JMI Journal of Medical Imaging

RESEARCH PAPER

Single slice thigh CT muscle group segmentation with domain adaptation and self-training

Qi Yang,[a,*] Xin Yu,[a] Ho Hin Lee,[a] Leon Y. Cai,[b] Kaiwen Xu,[a] Shunxing Bao,[c] Yuankai Huo,[a] Ann Zenobia Moore,[d] Sokratis Makrogiannis,[e] Luigi Ferrucci,[d] and Bennett A. Landman[c]

[a]Vanderbilt University, Department of Computer Science, Nashville, Tennessee, United States
[b]Vanderbilt University, Department of Biomedical Engineering, Nashville, Tennessee, United States
[c]Vanderbilt University, Department of Electrical and Computer Engineering, Nashville, Tennessee, United States
[d]National Institute on Aging, NIH, Translational Gerontology Branch, Baltimore, Maryland, United States
[e]Delaware State University, PEMACS Division, Dover, Delaware, United States

ABSTRACT. Purpose: Thigh muscle group segmentation is important for assessing muscle anatomy, metabolic disease, and aging. Many efforts have been put into quantifying muscle tissues with magnetic resonance (MR) imaging, including manual annotation of individual muscles. However, leveraging publicly available annotations in MR images to achieve muscle group segmentation on single-slice computed tomography (CT) thigh images is challenging.

Approach: We propose an unsupervised domain adaptation pipeline with self-training to transfer labels from three-dimensional MR to single CT slices. First, we transform the image appearance from MR to CT with CycleGAN and feed the synthesized CT images to a segmenter simultaneously. Single CT slices are divided into hard and easy cohorts based on the entropy of pseudo-labels predicted by the segmenter. After refining easy cohort pseudo-labels based on anatomical assumption, self-training with easy and hard splits is applied to fine-tune the segmenter.

Results: On 152 withheld single CT thigh images, the proposed pipeline achieved a mean Dice of 0.888 (0.041) across all muscle groups, including gracilis, hamstrings, quadriceps femoris, and sartorius muscle.

Conclusions: To our best knowledge, this is the first pipeline to achieve domain adaptation from MR to CT for thigh images. The proposed pipeline effectively and robustly extracts muscle groups on two-dimensional single-slice CT thigh images. The container is available for public use in GitHub repository available at: https://github.com/MASILab/DA_CT_muscle_seg.

Keywords: computed tomography; magnetic resonance; thigh muscle segmentation; single slice; domain adaptation; self-training

Paper 23042GR received Feb. 13, 2023; revised Jun. 9, 2023; accepted Jun. 20, 2023; published Jul. 12, 2023.

1 Introduction

Thigh muscle group segmentation is essential for assessing muscle anatomy, computing the muscle size/volume, and estimating muscle strength.[1] Quantitative thigh muscle assessment from segmentation can be a potential indicator of metabolic syndrome.[2] The loss of the thigh muscle

*Address all correspondence to Qi Yang, qi.yang@vanderbilt.edu

2329-4302/2023/$28.00 © 2023 SPIE

and associated functional capabilities are closely related to aging.[3] Accurate measurement of thigh muscle cross-sectional area, volumes, and mass can help researchers understand and study the effect of aging on the body composition of human body. Thus, extracting subject-specific muscle groups is an essential step.

Magnetic resonance (MR) imaging is the most common technique in previous muscle analyses, given its high contrast for soft tissue.[4] Many human efforts have been put into MR imaging for muscle analysis. Barnouin et al. optimized reproducible manual muscle segmentation.[5] Schlaeger et al. constructed a reference database (MyoSegmentTum) including the sartorius, hamstring, quadriceps femoris, and gracilis muscle groups for three-dimensional (3D) MR volume.[6] However, compared to MR imaging, the short acquisition time of computed tomography (CT) is better suited for routine clinical use.[4] In a longitudinal body composition study, single slice CT for each subject reduced unnecessary radiation.[7–9] Accurate segmentation of muscle groups on a single slice can aid in understanding thigh components and the effects of aging on muscle.[10]

Direct human manual annotation on the single slice CT is labor-intensive and challenging due to similar intensity among different muscle groups in CT. Leveraging publicly available annotation from existing MR resources (source domain) such as MyoSegmentTum for CT (target domain) is a promising direction to overcome the problem of muscle group segmentation. Methods handling domain shift or heterogeneity among modalities are called domain adaptation (DA).[11] DA aims to minimize differences among domains. In our case, DA has two challenging tasks that need to be addressed: (1) homogeneous intensity of different muscle groups of CT images and (2) inter-modality heterogeneity including contrast and anatomic appearance. The above two challenges are in Fig. 1. With the above challenges for thigh muscle segmentation problems, we propose a new DA pipeline to achieve CT thigh muscle segmentation. We build a segmenter trained with synthetic CT images in CycleGAN.[12] We infer segmentation maps from

