Automatic quadriceps and patellae segmentation of MRI with cascaded U 2 -Net and SASSNet deep learning model

Ruida Cheng, Marion Crouzier, François Hug, Kylie Tucker, Paul Juneau, Evan Mccreedy, William Gandler, Matthew Mcauliffe, Frances Sheehan

HHS Public Access

Author manuscript

Med Phys. Author manuscript; available in PMC 2023 January 01.

Published in final edited form as:

Med Phys. 2022 January ; 49(1): 443–460. doi:10.1002/mp.15335.

Automatic Quadriceps and Patellae Segmentation of MRI with Cascaded U²-Net and SASSNet Deep Learning Model

Ruida Cheng[1], **Marion Crouzier**[2,3], **François Hug**[4,5], **Kylie Tucker**[3], **Paul Juneau**[6], **Evan McCreedy**[1], **William Gandler**[1], **Matthew J. McAuliffe**[1], **Frances T. Sheehan**[7]

[1]Scientific Application Services (SAS), Office of Scientific Computing Services (OSCS), Office of Intramural Research, Center of Information Technology, NIH, Bethesda, MD, USA

[2]University of Nantes, Movement, Interactions, Performance, MIP, EA 4334, F-44000 Nantes, France

[3]The University of Queensland, School of Biomedical Sciences, Brisbane

[4]Institut Universitaire de France (IUF), Paris, France

[5]Université Côte d'Azur, LAMHESS, Nice, France.

[6]NIH Library, Office of Research Services, National Institutes of Health, Bethesda, MD, USA

[7]Rehabilitation Medicine Department, National Institutes of Health Clinical Center, Bethesda, MD, USA

Abstract

Purpose: Automatic muscle segmentation is critical for advancing our understanding of human physiology, biomechanics, and musculoskeletal pathologies, as it allows for timely exploration of large multi-dimensional image sets. Segmentation models are rarely developed/validated for the pediatric model. As such, auto-segmentation is not available to explore how muscle architectural changes during development and how disease/pathology affects the developing musculoskeletal system. Thus, we aimed to develop and validate an end-to-end, fully automated, deep learning model for accurate segmentation of the rectus femoris and vastus lateral, medialis, and intermedialis using a pediatric database.

Methods: We developed a two-stage cascaded deep learning model in a coarse-to-fine manner. In the first stage, the U^2-Net roughly detects the muscle sub-compartment region. Then, in the second stage, the Shape-aware 3D semantic segmentation method SASSNet refines the cropped target regions to generate the more finer and accurate segmentation masks. We utilized multi-feature image maps in both stages to stabilize performance and validated their use with an ablation study. The second stage SASSNet was independently run and evaluated with three different cropped

Corresponding author: Ruida Cheng (ruida@nih.gov) Phone:301-496.5363, 12/2009 National Institutes of Health, Bethesda, MD, 20892.

Conflict of Interest: None to report for any author.

Ethical Review: Ethical approval for this study was provided by The University of Queensland Institutional Human Research Ethics Committee #2018000159

The data of this paper is not available for sharing at the current time. This issue is due to the internal policy agreement of the NIH IRB and The University of Queensland IRB.

region resolutions: the original image resolution, and images down-sampled 2x & 4x (high, mid, and low). The relationship between image resolution and segmentation accuracy was explored. In addition, the patella was included as a comparator to past work. We evaluated segmentation accuracy using leave-one-out testing on a database of 3D MR images (0.43×0.43×2mm) from 40 pediatric participants (age 15.3±1.9years, 55.8±11.8kg, 164.2±7.9cm, 38F/2M).

Results: The mid-resolution second stage produced the best results for the vastus medialis, rectus femoris, and patella (Dice Similarity Coefficient = 95.0%, 95.1%, 93.7%), whereas the low-resolution second stage produced the best results for the vastus lateralis and vastus intermedialis (DSC = 94.5% and 93.7%). In comparing the low- to mid-resolution cases, the vasti intermedialis, vastus medialis, rectus femoris, and patella produced significant differences (p=0.0015, p=0.0101, p<0.0001, p=0.0003) and the vasti lateralis did not (p = 0.2177). The high-resolution Stage2 had significantly lower accuracy (1.0 to 4.4 Dice percentage points) compared to both the mid- and low-resolution routines (p ranged from <0.001 to 0.04). The one exception was the rectus femoris, where there was no difference between the low and high-resolution cases. The ablation study demonstrated that the multi-feature is more reliable than the single feature.

Conclusions: Our successful implementation of this two-stage segmentation pipeline provides a critical tool for expanding pediatric muscle physiology and clinical research. With a relatively small and variable dataset, our fully automatic segmentation technique produces accuracies that matched or exceeded the current state of the art. The two-stage segmentation avoids memory issues and excessive run times by using a first stage focused on cropping out unnecessary data. The excellent Dice similarity coefficients improve upon previous template-based automatic and semi-automatic methodologies targeting the leg musculature. More importantly, with a naturally variable dataset (size, shape, etc.), the proposed model demonstrates slightly improved accuracies, compared to previous neural networks methods.

Keywords

quadriceps muscle; segmentation; patella; magnetic resonance imaging; human; child; neural networks; cascaded; deep learning

1.0 Introduction

The ability to define muscular volume and morphology is critical to understanding human physiology, biomechanics, and musculoskeletal pathologies. As manual segmentation of individual muscles is time consuming, numerous semi- and fully-automatic muscle-segmentation are rapidly becoming available[1–4]. However, an interesting, and more challenging problem, has remained unanswered. These segmentation routines have rarely been developed or validated for the pediatric model[5,6]. The rapidly changing muscle volumes[7], muscle-fat ratios[8], and potentially shape during childhood development likely limits the ability to apply auto-segmentation routines developed for adult muscle to pediatric muscle. Thus, auto-segmentation is not available to explore how muscle physiology changes[9,10] with development nor how disease affects the developing musculoskeletal system (e.g., cerebral palsy[11,12], brachial plexus palsy[13], scoliosis[14], etc.). Studies exploring the ever-changing 3D pediatric musculoskeletal morphology[15] remain challenging and pediatric muscle volume is often approximated using surrogates (e.g., muscle cross-sectional

area[16]). When full volumes of individual muscles are quantified, the scope has been limited to small databases[9,11,12,14,15], examining total limb or muscle group volume[9,15], and typically to a single muscle[14].

Conventional auto- and semi-automatic segmentation methods (Table 1) for the leg musculature rely heavily on traditional algorithmic approaches, specifically image registration[17–22] and atlas-based methods[4,23–25]. The ambiguous boundaries, caused by homogeneity texture appearance, is a key source of error across all these algorithms. Even with a standardized imaging protocol, intensity ranges and the spatial location of the same anatomical regions can vary. Further, the size and shape of muscles varies across individuals and is affected by injury and pathology[9,10,13,14]. Le et al.[25] developed both single-atlas and multi-atlas fully-automatic segmentation methodologies based on non-linear registrations (e.g., free-form deformation and symmetric diffeomorphic normalization) to segment the quadriceps and each of its four individual heads. Baudin et al.[18] associated a statistical shape atlas and a random walks graph-based algorithm to automatically segment the individual quadriceps muscles. Ogier et al.[21] proposed semi-automatic segmentation of individual quadriceps muscles using automatic propagation on consecutive 2D slices through non-linear registration based on initial delineation of mask. This was expanded to non-linear registration via both the transversal and longitudinal propagations on water contrast MR images[26]. Prescott et al.[22] designed a semi-automated method of segmenting the four quadriceps muscles, a template selection method followed by a multi-phase level set contour evolution of a contraction phase and an expansion phase on pre-processed images to capture anatomical variations in a specific participant. These methodologies demonstrated an average Dice similarity score (DSC) of 0.86 (ranging from 0.75 for the rectus femoris[23] to DSC= 0.93 for the vastus lateralis[25] and vastus medialis[23]). Thus, there was still clearly room for improvement.

In recent years, convolutional neural networks (Appendix 1, Table 1), such as U-Net[27], have shown promising results in semantic segmentation for medical imaging. Deep learning models vastly improve segmentation accuracy, relative to their predecessors, becoming the primary tool in the auto-segmentation of both muscle[2,3] and bone[28,29] from MR images. Ding et al.[27] trained and validated a standard 2D U-Net to segment the quadriceps in its entirety using fat-water decomposition MRI, similar to earlier techniques[30–35]. Ni et al.[3] applied a two-staged cascaded 3-D DCNN[36] model on high-resolution MRI with 35 individual lower limb muscles. They used a two-staged cascaded model. Chen et al.[1] developed a CNN-based 3D U-Net to segment the thigh muscles on B1-corrected 3D Dixon fat-water decomposition MR images. Although these two latter routines, produced impressive DSC scores (0.93–0.98), these methods are limited in three key area. The dataset of Ni et al.[3] fostered increased DSC through data uniformity. All data were acquired on athletic college students and data from both legs, which have an inherent interdependence, were entered as independent samples. Chen et al.[1] focused the segmentation on the central thigh, avoiding the muscular origins and insertions, which are difficult to segment due the smaller muscular areas and the rapidly changing shape and size. Lastly, the overall number of testing cases was small.

One challenge in segmentation is that ground truth labels are typically available for one side (either left or right), while the MR images capture both legs (Appendix 1: Figure A1, GT). Applying conventional segmentation methods directly on the full-scale images, in general, will introduce symmetry errors due to the similar anatomic features on the mirroring side (Appendix 1: Figure A1, U-Net, U-Net++, DAF3D). A similar, but different issue is that of low-distribution errors when the object of interest is represented in only a small percentage of the images. Both of these issues have typically been handled by manually cropping the image prior to the auto-segmentation[1,3,4,17–22,25,27]. Lastly, as the resolution of MR images have rapidly improved, GPU bottlenecks have become an increasing problem[37].

To address the aforementioned problems, we set out to develop and validate a simple, yet effective cascaded U^2-Net and SASSNet model for individual quadriceps muscle (in their entirety) segmentation on MR images, specifically targeting a pediatric cohort. The model is prototyped in a cascaded fashion that generates refined segmentation masks in a coarse-to-fine manner. The MR data for training and testing are obtained from an inherently diverse adolescent population, in terms of developmental stage and health status (i.e., both typically developing adolescents[38] and adolescents with patellofemoral pain[39,40]). A leave-one-out strategy is used for testing, to maximize the data for training and testing. The segmentation includes the patella bone, as a comparator to our past work[5] and as a test for the models ability to handle low-distribution errors. As part of the validation, we test the stabilization effect of multi-feature filters and how alterations in the MR image resolution in the 2nd cascade affects accuracy. For a direct comparison to past auto-segmentation results, the same leave-one-out testing was conducted using four conventional architectures.

2.0 Methods

2.1 MRI datasets and ground truth segmentation

As part of an IRB-approved (Ethical approval for this study was provided by The University of Queensland Institutional Human Research Ethics Committee #2018000159) study on patellofemoral data, 40 adolescents (38 females and 2 males) were enrolled into this study. Prior to any data collection, all procedures were explained to the adolescents and their legal guardian. Written assent and consent were obtained. 3D axial steady state Vibe images (0.43mm × 0.43mm × 2.0mm, 1024 × 768 × 140–366 pixels, TR = 9 ms, TE= 2.26 ms, flip angle = 10^O) were acquired for each participant (3T, Magnetom Prisma, Siemens Healthcare, Germany). The scan area covered both legs from just below the tibial tuberosity to just above the anterior iliac spine. The first cohort (n=20, Table 1) were adolescents with diagnosed patellofemoral pain (19 F, age 12–18 years, weight 40–94kg, height 150–175cm). The 2nd cohort were typically developing adolescents without patellofemoral pain (19 F, age 12–18 years, weight 42–85kg, height 147–182cm), matched for gender, age, and body mass index (the latter two were within 15%).

The senior author (FG), with 20 years of experience creating musculoskeletal models from MR data, worked with author MC to define a set of criteria for manual segmentation. Then a single research assistant manually delineated the outer boundaries for the vastus lateralis, medialis, and intermedius (VL, VM, VI), the rectus femoris (RF), and the patella (PA) in a single leg for each participant using Mimics (Materialise, Belgium). The outlines (VOIs

– volumes of interest) were reviewed by the senior author. In preparation for the project's training step, we converted each individual VOI into a 3D boundary mask image set. The boundary mask and original image set were mirrored at the full-scaled image level without cropping for images with the right leg musculature delineated. This created a reference ("ground truth") dataset of all "left" legs, which was the basis for training and evaluating the auto-segmentation process.

