Assisted annotation in Deep LOGISMOS: Combining deep learning and graph optimization for simultaneous multi-compartment 3D segmentation of calf muscles on MRI

Lichun Zhang, Zhihui Guo, Honghai Zhang, Ellen van der Plas, Timothy R. Koscik, Peggy Nopoulos, Milan Sonka, *Fellow, IEEE*

***Abstract*—Automated segmentation of individual calf muscle compartments in 3D MR images is gaining importance in diagnosing muscle disease, monitoring its progression, and prediction of the disease course. Although deep convolutional neural networks have ushered in a revolution in medical image segmentation, the availability of sufficiently large annotated datasets still limits their applicability. In this paper, we present a novel approach for solving general segmentation problems in 3D, 4D, and generally n-D. Deep LOGISMOS combines deep-learning-based pre-segmentation of objects of interest provided by our convolutional neural network, FilterNet+, and our 3D multi-objects LOGISMOS framework (layered optimal graph image segmentation of multiple objects and surfaces) that uses newly designed machine-learned cost functions trained using the paradigm of assisted annotation. We have evaluated our method on 350 lower leg (left/right) T1-weighted MR images from 93 subjects (47 healthy, 46 patients with muscular morbidity) by 4-fold cross-validation, demonstrating that our approach not only dramatically reduces the expert's annotation efforts but also significantly improves the segmentation performance. Compared with the fully manual annotation approach, the annotation cost with assisted annotation is reduced by 95%, from 8 hours to 25 minutes in this study. When assessing the segmentation performance, our new Deep LOGISMOS approach improved the earlier state-of-the-art results as follows. The mean Dice similarity coefficient (DSC) was improved by 4.6% on average, from 88.0%–91.3% to 92.9%–95.9%. The mean absolute surface positioning error was improved by 47.5% on average, from 1.4–2.2 pixels to 0.7–1.2 pixels for the five 3D muscle compartments simultaneously segmented for each leg.**

***Index Terms*—Assisted annotation, Calf muscle compartment segmentation, Deep LOGISMOS, Deep neural convolutional network, MRI**

I. Introduction

In humans, the muscles of the lower leg between the knee joint and the ankle support weight-bearing activities such as walking, running and jumping. Anatomically, this group is composed of five individual muscle compartments shown in Fig. 1(a): Tibialis Anterior (TA), Tibialis Posterior (TP), Soleus (Sol), Gastrocnemius (Gas), and Peroneus Longus (PL) [1]. Structural and volumetric changes of these compartments provide valuable information for the diagnosis, severity, and progression evaluation for various muscular diseases such as myotonic dystrophy type 1 (DM1), an inherited disorder characterized by progressive muscle weakness, myotonia, and dystrophic changes [2]. DM1 is the most common form of muscular dystrophy that begins in adulthood and causes severe fatty degeneration of calf muscle in most patients [3]. Magnetic resonance (MR), which offers non-invasive imaging of muscles with high sensitivity to dystrophic changes, has been widely used in the clinic for muscular disease diagnosis and follow-up evaluation [3], [4]. Traditional structural assessment of multiple individual muscles invariably resorts to manual tracing [5], [6], which is arduous, time-consuming, and limiting in large research and clinical settings. Automated segmentation of multiple individual calf muscles is therefore essential for developing quantitative biomarkers of muscular disease diagnosis and progression.

Past calf muscle segmentation research is relatively sparse. Valentinitsch *et al.* [7] proposed a three-stage method using unsupervised multi-parametric k-means clustering to segment calf muscle regions and subcutaneous fat for determining subcutaneous adipose tissue (SAT) and inter-muscular adipose tissue (IMAT). Yao *et al.* [8] combined deep learning with a dual active contour model to accurately locate the fascia lata and segment multiple tissue types for quantifying calf muscle and fat volumes. Amer *et al.* [9] employed deep learning to segment the whole calf muscle region where IMAT and healthy muscle are classified afterward by deep convolutional auto-encoders. All these entire muscle-region segmentation methods are mainly proposed to separate muscle, SAT and IMAT for estimating fat infiltration into muscular dystrophies.

However, the segmentation of individual muscle compartments is more desirable for assessing the progression of different neuromuscular diseases [10]. For example, it has been shown that individual skeletal muscle may be affected differently by DM1 [11]. It is necessary to improve the efficiency and utility of muscle MRI as a marker of muscle pathology [12].

Automated 3D segmentation of individual calf muscle compartments is challenging and attempts in this field are rare. As shown in Fig. 1(b-e), muscular dystrophy introduces substan-

This work was supported in part by the NIH grants R01-EB004640 and R01-NS094387, and S10-OD025025. Human subject research approved by the University of Iowa IRB as part of NIH project R01-NS094387.

L. Zhang, Z. Guo, H. Zhang, and M. Sonka are with the Iowa Institute for Biomedical Imaging, The University of Iowa, Iowa City, IA 52242, USA (Correspondence: milan-sonka@uiowa.edu).

E. van der Plas, T. R. Koscik, and P. Nopoulos are with the Dept. of Psychiatry, The University of Iowa, Iowa City, IA 52242, USA.

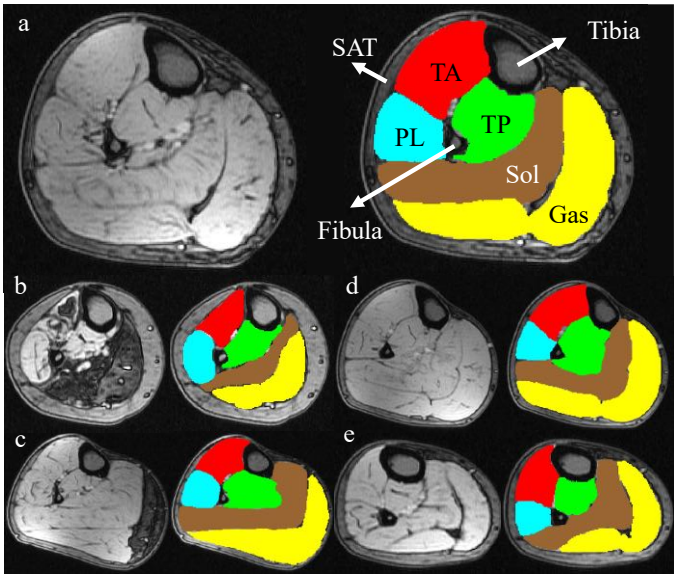

Fig. 1. Examples of T1-weighted MR images of calf muscle cross sections and corresponding expert segmentations of TA, TP, Sol, Gas and PL. (a) Normal subject. (b-c) Patients with severe DM1. (d) Patient at risk for DM1 (PreDM1). (e) Patient with juvenile onset DM1 (JDM). Best viewed in color.

tial variations of shape, texture and grayscale appearance to a part of or the entire calf region in addition to the already existing substantial variations due to the flexible nature of the muscles and leg's position in the scanner. Commean *et al.* [13] proposed a semi-automated method by thresholding and edge detection to segment bones, adipose tissue, and five individual muscle compartments. Ghosh *et al.* [14] fine-tuned a pre-trained AlexNet on 700 3D MR images to predict two parameters representing the contour of the leg muscles and achieved an average DSC (DICE Similarity Coefficient) of 0.85 ± 0.09. However, the network must be trained separately for each leg muscle and the whole method can not learn from the features while training all kinds of muscles together. More recently, Guo *et al.* [15] proposed a novel neighborhood relationship-aware network based on 3D U-Net [16], called FilterNet, for automated segmentation of individual calf muscle compartments and reached an average DSC of 0.90 ± 0.01 on 40 T1-weighted 3D MR images of 11 healthy and 29 diseased subjects. This approach was used in clinical research [12].