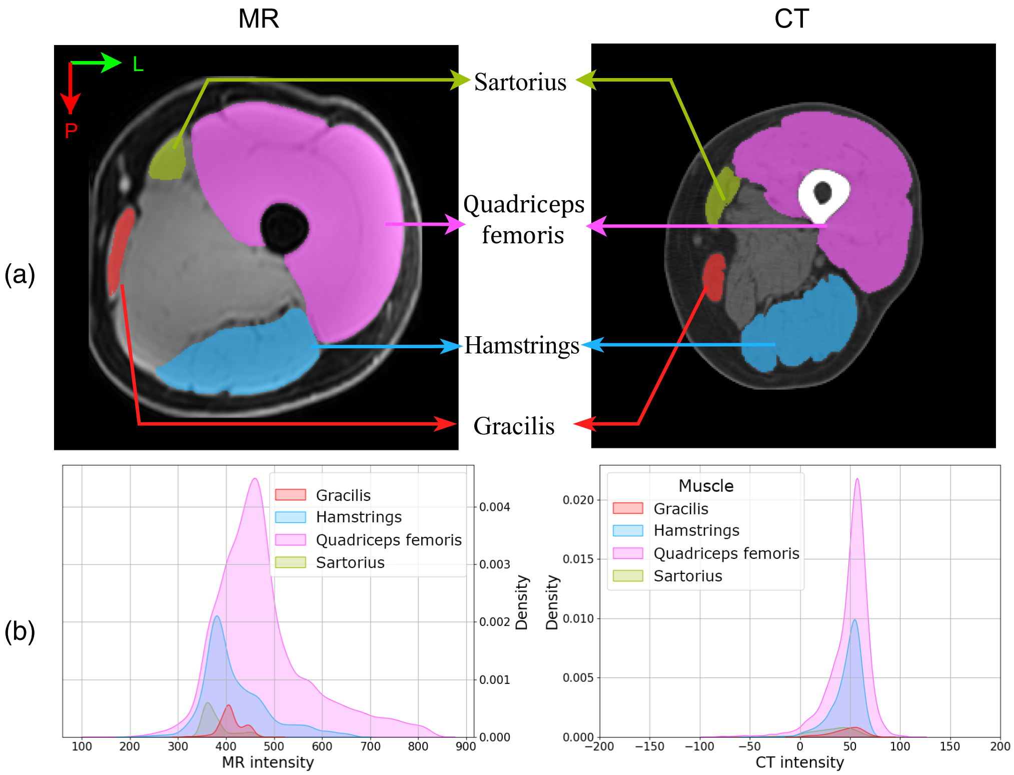


Fig. 1 A selective sample that highlights the inter-modality heterogeneity between MRI and CT and low-intensity difference among different muscle groups in CT. (a) The MR image is normalized by min-max. The original CT scale is clipped to $[-200, 500]$ and then normalized to [0,1]. (b) The intensity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

real CT images using the segmenter and divide the segmentation maps into two cohorts based on entropy. The entropy can work as an indicator for prediction map quality.[13] Based on the anatomic context, the whole muscle and bone masks of CT images are utilized to correct wrong predictions brought by domain shift. Self-training is applied on two cohorts to make the segmenter adapt to high entropy cohorts to enhance the robustness and preserve the segmentation performance on low entropy cohorts.

Our contributions can be summarized as the following:

- To our best knowledge, this is the first DA pipeline for thigh muscle group segmentation on CT thigh slices. The segmenter is trained with synthetic CT images learned from the unpaired MRI dataset to provide a coarse segmentation. Adversarial learning and anatomical processing of real CT images are combined to handle outliers and improve segmentation performance.
- We evaluate the pipeline with thigh CT single slices. The experiment shows that our proposed pipeline achieves state-of-the-art performance. The ablation study demonstrates the effectiveness of anatomical processing and self-training using real CT data.
- We release the source codes and models as a singularity[14] for researchers to use and apply to their existing data.

2 Related Work

2.1 Unsupervised DA

Unsupervised DA addresses the challenges of labeled data from the source domain and unlabeled data from the target domain for use during training. Unsupervised DA can reduce the labor required for annotation in the target domain, attracting more and more researchers, especially in medical imaging.[15] Dou et al. built an unsupervised DA framework for cardiac segmentation by only adapting low-level layers to reduce domain shift in the training stage.[16] CycleGAN[12] is known for translating image-to-image without needing pair samples, which has been applied extensively in DA. Huo et al. proposed SynSeg-Net to train CycleGAN and segmentor simultaneously to segment abdomen organs and brain image.[17] Zhou et al. extended the SynSeg-Net and combined contrastive learning generative model to preserve anatomical structure during image-to-image translation.[18] Based on the image-to-image translation, Chen et al. applied a synergistic method to adapt domain invariant features and image.[19] Disentanglement learning is used to learn two feature spaces: domain invariant structure and domain variant style. Yang et al. applied disentanglement learning to segment livers,[20] and Chang et al. embraced disentanglement learning to achieve semantic segmentation on natural images.[21]

2.2 Self-Training in DA

Self-training in DA generates pseudo-labels for the target domain, and the model is trained with pseudo-labels data to adapt to the target domain. However, directly using all pseudo-labels is risky due to the accumulation of errors and domain shift negatively impacting model performance. To overcome the challenge, Zou et al. proposed to apply a re-weighting class strategy to select high-quality pseudo-labels.[22] Zou et al. proposed to regularize the confidence of pixel pseudo-label to adapt to target domain.[23] Pan et al. proposed to reduce intra-domain gaps by dividing pseudo-labels into easy and hard splits based on entropy. The adversarial learning is applied on two splits to make the model adapt to the hard split without sacrificing the performance.[24] To further improve the quality of pseudo-labels, anatomical prior could be incorporated into self-training procedures to correct wrong predictions caused by domain shift.

3 Material and Method

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

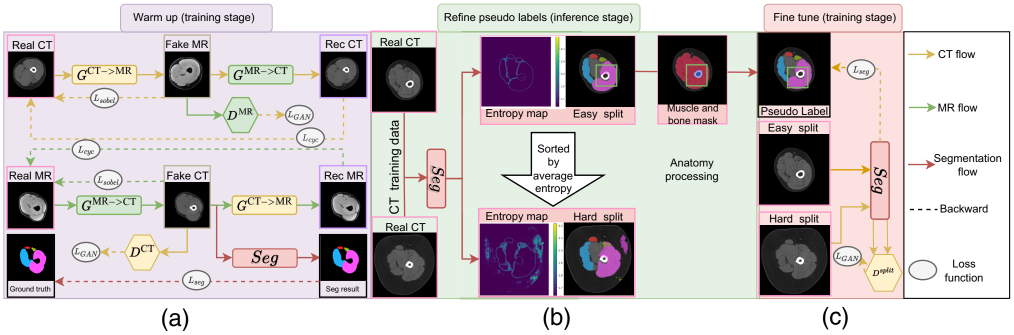


Fig. 2 Overview of proposed pipeline. (a) We adopt a CycleGAN design, including two generators and two discriminators for MR and CT, respectively. The segmentation module is trained by feeding synthetic CT images and corresponding MR ground truth. (b) The segmentation module from (a) is used to infer pseudo-labels divided into hard and easy cohorts based on entropy maps. Then, the easy cohort pseudo-labels are refined based on anatomy processing (muscle and bone bask). (c) Easy cohort pseudo-labels of CT images are used to fine-tune the segmentation module, and adversarial learning between easy and hard cohorts forces the segmentation module to adapt to hard cohorts simultaneously to increase segmentation module robustness.