2.2 Segmentation

2.2a Pipeline Overview—Our automated segmentation of the individual quadriceps' muscles and patella (QM&P) is based on a two-stage cascaded deep learning model (Figure 1). In the first stage, we localize the individual region of interest (VL, VM, VI, RF, or PA) as a coarse segmentation from the entire 3D MR image set. We first down-sample the original images into the low-resolution image space[27]. We utilize the multi-feature images (anisotropic diffusion, coherence enhanced diffusion, regularized diffusion, gradient magnitude) throughout the two-stage segmentation pipeline to stabilize the performance[5]. A U^2-Net[41] architecture generates the prediction probability maps and localizes the bounding box for the specific VOI. In the second stage, the segmentation crops based on the bounding box and integrates the semantic cues with shape aware 3D segmentation for boundary refinement[42]. The segmentation only focuses on the appropriately zoomed or cascaded QM&P region and spatial extents generated from the first stage in this second stage. In the testing phase, the final predicted high-resolution cropped 3D volumes are converted back to the original image space to compare with the ground truth binary masks. The proposed method generates boundary-preserving pixel-wise class label maps that result in the final segmentation.

2.2b Automatic Data Preparation—The high-resolution 3D volumes used in the current study will, in general, exhaust the deep learning model training with overwhelming computational overhead on GPUs. In addition, the intensity variation, noise artifacts, and low-contrast regions create major obstacles for the MR image segmentation[37]. For example, the edge boundaries might be missing or blended with the surrounding tissues. We compensate for this intensity variation by implementing a histogram equalization as an automatic preprocessing step in the auto-segmentation pipeline[43]. Then, we apply four edge enhanced filters to the normalized images, generating multi-featured image maps (Figure 1). This pre-processing step efficiently supports the proposed cascaded model in learning the semantic features and stabilizing the performance. To demonstrate the stabilization effect of the multi-feature filters, we conducted a leave-one-out cross-validation experiment, comparing the segmentation results of a single feature (original image) vs. multi-feature image set (Figure 1).

To curb the high 3D volume issue in the first stage, we down-sample each 2D image of the 3D set by a factor of 4. This creates a uniform dimensionality for all image sets across all participants ($256 \times 192 \times 140$–$366$) and allows for feasible multi-featured image processing and reasonable training time in the first stage.

2.2c Stage 1: U^2-Net for individual structure localization—Direct segmentation of QM&P on the full-scale image is problematic since the symmetric anatomical structure of the contralateral limb fosters noisy segmentation (Appendix 1: Figure A1, U-Net, U-Net++ and DAF3D, Appendix 1). Conventional deep learning-based object detection methods in medical imaging localization may not guarantee that the predicted bounding box contains the targeted object with high sensitivities on the pixel-level coverage[44]. To solve this problem, we use a simple, yet powerful deep learning architecture, U^2-Net[41] to automatically identify the most attractive region of interest on images. The saliency map[41] allows us to distinguish the important part of the image at the foreground from the background. The output (bounding box VOI for segmentation) from the U^2-Net localization phase is then fed into a more refined segmentation incorporating a shape-ware representation from the 3D semantic segmentation.

The architecture of U^2-Net (Figure 2) is a two-level nested U-structure[45]. At the bottom level, a residual U-block (RSU) extracts multi-scale features without reducing the resolution of the feature map[41]. At the top level, the building block is designed with a similar structure to U-Net, where each level is filled with RSU blocks. (Figure 2, colored U-Net alike structure). The critical difference between RSU and the conventional residual block is that RSU replaces the plain, single-stream convolutional with a structure like U-Net and replaces the original feature map with a local feature map transformed by a weighted layer. This architecture shift empowers the network to extract features from multiple scales directly from each residual block.

In the U^2-Net architecture, each stage of the encoder-decoder U-Net structure contains an RSU block, which is, in fact, a down-sampling/up-sampling encoder-decoder itself[41]. The RSU block uses multi-scale features as residuals, not just the original features, for identifying structures of interest. This keeps the fine-grained details and will force the network to extract multiple scale feature at every RSU block. The U^2-Net generates side output probability maps through a plain 3×3 convolution layer, followed by a bilinear up-sampling and sigmoid function.

In the training process, the U^2-Net uses the deep supervision similar to HED[46] to minimize the overall training loss[41] ($\mathscr{L}$, equations 1 & 2), which is defined as:

$$\mathscr{L} = \sum_{m-1}^{M} w_{side}^{(m)} \ell_{side}^{(m)} + w_{fuse} \ell_{fuse} \quad (1)$$

41

$\ell_{side}^{(m)}$: the loss of side output saliency map[41]

ℓ_{fuse}: the loss of the final fusion output saliency map[41]

$w_{side}^{(m)}$ & w_{fuse}: the weights of each loss term.

The fusion loss is imposed after each side-output layer (side loss, Figure 2) to guide the side-outputs to minimize the distance between the predictions and the ground truth label maps at different multi-scale levels. In most cases of thigh muscle segmentation, minimizing

the loss will instantly make the network converge to classifying every pixel as background due to the large dominance fraction of background class in the full-scale MR images. A class-balanced cross-entropy loss is utilized to combat the significant data imbalance between foreground/background pixels for different-sized thigh muscles.

For each term ℓ, a standard binary cross-entropy loss[47] is defined as:

$$\ell = -\sum_{(r,c)}^{(H,W)} \left[P_{G(r,c)} log\, P_{S(r,c)} + \left(1 - P_{G(r,c)}\right) \log\left(1 - P_{S(r,c)}\right) \right] \quad (2)$$

[41,47]

(r, c): the pixel coordinates

(H, W): the height (H) and width (W) of the cropped image

$P_{G(r, c)}$: the ground truth pixel values

$P_{S(r, c)}$: the predicted saliency probability map[41] pixel values

Upon completing Stage 1, the condensed muscle or patella VOI is mapped back to the original image space, creating a high-resolution image set covering the finer detailed structures and pruning the unnecessary background image volume. The second stage uses these cropped high-resolution images to create multi-featured maps, which are then used to define the muscle or bone boundary.

To evaluate the impact of image resolution on segmentation accuracy, the 2nd stage was trained and tested on three different image sets with in-plane pixels resolutions of 0.44×0.44mm, 0.88×0.88mm, and 1.76×1.76mm (high-res, mid-res, and low-res, respectively). We assumed the mid-res resolution was our standard algorithm, as the original resolution (high-res) had a resolution 2–4 times greater than all, but two, previous studies (Table 1).

2.2d Stage 2: Shape aware 3D semantic segmentation—In the second stage we use a 3D shape-aware semantic segmentation network (SASSNet[42], Figure 3) to refine the boundary of each quadriceps muscle and the patella after they have been appropriately localized in the first phase. Li et al.[42] proposed a similar shape-aware semi-supervised 3D segmentation model. The motivation is to use shape-aware semi-supervised learning strategy[42] to leverage abundant unlabeled data[42], enforcing a geometric shape constraint on the segmentation output[48]. In our work, we use the shape-aware semantic segmentation architecture, to learn the interior and depth maps[48] from the tightened individual muscle and patella VOI. Object-level interior signed distance map (SDM)[42] of object surfaces and the shape priors[49] among different SDMs (Appendix 2) can provide the intra-level visual cues to capture shape-aware features more effectively. Despite the semi-supervised setting with unlabeled data, the shape-aware 3D semantic segmentation model serves as a suitable deep representation to learning general raw pixel-in and label-out mapping functions[44].

The 3D shape aware semantic segmentation network is designed based on three key elements: 1) it enforces explicit 3D modeling of the geometric constraints[48] on labeled and unlabeled data[42]; 2) it uses a multi-task learning on both the binary segmentation map

and the SDM[42] predictions; and 3) it imposes implicit shape priors[49] among different SDMs and global consistency on object shape via adversarial loss[50] for semi-supervised volumetric segmentation. The 3D shape-aware method[42] takes 3D images as input and jointly predicts a binary segmentation map and SDM to train the network. The backbone is a 3D V-Net architecture[51] that consists of an encoder and decoder with two output branches, one for the segmentation map and the other for the SDM[42].

In addition to the 3D V-Net backbone, the light-weighted SDM head along with the original segmentation head act as two parallel prediction branches. The segmentation head generates the pixel-wise probability map and the SDM head produces the depth probability map of the segmentation target. The SDM head is composed by a 3D convolution block followed by *tanh* activation. The segmentation network uses a 3D image as input and predicts the pixel-wise probability map and depth map simultaneously. To handle the unlabeled data, the SASSNet leverages a semi-supervised learning strategy, which learns from unlabeled data by minimizing the difference between the predicted SDMs on the labeled and unlabeled dataset. To enforce the consistency, a discriminator network is used to distinguish the predicted SDMs from the labeled set and the ones from unlabeled set. The algorithm learns the effective shape-aware features that generalizes well to the unlabeled dataset by minimizing an adversarial loss induced by this discriminator. The 3D discriminator (3D GAN) is designed to encourage the smooth prediction on the unlabeled data and to improve the overall segmentation accuracy when generalized to unseen image datasets. In the 2nd stage training phase, we intentionally reduce the number of available ground truth labels to enforce the 3D GAN to generate the synthetic pseudo labels during the training process. Those pseudo-labels and the given ground truth labels together guide the 3D V-Net to retrain the model iteratively until the algorithm converge. This semi-supervised sense of training improves the overall generalizability of the trained segmentation model.

The network formulation of the SASSNet semi-supervised learning is constrained with a shape-aware regularization term, a multi-task loss[52] that consists of a supervised loss ($\mathscr{L}_s$) on the labeled set and an adversarial loss[50] ($\mathscr{L}_a$) on the entire set (equations 3 & 4). The training set contains N labeled data and M unlabeled data, where $N \ll M$. We denote the labeled set as $D^l = \{X_n, Y_n, Z_n\}_{n=1}^{N}$ and unlabeled data as $D^u = \{X_m\}_{m=N+1}^{N+M}$, where $X_n \in \mathbb{R}^{H \times W \times D}$ are the inputs 3D image volumes, $Y_n \in \{0, 1\}^{H \times W \times D}$ are the binary segmentation masks[42] and $Z_n \in \mathbb{R}^{H \times W \times D}$ are the ground truth SDMs[42] derived from Y_n.

$$\mathscr{L}_s(\theta) = \frac{1}{N}\sum_{i=1}^{N} l_{dice}\left(f_{seg}(X_i;\theta), Y_i\right) + \alpha\left(\frac{1}{N}\sum_{i=1}^{N} l_{mse}(f_{sdm}(X_i;\theta), Z_i)\right) \quad (3)$$

[42]

α: balancing factor

$f_{seg}(X_i; \theta)$: the predicted segmentation probability map[42]

$f_{sdm}(X_i; \theta)$: the signed depth map[42] respectively.

$\frac{1}{N}\sum_{i=1}^{N} l_{dice}\left(f_{seg}(X_i;\theta), Y_i\right)$: the segmentation loss[42]

$\alpha\left(\frac{1}{N}\sum_{i=1}^{N} l_{mse}(f_{sdm}(X_i;\theta), Z_i)\right)$: the SDM loss[42]

$\mathscr{L}_a$ is imposed to regularize the model learning with the unlabeled data. It enforces the consistency of SDM predictions on the labeled and unlabeled images. A discriminator network differentiates the predicted SMDs from labeled and unlabeled image data, minimizing the adversarial loss to learn compelling shape-aware features that generalize well to the unlabeled images. The discriminator network (D) consists of 5 convolutions layers followed by an MLP. The network takes a 3D SDM and corresponding 3D image volume as input, fused them through convolution layers, and predicts its class probability of being labeled data. Given D, the adversarial learning parameter is denoted as γ and the adversarial loss is defined as:

$$\mathscr{L}_a(\theta, \gamma) = \frac{1}{N}\sum_{n=1}^{N} log\ D(X_n, S_n; \gamma) + \frac{1}{M}\sum_{m=n+1}^{N+M} log(1 - D(X_m, S_m; \gamma)) \quad (4)$$

[42]

$S_n = f_{sdm}(X_n; \theta)$: the predicted SDMs for labeled image data

$S_m = f_{sdm}(X_m; \theta)$: the predicted SDMs for unlabeled image data.

