Published in final edited form as:

IEEE Trans Neural Netw Learn Syst. 2019 May ; 30(5): 1552–1564. doi:10.1109/TNNLS.2018.2870182.

STRAINet: Spatially-varying sTochastic Residual Adversarial Networks for MRI Pelvic Organ Segmentation

Dong Nie,

Department of Computer Science, Department of Radiology and BRIC, UNC-Chapel Hill.

Li Wang,

Department of Radiology and BRIC, UNC-Chapel Hill.

Yaozong Gao,

Shanghai United Imaging Intelligence Co., Ltd.

Jun Lian, and

Department of Radiation Oncology, UNC-Chapel Hill.

Dinggang Shen* [Fellow, IEEE]

Department of Radiology and BRIC, UNC-Chapel Hill, and also with the Department of Brain and Cognitive Engineering, Korea University, Seoul 02841, Republic of Korea;

Abstract

Accurate segmentation of pelvic organs is important for prostate radiation therapy. Modern radiation therapy starts to use magnetic resonance image (MRI) as an alternative to CT image because of its superior soft tissue contrast and also free of risk from radiation exposure. However, segmentation of pelvic organs from MRI is a challenging problem due to inconsistent organ appearance across patients and also large intra-patient anatomical variations across treatment days. To address such challenges, we propose a novel deep network architecture, called “Spatially-varying sTochastic Residual Adversarial Network” (STRAINet), to delineate pelvic organs from MRI in an end-to-end fashion. Compared to the traditional fully convolutional networks (FCN), the proposed architecture has two main contributions. 1) Inspired by the recent success of residual learning, we propose an evolutionary version of residual unit, i.e., stochastic residual unit, and use it to the plain convolutional layers in the FCN. We further propose long-range stochastic residual connections to pass features from shallow layers to deep layers. 2) We propose to integrate three previously proposed network strategies to form a new network for better medical image segmentation: a) We apply dilated convolution in the smallest resolution feature maps, so that we can gain larger receptive field without overly losing spatial information; b) We propose a spatially-varying convolutional layer that adapts convolutional filters to different regions of interest; c) An adversarial network is proposed to further correct the segmented organ structures. Finally, STRAINet is used to iteratively refine the segmentation probability maps in an auto-context manner. Experimental results show that our STRAINet achieved the state-of-the-art segmentation

accuracy. Further analysis also indicates that our proposed network components contribute most to the performance.

Index Terms—

Stochastic Residual Learning; Pelvic Organ Segmentation; Adversarial Learning; Dilation

I. INTRODUCTION

Prostate cancer is one of the most common cancers among American men. Magnetic resonance imaging (MRI) plays an emerging role in prostate cancer diagnosis and treatment as it can provide superior soft-tissue contrast compared to other modalities (e.g., CT and ultrasound) in a non-invasive and radiation-free way. Accurate segmentation of the prostate and its surrounding pelvic organs is critical in both guiding biopsy and cancer radiation therapy. Nowadays, manual segmentation of pelvic organs is still heavily adopted in clinical practice, which is time-consuming and often suffers from large intra- and inter-observer variability [1]. Therefore, it is highly desired to develop an accurate and reliable method to automatically segment pelvic organs from MRI.

The challenges of automatic pelvic organ segmentation in MRI come from three aspects. 1) Appearance patterns of pelvic organs vary largely across different patients. 2) Different regions of pelvic organs also exhibit inhomogeneous image contrast. 3) The shapes of pelvic organs (e.g., the bladder and rectum) could change significantly across different patients. Fig. 1 illustrates these challenges.

The current automatic methods for segmentation of pelvic organs in MRI can be categorized into three mainstreams, i.e., 1) multi-atlas-based methods [2], [3], [4], [5], 2) deformable-model-based methods [6], [7], [8], [9], and 3) learning-based methods [10], [11], [12], [13]. In 1), multiple images with expert segmentations, namely atlases, are first rigidly or non-rigidly aligned onto a target image under segmentation. Then, the aligned segmentation labels from different atlases are fused to obtain the segmentation of target image. Most atlas-based segmentation algorithms concentrate on the design of sophisticated atlas selection or the label fusion mechanism. For instance, Yan et al. [5] proposed an atlas selection with a label constraining and label fusion method to segment the prostate MRI. During the atlas selection, label images are used to constrain the manifold projection of intensity images, which can alleviate the misleading projection due to surrounding anatomical structures. Ou et al. [4] proposed to gradually improve the registration based on the prostate vicinity between the target and atlas images for iteratively carrying out the multi-atlas label fusion. In 2), the segmentation is formulated as an optimization problem to find the best shape model for fitting the target image. Both categories of methods need to define image-to-image similarity, which often requires careful feature engineering. Toth et al. [6] proposed to incorporate different features in the context of active appearance models (AAMs) to improve the prostate segmentation performance. In 3), Sciolla et al. [10] proposed a multi-pass 3D convolutional neural network to segment the prostate MRI. Zhu et al. [12] introduced deep supervision mechanism to the Unet-like structure for prostate MRI segmentation and

achieved reasonable performance. Milletari et al. [11] developed Vnet for MRI- based prostate segmentation using a Dice loss. Yu et al. [13] further introduced residual learning module to the Unet-like structures and achieved a very high score in the PROMISE12 challenge¹. Besides the above-mentioned methods, some unsupervised learning based methods are also used to segment prostate from MRI [14]. For example, Tian et al. [14] proposed a graph cut algorithm to obtain the prostate surface. Moreover, various supervised learning based (or example-based) methods are also proposed to segment prostate from MRI [15],[3]. It is worth noting that some of these methods usually extract self-designed features (e.g., Haar-like features [16]) from the MRI, and then apply suitable classifiers on the extracted features to complete the segmentation tasks. With the success of deep learning in computer vision and medical image fields [17], [18], many deep learning based approaches have been proposed to learn suitable feature representation for segmentation of prostate in MRI [19], [20], [21], [13], with promising performance. For instance, Guo et al. [20] proposed to learn effective features from data by the stacked auto-encoders, and achieved more accurate segmentations than the traditional methods using handcraft features. However, their feature learning stage is not optimized directly for the segmentation, which limits the effectiveness of those learned features. In addition, their methods were developed only for the segmentation of prostate, and thus may not work well for other pelvic organs (e.g., the bladder and rectum) due to different appearances and shape characteristics. Fully convolutional networks (FCN) [22], a variant of convolutional neural networks (CNN), is a recently common choice for semantic image segmentation in computer vision. FCN trains a neural network in an end-to-end fashion by directly optimizing intermediate feature layers for segmentation, which makes it outperform the traditional methods that often regard the feature learning and segmentation as two separate tasks. Following that, lots of works have been proposed to improve the FCN [23], [24], [25], [26], [27], [28]. For example, Li et al. [23] proposed contrast-oriented fully convolutional neural networks for dense prediction and salient object detection. In the ‘Deeplab’ [25], Chen et al. implemented an atrous spatial pyramid pooling module to increase the context information in a multi-scale manner and applied Dense Conditional Random Fields (Dense-CRF) [27] to refine the segmentation results. Lin et al. [29] introduced a well-designed encoder-decoder architecture to fuse the high-resolution feature maps from encoder pathway and the highly semantic feature maps from decoder pathway in the ‘RefineNet’. In the recent ‘PSPNet’, Zhao et al. [28] proposed a pyramid pooling module to aggregate the background information and auxiliary loss to intermediately supervise the segmentation task. FCN- based methods have also shown great success in medical image segmentation [11], [30], [31], [32], [33], [13], [34]. On the other hand, Unet [35], an evolutionary variant of FCN, has also achieved excellent performance by effectively combining high-level and low-level features in the network architecture. Compared to FCN, Unet can improve the localization accuracy near organ boundaries. Generally, while being effective, the performance of FCN (or Unet) is still limited by three factors. 1) The deep network architecture makes FCN (or Unet) difficult to train because of the gradient vanishing problem. 2) It is always difficult to tradeoff between receptive field and spatial information, since enlarging receptive field usually means losing more spatial

¹<https://promise12.grand-challenge.org/>

information, and vice versa. 3) Sharing convolutional filters limits the performance due to ignorance of large appearance variations across spatial domains.