3.1 Data and Preprocessing

We use two datasets in our study. One is the Baltimore Longitudinal Study of Aging (BLSA),[25] and the other one is MyoSegmenTUM.[6] The BLSA is a longitudinal dataset and collects 2D mid-thigh CT slices for each subject during the visit.[26] BLSA study protocols are approved by the National Institutes of Health Intramural Institutional Review Board, and all participants provided written informed consent. MyoSegmenTUM is a 3D MRI thigh dataset providing annotations for four muscle groups: the sartorius, hamstring, quadriceps femoris, and gracilis.

We used 1123 de-identified 2D low-dose single CT thigh slices of 763 participants from the BLSA. All data are de-identified under Institute Review Board approval. The slice has a size of 512×512 pixels. We split one single CT slice into left and right thigh images with size 256×256 pixels by following the pipeline in Ref. 27. During preprocessing steps, 11 images were discarded due to low quality or abnormal anatomic appearance. The CT images are the target domain in our case.

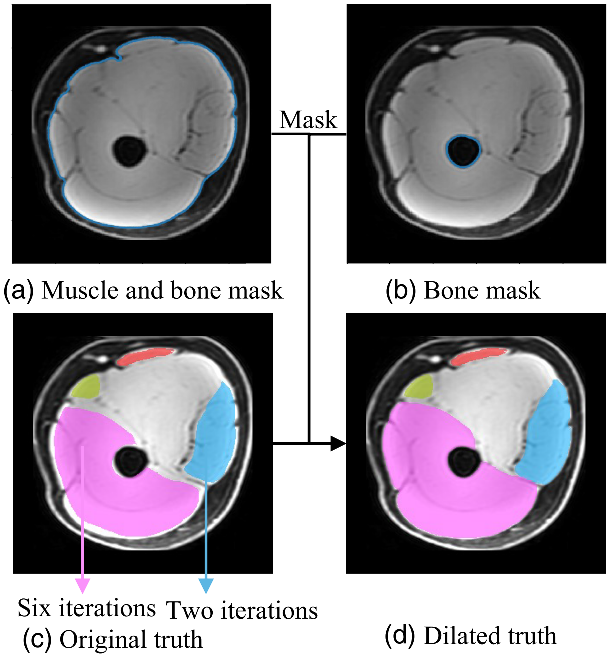
MyoSegmentTUM consists of waterfat MR images of 20 sessions of 15 healthy volunteers. The water protocol MR is selected as the source image. We select 1980 mid-thigh slices from MR volumes to reduce the anatomical gap between MR and CT slices at the mid-thigh position. The MR slices are divided into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

The original label of the MR slices is placed at each group with a margin of 2 mm to the outer boundary, as shown in Fig. 3(c). The incomplete ground truth makes the whole DA pipeline more challenging. To address this concern, we extract whole cross-sectional muscle and bone contour by using level set.[28] We use a binary 3×3 kernel to dilate the quadriceps femoris and hamstring muscle with six and two iterations, respectively. The complete muscle mask is obtained after performing the level set and dilation operation, as shown in Fig. 3(d).

We feed random pairs of CT and MR images to the proposed method. All 1980 MR images are fed into the training cohort. For CT, we divide all CT images into training, validation, and test cohorts based on participants. The training cohort includes 2044 CT thigh images from 669 participants. The validation cohort consists of 38 CT thigh images from 19 participants. The test cohort consists of 152 CT thigh images from 75 participants. Each CT image in the validation and test cohort has ground truth manually annotated from scratch to work for evaluation. The data distribution can be found in Table 1.

3.2 Train Segmentation Module from Scratch

Inspired by SynSeg-net,[17] we design a U-Net[29] segmentation module (*Seg*). We train the *Seg* with CycleGAN[12] in an end-to-end fashion as shown in Fig. 2(a). CycleGAN aims to solve the image-to-image translation problem unsupervised without requiring paired images. CycleGAN uses the idea of cycle consistency that we translate one image from one domain to the other and back



(a) Muscle and bone mask (b) Bone mask (c) Original truth (d) Dilated truth

Fig. 3 The preprocessing steps for dilating the ground truth of the MRI dataset. The blue contour in (a), (b) represents the muscle and bone boundaries extracted by level sets, and (c) represents the original ground truth. The quadriceps femoris muscle group is dilated in six iterations, and the hamstring muscle group is dilated in two iterations. (d) The final truth after preprocessing.

Table 1 Data distribution and image information for the whole pipeline.

