3-D RoI-Aware U-Net for Accurate and Efficient Colorectal Tumor Segmentation

Yi-Jie Huang, Qi Dou, *Member, IEEE*, Zi-Xian Wang, Li-Zhi Liu, Ying Jin, Chao-Feng Li, Lisheng Wang, Hao Chen, and Rui-Hua Xu

***Abstract*—Segmentation of colorectal cancerous regions from 3-D magnetic resonance (MR) images is a crucial procedure for radiotherapy. Automatic delineation from 3-D whole volumes is in urgent demand yet very challenging. Drawbacks of existing deep-learning-based methods for this task are two-fold: 1) extensive graphics processing unit (GPU) memory footprint of 3-D tensor limits the trainable volume size, shrinks effective receptive field, and therefore, degrades speed and segmentation performance and 2) in-region segmentation methods supported by region-of-interest (RoI) detection are either blind to global contexts, detail richness compromising, or too expensive for 3-D tasks. To tackle these drawbacks, we propose a novel encoder–decoder-based framework for 3-D whole volume segmentation, referred to as 3-D RoI-aware U-Net (3-D RU-Net). 3-D RU-Net fully utilizes the global contexts covering large effective receptive fields. Specifically, the proposed model consists of a global image encoder for global understanding-based RoI localization, and a local region decoder that operates on pyramid-shaped in-region global features, which is GPU memory efficient and thereby enables training and prediction with large 3-D whole volumes. To facilitate the global-to-local learning procedure and enhance contour detail richness, we designed a dice-based multitask hybrid loss function. The efficiency of the proposed framework enables an extensive model ensemble for further performance gain at acceptable extra computational costs. Over a dataset of 64 T2-weighted MR images, the experimental results of four-fold cross-validation show that our method achieved 75.5% dice similarity coefficient (DSC) in 0.61 s per volume on a GPU, which significantly outperforms competing methods in terms of accuracy and efficiency. The code is publicly available.**

***Index Terms*—3-D convolutional neural networks (CNN), colorectal cancer, multitask learning, region of interest (RoI), tumor segmentation.**

Manuscript received February 13, 2019; revised May 17, 2019 and January 14, 2020; accepted March 1, 2020. This work was supported in part by the Shanghai Intelligent Medicine Project under Grant 2018ZHYL0217, in part by the SJTU Translational Medicine Cross Research Funds under Grant YG2019ZDA26 and Grant ZH2018QNA05, in part by the Construction Project of Shanghai Key Laboratory of Molecular Imaging under Grant 18DZ2260400, in part by the Shanghai Municipal Education Commission (Class II Plateau Disciplinary Construction Program of Medical Technology of SUMHS, 2018–2020), and in part by the Shenzhen Science and Technology Program under Grant JCYJ20180507182410327. This article was recommended by Associate Editor I. Bukovsky. *(Corresponding authors: Lisheng Wang; Hao Chen; Rui-Hua Xu.)*

Yi-Jie Huang is with the Department of Automation, Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai 200240, China, and also with Department of Research and Development, Imsight Medical Technology Company Ltd., Hong Kong (e-mail: huangyj_wuhan@sjtu.edu.cn).

Qi Dou is with the Department of Computer Science and Engineering, Chinese University of Hong Kong, Hong Kong (e-mail: dqcarren@gmail.com).

Zi-Xian Wang, Li-Zhi Liu, Ying Jin, Chao-Feng Li, and Rui-Hua Xu are with the State Key Laboratory of Oncology in South China, Sun Yat-sen University Cancer Center, Guangzhou 510060, China, and also with the Collaborative Innovation Center for Cancer Medicine, Sun Yat-sen University Cancer Center, Guangzhou 510060, China (e-mail: wangzx@sysucc.org.cn; liulizh@sysucc.org.cn; jinying1@sysucc.org.cn; lichf@sysucc.org.cn; xurh@sysucc.org.cn).

Lisheng Wang is with the Department of Automation, Shanghai Jiao Tong University, Shanghai 200240, China, and also with the Shanghai Key Laboratory of Molecular Imaging, Shanghai University of Medicine and Health Sciences, Shanghai 201318, China (e-mail: lswang@sjtu.edu.cn).

Hao Chen is with Imsight Medical Technology Company Ltd., Hong Kong (e-mail: hchen@cse.cuhk.edu.hk).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCYB.2020.2980145

I. Introduction

COLORECTAL cancer strikes more than 1.4 million people and accounts for 694 000 deaths globally in 2012 [1]. It is more common in developed countries, for example, in the USA, colorectal cancer is the second leading cause of cancer-related mortalities [2]. In the current clinical routine of radiotherapy, due to the advantages of magnetic resonance (MR) imaging for soft tissue enhancement [3], colorectal cancer regions are manually recognized and delineated from volumetric MR images acquired for treatment, including surgery and radiation therapy. However, this procedure is laborious, time consuming, and observer dependent, thus suffers from the tedious effort and limited reproducibility. Therefore, automatic colorectal tumor detection and segmentation methods are highly demanded to improve the clinical routine.

Such demand defines a task of automatic detection and segmentation of the targets from whole 3-D image volumes. Compared to processing manually selected regions-of-interest (RoI) patches, the superiority of being fully automatic simplifies the workflow, excludes manual intervention, and enables fast processing of large amounts of image volumes. Taking initial works [4], [5] one step further, deep-learning-based methods dominate the state of the art of detection and segmentation field. Generally, deep-learning-based methods for this task are challenged by the following factors: weak intensity specificity, absence of shape characteristic, lacking positional priors (as is illustrated in Fig. 1), class imbalance, and long processing time of existing methods under inferior graphics processing unit (GPU) or CPU-only deployment environments.

Apart from the aforementioned challenges, a vital 3-D image-specific problem is not fully tackled by the community.

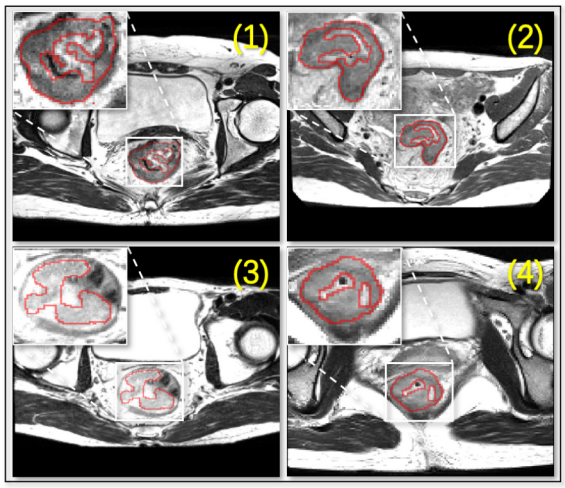


Fig. 1. Typical examples of MR slices with colorectal cancer. The cancer regions are delineated with red lines and zoomed in for clear illustration. It is clear that the target areas lack shape characteristic, intensity specificity, and positional priors.

Among existing methods for fully automatic image segmentation [6]–[13], though a plausible performance can be achieved by utilizing multilevel features (e.g., use skip connections) to gather fine-grained details that are lost in the downsampling process, the merit of maintaining a global understanding represented by deep features with a large receptive field is not fully enjoyed due to patch size limitation of GPU memory. As is supported by many researches for 2-D image processing, for example, dilated convolutions [14] and pyramid pooling schemes [15], enlarging receptive fields enables wide-range context utilization and makes further performance breakthroughs. In medical applications, global understanding is even more important since that the targets and the background are highly correlated, but not fully utilized due to input size limitation.

Generally, existing methods for lesion detection and segmentation from 3-D images can be divided into part-based models and in-region segmentation methods supported by RoI localization. Initially, as naive practices, part-based fully convolutional networks (FCNs) learn from local parts of 2-D slices [8], [16], [17], 2.5-D slices [18], [19], or small 3-D patches [11], [20], [21] and perform (often overlapped) part sliding for the whole volume inference, which is slow and prone to false positives and target incompleteness-related failures. More important, part-based methods suffer from limited effective receptive fields. V-Net [10], for example, claimed $551\times551\times551$ designed receptive field but used the $64\times128\times128$ patch sliding scheme, making the large designed receptive field not fully effective. To enlarge the effective receptive field under current part-based frameworks, Crossbar-Net [22] proposed to train segmentation networks using nonsquared patches with different aspect ratios to include more global contexts to local details.

More recently, trends highlight potential accuracy and speed benefits of adding RoI localization modules prior to FCNs. As a common practice, the RoI localization modules are individually designed as a standalone part of the full pipelines. Conventionally, RoIs are localized using prior knowledge such as multiatlas registration [23], [24], which is often used to localize normal organs. Apart from their inappropriateness for lesion localization, they are relatively slow. As is reported in [25], registration takes at least 20 s per patient using GPUs and typically hours per patient using CPUs. Learning-based RoI localization decouples RoI localization from prior knowledge [26]–[30]. Some of the related practices [26], [31] extract region proposals using external modules, such as selective search [32] or multiscale combinatorial grouping (MCG) [33], which are also well-known speed bottlenecks as is pointed out in [34] and replacing them with the region proposal network (RPN) accelerated a network from 0.5 frames/s to 5 frames/s. Later works adopt light convolutional neural-network (CNN) models, such as 2-D CNNs for RoI localization and 3-D FCNs for in-region segmentation [29], [35], [36]. Compared to part-based methods, these works tackle the tasks in more graceful manners. Still, using a standalone FCN for RoI segmentation requires repeated extraction of in-region features. However, repeated feature extraction is redundant since that it can be eliminated by feature sharing from the detection stage. As is reported in [37], feature sharing produces $146\times$ acceleration without truncated singular-value description [38], [39] and $213\times$ with it in the test phase for object detection, given large numbers of target candidates. In addition, using a standalone FCN for RoI segmentation leaves the problem of limited effective receptive fields unsolved and is therefore still blind to beneficial global contexts.