In this paper, we propose a novel network architecture, called “Spatially-varying sTochastic Residual Adversarial Networks” (STRAINet), to overcome the limitations of FCN (or Unet). Specifically, STRAINet has two major contributions over FCN (or Unet):

1. We replace all the plain convolutional layers in FCN with the proposed stochastic residual units (SRU). Also, with short and long skip connections in stochastic residual units, the gradient vanishing problem can be largely alleviated and thus the training can often converge faster than the networks without skip connections [36]; importantly, the overfitting phenomenon can also be largely mitigated by using the proposed SRU. To the best of our knowledge, stochastic residual connection is the first introduced for neural networks.
2. We propose to integrate three previously proposed network strategies to form a new network for better medical image segmentation.
 - a. Considering that pooling could result in spatial information loss, we propose to conduct dilated convolution only on the smallest resolution feature maps for gaining larger receptive field at the lowest cost of memory.
 - b. Given the fact that many medical images can just be roughly aligned before segmentation, we propose to use spatially-varying convolutional operations [37] in our proposed network. The proposed spatially-varying convolutions break the filter-sharing nature of traditional convolutions, and allow learning different filters at different spatial locations.
 - c. We further propose to use adversarial learning strategy to further correct the segmented organs, which is proved to be effective by our conducted experiments.

To better segment pelvic organs in MRI, we further use STRAINet as a basic unit in the auto-context model [38] to progressively refine the segmentation probability maps. Extensive experiments on 50 pelvic MRI show significant improvement with the above mentioned novel contributions.

II. METHODS

As mentioned in the introduction, the proposed STRAINet was inspired by FCN [22], Unet [35] and ResNet [36]. Accordingly, in the following subsections, we will first recap the network architectures of FCN, Unet and ResNet. Then, we will elaborate the design of the Segmenter (“Spatially-varying sTochastic Residual Networks”, called STRNet for convenience) as well as other components in our segmentation architecture, and further describe the adversarial learning over STRNet. Finally, we describe the implementation details, such as imbalance data processing and refinement with Auto-Context Model (ACM) [38]. We illustrate the two main networks of our proposed STRAINet in Fig. 2. The

‘Segmenter’ network works as a segmentation network which outputs a predicted mask by segmenting the input MRI, and the ‘Discriminator’ network works as the adversarial learning strategy to further improve the segmentation network. Details will be introduced in the following sections.

A. Stochastic Residual Unit for the FCN

FCN [22] has been widely adopted in various semantic segmentation tasks and achieved superior performance. FCN consists of two stages: 1) context understanding by using convolutions and pooling to condense the input image into high-level feature maps, and 2) label inference under the guidance from high-level feature maps. FCN trains an end-to-end CNN network for semantic image segmentation. Since both feature extractors and classifiers are embedded in the same CNN network, FCN is able to optimize intermediate feature layers directly for semantic image segmentation. This characteristic distinguishes FCN from other segmentation works that often learn features and classifiers separately (e.g., [20]). Thus, we adopt FCN as our basic architecture for the segmentation task (as shown in Fig. 3(a)).

The training dataset is composed of N pairs of input patches (i.e., slices from the MRI in our experiment) $\mathbf{X}^{(i)}$ and their corresponding label patches $\mathbf{Y}^{(i)}$, $i = 1, 2, \dots, N$. We suppose the patch size to be $[H, W]$, and the number of categories to be C , which means $\mathbf{X}$ is a $N \times C \times H \times W$ matrix and $\mathbf{Y}$ is a $N \times H \times W$ matrix. Then, the training objective can be formulated by the cross entropy loss, as in Eq. 1

$$L_{CE}(\mathbf{X}, \mathbf{Y}, \theta) = -\frac{1}{N} \left[\sum_{i=1}^N \sum_{p=1}^H \sum_{q=1}^W L'_{CE}(\mathbf{X}_{pq}^{(i)}, \mathbf{Y}_{pq}^{(i)}, \theta) \right] \quad (1)$$

$$L'_{CE}(\mathbf{x}, y, \theta) = \sum_{j=1}^C I(y, j) \left[\log \frac{e^{\theta_j^T \mathbf{x}}}{\sum_{l=1}^C e^{\theta_l^T \mathbf{x}}} \right] \quad (2)$$

$$I(y, j) = \begin{cases} 1, & y = j \\ 0, & y \neq j \end{cases} \quad (3)$$

where $\mathbf{x}$ and y denotes the data and label of one pixel in the patch, respectively, and the network parameters are represented by θ .

Very deep networks usually increase the discriminative ability [36]; however, their optimizations are often challenging due to the use of deep network architecture and complex network design [36], [39]. Residual learning has been proposed to better train the very deep networks, and achieved great success in both computer vision and medical image analysis

tasks [36],[13]. In this work, inspired by the bottleneck residual architecture from the recently popular residual networks ('ResNet')[36], we propose a stochastic version of residual connection to improve the plain convolutional layers in FCN.

Stochastic Residual Connection for FCN: The residual unit is a basic component in the residual learning [36], and it can be mathematically formulated as follows:

$$\mathbf{Z} = F(\mathbf{A}, \{\theta_i\}) + \mathbf{A} \quad (4)$$

where $\{\theta_i\}$ is the set of convolutional filters in the bottleneck residual unit, and $\mathbf{A}$ and $\mathbf{Z}$ are the input and output feature maps, respectively.

As demonstrated in [36], by introducing the skip connections, the bottleneck residual units can relieve the gradient vanishing problem and improve the convergence speed, especially for the very deep networks. Veit et al. [40] further analyzed that the residual connection behaved like ensembles of relatively shallow networks, which means the success of ResNet may be attributed to the ensemble models. Huang et al. [41] proposed using stochastic depth to decrease the training time of the deep residual networks without decreasing the testing performance. As a consequence, the residual module will be more powerful if we can enhance the ensemble behavior of the residual learning module. In this work, we propose a stochastic residual connection inspired by the efficient regularization 'Dropout' [42], the success of which can also be explained by its ensemble behaviors, to enhance the ensemble extent of the residual learning module. The proposed stochastic residual unit can be explained in Eq. 5.

$$\mathbf{Z} = F(\mathbf{A}, \{\theta_i\}) + \mathbf{R} \odot \mathbf{A} \quad (5)$$

where $\odot$ is element-wise multiplication, and $\mathbf{R}$ is a matrix of independent Bernoulli random variables and each element r has a probability p (determined experimentally as $p = 0.01$) to be set to 0:

$$r \sim \text{Bernoulli}(p) \quad (6)$$

Following Eq. 5, the residual connections are randomly built, which means part of the identity mappings are ignored at a random choice. Fig. 3(b) gives the visualization of a typical stochastic residual unit. Each plain convolutional layer can be replaced with the proposed stochastic residual unit, as shown in Fig. 3. Therefore, the proposed model efficiently ensembles exponentially many different residual connections, which can prevent overfitting and thus improve the network's performance on the test dataset.

Long-range Stochastic Residual Connection: While being successful, FCN cannot accurately localize object boundaries due to the lack of fine-level information during the label inference stage. To tackle this problem, Unet [35] was proposed to combine low-level

feature maps with high-level feature maps together for label inference during the condensing process. This combination effectively improves the localization accuracy. On the other hand, Yu et al. [13] applied the longrange residual units [36], [43] to the FCN and achieved similar effect as the Unet for medical image segmentation, and they argued that the long-range residual connection implicitly promoted the information exchange between feature learning and label inference. In this work, we also proposed to use longrange stochastic residual connections as a better alternative to improve the FCN.

