Convolutional Neural Network-Based Pelvic Floor Structure Segmentation Using Magnetic Resonance Imaging in Pelvic Organ Prolapse

Fei Feng
University of Michigan - Shanghai Jiao Tong University Joint Institute, Shanghai Jiao Tong University, Shanghai, 200240, China

James A. Ashton-Miller
Department of Mechanical Engineering, University of Michigan, Ann Arbor, MI 48109, USA

John O.L. DeLancey
Department of Obstetrics and Gynecology, University of Michigan, Ann Arbor, MI 48109, USA

Jiajia Luo
Biomedical Engineering Department, Peking University, Beijing, 100191, China

June 30, 2020

Corresponding author: Jiajia Luo. email: jiajia.luo@pku.edu.cn

Abstract

Purpose: Automated segmentation could improve the efficiency of modeling-based pelvic organ prolapse (POP) evaluations. However, segmentation performance is limited by the blurry soft tissue boundaries. In this study, we aimed to present a hybrid solution for uterus, rectum, bladder, and levator ani muscle segmentation by combining a convolutional neural network (CNN) and a level set method.
Methods: We used 24 sagittal pelvic floor magnetic resonance (MR) series from six anterior vaginal prolapse and six posterior vaginal prolapse subjects (a total 528 MR images). The stress MR images were performed both at rest and at maximal Valsalva. We assigned 264 images for training, 132 images for validation, and 132 images for testing. A CNN was designed by introducing a Multi-Resolution Features Pyramid module (MRFP) into an encoder-decoder model. Depth separable convolution and pre-training were used to improve model convergence. Multi-class cross entropy loss and multi-class Dice loss were used for model training. The Dice Similarity Coefficient (DSC) and average surface distance (ASD) were used for evaluating the segmentation results. To prove the effectiveness of our model, we compared it with advanced segmentation methods including Deeplabv3+, U-Net, and FCN-8s. The ablation study was designed to quantify the contributions of MRFP, the encoder network, and pre-training. Besides, we investigated the working mechanism of MRFP in the segmentation network by comparing our model with three of its variants. Finally, the level set method was

used to improve the CNN model further.

Results: Dice loss showed better segmentation performance than multi-class cross entropy loss. MRFP was efficacious for different encoder networks. With MRFP, U-Net and U-Net-X (X represents Xception encoder network) have improved the DSC, on average by 6.8 and 5.3 points. Compared with different CNN models, our model achieved the highest average DSC of 65.6 points and the lowest average ASD of 2.9 mm. With the level set method, the DSC of our model improved to 69.4 points.

Conclusions: MRFP proved to be effective in addressing the blurry soft tissue boundary problem on pelvic floor MR images. A hybrid solution based on CNN and level set method was presented for pelvic organ segmentation both at rest and at maximal Valsalva; with this method, we achieved state-of-the-art results.

Contents

I. Introduction

Pelvic organ prolapse (POP) is an abnormal caudal displacement and deformation of one or more female pelvic floor organs. POP can cause considerable discomfort to women both physically and mentally. In the United States, about 200,000 women undergo POP surgery every year, at a total cost of more than $1 billion[1,2]. The most common imaging techniques to evaluate POP include magnetic resonance (MR) and ultrasound imaging. Due to the good contrast of soft tissues, MR imaging has always been the golden standard for organ segmentation. Organ segmentation is crucial for three-dimensional (3D) geometric model reconstruction, finite element simulation of POP, and surgical planning[3,4]. Currently, manual organ segmentation is still the most widely used technique. However, the manual segmentation is not only time-consuming but also susceptible to large inconsistencies depending on the experience and skill of the evaluators and the quality of MR scans. To speed up the segmentation process, computer-aided diagnostic techniques may hold promise.

Several difficulties constrain the pelvic organ segmentation performance. First, MR images do not provide high enough contrast at the boundary of each organ, which makes segmentation particularly challenging for humans. Second, the occurrence rate is unbalanced between organs, which limits model convergence. For example, organs like the bladder are present in more MR images, whereas some organs, including rectum and uterus, may not be seen at all in many MR images, when viewed laterally. Adding to that challenge, some patients have undergone hysterectomy and lack a uterus. Third, large variations exist in these data. For instance, the shape and size of pelvic organs vary widely between resting and stressed (Valsalva) states (Fig. 1). Besides, the levator ani muscle exhibits a large inter-subject variance on MR images due to its structural complexity.

Computer-aided segmentation techniques include both deep learning and non-deep learning methods. The non-deep learning methods, including the deformable model and level set methods, have played an important role in the segmentation of the cardiac ventricle and other human body regions[5,6,7,8]. One limitation is that those methods often fail to converge for images with blurry boundaries. Besides, their segmentation speeds do not fulfill the current needs for rapid segmentation as they require much human interaction. Moreover, the poor generalization is a typical problem that both automatic and semi-automatic methods face. Generalization problems are usually related to generalization in new regions or on new data. The first generalization problem means that one organ segmentation algorithm is usually not suitable for another organ. This hampers POP analysis since we usually want to obtain a segmentation of the uterus, bladder, levator ani muscle, rectum, vaginal walls, and other tissues simultaneously. The second generalization problem is even more crucial for the clinical application of automatic segmentation tools. Since there are large variations in the structural profiles, it is challenging to find a solution that can adapt to inter-subject variability in MR images.

Recently, the convolutional neural network (CNN) has become the mainstream method for approaching many computer vision and medical imaging analysis problems. These include cell, lesion, tumor, retinal vessel, cardiac structure, and brain segmentation[9,10,11,12,13]. Compared with non-deep learning methods, CNN usually does not rely on much prior knowledge of the data[14,15], and it is trained with MR data from different subjects. Thus it has good generalization performance. The basic idea of the CNN method is that it uses several convolution layers to extract features so it can provide pixel-wise segmentation. Some researchers have proved that the sequentially stacked convolution layers are difficult to converge, so the residual connection and shortcut connection were proposed in ResNet[16] and U-Net[10] respectively, to smooth the model training process and preserve more detailed information.

Several CNN models were designed for different segmentation problems. U-Net[10] adopted the encoder-decoder network to accomplish neuronal structures segmentation and cell tracking tasks. V-net[17] used a 3D convolution to accomplish the volumetric segmentation task. DeepMedic[11] employed a dual-path 3D CNN based on dense patch ideas to deal with the high computational burden when training 3D CNN for brain lesion segmentation. UNet++[18] connected the encoder and decoder networks by a series of dense skip connections to avoid eliminating the gap between encoder and decoder networks and obtained better performance than U-Net and wide U-Net on four segmentation datasets.

However, these designs could not capture different scales of semantic information. Segmentation is a task that needs details at different scales. Coarse segmentation could be achieved from lower resolution feature maps, while the fine-grained boundary information must be detected from higher resolution feature maps. Therefore, different sizes of features may preserve different scales of context information[19]. Inspired by the image pyramid, an ensemble method of using different scales of features has been proposed to combine information from different scales of features to preserve different levels of image details. Initially, it was used for image classification and object detection. For example, spatial pyramid pooling[20] was proposed to deal with the variance in scale, size, and aspect ratio for the image classification problem. However, it was modified to detect objects with various scales, sizes, and aspect ratios. Single Shot MultiBox Detector[21] kept six different size feature maps for object detection and achieved a better detection performance. Feature Pyramid Networks[22] generated predictions at different feature levels for a single scale input image in order to take advantage of different levels of semantic information. Pyramid Scene Parsing Network[23] has been proposed for the pyramid pooling module to take advantage of prior global semantics and to capture different scales of contextual information by a parallel feature map stacking method. Deeplabv3+[24] used the atrous spatial pyramid pooling to replace the downsampling method to avoid the risk of potential information loss.

In this study, we present a CNN-based solution for segmenting four female pelvic organ structures from MR images both at rest and at maximal Valsalva. In the deep CNN model, a Multi-Resolution Feature Pyramid (MRFP)[24] module was inserted into the U-Net skip connections to capture the semantic information from different scales to improve segmentation performance in blurry regions. Depth separable convolution was used to improve the encoder network convergence. Transfer learning was applied to deal with inadequate training data. In post-processing, a level set method was used to further improve the CNN performance. The novelty of our work could be summarized in three areas. First, it represents a novel application for pelvic organ segmentation both at rest and at maximal Valsalva in women with and without POP, based on a deep learning method with MR images. Second, it is a novel design to combine MRFP with U-Net for blurry region segmentation of medical images. We proved its effectiveness in blurry pelvic organ segmentation of high-variance MR images in POP. Third, we applied a post-processing method to deal with the failure cases and further improve segmentation performance. As a result, compared with existing segmentation methods, our method achieves the best performance.

II. Materials and Methods

II.A. Data population and processing

We used 24 sagittal pelvic floor MR series of 12 subjects from the Michigan Pelvic Floor Research Collection that had been obtained with the approval of the institutional ethics review committee in case-control studies of POP. The subjects included six anterior vaginal prolapse and six posterior vaginal prolapse cases. Three women with and three women without a uterus were included per group. Supine, multi-planar MR imaging was performed in both resting and stressed states (maximal Valsalva when the patient attempts to increase the intra-abdominal pressure in order to push the pelvic organs out through the vaginal canal). All of the studies were scanned with a 3T superconducting magnet (Philips Medical Systems Inc, Bothell, WA, USA) with accompanying software (v. 2.5.1.0). In the sagittal plane, at rest, of each subject 30 slices were taken in a field of view of 200×200 mm, with a thickness of 4 mm per slice and a spacing between slices of 1 mm; at maximal Valsalva, due to the time limitation for the subjects to hold the stressed status, of each subject 14 slices were taken of scanning range 360×360 mm with a thickness of 6 mm per slice and a spacing of 1 mm[25]. The annotation of uterus, rectum, bladder, and levator ani muscle was accomplished based on previous anatomic work[26] using 3D Slicer software (v.3.4.2009-10-15). The annotation was accomplished by one expert and reviewed by another senior expert. Some pre-processing steps were applied to reduce the variance between these data. All of the slices were interpolated to the same interval in height and width dimensions. These images were then resampled into 256×256 pixel sizes for CNN model training. As there were a total of 24 sagittal pelvic floor MR series from 12 subjects and a total of 528 MR images, the different datasets were assigned as 12 3D MR series (264 images) from six subjects for training, six 3D MR series (132 images) from three subjects for validation, and six 3D MR series (132 images) from three subjects for testing. The organ occurrence rate in the training data is shown in Table 1. The uterus had the lowest occurrence rate, and the bladder had the highest occurrence rate.

II.B. Convolutional neural network structure

The main conceptual framework for our CNN model is illustrated in Fig. 2. The model had an encoder-decoder network structure[10,27]. When constructing the encoder network, we adopted the Xception[10,28] structure with residual connections. To extract different scales' context information, we used the MRFP module in the skip connections between the encoder and decoder, which will be introduced in the following subsection.

