M. Summers, “Deep learning-based muscle segmentation and quantification at abdominal CT: application to a longitudinal adult screening cohort for sarcopenia assessment,” *British journal of radiology*, vol. 92, no. 1100, 2019.
[99] Y. García-Alonso, A. M. Alonso-Martínez, A. García-Hermoso, G. Legarra-Gorgoñon, M. Izquierdo, and R. Ramírez-Vélez, “Centile reference curves of the ultrasound-based characteristics of the rectus femoris muscle composition in children at 4–11 years old,” *Front. Pediatr.*, vol. 11, p. 1168253, Aug. 2023, doi: 10.3389/fped.2023.1168253.
[100] R. Dotan, C. Mitchell, R. Cohen, P. Klentrou, D. Gabriel, and B. Falk, “Child—Adult Differences in Muscle Activation — A Review,” *Pediatric Exercise Science*, vol. 24, no. 1, pp. 2–21, Feb. 2012, doi: 10.1123/pes.24.1.2.
[101] A. Zapaishchykova *et al.*, “Automated temporalis muscle quantification and growth charts for children through adulthood,” *Nat Commun*, vol. 14, no. 1, p. 6863, Nov. 2023, doi: 10.1038/s41467-023-42501-1.
[102] N. Decaux *et al.*, “Semi-automatic muscle segmentation in MR images using deep registration-based label propagation,” *Pattern Recognition*, vol. 140, p. 109529, Aug. 2023, doi: 10.1016/j.patcog.2023.109529.
[103] S. Oliviero *et al.*, “Reproducibility of Densitometric and Biomechanical Assessment of the Mouse Tibia From In Vivo Micro-CT Images,” *Front. Endocrinol.*, vol. 13, p. 915938, Jun. 2022, doi: 10.3389/fendo.2022.915938.
[104] A. Fedorov *et al.*, “3D Slicer as an Image Computing Platform for the Quantitative Imaging Network,” 2013.
[105] P. Li, P. E. Highfield, Z.-Q. Lang, and D. Kell, “Cervical cancer prognosis and diagnosis using electrical impedance spectroscopy,” *Journal of Electrical Bioimpedance*, vol. 12, no. 1, pp. 153–162, Dec. 2021, doi: 10.2478/joeb-2021-0018.
[106] Z. Lin, E. Dall’Ara, and L. Guo, “A novel mean shape based post-processing method for enhancing deep learning lower-limb muscle segmentation accuracy,” *PLoS ONE*, vol. 19, no. 10, p. e0308664, Oct. 2024, doi: 10.1371/journal.pone.0308664.
[107] L. H. Clemmensen and R. D. Kjærsgaard, “Data Representativity for Machine Learning and AI Systems,” Feb. 03, 2023, *arXiv*: arXiv:2203.04706. Accessed: Sep. 13, 2024. [Online]. Available: http://arxiv.org/abs/2203.04706
[108] S. Bozkurt *et al.*, “Reporting of demographic data and representativeness in machine learning models using electronic health records,” *Journal of the American Medical*

Informatics Association, vol. 27, no. 12, pp. 1878–1884, Dec. 2020, doi: 10.1093/jamia/ocaa164.
[109] T. Maly, F. Zahalka, L. Mala, and P. Cech, "The bilateral strength and power asymmetries in untrained boys," *Open Medicine*, vol. 10, no. 1, Apr. 2015, doi: 10.1515/med-2015-0034.
[110] O. S. Al-Kadi and D. Watson, "Susceptibility of texture measures to noise: An application to lung tumor CT images," in *2008 8th IEEE International Conference on BioInformatics and BioEngineering*, Athens, Greece: IEEE, Oct. 2008, pp. 1–4. doi: 10.1109/BIBE.2008.4696789.
[111]J. Lexell and C. C. Taylor, "Variability in muscle fibre areas in whole human quadriceps muscle: effects of increasing age".
[112] M. S. Brook *et al.*, "Skeletal muscle homeostasis and plasticity in youth and ageing: impact of nutrition and exercise," *Acta Physiologica*, vol. 216, no. 1, pp. 15–41, Jan. 2016, doi: 10.1111/apha.12532.
[113] A. J. Létocart *et al.*, "Muscles adaptation to aging and training: architectural changes – a randomised trial," *BMC Geriatr*, vol. 21, no. 1, p. 48, Dec. 2021, doi: 10.1186/s12877-020-02000-0.
[114] E. O. Wesselink *et al.*, "Convolutional neural networks for the automatic segmentation of lumbar paraspinal muscles in people with low back pain," *Sci Rep*, vol. 12, no. 1, p. 13485, Aug. 2022, doi: 10.1038/s41598-022-16710-5.
[115] K. A. Weber *et al.*, "Multi-muscle deep learning segmentation to automate the quantification of muscle fat infiltration in cervical spine conditions," *Sci Rep*, vol. 11, no. 1, p. 16567, Aug. 2021, doi: 10.1038/s41598-021-95972-x.
[116] D. M. McSweeney *et al.*, "Transfer learning for data-efficient abdominal muscle segmentation with convolutional neural networks," *Medical Physics*, vol. 49, no. 5, pp. 3107–3120, May 2022, doi: 10.1002/mp.15533.
[117] E. A. Audenaert, C. Pattyn, G. Steenackers, J. De Roeck, D. Vandermeulen, and P. Claes, "Statistical Shape Modeling of Skeletal Anatomy for Sex Discrimination: Their Training Size, Sexual Dimorphism, and Asymmetry," *Front. Bioeng. Biotechnol.*, vol. 7, p. 302, Nov. 2019, doi: 10.3389/fbioe.2019.00302.
[118] C. Galleguillos, A. Rabinovich, and S. Belongie, "Object categorization using co-occurrence, location and appearance," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, AK, USA: IEEE, Jun. 2008, pp. 1–8. doi: 10.1109/CVPR.2008.4587799.
[119] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie, "Objects in Context," in *2007 IEEE 11th International Conference on Computer Vision*, Rio de Janeiro, Brazil: IEEE, 2007, pp. 1–8. doi: 10.1109/ICCV.2007.4408986.
[120] J. Lafferty, A. McCallum, and Fernando Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data," International Conference on Machine Learning, 2001.
[121] K. Nagawa *et al.*, "Texture analysis of muscle MRI: machine learning-based classifications in idiopathic inflammatory myopathies," *Sci Rep*, vol. 11, no. 1, p. 9821, May 2021, doi: 10.1038/s41598-021-89311-3.
[122] M. Rohm, M. Markmann, J. Forsting, R. Rehmann, M. Froeling, and L. Schlaffke, "3D Automated Segmentation of Lower Leg Muscles Using Machine Learning on a Heterogeneous Dataset," *Diagnostics*, vol. 11, no. 10, p. 1747, Sep. 2021, doi: 10.3390/diagnostics11101747.
[123] A. Alruwaili and M. Alsalim, "Data diversity and its impact on machine learning fairness," *Int. J. Cloud Comput. Database Manage.*, vol. 5, no. 1, pp. 38–41, Jan. 2024, doi: 10.33545/27075907.2024.v5.i1a.60.

[124] H. B. Braiek and F. Khomh, “Machine Learning Robustness: A Primer,” May 04, 2024, *arXiv*: arXiv:2404.00897. doi: 10.48550/arXiv.2404.00897.