Although the aforementioned approaches reported acceptable segmentation performance by applying deep learning methods, several critical issues remain to be settled. 1) Availability of sufficiently large annotated datasets represents a bottleneck limiting their application, especially in large clinical settings where new data accumulates continuously. Annotation (manual tracing) of medical images is not only arduous and time-consuming but also requires costly specialty-oriented knowledge and skills. 2) There is still room for improvement of deep learning-based calf segmentation approaches. 3) Undesirable regional inaccuracies remain in the deep learning segmentation due to the lack of global-information-aware optimization. Our work attempts to address all of these issues.

Compared with previously reported approaches, the contributions of our work can be summarized as follows.

1) *Assisted annotation* with efficient adjudication substantially decreased expert manual tracing effort when forming annotated training sets.
2) *FilterNet+* improved the performance of the underlying FilterNet approach and offered stable training, accelerated convergence, improved generalization, and – as a result – improved segmentation.
3) *Deep LOGISMOS* substantially improved the performance of 3D calf muscle compartment segmentation by utilizing FilterNet+ pre-segmentation and new machine-learned cost functions.

II. Methods

A. Assisted Annotation

Fig. 2 shows the workflow of our assisted annotation approach that employs the iterative loop to achieve the best use of the existing and efficient way of adding new annotated datasets. This approach a) starts with a small training set, b) uses it to create the initial version of an automated calf segmentation method, c) employs this method to automatically segment additional unannotated images, some of which are likely segmented inaccurately at first. These automated segmentations are d) expert-corrected using Just-Enough-Interaction (JEI) functionality of LOGISMOS [17] and combined with the previous training set, thus e) forming a new larger training set of expert-annotated images, which are iteratively used to create next versions of the automated calf segmentation method in step "b". The assisted annotation steps ("b–e") are repeated until the desired performance is achieved or all data are annotated.

The process of creating new versions of the automated calf segmentation (step "b" above) relies on the following substeps in each iteration of the assisted annotation loop: 1) deep-learning based approximate pre-segmentation of calf muscle compartments; 2) deep-learning based design of LOGISMOS cost functions; 3) design of multi-object JEI for efficient editing of automatically-segmented calf compartments.

B. FilterNet+: DL-Based Pre-Segmentation

Pre-processing: Bias field correction [18] is first applied to minimize intensity non-uniformity in MR images. The z-score normalization is applied to intensities of all images of individual legs to reduce inter-subject variations. Optimal thresholding and k-means clustering are used to extract the regions of interest (ROI) corresponding to left and right legs. All right legs are mirrored to conform to left legs to reduce the task complexity. All pre-processing steps are completely unsupervised and are automatically carried out without any user intervention.

FilterNet+, its novel training strategy: Our first-attempt FilterNet approach to calf segmentation was presented in [15], introducing a neighborhood-relationship-enhanced convolution neural network. Benefiting from the increased convolution receptive field, resolution-preserving skip connections, and explicitly edge-aware regulations by a kernel-based edge gate to

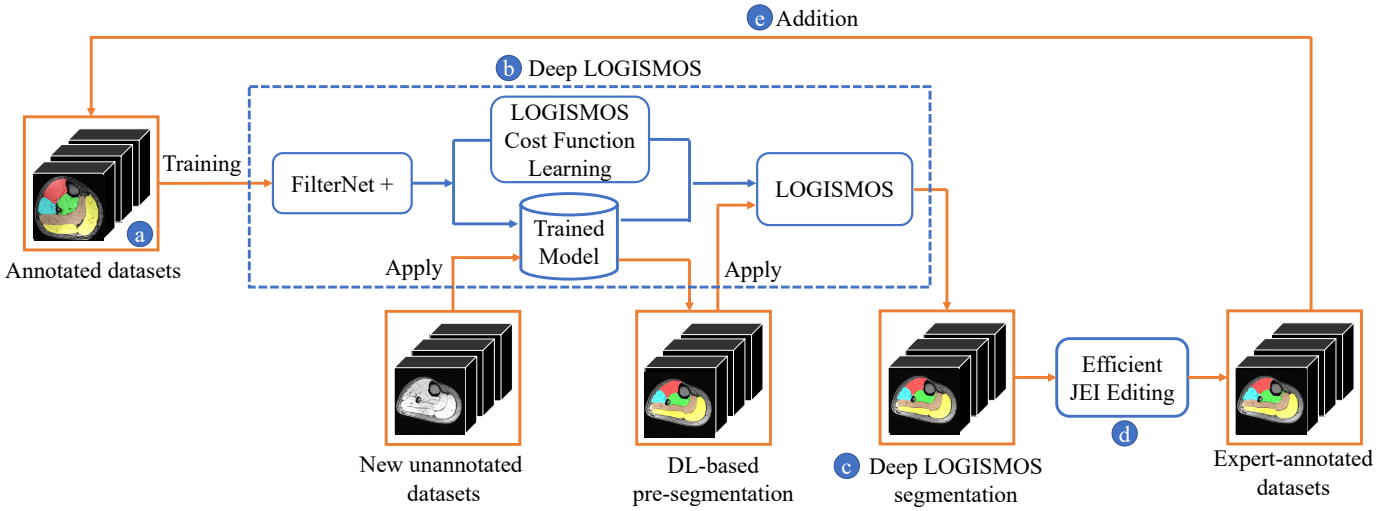

Fig. 2. Workflow of the proposed Deep LOGISMOS segmentation framework in the scheme of assisted annotation. Processing steps in blue, datasets in orange. Best viewed in color.

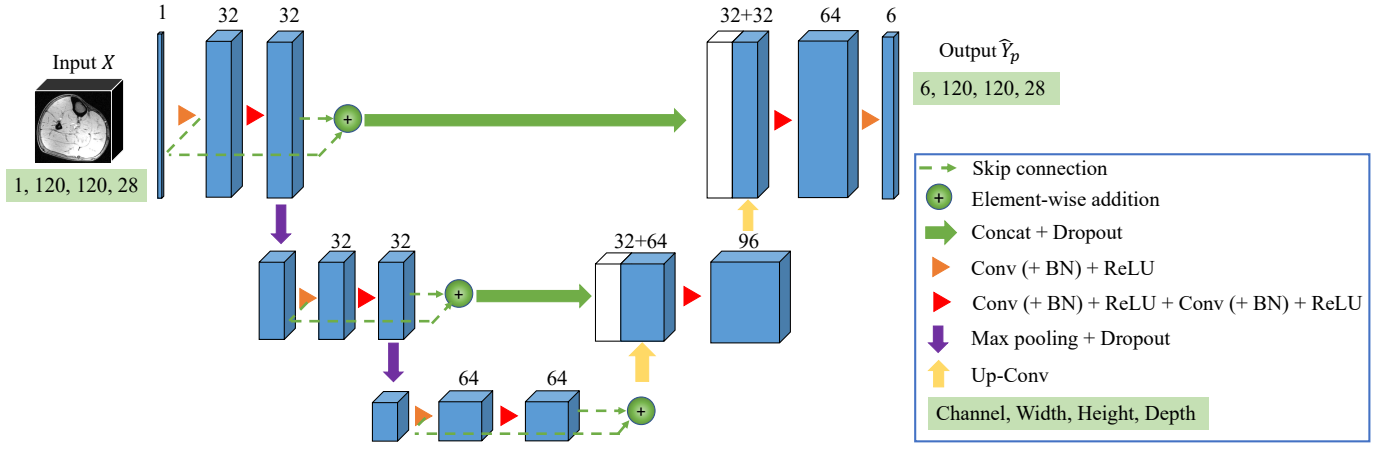

Fig. 3. The basic 3D FilterNet architecture. The input X is a $120 \times 120 \times 28$ 3D image patch cropped from the whole $160 \times 160 \times 28$ 3D pre-processed image of one leg. The size of output $\hat{Y}_p$ is $6 \times 120 \times 120 \times 28$ and it is further processed for loss calculation and the LOGISMOS cost function learning in Fig. 4. Best viewed in color.

constrain voxel-level probability values inside a neighborhood, our original FilterNet outperformed all other state-of-the-art deep-learning approaches tested in both voxel-level label predictions and 3D object surface positioning [15]. The newly designed and enhanced version, FilterNet+, overcomes several imperfect properties of the previous approach, namely the insufficient training strategy and the lack of optimization in deep learning due to underestimation of the impactful influence of non-architectural aspects. We also considered incorporation of rich network architecture extensions that were not incorporated such as attention mechanisms [19] and dense connections [20], which increased the number of network parameters and only offered marginal improvements. FilterNet+ improvements thus focus on two non-architectural aspects: the loss design and its training strategy, its architecture and training are shown in Figs. 3 and 4.