To tackle the aforementioned drawbacks, methods that jointly train RoI localization and in-region segmentation eliminate the repeated extraction of low-level features and pass global contexts to the in-region segmentation branch. Such methods achieve better speed and accuracy, as reported in, multitask network cascades (MNCs) [40] and its more recent competitor Mask R-CNN [41] with the feature pyramid network (FPN) [42]. However, an apparent drawback of Mask R-CNN is its accumulated detail losses introduced by both its heuristic feature level assignment and its RoI extracting scheme. Specifically, in Mask R-CNN, each proposal is predicted based on feature grids pooled from a single heuristically assigned feature level which is dominantly downsampled from the raw resolution. In addition, RoIAlign's bin-fitting scheme introduces resampling and distortion, which are also detail losing. To tackle this issue, the path aggregation network (PAN) [43] added another bottom-up path for multilevel feature utilization, but another path is too memory demanding for a 3-D task. Nevertheless, in a 3-D application, an anchor-based detector needs to define extra anchor boxes with additional aspect ratios along the z-axis, whereas the number of 3-D targets is very limited on the contrary. Fitting a small amount of 3-D objects to a large number of anchor boxes is problematic, making the acquired regressor prone to bad-shaped bounding box prediction. Such drawbacks motivate us to design an anchor free, memory efficient, and detail-preserving 3-D segmentation framework.

Apart from the way whole volume predictions are generated, recent works propose some strategies to further boost the performance of volumetric tasks. First, V-Net [10]

adopts parameter-free dice coefficient [44] loss to harness the class-imbalance issue. Second, inspired by the success of multitask learning [45], [46], deep contour-aware networks (DCANs) [47] and boundary-aware FCN [48] employ contour-aware loss functions for better discrimination between boundaries and the background. In addition, multilevel contextual 3-D CNNs [49], DeepMedic [50], orchestral FCNs (OFCNs) [51], and hybrid loss (HL)-guided FCNs (HL-FCNs) [52] adopt model ensemble for better robustness.

A part-based initial work to automatically segment colorectal cancer regions was published in ISBI 2018 [52]. In this article, we further aim to enjoy the benefits of fast global localization from 3-D whole volumes and global context sharing across global and local tasks while maintaining the easy to train and detail-preserving merits of popular volume-to-volume segmentation methods. Our implementation is publicly available at https://github.com/huangyjhust/3-D-RU-Net. Our main contributions are summarized as follows.

1) We propose a 3-D joint framework called 3-D RoI-aware U-Net (3-D RU-Net) for joint RoI localization and in-region segmentation. The proposed model consists of a shared global image encoder for global-understanding-based RoI localization, and a local region decoder working on pyramid-designed in-region global features for RoI segmentation. This design enables fast, memory efficient, and detail-preserving whole volume segmentation enhanced by the full utilization of global contexts from large receptive fields.
2) Considering automatic class rebalancing and better boundary discrimination, we propose a dice-based global-to-local multitask HL (MHL) function to further improve the accuracy. In addition, the accelerated framework encourages us to employ a multiple receptive field model ensemble strategy to suppress the false positives and refine the boundary details at an acceptable speed cost.
3) Extensive ablation studies are conducted to evaluate the contribution of each proposed component, and third-party methods were also compared to show the efficacy of our method.

The remainder of this article is organized as follows. We describe our method in Section II and report the experimental results in Section III. Section IV further discusses some insights as well as issues of the proposed method. The conclusions are drawn in Section V.

II. Method

In this section, to address slow prediction and limited effective receptive field issues of nonjoint models along with detail lossing and bad bounding box issues of joint models discussed in Section I, we propose a framework to effectively localize and segment colorectal tumors from whole volume 3-D images.

A. Construction of 3-D RU-Net

The proposed architecture is illustrated in Fig. 2. We input whole image volumes to global image encoder for multilevel feature encoding, employ an encoder-only RoI locator for RoI localization, crop in-region feature tensors from multiscale feature maps using the RoI pyramid layer, and design a local region decoder subnetwork to perform multilevel feature fusion for high-resolution cancerous tissue segmentation.

1) Global Image Encoder: Due to limited GPU memory of commonly used devices and dramatically increased parameters of 3-D convolution kernels, it is essential to carefully design the 3-D backbone feature extractor to avoid GPU memory overflow and overfitting.

Instead of constructing a full 3-D version of the encoder–decoder architecture like 3-D FPN, or directly extending popular backbones [53]–[55] to 3-D, a compact encoder-only network called the global image encoder is constructed to process whole volume images rather than dealing with context-limited small parts as common practices do. Specifically, the encoder employs a stack of ResBlocks [54] and MaxPooling layers to encode whole volume images. Each residual block has three convolutional layers, three normalization layers, three ReLU layers, and a skip connection for better gradient flowing. Constrained by memory consumption, it is necessary to set $batchsize = 1$ to make entire volumes trainable but batch normalization under a small batch size significantly degrades the performance. Therefore, we use instance normalization [56], a special case of group normalization [57], for replacement of batch normalization due to its insensitiveness to batch size.

2) RoI Locator: The RoI locator is a template where any method that employs encoder-only backbones for target detection can be employed. Due to the aspect ratio diversity of number-limited training samples, learning accurate bounding box regression can be difficult. For this specific 3-D semantic segmentation task, we recommend taking full advantage of available voxel-level masks as is discussed as follows for simplicity and more robust bounding box prediction.

The RoI Locator is trained to predict downsampled segmentation masks from global images instead of degrading voxelwise labels to objectwise labels to learn anchor fitting. Specifically, the locator is designed as a module taking feature map F_{III} as input, consisting of a convolutional layer with kernel size 1 and the *Sigmoid* activation function. To tackle the extremely imbalanced foreground-to-background ratio, instead of partial sampling, sampling a fixed proportion of foreground and background or employing online hard example mining (OHEM) [58], the locator is trained toward a global dice loss (DL), which will be introduced in Section II-B. Then, we perform a fast 3-D connectivity analysis to compute desired bounding boxes formulated as $\text{Bbox}_{III} = (z_3, y_3, x_3, d_3, h_3, w_3)$, where (z_3, y_3, x_3) denotes the starting coordinates and (d_3, h_3, w_3) denotes the depth, height, and width of Bbox_{III} in feature map F_{III}.

3) RoI Pyramid Layer: For full utilization of multilevel features and better mask details, we propose a novel layer called the RoI pyramid layer for replacement of the bin-fitting scheme of RoI alignment, which pools local RoI tensors from a heuristically selected single-scale feature map. As is illustrated in Fig. 2, the RoI pyramid layer extracts pyramid-shaped in-region features, forming tensor groups (f_I, f_{II}, f_{III}) from each

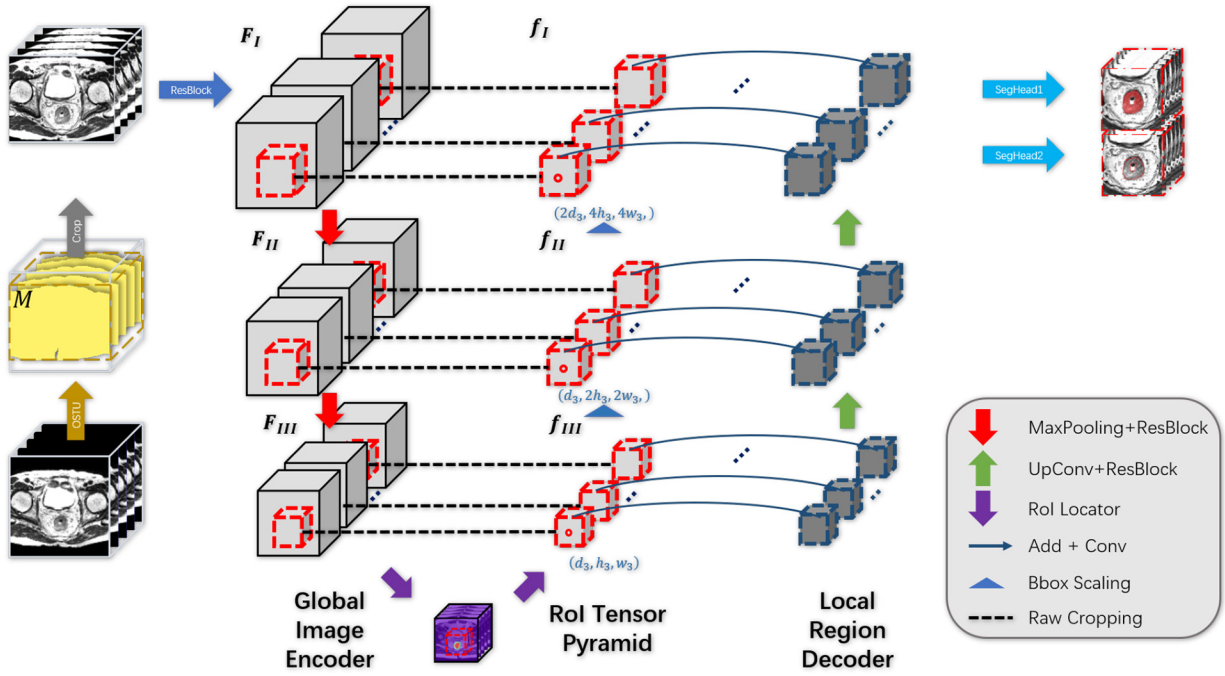


Fig. 2. Illustration of 3-D RU-Net. The network consists of the global image encoder, the RoI tensor pyramid, and the local region decoder. A bounding box is predicted using feature maps $\mathbf{F}_{III}$ and is extended as a bounding box pyramid, then the corresponding RoI tensor pyramid (f_I, f_{II}, f_{III}) is extracted from (F_I, F_{II}, F_{III}) and memory-efficient multilevel feature fusion for in-region segmention is performed in the decoder stage.

scale of globally encoded tensors (F_I, F_{II}, F_{III}) produced by the global image encoder.

