Automatic multi-organ segmentation in computed tomography images using hierarchical convolutional neural network

Sharmin Sultana, Adam Robinson, Daniel Y. Song, and Junghoon Lee*
Johns Hopkins University, Department of Radiation Oncology and Molecular Radiation Sciences, Baltimore, Maryland, United States

Abstract

Purpose: Accurate segmentation of treatment planning computed tomography (CT) images is important for radiation therapy (RT) planning. However, low soft tissue contrast in CT makes the segmentation task challenging. We propose a two-step hierarchical convolutional neural network (CNN) segmentation strategy to automatically segment multiple organs from CT.

Approach: The first step generates a coarse segmentation from which organ-specific regions of interest (ROIs) are produced. The second step produces detailed segmentation of each organ. The ROIs are generated using UNet, which automatically identifies the area of each organ and improves computational efficiency by eliminating irrelevant background information. For the fine segmentation step, we combined UNet with a generative adversarial network. The generator is designed as a UNet that is trained to segment organ structures and the discriminator is a fully convolutional network, which distinguishes whether the segmentation is real or generator-predicted, thus improving the segmentation accuracy. We validated the proposed method on male pelvic and head and neck (H&N) CTs used for RT planning of prostate and H&N cancer, respectively. For the pelvic structure segmentation, the network was trained to segment the prostate, bladder, and rectum. For H&N, the network was trained to segment the parotid glands (PG) and submandibular glands (SMG).

Results: The trained segmentation networks were tested on 15 pelvic and 20 H&N independent datasets. The H&N segmentation network was also tested on a public domain dataset ($N = 38$) and showed similar performance. The average dice similarity coefficients (mean $\pm$ SD) of pelvic structures are 0.91 ± 0.05 (prostate), 0.95 ± 0.06 (bladder), 0.90 ± 0.09 (rectum), and H&N structures are 0.87 ± 0.04 (PG) and 0.86 ± 0.05 (SMG). The segmentation for each CT takes <10 s on average.

Conclusions: Experimental results demonstrate that the proposed method can produce fast, accurate, and reproducible segmentation of multiple organs of different sizes and shapes and show its potential to be applicable to different disease sites.

 [DOI: 10.1117/1.JMI.7.5.055001]

Keywords: deep learning; segmentation; hierarchical convolutional neural network; radiotherapy.

Paper 20086RRR received Apr. 13, 2020; accepted for publication Sep. 28, 2020; published online Oct. 14, 2020.

1 Introduction

Radiation therapy (RT) is widely used for treating cancer patients where high-energy radiation is used to kill cancer cells. The efficacy of RT depends on accurate delivery of therapeutic radiation dose to the target while sparing adjacent healthy tissues for which accurate segmentation of the target tumor and organs at risk (OARs) is critical. Manual contouring by radiation oncologists is

*Address all correspondence to Junghoon Lee, junghoon@jhu.edu

2329-4302/2020/$28.00 © 2020 SPIE

still considered as the gold standard in current clinical practice, but it is very time consuming and the quality of the segmentation varies depending on the physician's knowledge and experience. Computed tomography (CT) is used as the reference image for radiotherapy planning as it offers the electron density information needed for dose calculation. However, poor soft-tissue contrast in CT images makes the contouring process challenging, thus yielding to large inter- and intra-observer contouring variability.[1–4]

Automatic organ segmentation has been an active research area for the last few decades. Among existing automatic segmentation methods, state-of-the-art methods include (but are not limited to) atlas-based, model-based, and learning-based methods. In atlas-based methods, atlas images are registered to the image to be segmented followed by atlas label propagation to the target image to get the final segmentation. Since a single atlas cannot perfectly fit every patient, use of multiple atlases has become a standard baseline for atlas-based segmentation.[5–9] Although multi-atlas-based segmentation has been widely adopted with state-of-the-art segmentation quality, it requires a significant amount of time as it involves multiple registrations between the atlas and the target volumes. Model-based segmentation utilizes a deformable model and/or *a priori* knowledge of the target such as shape, intensity, and texture to constrain the segmentation process.[10–15] It is often used in combination with another method, e.g., multi-atlas-based segmentation, to further improve the segmentation quality. Such hybrid approaches have been applied to the delineation of head and neck (H&N) structures on CT images, showing promising results.[10,16–18] These model-based methods require fine-tuning parameters for every structure to be segmented and are sensitive to structure and image quality variations. Learning-based methods train a classifier or regressor from a pool of training images. Then the segmentation is generated by predicting the likelihood map.[19,20] Conventional learning-based approaches require hand-crafted feature extraction and the segmentation quality significantly depends upon the extracted features.

In the last several years, deep learning-based automatic segmentation has demonstrated its potential in accurate and consistent organ segmentation. In particular, convolutional neural network (CNN) became state-of-the-art in solving challenging image classification and segmentation problems due to its capability of extracting deep image features.[21–23] CNN architecture consists of several hidden convolutional layers followed by an activation function and pooling layers that enable automatic feature learning to accomplish classification and segmentation tasks. CNN-based segmentation approaches have been widely applied to both normal organ and tumor segmentation problems,[23,24] significantly improving the segmentation performance over other state-of-the-art methods. The introduction of a fully convolutional network (FCN) especially enabled image segmentation with arbitrary image sizes through efficient and robust learning and inference.[25] One of the most successful FCN approaches in medical image segmentation is UNet, an FCN with skip connection and capability of extracting contextual features from contracting layers and structural information from expansion layers.[26] UNet and its variants have shown very promising results in automatic medical image segmentation.[27–33]

CNN-based automatic segmentation approaches have been widely used for multiple organ segmentations in CT images. Roth et al.[34] proposed a multi-level CNN model for pancreas segmentation. Wang et al.[32] developed an FCN combined with dilated convolution and deep supervision for prostate segmentation. Men et al.[35] used a deep dilated CNN to segment the clinical target volume and pelvic OARs. Kazemifar et al.[31] proposed a CNN-based segmentation of CT male pelvic organs using two-dimensional (2-D) UNet. Balagopal et al.[36] proposed a cascaded multi-channel 2-D and three-dimensional (3-D) UNet with aggregated residual networks to segment male pelvic CT images. Wang et al.[33] used UNet with boundary-sensitive information to segment male pelvic structures. Dong et al.[37] proposed a UNet combined with generative adversarial network (GAN)[38] to segment multiple organs in thorax CT. Recently, CNN has been used for organ segmentation in magnetic resonance imaging (MRI),[28,30,39–42] achieving promising results utilizing excellent soft tissue contrast in MRI that improves identification of organ boundaries. However, considering that CT is used as the reference images for RT planning, automatic segmentation of CT images is highly desired despite being more challenging than segmenting magnetic resonance (MR) images.

CNN-based models have also shown promising results in segmenting H&N structures. Existing patch-based networks to segment H&N anatomical structures used 2-D/3-D local patches in a sliding window to identify OARs, which are unable to capture global features.[43–46] These patch-based methods also require pre- and postprocessing steps with additional parameters tuning. Chan et al.[47] proposed a cascaded CNN for single- and multi-task learning through transfer learning for H&N anatomy segmentation using a limited number of training samples. Hänsch et al.[48] explored the potential of 2-D, 2-D ensemble, and 3-D UNet-based models in parotid gland segmentation. Zhu et al.[49] proposed a UNet-based model with residual blocks for the segmentation of OARs in H&N where the data imbalance problem is addressed using a combination of dice and focal loss function. Tong et al.[50] presented a CNN model with GAN and shape constraint to segment H&N structures in MR and CT images.

Although many CNN-based approaches require a preprocessing step that reshapes or crops input images to a proper size and resolution for the segmentation network and/or try to segment multiple organs simultaneously, several groups investigated hierarchical or multi-level CNN segmentation approaches to automatically segment target organ(s) of interest. This strategy has been used to segment the pancreas,[34,51,52] esophagus,[53] and also multiple organs in the abdomen,[54] H&N,[55] and male pelvic region[36] in CT images, demonstrating improved efficiency and segmentation performance by localizing the target region and letting the segmentation network focus on the local region around the target organ to be segmented.