FilterNet+ training uses a new loss function L that combines of L_{dice}, multi-class cross-entropy loss L_{CE} and the edge loss L_e as

$$L = L_{dice} + (1 - \lambda)L_{CE} + \lambda L_e \ , \qquad (1)$$

where λ is an adjustable weight reflecting the strength of edge-aware regularizations through training. L_{dice} originates from DSC as in [21]:

$$L_{dice} = -\frac{2}{|N|} \sum_{n \in N} \frac{\hat{Y}^n Y^n}{\sum_{i \in Y^n} \hat{Y}_i^n + \sum_{i \in Y^n} Y_i^n} \ , \qquad (2)$$

where $N = 6$, $\hat{Y}^n$ is the predicted label for class n from the softmax output of the network, Y is one-hot encoding of the ground truth, $i \in Y$ represents voxels of the foreground in the segmentation map. Incorporation of dice loss is beneficial for the model to consider the loss information both locally and globally and as a result, improve the edge continuity between calf muscles. L_{CE} is a multi-class cross-entropy loss and L_e represents the differences (L1-norm) between the derived edge maps and the true edge maps which are generated by our 2D trainable convolution kernels, edge gate $F_{L\rho G}$. Edge gate

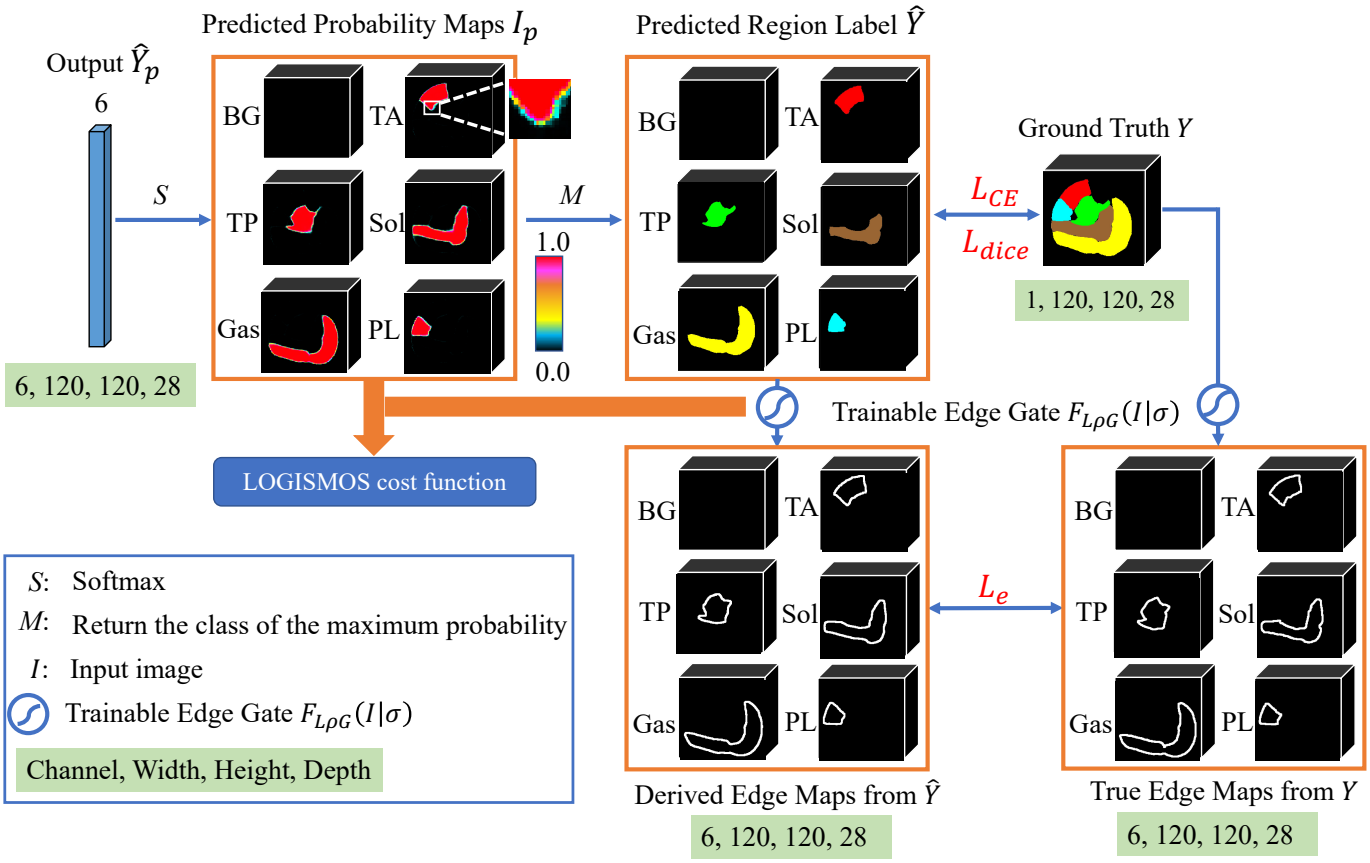

Fig. 4. Training phase of FilterNet+. Learned probability maps are optimized by L_{dice}, cross-entropy L_{CE} and edge constraints L_e. The trainable edge gate learns the muscle compartment boundary-related parameter σ from $\hat{Y}$ and Y, used later as an image-learned component of the LOGISMOS cost function (Section II-C and Fig. 5). Best viewed in color.

is a trainable variant of Laplacian of Gaussian filter and can effectively extract valuable edge information from predicted region label $\hat{Y}$ and ground truth Y to derive edge maps. It is updated while training with the trainable parameter σ, initially set as 1. More details about the edge gate can be found in our original FilterNet approach [15].

Benefiting from the new enhanced combined constraint, the network output – probability maps – are optimized to efficiently reflect both the regional and edge-based information as likelihood $[0, 1]$ of a voxel to be correctly classified, which contributes to the LOGISMOS cost function design as shown in Fig. 5.

Training of FilterNet+ was improved by introducing the following new strategies: a) dropout layers were added to the encoder path to prevent over-fitting and improve generalization [22]; b) Kaiming normalization of the initial trainable network parameters improved model fitting [23]; c) Adam optimization was employed instead of stochastic gradient descent for stochastic optimization [24], [25]; d) learning rate warmup heuristic for Adam was used to stabilize training and accelerate convergence [26]; and e) learning rate reduction was only allowed when the metric of validation stopped improving in two consecutive training epochs. These modifications resulted in stabilized training, accelerated convergence, improved generalization, and thus better segmentation performance.

Post-processing: Raw 3D object segmentation produced by the network shows local inaccuracies (small holes, coarse boundaries), which can be easily improved by simple post-processing refinement. Post-processing included two iterations of recursive Gaussian image filter ($\sigma = 2$) and hole filling by enforcing single-component connectivity of each segmented calf compartment. The refined FilterNet+ yielded approximate pre-segmentation of calf compartments, the performance of which was evaluated separately and was also further used for initialization and graph construction of the subsequent Deep LOGISMOS steps.