II.B.1. Multi-resolution feature pyramid

To merge context information at multiple scales, we needed these operations to have fields of view of different sizes. Larger kernel size and dilated convolution are two options. Since the parameter quantity increases drastically as the increase of kernel sizes, we adopted dilated convolution. Each MRFP module consists of four dilated convolutional layers and one average pooling layer (Fig. 2). We used 1×1 convolution with dilation 1, 3×3 convolution with dilation 1, 3×3 convolution with dilation 2, and 3×3 convolution with dilation 3 to perceive context information at scales of 1×1, 3×3, 5×5, and 7×7. All feature maps in different branches were concatenated together for the decoder network. A convolution layer was used to mix the feature maps from different scales. Therefore, the MRFP module is capable of capturing multi-scale contextual information. It was applied to all five shortcut connections in our model.

II.B.2. Encoder network structure

The encoder network (Fig. 3) is essential for feature extraction as well as for segmentation. Our encoder network adopted the Xception idea[28], which takes advantage of depth separable convolution to achieve the decomposition of ordinary convolution into channel-wise convolution and point-wise convolution. Customization of the model structure was proposed with modification on the downsampling operation. To preserve more detail, we replaced the pooling layers with a convolution of stride 2. Besides, we used fewer layers in the Middle Flow to avoid overfitting.

II.C. Post-processing method

The level set is a partial differential equation (PDE)-based method. A curve could be defined as $\phi(t, x, y)$, and after giving an initialization, the curve evolves based on image-driven forces. The PDE equation is as follows[29]:

$$\frac{\partial \phi}{\partial t} = \nabla\phi \cdot F, \phi(0, x, y) = \phi_0 \tag{1}$$

where t is the iteration times, x and y are image coordinates, $\phi_0 = 0$ defines the initial segmentation, and F is the velocity field. To be specific, in post-processing we used the level set method to improve the segmentation organ by organ. Using the bladder as an example, before applying the level set method, we first computed the minimum 3D boundary that includes the CNN-based bladder segmentation. This 3D boundary was then used to crop the 3D data including the bladder from the original 3D MR data. Finally, with CNN-based bladder segmentation as the initialization, we applied the level set method to the cropped MR data slice-by-slice for bladder segmentation. During model testing, compared with the ground truth, we evaluated our results using Dice Similarity Coefficient (DSC) metric and we kept the results of the level set method if they are better than the initial results. In practical applications, since ground truth values are not available, users need to determine whether the CNN model makes acceptable predictions. When users find the predictions provided by the CNN model to be unacceptable, such as the MR image segmentation is far beyond the normal range, the level set method will be applied for post-processing, although we will only keep the better final result. For convenience, we used the *morphological_chan_vese*[30] function in the scikit-image library[31].

II.D. Loss function and metrics

We investigated two different loss functions for model training, that is, pixelwise multi-class cross entropy loss (CE) and multi-class Dice loss (DL):

$$DL = 1 - 2\frac{\sum_{l=1}^{N}\sum_n t_{ln}p_{ln}}{\sum_{l=1}^{N}\sum_n (t_{ln} + p_{ln})} \tag{2}$$

$$CE = \sum_{l=1}^{N}\sum_n (-t_{ln}log(p_{ln})) \tag{3}$$

where $N = 5$ in our case, representing the background, uterus, rectum, bladder, and levator ani muscle classes, t_{ln} is the ground truth labeling on the nth pixel position for class l, and p_{ln} is the prediction result on the nth pixel position for class l.

Four metrics were used for individual organ segmentation evaluation, that is, the DSC, Average Symmetric Surface Distance (ASD), Relative Absolute Volume Difference (RAVD), and Organ Detection Recall (ODR). Following the definition of DL, the DSC is defined as follows:

$$DSC = 2\frac{\sum_{l=1}^{N}\sum_n t_{ln}p_{ln}}{\sum_{l=1}^{N}\sum_n (t_{ln}+p_{ln})} \times 100 \tag{4}$$

And the ASD is defined as follows:

$$ASD = 2\frac{1}{|S_T| + |S_P|}\left(\sum_{s_t\in S_T}\min_{s_p\in S_P}\|s_t - s_p\|_2 + \sum_{s_p\in S_P}\min_{s_t\in S_T}\|s_p - s_t\|_2\right) \tag{5}$$

where S_T and S_P are the surface of the ground truth and model prediction, respectively, and s_t and s_p are corresponding points in them. The RAVD is defined as follows:

$$RAVD = \frac{|V_T - V_P|}{V_T} \times 100 \tag{6}$$

where V_T and V_P are the volume of ground truth and model prediction, respectively. The ODR is defined as follows:

$$ODR = \frac{TP}{TP + FP} \times 100 \tag{7}$$

where TP is the number of images in which an organ is correctly detected and FP is the number of images in which the same organ is not correctly detected.

II.E. Experiments

The experiment setup was summarized as below. Experiments were implemented with Keras (v.2.2.0) using Python (v.3.5.0). Adam solver was used to minimize the loss function. Our choice for the learning rate was 0.0001, with a learning rate decay of 0.98 after each epoch. A total of 800 epochs were used for training. We used an NVIDIA 1080Ti graphic card to enable the parallel computing process, with a batch size of 4. To reduce overfitting because of insufficient data, we used data augmentation. The augmentation techniques included image rotation, shear and shift, sharpening, blurring, and contrast normalization. Before images were fed to the CNN model, they were set to zero mean and unit standard variance.

The Xception encoder network was trained on a cardiac structure segmentation dataset[32] for transfer learning.

Experiments were conducted as follows. First, we compared DL with the CE function. Second, we compared the proposed method's performance with three other advanced segmentation methods, that is, Deeplabv3+[33], U-Net[10], and FCN-8s[34]. Deeplabv3+[33] is a state-of-the-art semantic segmentation method, FCN-8s[34] has obtained state-of-the-art results on a PASCAL VOC 2012 Segmentation dataset, and U-Net[10] is a classical biomedical segmentation method which won a challenge competition in 2015. Third, we quantified the effectiveness of the Xception encoder network and the MRFP module using ablation studies. Compared with U-Net with the Xception (U-Net-X), and U-Net with MRFP (U-Net-M), our model used U-Net with both the Xception and MRFP (U-Net-XM). Fourth, we investigated the effects of the MRFP module among different skip connections between the encoder and decoder networks. In our model, as the encoder has five downsampling stages, there are five corresponding skip connections, which are the first to fifth skip connection from top to bottom in Fig. 2. Our model used MRFP in all the five connections so we called it U-Net-XM_{12345}. We compared our model with its three variants, that is, U-Net-XM_{123}, U-Net-XM_{135}, and U-Net-XM_{345}. Finally, we used the level set method to improve the results of all segmentation methods in the second experiment.

III. Results

III.A. Loss function comparison

The DL function obtained a much better segmentation result (Table 2), both with and without pre-training. Hence, in the following training, we compared different methods using the DL function. The model with pre-training showed better performance than without pre-training under both loss function configurations. The pre-training improved the average DSC from 64.0 to 65.6 when using DL. However, the pre-training operation exhibited the "butterfly effect", which means the model performance improved more in the post-processing step (Table 7), as discussed in Section III.E.

III.B. Performance comparison with other advanced segmentation methods

The proposed method yielded better results with respect to the DSC than the other three methods (Table 3). Our model without pre-training had an average DSC of 64.0, wining in three of four individual tasks (uterus, rectum, and bladder). FCN-8s showed better performance on the rectum, but its average DSC was only 58.2. Compared with Deeplabv3+ (60.2), FCN-8s (58.2), and U-Net (54.8), our model achieved an average DSC that was 3.8, 5.8, and 9.2 points higher than them, respectively. However, our model with pre-training did not exhibit better bladder segmentation performance than the model without pre-training because the bladder of one subject was outside the normal range (Fig. 5e). Segmentation of this subject was improved in the post-processing step (see Section III. E).

We also compared the model performances using the ODR and the RAVD (Table 4). Our model obtained the best RAVD, but did not show a distinct advantage with respect to the ODR. The ODR is the proportion of images with this organ that were correctly detected of the total number of images with this organ. The results indicate our model does not have a better organ detection ability. However, our model showed a markedly better segmentation performance (Table 3), which means that for the images that were correctly detected, our model had results closer to the ground truth. A comparison of the models' predictions is shown in 4.

III.C. Ablation study

Ablation experiments were performed to quantify the effectiveness of the MRFP and the encoder network. The difference between U-Net-M and U-Net is the use of MRFP. The difference between U-Net-X and U-Net is the use of Xception encoder network. Therefore, the difference between our model (U-Net-XM$_{12345}$) with U-Net-X or U-Net-M is the use of MRFP or Xception, respectively. The result is summarized in Table 5.

The DSC of U-Net-M, compared with U-Net, increased from 54.8 to 61.6, an increase of 6.8 points; the DSC of our model, compared with U-Net-X, increased from 58.7 to 64.0, an increase of 5.3 points; the DSC of U-Net-X, compared with U-Net, increased from 54.8 to 58.7, an increase of 3.9 points; the DSC of our model, compared with U-Net-M, increased from 61.6 to 64.0, an increase of 2.4 points. This proved the effectiveness of MRFP when used with U-Net or U-Net-X. Besides, MRFP made a larger contribution to the final segmentation performance. For each organ, with respect to the DSC, MRFP made a larger contribution to the uterus and the bladder than for the rectum and the levator.

III.D. Different MRFP combinations comparison

The detailed segmentation results are summarized in Table 6. For the average DSC, our model (U-Net-XM$_{12345}$) obtained almost the same results with U-Net-XM$_{345}$ and U-Net-XM$_{135}$, while it was 2.4 points higher than U-Net-XM$_{123}$. For individual organ segmentation, our model achieved almost the same results with U-Net-XM$_{345}$ and U-Net-XM$_{135}$ for the uterus and bladder, and slightly worse results for the rectum, and slightly better results for the levator. The rectum results improved using the post-processing technique in Section III.E (Table. 7). With respect to the ASD, our model obtained the best results. Besides, U-Net-XM$_{123}$ obtained better results than U-Net-XM$_{345}$ and U-Net-XM$_{135}$.

III.E. Post-processing improvement

We improved all CNN methods' results with the level set method. A comparison of the models' predictions is shown in Fig. 5. We demonstrated the re-segmentation results by organs. Since the levator and rectum were usually connected and showed no visible edges, it was difficult to segment them using the level set method. Therefore, the uterus (Fig 5a and b), rectum (Fig 5c and d), and bladder (Fig 5e and f) were used for comparison. With the deep learning model's prediction as prior knowledge, the level set method remedied the failure cases to a certain extent (Fig. 5a, c, and e). However, compared with the deep learning method, the level set method did not provide better segmentation results in some general cases (Fig. 5b, d, and f) even with the deep learning model's prediction as initialization.