As shown in Fig. 3(a), we build long-range stochastic residual connections to bridge shallow layers to their corresponding deep layers with same resolutions. The long-range stochastic skip connections are also built in a similar manner as in Eq. 5, but with more convolution operations and nonlinear mappings in the main path. Besides obtaining the localization information from the lower layers, the long-range stochastic residual connections can also help better propagate the gradients from higher layers to lower layers. Moreover, the stochastic connections can boost ensemble behaviors between lower layers and higher layers, which is supposed to improve the robustness of our model. The random probability (p in Eq. 6) is chosen at 0.1 for all the three long-range skip connections.

B. Dilated Convolution in the Smallest Resolution

As shown in Fig. 3, different from the classic segmentation architectures (e.g., FCN [22] and SegNet [44]) with four or five pooling layers, our designed network consists of only three pooling layers in order to avoid excessive spatial information loss due to over downsampling of the feature maps. However, our designed network would suffer from the problem of insufficient effective receptive field, compared to the classic ones.

As mentioned in [45], a wide effective receptive field is critical for the network, since it can take a wider context into consideration for segmentation/classification. To increase the effective receptive field, we adopt a dilated convolution [46]. However, the dilated convolution requires expensive computation and huge memory cost, since it needs to perform convolutions on a large number of detailed (high-resolution) feature maps. Thus, we use the dilation concept only on the layers with the smallest resolution (i.e., pure blue blocks in Fig. 3) to obtain more context information. In the implementation, we set the dilation as 2, which will achieve enough receptive field (as shown in Fig. 3(c)). It is worth noting that our model is different from the traditional dilated convolutional networks, which usually use dilated convolution through the whole network and consequently result in expensive computation and huge memory cost.

C. STRNet: Spatially-varying sTochastic Residual Network

The traditional convolutional layer in CNN works in a weight sharing scheme by assuming that the distribution of features is stationary in an image with respect to the location. While this assumption holds in the natural images, it is often not the case in the medical images. This is because medical images are often linearly or even non-linearly aligned before segmentation. After alignment, different image regions exhibit their unique image statistics. For example, within the MRI bladder region, the texture is homogeneous and simple, which is quite different from the texture of the prostate and rectum. Therefore, it is more

reasonable to learn different filters for adapting to different image regions than simply sharing the filters across the whole image domain. Inspired by this observation, we propose to use locally connected convolutions [37] in the second last layer (i.e., L19 as shown in Fig. 3(d)) of STRNet to learn more effective representations than the traditional convolutions. Because the filters in the locally connected convolutions vary in different locations, we call the second last layer as *spatially-varying convolution layer*. When combined with the stochastic residual FCN in Sec. II-A, the whole network architecture is named as “STRNet” (standing for Spatially-varying sTochastic Residual Networks), as shown in Fig. 3(a).

The spatially-varying convolution layer behaves in the same way as the traditional convolution layer. Their difference is that, instead of sharing filters across spatial domain, filters in the spatially-varying convolution layer can be different at different locations. Suppose that we divide the feature maps into R regions, with r indexing one region. Then, the spatially-varying convolution layer can be mathematically formulated as Eq. 7.

$$\mathbf{A}_{jr}^l = f \left(\sum_{i \in \mathbf{M}_j} \mathbf{A}_{ir}^{l-1} \odot \theta_{ijr}^l \right) \quad (7)$$

where i indexes input channels from the previous layer, and j indexes a filter in the current layer (i.e., the l^{th} layer). Correspondingly, $\mathbf{M}$ means the input feature maps, $\odot$ is an element-wise product operator, and f is the activation function. An illustration of Eq. 7 is given in Fig. 3(c).

With such spatial-varying convolution layers, the STRNet will be able to respect the fact that the pelvic MR images are just roughly aligned and thus need to learn adaptive filters for different regions. The number of local regions in L19 is finally set to 12 after trials. Specifically, we carry out a line search for the possible number of regions in the spatially-varying convolution filters. In particular, we investigate the following candidate numbers: 3 (3×1), 6 (3×2), 9 (3×3), 12 (3×4), 15 (3×5) and 18 (3×6), since we have 3 organs in MRI. Then, we choose the best number (i.e., 12) based on the performance on the validation set. Our experimental results show that the use of spatially-varying filters is important for improving the segmentation accuracy.

D. STRAINet: Adversarial Learning for Correcting Organs

The above-mentioned neural network is capable to achieve good segmentations for most patients. However, it may mismatch the organ shapes due to large shape variance and also violate the spatial constraint due to low tissue contrast between the pelvic organs. Thus, an additional network is needed to correct the segmented organs.

Adversarial learning was derived from the recently popular Generative Adversarial Network (GAN) [47]. It has achieved great success in the field of image generation and segmentation [47], [48], [49], [50], [51]. The GAN framework consists of two competing networks: a generator and a discriminator. Both of them are involved in an adversarial two-player game, in which the generator aims to learn the data distribution while the discriminator estimates

the probability of a sample coming from the training data or the generator. The objective of generator is to maximize the probability of the generated one following the distribution of real ones so that it can fool the discriminator; while, the objective of discriminator is to minimize the chance of mistakes in identifying the generated one from the real one.

It has been demonstrated that adversarial learning can help improve the segmentation models [49], [50]. Thus, we try to use adversarial learning to correct the segmented organs. The flowchart of whole segmentation networks is shown in Fig. 2, which we call “STRAINet” for short.

For convenience, we denote our Segmenter and Discriminator networks as S and D , respectively. And the proposed adversarial learning is a min-max two-player, which can be decomposed into two parts as described below.

Loss for the Discriminator: Loss for the Discriminator is represented by a binary cross entropy loss, as shown in Eq. 8.

$$L_D(\mathbf{X}, \mathbf{Y}, \theta_d) = L_{BCE}(D(\mathbf{Y}, \theta_d), 1) + L_{BCE}(D(S(\mathbf{X}), \theta_d), 0) \quad (8)$$

Where

$$L_{BCE}(\hat{\mathbf{z}}, \mathbf{z}) = -\frac{1}{N} \sum_{i=1}^N \mathbf{z}_i \log(\hat{\mathbf{z}}_i) + (1 - \mathbf{z}_i) \log(1 - \hat{\mathbf{z}}_i) \quad (9)$$

where N is the number of training samples, $\mathbf{Y}$ represents the manual label map of the input data, $\hat{\mathbf{Y}}$ is the predicted label map, and these label maps both take their values in $\{0, 1, 2, \dots, C-1\}$. θ_d denotes a set of network parameters for the discriminator z is the predicted probability in $[0, 1]$ that the discriminator assigns to the input of being drawn from the distribution of ground-truth maps, and i indexes the samples.

Loss for the Segmenter: In the case of S , besides the cross entropy loss defined by Eq. 1, there is another loss from the Discriminator which is used to correct the organ structure of predicted label map. In particular, the adversarial loss term (“ADV”) to improve S and fool D can be defined by Eq. 10.

$$L_{ADV}(\mathbf{X}) = L_{BCE}(D(S(\mathbf{X})), 1) \quad (10)$$

Thus, the total loss for the Segmenter in the context of adversarial learning for FCN can be concluded in Eq. 11.

$$L_S(\mathbf{X}, \mathbf{Y}, \theta_g) = L_{CE}(\mathbf{X}, \mathbf{Y}, \theta_g) + \lambda L_{ADV}(\mathbf{X}) \quad (11)$$

where Θ_g is the network parameters for the generator, λ is a scale parameter for adversarial loss, which is set as 0.5 in our experiments.