Specifically, to extract a feature group for a detected target, we pass the detected bounding box $\text{Bbox}_{III} = (z_3, y_3, x_3, d_3, h_3, w_3)$ to its former feature scales, constructing a pyramid-shaped bounding box set $(\text{Bbox}_I, \text{Bbox}_{II}, \text{Bbox}_{III})$. The bounding box set is computed iteratively by inverting the MaxPooling strides as is listed as follows:

$$\begin{aligned}\text{Bbox}_{i-1} = \big(&z_i \times \hat{z}_{i-1}, y_i \times \hat{y}_{i-1}, x_i \times \hat{x}_{i-1} \\ &d_i \times \hat{z}_{i-1}, h_i \times \hat{y}_{i-1}, w_i \times \hat{x}_{i-1}\big)\end{aligned} \tag{1}$$

where $(\hat{z}_{i-1}, \hat{y}_{i-1}, \hat{x}_{i-1})$ denotes the stride configuration of the MaxPooling_{i-1} layer along the z-, y-, and x-axes. Given the bounding box set $(\text{Bbox}_I, \text{Bbox}_{II}, \text{Bbox}_{III})$, we crop raw in-region features (f_I, f_{II}, f_{III}) from whole volume feature maps F_I, F_{II}, and F_{III} without applying any bin-fitting operation and pass it to the posterior local region decoder branch.

4) Local Region Decoder: With an in-region feature set (f_I, f_{II}, f_{III}) cropped from the encoder path, we construct a multilevel subnetwork for in-region segmentation called local region decoder by applying the successful multilevel feature fusion mechanism. The construction of the decoder is more or less symmetrical to the encoder part with skip connections to fuse feature maps of corresponding scales, while the beneficial difference lies on much smaller sizes of the decoder branch's feature tensors. Our initial experiments suggest that using pooling modules to bin-fit multilevel features before fusing them with convolutional layers is extremely harmful to the performance, therefore RoI Pooling and RoI Align are abandoned and raw features are processed by the local region decoder. Since no shape distortion or scale normalization is included in the RoI pyramid layer, this module restores the spatial dimension of the RoI region without losing details. The same set of decoder weights is used to iteratively process different RoIs if multiple RoIs are localized.

B. Dice-Based Multitask Hybrid Loss Function

In multitask learning practices, each task faces different challenges. In our case, the global image encoder mainly suffers from the class-imbalance issue, while the local region decoder has to focus on the exact boundaries of the target regions. Thus, we propose a dice-based MHL function to effectively learn these tasks.

1) Dice Loss Formulation: Inspired by the success of [10], we apply the DL function to formulate the optimization objective, since it serves as an effective hyperparameter free class balancer to help the network learn objects of small size and weak saliency. The DL is defined as

$$L_d(P, G) = 1 - 2 \times \frac{\sum_{i=1}^{N} p_i g_i + \epsilon}{\sum_{i=1}^{N} p_i + \sum_{i=1}^{N} g_i + \epsilon} \tag{2}$$

where the sums are computed over the N voxels of the predicted volume $p_i \in P$ and the ground-truth volume $g_i \in G$. ϵ is a minimal smoothness term set as 10^{-4} that avoids division by 0 when a proposal contains no target. In the optimization stage, the DL is minimized by gradient descent using the following derivative:

$$\frac{\partial L_d(P, G)}{\partial p_k} = -2 \times \frac{\sum_{i=1}^{N} p_i g_i - g_k \sum_{i=1}^{N} (p_i + g_i)}{\left[\sum_{i=1}^{N} (p_i + g_i)\right]^2}. \tag{3}$$

2) Dice Loss for Global Localization: To tackle the class-imbalance issue of the global image RoI localization task, we

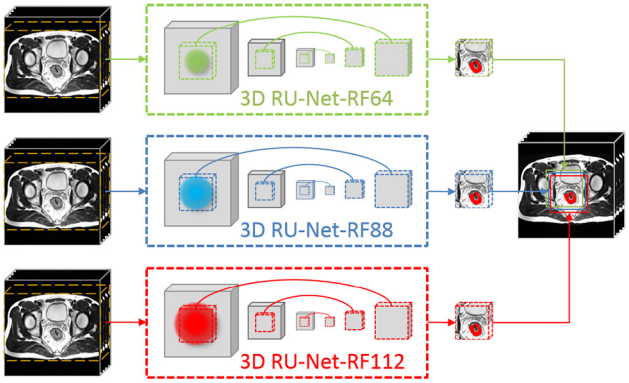


Fig. 3. 3-D RU-Net (RF64), 3-D RU-Net (RF88), and 3-D RU-Net (RF112) are of different dilation rates. The green, blue, and red spheres of different sizes indicate receptive fields of $26\times64\times64$, $26\times88\times88$, and $26\times112\times112$, respectively. In the output end, their predictions are averaged.

employ DL for global RoI localization

$$L_g = L_d\big(P_g, G_g\big) \tag{4}$$

where P_g and G_g denote predictions of *RoILocator* and downsampled annotations, respectively. The DL helps the global image encoder branch learn better discriminate foreground regions from the background and get rid of the influence of class imbalance.

3) Dice-Based Contour-Aware Loss for Local Segmentation: Compared to the localization task, the in-region segmentation branch needs multiple constraints to acquire better boundary-sensitive segmentation results. In semantic segmentation practices, the ambiguous borders are the most difficult to learn but learned with insufficient attention. Borrowing the insight of the previous exploration of adding an auxiliary contour-aware side task [47], we further formulate the side task using DL to help it tackle the extreme sparsity of contour labels in 3-D space. Practically, we add an extra output head called *SegHead*2 at the output terminal of the *Local Region Decoder* to predict the contour voxels, trained in parallel with the region segmentation head *SegHead*1. Taking the side task into account, the loss function of the segmentation branch L_{local} is denoted as following by summarizing the weighted losses:

$$L_l = L_d(P_r, G_r) + \lambda_c L_d(P_c, G_c) \tag{5}$$

where $\lambda_c = 0.5$ denotes the auxiliary task weight which is smaller than 1 to make the region task dominate. This weight is decided using the grid search.

Finally, the overall loss function is

$$L = L_g + L_l + \beta\|W\|_2^2 \tag{6}$$

where $\beta = 10^{-4}$ denotes the balance of the weight decay term and W denotes the parameters of the entire network.

C. Multiple Receptive Field Model Ensemble

Due to the limited accuracy of single models, the ensemble of multiple models is considered as an effective practice to perform robust inference, and is widely employed in practical cases, at a cost of computational expensiveness. Encouraged by the dramatically accelerated framework, the extra cost is acceptable.

To cover contexts of different scales, we construct the proposed 3-D RU-Net with different receptive fields by adding dilation to the convolutional layer. Specifically, as is illustrated in Table I, we first construct an original 3-D R-U-Net of receptive field $26 \times 64 \times 64$, called 3-D RU-Net (RF64). Next, we tune the dilation rate of *ResBlock*3 as 2, enlarging the receptive field to $26 \times 88 \times 88$ and formulate 3-D RU-Net (RF88); we further tune the dilation rates of *ResBlock*2, *ResBlock*3, and *ResBlock*4 as 2 and construct a 3-D R-U-Net of receptive field $26 \times 112 \times 112$ called 3-D RU-Net (RF112).

In the inference stage, as is shown in Fig. 3, three networks' outputs are averaged to generate the final prediction. Major voting produces similar scores and is therefore not discussed.

This is a generalization to the multiresolution strategy proposed in [52] that applies identical receptive field to images with different spatial resolutions, which is actually formulating different spatial receptive fields. Such generalization obtains rid of detail-losing downsampling and allows each model contribute to boundary details equally.

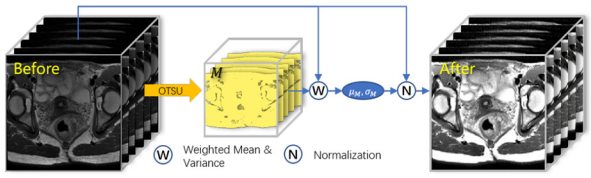


Fig. 4. Preprocessing: instead of ubiquitous mean and variance computing, in-body mean μ_M and in-body variance σ_M are computed inside the body mask M extracted using OTSU [59] thresholding. Then, μ_M and σ_M are used for normalization.

III. Experiments

A. Dataset and Preprocessing

1) Dataset: The dataset contains a total of 64 3-D MR images of the pelvic cavity of still T2 modality. The samples' spacing rates, which represent the physical dimensions of 3-D voxels along the z-, y-, and x-axes, range from 3.6 mm × 0.31 mm × 0.31 mm to 4.0 mm × 1.0 mm × 1.0 mm. The maximum volume dimension was $40 \times 512 \times 512$ voxels, whose spatial coverage is 160 mm × 512 mm × 512 mm and close to [60]'s spatial coverage. Target areas were labeled voxelwisely by one experienced radiologist yet the quality control was performed by three senior radiologists. Contour labels were automatically generated from the region labels of one-voxel thickness using erosion and subtraction operations. A 3-D image has mostly one and up to two RoIs containing cancerous tissues.

