Segmentation of male pelvic organs on Computed Tomography with a Deep Neural Network fine-tuned by a Level-set method

Gonçalo Almeida[a], Ana Rita Figueira[b], Joana Lencart[c] and João Manuel R. S. Tavares[d,*]

[a]*Instituto de Ciência e Inovação em Engenharia Mecânica e Engenharia Industrial, Faculdade de Engenharia, Universidade do Porto, Porto, Portugal*

[b]*Serviço de Radioterapia, Centro Hospitalar Universitário de São João, Porto, Portugal*

[c]*Serviço de Física Médica e Grupo de Física Médica Radiobiologia e Protecção Radiológica do Centro de Investigação - Instituto Português de Oncologia do Porto (CI-IPOP), Porto, Portugal*

[d]*Instituto de Ciência e Inovação em Engenharia Mecânica e Engenharia Industrial, Departamento de Engenharia Mecânica, Faculdade de Engenharia, Universidade do Porto, Porto, Portugal*

ARTICLE INFO

Keywords:
Deep Learning
Convolutional neural networks
Deformable model
Radiation therapy
Prostate cancer
Computed Tomography imaging

ABSTRACT

Computed Tomography (CT) imaging is used in Radiation Therapy planning, where the treatment is carefully tailored to each patient in order to maximize radiation dose to the target while decreasing adverse effects to nearby healthy tissues. A crucial step in this process is manual organ contouring, which if performed automatically could considerably decrease the time to starting treatment and improve outcomes. Computerized segmentation of male pelvic organs has been studied for decades and deep learning models have brought considerable advances to the field, but improvements are still demanded.

A two-step framework for automatic segmentation of the prostate, bladder and rectum is presented: a convolutional neural network enhanced with attention gates performs an initial segmentation, followed by a region-based active contour model to fine-tune the segmentations to each patient's specific anatomy. The framework was evaluated on a large collection of planning CTs of patients who had Radiation Therapy for prostate cancer.

The Surface Dice Coefficient improved from 79.41 to 81.00% on segmentation of the prostate, 94.03 to 95.36% on the bladder and 82.17 to 83.68% on the rectum, comparing the proposed framework with the baseline convolutional neural network. This study shows that traditional image segmentation algorithms can help improve the immense gains that deep learning models have brought to the medical imaging segmentation field.

1. Introduction

Computed Tomography (CT) is widely used in modern medical practice for screening, diagnosis, monitoring disease progression, cancer staging, treatment planning, and follow-up after successful treatment (15; 10; 21; 16; 2). This imaging modality is equivalent to taking hundreds of planar X-rays from different angles and directions, which are then reconstructed to generate the 3D image we are familiar with (20).

One application of CT is in radiation therapy planning, where the images are used as the base for three-dimensional arrangement of radiation beams to target the desired structures while at the same time minimizing radiation dose to normal, healthy structures (16). CT is the only image modality that can determine tissue electronic densities calibrated to Hounsfield units (34), which is essential for the estimation of dose absorption by the human body, and hence, why it is used for radiotherapy planning (27). The first step in the typical planning process is the manual contouring of anatomical structures, mainly treatment targets and organs-at-risk, performed by medical experts. This manual task is very time-consuming, prone to intra- and inter-observer variation and human error, typically taking 20-60 minutes per patient by a trained expert (12; 9).

The theoretical benefit of an accurate and reliable automated method for organ segmentation is immense, with specific advantages in radiotherapy planning, as it could allow faster time from first patient encounter to treatment start and, possibly, increase treatment quality owing to systematic observance of contouring guidelines (32). A computerized segmentation aid that requires expert validation is of value as the time for validation may be less than the time to perform fully manual segmentation (11).

*Corresponding author

galmeida@inegi.up.pt (G. Almeida);
ana.figueira@chsj.min-saude.pt (A.R. Figueira);
joana.lencart@ipoporto.min-saude.pt (J. Lencart); tavares@fe.up.pt (J.M.R.S. Tavares)

ORCID(s): 0000-0001-5746-0508 (G. Almeida); 0000-0001-7603-6526 (J.M.R.S. Tavares)

Prostate cancer is the most prevalent non-cutaneous cancer and the second leading cause of cancer death in men. It is estimated that about 1 in 9 men will be diagnosed with prostate cancer in their lifetime (40). However, with appropriate treatment, 5-year survival rate is 98.2% (31).