In this paper, we propose a hierarchical coarse-to-fine volumetric segmentation of CT where a coarse segmentation is produced using a multi-class 3-D UNet to determine organ-specific regions of interest (ROIs) and fine segmentation is performed utilizing GAN with a 3-D UNet as generator and FCN as discriminator. GAN contributes by providing learned parameters of accurate segmentation by distinguishing between real and generated segmentations and thus globally improving segmentation accuracy. Unlike many existing hierarchical/multi-level approaches that are based on 2-D/2.5-D images[34,53] and/or patches with sliding or tiling strategies,[34,51,52,54] our network is fully 3-D-based and processes the input 3-D CT volume and outputs associated multiple organ segmentations. Our method is similar to Balagopal et al.[36] (2-D localization network + modified 3-D UNet segmentation network) and Wang et al.[55] (3-D UNet bounding box network with sliding + 3-D UNet segmentation network) that first create bounding boxes for organs to be segmented followed by cropped image segmentation using (modified) UNet. Our approach creates organ-specific ROIs by 3-D multi-class segmentation, and the fine segmentation is performed using a 3-D UNet constrained by GAN, thus producing improved segmentation as reported in our experiments and results. Our initial approach and preliminary results on male pelvic CT segmentation were reported in a conference paper.[56] We further extended our approach to two different anatomical sites; male pelvic and H&N CT images, and trained and tested the proposed networks on a much larger cohort of data. The proposed approach showed its robust performance on both sites with state-of-the-art performances, demonstrating its generalizability to multiple sites and organs. H&N especially involves many OARs to be contoured requiring significant effort for contouring, therefore, it can benefit from the proposed method. We present complete methodology and extensive validation results using larger (internal and external) datasets in this paper.

2 Datasets and Preprocessing

We trained the proposed hierarchical CNN and tested to segment multiple organs in CT images of two different disease sites, male pelvic and H&N regions. Deidentified CT image data and associated contours drawn by the attending radiation oncologists were obtained from the patients' RT plan records under the approval of the institutional review board.

2.1 *Male Pelvic CT*

We obtained 290 pelvic CT images from prostate cancer patients who were treated by either external-beam RT (EBRT) or brachytherapy. Each patient had CT images of the pelvic region and manual contouring of the prostate, bladder, and rectum drawn by the attending radiation

oncologist. We used 275 datasets for training and the remaining 15 for testing. We augmented the training datasets to 1100 by applying random shifting, rotation, and flipping. The CT images have isotropic in-plane pixel sizes ranging from 1.17 to 1.36 mm and through-plane slice spacing of 3 mm. The sizes of the CT images range from $512 \times 512 \times [109 \text{ to } 284]$.

In these prostate cancer patients' CT images, there are fiducial markers (for EBRT cases) or brachytherapy seeds (for brachytherapy cases) implanted within the gland, creating very high-intensity values. We preprocessed CT images to remove such fiducials and seeds by automatically identifying them by thresholding and replacing their intensity values with a mean intensity of the neighboring gland voxels.

For coarse segmentation network training, we downsampled the original CT image to $5 \times 5 \times 5$ mm^3 voxel resolution with an image size of $118 \times 118 \times 72$. The coarse segmentation network produced a rough segmentation of the multiple organs of interest, i.e., prostate, bladder, and rectum, from which we extracted the ROI for each organ. The original CT was then cropped for each ROI, yielding cropped CT images of $96 \times 96 \times 32$, $112 \times 112 \times 64$, and $96 \times 96 \times 64$ voxels for the prostate, bladder, and rectum, respectively, with a voxel size of $2.3 \times 2.3 \times 3$ mm^3 for the successive fine segmentations. Notice that we did not use the original CT resolution due to varying in-plane resolutions and chose these ROIs and voxel sizes to sufficiently cover each organ, considering the graphics processing unit (GPU) memory for processing 3-D volume.

2.2 *H&N CT*

The performance of the proposed method was further evaluated for salivary glands segmentation in H&N CT. We collected 220 CT images from H&N cancer patients treated by EBRT. We used 200 datasets for training and 20 datasets for testing. The parotid glands (PG) and submandibular glands (SMG) were contoured by the attending radiation oncologists during the routine RT planning. The CT images had $512 \times 512 \times [100 \text{ to } 307]$ voxels with an isotropic in-plane pixel size of 0.94 to 1.36 mm and through-plane slice spacing of 3 to 3.13 mm.

Similar to the pelvic cases, we downsampled the original H&N CT to $5 \times 5 \times 5$ mm^3 voxel resolution with the volume dimension of $112 \times 112 \times 72$ voxels for the coarse segmentation network training. We augmented the training data by shifting, rotation, and flipping, yielding 800 training datasets. The coarse segmentation network produced a rough segmentation of the salivary glands from which we extracted the ROIs of the left and right side of PG and SMG. The original CT images were then cropped using the ROIs to produce a PG ROI image of $56 \times 56 \times 32$ voxels and SMG ROI image of $48 \times 48 \times 32$ voxels. Since salivary glands are much smaller than the pelvic organs, we were able process the fine segmentation at the original CT resolution. Also given the symmetric nature of both PG and SMG, we merged the flipped left PG and SMG with the right PG and SMG and trained only two networks: one for PG and the other for SMG for the fine segmentation. Since the coarse segmentation network segments left and right glands separately, laterality is known and can be restored after the fine segmentation.

We have also tested the trained (on our local data as described above) network on the Public Domain Database for Computational Anatomy (PDDCA) version 1.4.1.[57] This dataset includes 48 H&N CT images among which 38 datasets have manual segmentations of both left and right PG and SMG. These 38 CT images had $512 \times 512 \times [76 \text{ to } 263]$ voxels with a pixel size of 0.88 to 1.27 mm and through-plane slice spacing of 2 to 3 mm.

3 Methods

The proposed hierarchical CNN segmentation approach consists of two steps: coarse and fine segmentations as shown in Fig. 1. The coarse segmentation generates organ-specific ROIs that are used to crop the input CT for each organ. The fine segmentation then processes the cropped ROI volume to produce high-quality segmentation of each organ. Finally, the segmented organ masks are merged to form a multi-organ segmentation of the patient. The following sections describe each step in detail.

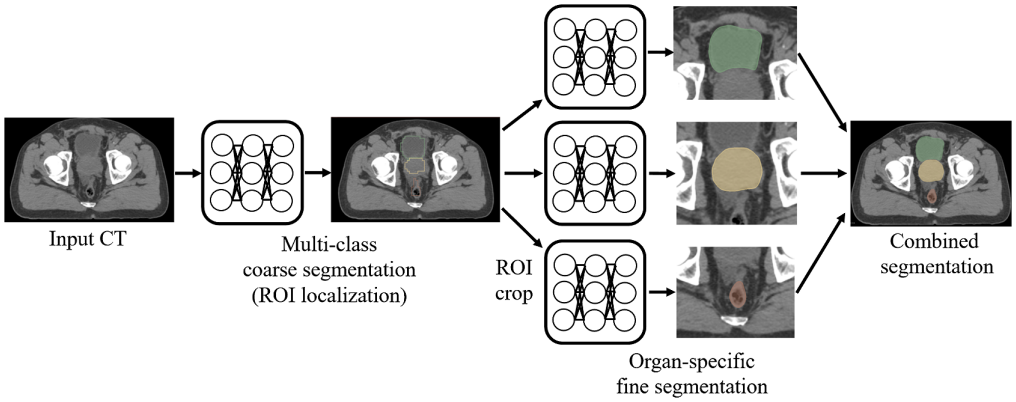

Fig. 1 Workflow of the proposed hierarchical segmentation method.

3.1 *Coarse Segmentation*

In the coarse segmentation step, multi-label segmentation is performed on the downsampled CT by a multi-class 3-D UNet. Organ-specific ROIs are extracted from the computed labels, each containing the corresponding organ of interest. The original CT volume contains a large background that carries contextual information irrelevant to the organs to be segmented. Including such background does not help each organ segmentation, but rather burdens in feature space and increases computational complexity. The computed ROIs allow us to crop the original CT to smaller volumes, compactly including the organs to be segmented, and therefore, to reduce computational complexity while improving the segmentation performance. Note that the coarse segmentation network is trained to segment all the organs of interest together instead of segmenting them separately. CT images are downsampled to a lower resolution to further save computational resources and time as only coarse-segmented labels are required in this step.

The modified 3-D UNet architecture used for coarse segmentation consists of contraction and expansion paths each with four layers. Each layer of the contraction path is composed of two $3 \times 3 \times 3$ convolutions ($3 \times 3 \times 3$ conv), followed by a leaky rectified linear unit (ReLU) activation function,[58] and $2 \times 2 \times 2$ max pooling. In each layer, batch normalization (BN) is added to speed up the learning process by reducing sensitivity to parameter initialization[59] and dropout to prevent overfitting.[60] The number of feature maps in the first layer is 16, which doubles at each successive layer. The expansion path has a similar architecture to the contraction path except that in each layer it has a $2 \times 2 \times 2$ up-convolution ($2 \times 2 \times 2$ up-conv) that halves the number of feature maps. The last $1 \times 1 \times 1$ convolution ($1 \times 1 \times 1$ conv) layer maps the output features to the desired number of labels. Skip connections are used to transfer features extracted from the early contraction path to the expansion path.