The overall training objective function $\mathscr{F}(\theta, \gamma)$ combines $\mathscr{L}_s$ and the $\mathscr{L}_a$ (equations 3 & 4) and the learning task ($\mathscr{F}(\theta, \gamma)$) is written as,

$$\mathscr{F}(\theta, \gamma) = \mathscr{L}_s(\theta) + \beta\mathscr{L}_a(\theta, \gamma) \quad (5)$$

[42]

β: a weight coefficient to balance the two losses.

Given a fixed discriminator $D(\cdot; \gamma)$, the objective function minimizes the binary cross entropy loss[47] induced using equation 3 to train the discriminator.

In the supervised training scenario with labeled image data, a supervised loss function (Eq. 3 & 4) employs a dice loss for the binary mask-based segmentation map and a mean square loss for the SDM based depth map of the multi-task[42] segmentation network. Irrespective of the loss function, most segmentation errors were located at the proximity of thigh muscle boundaries (Appendix 1, Figure A1). Nevertheless, the overall objective function (equation 5) proposes a simple strategy to penalize segmentation errors at the object boundaries utilizing SDM and binary mask with the adversarial mechanism. The proposed 3D V-Net backbone and the 3D GAN network are trained with distance-based loss penalty (SDM) and dice loss penalty (binary masks) to function as a fine-tuning strategy, effectively mitigating segmentation errors at the boundaries.

2.3 Training and testing

We train each cascaded stage separately (Figures 2 & 3). The training starts from scratch, as no existing backbone is used in our network to pre-trained model. In this first stage, we use the default setting of U^2-Net model hyper-parameter setting, batch size (12), learning rate (10−3), epsilon (10−7), weight decay (5 × 10−4), number of training iterations (50,000),

Adam optimizer. We omit the data augmentation step (i.e., rotation, flipping, transformation, and scaling) in the U^2-Net, utilizing the multi-feature maps as an alternative to data augmentation. In the second stage, the cropped higher-resolution 3D image volumes with paired ground truth labels are used to train the SASSNet model. For this stage, the hyper-parameters settings are: batch size (4), learning rate (10−3), discriminator learning rate (10−4), weight decay (0.0002), number of training iterations (6,000). Our experiments are conducted on leave-one-out cross-validation of the 40 MR images dataset. In each round of cross validation, we employ 39 images as the training set of the cascaded model, leaving one image out for testing. The single label-based training and testing scenarios are conducted individually with the cross-validation. The training and testing use the low-resolution images (resolution: 1.52 × 1.52 × 2 mm, image size: 256 × 192 × 140–366 pixels) in stage 1. The second cascade is trained and evaluated separately for three different resolution of the cropped images: low-, mid-, and high-res.

Both U^2-Net and SASSNet are implemented with PyTorch[53]. All the pre-processing (multi-featured maps), post-processing (3D morphology, transformation), 3D surface reconstruction, and 3D visualization are all implemented in the MIPAV[43] application.

To compare with the ground truth binary masks for each individual muscle and patella, the final predicted cropped 3D image masks from stage 2 are converted back to the original image space. The segmentation performance is evaluated with (1) DSC (%), (2) Jaccard (IoU, %), (3) Hausdorff distance (HD, mm), (4) average minimum surface-to-surface distance (ASD, mm), and (5) volumetric similarity coefficient (VSC). All metrics are calculated without trimming the ending contours and without cropping data to the probability interval (5% – 95% of the average value). Thus, outliers for a distance-based measure remain in the model. The mask-based performance measure uses the EvaluationSegmentation[54] tool to compare the ground truth and segmented masks.

To directly compare the current results with previous architectures, four other models (U-Net[45], DAF3D[55], U-Net++[56] with data augmentation, and U-Net++ without data augmentation[56]) were developed and validated on the same 40 datasets. We conducted the leave-one-out cross validation for all the four previous architectures.

Training the single U^2-Net model and single SASSNet model requires 24 hours and 12 hours, respectively, on the Nvidia p100 GPU card. The cascaded model takes up 2 to 5 minutes to produce a prediction map for a single image, depending on the size of the cropped region and resolution. The multi-feature map generation on the cropped images consumes most of the time in the testing phase. Stage 2 training requires 6–12 hours to train a single model with 6000 iterations, depending on the resolution (low-res: 6 hours, mid-res: 8 hours, high-res: 12 hours). In the testing phase, the stage 2 testing takes 10–40 seconds to generate the predicted result for a single cropped 3D region (low-res: 10 seconds, mid-res: 20 seconds, high-res: 40 seconds).

2.4 Statistics

We fit a linear mixed model, employing the participant as a random effect and the resolution as a fixed effect to the data to determine if the 3 resolution cases in stage 2 produced

different accuracies (DSC scores). A p-value less than 0.05 was considered significant. This model was deemed appropriate as the DSC scores met the normality and homoscedasticity criteria.

3.0 Results

The DSC scores, based on mid-res stage 2 (Table 2), were quite similar across all muscles and the patella, with the RF having the best score (DSC=95.3%) and the VI have the worst score (93.2%). When the low-resolution images were used in stage 2, there was a significant, but slight increase in accuracy for the VI and a significant, but slight loss in accuracy seen in the VM, RF, and patella (Figure 4). Except when compared to the low-res for the RF, the high-res stage 2 produced significantly lower accuracies (89.4%-94.1%) than either the low- or mid-resolution 2nd stage cases (Figure 4).

Across all participants, the distance between the automatically generated surface and the ground truth (Figures 5 & 6), typically fell within the mid-resolution pixel size (±0.88mm). The largest disagreements between the ground truth and the automatically segmented models were concentrated in the proximal and distal aspects of the muscles, where the muscle tapered, rapidly changing shape and size across images (Figures 5 & 6). These end regions also had high variability across participants (Figure 7).

The multi-feature approach produced consistently higher DICE scores (Table 3, 3–13 percentage points) and less variability across image sets (2–7 percentage points). The single-feature approach imposes apparent under-segmentation and over-segmentation errors, whereas the multi-feature approach does not. (Figure 8).

The four previous architectures were unable to produce the same segmentation accuracy as the current model (Figure A1, Appendix 1). U-Net++ without data augmentation was the best of these techniques, producing Dice scores 1.7–5.2 percentage points below the current method, whereas U-Net was the worst, producing Dice scores 53.6–87.9 percentage points below the current (Table 4).

4.0 Discussion

Our successful implementation of this two-stage segmentation pipeline provides a critical tool for expanding muscle physiology research into pediatric databases. The two-stage segmentation allows for the robustness afforded by high resolution images, but avoids memory issues and exorbitant run times by using a first stage with reduced resolution images to crop out unnecessary data. Two key findings for future algorithms are that highest resolution images did not lead improved results and that the multi-featured image maps significantly enhance accuracy by providing multiple boundary definitions. The excellent DSC scores clearly improve upon previous template based automatic and semi-automatic methodologies targeting the leg musculatures (Table 1). More importantly, with a variable pediatric dataset, the methodology also improves accuracies, relative to previous neural networks methodologies.

From the application standpoint, we developed a 2-stage cascaded U^2-Net[41] and SSASNet model as an end-to-end system for MR thigh muscle segmentation. This model tackles the general problems of medical imaging segmentation tasks, such as low-scale image dataset, highly imbalanced ground truth labels, anatomical constraints, and memory issues. CNN-based methods, like the conventional baselines (e.g., U-Net, 3D U-Net, and U-Net++), require a reasonably large training dataset to obtain high segmentation accuracy[56], which can lead to memory issues with the high-resolution images[37]. Moreover, imbalanced medical image data and high variability of target object shapes and locations often lead to unexpected segmentation results[57]. The anatomic symmetry and low distribution errors were ignored in previous models by a priori cropping[1,3,4,17–22,25,27]. The 2nd refinement stage alleviates the memory bottleneck by focusing only on the cropped high-resolution volume, provided by the low-resolution 1st stage. Ni et al.[3] employed a similar two-stage cascaded 3D U-Net. However, our upgrading to the cascaded U^2-Net and SSASNet improved stability and performance (Table 4), even with our inherently variable dataset. Previously, general data augmentation has provided improved performance, but produces highly correlated data, limiting the information it generates. Thus, we exploited the multi-feature maps, instead of general data augmentation, throughout the 2-stage cascade model to attain a stable performance. A vital component of our current model is the ability of the SASSNet to use SDM and adversarial learning in addition to the binary mask to enforce the geometric constraint and the generalizability, improving stability. With the 2nd stage mid-resolution testing scheme we achieved patellar segmentation accuracies (DSC = 93.75%) on par with the muscles, although the patella was only represented in 7% of the image set. This demonstrates that the proposed cascaded model can manage the low distribution label issue.

In comparing to previous studies, the current methodology clearly advances our ability to automatically segment muscle. After an extensive review of the literature, 15 manuscripts[1–4,17–26] focusing on segmentation of the lower leg musculature were found (Table 1). Half are not comparable, as only the central part of the muscle was segmented[1,17,20,21,23,26]. Using just the central muscle belly inflates that accuracy as automatic segmentation is most error-prone closest to its origins and insertions, where the muscle is rapidly changing size and shape (Figures 5&7). Further, most research questions based on muscle segmentation require the segmentation of the entire muscle. When comparing directly to previous techniques that segmented the entire muscle, the current technique clearly improves accuracy, relative to non-neural network techniques (e.g., template[22], atlas[24], and random walks[18]). In general, our results were slightly better than the remaining study[3]. Ni et al.[3] focused purely on college-age athletes (sex was not reported). This similarity in developmental stage and fitness may be associated with less variation in muscle-fat ratio, muscle size, and muscle shape across their participants, relative to our cohort of 12–18-year-old participants (Figure 7). This research team added further homogeneity by incorrectly assuming data from the paired legs of the same participant were independent samples. These two sources of uniformity in the previous study likely inflated the accuracies. Our direct comparison to four previous architectures (Table 4) clearly supports the advanced accuracy of the current, relative to previous models.

A secondary, but key finding, is that higher-resolution images did not always lead to improved accuracy. The smaller objects (the rectus femoris and patella) saw the largest

accuracy increases between the low-res and med-res stage 2 models, however using the high-res did not add to these gains. Thus, future implementations need to be keenly aware of the trade-offs between computational time, accuracy, and the inherent size of the object being segmented. Conversely, including multi-feature image maps in future models should improve prediction performance, based on the current ablation study.

The current results for the patellar segmentation were only slightly worse than our previous auto-segmentation of the patella (DSC score = 94.7%, tri-planar segmentation). This degraded performance is most likely due to the loss of resolution in the z-direction for the current study (2 vs. 1 mm). On average, the patella spanned 7% of the full image set, whereas the muscles spanned 57%-62%. Thus, training the current model to this previous dataset would likely produce improved results.

This study was limited by two inherent properties in our dataset. First, the data were collected for a project focused on the quadriceps musculature in adolescents. Thus, our algorithm focused on automatically and completely segmenting each of the 4 quadriceps muscles[22,25], as only the 4 quadriceps muscles and the patella were manually segmented. We were unable to develop an algorithm for the entire thigh or lower limb. As our results were a slight improvement from the full lower leg segmentation of Ni et al.[3], it is logical to assume that our methodology would produce robust results for the entire leg. Although our training and testing datasets were larger than nearly all previous studies (Table 1), in terms of auto-segmentation, the data set is still relatively small. As such, our results may underestimate the true accuracy of the technique.

In conclusion, this work presents a cascade U^2-Net and SASSNet model an end-to-end system for quadriceps muscle and patella segmentation. In overcoming key obstacles in medical imaging segmentation, our model improved segmentations accuracies over previous techniques, despite the focus on the novel and variable pediatric database. The proposed model alleviates the symmetric anatomy issue and achieves relatively high segmentation performance with the low-scale dataset (n=40). The cascaded U^2-Net and SSASNet coarse-to-fine segmentation mechanism effectively tackles the label imbalance problem. These contributions advance this work not only based on improved accuracy, but fundamentally from an applications perspective. A promising future direction is exploring transfer learning to generalize the model from low-scale dataset to large dataset, incorporating meta-learning and few-shot learning to further fine-tuning to different datasets.

Acknowledgements:

This work was funded by the Intramural Research Program of the National Institutes of Health Clinical Center and the National Institute of Health, and an Eventide Homes Grant and Arthritis South Australia and Western Australia Grant through Arthritis Australia.