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

again, and we should arrive at where we started.[12] Thus, we have two generators and discriminators in our framework. Generator $G^{X\to Y}$ represents the mapping function $X\colon \to Y$. Two generators $G^{MR\to CT}$ and $G^{CT\to MR}$ are utilized to synthesis fake CT $(G^{MR\to CT}(x_{MR}))$ and fake MR $(G^{CT\to MR}(x_{CT}))$ images, respectively. The discriminator D^{CT} and D^{MR} determine whether the input image (CT or MR) is synthetic or real. The adversarial loss is applied to generators and discriminators and is defined as

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

The above adversarial loss cannot guarantee that individual images are anatomically aligned to desired output since there are no constraints for the mapping function. Cycle loss[12] is introduced to reduce possible space for the mapping function by minimizing the difference between images and reconstructed images. The loss function is

$$L_{\mathrm{cyc}}^{\mathrm{CT}} = \|G^{\mathrm{MR}\to\mathrm{CT}}(G^{\mathrm{CT}\to\mathrm{MR}}(x_{\mathrm{CT}})) - x_{\mathrm{CT}}\|_1, \tag{2}$$

$$L_{\mathrm{cyc}}^{\mathrm{MR}} = \|G^{\mathrm{CT}\to\mathrm{MR}}(G^{\mathrm{MR}\to\mathrm{CT}}(x_{\mathrm{MR}})) - x_{\mathrm{MR}}\|_1. \tag{3}$$

To regularize the generator, we applied identity loss[12] to regularize generators. The identity loss is expressed as

$$L_{\text{Identity}} = \mathbb{E}[\|G^{\text{MR}\to\text{CT}}(x_{\text{MR}}) - x_{\text{MR}}\|_1] + \mathbb{E}[\|G^{\text{CT}\to\text{MR}}(x_{\text{CT}}) - x_{\text{CT}}\|_1]. \tag{4}$$

We further added an edge loss to preserve boundary information. Modified Sobel operator[30] is utilized to extract edge magnitude. The edge loss is calculated based on the difference in edge magnitude of two images. The edge loss is expressed as

$$\begin{gathered} v = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \quad h = \begin{bmatrix} 0 & 0 & 0 \\ -1 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} \\ \text{sobel}(x, y) = \left\| \sqrt{\|v * x\|_2 + \|h * x\|_2} - \sqrt{\|v * y\|_2 + \|h * y\|_2} \right\|_1 \\ L_{\text{edge}} = \text{sobel}(G^{\text{MR}\to\text{CT}}(x_{\text{MR}}), x_{\text{MR}}) + \text{sobel}(G^{\text{CT}\to\text{MR}}(x_{\text{CT}}), x_{\text{CT}}), \end{gathered} \tag{5}$$

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

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

$$L = \lambda_1(L_{\text{GAN}}^{\text{CT}} + L_{\text{GAN}}^{\text{MR}}) + \lambda_2(L_{\text{cyc}}^{\text{CT}} + L_{\text{cyc}}^{\text{MR}}) + \lambda_3 L_{\text{Identity}} + \lambda_4 L_{\text{edge}} + \lambda_5 L_{\text{seg}}. \tag{6}$$

3.3 Fine Tune Segmentation Module in Self-Training

Even though we train the segmenter from scratch by feeding synthesized CT images, the segmentation module is only robust to some CT cases, as shown in Fig. 2(b) (the segmentation map of hard split has incorrect prediction). The segmentation performance is still limited since synthetic data cannot transfer all information from real CT images. Inspired by Ref. 24, we adopted a self-training framework to handle this challenge. We infer all pseudo-labels and probability maps for real CT images in the training cohort. The entropy calculated based on probability for each class works as a measurement to evaluate the confidence of segmentation map in unsupervised DA.[13] All segmentation maps are ranked by average entropy map I^{CT} from low to high. The larger the entropy, the more potential error the segmentation map includes. Based on ranking order, all segmentation maps are divided into easy and hard splits. The first λ of training samples are split for the easy case, and the rest split is the hard case:

$$p^{\text{CT}} = \text{softmax}(\text{Seg}(x^{\text{CT}})) \quad I^{\text{CT}} = -\sum_{i=1}^{\text{class}} (p_i^{\text{CT}} \log_2(p_i^{\text{CT}})), \tag{7}$$

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

Anatomical context, such as spatial distribution, is an important prior to medical image segmentation. To reduce incorrect prediction induced by noise and appearance shift in synthetic images, we leverage muscle and bone masks from Ref. 27 to mask out erroneous predictions for the easy split as shown in Fig. 2(b).

As shown in Fig. 2(c), we construct the discriminator D^{split} from scratch. Different from Ref. 24, the segmentation module is further trained by aligning the entropy map of easy splits to ones of hard splits. At the same time, the segmentation module is fine-tuned by feeding rectified pseudo-labels of the easy split after anatomical processing and supervised by weighted cross entropy loss $L_{\text{seg}}^{\text{easy}}$. The loss function L^{finetune} can be expressed as

$$\begin{aligned} L_{\text{GAN}}^{\text{split}} &= \mathbb{E}_{x\sim X_{\text{CT}}^{\text{easy}}}[\log\, D^{\text{split}}(x)] + \mathbb{E}_{y\sim Y_{\text{CT}}^{\text{hard}}}[1 - \log\, D^{\text{split}}(y)] \\ L^{\text{finetune}} &= \lambda_6 L_{\text{GAN}}^{\text{split}} + L_{\text{seg}}^{\text{easy}}, \end{aligned} \tag{8}$$

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

4 Experimental Results

We compare the proposed pipeline with three state-of-the-art DA methods including SynSeg-net,[17] AccSeg-Net,[18] and DISE.[21] Then, we perform an ablation study to demonstrate the effectiveness of the fine-tuning stage and sensitivity analysis for the proposed method.

4.1 Implementation Details and Evaluation Metrics

We used Python 3.7.8 and Pytorch 1.10 to implement the whole framework. The baseline and proposed methods are run on Nvidia RTX 5000 16GB GPU. For training from scratch, we set $\lambda_1 = 1.0$, $\lambda_2 = 30.0$, $\lambda_3 = 0.5$, $\lambda_4 = 1.0$, $\lambda_5 = 1.0$. In the segmentation module, the weights for background, gracilis muscle, hamstring muscle, quadriceps femoris, and sartorius muscle are set as [1,10,1,1,10] in the weighted cross-entropy loss, respectively. For the training data divided into easy and hard cohorts, we set the first $\lambda = \frac{2}{3}$ as the easy cohort and the rest as hard cohort. For fine-tuning stage, we set $\lambda_6 = 0.001$. The initial learning rate for the training model from scratch is 0.0002. We set the maximal training epochs as 100. Before the first 50 epochs, the learning rate is constant as 0.0002, and then it decreases to 0 linearly. We clip the original CT intensity to $[-200, 500]$. For the MR images, we perform min–max normalization. All CT images and MR images are normalized to $[-1, 1]$.