The architecture of the Segmenter is described in Fig. 3. This network takes an MR image as the input, and tries to output the corresponding predicted label map. The Discriminator is a typical CNN architecture including three stages of convolution, BN, ReLU and max pooling, followed by one convolutional layer and three fully connected layers where the first two use ReLU as activation functions and the last one uses sigmoid (whose output represents the likelihood of the input data being drawn from the distribution of real masked image). The filter size is 3×3 , the numbers of the filters are 32, 64, 64 and 128 for the convolutional layers, respectively, and the numbers of the output nodes in the fully connected layers are 512, 128 and 1, respectively. The training of the two networks is performed in an alternating fashion. First, D is updated by taking a mini-batch of ground-truth data and a mini-batch of generated label maps corresponding to the outputs of S (Note that both of them are concatenated with the original input MRI to provide more background information). Then, S is updated by using another mini-batch of samples, including sources and their corresponding ground-truth masked images.

E. Iterative Auto-Context Refinement

With the proposed model in Fig. 2, the networks can be trained to learn the organ label for each voxel in the image. However, the obtained 3D label probability maps are often spatially inconsistent, since the context information available for each voxel is limited, especially across slices. To ensure the spatial consistency, it is necessary to consider the predictions of neighboring voxels during voxel-wise label prediction. Auto-context model (ACM) [38] provides an iterative framework that can incorporate the neighboring prediction information to compensate for the limitation of independent voxel-wise estimations. Specifically, we concatenate the prediction maps generated by the Segmenter of STRAINet in the k^{th} iteration with the original MRI data, and feed them as the overall input to the Segmenter of STRAINet in the $(k + 1)^{th}$ iteration. Note, the training in the next iteration for auto-context refinement has to wait until the completion of the current iteration.

F. Algorithm Details

Weighted Loss Function: In our applications, the numbers of data samples for different organ labels are often different. This may cause over-fitting for the organ(s) with more samples, which is known as the class imbalance problem. To avoid this, we apply a class balance strategy during the training. Specifically, we adopt a weighting scheme in the loss function to address this class imbalance problem, i.e., assigning larger weight for the class(es) of organs with less samples, e.g., the prostate, and vice versa. In other words, we apply a weighted indicator function in Eq. 2, and it can be formulated as Eq. 12.

$$I\left\{y, j\right\} = \begin{cases} \mathbf{W}_j, y = j \\ 0, y \neq j \end{cases} \quad (12)$$

where $\mathbf{W}_j$ is a weight given for the j^{th} category.

In our experiment, the weights for background, bladder, prostate and rectum are set to 1, 4, 8 and 6 respectively after trials.

Consecutive Slices as Input: The proposed STRNet is implemented in a 2D fashion, which may cause inconsistency across slices. To alleviate this problem, we utilize 3 continuous slices as input to the network for capturing the certain level of spatial information across slices.

Implementation Details: The pseudo-code of the proposed framework is given in Algorithm 1. Pytorch² is adopted to implement the proposed STRNet shown in Fig. 2. The respective code can be publicly accessed through this link³. Specifically, we adopt stochastic gradient descent to optimize the network. The network weights are initialized by Xavier algorithm [39] to automatically determine the scale of initialization. The weight decay is set to be 1e-4. For the network bias, we initialize it to be 0. The learning rates for the Segmenter and the Discriminator are initialized as 1e-2 and 1e-3, respectively, followed by decreasing the learning rates during the training. Four Titan X GPUs are utilized to train the networks.

III. EXPERIMENTS AND RESULTS

Our pelvic dataset consists of 50 prostate cancer patients from the University of North Carolina Cancer Hospital, each with one T2-weighted MR image and its corresponding manually-labeled map by a medical expert (Jun Lian). The images were acquired with 3T magnetic field strength, while different patients were scanned with different MR image scanners (i.e., Siemens Medical Systems and Philips Medical Systems). Under such a situation, the challenge for the segmentation task increases since both shape and appearance differences are large. The prostate, bladder and rectum in all MRI scans have been manually segmented, which serve as the ground truth for evaluating our segmentation method. The image size is mostly $256 \times 256 \times (120 \sim 176)$, and the image spacing is $1 \times 1 \times 1 \text{ mm}^3$. A typical example of the MR image and its corresponding label map are given in Fig. 1(a).

²<https://github.com/pytorch/pytorch>

³<https://github.com/ginobilinie/Strainet>

Algorithm 1: Auto-context Model based training of STRAINet. We denote Algorithm 2 as *STRAINet* for convenience, and we denote the generator network as *STRNet*.

input : Input training samples $\mathbf{X}$ and its corresponding label maps $\mathbf{Y}$
output: Network parameters $\{\theta_g\}$ for the generators
 Initialization;
 Train the STRAINet with Algorithm 2:
 $\theta_g^{(0)} \leftarrow \text{STRAINet}(\mathbf{X}, \mathbf{Y})$

for $t \leftarrow 1$ **to** maxIter **do**
 Obtain the predicted probability map through the $(t-1)^{\text{th}}$ trained generator network:
 $\hat{\mathbf{Y}}^{(t-1)} \leftarrow \text{STRNet}(\mathbf{X}, \theta_g^{(t-1)})$
 Concatenate $\mathbf{X}$ and $\hat{\mathbf{Y}}^{(t-1)}$ to be $\mathbf{X}^{\text{cat}}$;
 Train the STRAINet with Algorithm 2:
 $\theta_g^{(t)} \leftarrow \text{STRAINet}(\mathbf{X}^{\text{cat}}, \mathbf{Y})$
end

Algorithm 2: Minibatch training of our proposed STRAINet. The number of steps to apply to the discriminator, K , is a hyperparameter. We use $K = 1$ in our experiments.

input : Input training samples $\mathbf{X}$ and its corresponding label maps $\mathbf{Y}$
output: Network parameters θ_g for the generator
 Initialization;
for $t \leftarrow 0$ **to** numOfIter **do**
for $k \leftarrow 0$ **to** K **do**
 Sample a mini-batch of M training samples:
 $\{\mathbf{X}^{(i)}, \mathbf{Y}^{(i)}\}, i = 1, 2, \dots, M$
 Update the discriminator by ascending the stochastic gradient:
 $\nabla_{\theta_d} L_D(\{\mathbf{X}^{(i)}\}, \{\mathbf{Y}^{(i)}\}, \theta_d)$
end
 Sample a mini-batch of M training samples:
 $\{\mathbf{X}^{(i)}, \mathbf{Y}^{(i)}\}, i = 1, 2, \dots, M$
 Update the generator by descending the stochastic gradient:
 $\nabla_{\theta_g} L_S(\{\mathbf{X}^{(i)}\}, \{\mathbf{Y}^{(i)}\}, \theta_g)$
end

Five-fold cross validation is used to evaluate our method. Specifically, in each fold of cross validation, we randomly chose 35 subjects as training set, 5 subjects as validation set, and the remaining 10 subjects as testing set. All the reported performances are evaluated on the testing set (unless indicated specially). Two iterations are used in the auto-context refinement stages. As for the evaluation metrics, we utilize Dice Similarity Coefficient (DSC) to measure the agreement between manually and automatically segmented label maps:

$$DSC = 2 \frac{|A \cap B|}{|A| + |B|}, \quad (13)$$

where A and B denote the binary manual segmentation and automatic segmentation, respectively, $|A|$ denotes the number of positive elements in the binary segmentation A , and $|A \cap B|$ is the number of shared positive elements by A and B . We further evaluate the accuracy by measuring Average Surface Distance (ASD) [52], as defined below:

$$ASD = \frac{1}{2} \left(\frac{\sum_{V_i \in S_A} \min_{V_j \in S_B} d(V_i, V_j)}{\sum_{V_i \in S_A} 1} + \frac{\sum_{V_j \in S_B} \min_{V_i \in S_A} d(V_j, V_i)}{\sum_{V_j \in S_B} 1} \right), \quad (14)$$

where S_A is the surface of the manual segmentation map, S_B is the surface of the automatic segmentation map, and $d(V_j, V_i)$ indicates the Euclidean distance from vertex V_j to the vertex V_i .