Appendix 1: Comparison of conventional neural network architectures

The U-Net[45] is the most popular deep learning model for many medical imaging segmentation tasks. The U-Net comprises a contracting path (down-sampling encoder) to capture contextual information through a compact feature map; and a symmetric expanding

path (up-sampling decoder) which allows precise localization to retain spatial information. The skip connections in U-Net connect the encoder to the decoder at each multi-scale level and fuses only the same-scale feature maps from down-sampling and up-sampling paths. The motivation behind the U-Net is to bridge the semantic gap between encoder and decoder prior to concatenation.

Adapted the U-Net architecture, the U-Net++[56] replaces the direct skip connections in U-Net with the nested dense skip connections and generates full resolution feature maps at multiple semantic levels. The UNet++ consists of an encoder and decoder connected through a series of nested, dense skip pathways. The main idea behind UNet++ is the re-designed skip pathways that narrow the semantic gap between the feature maps of the encoder and decoder subnetworks. As a result, the U-Net++ can effectively capture fine-grained details of the foreground objects when high-resolution feature maps from the encoder and decoder networks are semantically similar. In contrast, U-Net uses direct skip connections, in which high-resolution feature maps are directly passed from the encoder to the decoder, resulting in the fusion of semantically dissimilar feature maps. This architecture-wise change in U-Net++ yields significant performance gain over the vanilla U-Net.

DAF3D[55] is a 3D feature pyramid network equipped with attention modules to generate the deep attentive feature (DAF) for medical image segmentation, such as prostate segmentation on transrectal ultrasound (TRUS) images. The DAF3D architecture lands in the completely different domain from the U-Net derived architectures. DAF3D is a bit complicated end-to-end system composed of three modules. The first module utilizes the 3D feature pyramid network (3D-FPN) architecture to combine multi-level features via a top-down pathway with deep supervision to extract more representative features. The 3D FPN produces multi-level single-layer features (SLF) and fused multi-layer features (MLF). The MLF encodes the low-level detail information as well as the high-level semantic context. It was also inevitably incorporating noise due to the coarse features at deep layers. The 2nd attention module leverages the MLF and the SLF as inputs. It produces the refined, attentive feature maps for each layer by adding attention gates to disambiguate irrelevant and noises responses in the background. The 3rd module employs a 3D atrous spatial pyramid pooling (ASPP) to resample attentive features at different scale levels for more accurate segmentation. Overall, the DAF3D intends to utilize the 3D attention mechanism to constrain the specific segmentation region, filter out the noise from the background, and enhance the boundary identification. The key idea is to select the useful complementary information from the multi-level features to refine features at each individual layer.

The proposed two phases cascaded model inspired from both U^2-Net and SASSNet[42] to targeting and refining the muscle thigh in a coarse-to-fine manner. As we explained in Section 2.2.c, the U^2-Net is a novel and simple network architecture with a two-level nested U-Shaped structure. In this architecture each stage of encoder-decoder U-Net structure contains a the newly proposed Residual U-block (RSU), which is, in fact, a down-sampling and up-sampling encoder-decoder itself. The idea behind the RSU block is to use multi-scale features as residuals instead of the original features. The nested U-structure will effectively maintain the fine-grained details and force the network to extract features from multiple-scales at every residual block and aggregate multi-level features across levels. SASSNet

based on 3D V-Net[42] imposes the shape constraint with a multi-task deep networks that jointly predict the 3D distance maps and 3D binary masks. A 3D discriminator (3D GAN) predicts the adversarial loss between predicted and unlabeled 3D distance maps to capture shape-aware features more effectively. In the cascaded model, U^2-Net can functionally discriminate better features from the nested U-structure and RSU blocks. However, we only utilize U^2-Net as the coarser segmentation step to mining the synonymous attention mechanism to make bold part identification for the thigh muscle. The SASSNet is motivated to enhance the finer-grained shape boundary identification into the cropped 3D region as the refinement step.

The baseline architectures, U-Net, U-Net++, DAF3D, and U^2-Net, all intended to capture the fine-grained feature maps via different deep learning pathways mechanisms. Systematically, the U^2-Net alone should yield comparable, even more, advanced segmentation performance than the baseline U-Net and U-Net++. Since the nested U-structure with RSU blocks is more complicated than the nested with dense skip connections in U-Net++. Practically, the proposed cascaded model outperforms the baselines as an end-to-end system, shown in Figure A1 (Appendix 1). One primary obstacle of this study is the specific problem setting of the thigh muscle segmentation. We are only given a single side (left side) ground truth muscle labels with the full-scale images. The anatomical symmetry issue in the full-scale images raises the difficulty in this specific segmentation task. As demonstrated in Figure A1 (Appendix 1), the baseline U-Net illustrates the unstable performance with noise. Some cases are entirely missing interpretation. DAF3D can perceptually reduce the background noise; however, over-segmented the thigh muscle regions failed to refine the boundary. U-Net++ with data augmentation (flipping, rotation, scale, and transform) substantially downgrades the performance due to the symmetric issue. U-Net++ (without augmentation) yields perceptually stable performance better than the rest baselines. However, some cases are entirely missing the segmentation, consequently, lower the Dice score. Overall, the proposed cascaded model obtains the stable Dice score performance by surpassing the baseline U-Net++ (with augmentation) from 0.5% to 5%.

Appendix 2: Statistical Shape Modeling

Statistical shape models play an important role in MR image segmentation. Toth et al.[58] extended the traditional active appearance model with principal component analysis (PCA) to include intensity and gradient information. They used level-set methods to capture the shape statistical model information with a multi-feature landmark-free framework. Aswani et al.[59] proposed a dual path U-Net based autoencoder with singular value decomposition (SVD) as the geometric constraint in latent space optimization for brain tumor segmentation from MR images. In addition to the shape-based models, the shape-prior has also emerged as an effective method of image segmentation in many medical imaging segmentation contexts. Shigwan et al.[60] proposed to couple deep neural networks with pointset-based shape prior that can be learned effectively despite the training sets having small size and imperfections in expert segmentation. The shape prior relies on sparse Riemannian modeling in Kendall shape space within the proposed Bayesian inference framework, performs optimal alignment of the shape model to object boundary in the image. The proposed method showed improved segmentation performance of the thalamus and the caudate on MR images. Kruger et

al.[61] proposed a probabilistic approach for statistical appearance models in a maximum a-posteriori framework to segment the hand shape with 2D hand X-rays. Zheng et al.[62] introduced a semi-supervised adversarial learning model with the Deep Atlas Prior (DAP) to segment the liver on CT images. The proposed model encodes a probabilistic shape prior to its loss design. He et al.[63] proposed an auto-encoder that embeds prior anatomical features on the unlabeled dataset to segment the renal artery on abdominal CT images. However, most shape priors typically assumes properly aligned input images, which is difficult to achieve in practice for objects with large variation in shape. Traditional shape-based analysis models, such as PCA and SVD, require a fair amount of computation overhead. For example, solving the SVD of a matrix needs a time complexity of $O(n^3)$ and the processing time increasing dramatically when both number images and image size are large. Automated muscle segmentation on MR images is challenging due to the noise and low contrasts between different anatomic structures and the large variability of muscle shape[64]. We do not explicitly apply the shape prior with statistical aligned mean shape models to thigh muscle segmentation in the proposed work. However, the signed distance map (SDM) in the second phase refinement step, assigns values to points according to their distance to the boundary curve, is a decent representation of the shape property. Furthermore, the SDM combined with binary masks avoids the calculation of SVD or PCA, while the second phase adversarial network's ability to discriminate muscle regions on the shape prior among different SDMs is still maintained.

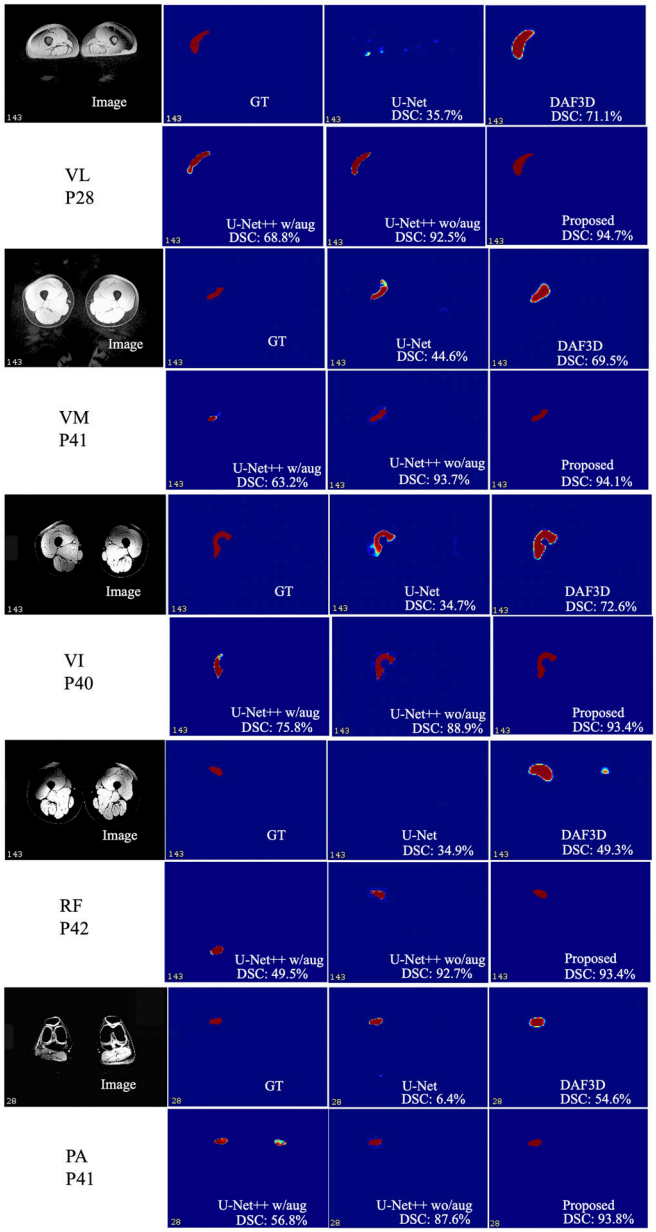

Figure A1:

Comparison to previous architectures. The Dice score (DSC) indicates the prediction of each patient image, not just for each slice. Abbreviations: **VL, VM, VI**: vastus lateralis, medialis, intermedius; **RF**: rectus femoris; **PA**: patella; **P#:** participant number

References:

1. Chen Y, Moiseev D, Kong WY, et al. Automation of Quantifying Axonal Loss in Patients with Peripheral Neuropathies through Deep Learning Derived Muscle Fat Fraction. Journal of magnetic resonance imaging : JMRI. 2021. doi: 10.1002/jmri.27508.
2. Hiasa Y, Otake Y, Takao M, et al. Automated Muscle Segmentation from Clinical CT Using Bayesian U-Net for Personalized Musculoskeletal Modeling. IEEE transactions on medical imaging. 2020;39(4):1030–1040. [PubMed: 31514128]