Dice similarity coefficient (DSC)[31] is used to evaluate the overlap between segmentation and ground truth. Briefly, we consider S as the segmentation, G as the ground truth, and $||$ as the L^1 norm operation:

$$\mathrm{DSC}(S, G) = \frac{2|S \cap G|}{|S| + |G|}. \tag{9}$$

4.2 Qualitative and Quantitative Results

A detailed comparison of quantitative performance is shown in Table 2 and Fig. 4. All methods are trained with the same training dataset, and inference is performed on the same testing dataset. In Table 2, the proposed method achieves the highest mean DSC of 0.888 with the lowest standard deviation of 0.041. The proposed method significantly differed from all baseline methods with $p < 0.05$ under Wilcoxon signed-rank test. The proposed method achieves the best DSC for each muscle group and the lowest standard deviation except for the quadriceps femoris. Compared to AccSeg-Net, the proposed method makes the largest improvement in the sartorius muscle increasing mean DSC from 0.708 to 0.837, and decreasing standard deviation from 0.176 to 0.099. In Fig. 4, compared to the second best-performing method SynSeg-net, our method further reduces outliers and has a tighter and better DSC distribution. In Fig. 5, while the baseline methods make incorrect predictions on bone, our method is more robust and has fewer incorrect predictions.

Table 2 The mean DSC and standard deviation for each muscle group and average performance across from whole test dataset. Included outliers may impact the standard deviation. Bold values represent the best result.

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

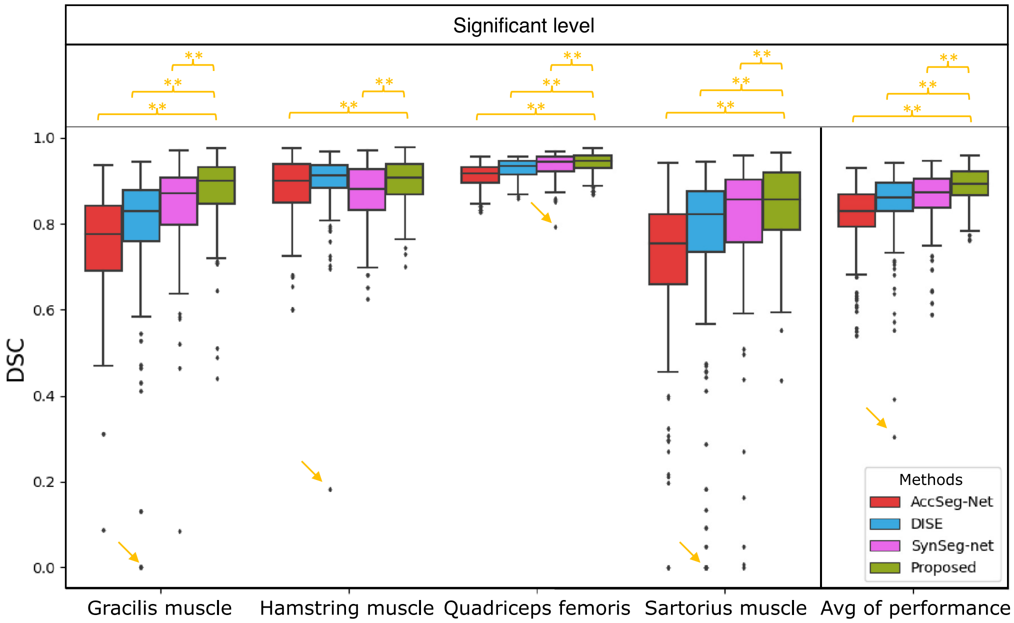


Fig. 4 Quantitative results of DSC of baseline methods and the proposed method. * indicates ($p < 0.05$) significant difference between by Wilcoxon signed-rank test and ** indicates ($p < 0.02$ corrected by Bonferroni method.[32] The yellow arrows indicate outliers that are located at a far distance from the distribution, spanning from the 25th percentile to the 75th percentile, among the four methods. When calculating the standard deviation, these outliers are included in the calculation and can potentially influence the resulting standard deviation. Therefore, the box represents the data distribution from 25th percentile to 75th percentile rather than the standard deviation of the entire test dataset.

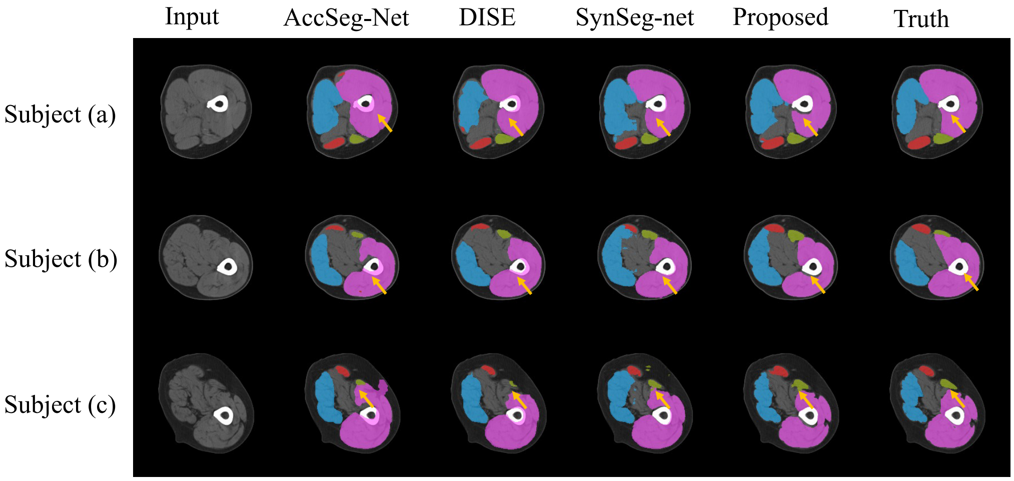


Fig. 5 Representation results of the proposed methods and baseline methods. Each row represents one subject. The proposed method reduces prediction errors on bones and around muscle group boundaries. The yellow arrows indicate differences between the proposed method and AccSeg-Net, DISE, and SynSeg-Net. The input column images are rescaled for visualization purposes.