A. Comparison with State-of-the-art Methods

To demonstrate the advantage of the proposed method, we also compare our method with other six widely used methods on the same dataset as shown in Table I: 1) multi-atlas label fusion (MALF), 2) SSAE [20] and 3) Unet [35], 4) DeepLab v2 [25], 5) RefineNet [29] and 6) PSPNet [28]. Also, we present the performance of our proposed STRAINet after two iterations of ACM based refinement (denoted as STRAINet+ACM).

Table I quantitatively compares our method with six state-of-the-art segmentation methods. With iterative refinement by ACM, the performance in terms of DSC can be improved by approximately 0.5%, which shows the usefulness of context features in segmentation refinement. As the input to our networks is the 3 consecutive slices (which can provide some context information), the performance cannot be improved as more as declared in [38]. We can see that our methods (STRAINet and STRAINet+ACM) achieve better performance than the state-of-the-art methods in terms of both DSC and ASD. The state-of-the-art segmentation algorithms developed for natural image segmentation work well in segmenting bladder and prostate, but not rectum, due to the missing of detailed information or important structure by over downsampling (i.e., more than 5 times) on the long and narrow rectum. It is worth noting that, although our methods don't apply any deformable model to refine the shape, our methods still perform much better than SSAE that uses shape prior for segmentation refinement based on the deep learning based results. This further shows the effectiveness of our proposed methods. Moreover, it is interesting to observe that our proposed methods have the smallest standard deviation, indicating the robustness of our methods in MRI pelvic organ segmentation.

IV. ANALYSIS AND DISCUSSION

A. Analysis of Components in the Proposed Networks

As our proposed networks consist of several designed components, we conduct some empirical studies below to analyze them.

Impact of Stochastic Residual Unit (SRU): As mentioned in Sec. II-A, we adopt a novel idea of residual learning with stochastic connections to address the convergence issue of the proposed deep network. To validate the effectiveness of this proposed long-range and short-range stochastic residual FCN (LSSRN), we carry out experiments to compare the convergence speeds and the segmentation performance with the following networks:

- FCN: It shares the same contracting and expanding paths as Fig. 3(a), but with the plain convolutional layers and without any residual connections.
- LRN: We add long-range residual connections as described in [43], [13] to the above-described FCN.
- LSRN: We add both long-range and short-range residual connections to the above-described FCN.
- LSSRN: We add both long-range and short-range stochastic residual *units* to the above-described FCN.

The experimental results in terms of convergence and performance are presented in Fig. 4.

In particular, the loss and segmentation accuracy are shown in the left and right part of Fig. 4, respectively. It can be seen that the residual learning can provide a better convergence of the deep network, and the proposed stochastic residual connections can further improve the training and achieve better segmentation accuracy. It is interesting to note that the long-range residual connection has similar effect as the skip connection used in Unet [35].

It is also worth noting that the testing loss of LSSRN is at a similar level as the training loss, while the testing losses of three other networks are all bigger than the training losses. This proves the effectiveness of the proposed stochastic residual connection in preventing overfitting issue and also optimizing the networks. In this sense, the SRU is a combination of residual unit and spatial dropout.

Analysis of Dilation in the Smallest-Resolution Feature Maps: As mentioned in Sec. II-B, we use two dilated stochastic residual units in the feature maps with the smallest resolution. The receptive field has been increased from 5 ($3+3-1$) to 9 ($5+5-1$), which can be considered as an additional pooling layer without losing spatial information, especially considering the smallest resolution in these feature maps. The effectiveness of the dilation in the smallest-resolution feature maps is further confirmed by the improved performance, e.g., the performance in terms of DSC is improved by 2.38%, 2.60% and 2.16% for bladder, prostate and rectum respectively.

Impact of Spatial-varying Convolution: As shown in Fig. 3, the second last convolutional layer (L19) is designed in a spatial-varying fashion for adapting to the non-

stationary features. To fairly evaluate the role of the spatial-varying convolution in STRNet, we compare it with ‘LSSRN’, in which the number of filters in L19 is set to be the same as that in each convolution region of STRNet; we also compare it with a wider version of LSSRN (‘WLSSRN’), in which the number of filters in L19 is set to be the summation of numbers of filters in all the convolution regions of STRNet. The visualization shown in Fig. 5(a) indicates that the performance of STRNet is better than the other two networks, especially for segmentations of prostate and rectum. Fig. 5(b) shows the 3D renderings of segmentations produced by STRNet, which are quite close to the 3D renderings of the manual ground truth. The Dice ratios for the three networks (LSSRN, WLSSRN and STRNet) are also provided in Table II. The WLSSRN performs quite similarly as the LSSRN, while STRNet works much better than WLSSRN even if they use the same number of filters and network variables. This is because the simple increase of width in the networks cannot provide extra information, and thus limits the improvement of WLSSRN. In contrast, the spatially-varying convolution injected in STRNet uses different convolution filters in different regions which would lead to different feature distributions for different regions. This is beneficial for the MR images since they usually have the same structures in the similar regions (with the feature representation being more stationary and close to the real situation).

Impact of Adversarial Learning: To show the contribution of the adversarial learning, we conduct a comparison experiment among STRNet (shown in Fig. 3), STRAINet (STRNet with adversarial learning, as shown in Fig. 2) and STRNet with context model based refinements (STRNet with ACM based refinement (STRNet+ACM) and STRNet with fully connected conditional random field [27] (STRNet+DenseCRF)). As shown in Table III, with the adversarial learning, the DSC values are improved by 0.3%, 0.9% and 1.2% for bladder, prostate and rectum, respectively, without increasing the standard deviation. On the other hand, without adversarial learning, the context model based refinements can improve some performance, especially for bladder, but not as large as the case using the adversarial learning (i.e., for the prostate and rectum). The experimental results also indicate that 1) adversarial learning can contribute to performance improvement and 2) adversarial learning contributes differently from the context model based refinements. That is, adversarial learning provides a way of minimizing the “variational” loss by enforcing higher-order consistency between ground-truth segmentations and automatic segmentations. But, the context model based refinements try to improve the performance in a different way, i.e., combining context information from initial segmentation maps in the inference.

We further investigate how adversarial learning help the Segmenter by visually checking two typical subjects in Fig. 6. In Fig. 6(a), STRNet gives similar (or a little better) segmentation results as STRAINet, which means adversarial learning can still provide subtle improvement even if STRNet has already produced very similar organs to the manual ground truth. In Fig. 6(b), we can clearly see that adversarial learning has corrected the obvious errors in the segmented organs by STRNet. To summarize, adversarial learning can supply subtle improvement if the Segmenter network (without adversarial learning) has already achieved good results, while it can provide huge improvement to correct the segmented organs if the

results from the Segmenter have obvious structural errors. Actually, in our case, the adversarial network is capable to recognize and correct the obvious defects in segmentation.

B. Impact of Refinement Strategy

Besides the Auto-Context Model (ACM) based refinement, Dense Conditional Random Field (Dense CRF) [27] is also widely used for segmentation refinement and shows great potential for structure learning with small time cost. Thus, in this paper, we have also adopted 3D Dense CRF [53] to refine the segmentation results from the deep neural network. Experimental results in Table IV show that 3D Dense CRF can achieve similar performance gain as the ACM for segmenting bladder and prostate (i.e., around 0.5 ~ 1% in terms of DSC), but the performance for segmenting rectum decreases by 1% when using 3D Dense CRF. This is mainly because the rectum's shape is very complicated (i.e., narrow, long and curved) and difficult for inference. Different from 3D Dense CRF, the ACM based refinement involves in a model learning process, with two kinds of benefits. 1) The ACM can borrow context information beyond the input patch itself; 2) it can also provide higher-level and complementary information for easier training of the network to achieve higher performance. The results also indicate that it is more reasonable to use ACM based refinement in our pelvic organ segmentation task.