2) Preprocessing: Different spacing rates are normalized to 4.0 mm × 1.0 mm × 1.0 mm as default called HighRes. Some methods listed in Table II employ downsampled image sets, namely, LowRes set of 4.0 mm × 2.0 mm × 2.0 mm spacing and MidRes set of 4.0 mm × 1.5 mm × 1.5 mm spacing.

As is illustrated in Fig. 4, to normalize the intensities of input images acquired under different imaging configurations and field of views, we perform a body mask-weighted intensity

TABLE I
PARAMETERS AND GPU MEMORY FOOTPRINT TRACKING GIVEN AN INPUT VOLUME OF SIZE $40 \times 256 \times 320$ AND ROI SIZE OF $24 \times 96 \times 96$

Model Name	Part Name	Layer Name	Kernel/Stride	Nodes	Channels	Shape	Receptive Field 1	Receptive Field 2	Receptive Field 3	Memory Footprint	Total Footprint
3D RU-Net	Global Image Encoder	ResBlock1	$1 \times 3 \times 3$	9	48	$40 \times 256 \times 320$	$1 \times 7 \times 7$	$1 \times 7 \times 7$	$1 \times 7 \times 7$	5400.01 MB	9698.01 MB
		MaxPooling1	$1 \times 1/2 \times 1/2$	1	48	$40 \times 128 \times 160$	-	-	-	150.01 MB	
		ResBlock2	$3 \times 3 \times 3$	9	96	$40 \times 128 \times 160$	$7 \times 20 \times 20$	$7 \times 20 \times 20$	$7 \times 34 \times 34$	2699.38 MB	
		MaxPooling2	$1/2 \times 1/2 \times 1/2$	1	96	$20 \times 64 \times 80$	-	-	-	37.50 MB	
		ResBlock3	$3 \times 3 \times 3$	9	192	$20 \times 64 \times 80$	$20 \times 46 \times 46$	$20 \times 70 \times 70$	$20 \times 82 \times 82$	674.99 MB	
		Locator (sigmoid)	$1 \times 1 \times 1$	1	1	$32 \times 64 \times 80$	$20 \times 46 \times 46$	$20 \times 70 \times 70$	$20 \times 82 \times 82$	0.38MB	
	RoI Pyramid Layer	RoI Tensor1	-	1	48	$24 \times 96 \times 96$	$1 \times 7 \times 7$	$1 \times 7 \times 7$	$1 \times 7 \times 7$	40.50 MB	
		RoI Tensor2	-	1	96	$24 \times 48 \times 48$	$7 \times 20 \times 20$	$7 \times 20 \times 20$	$7 \times 34 \times 34$	20.25 MB	
		RoI Tensor3	-	1	192	$12 \times 24 \times 24$	$20 \times 46 \times 46$	$20 \times 70 \times 70$	$20 \times 82 \times 82$	5.06 MB	
	Local Region Decoder	UpConv1	$3 \times 3 \times 3$	1	96	$24 \times 48 \times 48$	-	-	-	20.25 MB	
		Add1	-	1	96	$24 \times 48 \times 48$	-	-	-	20.25 MB	
		ResBlock4	$2 \times 2 \times 2$	9	96	$24 \times 48 \times 48$	$26 \times 58 \times 58$	$26 \times 82 \times 82$	$26 \times 106 \times 106$	182.25 MB	
		UpConv2	$2 \times 2 \times 2$	1	48	$24 \times 96 \times 96$	-	-	-	40.50 MB	
		Add2	-	1	48	$24 \times 96 \times 96$	-	-	-	40.50 MB	
		ResBlock5	$1 \times 3 \times 3$	9	48	$24 \times 96 \times 96$	$26 \times 64 \times 64$	$26 \times 88 \times 88$	$26 \times 112 \times 112$	364.50 MB	
		SegHead1 (sigmoid)	$1 \times 1 \times 1$	1	1	$24 \times 96 \times 96$	$26 \times 64 \times 64$	$26 \times 88 \times 88$	$26 \times 112 \times 112$	0.84 MB	
		SegHead2 (sigmoid)	$1 \times 1 \times 1$	1	1	$24 \times 96 \times 96$	$26 \times 64 \times 64$	$26 \times 88 \times 88$	$26 \times 112 \times 112$	0.84 MB	
3D U-Net	Encoder	ResBlock1	$1 \times 3 \times 3$	9	48	$40 \times 256 \times 320$	$1 \times 7 \times 7$	$1 \times 7 \times 7$	$1 \times 7 \times 7$	5400.01 MB	18874.43 MB
		MaxPooling1	$1 \times 1/2 \times 1/2$	1	48	$40 \times 128 \times 160$	-	-	-	150.01 MB	
		ResBlock2	$3 \times 3 \times 3$	9	96	$40 \times 128 \times 160$	$7 \times 20 \times 20$	$7 \times 20 \times 20$	$7 \times 34 \times 34$	2699.38 MB	
		MaxPooling2	$1/2 \times 1/2 \times 1/2$	1	96	$20 \times 64 \times 80$	-	-	-	37.50 MB	
		ResBlock3	$3 \times 3 \times 3$	9	192	$20 \times 64 \times 80$	$20 \times 46 \times 46$	$20 \times 70 \times 70$	$20 \times 82 \times 82$	674.99 MB	
	Decoder	UpConv1	$2 \times 2 \times 2$	1	96	$40 \times 90 \times 160$	-	-	-	300.00 MB	
		Add1	-	1	96	$40 \times 128 \times 160$	-	-	-	300.00 MB	
		ResBlock4	$3 \times 3 \times 3$	9	96	$40 \times 128 \times 160$	$26 \times 58 \times 58$	$26 \times 82 \times 82$	$26 \times 106 \times 106$	2700.01 MB	
		UpConv2	$2 \times 2 \times 2$	1	48	$40 \times 256 \times 320$	-	-	-	600.01 MB	
		Add2	-	1	48	$40 \times 256 \times 320$	-	-	-	600.01 MB	
		ResBlock5	$1 \times 3 \times 3$	9	48	$40 \times 256 \times 320$	$26 \times 64 \times 64$	$26 \times 88 \times 88$	$26 \times 112 \times 112$	5400.01 MB	
		SegHead1 (sigmoid)	$1 \times 1 \times 1$	1	1	$40 \times 256 \times 320$	$26 \times 64 \times 64$	$26 \times 88 \times 88$	$26 \times 112 \times 112$	12.50 MB	

normalization to exclude the affect of inconsistent body-to-background ratios and intensity ranges. By OTSU [59] thresholding, body masks M are extracted where in-body $m_i = 1$ and other $m_i = 0$. The in-body mean intensity and standard deviation are computed according to the following formulas:

$$\mu_M = \frac{1}{\sum_i m_i} \sum_i m_i x_i \tag{7}$$

$$\sigma_M = \sqrt{\frac{1}{\sum_i m_i} \sum_i m_i (x_i - \mu_M)^2} \tag{8}$$

where $x_i \in X$ denotes the intensity of the ith voxel from image X and $m_i \in M$ denotes the ith value of body mask M. Then, the image is normalized using aforementioned μ_M and σ_M according to standard normalization criterion.

3) Body Cropping: Before feeding the images to the network, we crop the input image according to the bounding box of the body mask M to further reduce the GPU memory footprint, as is illustrated in Fig. 2.

4) Augmentation: In addition, in the training stage, we performed on-the-fly data augmentation when feeding training samples. Applied random operations include from $0.9\times$ to $1.1\times$ scaling, flipping with respect to the x-axis, $0.9\times$ to $1.1\times$ intensity jittering, and RoI translation that shifts the RoI center by -50% to 50% width along each axis.

B. Implementation Details

1) Hyperparameters: The network's detailed connectivity and kernel configuration are illustrated in Table I. Specifically, to fit the anisotropic spacing of the acquired dataset which has larger spacing along the z-axis, flat kernels of $1\times3\times3$, pooling rate of $1 \times 1/2 \times 1/2$, and upsampling rate of $1 \times 2 \times 2$ are employed by the input and output blocks, that is, ResBlock1, MaxPooling1, UpConv2, and ResBlock5. Initial experiments demonstrate that adding MaxPoolings, ResBlocks, or channels does not improve the performance, hence we tune receptive field setting by applying dilated convolution rather than adding layers.

2) Training Process: The backbone network was initialized using criterion proposed in [61], then pretrained using our previous work's patchwise HL-FCN [52]. We used the Adam [62] optimizer at a learning rate of 10^{-4}. The weights of convolution kernels were penalized with the 10^{-4} L2 norm for better generalization capability. Then, we first train the RoI locator with loss L_g until evaluation loss no longer decrease, then jointly train the full model with loss L.

C. Evaluation Metrics

1) Dice Similarity Coefficient: The dice similarity coefficient (DSC) measures a general overlap rate that equally assigns significance to recall rate and false-positive rate. DSC is denoted as

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

where the metric is scored in [0, 1]. Better prediction generates a score closer to 1.0.