Final segmentation results of CNN methods after post-processing are summarized in Table 7. Our model obtained the best DSC and ASD results for both individual organs and the overall average. The model without pre-training achieved an average DSC of 66.1 points, outperforming other methods with 4.0 to 9.7 points. Our model with pre-training obtained the highest average DSC (69.4 points) and best average ASD (2.9 mm).

IV. Discussion

IV.A. CNN application to POP analysis

Our work represents a novel application for female pelvic organ segmentation both at rest and at maximal Valsalva in women with and without POP, using a CNN method with MR images. In the end, we presented a hybrid solution for simultaneous uterus, rectum, bladder, and levator ani muscle segmentation and showed good results qualitatively and quantitatively. There are some differences with previous investigations[35,36,37,38,39,40]. Different modalities of medical imaging techniques have their own advantages. Two groups used ultrasound images to accomplish levator hiatus segmentation using the fully CNN (FCN) and U-Net[37,41]. Wang et al.[38] and He et al.[39] investigated prostate, rectum and bladder segmentation using axial view computed tomography based on a multi-stage FCN. Techniques including dilated convolution[42] and full-resolution residual network[43] were also investigated to deal with the blurry edges of objects by capturing a larger field of view information. The level set technique as a shape prior has been considered previously for natural image segmentation[44].

Although MR imaging is the golden standard for analyzing POP, it is quite challenging, even for clinical experts, to segment pelvic organs in MR images at rest and at maximal Valsalva of women with and without POP. Our deep learning model's performance is also limited by the imaging quality, the stress state, the prolapse status, and the training set size, etc. For example, the difficulty changes with segmentation from different views[35,39]. Prolapse is a downward displacement and deformation of pelvic organs, and thus its analysis is usually done from sagittal views. However, it might be more difficult for both humans and computer models to segment the uterus, levator, and rectum in the sagittal view compared with the axial view, in which the smaller organs have a higher occurrence rate. For the MR images in the sagittal view, the rest images have a thickness of 4 mm and 1 mm spacing. At maximal Valsalva, the stress images have a thickness of 6 mm and 1 mm spacing. The difficulty increases when segmenting small or thin organs, such as the levator ani and the rectum. The organs of women with POP also showed more variance than those of healthy women at maximal Valsalva compared to resting state, i.e., bladders of prolapsed women might become longer at maximal Valsalva, which is very different from the bladder segmentation of men. Besides, we only included 24 sagittal MR series of 12 subjects, and images of six subjects were used for model training, limiting the deep learning model's performance. Despite these challenges, nevertheless, our deep learning model still obtained the best performance compared with other methods (Table 7).

IV.B. Effectiveness analysis of different components

The effectiveness of the MRFP module is illustrated by the ablation experiments. As shown in Table 5, the average DSC of U-Net-M improved by 6.8 points compared with U-Net. The average DSC of our model improved by 5.3 points compared with U-Net-X. These results suggest that MRFP is efficacious for different encoder networks. Comparing the DSC for individual organs (Table 5), MRFP made larger improvements for the uterus and bladder than for the rectum and levator, because no information is obtained on the edge between the levator and rectum, as shown in Figs. 4 and 5. It is even tricky for humans to segment the rectum and levator. Models with different MRFP combinations (Table 6) revealed that our model (U-Net-XM_{12345}) had almost the same average DSC as U-Net-XM_{345} and U-Net-XM_{135}, but a better result on average ASD. U-Net-XM_{123} achieved a lower average DSC than U-Net-XM_{345} and U-Net-XM_{135}, but a better average ASD. A possible explanation for these observations is that MRFP on higher-order (fourth and fifth) skip connections could improve model convergence, while MRFP on lower-order (first and second) skip connections could smooth the segmentation results. In the end, our model U-Net-XM_{12345}, achieved the best results for both average DSC and ASD, and it is therefore the recommended design.

The effectiveness of the Xception encoder network is shown in Table 5. The average DSC of U-Net-X was 3.9 points higher than that of U-Net. The average DSC of our model was 2.4 points higher than that of U-Net-M on average DSC. This proved the importance of an encoder network, and a better encoder network is useful to improve segmentation.

The effectiveness of pre-training was proved in Tables 3 and 7. We can conclude the pre-training made a larger contribution to the uterus and levator segmentation than to the rectum and bladder segmentation. We used a cardiac MR dataset for pre-training, but a larger pelvic MR dataset might give better results. It also means more training data could be helpful to improve segmentation.

The effectiveness of the post-processing method is shown in Tables 3 and 7. It also proved useful for all the CNN methods in our experiments. However, these improvements were based on using the CNN model prediction as prior knowledge. The level set method made improvements for some failure cases, such as for the examples in Fig. 5a, c, and e. However, for general cases, the level set method did not provide better segmentation than the CNN method even with the CNN prediction as initialization, such as for the examples in Fig. 5b, d, and f. This suggests that the CNN method has an advantage in blurry region segmentation due to training with "big data". On the contrary, since it is often challenging to collect medical imaging data and to label them, the non-deep learning method could be useful to improve the model performance to some extent. So far, whether post-processing has improved the results needs to be compared with the ground truth. This means that it is up to the user to determine whether or not post-processing is needed. Fortunately, comparison is a much easier task than manual segmentation. But it points to the fact that we can integrate the level set method into the CNN workflow to achieve better and faster segmentation.

IV.C. Segmentation performance analysis

We improved the segmentation performance from three aspects. First, we used the MRFP module to improve the blurry region segmentation on pelvic MR images. The average DSC when using MRFP increased from 54.8 to 61.6 points (Table 5). Second, we built the encoder network based on the Xception idea and transfer learning technique. With the Xception, our model's performance increased from 61.6 to 64.0 points (Table 5). Pre-training process improved the average DSC from 64.0 to 65.6 points (Table 3). However, the pre-training operation contributed to more improvements (3.8 points) in the post-processing step (Table 7). Third, we introduced the level set method as a post-processing technique to deal with the limited training data and high-variance problems. Using post-processing, our model with pre-training improved from 65.6 to 69.4 points on average DSC (Table 7). With respect to the DSC, our model outperformed other methods with 7.3 to 13.0 points. Additionally, we compared the models' performances using the ODR and the RAVD (Table 4). Our model did not show a distinct advantage with respect to the ODR, which means our model does not detect more organs than other methods. Nevertheless, our model showed better segmentation performance (Table 3), suggesting that with respect to the organs that were correctly detected, our model's results are closer to the ground truth.

The segmentation performance was ordered as follows: bladder > rectum > uterus > levator. The results of the bladder were markedly better, because the bladder has larger size, and clearer boundary than that of other organs. The rectum is easy to detect since its ODR results were higher compared to the levator and uterus (Table 4). Half of the subjects did not have a uterus, which further exacerbated the shortage of training data and the imbalance of the data, resulting in a low ODR. However, our model could predict whether there is a uterus from the subject level evaluation. After post-processing, the highest DSC for the uterus was 65.3, which exhibited the largest improvement, as shown in Tables 3 and 7. The levator ani had the worst segmentation results, since it has the smallest size and does not have a clear boundary; identifying the levator ani is always a challenge, even for experienced clinicians.

V. Conclusions

To segment pelvic organs at rest and at maximum Valsalva (stress), we proposed a novel CNN design by integrating the MRFP module into an encoder-decoder model. This proved useful to address the blurry soft tissue boundary problem on MR images in POP. Together with the Xception encoder network and model pre-training, our model obtained better segmentation results than Deeplabv3+, FCN-8s, and U-Net. Moreover, due to the limited training data problem, a level set method was used to improve the segmentation of failure cases. Future directions include feature fusion between 2D and 3D CNNs to exploit spatial context information as discussed by Isense et al.[9]. Model pre-training with unlabeled data using unsupervised or self-supervised methods, which could take advantage of more data, can also potentially improve the segmentation quality.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation of China General Program grant 31870942, US Public Health Service grants R01 HD038665 and P50 HD044406.

CONFLICT OF INTEREST

The authors have no relevant conflict of interest to disclose.

References

1 S. H. Boyles, A. M. Weber, and L. Meyn, Procedures for pelvic organ prolapse in the United States, 1979-1997, American Journal of Obstetrics and Gynecology **188**, 108–115 (2003).

2 L. L. Subak, L. E. Waetjen, S. van den Eeden, D. H. Thom, E. Vittinghoff, and J. S. Brown, Cost of pelvic organ prolapse surgery in the United States, Obstetrics and Gynecology **98**, 646–651 (2001).

3 L. Chen, J. A. Ashton-Miller, and J. O. L. DeLancey, A 3D finite element model of anterior vaginal wall support to evaluate mechanisms underlying cystocele formation, Journal of Biomechanics **42**, 1371–1377 (2009).

4 J. Luo, L. Chen, D. E. Fenner, J. A. Ashton-Miller, and J. O. L. DeLancey, A multi-compartment 3-D finite element model of rectocele and its interaction with cystocele, Journal of Biomechanics **48**, 1580–1586 (2015).

5 L. Hoyte, W. Ye, L. Brubaker, J. R. Fielding, M. E. Lockhart, M. E. Heilbrun, M. B. Brown, and S. K. Warfield, Segmentations of MRI images of the female pelvic floor: A study of inter- and intra-reader reliability, Journal of Magnetic Resonance Imaging **33**, 684–691 (2011).

6 Z. Ma, R. N. M. Jorge, T. Mascarenhas, and J. M. R. S. Tavares, Segmentation of female pelvic cavity in axial T2-weighted MR images towards the 3D reconstruction, International Journal for Numerical Methods in Biomedical Engineering **28**, 714–726 (2012).

7 Z. Ma, J. M. R. S. Tavares, R. N. Jorge, and T. Mascarenhas, A review of algorithms for medical image segmentation and their applications to the female pelvic cavity, Computer Methods in Biomechanics and Biomedical Engineering **13**, 235–246 (2010).

[8] H. R. Malone, O. N. Syed, M. S. Downes, A. L. D'Ambrosio, D. O. Quest, and M. G. Kaiser, Simulation in Neurosurgery: A Review of Computer-Based Simulation Environments and Their Surgical Applications, Neurosurgery **67**, 1105–1116 (2010).

[9] F. Isensee, P. F. Jaeger, P. M. Full, I. Wolf, S. Engelhardt, and K. H. Maier-Hein, Automatic Cardiac Disease Assessment on cine-MRI via Time-Series Segmentation and Domain Specific Features, in *Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges*, edited by M. Pop, M. Sermesant, P.-M. Jodoin, A. Lalande, X. Zhuang, G. Yang, A. Young, and O. Bernard, pages 120–129, Springer International Publishing.