C. Deep LOGISMOS

LOGISMOS (Layered Optimal Graph Image Segmentation for Multiple Objects and Surfaces) is a general approach for optimally segmenting multiple n-D surfaces that mutually interact within and/or between objects [27], [28]. Columns of interconnected graph nodes are used to cover the search region for target surfaces. After assigning a cost to each node, multisurface segmentation is achieved by finding the set of nodes, one node per column, with globally optimal total cost. Additional context-specific graph arcs can be used to enforce geometric constraints that represent prior shape and anatomy knowledge. The efficiency of LOGISMOS is mainly determined by good target-object shape priors as the

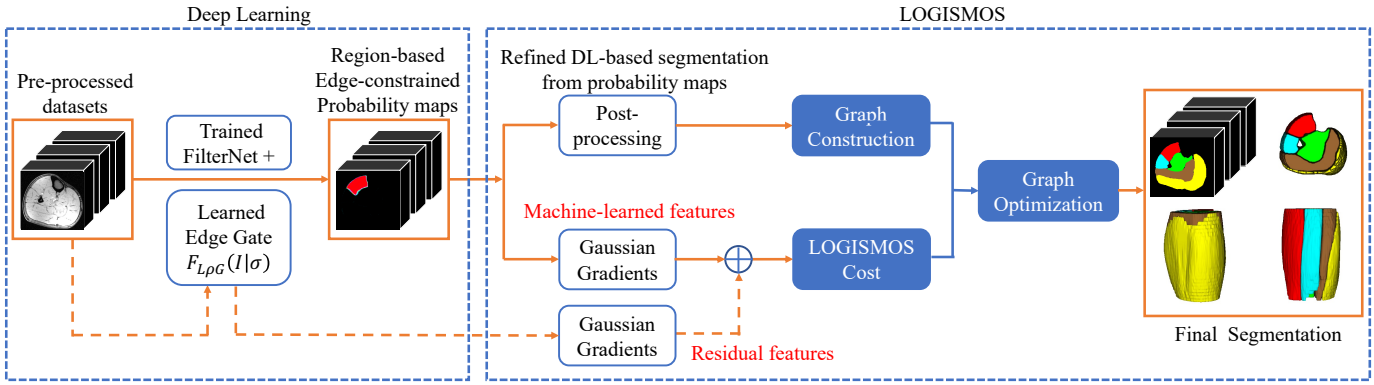

Fig. 5. Schematic diagram of the Deep LOGISMOS method. Best viewed in color.

initialization and a relevant cost function that yields the desired image segmentation. The traditional implementation relies on interactively defined initial approximate segmentation and the human-expert designed cost functions. In this work, we overcame these manual-design limitations by using the above FilterNet+ segmentation to initialize LOGISMOS while the cost functions were jointly learned from segmentation examples in combination with utilizing the independently learned FilterNet+ parameters, yielding the overall Deep LOGISMOS approach (Fig. 5).

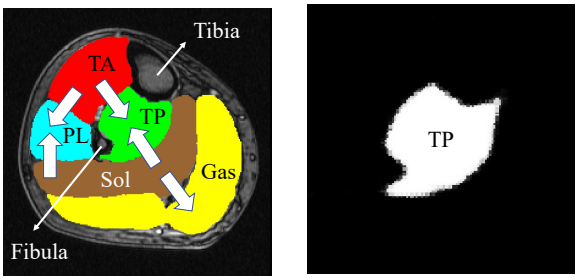

Fig. 6. (a) Graph column orientations. (b) Probability map of TP compartment.

Graph construction: FilterNet+ pre-segmentation provides approximate segmentation of each calf muscle compartment as 3D mesh surfaces, defines their topology, and mutual relationships. Graph columns are constructed along the directions normal to the mesh surfaces. More detailed information about graph construction can be found in [17], [27], [28]. To incorporate the spatial relationships between muscle compartments as object separation constraints, the column orientations of subgraphs associated with individual compartments are specially designed as shown in Fig. 6(a), where the columns are built from inside to outside for TA and Sol, and outside to inside for TP, Gas and PL. This special orientation scheme utilizes anatomical prior knowledge about the muscle compartments to avoid formation of frustrating cycles [29].

Machine learning cost design: As shown in Fig. 6(b), the appearance of the probability map of a muscle compartment is very similar to a clearly defined bright object. Therefore, the gradient of the probability along the column directions is chosen as a machine-learned feature in the trained LOGISMOS cost function. In Section II-B, the edge gate is trained globally on the predicted labels $\hat{Y}$ and the ground truth Y to derive edge maps (Fig. 4). Since $\hat{Y}$ and Y represent calf muscle compartments and hold both the region and edge information, we utilized the edge gate learned on the input calf images to derive residual features to be combined with the machine-learned features from the probability maps (Fig. 5). Contribution from the added residual features improve the proipertires of the learned cost functions.

Deep LOGISMOS segmentation: The constructed graph in the LOGISMOS system integrates shape prior from the refined FilterNet+ pre-segmentation, object separation constraints, geometric smoothness constraints, and learned costs for each node by the newly machine-learned cost function design, and the globally optimized segmentation is guaranteed by the graph optimization. The final simultaneous segmentation of all 5 calf-muscle compartments is obtained by optimal hypersurface detection in polynomial time as described in [27] .

Just-Enough Interaction – Deep LOGISMOS-JEI: The dynamic nature of the underlying algorithm is utilized to edit the segmentation result via interactive modification of local costs. Since JEI modification is directly applied to the graph, the updated result is still globally optimal (with respect to the modified costs) and satisfies existing geometric constraints. In practice, user interaction on one 2D slice is often enough to correct segmentation errors in its neighboring 2D slices and thus reduce the amount of human effort. In addition, due to the existence of embedded inter-object constraints, in regions where multiple compartments are close to each other, editing is only needed on one compartment.

III. Experimental Methods

A. Data

Only 40 lower leg T1-weighted MR images from 40 subjects were initially available (11 healthy, 23 DM1, 2 pre-DM1, 4 Juvenile Onset DM1 or JDM), the same data set as reported in [15]. Over the course of a longitudinal DM1 study, some of the initial subjects were re-scanned and new subjects added, with additional 135 MR images acquired with the same scanning parameters, increasing the annotated set size to 175 images of 350 lower legs from 93 subjects (47 healthy, 35 DM1, 6 Pre-DM1, 5 JDM). MR image size was 512×512×30, voxel

TABLE I

EVALUATION INDICES FOR FIVE CALF MUSCLE COMPARTMENTS FROM DIFFERENT SEGMENTATION METHODS AND TRAINING DATASETS. $*$ AND $**$ DENOTE RESULTS OF PAIRED t-TESTS VS. FILTERNET_80, AND FILTERNET+_350, RESPECTIVELY. SEE SECTION III-C FOR DETAILS OF THE COMPARED METHODS. BOLDED VALUES REPRESENT STATISTICALLY SIGNIFICANT IMPROVEMENTS IN COMPARISON WITH THE COMPARED APPROACHES.