2) Voxelwise Recall Rate: We also employ the voxelwise recall rate to evaluate the recall capability of different methods

$$\mathrm{Recall} = \frac{|P \cap G|}{|G|}. \tag{10}$$

3) Average Symmetric Surface Distance: The shortest distance between an arbitrary voxel of one volume's surface and another volume's surface is defined as

$$d(a_k, B) = \min_{b_i \in S(B), a_k \in S(A)} d(a_k, b_i) \tag{11}$$

where a_k denotes the kth voxel from extracted surface $S(A)$ of volume A, b_i denotes the ith voxel from extracted surface $S(B)$ of volume B, and $d(a_k, b_i)$ denotes the Euclidean distance between a_k and b_i. Then, the evaluation value is defined as

$$\mathrm{ASD} = \frac{\sum_{p_k \in S(P)} d(p_k, G) + \sum_{g_k \in S(G)} d(g_k, P)}{|S(P)| + |S(G)|} \tag{12}$$

TABLE II
COMPARING DIFFERENT METHODS OVER DIFFERENT METRICS REGARDING ACCURACY (DSC, RECALL RATE, ASD, AND FALSE-POSITIVE-RoI TO RoI-NUMBER RATIO FP_{RoIs}/N_{RoIs}) AND EFFICIENCY (TIME CONSUMPTION FOR RoI LOCALIZATION AND IN-REGION SEGMENTATION USING GPU OR CPU). THE METHODS ARE DIVIDED INTO THIRD-PARTY METHODS AND ABLATION STUDIES. IN ADDITION, MODEL ENSEMBLE STRATEGIES, NAMELY, MULTIRESOLUTION (MULTIRES) AND MULTIRECEPTIVE FIELD (MULTIRF), ARE COMPARED

Contribution	Method	Resolution	DSC[%]	Recall[%]	ASD[mm]	FP_{RoIs}/N_{RoIs}	Loc/Seg GPU Time[s]	Loc/Seg CPU Time[s]
	3D RU-Net+MHL+MultiRF	HighRes	**75.5±10.7**	77.8±14.8	**2.45±3.26**	-	0.61	38.10
+Dilation	3D RU-Net (RF112)+MHL	HighRes	74.2±10.6	78.6±13.9	3.02±3.55	0.7/1.8	0.22/0.01	15.30/0.79
+Global Context	3D RU-Net (RF88)+MHL	HighRes	73.7±10.5	75.4±14.8	2.83±3.57	1.1/2.2	0.17/0.01	10.44/0.45
	3D RU-Net (RF64)+MHL	HighRes	72.7±12.5	76.2±17.2	2.62±3.05	1.6/2.7	0.15/0.01	8.98/0.34
	3D RU-Net+MHL+MultiRes	-	74.9±10.3	76.1±14.8	2.46±3.35	-	0.33	17.29
+Resampling	3D RU-Net (RF64)+MHL	LowRes	73.5±10.4	75.7±14.4	2.46±2.98	0.7/1.8	0.04/0.01	2.11/0.08
+Global Context	3D RU-Net (RF64)+MHL	MidRes	73.1±10.2	74.0±14.5	2.91±3.65	1.4/2.5	0.07/0.01	3.70/0.14
	3D RU-Net (RF64)+MHL	HighRes	72.7±12.5	76.2±17.2	2.62±3.05	1.6/2.7	0.15/0.01	8.98/0.34
	3D FCN+3D U-Net+HL+MultiRF	HighRes	73.4±11.6	78.5±15.3	3.10±3.71	-	0.72	44.45
+RoI Localization	3D FCN+3D U-Net (RF112)+HL	HighRes	72.0±12.1	78.2±15.5	3.41±4.14	1.1/2.2	0.22/0.03	15.30/1.77
	3D FCN+3D U-Net (RF88)+HL	HighRes	71.6±12.5	75.9±16.6	3.42±3.73	1.5/2.6	0.17/0.02	10.44/1.12
	3D FCN+3D U-Net (RF64)+HL	HighRes	71.7±11.9	79.1±14.9	3.56±4.07	2.0/3.1	0.15/0.02	8.98/0.96
	3D U-Net+HL+MultiRF [52]	HighRes	70.6±12.5	72.5±13.2	9.47±10.51	-	37.35	1533.2
+Dilation	3D U-Net (RF112)+HL [52]	HighRes	67.8±13.8	70.5±14.8	12.44±13.70	-	14.62	827.57
	3D U-Net (RF88)+HL [52]	HighRes	66.9±17.4	71.8±15.7	14.16±11.54	-	12.47	358.88
	3D U-Net (RF64)+HL [52]	HighRes	67.7±18.4	69.2±21.3	10.24±14.59	-	10.26	346.72
	3D U-Net (RF64)+HL+MultiRes [52]	-	72.1±13.9	72.2±17.2	3.83±4.95	-	18.11	616.70
+Contour Extraction	3D U-Net (RF64)+HL [52]	LowRes	69.9±12.5	70.2±14.9	3.90±4.43	-	2.25	88.90
	3D U-Net (RF64)+HL [52]	MidRes	70.0±14.5	72.1±17.3	5.48±7.06	-	5.60	180.62
	3D U-Net (RF64)+HL [52]	HighRes	67.7±18.4	69.2±21.3	10.24±14.59	-	10.26	346.72
	3D U-Net (RF64)+DL+MultiRes [10]	-	69.9±13.7	72.4±18.0	4.18±5.89	-	18.11	616.70
	3D U-Net (RF64)+DL [10]	LowRes	68.5±13.8	68.5±19.5	4.19±5.75	-	2.25	88.90
Baseline	3D U-Net (RF64)+DL [10]	MidRes	67.3±15.3	70.2±17.7	5.70±7.31	-	5.60	180.62
	3D U-Net (RF64)+DL [10]	HighRes	66.0±18.2	70.9±22.0	10.32±12.11	-	10.26	346.40
	3D U-Net (RF64)+CE [9]	HighRes	61.7±19.2	57.3±23.9	4.26±4.35	-	10.26	346.40
	2D U-Net+3D U-Net+MultiRes	-	72.0±13.6	76.1±18.3	3.86±5.46	-	1.021	79.70
	2D U-Net+3D U-Net [36]	LowRes	70.2±12.4	74.5±15.8	4.11±5.03	4.2/5.3	0.15/0.02	7.68/0.53
	2D U-Net+3D U-Net [36]	MidRes	69.1±17.7	73.7±21.0	6.05±9.53	5.0/6.1	0.18/0.02	16.95/0.87
	2D U-Net+3D U-Net [36]	HighRes	69.4±14.1	76.2±18.2	6.23±8.77	6.0/7.1	0.25/0.03	38.29/1.22
Third-Party Methods	2D kU-Net+BDC-LSTM [17]	HighRes	69.3±13.1	79.1±16.7	7.81±6.88	-	0.51	39.22
	3D Mask R-CNN+MultiRes [41]	-	53.6±18.9	53.5±24.9	5.88±6.53	-	1.11	65.56
	3D Mask R-CNN [41]	HighRes	56.4±19.0	58.5±25.6	7.93±10.33	-	0.55	35.88
	3D Mask R-CNN [41]	MidRes	54.6±17.3	61.0±24.5	9.05±8.53	-	0.32	18.07
	3D Mask R-CNN [41]	LowRes	52.0±16.8	55.0±24.1	7.02±7.24	-	0.24	11.61
	Super-Voxel Clustering [5]	HighRes	62.6±14.9	60.2±18.2	6.54±5.96	-	-	15.13

where $|S(P)|$ and $|S(G)|$ denote the number of prediction volume's surface voxels and the number of ground-truth volume's surface voxels.

Specifically, this metric measures boundary fitness and is sensitive to failures such as debris outliers predicted far away from the colon region because the long distance makes up for the small size of the debris and produces a large error penalty. If a failure result has 0 recall rate, average symmetric surface distance (ASD) is not calculable. Instead, we cast a 50-mm penalty to it since that the statistical maximum radius of the tumors is smaller than 50 mm.

4) Average Inference Time: We include average inference time to evaluate speed in the inference stage using a GPU or CPU only. Since this metric is decided by the size of the input volume, the standard deviation is not evaluated. The tested methods are all performed on a workstation platform with 2x Xeon E5 CPU (8C16T) @ 2.4 GHz, 128-GB RAM, and an NVIDIA Titan Xp GPU with 12-GB GPU memory using the Ubuntu 16.04 system. The code is implemented with Python 3.6 and PyTorch and the inference speed is evaluated under the volatile mode. For methods combining RoI localization and in-region segmentation, the time consumptions of two stages are reported individually.

5) Average Number of False-Positive RoIs: RoI localization-based methods produce false-positive proposals. By the average number of false-positive RoIs (FP_{RoIs}) out of the number of detected RoIs (N_{RoIs}), we evaluate the precision of RoI localization and its effect to speed. This metric is not evaluated for anchor-based methods (3-D Mask R-CNN) because for each RoI, multiple bounding boxes are predicted and suppressed.

D. Results

For evaluation, four-fold cross-validation was conducted on 64 scans and their mean DSC scores are reported in Table II. Over two typical test samples, we compare segmentation results predicted by different methods and illustrate them in Fig. 5. While the proposed method presented a sensitive response to boundary details and retained general correctness, competing methods presented inferior boundary details (3-D U-Net, 3-D U-Net+DL, 3-D FCN+3-D U-Net, 3-D Mask R-CNN, and 2-D kU-Net+LSTM) or limited correctness (supervoxel clustering). Also, the proposed method's predicted regions matched its predicted contours, and the missegmented region predictions also associated with the failed parts of contour predictions, which highlight the mutual benefit of the regional task and the contour task. In addition, eight cancerous volumes predicted by the proposed method are visualized in Fig. 6 using the 3-D rendering module of SimpleITK [63]. Despite the background complexity, our method correctly located and segmented targets without being significantly misguided by nearby distractions thanks to the fully utilized global contexts. The observed major pattern of mistakes is that the model is sometimes confused about which Z slice to start or end.