Once this first network is trained, it produces a coarse segmentation map of the organs of interest, which is then used to automatically extract ROIs for every organ. Based on the coarse segmentation, the centroid of each organ is calculated and then the ROI of each organ is cropped from the original image for which the centroid of each segmented organ is the center of the cropped ROI. The size of ROIs was determined to be large enough to cover each organ and enough surrounding background context.

3.2 *Fine Segmentation*

The fine segmentation network takes the organ-specific cropped CT images obtained from the coarse segmentation as input and segments of each organ of interest. Organ-specific fine segmentation CNNs are designed using 3-D UNet and GAN as shown in Fig. 2. The GAN network consists of generator and discriminator networks where the two networks compete with each other to produce an accurate segmentation. A modified 3-D UNet is used as a generator and trained using the cropped CT images and manual segmentations. The 3-D UNet has similar architecture as described for the coarse segmentation network as shown in Fig. 3. The generator produces the predicted label of the organ of interest. On the other hand, a standard FCN

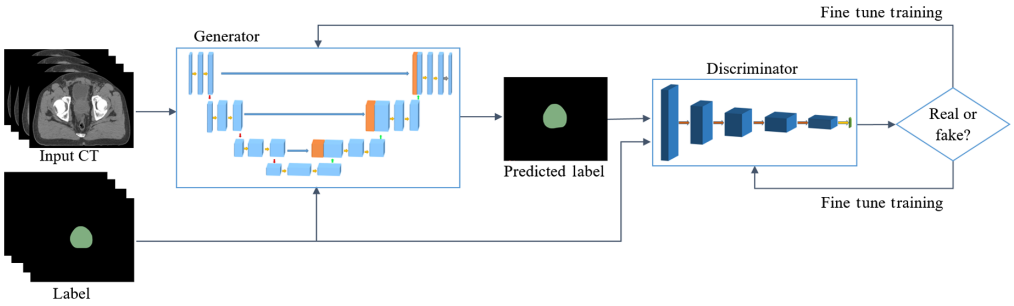

Fig. 2 The UNet-GAN architecture for fine segmentation.

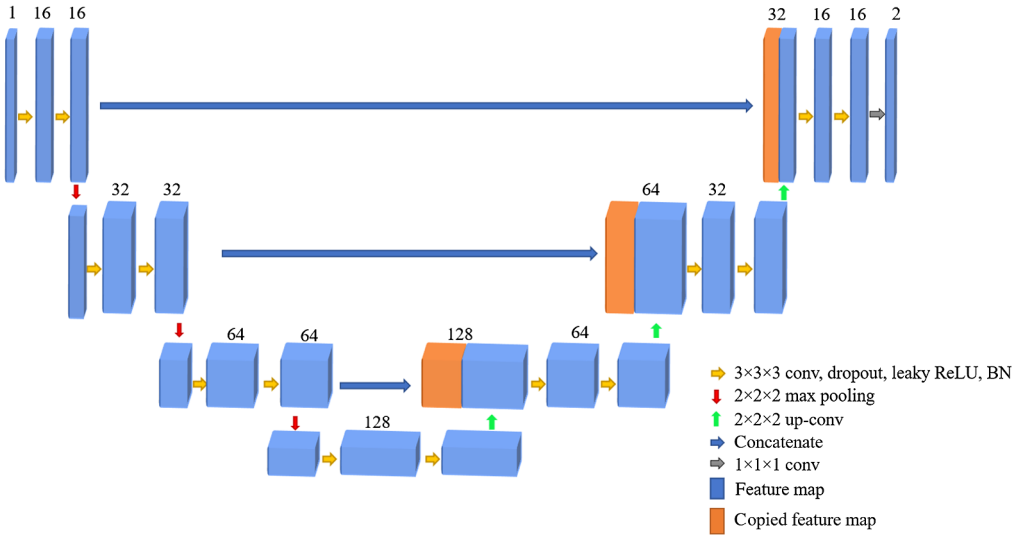

Fig. 3 Generator network.

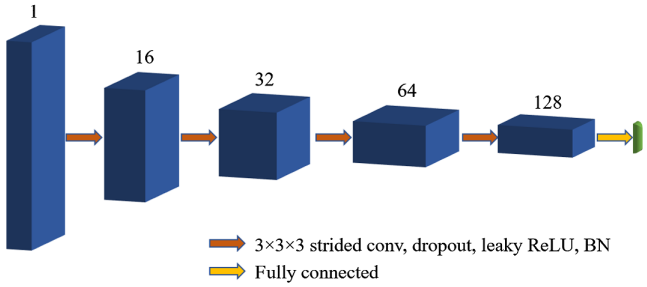

Fig. 4 Discriminator network.

composed of four strided convolutions each with dropout and BN followed by a fully connected layer is used as the discriminator (Fig. 4).The discriminator is trained using manually contoured labels to determine if the generator-predicted labels are real or fake.

For GAN, the generator network G and discriminator network D are trained simultaneously. The objective of G is to learn the distribution p_x from the dataset x and then sample a variable z from the uniform or Gaussian distribution $p_z(z)$. The purpose of D is to classify whether an image comes from the training dataset or from G. To define the cost function of the GAN, let l_{fake} and l_{real} denote labels for fake (generator-produced labels) and real data (labels from training data), respectively. Then the cost function for D and G are defined using a least squares loss function[61] as follows:

$$\text{loss}_D = \frac{1}{2}\text{E}_{x\sim p_x(x)}[(D(x) - l_{\text{real}})^2] + \frac{1}{2}\text{E}_{z\sim p_z(z)}[(D(G(z)) - l_{\text{fake}})^2], \tag{1}$$

$$\text{loss}_G = \frac{1}{2}\text{E}_{z\sim p_z(z)}[(D(G(z)) - l_{\text{real}})^2]. \tag{2}$$

To maximize the similarity between the generator-produced and the ground truth segmentations, we also compute a weighted dice loss defined as

$$\text{loss}_{\text{dice}} = \sum_i w_i \left[1 - \frac{\sum_x b_i(x)p(x,i)}{\sum_x b_i(x) + p(x,i)}\right], \tag{3}$$

where for each class i, w_i is the class weight, $b_i(x)$ is the binary label at each pixel x, and $p(x,i)$ is the predicted binary label. In case of multi-class coarse segmentation, i in Eq. (3) represents the class label for each organ to be segmented. For the fine segmentation, $i = w_i = 1$ as only one class needs to be predicted. The final objective function for the generator is defined as the sum of the least squares generator loss and the weighted dice loss as follows:

$$\text{loss} = \text{loss}_{\text{dice}} + \lambda\,\text{loss}_G, \tag{4}$$

where $\lambda \in [0,1]$.

3.3 *Network Implementation*

We implemented the proposed CNNs using Tensorflow[62] and used the following settings for both coarse and fine segmentation networks. The initial learning rate was set to 5×10^{-4}. We used an Adam optimizer, a stochastic gradient descent-based optimizer that adaptively estimates the lower order moments and automatically adjusts step size during optimization.[63] Dropout rate was 0.25 and the mini-batch size was 16. We trained and tested our network on a workstation with an Intel Xeon processor with 32 GB RAM and NVIDIA GeForce GTX TITAN X GPU with 12 GB memory.

4 Experiments and Results

4.1 *Network Training and Computation Time*

For pelvic images, training for the coarse and fine segmentation networks took 15 and 32 h, respectively. In the testing phase, coarse segmentation took 3 s and fine segmentation for each organ took 4 s. For the H&N images, training of the coarse and fine segmentation networks took 15 and 28 h, respectively. The computation time for coarse segmentation was 3 s and fine segmentation for each organ took 4 s.

4.2 *Evaluation Criteria*

We quantitatively assessed the automatic segmentation quality using six metrics in comparison to the ground truth segmentation.[64] To measure the degree of overlap between the automatic and ground truth segmentations, we computed the dice similarity coefficient (DSC). To compare the distances between the surfaces of the automatic and ground truth segmentations, we computed mean surface distance (MSD) and 95% Hausdorff distance (HD95), i.e., the 95th percentile of the distances between the surface points of the automatic and ground truth segmentations. Instead of maximum HD, HD95 was computed to discard the impact of a small subset of inaccurate segmentation while evaluating the overall segmentation quality. Finally, to measure the accurately segmented portion among the automatic segmentation, we computed positive predictive value (PPV) and sensitivity (SEN) defined as $\text{PPV} = |V_{\text{gt}} \cap V_{\text{seg}}|/|V_{\text{seg}}|$ and $\text{SEN} = |V_{\text{gt}} \cap V_{\text{seg}}|/|V_{\text{gt}}|$, where V_{gt} and V_{seg} are the ground truth and automatic segmentations, respectively.