		FilterNet_80	FilterNet+_80		DeepLOGISMOS_80		FilterNet_350	DeepLOGISMOS_350	
		Mean±STD	Mean±STD	p value*	Mean±STD	p value*	Mean±STD	Mean±STD	p value**
TA	DSC (%)	91.29±0.10	94.45±0.04	**0.014**	94.58±0.03	**0.005**	95.97±0.03	95.94±0.03	0.896
	JSC (%)	85.23±0.14	89.73±0.06	**0.012**	89.9±0.06	**0.004**	92.41±0.05	92.35±0.05	0.849
	ASSD (pixel)	1.42±1.25	1.02±0.59	**0.014**	1.07±1.56	0.135	0.76±0.45	0.70±0.35	**0.043**
	Max ASSD (pixel)	12.73±7.43	9.46±5.94	**0.004**	8.32±5.45	**≪0.001**	6.59±3.38	6.09±2.97	**0.024**
	ASSD score	86.81±0.10	90.01±0.05	**0.017**	90.19±0.09	**0.029**	92.42±0.04	92.97±0.03	**0.033**
	RSSD (pixel)	-0.44±1.17	0.04±0.53	**0.001**	0.24±1.58	**0.003**	0.03±0.30	0.00±0.31	0.136
	RSSD score	93.17±0.09	96.85±0.04	**0.002**	95.97±0.09	0.054	98.12±0.02	97.88±0.02	0.157
	Final score	89.12±0.10	92.76±0.04	**0.005**	92.66±0.06	**0.007**	94.73±0.03	94.79±0.03	0.818
TP	DSC (%)	89.46±0.04	91.17±0.04	**0.005**	91.19±0.04	**0.004**	93.02±0.04	93.01±0.03	0.972
	JSC (%)	81.18±0.07	83.95±0.06	**0.005**	84.00±0.06	**0.003**	87.19±0.06	87.12±0.05	0.865
	ASSD (pixel)	1.92±1.25	1.78±0.96	0.361	1.40±0.67	**0.001**	1.39±0.89	1.13±0.67	**≪0.001**
	Max ASSD (pixel)	18.78±6.45	16.80±5.88	**0.022**	14.23±5.25	**≪0.001**	14.72±6.16	13.05±5.50	**≪0.001**
	ASSD score	82.33±0.10	83.32±0.08	0.428	86.53±0.06	**0.001**	86.69±0.06	88.94±0.05	**≪0.001**
	RSSD (pixel)	0.00±1.61	0.21±1.23	0.381	-0.01±0.93	0.942	0.06±0.80	-0.14±0.67	**≪0.001**
	RSSD score	90.62±0.10	92.09±0.08	0.273	93.67±0.06	**0.029**	94.51±0.05	95.19±0.05	0.059
	Final score	85.90±0.07	87.63±0.06	0.060	88.85±0.05	**0.002**	90.35±0.05	91.06±0.04	**0.034**
Sol	DSC (%)	87.99±0.08	89.06±0.07	0.373	89.17±0.07	0.303	92.84±0.04	92.91±0.04	0.790
	JSC (%)	79.30±0.11	80.86±0.10	0.351	81.03±0.10	0.277	86.85±0.06	86.96±0.06	0.807
	ASSD (pixel)	2.24±1.57	2.28±1.37	0.895	1.87±1.02	0.063	1.48±0.74	1.23±0.56	**≪0.001**
	Max ASSD (pixel)	23.73±9.60	20.00±8.84	**0.004**	18.19±7.88	**≪0.001**	16.39±8.29	14.96±7.43	**0.010**
	ASSD score	79.95±0.12	79.48±0.11	0.788	82.55±0.08	0.087	85.82±0.06	87.97±0.05	**≪0.001**
	RSSD (pixel)	0.61±1.35	0.45±1.16	0.366	0.36±0.88	0.147	0.21±0.64	0.17±0.55	0.341
	RSSD score	90.76±0.09	92.15±0.08	0.315	93.57±0.06	**0.022**	95.43±0.05	96.03±0.04	0.062
	Final score	84.5±0.08	85.39±0.08	0.480	86.58±0.07	0.066	90.24±0.04	90.97±0.04	**0.014**
Gas	DSC (%)	89.5±0.08	90.50±0.08	0.448	90.66±0.08	0.356	94.43±0.05	94.64±0.05	0.532
	JSC (%)	81.79±0.11	83.38±0.11	0.394	83.65±0.11	0.287	89.72±0.06	90.1±0.06	0.428
	ASSD (pixel)	1.71±3.03	1.23±1.05	0.193	1.15±0.70	0.105	1.00±0.63	0.82±0.47	**≪0.001**
	Max ASSD (pixel)	29.01±18.67	22.82±17.18	**0.043**	17.80±14.75	**≪0.001**	17.73±13.97	13.77±9.96	**≪0.001**
	ASSD score	86.4±0.16	88.32±0.09	0.371	88.80±0.06	0.205	90.21±0.06	91.88±0.04	**≪0.001**
	RSSD (pixel)	0.38±2.42	-0.08±0.62	0.104	-0.14±0.44	0.060	-0.03±0.45	-0.11±0.36	**0.013**
	RSSD score	93.61±0.13	96.71±0.05	0.063	96.80±0.03	**0.040**	97.38±0.04	97.5±0.03	0.616
	Final score	87.82±0.11	89.73±0.06	0.208	89.98±0.05	0.124	92.93±0.04	93.53±0.04	**0.037**
PL	DSC (%)	89.51±0.12	91.92±0.06	0.138	92.26±0.06	0.066	94.37±0.03	94.3±0.05	0.845
	JSC (%)	82.44±0.13	85.53±0.08	0.090	86.11±0.08	**0.024**	89.47±0.05	89.56±0.07	0.832
	ASSD (pixel)	1.46±1.46	1.32±0.88	0.479	1.01±0.57	**0.012**	0.89±0.47	0.75±0.45	**≪0.001**
	Max ASSD (pixel)	11.97±6.55	9.70±4.61	**0.019**	8.58±4.10	**≪0.001**	7.74±4.13	6.86±3.84	**0.003**
	ASSD score	86.61±0.10	87.37±0.07	0.586	90.03±0.05	**0.006**	91.13±0.04	92.54±0.04	**≪0.001**
	RSSD (pixel)	0.06±1.34	-0.04±0.63	0.545	-0.19±0.46	0.099	0.06±0.45	-0.12±0.46	**≪0.001**
	RSSD score	93.53±0.09	95.55±0.04	0.087	96.27±0.03	**0.014**	97.65±0.04	97.35±0.04	0.298
	Final score	88.02±0.10	90.09±0.06	0.124	91.17±0.05	**0.009**	93.15±0.03	93.44±0.04	0.343

size 0.7×0.7×7 mm, acquisition used the first echo of a 3-point Dixon gradient echo sequence, TR=150 ms, TE=3.5 ms, FOV=36 cm, bandwidth 224 Hz/pixel, scan time 156 s.

B. Independent standard

The initial set of 40 annotated MR datasets (80 legs) was fully manually traced by experts in 3D Slicer and each annotation took approximately 8 hours on average [30]. This set was used for the initial stage of training Deep LOGISMOS to deliver decent automated segmentations of the five calf compartments. The remaining 135 MR datasets were sequentially segmented, their segmentations reviewed and – if needed – interactively corrected by experts using Deep-LOGISMOS+JEI (Section II-C), and served as additional training data used in the assisted annotation training loop (Fig. 2). The average time of reviewing and editing each 3D MR image used in the assisted annotation loop steps was approximately 25 minutes – expert effort decreased by 95%.

C. Experimental Setting

Multiple experiments were designed to compare the performance of our original FilterNet [15] that served as a baseline method with the newly developed methods and to demonstrate the contribution of our new approaches. Similarly, to demonstrate the improvements achieved by assisted annotation, we compared performance on differently sized datasets (fully traced or assist-annotated). The following methods were compared, the numeric index specifies the number of training datasets used:

- FilterNet_80: The original deep-learning baseline approach reported in [15], 40 subjects, 80 legs.