1) GPU Footprint Tracking: In Table I, we compare the proposed 3-D RU-Net to vanilla 3-D U-Net under parameter

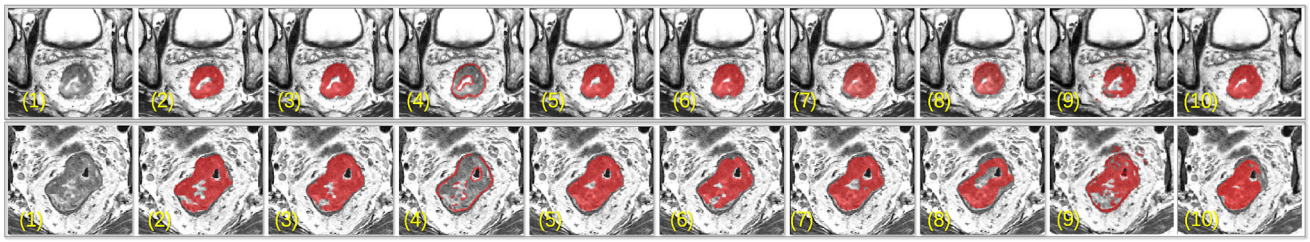


Fig. 5. Illustrations inside a chosen RoI of (1) cancerous region, (2) expert delineation, (3) proposed method (predicted regions), (4) proposed method (predicted contours), (5) 3-D U-Net + DL [10] (ensemble), (6) 3-D U-Net [9], (7) 3-D FCN + 3-D U-Net, (8) 3-D Mask R-CNN [41], (9) supervoxel clustering [5], and (10) 2-D kU-Net + LSTM [17].

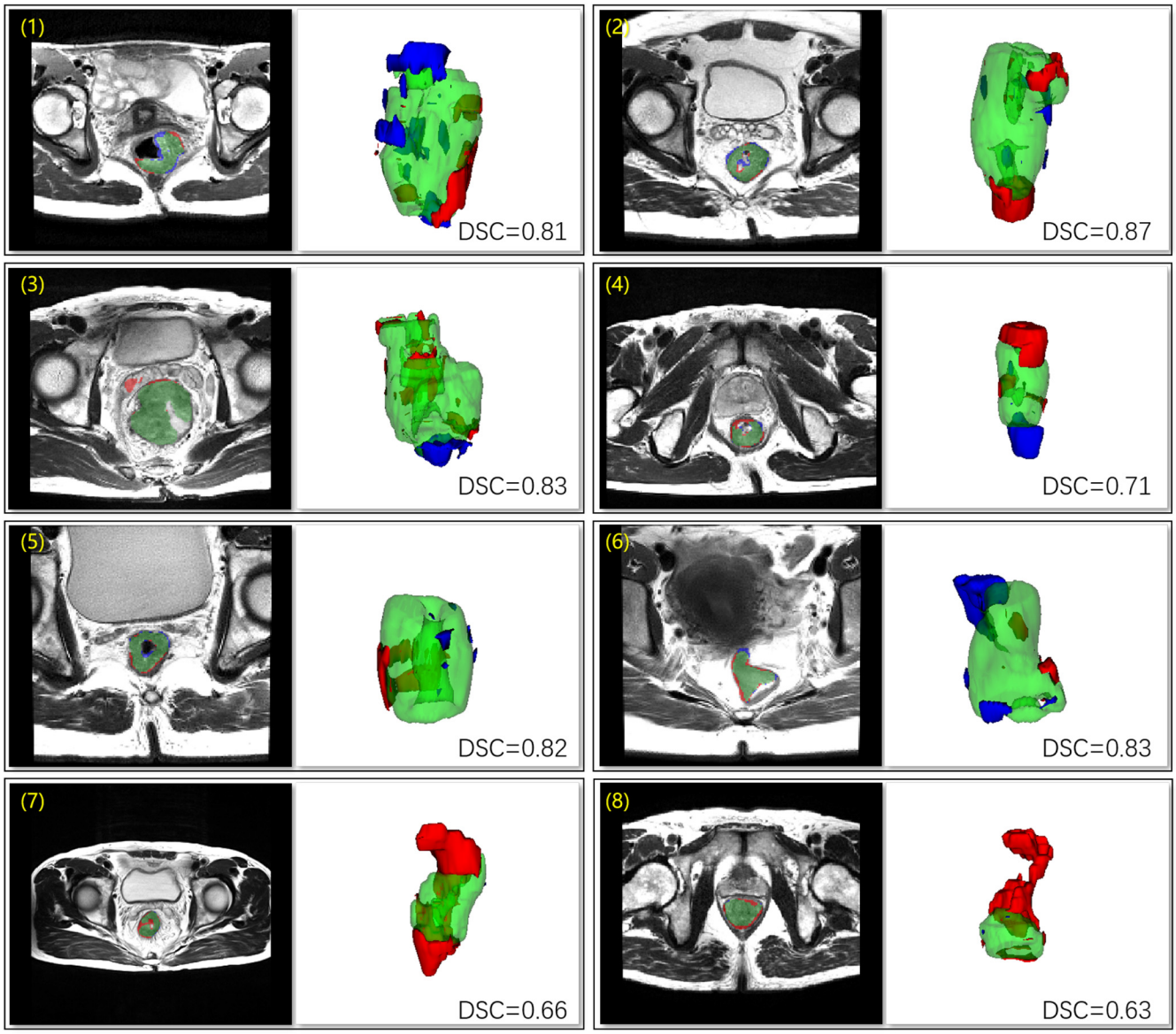


Fig. 6. Illustration of selected 2-D key slices and 3-D segmentation results from different patient cases numbered from (1) to (8). Semitransparent green indicates true positives; red indicates false positives; and blue indicates false negatives.

settings in Table I. With an input volume of size 160 × 256 × 320 mm, the 3-D RU-Net takes 9.7 GB while the 3-D U-Net takes 18874.81 MB. By enabling in-place computing, the ReLU activations become memory free. The memory footprint of standard U-Net further drops from 18.9 to 13.3 GB and the footprint of 3-D RU-Net drops from 9.7 to 6.5 GB.

Though not strictly defined, common colorectal MR imaging covers a maximum spatial range of 200 mm × 512 mm × 512 mm, as is the case of our dataset and [60]. By performing body cropping described in Section III-A3, this range can be reduced to 200 mm × 256 mm × 320 mm. According to our experiment, the largest trainable volume size

using a device with 12-GB GPU memory is increased from $48 \times 168 \times 168$ to $48 \times 256 \times 320$, whose spatial coverage is 192 mm $\times$ 256 mm $\times$ 320 mm and is close to the common range.

2) Ablation Studies: We conduct a full ablation study to evaluate the contribution of each proposed component, listed in the upper section of Table II.

First, the baseline of this article, namely, 3-D U-Net (RF64), is a part-based 3-D U-Net of $26 \times 64 \times 64$ receptive field trained toward cross-entropy (CE) loss [9] or DL [10] which shares the same encoder and decoder settings of the proposed method. Since that using DL demonstrated clearly better performance, 3-D U-Net+DL is used for further evaluation. Specifically, we acquired $(d, h, w) = (24, 96, 96)$ patches at a stride of 50% window overlapping for training and predicting. Three of this model are trained under different resolution settings, HighRes, MidRes, and LowRes, respectively, as were mentioned in Section III-A2, and the multiresolution ensemble model 3-D U-Net (RF64) + DL + MultiRes produces the final prediction. Clearly, downsampling the input image harms the predictions' detail richness, but this operation enlarges the physical receptive field without an extra computational burden and therefore produced better performance.

Second, we evaluate the contribution of DL formulated contour extraction side task. In the experiment, the same 3-D U-Net trained toward a dice formulated HL called 3-D U-Net (RF64) + HL [52] is trained following the baseline model's scheme and outperformed it by 2%. While enlarging the receptive field by downsampling worked well, by enlarging the receptive field by dilation without enlarging the patch size, the formed 3-D U-Net (RF88) + HL and 3-D U-Net (RF112) + HL did improve the performance. This suggests that the patch size limited the effective receptive field.

Third, we evaluate the contribution of performing 3-D RoI localization prior to in-region segmentation. To prove this contribution, we form an intermediate method called 3-D FCN + 3-D U-Net (RF64) + HL. This method consists of a standalone 3-D global image encoder trained toward global loss L_g for RoI localization and a full 3-D-U-Net of $26 \times 64 \times 64$ receptive field trained toward the HL L_l for in-region segmentation. These two networks work in a cascaded manner, producing 1% better result by eliminating false positives and much faster speed. We also tuned the receptive field from $26 \times 64 \times 64$ to $26 \times 88 \times 88$ and $26 \times 122 \times 122$, and did not notice a significant performance difference ($< 0.4\%$).

Finally, we evaluate the contribution of passing global context to the in-region segmentation branch. By connecting the 3-D FCN to the 3-D U-Net of the intermediate method, the proposed method 3-D RU-Net (RF64) + MHL is jointly trained toward the proposed MHL. Compared to 3-D FCN + 3-D U-Net (RF64) + HL, the proposed 3-D RU-Net (RF64) + MHL enjoyed a higher DSC score from 3-D FCN + 3-D U-Net (RF64) + HL's 71.7% to 72.7% due to the merit of passing global contexts from the global image encoder to the segmentation branch. Nevertheless, by enlarging the receptive field, the DSC score further increases from 72.7% to 74.2% while the nonjoint method witnessed very limited differences. Thanks to the elimination of redundant feature extraction, in-region segmentation is accelerated from 20 ms/RoI to less than 10 ms/RoI. In addition, compared to multiresolution ensemble's 74.9% DSC, enlarging receptive field by dilation produced higher 75.5% DSC due to better detail richness.

3) Third-Party Comparison: Next, we conducted further evaluation by comparing the proposed method to other third-party methods.

First, 2-D U-Net + 3-D U-Net proposed by [36] is another version of model cascading. This method employs a 2-D U-Net to coarsely segment the target for RoI localization, and a 3-D U-Net for fine in-region segmentation. Compared to 3-D FCN + 3-D U-Net and 3-D RU-Net, a 2-D U-Net produces significantly more false-positive candidates (larger $\mathrm{FP}_{\mathrm{RoIs}}$) with larger RoI length along the z-axis, which degrades the performance and speed.