4.3 Ablation Study

To investigate the effectiveness of the anatomical processing step and adversarial learning in fine tune stage, we designed (1) “from scratch,” (2) “from scratch + fine tune,” (3) “from scratch + muscle mask,” and (4) “from scratch + muscle mask + fine tune” pipelines by modifying the procedures of the proposed pipeline. “From scratch” represents the result directly from method section B. “From scratch + fine tune” means splitting pseudo-labels from scratch into the easy and hard cohorts and performing adversarial learning between two cohorts. “From scratch + muscle mask” represents that the muscle mask derived from Ref. 27 is used to mask out noise for the final prediction map. “From scratch + muscle mask + fine tune” represents the proposed pipeline. The graphic description for each pipeline is shown in Fig. 6.

As shown in Fig. 7, compared to “from scratch,” the proposed pipeline significantly increases mean DSC from 0.870 to 0.888. It demonstrates that the anatomical processing step plus fine-tuning stage can improve segmentation performance. Compared to “from scratch + fine tune,” the proposed pipeline significantly increased mean DSC from 0.878 to 0.888, which shows that the muscle mask can help the segmentation module discriminate noise outside the muscle mask. Compared to “from scratch + muscle mask,” the pipeline shows that adversarial learning can make the segmentation module adapt to the hard split improving DSC from 0.878 to 0.888 on the test dataset instead of only relying on the muscle mask.

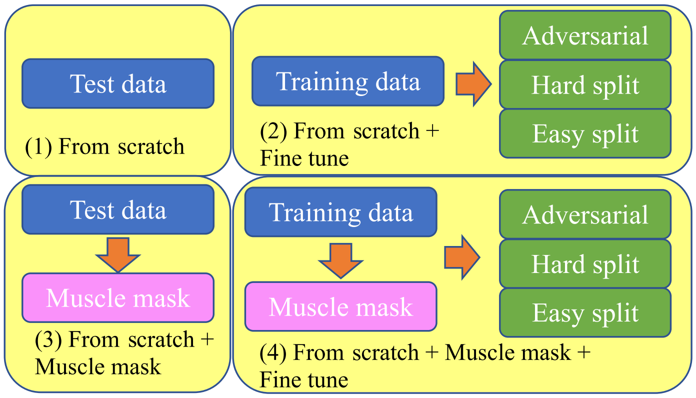


Fig. 6 Graphic visualization for the four pipelines designed for the ablation study. (1) represents segmentation maps inferred by the segmentation module trained from scratch. (2) The pseudo-labels of the training data are inferred by the segmentation module from scratch and then divided into two cohorts for fine-tuning. (3) The prediction map inferred by the segmentation module from scratch is masked by a muscle mask. (4) Proposed pipeline. The pseudo-labels of the training data are inferred by the segmentation module from scratch and then masked by a muscle mask for fine-tuning.

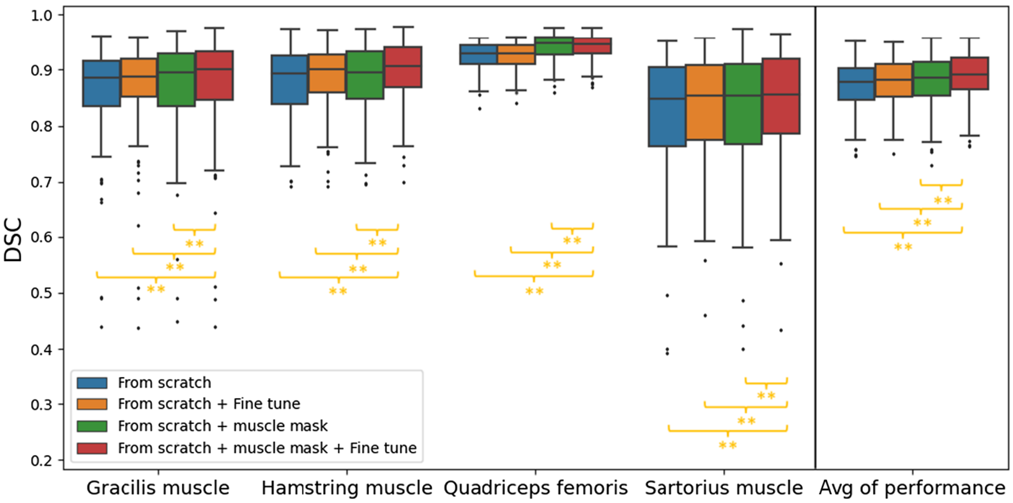


Fig. 7 The quantitative results for four pipelines were used in the ablation study. * indicates ($p < 0.05$) significant difference between by Wilcoxon signed-rank test and ** indicates $p < 0.02$ corrected by Bonferroni method.