[10] O. Ronneberger, P. Fischer, and T. Brox, U-Net: Convolutional Networks for Biomedical Image Segmentation, in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, edited by N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, pages 234–241, Springer International Publishing.

[11] K. Kamnitsas, C. Ledig, V. F. J. Newcombe, J. P. Simpson, A. D. Kane, D. K. Menon, D. Rueckert, and B. Glocker, Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation, Medical Image Analysis **36**, 61–78 (2017).

[12] M. Havaei, A. Davy, D. Warde-Farley, A. Biard, A. Courville, Y. Bengio, C. Pal, P. M. Jodoin, and H. Larochelle, Brain tumor segmentation with Deep Neural Networks, Medical Image Analysis **35**, 18–31 (2017).

[13] P. Liskowski and K. Krawiec, Segmenting Retinal Blood Vessels With Deep Neural Networks, IEEE Transactions on Medical Imaging **35**, 2369–2380 (2016).

[14] Y. Lecun and Y. Bengio, *Convolutional networks for images, speech, and time-series*, MIT Press, 1995.

[15] Y. LeCun, Y. Bengio, and G. Hinton, Deep learning, Nature **521**, 436–444 (2015).

[16] K. He, X. Zhang, S. Ren, and J. Sun, Deep Residual Learning for Image Recognition, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016)*, pages 770–778.

[17] F. Milletari, N. Navab, and S. Ahmadi, V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation, in *2016 Fourth International Conference on 3D Vision (3DV)*, pages 565–571.

[18] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, UNet++: A Nested U-Net Architecture for Medical Image Segmentation, in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, edited by D. Stoyanov, Z. Taylor, G. Carneiro, T. Syeda-Mahmood, A. Martel, L. Maier-Hein, J. M. R. S. Tavares, A. Bradley, J. P. Papa, V. Belagiannis, J. C. Nascimento, Z. Lu, S. Conjeti, M. Moradi, H. Greenspan, and A. Madabhushi, pages 3–11, Springer International Publishing, 2018.

[19] M. D. Zeiler and R. Fergus, Visualizing and Understanding Convolutional Networks, in *European Conference on Computer Vision – ECCV 2014*, edited by D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, pages 818–833, Springer International Publishing.

[20] K. He, X. Zhang, S. Ren, and J. Sun, Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence **37**, 1904–1916 (2015).

[21] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, SSD: Single Shot MultiBox Detector, in *European Conference on Computer Vision 2016*, edited by B. Leibe, J. Matas, N. Sebe, and M. Welling, pages 21–37, Springer International Publishing.

[22] T. Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, Feature Pyramid Networks for Object Detection, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017)*, pages 936–944.

[23] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, Pyramid Scene Parsing Network, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017)*, pages 6230–6239.

[24] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully

Connected CRFs, IEEE Transactions on Pattern Analysis and Machine Intelligence **40**, 834–848 (2018).

25 J. Luo, K. A. Larson, D. E. Fenner, J. A. Ashton-Miller, and J. O. L. DeLancey, Posterior vaginal prolapse shape and position changes at maximal Valsalva seen in 3-D MRI-based models, International Urogynecology Journal **23**, 1301–1306 (2012).

26 J. Luo, J. A. Ashton-Miller, and J. O. L. DeLancey, A model patient: Female pelvic anatomy can be viewed in diverse 3-dimensional images with a new interactive tool, American Journal of Obstetrics and Gynecology **205**, 391.e1–391.e2 (2011).

27 V. Badrinarayanan, A. Kendall, and R. Cipolla, SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence **39**, 2481–2495 (2017).

28 F. Chollet, Xception: Deep Learning with Depthwise Separable Convolutions, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017)*, pages 1800–1807.

29 S. Osher and J. A. Sethian, Fronts propagating with curvature-dependent speed: Algorithms based on Hamilton-Jacobi formulations, Journal of Computational Physics **79**, 12–49 (1988).

30 P. Márquez-Neila, L. Baumela, and L. Alvarez, A Morphological Approach to Curvature-Based Evolution of Curves and Surfaces, IEEE Transactions on Pattern Analysis and Machine Intelligence **36**, 2–17 (2014).

31 S. van der Walt, J. L. Schönberger, J. Nunez-Iglesias, F. Boulogne, J. D. Warner, N. Yager, E. Gouillart, and T. Yu, scikit-image: image processing in Python, PeerJ **2**, e453 (2014).

32 O. Bernard et al., Deep Learning Techniques for Automatic MRI Cardiac Multi-Structures Segmentation and Diagnosis: Is the Problem Solved?, IEEE Transactions on Medical Imaging **37**, 2514–2525 (2018).

33 L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation, in *European Confer-*

ence on Computer Vision – ECCV 2018, edited by V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, pages 833–851, Springer International Publishing.

[34] E. Shelhamer, J. Long, and T. Darrell, Fully Convolutional Networks for Semantic Segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence **39**, 640–651 (2017).

[35] S. Zhou, D. Nie, E. Adeli, Y. Gao, L. Wang, J. Yin, and D. Shen, Fine-Grained Segmentation Using Hierarchical Dilated Neural Networks, in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, edited by A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, and G. Fichtinger, pages 488–496, Springer International Publishing.

[36] D. Nie, Y. Gao, L. Wang, and D. Shen, ASDNet: Attention Based Semi-supervised Deep Networks for Medical Image Segmentation, in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, edited by A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, and G. Fichtinger, pages 370–378, Springer International Publishing.

[37] E. Bonmati, Y. Hu, N. Sindhwani, H. P. Dietz, J. D'hooge, D. Barratt, J. Deprest, and T. Vercauteren, Automatic segmentation method of pelvic floor levator hiatus in ultrasound using a self-normalizing neural network, Journal of Medical Imaging **5**, 1–8, 8 (2018).

[38] S. Wang, K. He, D. Nie, S. Zhou, Y. Gao, and D. Shen, CT male pelvic organ segmentation using fully convolutional networks with boundary sensitive representation, Medical Image Analysis **54**, 168–178 (2019).

[39] K. He, X. Cao, Y. Shi, D. Nie, Y. Gao, and D. Shen, Pelvic Organ Segmentation Using Distinctive Curve Guided Fully Convolutional Networks, IEEE Transactions on Medical Imaging **38**, 585–595 (2019).

[40] D. Nie, L. Wang, Y. Gao, J. Lian, and D. Shen, STRAINet: Spatially Varying sTochastic Residual AdversarIal Networks for MRI Pelvic Organ Segmentation, IEEE Transactions on Neural Networks and Learning Systems **30**, 1552–1564 (2019).

[41] N. Wang, Y. Wang, H. Wang, B. Lei, T. Wang, and D. Ni, Auto-context fully convolutional network for levator hiatus segmentation in ultrasound images, in *ISBI 2018*, pages 1479–1482.

[42] H. Y. Xia, W. F. Sun, S. X. Song, and X. W. Mou, Md-Net: Multi-scale Dilated Convolution Network for CT Images Segmentation, Neural Processing Letters **51**, 2915–2927 (2020).

[43] M. P. Shah, S. N. Merchant, and S. P. Awate, MS-Net: Mixed-Supervision Fully-Convolutional Networks for Full-Resolution Segmentation, in *Medical Image Computing and Computer Assisted Intervention - Miccai 2018*, edited by A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. AlberolaLopez, and G. Fichtinger, volume 11073 of *Lecture Notes in Computer Science*, pages 379–387, Cham, 2018, Springer.

[44] Y. M. Han, S. H. Zhang, Z. Q. Geng, W. Qin, and O. Y. Zhi, Level set based shape prior and deep learning for image segmentation, IET Image Processing **14**, 183–191 (2020).

615 # Figures

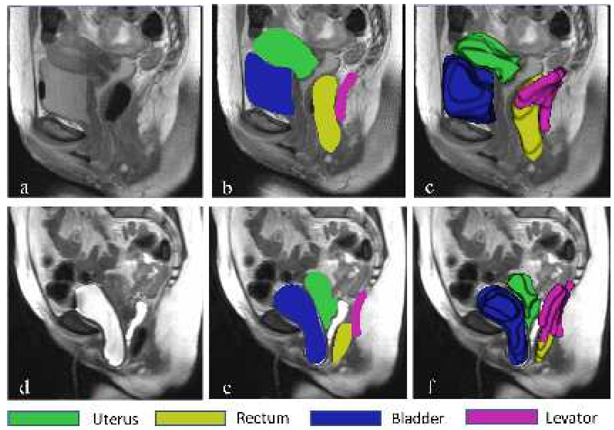


Figure 1: Left lateral views of a patient with anterior vaginal wall prolapse. (a and d) Midsagittal MR images at rest and at maximum Valsalva. (b and e) Similar images of the pelvic floor organs, including the uterus, rectum, bladder, and levator ani muscle, shown at rest and at maximum Valsalva. (c and f) Views of the 3D models of the pelvic floor organs.