Next, a 2-D U-Net+BDC-LSTM [17] is evaluated, whose kU-Net is employed for intraslice feature extraction and a bidirectional convolutional LSTM is used to explore intraslice features. Since patch size no longer limits the effective receptive field, we evaluated this method only using the HighRes dataset with a large designed receptive field as is proposed in [17]. It scored similarly compared to a 3-D U-Net+HL trained with the HighRes dataset, highlighting the effectiveness of intraslice LTSMs of identifying interslice connectivity. However, the limited utilization of 3-D context produced higher recall along with more false positives, and its single-model speed is significantly slower compared to the proposed method.

In addition, a 3-D-FPN-based Mask R-CNN is evaluated and got reasonable but inferior scores. First, region proposals are misclassified or poorly regressed, resulting in missing targets, false-positive targets, and incomplete targets. Second, for true positives, bin-fitting in-region features from heuristically selected feature maps degraded the detail richness and produced coarser masks.

Finally, we also evaluated a supervoxel clustering-based [5] method. Without the merit of discriminative 3-D deep features, supervoxels are inevitably oversegmented, undersegmented, and misclassified. In our experiments, one of the 64 targets went completely missing and significantly lowered the dice score, while some wrong supervoxels were chosen as the output mask.

IV. Discussion

In this article, we proposed a method to inherit easy to train and detail-preserving merits of volume-to-volume 3-D FCNs while acquiring fast RoI localization, target completeness, and global understanding. We combined a whole volume RoI localization model called a global image encoder and the in-region segmentation model called a local region decoder as a joint model called 3-D RoI-aware U-Net (3-D RU-Net). By sharing global context across the RoI localization and in-region segmentation tasks and elimination of redundant feature extraction, the proposed method demonstrates faster and more accurate performance.

Although our method achieved competitive results, there are several limitations.

First, as is illustrated in Fig. 6, the model is often confused about which slice to start or end, thus this significantly affects the score. As is illustrated in Table II, all competing methods including applying a bidirectional convolutional LSTM [17] did not thoroughly tackle this issue. As an explanation, this difficulty is data related and decision about starting and ending slice index can be observer dependent due to weak contrast in the border of cancerous tissues and low resolution along the z-axis.

Second, for this specific task without the need of discriminating different instances of tumors, we did not include instance separation capability in our design. If multiple adjacent targets are detected, there is no guarantee that they are properly separated. When applied to tasks requiring instance discrimination, extra modules, such as anchor classifier and regressor, should be added to the global image encoder template. However, due to the challenges to the anchor-based method, careful refinement to the bounding box predictions should be made with the potential help of the predicted masks.

Further works include cancer subtype classification, application to real-time 3-D imaging, and further memory optimization. First, as is suggested by clinical guideline [64], T1/T2 and T3/T4 colorectal cancer are of different danger extent and should follow different treatment routines. Therefore, it is important to distinguish T3 cases from T1/T2. From T1 to T4, the tumor gradually passes through the rectal wall, therefore an RoI feature describing the extent of wall passing should be acquired by accurate segmentation of colorectal tumors and rectal walls. Second, due to the proposed method's speed advantage, it is possible to extend it to real-time 3-D imaging applications, yet some further adjustments should be included to utilize interframe context for motion management. In addition, to further facilitate the training of large 3-D whole volumes, trainable volume size, and batch size can be further enlarged by reducing GPU memory footprint, for example, adopting mixed-precision training [65] and virtualized deep neural networks [66].

V. Conclusion

In this article, we proposed a joint framework for fully automatic whole volume colorectal cancer segmentation referred to as 3-D RoI-aware U-Net (3-D RU-Net). We emphasized the importance and effectiveness of integrating RoI localization and in-region segmentation fed with globally encoded features to perform fast and accurate whole volume segmentation. The proposed method enables the merit of enlarging receptive fields originally limited by GPU memory capacity and ensembles models with different receptive field settings. A dice-based Multitask HL function is present to smoothen the training process. The experimental results evaluated each proposed component's contribution and demonstrated the proposed method's advantage over competing methods in terms of accuracy and speed. In principle, the proposed framework is scalable enough to be adopted to other medical image segmentation tasks.

References

[1] *World Cancer Report 2014*, World Health Org., Geneva, Switzerland, Feb. 2015.

[2] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2017," *CA Cancer J. Clin.*, vol. 67, no. 1, pp. 7–30, 2017, doi: 10.3322/caac.21387.

[3] L. Kavalcova, R. Skaba, M. Kyncl, B. Rouskova, and A. Prochazka, "The diagnostic value of MRI fistulogram and MRI distal colostogram in patients with anorectal malformations," *J. Pediatric Surg.*, vol. 48, no. 8, pp. 1806–1809, 2013.

[4] S. Rathore, M. Hussain, A. Ali, and A. Khan, "A recent survey on colon cancer detection techniques," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 10, no. 3, pp. 545–563, May/Jun. 2013.

[5] B. Irving *et al.*, "Automated colorectal tumour segmentation in DCE-MRI using supervoxel neighbourhood contrast characteristics," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2014, pp. 609–616.

[6] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder–decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intel.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[7] H. Chen, Q. Dou, X. Wang, J. Qin, J. C. Y. Cheng, and P. A. Heng, "3D fully convolutional networks for intervertebral disc localization and segmentation," in *Proc. Int. Conf. Med. Imaging Virtual Reality*, 2016, pp. 375–382.

[8] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2015, pp. 234–241.

[9] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning dense volumetric segmentation from sparse annotation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2016, pp. 424–432.

[10] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. IEEE Int. Conf. 3D Vis.*, 2016, pp. 565–571.

[11] L. Yu, X. Yang, H. Chen, J. Qin, and P.-A. Heng, "Volumetric ConvNets with mixed residual connections for automated prostate segmentation from 3D MR images," in *Proc. AAAI Conf. Artif. Intel.*, 2017, pp. 66–72.

[12] H. Chen, Q. Dou, L. Yu, J. Qin, and P. A. Heng, "VoxResNet: Deep voxelwise residual networks for brain segmentation from 3D MR images," *Neuroimage*, vol. 170, pp. 446–455, Apr. 2018.

[13] Q. Dou *et al.*, "3D deeply supervised network for automated segmentation of volumetric medical images," *Med. Image Anal.*, vol. 41, pp. 40–54, Oct. 2017.

[14] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015. [Online]. Available: arXiv:1511.07122.

[15] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Soc. Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2881–2890.

[16] J. Wang *et al.*, "A deep learning based auto segmentation of rectal tumors in MR images," *Med. Phys.*, vol. 45, no. 6, pp. 3532–3542, 2018.

[17] J. Chen, L. Yang, Y. Zhang, M. S. Alber, and D. Z. Chen, "Combining fully convolutional and recurrent neural networks for 3D biomedical image segmentation," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 3036–3044.

[18] H. R. Roth *et al.*, "DeepOrgan: Multi-level deep convolutional networks for automated pancreas segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2015, pp. 556–564.

[19] H. R. Roth *et al.*, "A new 2.5D representation for lymph node detection using random sets of deep convolutional neural network observations," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, vol. 17, 2014, pp. 520–527.

[20] Q. Dou *et al.*, "Automatic detection of cerebral microbleeds from MR images via 3D convolutional neural networks," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1182–1195, May 2016.

[21] D. Nie, L. Wang, E. Adeli, C. Lao, W. Lin, and D. Shen, "3-D fully convolutional networks for multimodal isointense infant brain image segmentation," *IEEE Trans. Cybern.*, vol. 49, no. 3, pp. 1123–1136, Mar. 2019.

[22] Q. Yu, Y. Shi, J. Sun, Y. Gao, Y. Dai, and J. Zhu, "Crossbar-Net: A novel convolutional network for kidney tumor segmentation in ct images," 2018. [Online]. Available: arXiv:1804.10484.

[23] T. Rohlfing, D. B. Russakoff, and C. R. Maurer, "Performance-based classifier combination in atlas-based image segmentation using expectation-maximization parameter estimation," *IEEE Trans. Med. Imag.*, vol. 23, no. 8, pp. 983–994, Aug. 2004.

[24] S. Klein, U. Van-Der-Heide, I. Lips, M. Van-Vulpen, M. Staring, and J. Pluim, "Automatic segmentation of the prostate in 3D MR images by atlas matching using localized mutual information," *Med. Phys.*, vol. 35, no. 4, pp. 1407–1417, 2008.

[25] S. Murphy, B. Mohr, Y. Fushimi, H. Yamagata, and I. Poole, "Fast, simple, accurate multi-atlas segmentation of the brain," in *Proc. Int. Workshop Biomed. Image Reg.*, 2014, pp. 1–10.

[26] B. Hariharan, P. A. Arbeláez, R. B. Girshick, and J. Malik, "Hypercolumns for object segmentation and fine-grained localization," in *Proc. IEEE Soc. Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 447–456.

[27] J. Dai, K. He, and J. Sun, "Convolutional feature masking for joint object and stuff segmentation," in *Proc. IEEE Soc. Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3992–4000.

[28] P. H. O. Pinheiro, R. Collobert, and P. Dollar, "Learning to segments objects candidates," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2015, pp. 1990–1998.

[29] X. Li, H. Chen, X. Qi, Q. Dou, C. W. Fu, and P. A. Heng, "H-DenseUNet: Hybrid densely connected U-Net for liver and tumor segmentation from ct volumes," *IEEE Trans. Med. Imag.*, vol. 37, no. 12, pp. 2663–2674, Dec. 2018.