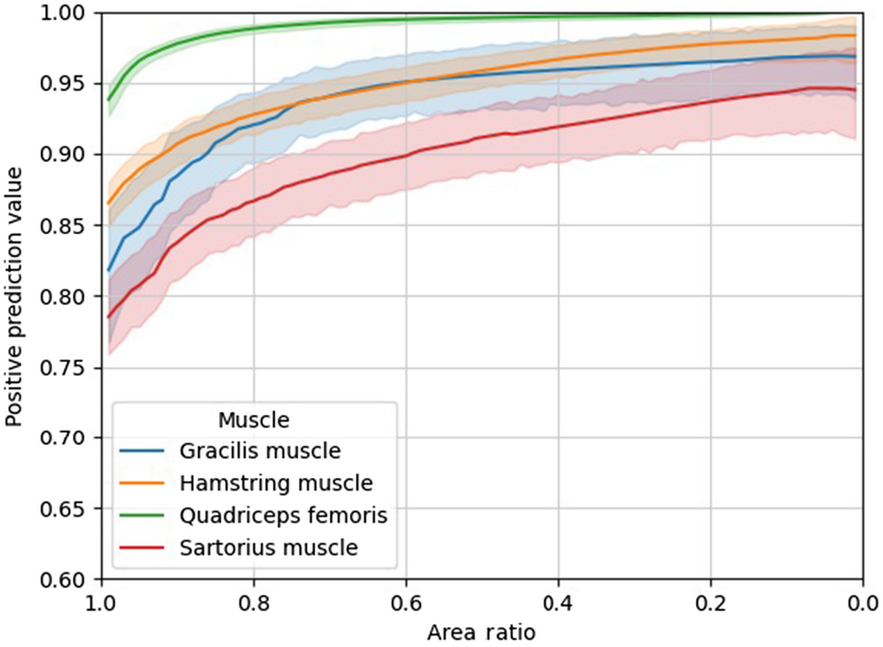


Fig. 8 The sensitivity plot of the proposed pipeline result. The *x*-axis represents the ratio between eroded area and muscle ground truth. The PPV is calculated based on Eq. (10).

4.4 Sensitivity Analysis

As shown in Fig. 1, the thigh muscle is homogeneous, and it is hard to discriminate the muscle group based on intensity alone. Furthermore, it is difficult to delineate the boundary of muscle groups by visual assessing of CT images. We perform a sensitivity analysis of the proposed method to check whether prediction maps cover central areas of muscle groups. For each muscle group, we apply a binary 3×3 kernel to erase every muscle group iteratively until the predicted muscle group is empty. The area ratio is defined as the rate between eroded muscle mask and manual ground truth. The positive predictive value (PPV) is defined as

$$\mathrm{PPV} = \frac{|S \cap G|}{|S|}, \tag{10}$$

where S represents the segmentation and G represents the ground truth. $||$ represents the L^1 norm operator. In Fig. 8, the quadriceps femoris has the highest initial PPV of 0.94, and sartorius has the lowest PPV of 0.78. The PPV of all four muscle groups is more than 0.85 when the area ratio is 0.8. The quadriceps femoris, hamstring, gracilis, and sartorius muscle have a final PPV of 1.0, 0.97, 0.96, and 0.95, respectively.

5 Discussion

In this work, we study thigh CT and achieve single slice muscle group segmentation by proposing a two stages pipeline to leverage manual labels from MR 3D volume. In the first stage, we selected single thigh CT slices from 3D volumes and split slices into left and right thigh images. The real MR and CT images were fed into a CycleGAN framework to generate synthetic CT images. The generated synthesized images were input into the segmentation module. We used the original annotation from the MR volumes to supervise the segmentation module. In the second stage, the pseudo-labels of CT images in the training cohort are inferred by the segmentation module (Seg). Based on the assumption that uncertainty is related to wrong predictions, we divided the training cohort into easy and hard splits based on inference entropy. We observed that the bone in MR is dark. However, bone in CT is bright. The significant contrast between MR and CT causes domain shift during CycleGAN incurring the wrong prediction on the bones area. To address the domain shift problem, the muscle mask from Ref. 27 is used to correct the noise map. Finally, adversarial learning aligns the prediction map between easy and hard split to make the segmentation module robust to real CT images. The thigh muscle is homogeneous and recognition of the boundary of each muscle group is difficult if it is only based on visual assessment. In order words, some boundary of one muscle group is not reproducible even with human annotation. Sensitive analysis is performed by eroding the predicted map to investigate how much eroded prediction map can cover the central area of the muscle group. As shown in Fig. 8, the quadriceps femoris has the highest PPV of 0.94, and the sartorius muscle has the lowest PPV of 0.78. The quadriceps femoris has the largest area compared with the other three muscles and is easier to predict. This can also explain why the PPV of the quadriceps femoris reaches 1.0 quickly. However, the sartorius muscle is hard to annotate and has a smaller area than the quadriceps femoris. It is hard to predict among four muscle groups, as the DSC of Table 2 indicates. When the ratio between the predicted area and ground truth is 0.6, the PPV of the sartorius muscle reaches 0.9 and the PPV of the other three muscles group reaches 0.95 or more than 0.95. The user should pay greater attention to sartorius when applying the method in a clinical scenario. To our best knowledge, this is the first pipeline to perform DA on thigh CT images. We collect all modules into one container to let the public and more researchers take advantage of our contribution (GitHub repository, https://github.com/MASILab/DA_CT_muscle_seg). The segmentation module can be directly used for single-slice CT muscle group segmentation.

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

6 Conclusion

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

Disclosures

The authors of the paper are directly employed by institutes or companies provided in this paper. No conflicts of interest exist in the submission of this paper.

Acknowledgments

This research was supported by the National Institutes of Health (NIH) under award numbers R01EB017230, R01EB006136, R01NS09529, T32EB001628, 5UL1TR002243-04, 1R01MH121620-01, and T32GM007347; by ViSE/VICTR VR3029; and by the National Center for Research Resources, grant UL1RR024975-01. It is now at the National Center for Advancing Translational Sciences, grant 2UL1TR000445-06. This project was also supported by the National Science Foundation under award numbers 1452485 and 2040462. This research was conducted with the support from the Intramural Research Program of the National Institute on Aging of the NIH. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

1. M. Hudelmaier et al., “Effect of exercise intervention on thigh muscle volume and anatomical cross-sectional areas: quantitative assessment using MRI,” *Magn. Reson. Med.* **64**(6), 1713–1720 (2010).
2. K. I. Lim et al., “The association between the ratio of visceral fat to thigh muscle area and metabolic syndrome: the Korean Sarcopenic Obesity Study (KSOS),” *Clin. Endocrinol.* **73**(5), 588–594 (2010).
3. P. Francis et al., “Measurement of muscle health in aging,” *Biogerontology* **18**(6), 901–911 (2017).