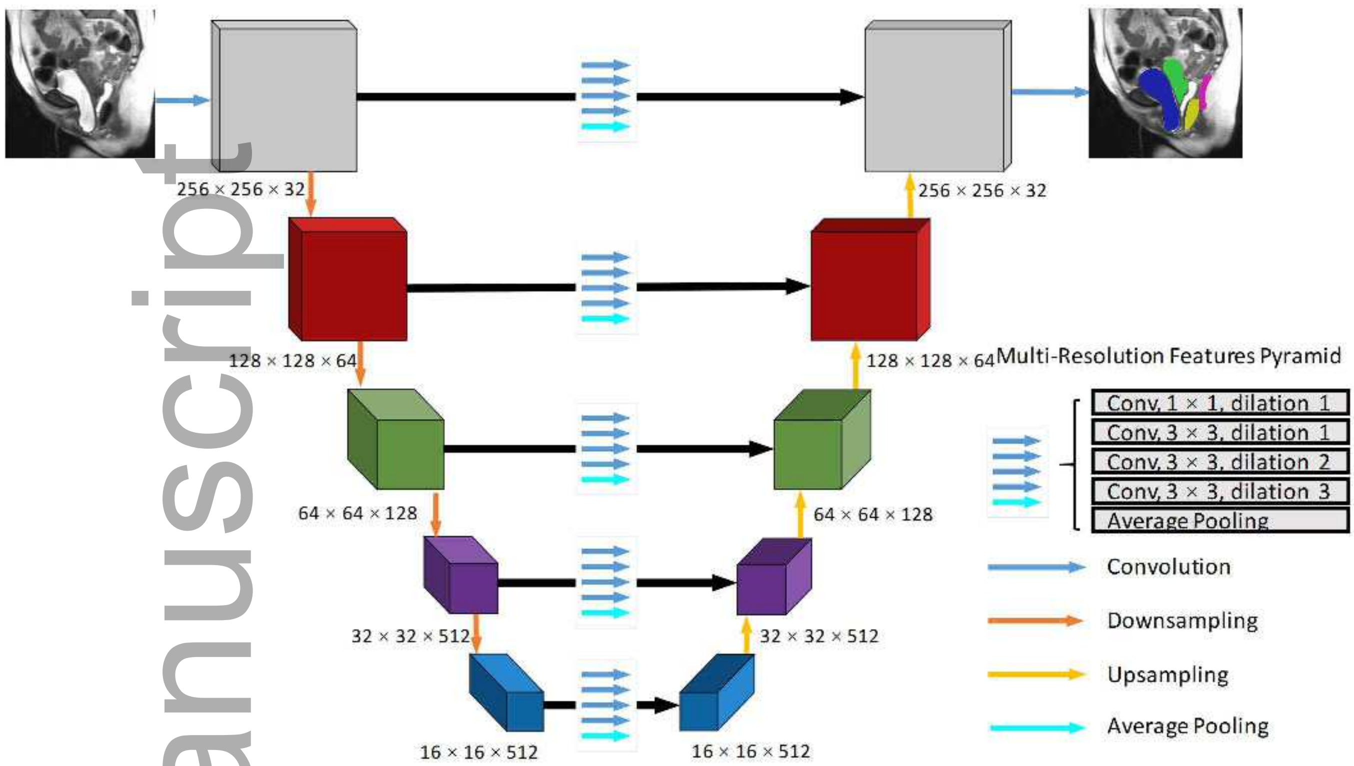


Figure 2: CNN model structure. Feature maps of skip connection and upsampling branches were combined using a concatenation method.

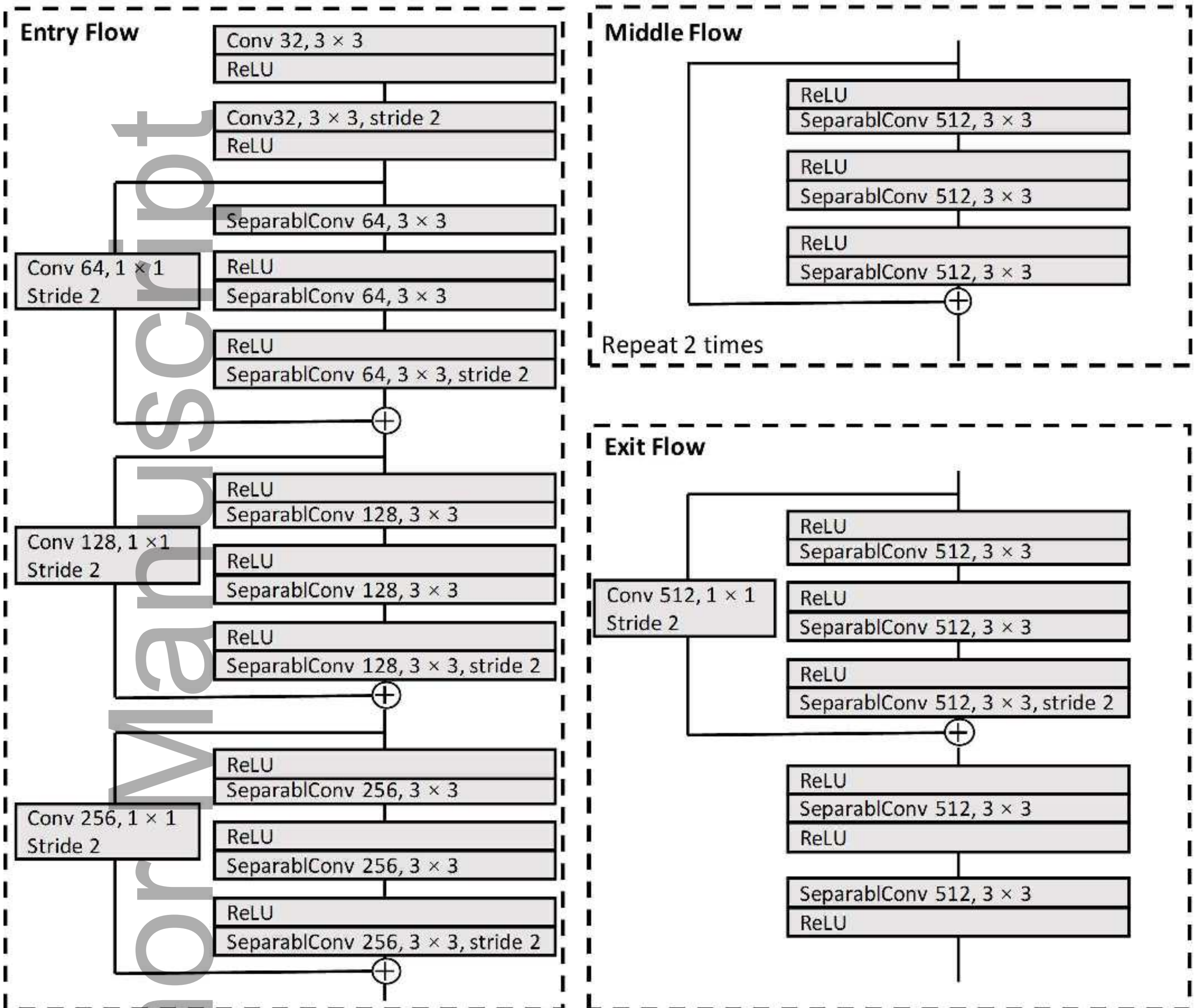


Figure 3: Diagram illustrating the structure of the encoder network.

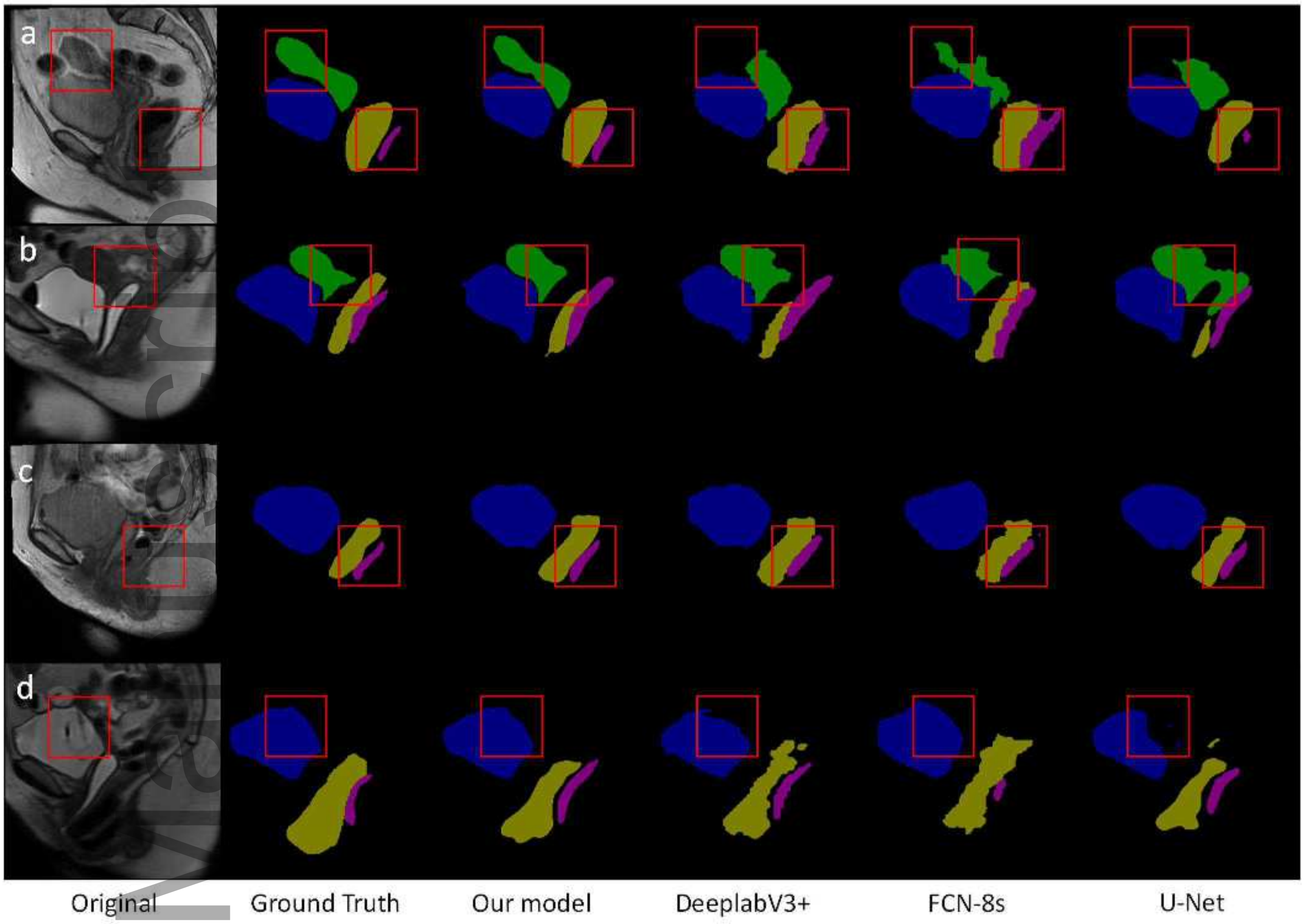


Figure 4: A comparison of segmentation results among our model, Deeplabv3+, FCN-8s, and U-Net . (a) Resting example with uterus. (b) Stressed example with uterus. (c) Resting example without uterus. (d) Stressed example without uterus. Results of different methods were compared with the ground truth labeling.

Figure caption

- Figure 1. Left lateral views of a patient with anterior vaginal wall prolapse. (a and d) Midsagittal MR images at rest and at maximum Valsalva. (b and e) Similar images of the pelvic floor organs, including the uterus, rectum, bladder, and levator ani muscle, shown at rest and at maximum Valsalva. (c and f) Views of the 3D models of the pelvic floor organs.

- Figure 2. CNN model structure. Feature maps of skip connection and upsampling branches were combined using a concatenation method.

- Figure 3. Diagram illustrating the structure of the encoder network.

- Figure 4. A comparison of segmentation results among our model, Deeplabv3+, FCN-8s, and U-Net . (a) Resting example with uterus. (b) Stressed example with uterus. (c) Resting example without uterus. (d) Stressed example without uterus. Results of different methods were compared with the ground truth labeling.

- Figure 5. Examples of re-segmentation results using the level set method. (a and b) Uterus re-segmentation. (c and d) Rectum re-segmentation, (e and f) Bladder re-segmentation. The composite results were obtained by replacing the models' predictions with the level set results on the corresponding organ. Results were compared with the ground truth labeling.