3. Ni R, Meyer CH, Blemker SS, et al. Automatic segmentation of all lower limb muscles from high-resolution magnetic resonance imaging using a cascaded three-dimensional deep convolutional neural network. Journal of medical imaging (Bellingham, Wash). 2019;6(4):044009.
4. Yokota F, Otake Y, Takao M, et al. Automated muscle segmentation from CT images of the hip and thigh using a hierarchical multi-atlas method. International journal of computer assisted radiology and surgery. 2018;13(7):977–986. [PubMed: 29626280]
5. Cheng R, Alexandridi NA, Smith RM, et al. Fully automated patellofemoral MRI segmentation using holistically nested networks: Implications for evaluating patellofemoral osteoarthritis, pain, injury, pathology, and adolescent development. Magnetic resonance in medicine. 2020;83(1):139–153. [PubMed: 31402520]
6. Conze PH, Brochard S, Burdin V, et al. Healthy versus pathological learning transferability in shoulder muscle MRI segmentation using deep convolutional encoder-decoders. Computerized medical imaging and graphics : the official journal of the Computerized Medical Imaging Society. 2020;83:101733. [PubMed: 32505943]
7. Tonson A, Ratel S, Le Fur Y, et al. Effect of maturation on the relationship between muscle size and force production. Medicine and science in sports and exercise. 2008;40(5):918–925. [PubMed: 18408605]
8. Weber DR, Leonard MB, Zemel BS. Body composition analysis in the pediatric population. Pediatr Endocrinol Rev. 2012;10(1):130–139. [PubMed: 23469390]
9. De Ste Croix MB, Armstrong N, Chia MY, et al. Changes in short-term power output in 10- to 12-year-olds. Journal of sports sciences. 2001;19(2):141–148. [PubMed: 11217012]
10. Sheehan FT, Brochard S, Behnam AJ, et al. Three-dimensional humeral morphologic alterations and atrophy associated with obstetrical brachial plexus palsy. J Shoulder Elbow Surg. 2014;23(5):708–719. [PubMed: 24291045]
11. Gillett JG, Lichtwark GA, Boyd RN, et al. Functional Anaerobic and Strength Training in Young Adults with Cerebral Palsy. Medicine and science in sports and exercise. 2018;50(8):1549–1557. [PubMed: 29557839]
12. Theis N, Brown MA, Wood P, et al. Leucine Supplementation Increases Muscle Strength and Volume, Reduces Inflammation, and Affects Wellbeing in Adults and Adolescents with Cerebral Palsy. The Journal of nutrition. 2021;151(1):59–64. [PubMed: 31965179]
13. Brochard S, Mozingo JD, Alter KE, et al. Three dimensionality of gleno-humeral deformities in obstetrical brachial plexus palsy. J Orthop Res. 2016;34(4):675–682. [PubMed: 26363273]
14. Zoabli G, Mathieu PA, Aubin CE. Back muscles biometry in adolescent idiopathic scoliosis. Spine J. 2007;7(3):338–344. [PubMed: 17482118]
15. Tomlinson OW, Barker AR, Fulford J, et al. Quantification of thigh muscle volume in children and adolescents using magnetic resonance imaging. European journal of sport science. 2020;20(9):1215–1224. [PubMed: 31928202]
16. Rinaldo N, Gualdi-Russo E, Zaccagni L. Influence of Size and Maturity on Injury in Young Elite Soccer Players. Int J Environ Res Public Health. 2021;18(6).
17. Andrews S, Hamarneh G. The Generalized Log-Ratio Transformation: Learning Shape and Adjacency Priors for Simultaneous Thigh Muscle Segmentation. IEEE transactions on medical imaging. 2015;34(9):1773–1787. [PubMed: 25700442]
18. Baudin PY, Azzabou N, Carlier PG, et al. Prior knowledge, random walks and human skeletal muscle segmentation. Medical image computing and computer-assisted intervention : MICCAI International Conference on Medical Image Computing and Computer-Assisted Intervention. 2012;15(Pt 1):569–576.
19. Gilles B, Magnenat-Thalmann N. Musculoskeletal MRI segmentation using multi-resolution simplex meshes with medial representations. Medical image analysis. 2010;14(3):291–302. [PubMed: 20303319]
20. Molaie M, Zoroofi RA. A Knowledge-Based Modality-Independent Technique for Concurrent Thigh Muscle Segmentation: Applicable to CT and MR Images. Journal of digital imaging. 2020. doi: 10.1007/s10278-020-00354-w.
21. Ogier A, Sdika M, Foure A, et al. Individual muscle segmentation in MR images: A 3D propagation through 2D non-linear registration approaches. Conference proceedings : Annual

International Conference of the IEEE Engineering in Medicine and Biology Society IEEE Engineering in Medicine and Biology Society Annual Conference. 2017;2017:317–320.

22. Prescott JW, Best TM, Swanson MS, et al. Anatomically anchored template-based level set segmentation: application to quadriceps muscles in MR images from the Osteoarthritis Initiative. Journal of digital imaging. 2011;24(1):28–43. [PubMed: 20049623]
23. Andrews S, Hamarneh G, Yazdanpanah A, et al. Probabilistic multi-shape segmentation of knee extensor and flexor muscles. Medical image computing and computer-assisted intervention : MICCAI International Conference on Medical Image Computing and Computer-Assisted Intervention. 2011;14(Pt 3):651–658.
24. Karlsson A, Rosander J, Romu T, et al. Automatic and quantitative assessment of regional muscle volume by multi-atlas segmentation using whole-body water-fat MRI. *Journal of magnetic resonance imaging* : JMRI. 2015;41(6):1558–1569. [PubMed: 25111561]
25. Le Troter A, Fouré A, Guye M, et al. Volume measurements of individual muscles in human quadriceps femoris using atlas-based segmentation approaches. Magma (New York, NY). 2016;29(2):245–257.
26. Ogier AC, Heskamp L, Michel CP, et al. A novel segmentation framework dedicated to the follow-up of fat infiltration in individual muscles of patients with neuromuscular disorders. Magnetic resonance in medicine. 2020;83(5):1825–1836. [PubMed: 31677312]
27. Ding J, Cao P, Chang HC, et al. Deep learning-based thigh muscle segmentation for reproducible fat fraction quantification using fat-water decomposition MRI. Insights Imaging. 2020;11(1):128. [PubMed: 33252711]
28. Deniz CM, Xiang S, Hallyburton RS, et al. Segmentation of the Proximal Femur from MR Images using Deep Convolutional Neural Networks. Scientific reports. 2018;8(1):16485. [PubMed: 30405145]
29. Liu F, Zhou Z, Jang H, et al. Deep convolutional neural network and 3D deformable approach for tissue segmentation in musculoskeletal magnetic resonance imaging. Magnetic resonance in medicine. 2018;79(4):2379–2391. [PubMed: 28733975]
30. Barra V, Boire JY. Segmentation of fat and muscle from MR images of the thigh by a possibilistic clustering algorithm. Computer methods and programs in biomedicine. 2002;68(3):185–193. [PubMed: 12074845]
31. Brunner G, Nambi V, Yang E, et al. Automatic quantification of muscle volumes in magnetic resonance imaging scans of the lower extremities. Magnetic resonance imaging. 2011;29(8):1065–1075. [PubMed: 21855242]
32. Mandić M, Rullman E, Widholm P, et al. Automated assessment of regional muscle volume and hypertrophy using MRI. Scientific reports. 2020;10(1):2239. [PubMed: 32042024]
33. Mesbah S, Shalaby AM, Stills S, et al. Novel stochastic framework for automatic segmentation of human thigh MRI volumes and its applications in spinal cord injured individuals. PloS one. 2019;14(5):e0216487. [PubMed: 31071158]
34. Middleton MS, Haufe W, Hooker J, et al. Quantifying Abdominal Adipose Tissue and Thigh Muscle Volume and Hepatic Proton Density Fat Fraction: Repeatability and Accuracy of an MR Imaging-based, Semiautomated Analysis Method. Radiology. 2017;283(2):438–449. [PubMed: 28278002]
35. Schlaeger S, Freitag F, Klupp E, et al. Thigh muscle segmentation of chemical shift encoding-based water-fat magnetic resonance images: The reference database MyoSegmenTUM. PloS one. 2018;13(6):e0198200. [PubMed: 29879128]
36. Ji S, Yang M, Yu K. 3D convolutional neural networks for human action recognition. IEEE Trans Pattern Anal Mach Intell. 2013;35(1):221–231. [PubMed: 22392705]
37. Cheng R, Lay N, Roth HR, et al. Fully automated prostate whole gland and central gland segmentation on MRI using holistically nested networks with short connections. Journal of medical imaging (Bellingham, Wash). 2019;6(2):024007.
38. Maffulli N, Longo UG, Spiezia F, et al. Aetiology and prevention of injuries in elite young athletes. Medicine and sport science. 2011;56:187–200. [PubMed: 21178374]

39. Fick CN, Grant C, Sheehan FT. Patellofemoral Pain in Adolescents: Understanding Patellofemoral Morphology and Its Relationship to Maltracking. Am J Sports Med. 2020;48(2):341–350. [PubMed: 31834811]
40. Shen A, Boden BP, Grant C, et al. Adolescents and adults with patellofemoral pain exhibit distinct patellar maltracking patterns. Clinical Biomechanics. 2021;90:105481. [PubMed: 34562716]
41. Qin X, Zhang Z, Huang C, et al. U2-Net: Going deeper with nested U-structure for salient object detection. Pattern Recognition. 2020;106.
42. Li S, Zhang C, He X. Shape-Aware Semi-supervised 3D Semantic Segmentation for Medical Images. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. doi: 10.1007/978-3-030-59710-8_542020:552-561.
43. McAuliffe MJ, Lalonde FM, McGarry D, et al. Medical Image Processing, Analysis and Visualization in clinical research. Proceedings 14th IEEE Symposium on Computer-Based Medical Systems. CBMS 2001; 2001.
44. Farag A, Le L, Roth HR, et al. A Bottom-Up Approach for Pancreas Segmentation Using Cascaded Superpixels and (Deep) Image Patch Labeling. IEEE Trans Image Process. 2017;26(1):386–399. [PubMed: 27831881]
45. Alom MZ, Yakopcic C, Hasan M, et al. Recurrent residual U-Net for medical image segmentation. Journal of medical imaging (Bellingham, Wash). 2019;6(1):014006.
46. Xie S, Tu Z. Holistically-Nested Edge Detection. 2015 IEEE International Conference on Computer Vision (ICCV); 2015.
47. Ho Y, Wookey S. The Real-World-Weight Cross-Entropy Loss Function: Modeling the Costs of Mislabeling. IEEE Access. 2020;8:4806–4813.
48. Liu L, Wolterink JM, Brune C, et al. Anatomy-aided deep learning for medical image segmentation: a review. Physics in medicine and biology. 2021;66(11).
49. Chen F, Yu H, Hu R, et al. Deep Learning Shape Priors for Object Segmentation. 2013 IEEE Conference on Computer Vision and Pattern Recognition; 2013.
50. Aggarwal A, Mittal M, Battineni G. Generative adversarial network: An overview of theory and applications. International Journal of Information Management Data Insights. 2021;1(1).
51. Baur C, Milletari F, Belagiannis V, et al. Automatic 3D reconstruction of electrophysiology catheters from two-view monoplane C-arm image sequences. International journal of computer assisted radiology and surgery. 2016;11(7):1319–1328. [PubMed: 26615429]
52. Gong T, Lee T, Stephenson C, et al. A Comparison of Loss Weighting Strategies for Multi task Learning in Deep Neural Networks. IEEE Access. 2019;7:141627–141632.
53. Paszke A, Gross S, Massa F, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. Adv Neur In. 2019;32.
54. Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. BMC Med Imaging. 2015;15:29. [PubMed: 26263899]
55. Wang Y, Dou H, Hu X, et al. Deep Attentive Features for Prostate Segmentation in 3D Transrectal Ultrasound. IEEE transactions on medical imaging. 2019;38(12):2768–2778. [PubMed: 31021793]
56. Zhou Z, Siddiquee MMR, Tajbakhsh N, et al. UNet++: A Nested U-Net Architecture for Medical Image Segmentation. Deep Learn Med Image Anal Multimodal Learn Clin Decis Support *(2018)*. 2018;11045:3–11.
57. Rezaei M, Nappi JJ, Lippert C, et al. Generative multi-adversarial network for striking the right balance in abdominal image segmentation. International journal of computer assisted radiology and surgery. 2020;15(11):1847–1858. [PubMed: 32897490]
58. Toth R, Madabhushi A. Multifeature landmark-free active appearance models: application to prostate MRI segmentation. IEEE transactions on medical imaging. 2012;31(8):1638–1650. [PubMed: 22665505]
59. Aswani K, Menaka D. A dual autoencoder and singular value decomposition based feature optimization for the segmentation of brain tumor from MRI images. BMC Med Imaging. 2021;21(1):82. [PubMed: 33985449]
60. Shigwan SJ, Gaikwad AV, Awate SP. Object Segmentation with Deep Neural Nets Coupled with a Shape Prior, When Learning From a Training set of Limited Quality and Small Size. Paper

presented at: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI); 3–7 April 2020, 2020.