4. F. Yokota et al., “Automated muscle segmentation from CT images of the hip and thigh using a hierarchical multi-atlas method,” *Int. J. Comput. Assist. Radiol. Surg.* **13**, 977–986 (2018).
5. Y. Barnouin et al., “Manual segmentation of individual muscles of the quadriceps femoris using MRI: a reappraisal,” *J. Magn. Reson. Imaging* **40**(1), 239–247 (2014).
6. S. Schlaeger et al., “Thigh muscle segmentation of chemical shift encoding-based water-fat magnetic resonance images: the reference database MyoSegmenTUM,” *PLoS One* **13**(6), e0198200 (2018).
7. X. Yu et al., “Reducing positional variance in cross-sectional abdominal CT slices with deep conditional generative models,” *Lect. Notes Comput. Sci.* **13437**, 202–212 (2022).
8. Q. Yang et al., “Quantification of muscle, bones, and fat on single slice thigh CT,” *Proc. SPIE* **12032**, 120321K (2022).
9. X. Yu et al., “Longitudinal variability analysis on low-dose abdominal CT with deep learning-based segmentation,” *Proc. SPIE* **12464**, 1246423 (2023).
10. T. Overend et al., “Thigh composition in young and elderly men determined by computed tomography,” *Clin. Physiol.* **12**(6), 629–640 (1992).
11. H. Guan and M. Liu, “Domain adaptation for medical image analysis: a survey,” *IEEE Trans. Biomed. Eng.* **69**(3), 1173–1185 (2021).
12. J.-Y. Zhu, et al., “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in *Proc. IEEE Int. Conf. Comput. Vision*, pp. 2223–2232 (2017).
13. T.-H. Vu et al., “Advent: adversarial entropy minimization for domain adaptation in semantic segmentation,” in *Proc. IEEE/CVF Conf. Comput. Vision and Pattern Recognit.*, pp. 2517–2526 (2019).
14. G. M. Kurtzer, V. Sochat, and M. W. Bauer, “Singularity: scientific containers for mobility of compute,” *PLoS One* **12**(5), e0177459 (2017).
15. G. Wilson and D. J. Cook, “A survey of unsupervised deep domain adaptation,” *ACM Trans. Intell. Syst. Technol.* **11**(5), 1–46 (2020).
16. Q. Dou et al., “PnP-AdaNet: plug-and-play adversarial domain adaptation network at unpaired cross-modality cardiac segmentation,” *IEEE Access* **7**, 99065–99076 (2019).
17. Y. Huo et al., “SynSeg-Net: synthetic segmentation without target modality ground truth,” *IEEE Trans. Med. Imaging* **38**(4), 1016–1025 (2018).
18. B. Zhou, C. Liu, and J. S. Duncan, “Anatomy-constrained contrastive learning for synthetic segmentation without ground-truth,” *Lect. Notes Comput. Sci.* **12901**, 47–56 (2021).
19. C. Chen et al., “Unsupervised bidirectional cross-modality adaptation via deeply synergistic image and feature alignment for medical image segmentation,” *IEEE Trans. Med. Imaging* **39**(7), 2494–2505 (2020).
20. J. Yang et al., “Unsupervised domain adaptation via disentangled representations: application to cross-modality liver segmentation,” *Lect. Notes Comput. Sci.* **11765**, 255–263 (2019).
21. W.-L. Chang et al., “All about structure: adapting structural information across domains for boosting semantic segmentation,” in *Proc. IEEE/CVF Conf. Comput. Vision and Pattern Recognit.*, pp. 1900–1909 (2019).
22. Y. Zou et al., “Unsupervised domain adaptation for semantic segmentation via class-balanced self-training,” in *Proc. Eur. Conf. Comput. Vision (ECCV)*, pp. 289–305 (2018).
23. Y. Zou et al., “Confidence regularized self-training,” in *Proc. IEEE/CVF Int. Conf. Comput. Vision*, pp. 5982–5991 (2019).
24. F. Pan et al., “Unsupervised intra-domain adaptation for semantic segmentation through self-supervision,” in *Proc. IEEE/CVF Conf. Comput. Vision and Pattern Recognit.*, pp. 3764–3773 (2020).
25. L. Ferrucci, “The Baltimore Longitudinal Study of Aging (BLSA): a 50-year-long journey and plans for the future,” *J. Gerontol. Biol. Sci. Med. Sci.* **63**(12), 1416–1419 (2008).
26. X. Yu et al., “Accelerating 2D abdominal organ segmentation with active learning,” *Proc. SPIE* **12032**, 120323F (2022).
27. Q. Yang et al., “Label efficient segmentation of single slice thigh CT with two-stage pseudo labels,” *J. Med. Imaging* **9**(5), 052405 (2022).
28. C. Li et al., “Distance regularized level set evolution and its application to image segmentation,” *IEEE Trans. Image Process.* **19**(12), 3243–3254 (2010).
29. O. Ronneberger, P. Fischer, and T. Brox, “U-Net: convolutional networks for biomedical image segmentation,” *Lect. Notes Comput. Sci.* **9351**, 234–241 (2015).
30. N. Kanopoulos, N. Vasanthavada, and R. L. Baker, “Design of an image edge detection filter using the sobel operator,” *IEEE J. Solid-State Circuits* **23**(2), 358–367 (1988).
31. L. R. Dice, “Measures of the amount of ecologic association between species,” *Ecology* **26**(3), 297–302 (1945).
32. M. Hollander, D. A. Wolfe, and E. Chicken, *Nonparametric Statistical Methods*, John Wiley & Sons (2013).

Qi Yang is a graduate student at Vanderbilt University. He received his BS and MS degrees from Beijing Institute of Technology in 2016 and 2019, respectively. He is a member of SPIE.

Biographies of the other authors are not available.