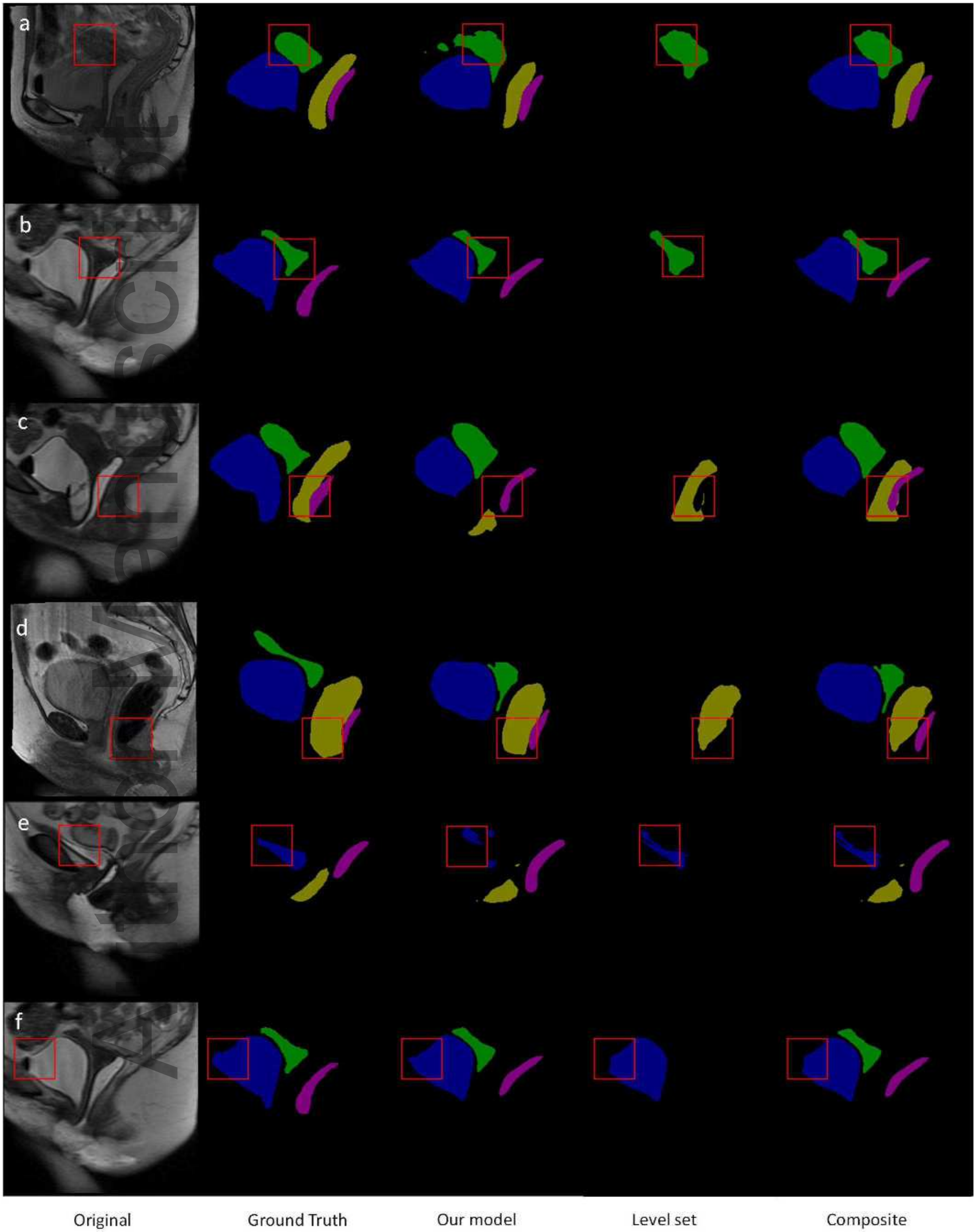


Figure 5: Examples of re-segmentation results using the level set method. (a and b) Uterus re-segmentation. (c and d) Rectum re-segmentation, (e and f) Bladder re-segmentation. The

Tables

Table 1: Organ occurrence rate in training data.

Organ	Uterus	Rectum	Bladder	Levator
Number of occurrence	103	152	197	112
Number of total images	256	256	256	256
Presence rate	0.40	0.59	0.77	0.44

Table 2: Model performance comparison using different loss functions. Units: DSC in %, and ASD in mm. (+) means with pre-training, and (*) means without pre-training. Number in the () is the standard deviation.

Methods	Uterus		Rectum		Bladder		Levator		Average	
	DSC	ASD	DSC	ASD	DSC	ASD	DSC	ASD	DSC	ASD
DL(+)	**55.0 (9.3)**	**5.2 (1.7)**	**64.1 (17.6)**	**2.5 (1.3)**	82.7 (16.5)	1.6 (0.5)	**60.8 (7.4)**	**2.3 (1.4)**	**65.6**	**2.9**
DL(*)	53.5 (18.3)	6.6 (4.6)	62.0 (17.9)	2.7 (1.1)	**84.8 (10.0)**	1.6 (0.5)	55.6 (9.4)	3.6 (2.5)	64.0	3.6
CE(+)	37.3 (9.8)	7.8 (2.6)	57.7 (21.6)	3.3 (1.7)	84.5 (11.7)	1.6 (0.5)	50.6 (8.3)	10.1 (11.1)	57.6	5.7
CE(*)	40.4 (19.4)	10.8 (5.6)	56.4 (16.7)	3.5 (1.4)	84.4 (10.9)	**1.6 (0.3)**	45.3 (13.7)	10.1 (13.6)	56.6	6.5

Table 3: Models' performance comparison with other advanced segmentation methods. Units: DSC in %, and ASD in mm. (+) means with pre-training, and (*) means without pre-training. Number in the () is the standard deviation.

Methods	Uterus		Rectum		Bladder		Levator		Average	
	DSC	ASD	DSC	ASD	DSC	ASD	DSC	ASD	DSC	ASD
Proposed (+)	**55.0 (9.3)**	**5.2 (1.7)**	64.1 (17.6)	2.5 (1.3)	82.7 (16.5)	**1.6 (0.5)**	**60.8 (7.4)**	**2.3 (1.4)**	**65.6**	**2.9**
Proposed (*)	53.5 (18.3)	6.6 (4.6)	62.0 (17.9)	2.7 (1.1)	**84.8 (10.0)**	**1.6 (0.5)**	55.6 (9.4)	3.6 (2.5)	64.0	3.6
Deeplabv3+	45.0 (11.2)	7.3 (3.3)	58.9 (16.8)	3.0 (1.2)	83.3 (10.7)	1.9 (0.5)	53.4 (13.1)	3.8 (2.0)	60.2	4.0
FCN-8s	39.8 (14.9)	6.9 (4.8)	**65.6 (11.7)**	**2.5 (1.0)**	80.0 (13.7)	1.9 (0.7)	47.4 (16.9)	9.5 (11.6)	58.2	5.2
U-Net	45.0 (16.0)	14.3 (7.6)	42.0 (27.2)	4.7 (2.6)	77.2 (23.2)	2.9 (2.2)	54.6 (11.0)	5.2 (5.6)	54.8	6.8

Table 4: Models' performance comparison using other metrics. Units: ODR in %, and RAVD in %. (+) means with pre-training, and (*) means without pre-training. Number in the () is the standard deviation.

Methods	Uterus		Rectum		Bladder		Levator		Average	
	ODR	RAVD	ODR	RAVD	ODR	RAVD	ODR	RAVD	ODR	RAVD
Proposed (+)	84.8 (16.2)	**34.5 (14.9)**	**100 (0.0)**	41.0 (31.6)	91.6 (4.6)	10.8 (8.8)	**95.1 (5.8)**	**19.9 (16.0)**	**92.9**	**26.6**
Proposed (*)	84.5 (14.1)	43.3 (7.0)	94.6 (8.0)	37.6 (21.0)	98.0 (4.4)	8.6 (5.0)	91.6 (8.6)	22.2 (19.4)	92.1	28.0
Deeplabv3+	87.2 (15.7)	52.7 (35.6)	94.5 (8.0)	27.8 (16.4)	94.7 (5.4)	**6.2 (5.4)**	80.1 (16.2)	30.4 (16.3)	89.1	29.3
FCN-8s	84.1 (14.9)	61.9 (34.9)	96.7 (7.5)	**21.0 (8.2)**	**98.0 (2.8)**	11.9 (17.9)	91.7 (8.6)	52.8 (29.3)	92.6	36.9
U-Net	**94.2 (9.9)**	47.5 (33.5)	82.5 (21.6)	55.2 (26.7)	90.6 (5.7)	23.6 (24.2)	91.4 (8.9)	20.5 (16.3)	90.0	36.6

Table 5: Ablation study results. Units: DSC in %, and ASD in mm. Proposed model is the U-Net-XM_{12345}. (*) means without pre-training. Number in () is the standard deviation.