61. Kruger J, Ehrhardt J, Handels H. Statistical appearance models based on probabilistic correspondences. Medical image analysis. 2017;37:146–159. [PubMed: 28219833]
62. Zheng H, Lin L, Hu H, et al. Semi-supervised Segmentation of Liver Using Adversarial Learning with Deep Atlas Prior. Paper presented at: Medical Image Computing and Computer Assisted Intervention – MICCAI 2019; 2019//, 2019; Cham.
63. He Y, Yang G, Chen Y, et al. DPA-DenseBiasNet: Semi-supervised 3D Fine Renal Artery Segmentation with Dense Biased Network and Deep Priori Anatomy. Paper presented at: Medical Image Computing and Computer Assisted Intervention – MICCAI 2019; 2019//, 2019; Cham.
64. Barnouin Y, Butler-Browne G, Voit T, et al. Manual segmentation of individual muscles of the quadriceps femoris using MRI: a reappraisal. Journal of magnetic resonance imaging : JMRI. 2014;40(1):239–247. [PubMed: 24615897]

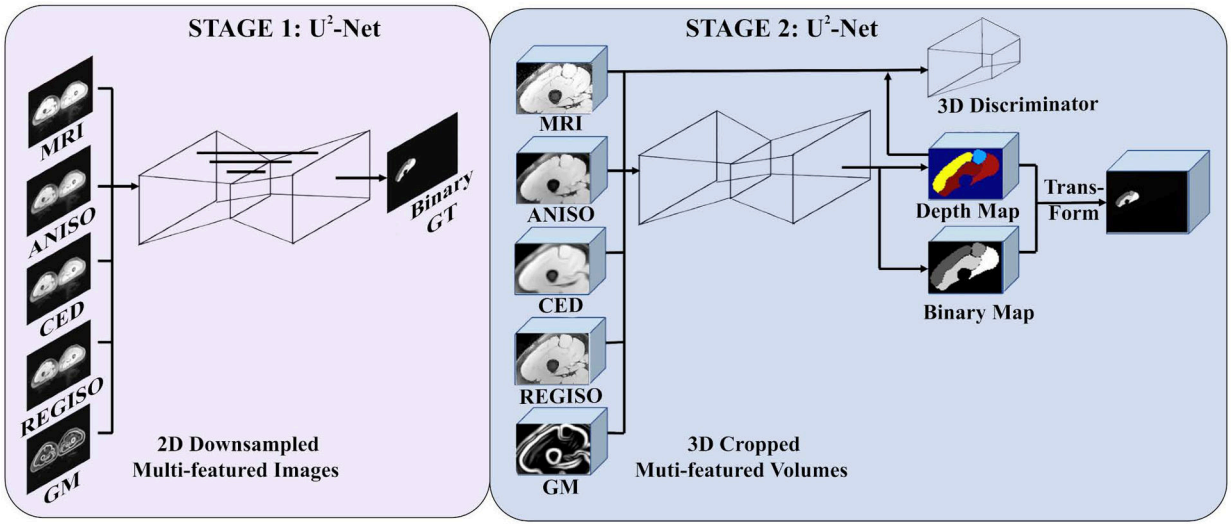

Figure 1.

Feed-forward path of the segmentation pipeline. The input of the two-stage segmentation pipeline is the down-sampled multi-feature slices of the full-scale images and paired binary ground truth labels (not shown). These multi-featured image sets are produced by filtering the down-sampled MR images using coherence enhanced diffusion (CED), anisotropic diffusion (ANISO), regularized anisotropic diffusion (REGISO), and gradient magnitude (GM)[43]. In the 1st stage, the U^2-Net architecture generates the coarse level predicted probability map. A 3D morphology applies to the coarse predictions to remove noise and extract the tightened bounding box. 3D cropping is applied to the high-resolution images to generate the multi-featured 3D image volumes centered on the region of interest. In the 2nd stage, multi-featured image sets (CED, ANISO, REGISO, GM) are created using with the cropped high-resolution images or the cropped down-sampled (2x, 4x) images. In stage 2, a 3D V-Net and 3D GAN based SASSNet refines the boundary. The predicted binary mask volume transforms back to the original image space as the final result.

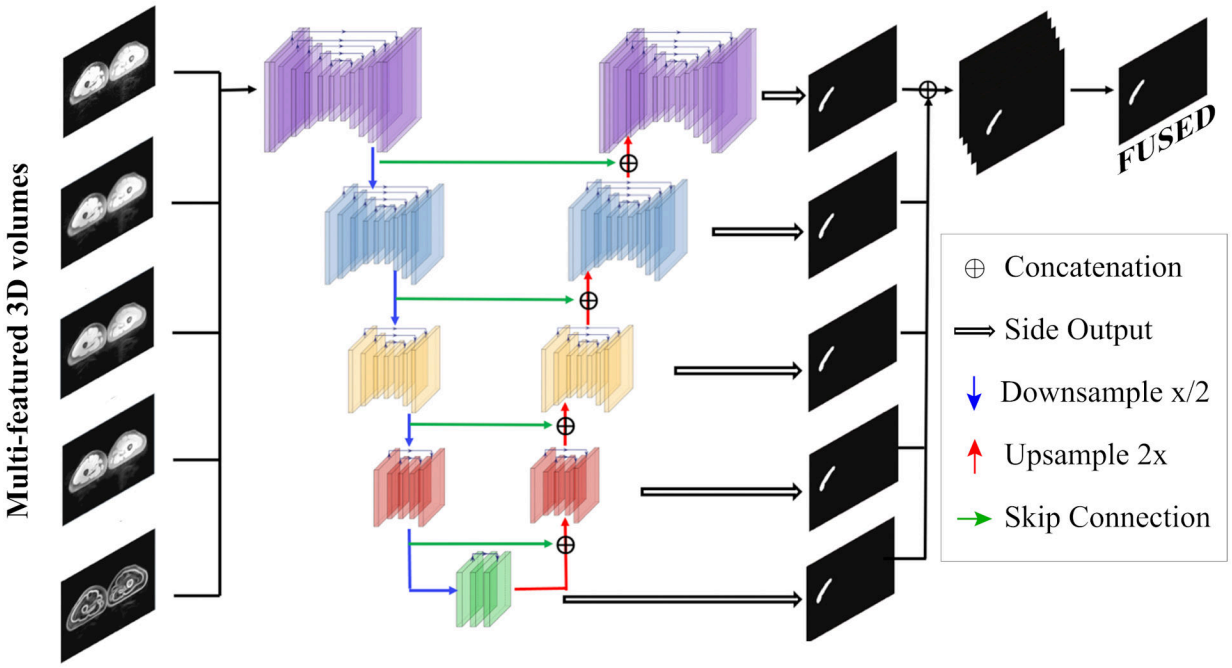

Figure 2.

Feed-forward Path of the U^2-Net Training Process. The U^2-Net conducts the coarse level segmentation on the full scaled images. The input of the U^2-Net architecture is the full scaled multi-featured 2D image slices (Figure 1) with corresponding ground truth labels (not shown). The latter is used to supervise the training process. The leading architecture is the RSU block (colored blocks), a U-Net like encoder-decoder. The RSU block colors reflect the different scale level feature maps. Within each block, the residual connections (the thinner black arrow links on top of each colored block) enable focus on local details while the overall residual U-Net architecture (inside RSU block) enable fusing these local details with global (multiscale) contextual information to improve performance. On the encoder path, the convolution generated feature maps are down-sample by a factor of 2 (blue arrow) between each multi-scale level. Along the decoder path, the convolution generated feature maps are up-sampled by a factor of 2 (red arrow) between each multi-scale level. The skip connection (green arrow) propagates the feature map from the encoder to the decoder at each multiscale level to enforce the local contextual information passing. Each multiscale level also side outputs (white arrow) the predicted feature map. These feature maps are up-sampled, concatenated, and fused to the original image size, constituting the final segmentation prediction map, and compared with the ground truth label. The predicted target object probability map is used to extract the tightened bounding box (centered on the predicted binary map) on the high-resolution images. By doing so, it allows the algorithm to prune the unrelated background volumes. The predicted output binary mask in this diagram is the vastus lateralis.

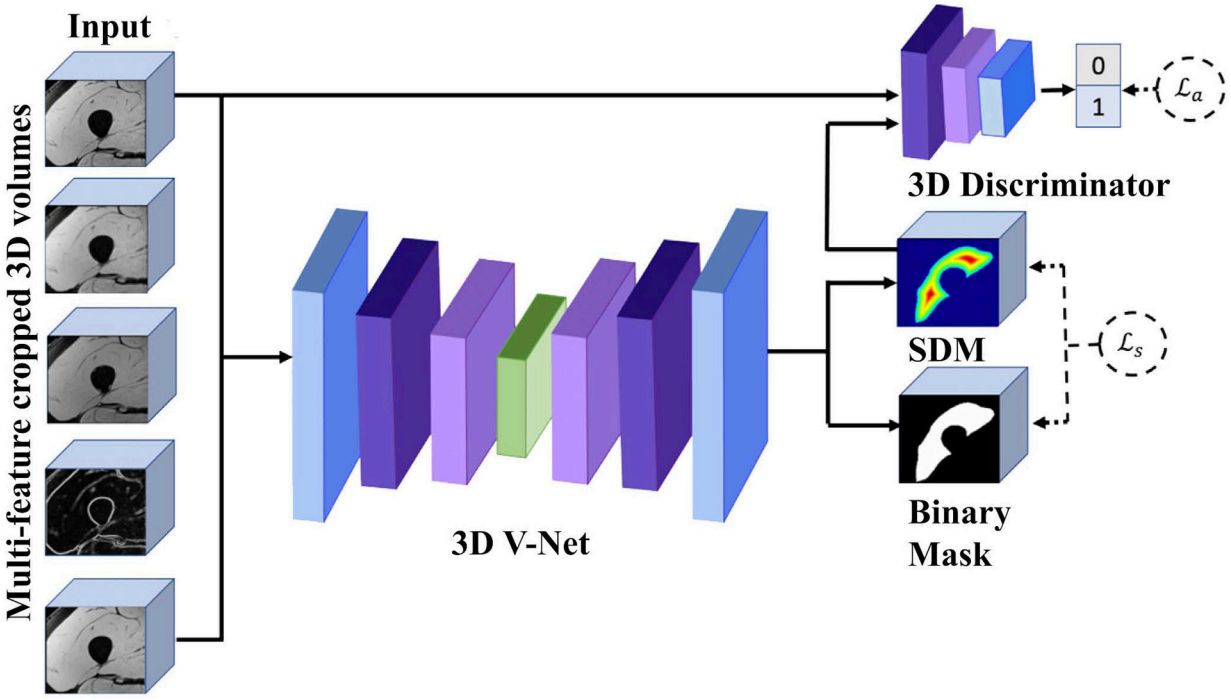

Figure 3.

SASSNet architecture. The input is the multi-features cropped 3D volumes (Figure 1) with corresponding binary ground truth map (not shown) and signed distance map (SDM). The 3D V-Net and 3D Discriminator together emulate the 3D cycle GAN alike structure to generate the synthetic images and sign distance map from unlabeled image datasets in a semi-supervised manner. The diagram only illustrates the feedforward path of the training process. **Abbreviations**: $\mathscr{L}_a$: adversarial loss, $\mathscr{L}_s$: supervised loss. This figure is modified from the first figure of Li et al[42].

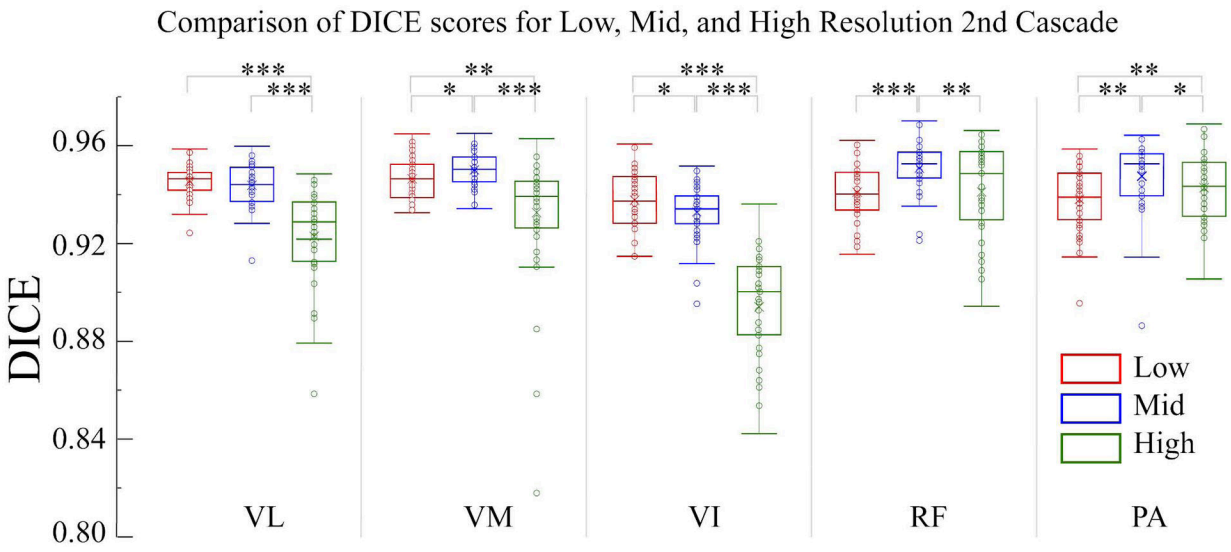

Figure 4:

Box and Whiskers comparison of Dice scores. Box and Whiskers comparison of Dice scores for the 3 resolution cases (red: low-resolution, blue: mid-resolution, green: high-resolution in the 2nd stage cascade). Abbreviations: **VL, VM, VI**: vastus lateralis, medialis, intermedius; **RF**: rectus femoris; **PA**: patella; * $p<0.05$; ** $p<0.01$; ***$p<0.001$.