4.3 *Pelvic Segmentation*

Example segmentations of two test cases are shown in Fig. 5 and the quantitative segmentation accuracy of all 15 test cases in terms of DSC is shown in Fig. 6. Our training datasets included full, half, and empty bladder so that the trained network was capable of segmenting them correctly as shown in Fig. 5. It should be noted that we used contours drawn by the attending radiation oncologist for the patients' RT planning to train and test the proposed network. There was slight variation among physicians regarding where to stop contouring the rectum superiorly (rectum-sigmoid boundary), which caused a slightly reduced autosegmentation performance for the rectum. We did not attempt to modify the rectum contour from what was defined by the attending radiation oncologist to reflect the real clinical scenario. Notice that the network produced similar segmentation as the physician's manual segmentation.

To demonstrate the benefit of incorporating GAN and performing organ-specific fine segmentation in the proposed hierarchical approach, we compared the proposed hierarchical UNet-GAN segmentation with two other CNN approaches; (1) multi-class UNet and (2) multi-class

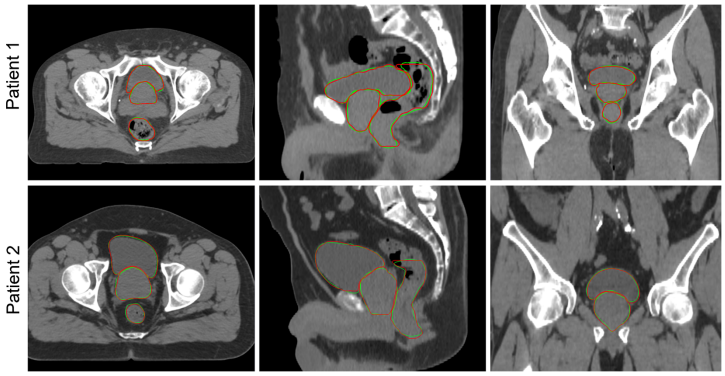

Fig. 5 Example segmentations of the prostate, bladder, and rectum shown in axial, sagittal, and coronal planes. Each row represents a different patient. Note that the bladder shapes are significantly different between these two cases. Green, automatic segmentation and red, manual segmentation.

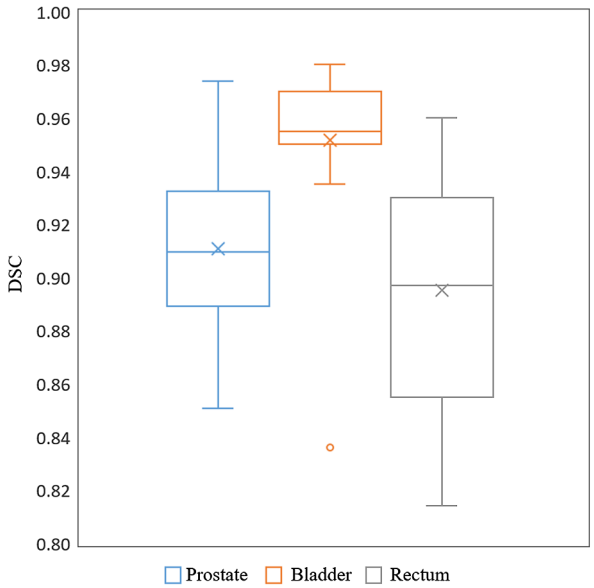

Fig. 6 Box and whisker plots for DSC of the three segmented pelvic organs. The boxes show 25th and 75th percentiles and the centerline inside each box indicates the median value. "×" marks indicate mean values over 15 cases.

UNet-GAN where both networks were trained to simultaneously segment the prostate, bladder, and rectum together. Both of the multi-class CNNs were trained using 275 CT with augmentation where the ROI of dimensions $112 \times 112 \times 64$ was cropped from original CT with a voxel size of $2.3 \times 2.3 \times 3$ mm^3. The cropped ROI includes the prostate, bladder, rectum, and enough background contexts.

Quantitative and qualitative comparisons are reported in Table 1 and shown in Fig. 7, respectively. In general, multi-class UNet achieved a reasonable segmentation performance as reported in the prior studies.[19,56,65] We observed that erroneously segmented regions in the multi-class UNet were often improved when GAN was incorporated as shown in Fig. 7. We believe that GAN constrained the network to produce segmentations of which shapes vary within a reasonable range of variation and are close to human experts' manual segmentations. The proposed hierarchical UNet-GAN achieved superior performance to the multi-class UNet and UNet-GAN approaches with an overall DSC (mean $\pm$ SD) of 0.91 ± 0.05, 0.95 ± 0.06 and 0.90 ± 0.09 for the prostate, bladder and rectum, respectively. Although the number of test cases was small, we performed Wilcoxon signed rank test for the DSC scores to assess the statistical significance of the performance difference between the proposed hierarchical UNet-GAN and multi-class UNet/UNet-GAN (Table 1). It was observed that there were significant improvements in prostate and bladder segmentations with $p = 0.0009/0.031$, $p = 0.0002/0.028$ (vs UNet/UNet-GAN), respectively, whereas the difference was not statistically significant for the rectum segmentation ($p = 0.155/0.429$). The proposed method also achieved lower MSD and HD95 than the other two methods for all three structures. Note that MSD is less than 1.8 mm for all structures, which can be considered excellent segmentation performance given the original CT image resolution

Table 1 Quantitative comparison of pelvic CT segmentation performance of different methods (mean $\pm$ SD).

Metrics	Method	Prostate	Bladder	Rectum
DSC	Hierarchical UNet-GAN	0.91 ± 0.05	0.95 ± 0.06	0.90 ± 0.09
	UNet-GAN	0.86 ± 0.07 ($p = 0.031$)	0.92 ± 0.05 ($p = 0.028$)	0.87 ± 0.13 ($p = 0.429$)
	UNet	0.84 ± 0.05 ($p = 0.0009$)	0.88 ± 0.06 ($p = 0.0002$)	0.83 ± 0.16 ($p = 0.155$)
MSD (mm)	Hierarchical UNet-GAN	1.56 ± 0.37	0.95 ± 0.15	1.78 ± 1.13
	UNet-GAN	2.28 ± 0.78	2.11 ± 0.45	3.45 ± 0.95
	UNet	2.89 ± 1.15	2.34 ± 0.91	3.91 ± 0.47
HD95 (mm)	Hierarchical UNet-GAN	5.21 ± 1.17	4.37 ± 0.56	6.11 ± 1.47
	UNet-GAN	6.55 ± 2.87	5.83 ± 1.53	7.11 ± 3.42
	UNet	7.20 ± 1.90	7.20 ± 1.90	9.20 ± 1.90
PPV	Hierarchical UNet-GAN	0.90 ± 0.06	0.94 ± 0.02	0.87 ± 0.09
	UNet-GAN	0.84 ± 0.50	0.86 ± 0.63	0.83 ± 0.25
	UNet	0.86 ± 0.04	0.88 ± 0.45	0.85 ± 0.31
SEN	Hierarchical UNet-GAN	0.84 ± 0.07	0.97 ± 0.13	0.88 ± 0.16
	UNet-GAN	0.90 ± 0.06	0.92 ± 0.08	0.89 ± 0.01
	UNet	0.89 ± 0.82	0.93 ± 0.63	0.90 ± 0.96

Note: p-values were computed between UNet-GAN/UNet and hierarchical UNet-GAN through Wilcoxon signed rank test. $p < 0.05$ is considered statistically significant.