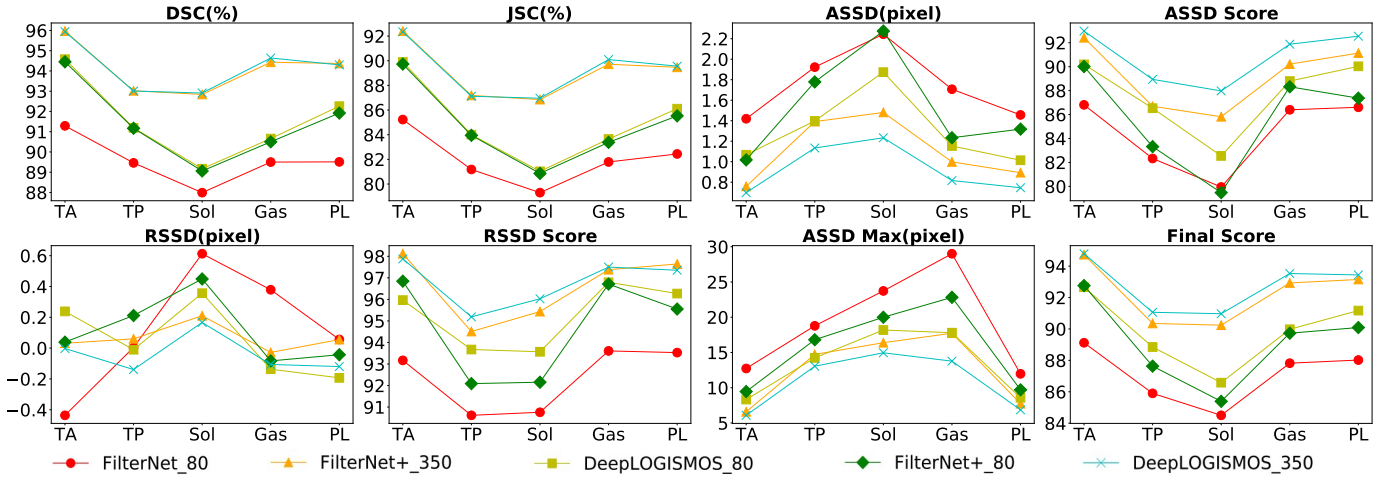

Fig. 7. Performance comparison for the segmentation of five calf muscle compartments from different experiments. Best viewed in color.

- FilterNet+_80: The deep learning method resulting from performance extensions of the original FilterNet approach, 40 subjects, 80 legs.
- FilterNet+_350: FilterNet+ approach using the full assisted-annotation datasets of 93 subjects, 350 legs.
- DeepLOGISMOS_80: Deep LOGISMOS method using FilterNet+_80 results as pre-segmentation.
- DeepLOGISMOS_350: Deep LOGISMOS method using FilterNet+_350 results as pre-segmentation.

Given a limited-size dataset, 4-fold cross-validation was used to evaluate the performance of each tested approach with the 4 groups created randomly at the subject level so that data (legs) from the same subject were never simultaneously used for both training and testing. The data were split 65%–10%–25% to form the training, validation, and testing sets. That means that for a dataset of 80 (350) legs, training+validation was based on 60 (262) legs and testing was done in 20 (88) legs, repeated 4 times.

Each image segmentation method design uses specific parameters that influence its behavior. In all tests, the same parameters were used in the corresponding steps of each method. In FilterNet+, to increase the robustness and generalization of the network, the input image patches were sized 120×120×28, cropped from the localized leg-areas, overlapping with a step size of 20 voxels along the x and y directions, yielding 9 times as many training patches as the number of available leg images. Data augmentation was performed on each training patch with random rotation, scaling, etc. The learning rate was halved throughout the training process if the combination of the loss L on the validation dataset did not decrease in 2 consecutive training epochs. FilterNet+ training loss parameter λ was initially set as 0.001 and increased ten fold every 10 epochs, the batch size was 16. FilterNet+ was implemented using PyTorch platform [31] and trained on Nvidia Tesla V100 GPU with 32 GB memory. LOGISMOS graph columns consisted of 49 nodes spaced 0.35 mm apart. LOGISMOS smoothness constraints were set as 6 node-to-node distances, corresponding to 2.1 mm.

D. Quantitative Analysis

To comprehensively evaluate the segmentation performance and allow method-to-method comparisons, DSC and Jaccard Similarity Coefficient (JSC) evaluated region-based accuracy, absolute surface-to-surface distance (ASSD, in pixels) and relative surface-to-surface distance (RSSD, in pixels) assessed boundary-based accuracy. ASSD and RSSD measure the distances between the surface of the automated segmentation and the independent standard. Because of the order-of-magnitude difference between the XY plane in-slice resolution (0.7 mm) and the Z plane slice distance (7 mm), the surface-to-surface distances were calculated on the XY plane slice-by-slice. To allow meaningful comparisons, scores for ASSD and RSSD were calculated as

$$S_{ASSD} = \alpha^{-ASSD} \times 100\ ,\ S_{RSSD} = \alpha^{-|RSSD|} \times 100\ , \quad (3)$$

where α is an application-specific parameter empirically chosen as $\alpha = 1.111$ to reach approximate linearity and maximum score of 1.0 when the two surfaces match (at zero distance). Given the four indices: DSC, JSC, S_{ASSD}, and S_{RSSD}, the final comprehensive score was defined as

$$S_{final} = 0.25 \times (DSC + JSC + S_{ASSD} + S_{RSSD})\ . \quad (4)$$

Higher S_{final} indicates better comprehensive performance in the combined regional and boundary-positioning respect. Additionally, $ASSD_{max}$, was evaluated as the maximum absolute distance between two surfaces of each compartment. Performance indices were averaged for left and right calf muscle compartments and reported as mean±standard deviation. For statistical comparisons between methods, paired t-tests were used, p value < 0.05 denoted statistical significance.

IV. Results

The performance comparisons of the five tested methods are listed in Table I and also visualized in Fig. 7. Compared with the original FilterNet_80 [15], FilterNet+_80 achieved significantly better results for each compartment in terms of DSC, JSC, ASSD, RSSD, $ASSD_{max}$, as well as the comprehensive S_{final} for at least some of the compared quantitative indices