C. Time and Computational Complexity

As the model training can be completed off-line, the testing time cost is often more crucial for a segmentation task. Thus, in Table V, we provide the average time costs for segmenting one testing subject by different segmentation methods. Note that this experiment is performed on the same PC with the following settings – Memory: 16GB Quad Channel DDR4; Video Card: Nvidia Titan XP; Processor: Intel i7-5820K; Operation System: Ubuntu 14.04. The running time in minutes are listed in Table V for different segmentation methods. As can be seen, our proposed STRAINet is among the fastest methods, and is significantly faster than two conventional methods. The FCN-like and Unet-like methods cost similar testing time. With two extra iterations for auto-context refinement, our proposed framework needs to segment one subject in 2–3 minutes.

D. Model Robustness to Noise

To explore the robustness of our model to noise, we generated five sets of extra images by adding different amounts of Gaussian noise to each of the original testing MR images. As the intensities of each testing image have been basically normalized to $[-1, 5]$ (with actually more than 99% of the normalized intensities being ranged in $[-0.6, 1.4]$), we use the following levels of Gaussian noise, i.e., $std = 0, 0.01, 0.1, 0.2, 0.3, 0.4, 0.5$ or 0.6 , and the fixed mean 0 . We directly apply the model trained on the original dataset (without adding noise) onto the noise-added images, and report the performance in terms of DSC in Table VI. The results show that our proposed model can work very well with a small amount of noise (i.e., $std = 0.1$), and can still present decent results when increasing the amount of noise to some extent (i.e., $std = 0.3$). Considering most of the intensities are within $[-0.6, 1.4]$, we can argue that our model is robust to certain level of noise. However, with the increased amount of noise (i.e., $std = 0.4$), our model trained without considering noise gets worse and worse performance.

E. Generalization Ability: Validation on Another Dataset

To show the generalization ability of our proposed algorithm, we further conduct experiments on the PROMISE12- challenge dataset [54]. This dataset contains 50 subjects, each with a pair of MRI and its manually labeled map (Note, only prostate is annotated in this dataset). Again, five-fold cross validation is performed to evaluate the performances of the methods on this dataset. Note that, in each fold, it takes approximately 12 hours to train our designed framework on a Titan XP GPU cluster. In Table VII, we report performance in terms of DSC and ASD.

As shown in Table VII, our proposed algorithm again achieves very competitive performance in segmenting prostate (0.885 ± 0.015) on the training dataset in a five-fold cross validation manner, compared to the comparison methods. It is worth noting that both Deeplab V2 and Refinenet which are the state-of-the-art methods in natural image segmentation tasks cannot work quite well on this dataset either. Our method is still very competitive compared to the state-of-the-art methods on this dataset [11], [13]. These comparison experimental results indicate that our proposed STRAINet (STRAINet+ACM) can be generalized well to other segmentation datasets.

We also evaluate our proposed method on the testing dataset whose ground-truth label maps are hidden from the participants. The official evaluation metrics used in this challenge include the DSC, the average over the shortest distance between the boundary (surface) points of the volumes (ABD or ASD), the percentage of the absolute difference between the volumes (aRVD), and the 95% Hausdorff distance (95HD). It is worth noting that the organizers not only calculate the evaluation metrics on the whole prostate, but also on the apex and base parts of the prostate that are believed to be the most difficult case for segmentation. In addition, an overall score (shown in the last column) combining the above-mentioned evaluation metrics is also provided to rank the submitted methods (Please refer to [54] for the details about the evaluation metrics). The quantitative results of our method and our competitors are shown in Table VIII (Note, the results were directly obtained from the organizers). Currently, there are totally 50 teams submitting their results and only top 10 teams are listed in the Table. Our proposed method ranks 6th in terms of the overall score among all the participants. More importantly, our proposed method presents much lower standard deviation value compared to all other top 9 methods, which further indicates the effectiveness and robustness of our proposed method.

V. CONCLUSIONS

In this paper, we have presented a *novel* spatial-varying stochastic residual FCN with adversarial learning (STRAINet) to jointly segment pelvic organs from MRI. Specifically, the stochastic residual learning strategy is used to effectively address the optimization problem of the complex network in the FCN, and also the long-range stochastic residual connections are proposed to boost spatial information exchange between lower layers and higher layers. Furthermore, the spatially-varying convolution is also employed to consider different anatomical appearances of different organs in the pelvic MR images. Moreover, adversarial learning strategy is further used to correct the segmented organs for improvement of the segmentation accuracy. By integrating all these components into the FCN, our

proposed STRAINet has achieved significant improvement in terms of both accuracy and robustness.

ACKNOWLEDGMENT

This work was supported by the National Institutes of Health grant R01 CA206100. We thank our colleague Dr. Shaun Stone for helping check the grammar issues.

REFERENCES

- [1]. Fiorino C, Reni M, Bolognesi A, Cattaneo GM, and Calandrino R, "Intra- and inter-observer variability in contouring prostate and seminal vesicles: implications for conformal treatment planning," *Radiotherapy and oncology*, 47, no. 3, pp. 285–292, 1998. [PubMed: 9681892]
- [2]. Klein S, Van Der Heide UA, Lips IM, Van Vulpen M, Staring M, and Pluim JP, "Automatic segmentation of the prostate in 3d mr images by atlas matching using localized mutual information," *Medical physics*, 35, no. 4, pp. 1407–1417, 2008. [PubMed: 18491536]
- [3]. Liao S, Gao Y, Shi Y, Yousuf A, Karademir I, Oto A, and Shen D, "Automatic prostate mr image segmentation with sparse label propagation and domain-specific manifold regularization," in *International Conference on Information Processing in Medical Imaging*, 23 NIH Public Access, 2013, p. 511.
- [4]. Ou Y, Doshi J, Erus G, and Davatzikos C, "Multi-atlas segmentation of the prostate: A zooming process with robust registration and atlas selection," *Medical Image Computing and Computer Assisted Intervention (MICCAI) Grand Challenge: Prostate MR Image Segmentation*, 7, pp. 1–7, 2012.
- [5]. Yan P, Cao Y, Yuan Y, Turkbey B, and Choyke PL, "Label image constrained multiatlas selection," *IEEE Transactions on Cybernetics*, 45, no. 6, pp. 1158–1168, 2015. [PubMed: 25415994]
- [6]. Toth R and Madabhushi A, "Multifeature landmark-free active appearance models: application to prostate mri segmentation," *IEEE Transactions on Medical Imaging*, 31, no. 8, pp. 1638–1650, 2012. [PubMed: 22665505]
- [7]. Toth R, Bloch BN, Genega EM, Rofsky NM, Lenkinski RE, Rosen MA, Kalyanpur A, Pungavkar S, and Madabhushi A, "Accurate prostate volume estimation using multifeature active shape models on t2-weighted mri," *Academic radiology*, 18, no. 6, pp. 745–754, 2011. [PubMed: 21549962]
- [8]. Zhan Y and Shen D, "Automated segmentation of 3d us prostate images using statistical texture-based matching method," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2003, pp. 688–696.
- [9]. Feng Q, Foskey M, Chen W, and Shen D, "Segmenting ct prostate images using population and patient-specific statistics for radiotherapy," *Medical Physics*, 37, no. 8, pp. 4121–4132, 2010. [PubMed: 20879572]
- [10]. Sciolla B, Martin M, and Delachartre P, "Multi-pass 3d convolutional neural network segmentation of prostate mri images," 2017.
- [11]. Milletari F, Navab N, and Ahmadi S-A, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *Fourth IEEE International Conference on 3D Vision (3DV)*. IEEE, 2016, pp. 565–571.
- [12]. Zhu Q, Du B, Turkbey B, Choyke PL, and Yan P, "Deeply-supervised cnn for prostate segmentation," in *International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2017, pp. 178–184.
- [13]. Yu L, Yang X, Chen H, Qin J, and Heng P-A, "Volumetric convnets with mixed residual connections for automated prostate segmentation from 3d mr images" in *AAAI*, 2017, pp. 66–72.
- [14]. Tian Z, Liu L, Zhang Z, and Fei B, "Superpixel-based segmentation for 3d prostate mr images," *IEEE Transactions on Medical Imaging*, 35, no. 3, pp. 791–801, 2016. [PubMed: 26540678]
- [15]. Zheng Y and Comaniciu D, *Marginal Space Learning for Medical Image Analysis: Efficient Detection and Segmentation of Anatomical Structures*. Springer Science & Business, 2014.