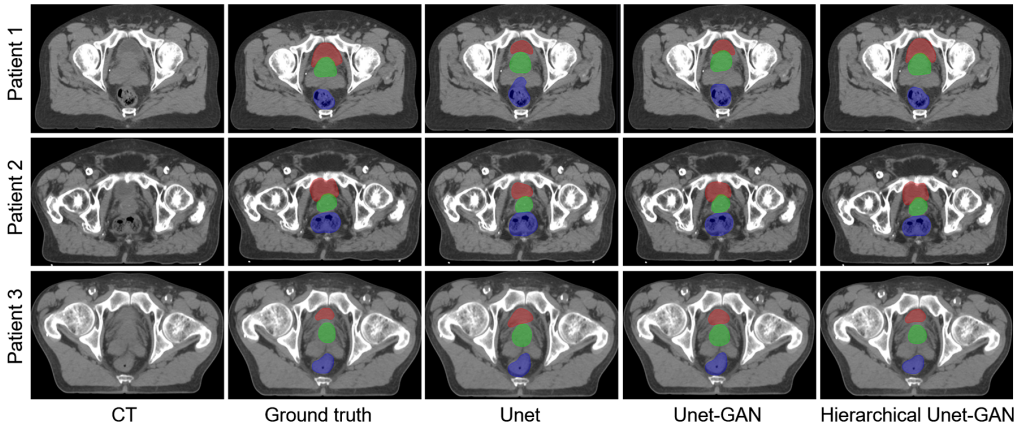

Fig. 7 Examples of segmentations for (green) prostate, (red) bladder, and (blue) rectum using three different approaches. Each row shows an axial view of a different patient.

Table 2 Pelvic CT segmentation performance comparison with other state-of-the-art methods (DSC, mean ± SD).

Method		Prostate	Bladder	Rectum
Martinez[15]	Model-based	0.87 ± 0.07	0.89 ± 0.08	0.82 ± 0.06
Gao[19]	Regression forest	0.87 ± 0.04	0.92 ± 0.05	0.88 ± 0.05
Shao[20]		0.88 ± 0.02	—	0.84 ± 0.05
Kazemifar[31]	CNN	0.88 ± 0.10	0.95 ± 0.04	0.92 ± 0.10
Wang[33]		0.89 ± 0.03	0.94 ± 0.03	0.89 ± 0.04
Balagopal[36]		0.90 ± 0.02	0.95 ± 0.02	0.84 ± 0.04
Proposed		0.91 ± 0.05	0.95 ± 0.06	0.90 ± 0.09

($1.30 \times 1.30 \times 3$ mm^3 on average). The main reason for the superior performance of the proposed method to the other methods is twofold. First, the fine segmentation focuses only on a specific organ confined within a compact ROI that contains enough contextual information around the target organ while excluding unrelated (or very weakly related) background information outside the ROI. Second, the incorporation of GAN enables adversarial training, which further strengthens the segmentation network to produce accurate segmentation by penalizing segmentation with irregular shapes that are inconsistent with the experts' manual segmentation.

Table 2 shows quantitative comparison between the proposed method and existing state-of-the-art methods. Although these methods used different datasets, this comparison allows us to assess the performance of the proposed methods in (indirect) comparison to other methods. In terms of DSC, our method outperformed both model-based and regression forest-based machine learning methods.[15,19,20] In comparison to three CNN-based approaches,[31,33,36] our method showed the best performance overall, outperforming all three for the prostate segmentation while being comparable or slightly better for the bladder and rectum segmentations.

4.4 *H&N Segmentation*

We have segmented PG and SMG for 20 H&N CT test datasets. Two example cases are shown in Figs. 8 and 9. The distribution of DSC over 20 cases is shown in Fig. 10(a). The performance of the proposed method is compared with multi-class UNet and multi-class UNet-GAN segmentations similar to the pelvic cases, and the quantitative comparison is reported in Table 3. It is

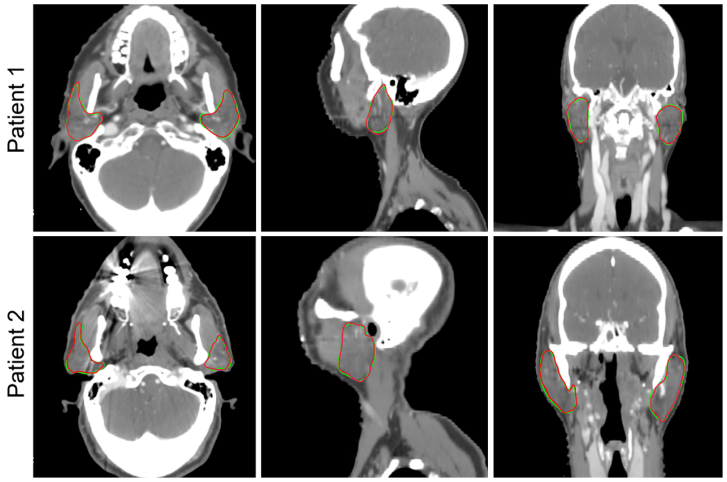

Fig. 8 Examples of PG segmentations. Each row shows axial, sagittal, and coronal views of a different patient. Green, automatic segmentation and red, manual segmentation.

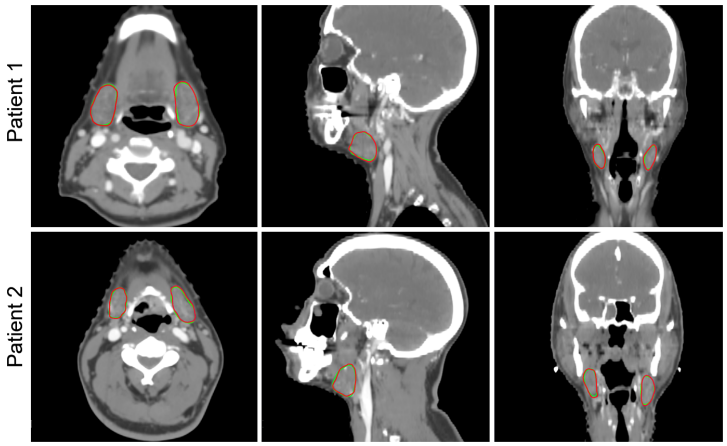

Fig. 9 Examples of SMG segmentations. Each row shows axial, sagittal, and coronal views of a different patient. Green, automatic segmentation and red, manual segmentation.

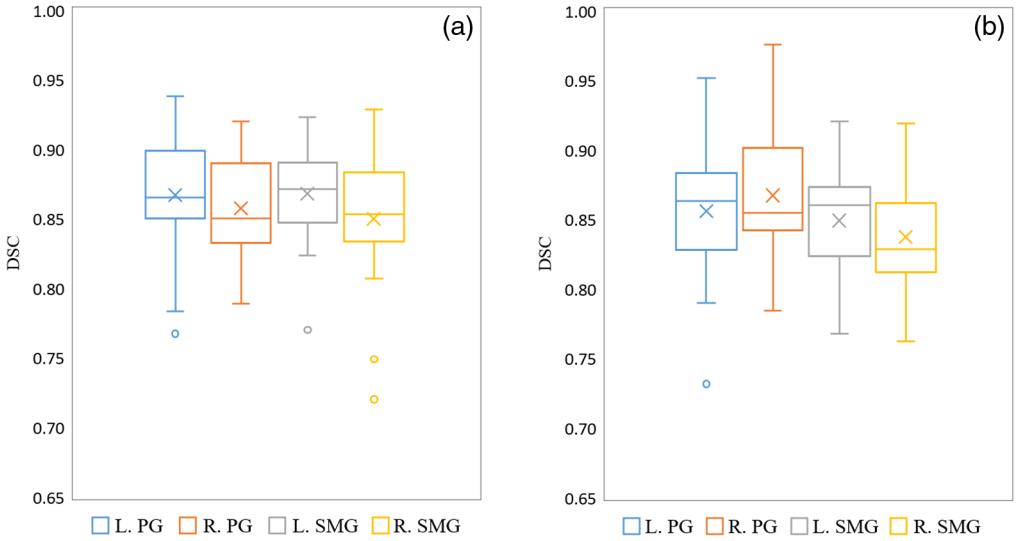

Fig. 10 Box and whisker plots for DSC of PG and SMG segmentations: (a) internal datasets with 20 cases and (b) PDDCA dataset with 38 cases. The boxes show 25th and 75th percentiles and the centerline inside each box indicates the median value. "×" marks indicate mean values.

Table 3 Quantitative comparison of H&N CT segmentation performance of different methods (mean ± SD).