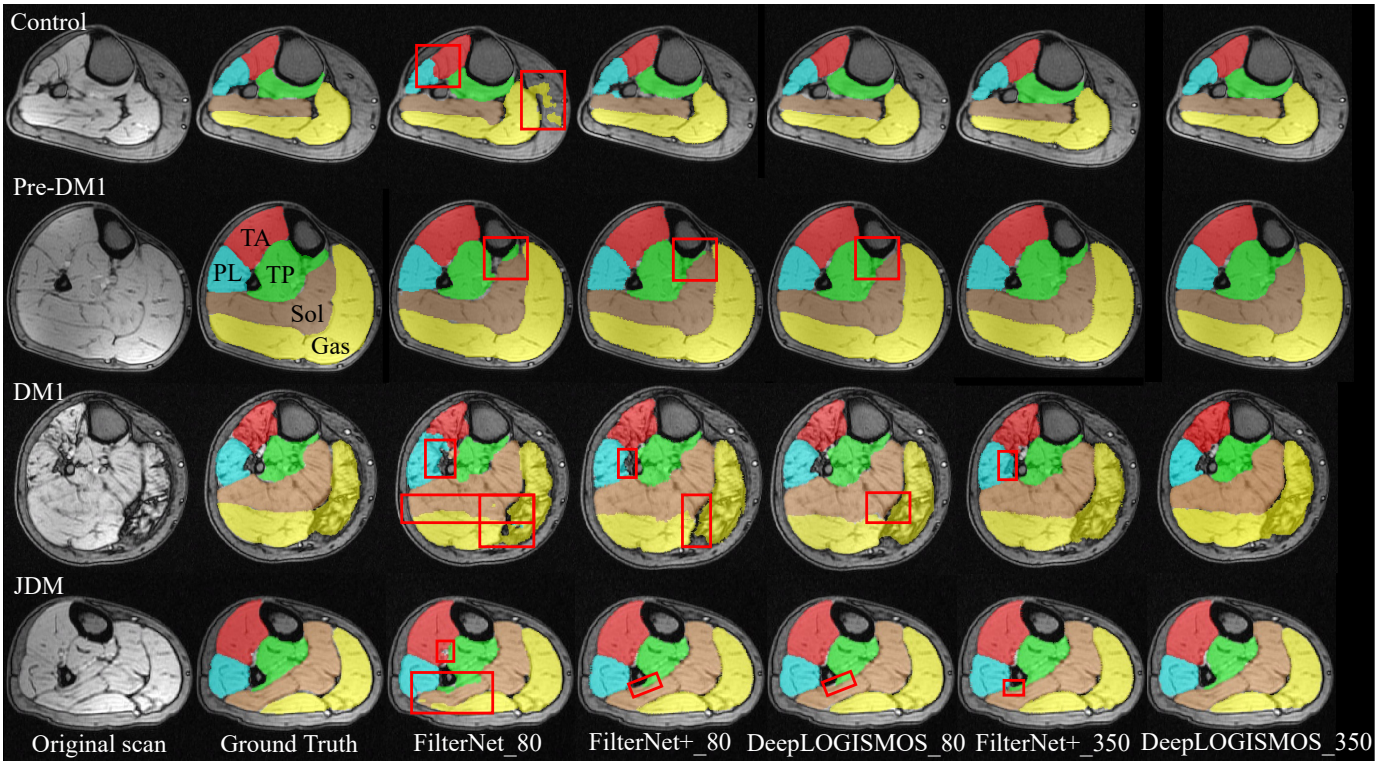

Fig. 8. Examples of segmentation overlaid with MR images. Each row shows a representative 2D cross-sectional slice from the image of subject with given status (healthy/diseased). The red rectangles highlight regions with segmentation errors. In the control subject, the infiltration of Gas into SAT in FilterNet_80 is corrected in FilterNet+_80. In the Pre-DM1 example, the disconnected TP in FilterNet_80 is significantly alleviated in FilterNet+_80. In the severe DM1 and JDM case, FilterNet_80 performs poorly while other methods obtain satisfactory results. In the DM1 and JDM example, segmentation improvements are noticeable in each method from the FilterNet+_80 to DeepLOGISMOS_350, for which muscle compartments segmented are increasingly better agreeing with the ground truth. Throughout all the examples, the errors in FilterNet_80 tend to be corrected in FilterNet+_80 and FilterNet+_350, and are further topologically optimized by DeepLOGISMOS methods. Best viewed in color.

with the remaining indices showing statistically comparable performance. The notable performance increase attests to the effectiveness of our proposed deep learning method improvements. Compared to the original FilterNet_80, DeepLOGISMOS_80 offered yet additional improvement of at least some compared indices for all muscle compartments, the other indices were statistically comparable. In particular, ASSD and RSSD scores were increased substantially after the LOGISMOS steps resulting in improved values of S_{final}. On the 350-leg (93-subject) dataset obtained by assisted annotation, both the FilterNet+_350 and DeepLOGISMOS_350 outperformed the methods using 80-leg dataset. Similarly, compared to FilterNet+_350 alone, DeepLOGISMOS_350 demonstrated overall improvements for all muscle compartments, many statistically significant differences, especially in terms of ASSD and RSSD. Notably, the maximum ASSD values, representing the locally most severe segmentation inaccuracies, decreases significantly for all 5 segmented muscle compartments.

Fig. 8 displays four cross-sectional segmentation examples from images of four subjects – healthy control, Pre-DM1, DM1, and JDM. The comparisons show that the most advanced DeepLOGISMOS_350 avoids almost all of the segmentation inaccuracies present in the results of the other methods.

V. Discussion

A. Ablation Study

Table I and Fig. 7 show the superiority of our deep learning method in comparison with our previous FilterNet approach. In agreement with the ablation study principles, the results are methodically ordered from the simplest and earliest original FilterNet_80 in increasing complexity by first introducing improvement in the FilterNet+_80 approach, then combining FilterNet+ with LOGISMOS optimization, proceeding further to employing assisted annotation to increase the training sizes in FilterNet+_350 and DeepLOGISMOS_350 approaches, Fig. 8 demonstrates that segmentation inaccuracies in FilterNet_80 (tunnels, mis-classifications, undesired disjoint objects) are successfully resolved by the improvements designed for the FilterNet+_80 approach. DeepLOGISMOS_80 segmentations exhibit additional increases in accuracy, surface smoothness, and topologic superiority as shown in Fig. 8. For the DM1 subject, while the PL compartment segmented by FilterNet+_80 spreads into the surrounding tissue, this problem is resolved by DeepLOGISMOS_80 due to the addition of machine-learned residual features to the cost function (Section II-C). Similarly, benefiting from LOGISMOS graph optimization, a region of mis-classified PL around the boundary of Sol and mis-classified Gas in the control subject by FilterNet_80 is corrected by FilterNet+_80 and DeepLOGISMOS_80. At the

same time, the importance of topologic correctness of Pre-DM1 and DM1 pre-segmentation can be seen in FilterNet+_80 and DeepLOGISMOS_80, where the attractive edge-costs at falsely pre-segmented locations did not allow LOGISMOS to properly reposition the Sol surface.

The superiority of training on a larger dataset, indicating the effectiveness of assisted annotation, is further shown in FilterNet+_350 and DeepLOGISMOS_350 (Table I, Fig. 7). LOGISMOS-JEI based assisted annotation, in the process of providing larger training datasets, dramatically reduces the annotation effort of human experts. In all four examples in Fig. 8, most of the errors in the earlier segmentation approaches are successfully resolved by FilterNet+_350 and DeepLOGISMOS_350.

B. Generalizability of combining Deep LOGISMOS and assisted annotation

The power of our work in combining deep learning pre-segmentation and graph-optimality seeking Deep LOGISMOS trained on data produced by efficient assisted annotation was demonstrated in the case of segmenting human calf muscle compartments on MRI. Alternatively, other deep convolutional neural network architectures can be integrated into the Deep LOGISMOS framework to utilize information linkages between deep learning and graph optimization. Further strengthened by the inherently incorporated Deep-LOGISMOS+JEI based assisted annotation (Fig. 2), its effectiveness and efficiency in reducing the annotation effort and optimizing the segmentation model are clearly visible from the achieved segmentation improvements (Section IV). Note of course, that the Deep-LOGISMOS+JEI method used here for assisted annotation is not the only one applicable. The idea of training-segmentation-annotation iterative epochs can be generically incorporated into supervised learning methods or one can elect to employ suggested annotation approaches [32]. Given this inherent generalizability of Deep LOGISMOS and the assisted annotation paradigm, these strategies can be further integrated and the machine-learned deep segmentation features and the machine-learned LOGISMOS cost functions applied to various segmentation tasks to benefit both the segmentation processes and those leading to assisted annotations.

C. Future work

Although we showed that assisted annotation helps experts reduce the effort of manual tracing substantially (from 8 hours to 25 minutes per 3D image), the total time and effort of reviewing and editing a large dataset can not be neglected either. There are two promising directions to further relieve the annotation effort problem: active learning [33] and quality assessment without ground truth. The approach of quality assessment without the ground truth focuses on further reducing the human effort in searching for small segmentation errors in a large 3D image by automatically locating likely segmentation errors on the volumetrically visualized object surfaces. Afterward, the identified likely-erroneous locations can be used as feedback to guide the network to prevent similar errors. As a result, the time of reviewing and editing the segmentations to produce new annotations can be significantly reduced.

VI. Conclusion

A hybrid framework combining the main advantages of our convolutional neural network FilterNet+ with those of our graph-based LOGISMOS approach, further supported by Deep-LOGISMOS+JEI assisted annotation, was reported. The presented comparative performance assessment demonstrated an improved performance obtained during simultaneous multi-compartment 3D segmentation of calf muscle compartments on 3D MRI. By maximizing the value of an original small dataset of fully annotated MR images of 80 lower legs, and by initially training a Deep LOGISMOS segmentation method on this small dataset, we have designed and employed an efficient assisted annotation strategy that decreased the average annotation time required to 3D-annotate 5 calf-muscle compartments on a volumetric $512\times512\times30$ MR image from 8 hours to 25 minutes – a 95% reduction of human expert effort. Our Deep LOGISMOS method trained on a larger dataset of 350 assisted-annotated legs then outperformed all other tested deep learning and graph-optimization approaches in the region-based voxel labeling, boundary-based surface positioning, and the final comprehensive performance score. Compared with our previously reported FilterNet method, mean DSC was improved by 4.6% on average, from 88.0–91.3% to 92.9–95.9%. The mean absolute surface positioning errors were improved by 47.5% on average, from 1.4–2.2 pixels to 0.7–1.2 pixels. The mean comprehensive final score was improved by 6.5 on average, from 84.5–89.1 to 91.0–94.8 for the five 3D muscle compartments per leg. The reduction of local maximum segmentation errors (Max ASSD) was even more pronounced. The striking performance improvements suggest the clinical-use potential of our new fully automated simultaneous segmentation of calf muscle compartments.