- [16]. Lienhart R and Maydt J, "An extended set of haar-like features for rapid object detection," in Proceedings of International Conference on Image Processing, 1 IEEE, 2002, pp. I-I.
- [17]. LeCun Y, Bengio Y, and Hinton G, "Deep learning," *Nature*, 521, no. 7553, pp. 436–444, 2015. [PubMed: 26017442]
- [18]. Shen D, Wu G, and Suk H-I, "Deep learning in medical image analysis," *Annual Review of Biomedical Engineering*, no. 0.
- [19]. Liao S, Gao Y, Oto A, and Shen D, "Representation learning: a unified deep learning framework for automatic prostate mr segmentation," in *International Conference on Medical Image Computing and Computer- Assisted Intervention*. Springer, 2013, pp. 254–261.
- [20]. Guo Y, Gao Y, and Shen D, "Deformable mr prostate segmentation via deep feature learning and sparse patch matching," *IEEE Transactions on Medical Imaging*, 35, no. 4, pp. 1077–1089, 2016. [PubMed: 26685226]
- [21]. Cheng R, Roth HR, Lu L, Wang S, Turkbey B, Gandler W, McCreedy ES, Agarwal HK, Choyke PL, Summers RM et al., "Active appearance model and deep learning for more accurate prostate segmentation on mri" in *Medical Imaging: Image Processing*, 2016, p. 978421.
- [22]. Long J, Shelhamer E, and Darrell T, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.
- [23]. Li G and Yu Y, "Contrast-oriented deep neural networks for salient object detection," *IEEE Transactions on Neural Networks and Learning Systems*, 2018.
- [24]. Ni B, Li T, and Yang X, "Learning semantic-aligned action representation," *IEEE Transactions on Neural Networks and Learning Systems*, 2017.
- [25]. Chen L-C, Papandreou G, Kokkinos I, Murphy K, and Yuille AL, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40, no. 4, pp. 834–848, 2018. [PubMed: 28463186]
- [26]. Zhang Y, Zhou G, Jin J, Zhao Q, Wang X, and Cichocki A, "Sparse bayesian classification of eeg for brain–computer interface," *IEEE transactions on neural networks and learning systems*, 27, no. 11, pp. 2256–2267, 2016. [PubMed: 26415189]
- [27]. Krähenbühl P and Koltun V, "Efficient inference in fully connected crfs with gaussian edge potentials," in *Advances in Neural Information Processing Systems (NIPS)*, 2011, pp. 109–117.
- [28]. Zhao H, Shi J, Qi X, Wang X, and Jia J, "Pyramid scene parsing network," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2881–2890.
- [29]. Lin G, Milan A, Shen C, and Reid I, "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1925–1934.
- [30]. Nie D, Wang L, Gao Y, and Sken D, "Fully convolutional networks for multi-modality isointense infant brain image segmentation," in *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2016, pp. 1342–1345.
- [31]. Chen H, Dou Q, Wang X, Qin J, Cheng JC, and Heng P-A, "3d fully convolutional networks for intervertebral disc localization and segmentation," in *International Conference on Medical Imaging and Virtual Reality*. Springer, 2016, pp. 375–382.
- [32]. Liu J, Wang Y, Li Y, Fu J, Li J, and Lu H, "Collaborative deconvolutional neural networks for joint depth estimation and semantic segmentation," *IEEE Transactions on Neural Networks and Learning Systems*, 2018.
- [33]. Xing F, Xie Y, Su H, Liu F, and Yang L, "Deep learning in microscopy image analysis: A survey," *IEEE Transactions on Neural Networks and Learning Systems*, 2017.
- [34]. Nie D, Wang L, Adeli E, Lao C, Lin W, and Shen D, "3-d fully convolutional networks for multimodal isointense infant brain image segmentation," *IEEE Transactions on Cybernetics*, 2018.
- [35]. Ronneberger O, Fischer P, and Brox T, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 234–241.

- [36]. He K, Zhang X, Ren S, and Sun J, “Deep residual learning for image recognition,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.
- [37]. Taigman Y, Yang M, Ranzato M, and Wolf L, “Deepface: Closing the gap to human-level performance in face verification,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1701–1708.
- [38]. Tu Z and Bai X, “Auto-context and its application to high-level vision tasks and 3d brain image segmentation,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 32, no. 10, pp. 1744–1757, 2010. [PubMed: 20724753]
- [39]. Glorot X and Bengio Y, “Understanding the difficulty of training deep feedforward neural networks,” in Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, 2010, pp. 249–256.
- [40]. Veit A, Wilber MJ, and Belongie S, “Residual networks behave like ensembles of relatively shallow networks,” in Advances in Neural Information Processing Systems, 2016, pp. 550–558.
- [41]. Huang G, Sun Y, Liu Z, Sedra D, and Weinberger KQ, “Deep networks with stochastic depth,” in European Conference on Computer Vision. Springer, 2016, pp. 646–661.
- [42]. Srivastava N, Hinton GE, Krizhevsky A, Sutskever I, and Salakhutdinov R, “Dropout: a simple way to prevent neural networks from overfitting,” Journal of Machine Learning Research, 15, no. 1, pp. 1929–1958, 2014.
- [43]. Kim J, Kwon Lee J, and Mu Lee K, “Accurate image super-resolution using very deep convolutional networks,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1646–1654.
- [44]. Badrinarayanan V, Kendall A, and Cipolla R, “Segnet: A deep convolutional encoder-decoder architecture for image segmentation,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 39, no. 12, pp. 2481–2495, 2017. [PubMed: 28060704]
- [45]. Luo W, Li Y, Urtasun R, and Zemel R, “Understanding the effective receptive field in deep convolutional neural networks,” in Advances in Neural Information Processing Systems, 2016, pp. 4898–4906.
- [46]. Yu F and Koltun V, “Multi-scale context aggregation by dilated convolutions,” arXiv preprint arXiv:1511.07122, 2015.
- [47]. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, and Bengio Y, “Generative adversarial nets,” in Advances in neural information processing systems (NIPS), 2014, pp. 2672–2680.
- [48]. Nie D, Trullo R, Petitjean C, Ruan S, and Shen D, “Medical image synthesis with context-aware generative adversarial networks,” arXiv preprint arXiv:1612.05362, 2016.
- [49]. Kohl S, Bonekamp D, Schlemmer H-P, Yaqubi K, Hohenfellner M, Hadaschik B, Radtke J-P, and Maier-Hein K, “Adversarial networks for the detection of aggressive prostate cancer,” arXiv preprint arXiv:1702.08014, 2017.
- [50]. Moeskops P, Veta M, Lafarge MW, Eppenhof KA, and Pluim JP, “Adversarial training and dilated convolutions for brain mri segmentation,” arXiv preprint arXiv:1707.03195, 2017.
- [51]. Nie D, Trullo R, Lian J, Wang L, Petitjean C, Ruan S, Wang Q, and Shen D, “Medical image synthesis with deep convolutional adversarial networks,” IEEE Transactions on Biomedical Engineering, 2018.
- [52]. Yeghiazaryan V and Voiculescu I, “An overview of current evaluation methods used in medical image segmentation,” Tech. Rep CS-RR-1508, Department of Computer Science, University of Oxford, Oxford, UK, Tech. Rep., 2015.
- [53]. Kamnitsas K, Ledig C, Newcombe VF, Simpson JP, Kane AD, Menon DK, Rueckert D, and Glocker B, “Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation,” Medical Image Analysis, 36, pp. 61–78, 2017. [PubMed: 27865153]
- [54]. Litjens G, Toth R, van de Ven W, Hoeks C, Kerkstra S, van Gin-neken B, Vincent G, Guillard G, Birbeck N, Zhang J et al., “Evaluation of prostate segmentation algorithms for mri: the promise12 challenge,” Medical image analysis, 18, no. 2, pp. 359–373, 2014. [PubMed: 24418598]

Fig. 1.