Metrics	Method	Left PG	Right PG	Left SMG	Right SMG
DSC	Hierarchical UNet-GAN	0.87 ± 0.04	0.86 ± 0.03	0.87 ± 0.04	0.85 ± 0.05
	UNet-GAN	0.83 ± 0.06 ($p = 0.033$)	0.83 ± 0.05 ($p = 0.170$)	0.81 ± 0.15 ($p = 0.145$)	0.80 ± 0.13 ($p = 0.130$)
	UNet	0.83 ± 0.08 ($p = 0.090$)	0.82 ± 0.06 ($p = 0.032$)	0.79 ± 0.10 ($p = 0.012$)	0.78 ± 0.14 ($p = 0.036$)
MSD (mm)	Hierarchical UNet-GAN	1.55 ± 0.97	1.21 ± 0.35	1.81 ± 0.60	2.11 ± 0.31
	UNet-GAN	3.15 ± 0.92	2.13 ± 0.44	4.15 ± 0.85	3.97 ± 1.12
	UNet	3.80 ± 1.15	4.08 ± 0.32	4.18 ± 1.13	3.55 ± 2.50
HD95 (mm)	Hierarchical UNet-GAN	3.49 ± 2.25	4.25 ± 1.11	4.22 ± 1.16	4.01 ± 0.20
	UNet-GAN	4.53 ± 1.04	3.88 ± 2.33	5.53 ± 1.06	5.11 ± 1.58
	UNet	4.89 ± 2.15	4.20 ± 2.21	5.56 ± 2.15	6.05 ± 2.03
PPV	Hierarchical UNet-GAN	0.87 ± 0.10	0.86 ± 0.15	0.81 ± 0.28	0.78 ± 0.15
	UNet-GAN	0.80 ± 0.42	0.83 ± 0.10	0.81 ± 0.45	0.80 ± 0.35
	UNet	0.78 ± 0.55	0.79 ± 0.83	0.78 ± 0.37	0.76 ± 0.26
SEN	Hierarchical UNet-GAN	0.85 ± 0.63	0.88 ± 0.52	0.80 ± 0.14	0.84 ± 0.42
	UNet-GAN	0.82 ± 0.31	0.84 ± 0.78	0.81 ± 0.19	0.80 ± 0.12
	UNet	0.78 ± 0.35	0.82 ± 0.23	0.79 ± 0.65	0.80 ± 0.25

Note: p-values were computed between UNet-GAN/UNet and hierarchical UNet-GAN through Wilcoxon signed rank test. $p < 0.05$ is considered statistically significant.

observed that incorporating GAN to UNet improves the segmentation performance, and the proposed hierarchical UNet-GAN achieves better segmentation performance compared to the other two methods. These performance improvements are consistent with the trend in the pelvic cases and demonstrate the benefit of incorporating GAN and utilizing a hierarchical approach. The average DSCs for PG and SMG are 0.87 and 0.86, respectively. We also performed Wilcoxon signed rank test for the DSC scores to assess the statistical significance of the performance difference between the proposed hierarchical UNet-GAN and multi-class UNet/UNet-GAN (Table 3). It was observed that there were significant improvements in both glands with $p = 0.005/0.012$ and $p = 0.001/0.038$ (versus UNet/UNet-GAN) for the PG and SMG (left–right combined), respectively.

To assess the generalizability of the proposed method, we applied our trained network to 38 H&N CTs in the PDDCA dataset that has both PG and SMG segmentations. Although trained on a different dataset, the proposed hierarchical UNet-GAN was able to achieve similar segmentation performance for both PG and SMG. The distribution of DSC over 38 cases is shown in Fig. 10(b) and the quantitative performance is reported in Table 4.

We also compared the proposed hierarchical UNet-GAN with seven existing state-of-the-art methods as shown in Table 5. These methods are based on multi-atlas,[7] deformable model using landmarks,[10] hierarchical vertex regression to learn shape and appearance,[66] patch-based CNN,[44] convolutional dense-net with shape constraint and GAN,[50] two-stage 3-D Unets (3-D UNet bounding box network with sliding + 3-D UNet segmentation network),[55] and 3-D UNet with residual blocks.[49] It should be noted that some of these methods[50,66,67] used the PDDCA dataset to evaluate segmentation performance. Evaluation using the PDDCA dataset shows a direct and fair comparison among the competing methods. As reported in Table 5, the proposed method

Table 4 Quantitative results on PDDCA dataset (mean ± SD).

	Left PG	Right PG	Left SMG	Right SMG
DSC	0.86 ± 0.05	0.87 ± 0.06	0.85 ± 0.07	0.84 ± 0.08
MSD (mm)	1.27 ± 0.78	1.12 ± 0.57	1.41 ± 0.76	1.35 ± 0.85
HD95 (mm)	2.84 ± 1.52	2.23 ± 1.70	2.97 ± 1.42	3.15 ± 1.65
PPV	0.88 ± 0.20	0.86 ± 0.10	0.83 ± 0.27	0.85 ± 0.35
SEN	0.84 ± 0.05	0.87 ± 0.06	0.84 ± 0.22	0.82 ± 0.10

Table 5 H&N CT segmentation performance comparison with other state-of-the-art methods (DSC).

Method		Left PG	Right PG	Left SMG	Right SMG
Han[7]	Multi-atlas	0.82	0.82	0.69	0.69
Qazi[10]	Model-based	0.83	0.83	0.84	0.81
Wang[66]	Regression-based	0.82 ± 0.05	0.83 ± 0.06	—	—
Ibragimov[44]	CNN	0.77 ± 0.06	0.78 ± 0.05	0.70 ± 0.13	0.73 ± 0.09
Tong[50]		0.85 ± 0.02	0.86 ± 0.02	0.81 ± 0.05	0.82 ± 0.05
Wang[55]		0.86 ± 0.03	0.85 ± 0.07	0.76 ± 0.15	0.73 ± 0.10
Zhu[49]		0.88 ± 0.02	0.87 ± 0.04	0.81 ± 0.04	0.81 ± 0.04
Proposed		0.87 ± 0.04	0.86 ± 0.03	0.87 ± 0.04	0.85 ± 0.05

outperformed these state-of-the-art methods in terms of DSC. The only exception is the method using 3-D UNet with a residual block,[49] which achieved slightly better DSC (average of left and right is 0.875) for PG compared to the proposed method (average of left and right is 0.865). However, the proposed method achieved the best performance for SMG segmentation among all the methods compared.

5 Discussion

In this paper, we proposed a hierarchical coarse-to-fine segmentation network to automatically segment multiple organs from CT images for RT planning. In the coarse segmentation stage, a less time-consuming multi-class coarse segmentation of multiple organs of interest is performed using whole CT images. This multi-class segmentation is used to localize the ROI of each organ, which is subsequently used in the fine segmentation stage. This organ localization network helps to remove less important background and improve the efficiency of the fine segmentation by constraining the segmentation within the specific region of the organ. The fine segmentation network is designed with a modified 3-D UNet combined with GAN. GAN performs adversarial training by distinguishing between ground truth and predicted segmentations. The combined dice and GAN loss guides the training process to better handle inconsistent and irregular shapes, thus producing more accurate segmentations. Segmentation comparisons with other variations of UNet reported in Tables 1 and 3 show the contribution of GAN-based adversarial training integrated with UNet.

In the multi-class segmentation, class imbalance is a common problem where small structures are prone to being underrepresented compared to the bigger structures.[68] In the proposed method, we perform single-class segmentation of each organ instead of multi-class segmentation, which

potentially improves the quality of individual organs' segmentation regardless of their size. Tables 1 and 3 show that the single-class segmentation always outperforms one-step multi-class segmentation. Another advantage of the single-class segmentation network is that the network training can utilize all available data even if there are missing labels while the multi-class segmentation network typically requires a complete set of labels for all structures unless additional constraints to handle missing labels are incorporated.[23] However, one limitation of the single-class segmentation is that it requires more networks to be trained, one for each organ. In the proposed method at the fine segmentation step, each organ has a much smaller ROI than the combined ROI for multi-class segmentation, which enables much more efficient network training and faster execution.

An extensive validation of the proposed method was performed using two disease sites: pelvic and H&N regions. Such an extensive validation in multiple disease sites proves that the proposed method is versatile and generalizable to segment organs of diverse shape and size.

Automatic segmentation of pelvic organs is a crucial step for the effective treatment of prostate cancer using RT. Over the past years, several prostate segmentation methods have been proposed using atlas-based, model-based and most recently, deep-learning-based approaches.[19,20,31,33,36,69] We have compared our proposed method with these state-of-the-art methods and showed that the proposed method outperforms them with reliable, accurate, and reproducible organ segmentation performance.