Acknowledgment

Eric Axelson's contributions to data preparation and management are gratefully acknowledged.

References

[1] A. Yaman, C. Ozturk, P. A. Huijing, and C. A. Yucesoy, "Magnetic resonance imaging assessment of mechanical interactions between human lower leg muscles in vivo," *Journal of biomechanical engineering*, vol. 135, no. 9, 2013.

[2] L. P. Ranum and J. W. Day, "Myotonic dystrophy: clinical and molecular parallels between myotonic dystrophy type 1 and type 2," *Current neurology and neuroscience reports*, vol. 2, no. 5, pp. 465–470, 2002.

[3] M. P. Wattjes, R. A. Kley, and D. Fischer, "Neuromuscular imaging in inherited muscle diseases," *European radiology*, vol. 20, no. 10, pp. 2447–2460, 2010.

[4] A. C. Ogier, M.-A. Hostin, M.-E. Bellemare, and D. Bendahan, "Overview of MR image segmentation strategies in neuromuscular disorders," *Frontiers in Neurology*, vol. 12, p. 255, 2021.

[5] L. Heskamp, M. van Nimwegen, M. J. Ploegmakers, G. Bassez, J.-F. Deux, S. A. Cumming, D. G. Monckton, B. G. van Engelen, and A. Heerschap, "Lower extremity muscle pathology in myotonic dystrophy type 1 assessed by quantitative MRI," *Neurology*, vol. 92, no. 24, pp. e2803–e2814, 2019.

[6] L. Maggi, M. Moscatelli, R. Frangiamore, F. Mazzi, M. Verri, A. De Luca, M. B. Pasanisi, G. Baranello, I. Tramacere, L. Chiapparini *et al.*, "Quantitative muscle MRI protocol as possible biomarker in Becker muscular dystrophy," *Clinical neuroradiology*, pp. 1–10, 2020.

[7] A. Valentinitsch, D. C. Karampinos, H. Alizai, K. Subburaj, D. Kumar, T. M. Link, and S. Majumdar, "Automated unsupervised multi-parametric classification of adipose tissue depots in skeletal muscle," *Journal of Magnetic Resonance Imaging*, vol. 37, no. 4, pp. 917–927, 2013.

[8] J. Yao, W. Kovacs, N. Hsieh, C.-Y. Liu, and R. M. Summers, "Holistic segmentation of intermuscular adipose tissues on thigh MRI," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2017, pp. 737–745.

[9] R. Amer, J. Nassar, D. Bendahan, H. Greenspan, and N. Ben-Eliezer, "Automatic segmentation of muscle tissue and inter-muscular fat in thigh and calf MRI images," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 219–227.

[10] H. Alizai, L. Nardo, D. C. Karampinos, G. B. Joseph, S. P. Yap, T. Baum, R. Krug, S. Majumdar, and T. M. Link, "Comparison of clinical semi-quantitative assessment of muscle fat infiltration with quantitative assessment using chemical shift-based water/fat separation in MR studies of the calf of post-menopausal women," *European radiology*, vol. 22, no. 7, pp. 1592–1600, 2012.

[11] C. Kornblum, G. Lutterbey, M. Bogdanow, K. Kesper, H. Schild, R. Schröder, and M. P. Wattjes, "Distinct neuromuscular phenotypes in myotonic dystrophy types 1 and 2," *Journal of neurology*, vol. 253, no. 6, pp. 753–761, 2006.

[12] E. van der Plas, L. Gutmann, D. Thedens, R. K. Shields, K. Langbehn, Z. Guo, M. Sonka, and P. Nopoulos, "Quantitative muscle MRI as a sensitive marker of early muscle pathology in myotonic dystrophy type 1," *Muscle & Nerve*, vol. 63, pp. 553–562, 2021.

[13] P. K. Commean, L. J. Tuttle, M. K. Hastings, M. J. Strube, and M. J. Mueller, "Magnetic resonance imaging measurement reproducibility for calf muscle and adipose tissue volume," *Journal of Magnetic Resonance Imaging*, vol. 34, no. 6, pp. 1285–1294, 2011.

[14] S. Ghosh, N. Ray, and P. Boulanger, "A structured deep-learning based approach for the automated segmentation of human leg muscle from 3D MRI," in *2017 14th Conference on Computer and Robot Vision (CRV)*. IEEE, 2017, pp. 117–123.

[15] Z. Guo, H. Zhang, Z. Chen, E. van der Plas, L. Gutmann, D. Thedens, P. Nopoulos, and M. Sonka, "Fully automated 3D segmentation of mr-imaged calf muscle compartments: neighborhood relationship enhanced fully convolutional network," *Computerized Medical Imaging and Graphics*, vol. 87, p. 101835, 2021.

[16] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: learning dense volumetric segmentation from sparse annotation," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2016, pp. 424–432.

[17] H. Zhang, K. Lee, Z. Chen, S. Kashyap, and M. Sonka, "LOGISMOS-JEI: Segmentation using optimal graph search and just-enough interaction," in *Handbook of Medical Image Computing and Computer Assisted Intervention*. Elsevier, 2020, pp. 249–272.

[18] N. Tustison and J. Gee, "N4ITK: Nick's N3 ITK implementation for MRI bias field correction," *Insight Journal*, vol. 9, 2009.

[19] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz *et al.*, "Attention U-Net: learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.

[20] S. Jégou, M. Drozdzal, D. Vazquez, A. Romero, and Y. Bengio, "The one hundred layers tiramisu: Fully convolutional DenseNets for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 11–19.

[21] M. Drozdzal, E. Vorontsov, G. Chartrand, S. Kadoury, and C. Pal, "The importance of skip connections in biomedical image segmentation," in *Deep learning and data labeling for medical applications*. Springer, 2016, pp. 179–187.

[22] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[24] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[25] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of COMPSTAT'2010*. Springer, 2010, pp. 177–186.

[26] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han, "On the variance of the adaptive learning rate and beyond," *arXiv preprint arXiv:1908.03265*, 2019.

[27] K. Li, X. Wu, D. Z. Chen, and M. Sonka, "Optimal surface segmentation in volumetric images – a graph-theoretic approach," *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, no. 1, pp. 119–134, 2005.

[28] Y. Yin, X. Zhang, R. Williams, X. Wu, D. D. Anderson, and M. Sonka, "LOGISMOS – layered optimal graph image segmentation of multiple objects and surfaces: cartilage segmentation in the knee joint," *IEEE transactions on medical imaging*, vol. 29, no. 12, pp. 2023–2037, 2010.

[29] A. Delong and Y. Boykov, "Globally optimal segmentation of multi-region objects," in *2009 IEEE 12th International Conference on Computer Vision*. IEEE, 2009, pp. 285–292.

[30] A. Fedorov, R. Beichel, J. Kalpathy-Cramer, J. Finet, J.-C. Fillion-Robin, S. Pujol, C. Bauer, D. Jennings, F. Fennessy, M. Sonka *et al.*, "3D Slicer as an image computing platform for the quantitative imaging network," *Magnetic resonance imaging*, vol. 30, no. 9, pp. 1323–1341, 2012.

[31] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in PyTorch," 2017.

[32] L. Yang, Y. Zhang, J. Chen, S. Zhang, and D. Z. Chen, "Suggestive annotation: A deep active learning framework for biomedical image segmentation," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2017, pp. 399–407.

[33] D. A. Cohn, Z. Ghahramani, and M. I. Jordan, "Active learning with statistical models," *Journal of artificial intelligence research*, vol. 4, pp. 129–145, 1996.