(a) A typical pelvic MRI and its manual segmentations of bladder (orange), prostate (silver), and rectum (pink), where the three columns show a MRI slice, the same slice overlaid with manual segmentations, and 3D renderings of manual segmentations, respectively. (b) Another subject's 3D renderings of manual segmentations, showing large inter-subject variations of pelvic organs, compared to the one shown in (a).

Fig. 2. Pipeline of our segmentation networks with adversarial learning (STRAINet). One is the Segmenter (S) network to generate the predicted mask, and the other is Discriminator (D) network for adversarial learning.

Fig. 3.

Illustration of two contributions in architectures of our proposed network for segmenter, STRNet. In STRNet, the second last convolutional layer (L19) is the spatial-varying convolutional layer, which does not share filters at different spatial locations (as shown in (d)). All other plain convolutional layers are replaced by a (dilated) stochastic residual learning module (as shown in (a)). The long-range stochastic residual connection is used to help spatial information exchange between shallow layers and deep layers with the same resolution.

Fig. 4. Comparison of different networks. Left: Training and testing losses of the 4 networks; Right: Dice of the 4 networks on the testing dataset.

Fig. 5.

(a) Typical pelvic organ segmentation results of a typical subject by three different networks. Orange, silver and pink contours indicate the manual ground-truth segmentations, and yellow, red and cyan contours indicate automatic segmentations. (b) 3D renderings of segmentations by STRNet and manual expert for the same subject in (a).

Fig. 6.

Visual inspection of segmentation improvements by adversarial learning on two different cases. Here, STRNet means our proposed networks without adversarial learning, and STRAINet means our proposed networks with adversarial learning. In (a), adversarial learning does not help much, as STRNet already gives good results. In (b), adversarial learning can help to correct the segmented organs obviously, due to large segmentation errors by STRNet.

TABLE I

DSC and ASD by different methods on the pelvic image dataset.

Method	DSC			ASD		
	Bladder	Prostate	Rectum	Bladder	Prostate	Rectum
MALF	.867(.068)	.793(.087)	.764(.119)	1.641(.360)	2.791(.930)	3.210(2.112)
SSAE	.918(.031)	.871(.042)	.863(.044)	1.089(.231)	1.660(.490)	1.701(.412)
Unet	.896(.028)	.822(.059)	.810(.053)	1.214(.216)	1.917(.645)	2.186(0.850)
DeepLab v2	.904(.024)	.805(.048)	.0737(.072)	1.201(.210)	1.985(.588)	3.060(1.290)
RefineNet	.940(.022)	.821(.034)	.816(.080)	1.042(.195)	1.736(.504)	2.055(.954)
PSPNet	.938(.013)	.883(.046)	.0.794(.071)	1.058(.168)	1.404(.477)	2.159(1.091)
STRAINet	.968(.007)	.907(.015)	.905(.028)	.884(.148)	1.317(.301)	1.386(.350)
STRAINet+ACM	.972(.006)	.914(.014)	.910(.026)	.851(.143)	1.312(.303)	1.380(.347)

TABLE II

Comparison of the performances of three networks on the pelvic dataset in terms of DSC.

	Bladder	Prostate	Rectum
LSSRN	.943(.021)	.888(.032)	.874(.033)
WLSSRN	.944(.021)	.888(.032)	.875(.033)
STRNet	.965(.007)	.898(.015)	.893(.028)

TABLE III

Comparison of the performances of the four methods on the pelvic dataset in terms of DSC.

Method	Bladder	Prostate	Rectum
STRNet	.965(.007)	.898(.015)	.893(.028)
STRAINet	.968(.007)	.907(.015)	.905(.028)
STRNet+ACM	.971(.006)	.903(.014)	.899(.026)
STRNet+dense CRF	.970(.007)	.902(.016)	.887(.031)

TABLE IV

Comparison of the performances with different refinement strategies on the pelvic image dataset (in terms of DSC).

Method	Bladder	Prostate	Rectum
STRAINet	.968(.007)	.907(.015)	.905(.028)
STRAINet+ACM	.972(.006)	.914(.014)	.910(.026)
STRAINet+dense CRF	.974(.006)	.912(.014)	.893(.033)

Average time cost (in minutes) and standard deviation on each testing subject of the pelvic MRI dataset by the baseline comparison methods and our proposed method.

TABLE V

Method	MALF	SSAE	Unet	STRAINet	STRAINet+ACM
Average Time Cost	108(13.12)	47(6.65)	0.91(0.17)	0.89(0.17)	2.69(0.49)

TABLE VI

Segmentation performance in terms of DSC on images with different amount of Gaussian noise.

Noise	0	.01	.1	.2	.3	.4	.5	.6
Bladder	.972(.006)	.967(.007)	.893(.019)	.823(.093)	.752(.250)	.417(.145)	.130(.088)	.003(.001)
Prostate	.914(.014)	.910(.015)	.876(.024)	.7719(.114)	.605(.179)	.124(.082)	.001(0)	0(0)
Rectum	.910(.026)	.902(.028)	.863(.040)	.7882(.120)	.643(.241)	.081(.0612)	.001(0)	0(0)

TABLE VII

Comparison of performances of four methods on a second dataset for prostate segmentation.

Method	DSC	ASD
Unet	.837(.025)	2.778(.465)
Deeplab V2	.796(.075)	3.291(1.110)
RefineNet	.811(.064)	3.279(.997)
PSPNet	.864(.024)	2.070(.428)
STRAINet	.876(.017)	1.893(.330)
STRNet+ACM	.885(.015)	1.855(.311)

TABLE VIII

Quantitative comparison between our proposed method and other methods on the testing dataset.

Method	DSC(%)			ASD			95HD			aRVD			Score(std)
	whole	base	apex	whole	base	apex	whole	base	apex	whole	base	apex	
Philips-DLMBS	90.46	88.51	85.29	1.71	1.92	1.91	4.94	5.09	4.08	2.14	7.22	-4.93	87.21(4.49)
AutoDenseSeg	90.14	88.09	86.79	1.83	1.94	1.79	5.36	5.13	4.32	4.53	5.19	2.05	87.19(4.25)
WHU-CS-RPI	90.26	89.15	88.36	1.86	1.79	1.62	5.57	4.83	3.90	9.74	10.73	9.64	87.04(5.79)
CUMED	89.43	86.42	86.81	1.95	2.13	1.74	5.54	5.41	4.29	6.95	11.04	15.18	86.65(4.42)
SCIRESU	90.24	88.98	83.30	1.74	1.81	2.11	4.93	4.51	5.34	6.01	8.18	-7.33	86.41 (3.49)
Proposed	89.25	88.21	86.79	1.98	1.89	1.81	5.70	4.82	4.27	9.19	4.94	11.18	86.15(2.93)
QUILL(Method2)	88.81	87.39	85.46	1.97	2.01	1.91	5.29	5.07	4.35	6.97	4.76	5.85	85.93(4.97)
RUCIMS	88.75	86.04	86.11	2.046	2.220	1.844	5.59	5.68	4.32	5.91	3.80	5.15	85.78(5.09)
CREATIS	89.33	86.60	86.77	1.93	2.14	1.74	5.59	5.62	4.22	5.21	4.20	5.82	85.74(4.96)
methinks	87.91	86.78	84.58	2.06	2.11	2.01	5.53	5.45	4.62	5.73	1.09	11.95	85.41(5.47)