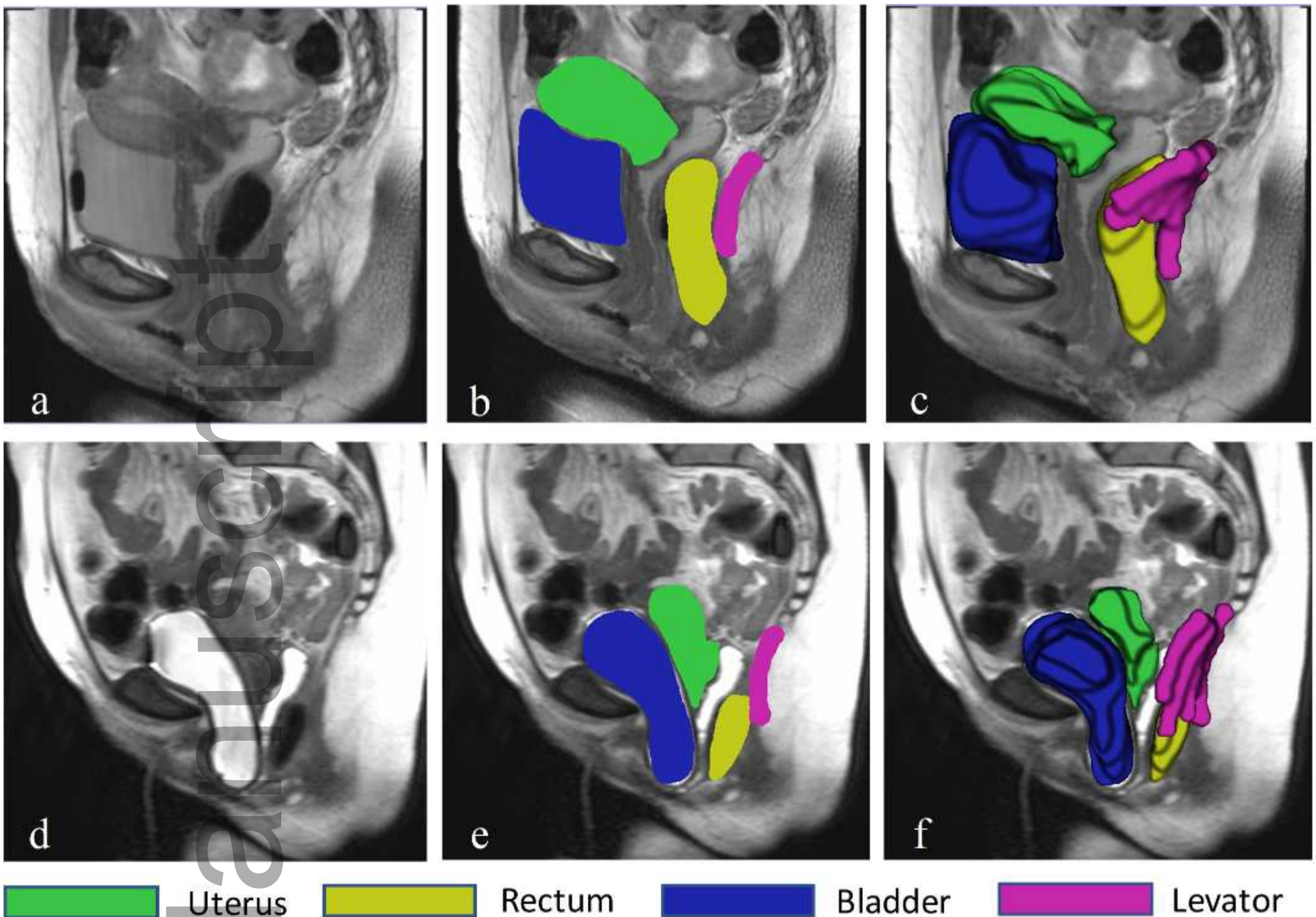
Methods	Uterus		Rectum		Bladder		Levator		Average	
	DSC	ASD	DSC	ASD	DSC	ASD	DSC	ASD	DSC	ASD
Proposed (*)	**53.5 (18.3)**	**6.6 (4.6)**	62.0 (17.9)	**2.7 (1.1)**	**84.8 (10.0)**	**1.6 (0.5)**	**55.6 (9.4)**	3.6 (2.5)	**64.0**	**3.6**
U-Net-M	49.5 (11.3)	9.8 (6.8)	**63.6 (16.6)**	3.3 (2.2)	79.4 (18.4)	2.0 (0.8)	52.8 (10.0)	6.8 (10.0)	61.6	4.9
U-Net-X	41.2 (13.4)	11.0 (6.5)	63.4 (14.8)	2.9 (1.5)	76.1 (28.3)	2.9 (2.8)	54.2 (8.0)	**3.3 (1.8)**	58.7	5.0
U-Net	45.0 (16.0)	14.3 (7.6)	42.0 (27.2)	4.7 (2.6)	77.2 (23.2)	2.9 (2.2)	54.6 (11.0)	5.2 (5.6)	54.8	6.8

Table 6: Models' performance comparison for different MRFP configurations. Units: DSC in %, ASD in mm. Proposed model is the U-Net-XM$_{12345}$. (*) means without pre-training. Number in () is the standard deviation.

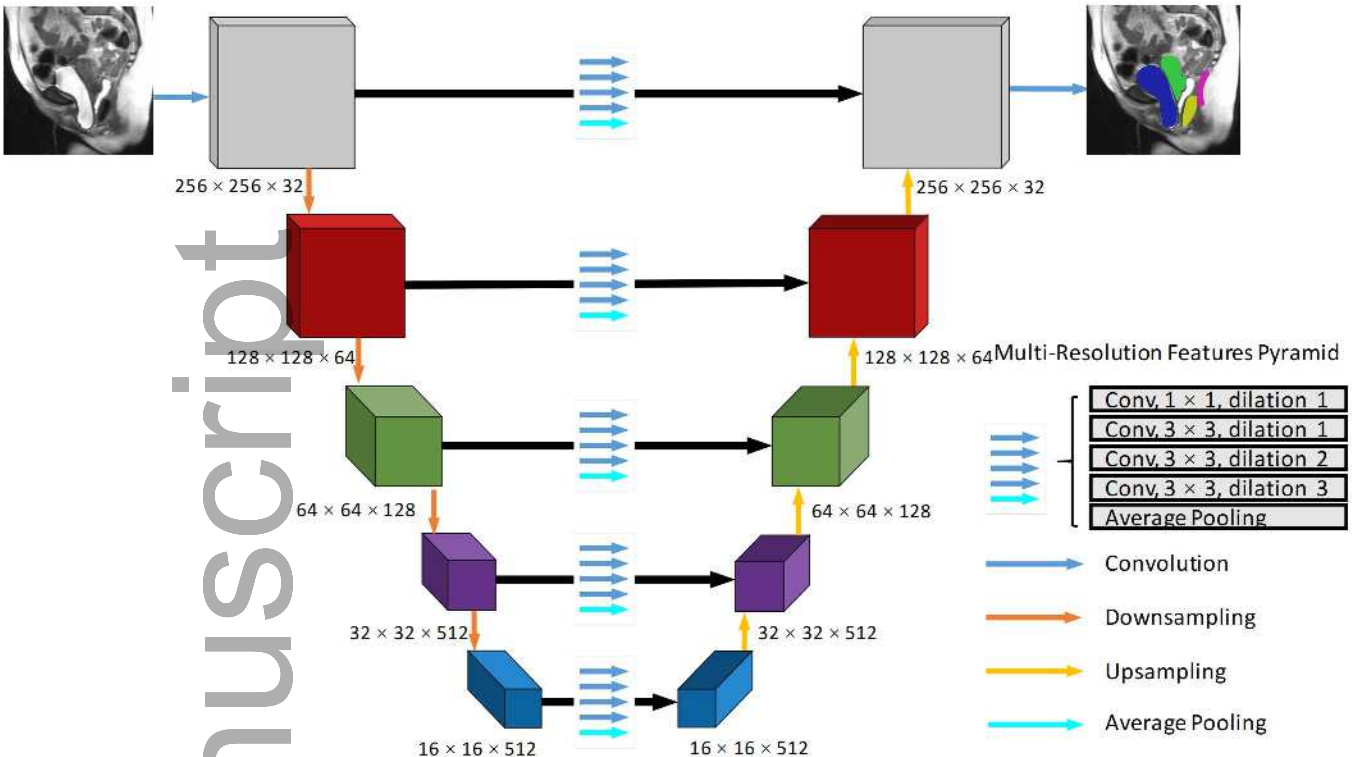
Methods	Uterus		Rectum		Bladder		Levator		Average	
	DSC	ASD	DSC	ASD	DSC	ASD	DSC	ASD	DSC	ASD
Proposed (*)	**53.5 (18.3)**	**6.6 (4.6)**	62.0 (17.9)	2.7 (1.1)	**84.8 (10.0)**	**1.6 (0.5)**	**55.6 (9.4)**	3.6 (2.5)	64.0	**3.6**
U-Net-XM$_{123}$	50.9 (14.0)	8.3 (6.5)	**66.1 (12.4)**	**2.6 (1.0)**	83.7 (12.1)	2.0 (0.7)	45.8 (11.0)	4.7 (3.3)	61.6	4.4
U-Net-XM$_{135}$	54.1 (10.6)	7.5 (5.6)	65.6 (14.5)	2.8 (1.4)	84.6 (12.5)	1.6 (0.7)	52.3 (9.6)	6.5 (4.3)	**64.2**	4.6
U-Net-XM$_{345}$	53.6 (16.2)	10.8 (6.3)	65.5 (12.8)	3.5 (2.2)	84.8 (11.8)	1.6 (0.7)	52.6 (12.1)	**3.3 (1.8)**	64.1	4.7
U-Net-X	41.2 (13.4)	11.0 (6.5)	63.4 (14.8)	2.9 (1.5)	76.1 (28.3)	2.9 (2.8)	54.2 (8.0)	**3.3 (1.8)**	58.7	5.0

Table 7: Model performance comparison after using the level set method. Units: DSC in %, and ASD in mm. (+) means with pre-training, and (*) means without pre-training. Number in () is the standard deviation.