The proposed hierarchical UNet-GAN method was also employed to segment salivary glands in H&N CT images. PG and SMG are responsible for producing saliva. Excessive irradiation of these organs may cause side effects such as xerostomia that may lead to late complications including poor dental hygiene, oral infections, sleep disturbances, and difficulty in swallowing.[70] Therefore, these glands are routinely contoured for an effective RT planning to minimize their RT-induced toxicities. The segmentation results of PG and SMG using our proposed method show promising outcome when compared with the performance of existing methods. This segmentation performance is comparable to human experts' considering the significant inter-observer variability between experts' manual segmentations of H&N structures.[4,71]

Performing a fair comparison with different segmentation methods is difficult as they are often designed and optimized for a specific problem as well as tuned and tested on a specific dataset. Reimplementing or utilizing available open-source tools may allow us to directly compare their performances on the same data, but this may lead to unfair comparison as given parameters may not be optimal for other data. A slight change of parameters may also cause significant performance change. Instead, we used the publicly available PDDCA dataset that was used in MICCAI 2015 H&N autosegmentation grand challenge.[57] Such a validation using independent public datasets provides a frame of reference to compare results from different competing segmentation methods. Quantitative results using the PDDCA dataset as reported in Table 4 shows consistent performance as with the internal dataset, demonstrating the generalizability of our network to other datasets. Furthermore, indirect comparison to other state-of-the-art segmentation methods shows the excellent performance of the proposed method over existing methods.

We used 275 prostate and 200 H&N data sets to train our networks. These data sets were augmented by rotation, translation, and lateral flipping to yield 1100 prostate and 800 H&N training data sets. These numbers are comparable to or exceed the number of training sets used in other CNN-based segmentation approaches in the literature[31,33,36,44,49,50,55] and allowed us to train our networks to produce the reported promising performance. We observed that the segmentation performance degraded when we reduced the number of training data sets. Given the limited set of available data, we had to maximize the training data, leaving a small number of testing data sets. Although the numbers of test cases (15 prostate and 58 [20 internal and 38 external] H&N cases) are also comparable to other studies, extended testing on a much larger cohort of test cases may be needed to perform a rigorous statistical analysis.

In this paper, we have presented segmentation results of a limited number of organs in each disease site. In addition to the prostate, bladder, and rectum, our method can be extended to include more structures such as the bowel, femoral head, seminal vesicle, and sigmoid that are commonly contoured for prostate RT planning. For H&N, including other structures such as the brain, brainstem, eyes, optic nerve, optical chiasm, mandible, pituitary gland, thyroid, and larynx

are our next steps toward more efficient H&N RT planning. We are confident that the current network can be readily used to segment these organs without much modification.

Finally, we have used CT images to train and test the proposed method in this study. The same network can be used with other image modalities such as MRI, and may lead to similar or higher accuracy depending on the organ visibility in those image modalities.

6 Conclusion

This paper presented an end-to-end, CNN-based automatic multi-organ segmentation in CT images using a hierarchical UNet-GAN with automatic ROI localization of the organs to be segmented. The automatic ROI extraction improved computational efficiency and the segmentation accuracy by allowing the fine segmentation network to focus only on the region of each organ. The fine segmentation of each organ is performed using UNet-GAN where the generator and discriminator compete with each other to improve segmentation accuracy. To avoid a class imbalance problem of multi-class segmentation, we performed single-class training, which improved segmentation accuracy over the multi-class segmentation. We performed extensive experimental validation using clinical data and showed that the proposed method outperformed other state-of-the-art methods. The proposed method can potentially improve the efficiency of RT planning of cancer treatment by reducing the burden of tedious manual contouring.

Disclosures

No conflicts of interest to report.

Acknowledgments

This work was supported by the National Cancer Institute (NCI), National Institutes of Health (NIH) under Grant No. R01CA151395.

References

1. P. Steenbergen et al., "Prostate tumor delineation using multiparametric magnetic resonance imaging: inter-observer variability and pathology validation," *Radiother. Oncol.* **115**(2), 186–190 (2015).
2. C. Fiorino et al., "Intra-and inter-observer variability in contouring prostate and seminal vesicles: implications for conformal treatment planning," *Radiother. Oncol.* **47**(3), 285–292 (1998).
3. W. R. Lee et al., "Interobserver variability leads to significant differences in quantifiers of prostate implant adequacy," *Int. J. Radiat. Oncol. Biol. Phys.* **54**(2), 457–461 (2002).
4. B. E. Nelms et al., "Variations in the contouring of organs at risk: test case from a patient with oropharyngeal cancer," *Int. J. Radiat. Oncol. Biol. Phys.* **82**(1), 368–378 (2012).
5. C. Sjöberg et al., "Clinical evaluation of multi-atlas based segmentation of lymph node regions in head and neck and prostate cancer patients," *Radiat. Oncol.* **8**(1), 229 (2013).
6. O. Acosta et al., "Evaluation of multi-atlas-based segmentation of CT scans in prostate cancer radiotherapy," in *IEEE Int. Symp. Biomed. Imaging: From Nano to Macro*, pp. 1966–1969 (2011).
7. X. Han et al., "Atlas-based auto-segmentation of head and neck CT images," *Lect. Notes Comput. Sci.* **5242**, 434–441 (2008).
8. Y. Wang et al., "A quality control model that uses PTV-rectal distances to predict the lowest achievable rectum dose, improves IMRT planning for patients with prostate cancer," *Radiother. Oncol.* **107**(3), 352–357 (2013).
9. A. J. Asman and B. A. Landman, "Non-local statistical label fusion for multi-atlas segmentation," *Med. Image Anal.* **17**(2), 194–208 (2013).
10. A. A. Qazi et al., "Auto-segmentation of normal and target structures in head and neck CT images: a feature-driven model-based approach," *Med. Phys.* **38**(11), 6160–6170 (2011).

11. J. Huang et al., "An improved level set method for vertebra CT image segmentation," *Biomed. Eng. Online* **12**(Suppl 1), S1–16 (2013).
12. X. Qian et al., "An active contour model for medical image segmentation with application to brain CT image," *Med. Phys.* **40**(2), 21911 (2013).
13. M. R. Kaus, T. McNutt, and V. Pekar, "Automated 3D and 4D organ delineation for radiation therapy planning in the pelvic area," *Proc. SPIE* **5370**, 346–356 (2004).
14. M. J. Costa et al., "Automatic segmentation of bladder and prostate using coupled 3D deformable models," *Lect. Notes Comput. Sci.* **4791**, 252–260 (2007).
15. F. Martinez et al., "Segmentation of pelvic structures for planning CT using a geometrical shape model tuned by a multi-scale edge detector," *Phys. Med. Biol.* **59**(6), 1471 (2014).
16. A. Chen et al., "Combining registration and active shape models for the automatic segmentation of the lymph node regions in head and neck CT images," *Med. Phys.* **37**(12), 6338–6346 (2010).
17. S. Gorthi et al., "Segmentation of head and neck lymph node regions for radiotherapy planning using active contour-based atlas registration," *IEEE J. Sel. Top. Signal Process.* **3**(1), 135–147 (2009).
18. V. Pekar et al., "Head and neck auto-segmentation challenge: segmentation of the parotid glands," in *Med. Image Comput. and Comput. Assist. Intervention (MICCAI)*, pp. 273–280 (2010).
19. Y. Gao et al., "Accurate segmentation of CT male pelvic organs via regression-based deformable models and multi-task random forests," *IEEE Trans. Med. Imaging* **35**(6), 1532–1543 (2016).
20. Y. Shao et al., "Locally-constrained boundary regression for segmentation of prostate and rectum in the planning CT images," *Med. Image Anal.* **26**(1), 345–356 (2015).
21. D. Shen, G. Wu, and H.-I. Suk, "Deep learning in medical image analysis," *Annu. Rev. Biomed. Eng.* **19**, 221–248 (2017).
22. W. Rawat and Z. Wang, "Deep convolutional neural networks for image classification: A comprehensive review," *Neural Comput.* **29**(9), 2352–2449 (2017).
23. N. Tajbakhsh et al., "Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation," *Med. Image Anal.* **63**, 101693 (2020).
24. G. Litjens et al., "A survey on deep learning in medical image analysis," *Med. Image Anal.* **42**, 60–88 (2017).
25. J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 3431–3440 (2015).
26. O. Ronneberger, P. Fischer, and T. Brox, "U-net: convolutional networks for biomedical image segmentation," *Lect. Notes comput. Sci.* **9351**, 234–241 (2015).
27. Ö. Çiçek et al., "3D U-Net: learning dense volumetric segmentation from sparse annotation," *Lect. Notes Comput. Sci.* **9901**, 424–432 (2016).
28. F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: fully convolutional neural networks for volumetric medical image segmentation," in *Fourth Int. Conf. 3D Vision (3DV)*, pp. 565–571 (2016).
29. X. Li et al., "H-DenseUNet: hybrid densely connected UNet for liver and tumor segmentation from CT volumes," *IEEE Trans. Med. Imaging* **37**(12), 2663–2674 (2018).
30. Q. Zhu et al., "Deeply-supervised CNN for prostate segmentation," in *Int. Joint Conf. Neural Networks (IJCNN)*, pp. 178–184 (2017).
31. S. Kazemifar et al., "Segmentation of the prostate and organs at risk in male pelvic CT images using deep learning," *Biomed. Phys. Eng. Express.* **4**(5), 55003 (2018).
32. B. Wang et al., "Automated prostate segmentation of volumetric CT images using 3D deeply supervised dilated FCN," *Proc. SPIE* **10949**, 109492S (2019).
33. S. Wang et al., "CT male pelvic organ segmentation using fully convolutional networks with boundary sensitive representation," *Med. Image Anal.* **54**, 168–178 (2019).
34. H. R. Roth et al., "Deeporgan: multi-level deep convolutional networks for automated pancreas segmentation," *Lect. Notes Comput. Sci.* **9349**, 556–564 (2015).
35. K. Men, J. Dai, and Y. Li, "Automatic segmentation of the clinical target volume and organs at risk in the planning CT for rectal cancer using deep dilated convolutional neural networks," *Med. Phys.* **44**(12), 6377–6389 (2017).