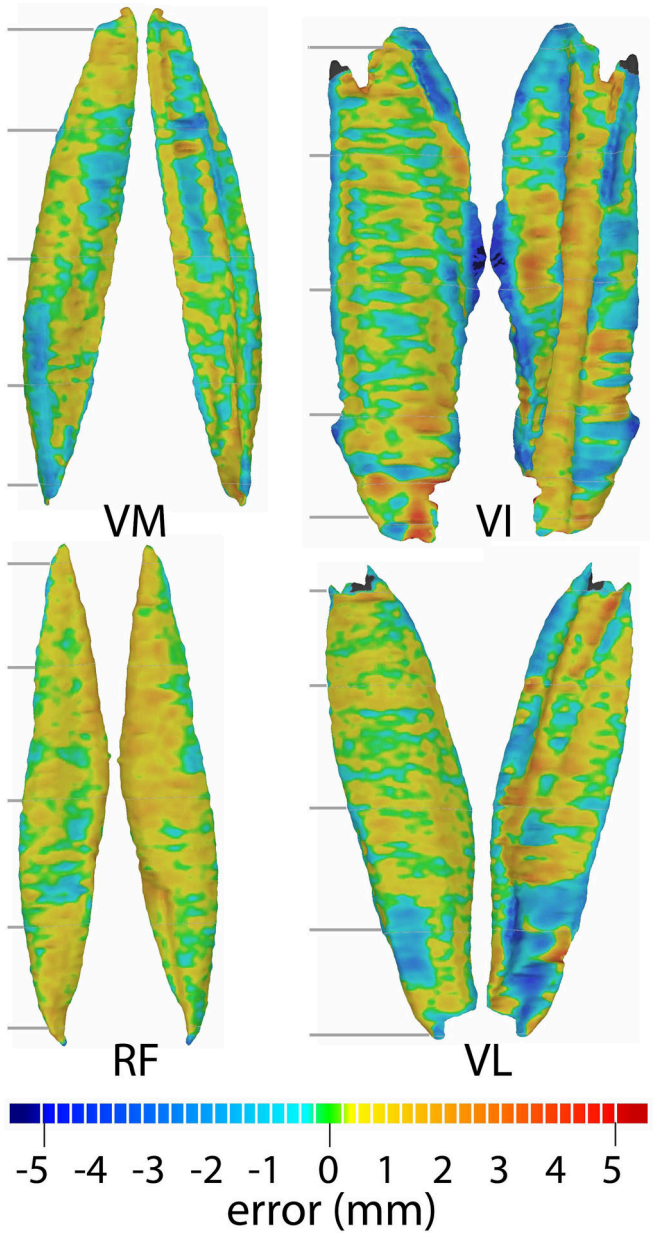

Figure 5.

Visual evaluation of differences between the auto- and manual segmentation. For each muscle, the model having an error that matched the average dice similarity score (Dice, Table 2) was selected for display [P02: vastus medialis (VM)=95.2%; P27: vastus intermedialis (VI)=93.3%, P02: rectus femoris (RF)=94.4%, and P12: vasti lateralis (VL)=94.4%]. The maximum absolute errors for VM, VI, RF, and VL models displayed were 2.0, 7.7, 1.7, 4.5 mm. Based these errors, the colormap scale (bottom), displaying the distance from the automatically generate surface to the ground truth model, was set from −5 (deep blue) to 5mm (deep red). The errors between surfaces typically remained within the mid-resolution pixel size (±0.88mm, aqua-blue to orange-yellow). Two viewpoints are provided anterior to posterior and posterior to anterior. To demonstrate the curvature around

the bone, a slightly oblique view was used in some cases. The grey lines indicate the imaging planes used to visualize the accuracy in 2D (Figure 7).

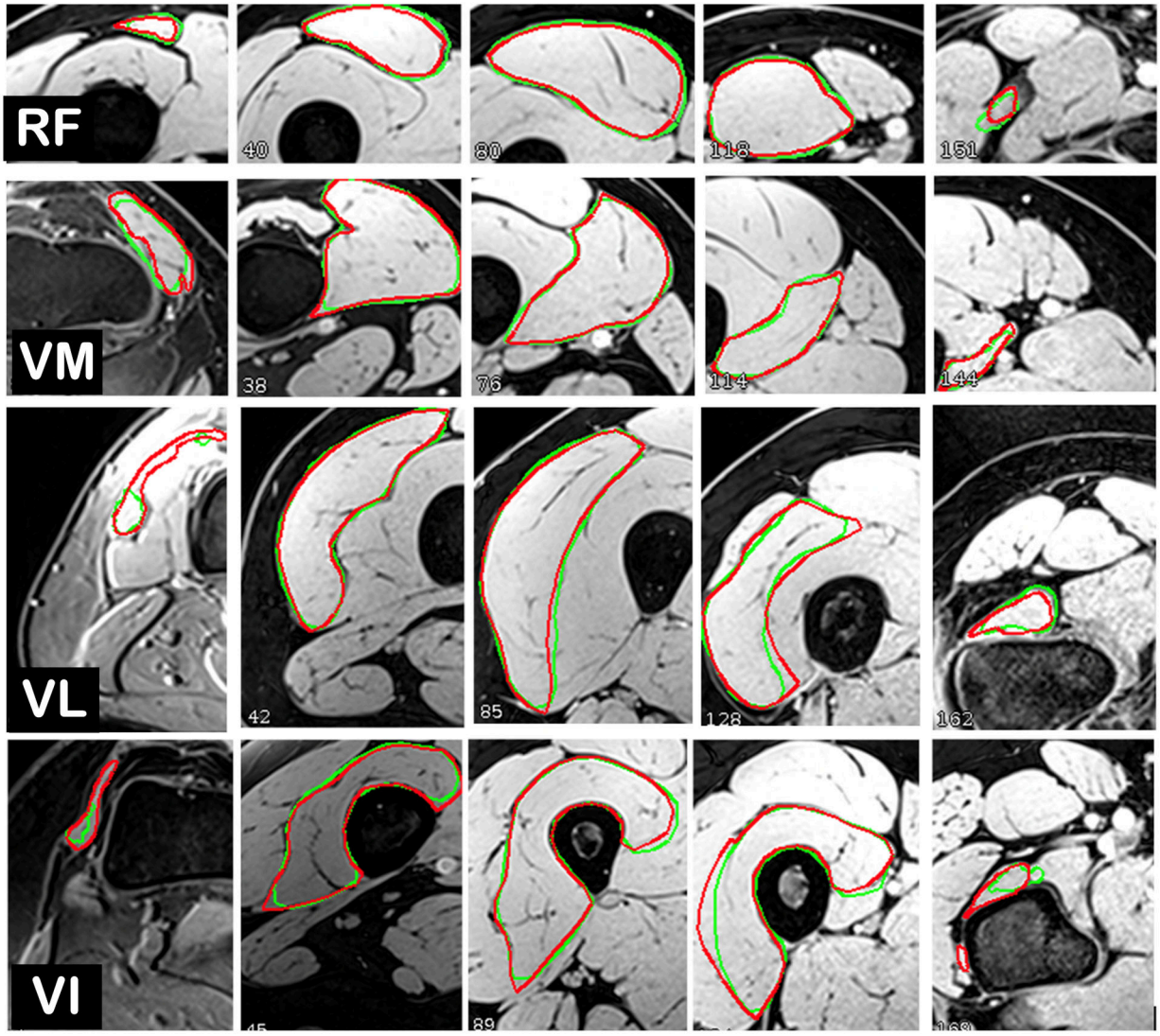

Figure 6.

2D Comparison between Segmentation and Ground Truth. The participant data are selected as previously (Figure 5). To prevent image selection bias, the image slices are selected at the 5%, 25%, 50%, 75%, and 95% of the whole muscle (indicated by grey slice numbers in Figure 6). The manual outline (ground-truth) is in red, with the auto-segmentation in green. Abbreviations: **VL, VM, VI**: vastus lateralis, medialis, intermedius; **RF**: rectus femoris.

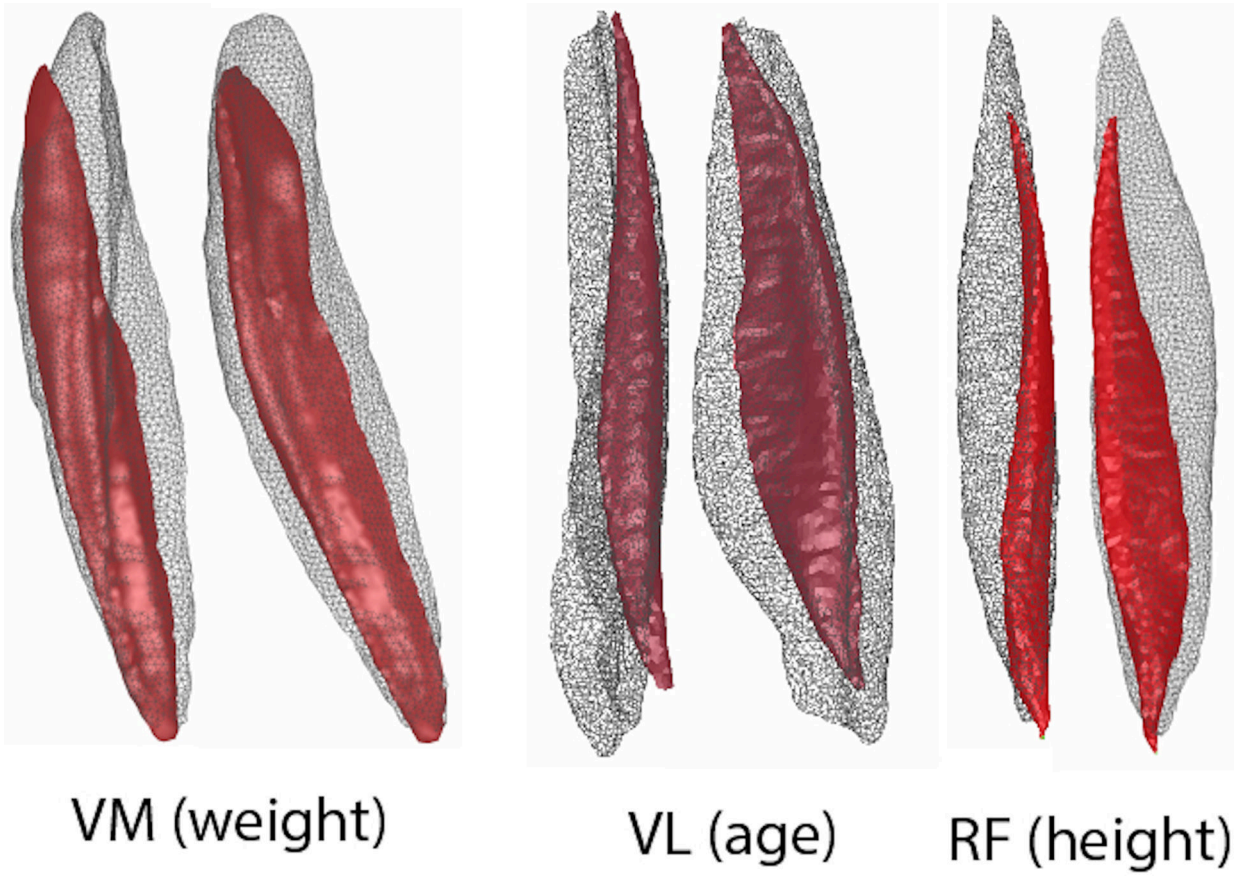

Figure 7.