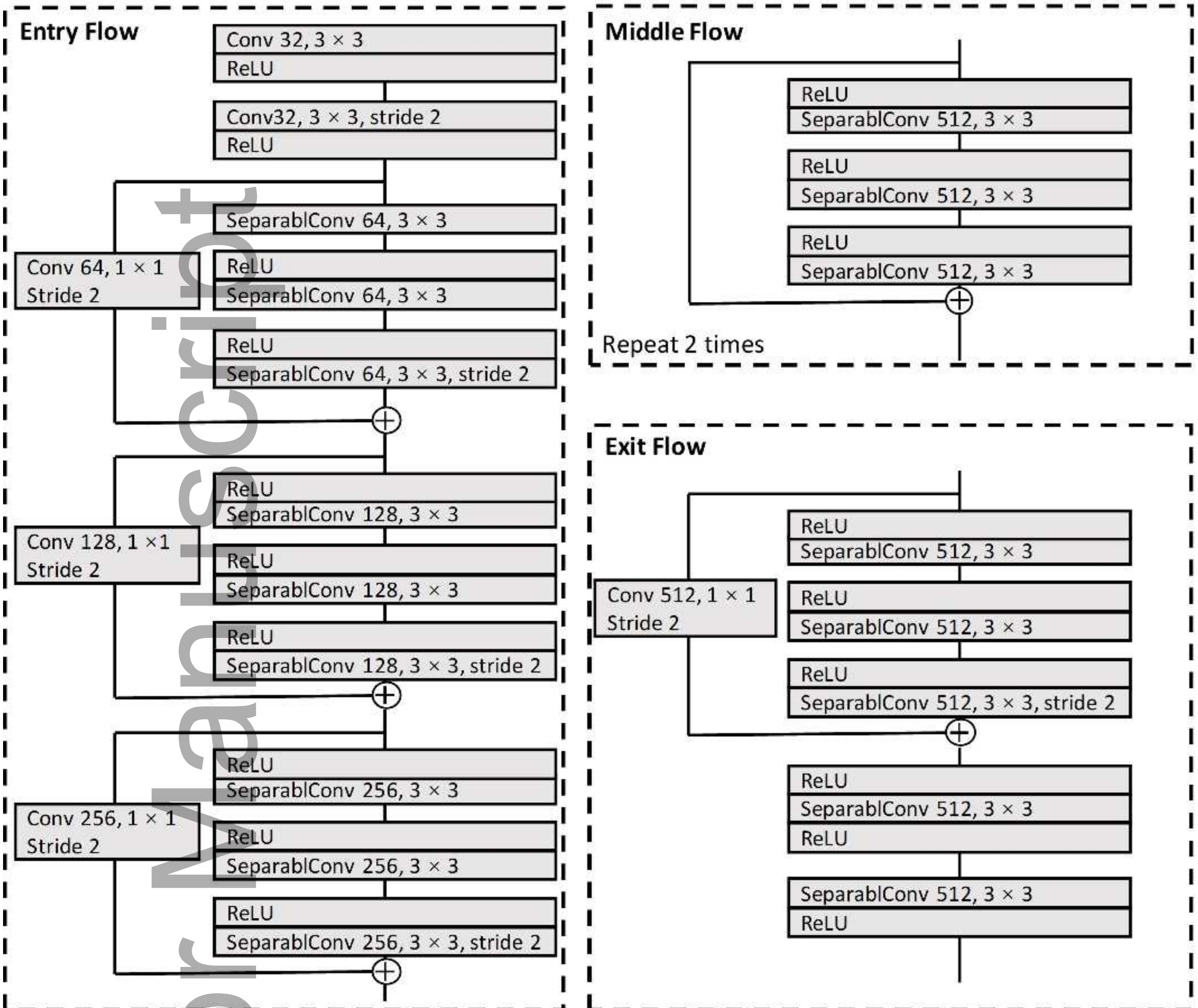
Methods	Uterus		Rectum		Bladder		Levator		Average	
	DSC	ASD	DSC	ASD	DSC	ASD	DSC	ASD	DSC	ASD
Proposed (+)	**65.3 (3.8)**	**5.4 (1.9)**	**66.3 (15.0)**	**2.1 (0.9)**	**85.6 (10.0)**	**1.6 (0.4)**	**60.8 (7.4)**	**2.3 (1.4)**	**69.4**	**2.9**
Proposed (*)	58.3 (18.6)	6.6 (5.1)	65.8 (14.4)	2.4 (1.1)	84.8 (10.1)	1.7 (0.5)	55.6 (9.4)	3.6 (2.5)	66.1	3.6
Deeplabv3+	52.0 (14.8)	9.2 (5.5)	59.8 (17.0)	3.3 (2.0)	83.3 (10.7)	1.9 (0.5)	53.4 (13.1)	3.8 (2.0)	62.1	4.5
FCN-8s	46.0 (18.3)	8.3 (6.2)	66.0 (11.8)	2.3 (0.8)	80.7 (12.2)	1.9 (0.5)	47.4 (16.9)	9.5 (11.6)	60.0	5.6
U-Net	47.6 (15.0)	11.8 (16.3)	47.6 (22.8)	7.3 (4.4)	80.8 (15.5)	2.6 (1.2)	54.6 (11.0)	5.2 (5.6)	56.4	5.3



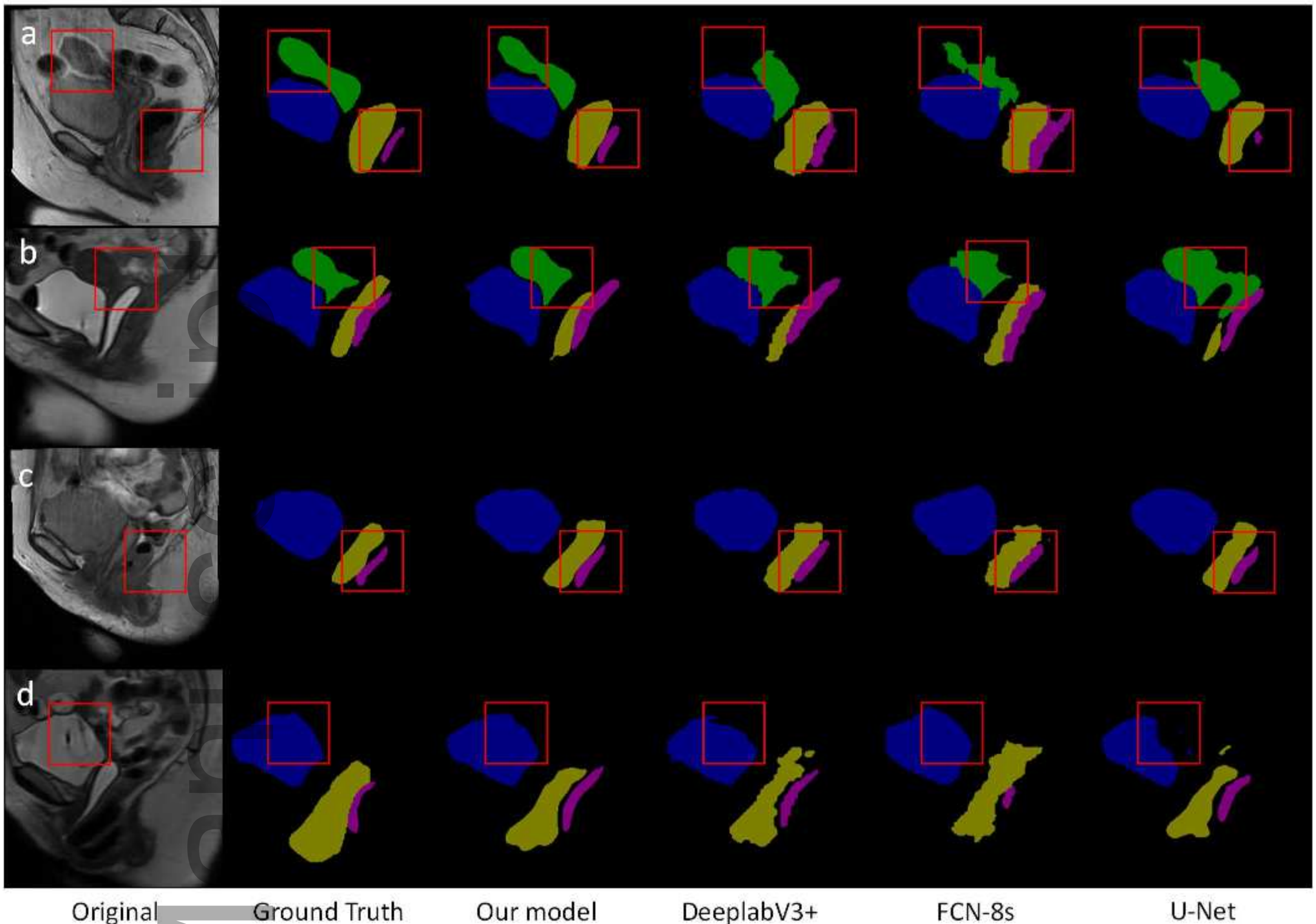
mp_14377_f1.jpg



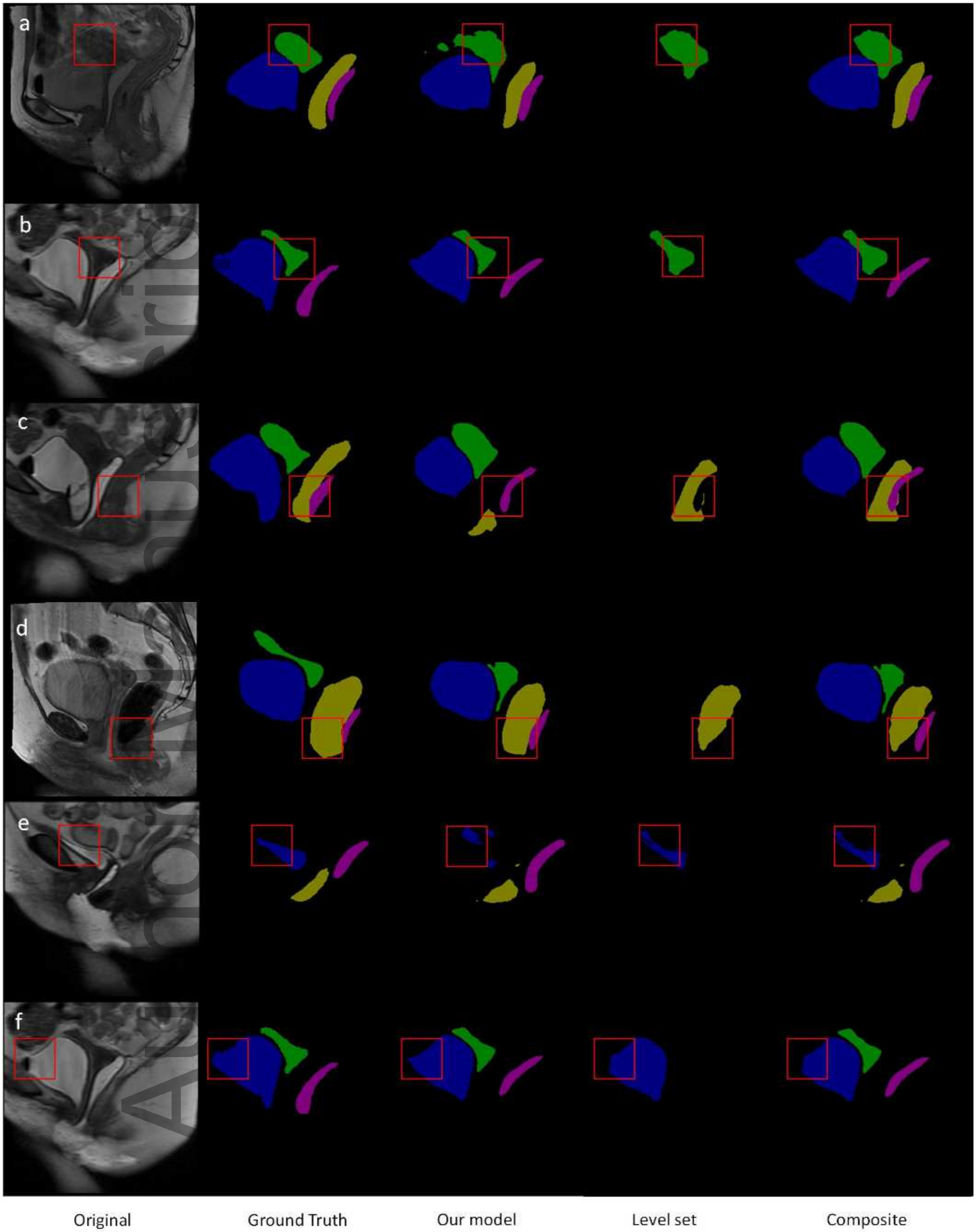
mp_14377_f2.jpg



mp_14377_f3.jpg



mp_14377_f4.jpg



mp_14377_f5.jpg