[30] F. Liao, X. Chen, X. Hu, and S. Song, "Estimation of the volume of the left ventricle from MRI images using deep neural networks," *IEEE Trans. Cybern.*, vol. 49, no. 2, pp. 495–504, Feb. 2019.

[31] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, "Simultaneous detection and segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 297–312.

[32] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *Int. J. Comput. Vis.*, vol. 104, no. 2, pp. 154–171, 2013.

[33] P. Arbeláez, J. Pont-Tuset, J. T. Barron, F. Marques, and J. Malik, "Multiscale combinatorial grouping," in *Proc. IEEE Soc. Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 328–335.

[34] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[35] M. Tang, Z. Zhang, D. Cobzas, M. Jagersand, and J. L. Jaremko, "Segmentation-by-detection: A cascade network for volumetric medical image segmentation," in *Proc. IEEE Int. Symp. Biomed. Imag.*, 2018, pp. 1356–1359.

[36] A. Balagopal *et al.*, "Fully automated organ segmentation in male pelvic CT images," *Phys. Med. Biol.*, vol. 63, no. 24, 2018, Art. no. 245015.

[37] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1440–1448.

[38] E. L. Denton, W. Zaremba, J. Bruna, Y. LeCun, and R. Fergus, "Exploiting linear structure within convolutional networks for efficient evaluation," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2014, pp. 1269–1277.

[39] J. Xue, J. Li, and Y. Gong, "Restructuring of deep neural network acoustic models with singular value decomposition," in *Proc. Interspeech*, 2013, pp. 2365–2369.

[40] J. Dai, K. He, and J. Sun, "Instance-aware semantic segmentation via multi-task network cascades," in *Proc. IEEE Soc. Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3150–3158.

[41] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.

[42] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 1, 2017, p. 4.

[43] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proc. IEEE Soc. Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8759–8768.

[44] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.

[45] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Facial landmark detection by deep multi-task learning," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 94–108.

[46] H. Chen *et al.*, "Ultrasound standard plane detection using a composite neural network framework," *IEEE Trans. Cybern.*, vol. 47, no. 6, pp. 1576–1586, Jun. 2017.

[47] H. Chen, X. Qi, L. Yu, Q. Dou, J. Qin, and P. A. Heng, "DCAN: Deep contour-aware networks for object instance segmentation from histology images." *Med. Image Anal.*, vol. 36, pp. 135–146, Feb. 2017.

[48] H. Shen, R. Wang, J. Zhang, and S. J. Mckenna, "Boundary-aware fully convolutional network for brain tumor segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2017, pp. 433–441.

[49] D. Qi, C. Hao, L. Yu, Q. Jing, and P. A. Heng, "Multilevel contextual 3-D CNNs for false positive reduction in pulmonary nodule detection," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 7, pp. 1558–1567, Jul. 2017.

[50] K. Kamnitsas *et al.*, "Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation," *Med. Image Anal.*, vol. 36, pp. 61–78, Feb. 2017.

[51] B. Xu *et al.*, "Orchestral fully convolutional networks for small lesion segmentation in brain MRI," in *Proc. IEEE Int. Symp. Biomed. Imag.*, 2018, pp. 889–892.

[52] Y.-J. Huang *et al.*, "HL-FCN: Hybrid loss guided FCN for colorectal cancer segmentation," in *Proc. IEEE Int. Symp. Biomed. Imag.*, 2018, pp. 195–198.

[53] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014. [Online]. Available: arXiv:1409.1556.

[54] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Soc. Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[55] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten, "Densely connected convolutional networks," in *Proc. IEEE Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 1, 2017, p. 3.

[56] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," 2016. [Online]. Available: arXiv:1607.08022.

[57] Y. Wu and K. He, "Group normalization," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.

[58] A. Shrivastava, A. Gupta, and R. B. Girshick, "Training region-based object detectors with online hard example mining," in *Proc. IEEE Soc. Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 761–769.

[59] N. Otsu, "A threshold selection method from gray-level histogram," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. SMC-9, no. 1, pp. 62–66, Jan. 1979.

[60] D. Mahapatra *et al.*, "Automatic detection and segmentation of Crohn's disease tissues from abdominal MRI," *IEEE Trans. Med. Imag.*, vol. 32, no. 12, pp. 2332–2347, Dec. 2013.

[61] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1026–1034.

[62] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, 2015.

[63] B. C. Lowekamp, D. T. Chen, L. Ibáñez, and D. Blezek, "The design of simpleitk," *Front. Neuroinformat.*, vol. 7, p. 45, Dec. 2013.

[64] B. Glimelius *et al.*, "Rectal cancer: ESMO clinical practice guidelines for diagnosis, treatment and follow-up," *Ann. Oncol.*, vol. 24, no. 6, pp. 22–40, 2013.

[65] P. Micikevicius *et al.*, "Mixed precision training," in *Proc. Int. Conf. Learn. Represent.*, 2018.

[66] M. Rhu, N. Gimelshein, J. Clemons, A. Zulfiqar, and S. W. Keckler, "VDNN: Virtualized deep neural networks for scalable, memory-efficient neural network design," in *Proc. Ann. IEEE/ACM Int. Symp. Microarchit.*, 2016, p. 18.

Yi-Jie Huang received the bachelor's degree in automation from the Huazhong University of Science and Technology, Wuhan, China, in 2015. He is currently pursuing the Ph.D. degree with the Department of Automation, Institute of Image Processing and Pattern Recognition, Shanghai, China.

He is a Research Intern with the Department of Research and Development, Imsight Medical Technology Company Ltd., Hong Kong. His research interests include tumor detection, segmentation, and annotation-efficient deep learning for medical imaging.

Qi Dou (Member, IEEE) received the bachelor's degree in biomedical engineering from Beihang University, Beijing, China, in 2014, and the Ph.D. degree from the Department of Computer Science and Engineering, Chinese University of Hong Kong, Hong Kong, in 2018.

She is currently an Assistant Professor with the Department of Computer Science and Engineering, Chinese University of Hong Kong. Her research interests are in the interdisciplinary fields of medical image analysis and artificial intelligence, for improving lesion detection, anatomical structure computation, and surgical robotics perception, with an impact to advance disease diagnosis and robot-assisted intervention via machine intelligence.

Zi-Xian Wang received the bachelor's degree in clinical medicine from Sun Yat-sen University, Guangzhou, China, in 2015, and the M.D. degree in oncology from Sun Yat-sen University Cancer Center, Guangzhou, in 2017.

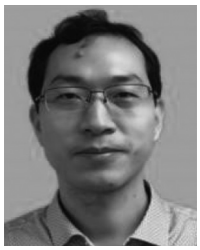
He is currently an Attending Physician with the Department of Medical Oncology, Sun Yat-sen University Cancer Center. His research interests are medical image analysis, cancer biomarkers, novel anticancer therapeutic targets, and clinical trial design.

Li-Zhi Liu received the bachelor's degree in clinical medicine from Nanchong University, Nanchang, China, in 1996, and the M.D. and Ph.D. degrees from Sun Yat-sen University Cancer Center, Guangzhou, China, in 2016.

He is currently a Senior Radiologist with Sun Yat-sen University Cancer Center. His research interests are medical image analysis and artificial intelligence.

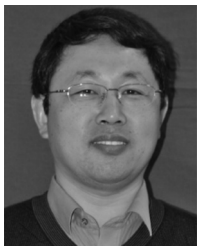
Ying Jin received the bachelor's degree in clinical medicine from Sun Yat-sen University, Guangzhou, China, in 2011, and the M.D. degree in oncology from Sun Yat-sen University Cancer Center, Guangzhou, in 2013.

She is currently an Attending Physician with the Department of Medical Oncology, Sun Yat-sen University Cancer Center. Her research interests are medical image analysis, cancer biomarkers, novel anticancer therapeutic targets, and clinical trial design.



Chao-Feng Li received the bachelor's degree in information and computing science from Henan University, Kaifeng, China, in 2004, and the Ph.D. degree in epidemiology and biostatistics from Sun Yat-sen University, Guangzhou, China, in 2014.

He is currently an Engineer with the Department of Information and Technology, Sun Yat-sen University Cancer Center, Guangzhou. His research interests are medical image analysis and artificial intelligence.



Lisheng Wang received the M.S. degree in mathematics and the Ph.D. degree in electronic and information engineering from Xi'an Jiaotong University, Xi'an, China, in 1993 and 1999, respectively.

He is currently a Professor with the Department of Automation, Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai, China. He is also an Adjunct Professor with the Shanghai Key Laboratory of Molecular Imaging, Shanghai University of Medicine and Health Sciences, Shanghai. His research interests include analysis and visualization of 3-D biomedical images, computer-aided imaging diagnosis, and surgery planning.

Hao Chen received the bachelor's degree in information engineering from Beihang University, Beijng, China, in 2013, and the Ph.D. degree in computer science from the Chinese University of Hong Kong, Hong Kong, in 2017.

His research interests include medical image analysis, deep learning, object detection, and segmentation.

Dr. Chen received the Hong Kong Ph.D. Fellowship in 2013 and the MICCAI Young Scientist Publication Impact Award in 2019. Three of his works have been received the best paper awards.

Rui-Hua Xu received the bachelor's degree in clinical medicine from Nanchang University, Nanchang, China, in 1988, and the M.D. and Ph.D. degrees in oncology from Sun Yat-sen University Cancer Center, Guangzhou, China, in 2000.

He is currently the President of Sun Yat-sen University Cancer Center, where he is the Director of the State Key Laboratory of Oncology in South China. His research interests are medical image analysis, cancer biomarkers, novel anticancer therapeutic targets, and clinical trial design.