Qualitative muscle shape variability. To qualitatively demonstrate the variability is shape and size of the muscles across our adolescent study population, the vastus medialis (VM) was modeled for the youngest (12 years) and oldest (18 years) participants, the vastus lateralis (VL) was modeled for the shortest (147 cm) and tallest (182cm) participants, and the rectus femoris (RF) was modeled for the lightest (42kg) and heaviest (94 kg). The muscle for each case was picked randomly to show the variation across the muscles. For each pair, the models for the oldest, tallest, or heaviest participants were modeled using a mesh, so that the models for the youngest, shortest, and lightest participants could be seen through it. Each pair is displayed with the same resolution, but the resolution is not consistent across pairs. To remove variation due to the relative positioning of the participant in the scanner, the smaller model was manually best fit to the largest model and then the best fit algorithm, restricted to fine adjustments.

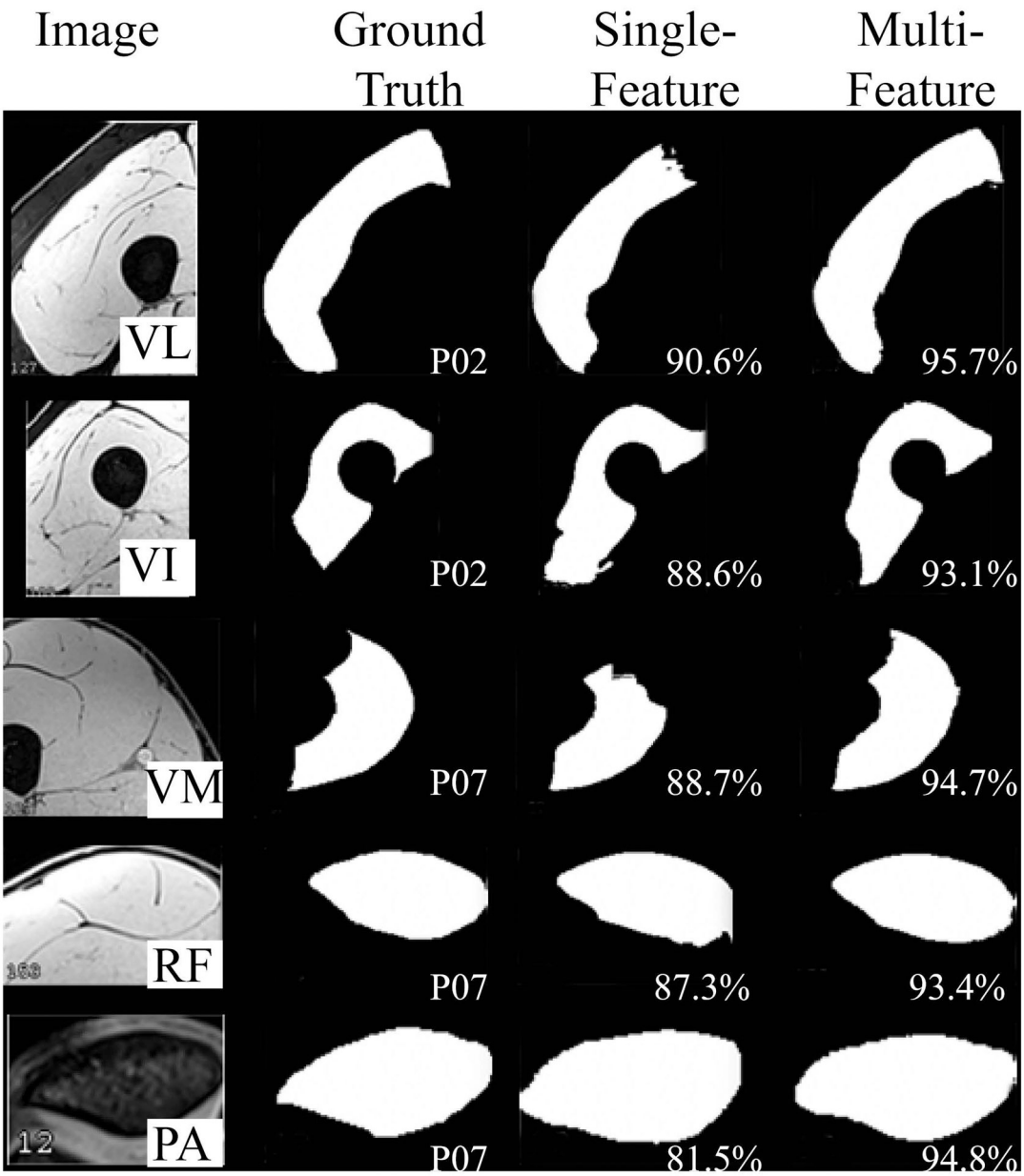

Figure 8:
Visual comparison of single-feature vs. multi-feature. The Dice score reflects the segmentation accuracy of each case in the 2nd stage testing phase.

Table 1:

Literature Summary

	Population					Images			Technique					Di	
	gender	**type**	**age**	**height**	**weight**	**image**	**resolution (mm)**	**#images**	**F/P**	**Algorithm**		**Train**	**Test**	**RF**	**VI**
Giles[19]	7M&6F	C	"young"	NR	NR	MRI	0.78×0.78×(2–10)	250	**full**	A	SSM	NR	NR		NI
Andrews[23]	20/20	C&P:COPD	64(8)/68(10)	NR	NR	MRI	0.78×0.78×5	~100	part	SA	Atlas	39	40	0.75	0.78
Prescott[22]	53M/50F	P:OA	61(10)	NR	NR	MRI	.98×.98×5	17	part	A	Temp	50	50	0.78	0.79
Baudin[18]	14 (NR)	C	NR	NR	NR	MRI	1×1×5	NR	NR	A	RW	13	14	0.92	0.81
Andrews[17]	same as Andrews 2011								part	SA	SSM	20	20		NI
Karlsson[24]	4M/7F	C	33–54	BMI = 26(4)		MRI	1.75×1.75×1.75	NR	**full**	A	Atlas	10	11		NI
LeTroter[25]	25 M	C	22(1)	178(6)	68(7)	MRI	0.38×0.38×12	20	part	A	Atlas	25	7	0.86	0.78
Ogier[21]	same as Le Trotter								part	SA	Prop	NA	25	0.90	0.93
Yokota[4]	20F	P:UHD	NR	NR	NR	CT	1.4×1.4×(1–10)	NR	**full**	A	Atlas	38	19	0.80 (ave of	
Hiasa[2]	same as Yokota								**full**	A	CNN	19	20	0.92	0.9
Molaie[20]	10M/12F	C	51(13)	NR	NR	CT	?×?×1.25	200–240	part	SA	FRFCM	NA	22	0.90	0.9
Ni[3]	64 (NR)	A	college	NR	NR	MRI	1×1×5	200–240	**full**	A	DCNN	51	13	0.96	0.88
Ogier[26]	12F	P:MD	46(12)	NR	NR	MRI	1×1×5	32	part	SA	Prop	NA	10	0.93	0.93
Chen[1]	40 (NR)	C&P:PND	NR	NR	NR	MRI	0.6×0.6×3.0	40	part	A	CNN	23	17	0.93	0.95
Ding[27]	41	NR	NR	NR	NR	MRI	0.74×.074×6	40	**full**	A	CNN	30	11		.094±(
Current	40	C&P:PFP	12–18	147–182	40–94	MRI	0.43 ×0.43× 2	288	**full**	A	Cascade	39	40	0.96	0.94

Literature Summary of Semi- and Fully Automatic Segmentation methodologies available for segmenting the individual muscles of the quadriceps (at a minimum). Age, height, and weight of participants is provided as an average and standard deviation [ave(sd)] or as a range. Listed earliest to latest publication. **Abbreviations**: male (M), female (F), C (control), not applicable (NA), P (patient), COPD (chronic obstructive pulmonary disease), unilateral hip disease (UHD), muscular dystrophy (MD), peripheral nerve disease (PND), patellofemoral pain (PFP), not reported (NR), magnetic resonance imaging (MRI), computer tomography (CT), segmentation completed on full or partial muscle (F/P), automatic-segmentation (A), semi-automatic segmentation (SA), template (temp), (RW), statistical shape matching (SSM), propagation (Prop), convolutional neural networks (CNN), fast and robust fuzzy C-means Clustering (FRCM), deep convolutional neural networks (DCNN), number of training datasets (train), number of testing datasets (test), vastus intermedialis, lateralis, & medialis (VI, VL, VM), and rectus femoris (RF).

Table 2.

Testing phase validation

	VL					VM				
	Dice	**IoU**	**HD**	**ASD**	**VSC**	**Dice**	**IoU**	**HD**	**ASD**	**VSC**
mean	94.4%	89.6%	12.1	0.10	99.2%	95.1%	90.7%	9.4	0.07	98.6%
std	0.9%	1.8%	4.2	0.05	0.9%	0.7%	1.2%	3.0	0.03	0.9%
min	91.3%	84.0%	6.5	0.04	96.8%	93.0%	87.0%	4.6	0.03	96.3%
max	96.0%	94.4%	22.9	0.32	100.0%	96.6%	93.4%	16.6	0.21	100.0%
median	94.4%	89.5%	11.2	0.08	99.4%	95.2%	90.8%	8.8	0.07	98.7%

	VI					RF				
	Dice	**IoU**	**HD**	**ASD**	**VSC**	**Dice**	**IoU**	**HD**	**ASD**	**VSC**
mean	93.2%	87.3%	14.3	0.12	98.3%	95.3%	91.0%	11.1	0.07	98.3%
std	1.2%	2.0%	4.8	0.06	1.1%	1.0%	1.8%	4.3	0.03	1.1%
min	89.8%	81.5%	6.6	0.07	94.6%	92.1%	85.4%	4.1	0.02	96.4%
max	95.1%	90.6%	29.1	0.32	100.0%	97.1%	94.4%	19.1	0.16	99.9%
median	93.4%	87.7%	13.9	0.11	98.3%	95.3%	91.0%	10.3	0.06	98.4%

	PAT				
	Dice	**IoU**	**HD**	**ASD**	**VSC**
mean	93.8%	88.3%	2.9	0.07	98.3%
std	1.5%	2.7%	1.1	0.02	1.2%
min	87.7%	78.0%	1.4	0.05	94.5%
max	95.4%	91.1%	5.5	0.16	100.0%
median	94.3%	89.3%	2.8	0.06	98.6%

Performance results of the proposed U^2-Net+SASSNet method. Abbreviations: **Dice** (%) dice similarity coefficient; **IoU** (%): Jaccard index; **HD** (mm): Hausdorff distance; **ASD** (mm): average minimum surface-to-surface distance; **VSC** (%): volumetric similarity coefficient; **VL, VM, VI**: vastus lateralis, medialis, intermedius; **RF**: rectus femoris; and **PA**: patella.

Table 3:

DICE-score comparison between single-feature and multi-feature in the second stage.

Mean Dice score	test	PA (%)	RF (%)	VL (%)	VM (%)	VI (%)
single-feature	leave-one-out	81.3±8.2	88.0±4.9	88.7±4.9	91.9±2.7	88.2±3.4
multi-feature	leave-one-out	93.8±1.5	95.3±1.0	94.4±0.9	95.1±0.7	93.2±1.2

Abbreviations: **PA**: patella, **RF**: rectus femoris, **VL, VM, VI**: vastus lateralis, medialis, intermedialis

Table 4:

Quantitative comparison to previous architectures

Dice score	PA(%)	RF(%)	VL(%)	VM(%)	VI(%)
U-Net[45]	5.9±1.8	26.4±15.3	33.9±14.8	41.5±14.2	33.4±11.5
DAF3D[55]	50.1±12.9	57.7±6.8	71.5±7.9	66.4±12.7	71.3±13.5
U-Net++ w/data aug[56]	53.6±4.0	49.5±4.2	72.7±7.8	58.1±7.9	72.9±4.1
U-Net++ w/o data aug[56]	88.6±7.8	93.1±1.0	90.4±4.1	93.4±2.7	88.4±1.6
current	93.8±1.5	95.2±1.0	94.1±0.9	95.1±0.7	93.2±1.2