Computerized segmentation of the prostate in medical images has been researched for several years, first with image filters followed by intensity thresholding, then statistical and active shape models and, more recently, with deep learning (DL) models (19; 38). There have been several challenges organized to allow for direct comparison between methods for prostate segmentation (23; 22). However, these are focused on Magnetic Resonance Imaging (MRI), which is more commonly used in clinical practice for diagnosis and cancer staging rather than radiation therapy planning. Given that the tissues of interest are mostly soft tissues, MRI offers better image quality, with higher contrast between structures. Hence, segmentation of pelvic organs on CT is considered more challenging (43). Figure 1 depicts the low contrast problem in unprocessed CT pelvic images.

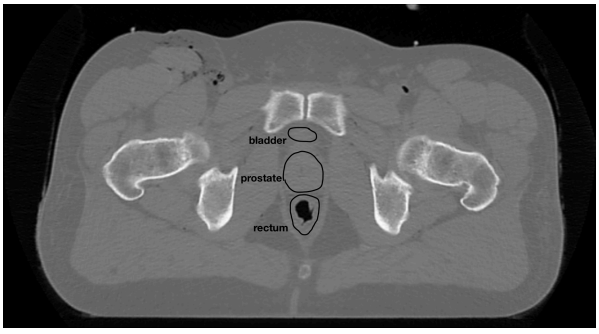

Fig. 1: Example of a CT slice of the male pelvis, where the bladder, prostate and rectum are labelled.

CT scans are composed of a series of planar images, i.e., slices, forming a three-dimensional volume representing the correct anatomy of the patient. In order to perform accurate image segmentation, the use of a model which can take the whole three-dimensional scan into account is preferable. However, this can take very high computational capability. Instead, many methods take each slice as a separate sample and perform segmentation on a 2D basis with good results (39; 25; 26; 46). Alternatively, some authors have experimented with so-called 2.5D frameworks where the two immediately adjacent slices are also processed to aid in the segmentation of the target (centre) slice (1; 41). However, inserting the whole 2D CT scan into a DL model is challenging due to the sheer size of the tensors, which may not fit in the available computational memory, much less using appropriate batch sizes. To avoid this problem, some authors have performed segmentation on a fully three-dimensional basis, using smaller 3D patches of the imaging scans randomly sampled from the whole set (24; 17). This offers performance gains compared to the earlier approaches, and offers support to the hypothesis that a volume which encompasses the whole region of interest could further improve the segmentation results. The reasoning is intuitive: the whole set of information is available for the model to learn from and utilize during inference. An additional strategy is to downsample the images, which is our approach here.

In this study, a state-of-the-art DL model is used for the main segmentation task, followed by a fine-tuning, i.e. refining, step using an active contour model based on the level-set method: the Chan-Vese implementation. The fully convolutional neural network uses CT scans that have been downsized to one-eighth the volume (half-size at each of the three axes), thus avoiding computational memory limitations. The fine-tuning step is performed on the full-resolution images on a slice-by-slice basis, and offers considerable improvement to the DL segmentation for the used dataset.

The contribution of this work is three-fold: collection of the largest dataset of male pelvic CTs for organ segmentation for radiotherapy planning; modification of the U^2-net neural network architecture with attention gates, which improves the performance slightly; and use of a fine-tuning step after the DL segmentation applied at full-resolution, with further performance gains.

Following this introduction, section 2 details the developed approach, including the dataset that was newly collected and the used preprocessing steps. Section 3 describes the results of the performed experiments including a comparison with the baseline U^2-net DL architecture. A discussion of those results is also provided in the same section, and section 4 summarizes the contributions of this work and our future planned work on this relevant field of medical image analysis.

2. Methods

This section describes the dataset and the methods used in the proposed fully automated framework for organ segmentation. Image preprocessing is the first step, followed by the use of a three dimensional fully convolutional neural network, and finally the segmentation fine-tuning using an active contour model: the Chan-Vese approach to a level-set method.