36. A. Balagopal et al., "Fully automated organ segmentation in male pelvic CT images," *Phys. Med. Biol.* **63**(24), 245015 (2018).
37. X. Dong et al., "Automatic multiorgan segmentation in thorax CT images using U-net-GAN," *Med. Phys.* **46**(5), 2157–2168 (2019).
38. I. Goodfellow, "NIPS 2016 tutorial: generative adversarial networks," arXiv:1701.00160 (2016).
39. Z. Tian et al., "PSNet: prostate segmentation on MRI based on a convolutional neural network," *J. Med. Imaging* **5**(2), 21208 (2018).
40. M. N. N. To et al., "Deep dense multi-path neural network for prostate segmentation in magnetic resonance imaging," *Int. J. Comput. Assist. Radiol. Surg.* **13**(11), 1687–1696 (2018).
41. S. Elguindi et al., "Deep learning-based auto-segmentation of targets and organs-at-risk for magnetic resonance imaging only planning of prostate radiotherapy," *Phys. Imaging Radiat. Oncol.* **12**, 80–86 (2019).
42. M. H. F. Savenije et al., "Clinical implementation of MRI-based organs-at-risk auto-segmentation with convolutional networks for prostate radiotherapy," *Radiat. Oncol.* **15**, 1–12 (2020).
43. K. Fritscher et al., "Deep neural networks for fast segmentation of 3D medical images," *Lect. Notes Comput. Sci.* **9901**, 158–165 (2016).
44. B. Ibragimov and L. Xing, "Segmentation of organs-at-risks in head and neck CT images using convolutional neural networks," *Med. Phys.* **44**(2), 547–557 (2017).
45. D. Močnik et al., "Segmentation of parotid glands from registered CT and MR images," *Phys. Med.* **52**, 33–41 (2018).
46. X. Ren et al., "Interleaved 3D-CNN s for joint segmentation of small-volume structures in head and neck CT images," *Med. Phys.* **45**(5), 2063–2075 (2018).
47. J. W. Chan et al., "A convolutional neural network algorithm for automatic segmentation of head and neck organs at risk using deep lifelong learning," *Med. Phys.* **46**(5), 2204–2213 (2019).
48. A. Hänsch et al., "Comparison of different deep learning approaches for parotid gland segmentation from CT images," *Proc. SPIE* **10575**, 1057519 (2018).
49. W. Zhu et al., "AnatomyNet: deep learning for fast and fully automated whole-volume segmentation of head and neck anatomy," *Med. Phys.* **46**(2), 576–589 (2019).
50. N. Tong et al., "Shape constrained fully convolutional DenseNet with adversarial training for multiorgan segmentation on head and neck CT and low-field MR images," *Med. Phys.* **46**(6), 2669–2682 (2019).
51. Z. Zhu et al., "A 3D coarse-to-fine framework for volumetric medical image segmentation," in *Int. Conf. 3D Vision (3DV)*, pp. 682–690 (2018).
52. Z. Zhu et al., "Multi-scale coarse-to-fine segmentation for screening pancreatic ductal adenocarcinoma," *Lect. Notes Comput. Sci.* **11769**, 3–12 (2019).
53. R. Trullo et al., "Fully automated esophagus segmentation with a hierarchical deep learning approach," in *IEEE Int. Conf. Signal and Image Process. Appl. (ICSIPA)*, pp. 503–506 (2017).
54. H. R. Roth et al., "Hierarchical 3D fully convolutional networks for multi-organ segmentation," arXiv:1704.06382 (2017).
55. Y. Wang et al., "Organ at risk segmentation in head and neck ct images using a two-stage segmentation framework based on 3D U-Net," *IEEE Access* **7**, 144591–144602 (2019).
56. S. Sultana et al., "CNN-based hierarchical coarse-to-fine segmentation of pelvic CT images for prostate cancer radiotherapy," *Proc. SPIE* **11315**, 113151I (2020).
57. P. F. Raudaschl et al., "Evaluation of segmentation methods on head and neck CT: auto-segmentation challenge 2015," *Med. Phys.* **44**(5), 2020–2036 (2017).
58. B. Xu et al., "Empirical evaluation of rectified activations in convolutional network," arXiv:1505.00853 (2015).
59. S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," arXiv:1502.03167 (2015).
60. N. Srivastava et al., "Dropout: a simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.* **15**(1), 1929–1958 (2014).

61. X. Mao et al., "Least squares generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vision*, pp. 2794–2802 (2017).
62. M. Abadi et al., "Tensorflow: large-scale machine learning on heterogeneous distributed systems," arXiv:1603.04467 (2016).
63. D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," arXiv:1412.6980 (2014).
64. A. A. Taha and A. Hanbury, "Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool," *BMC Med. Imaging* **15**(1), 29 (2015).
65. S. Sultana, D. Y. Song, and J. Lee, "Deformable registration of PET/CT and ultrasound for disease-targeted focal prostate brachytherapy," *J. Med. Imaging* **6**(3), 35003 (2019).
66. Z. Wang et al., "Hierarchical vertex regression-based segmentation of head and neck CT images for radiotherapy planning," *IEEE Trans. Image Process.* **27**(2), 923–937 (2017).
67. X. Han et al., "Automatic segmentation of head and neck CT images by GPU-accelerated multi-atlas fusion," in *Proc. Head Neck Auto-Segment. Challenge Workshop*, p. 219 (2009).
68. C. Wachinger, M. Reuter, and T. Klein, "DeepNAT: deep convolutional neural network for segmenting neuroanatomy," *Neuroimage* **170**, 434–445 (2018).
69. O. Acosta et al., "Multi-atlas-based segmentation of pelvic structures from CT scans for planning in prostate cancer radiotherapy," in *Abdomen and Thoracic Imaging*, A. S. El-Baz, L. Saba, and J. Suri, Eds., pp. 623–656, Springer, Boston, Massachusetts (2014).
70. P. Dirix and S. Nuyts, "Evidence-based organ-sparing radiotherapy in head and neck cancer," *Lancet Oncol.* **11**(1), 85–91 (2010).
71. C. L. Brouwer et al., "3D variation in delineation of head and neck organs at risk," *Radiat. Oncol.* **7**(1), 32 (2012).

Sharmin Sultana is a postdoctoral researcher in the Department of Radiation and Oncology at Johns Hopkins University. She received her BS and MS degrees in computer science and engineering from the University of Dhaka in 2009 and 2011, respectively, and her PhD in computational modeling and simulation engineering from Old Dominion University in 2017. Her current research interests include medical image analysis, computer vision, and deep learning

Adam Robinson is a research assistant in the Department of Radiation Oncology at Johns Hopkins University. He received his BS degree in physics and mathematics from the University of Maryland Baltimore County in 2012, and MS degree in applied and computational mathematics in 2016. His current research interests include medical image analysis and machine learning applied to radiation therapy.

Daniel Y. Song serves as a professor in the Department of Radiation Oncology at Johns Hopkins University. His research focus is on technological innovations for improving the practice of prostate brachytherapy, as well as the conduct of clinical trials in innovative methods of radiotherapy for prostate cancer and other genitourinary malignancies. He performed some of the original research testing the feasibility of hydrogel spacers and establishing their benefit in reducing dose to the rectum.

Junghoon Lee is an associate professor in the Department of Radiation Oncology at Johns Hopkins University. He received his BS in electrical engineering and MS in biomedical engineering in 1997 and 1999, respectively, from Seoul National University, Republic of Korea, and his PhD in electrical and computer engineering from Purdue University in 2006. His research interests are in image processing and computer vision with applications to medical imaging problems.