2.1. Dataset

A new dataset of treatment planning CT scans of patients who underwent radiation therapy for prostate cancer

between 2012 and 2020 in two accredited centers in Porto, Portugal: Centro Hospitalar Universitário São João and Instituto Português de Oncologia do Porto, was collected. The full dataset is composed of nearly 4000 scans, of which 2266 were used in this study. This subset contains all patients which performed definitive radiotherapy, meaning the images contain the prostate gland as they had not been subject to surgical removal of the gland. Most of the remaining scans comprise patients who underwent radiotherapy after surgical removal of the prostate. We plan on using those for further work.

The ground-truth manual segmentations were performed and validated by radiation oncology specialists, and were the same that were used for the actual treatment of these patients. Both centers follow identical contouring guidelines (13), and although individual variation is expected (as detailed in the introduction), the relatively large amount of data can compensate for it, particularly because the neural network can perceive the common features from the data and implicitly extract an 'average' between all experts.

The collected scans have varying sizes and resolutions: pixel spacing varies between 0.937 and 1.269 mm, slice thickness is 1.25, 2, 2.5 or 3 mm; number of slices varies between 105 and 345. The manual segmentations include femoral heads and penile bulb for most patients, but, for this study, only the ground-truths for the prostate, bladder and rectum were used. This decision was made after noting that there was little consistency across different patients of the manual segmentations of the penile bulb and femoral heads, due to the scans coming from different institutions and contoured by different experts which do not follow the exact same guidelines. This does not happen in the case of the prostate, bladder and rectum, which are also the most important organs for radiotherapy planning for prostate cancer. The dataset was randomly divided into a training set of 2066 samples, for neural network training; a validation set of 150 samples, for DL hyperparameter search and to determine the parameters for the Level-set fine tuning method; and a test set of 50 samples for final metric evaluation. This split ensures there is no adaptation of the model to the samples that will be used for determining the final results and correctly assess generalizability of the overall framework.

All patient-specific information was anonymized during the data collection process. Approval for this project was received from each of the centers' ethical committees.

2.2. Image Preprocessing

Prior to any computerized image processing and analysis, a few preprocessing tasks ought to be performed, in order to organize and standardize the imaging data and facilitate the subsequent tasks.

As aforementioned, analysis of the dataset revealed a range of pixel spacings and slice thicknesses. An isotropic resampling method was employed to bring all scans to a voxel resolution of 1x1x1 mm^3. This greatly improves DL processing ability, as the tensor arrays that are input into the neural network do not have spacing information, so the model always works under the principle that every image has the same resolution.

Because there is a fixed protocol of image acquisition for treatment planning purposes, in any given patient the prostate is always fairly close to the center of the acquisition volume. Thus, the center region of the scan was cropped to 192x352x192 which was found to be large enough to capture all organs in all samples of this dataset.

Intensity windowing, using the expertise developed by radiologists for years, selects a window appropriate for pelvic CT scans: between -140 and 210 Houndsfield units. This maximizes the contrast between soft tissue structures, such as skeletal and smooth muscle, fat, mucous membranes and prostate gland tissue, important for this specific task. There is also the presence of bone, which becomes overwhelmingly white, but this does not present a problem because it is not the focus in this study. This method reduces the range of intensity values in each sample, facilitating model convergence (18). Intensity normalization to zero mean and unit variance is also performed to improve neural network training.

Lastly, the scans are downscaled to half-size in each axis resulting in a volume that is one-eighth the total size, avoiding computational memory issues with DL model training. Now the volumetric array, and corresponding ground-truth segmentation mask, is ready for neural network processing.

2.3. Fully Convolutional Neural Network

A fully convolutional neural network was employed for the second step in the proposed framework of automatic segmentation. This is a three-dimensional model, based on the U-net architecture, which was presented in 2015 by Ronneberger et al. (37). The used model is the U^2-net, developed by Qin et al. (36), adapted to work with a 3D input, and improved with attention gates similar to those proposed by Oktay et al. for segmentation of the pancreas (33).

The U^2-net is a clever evolution of the U-net, where each processing level is itself a mini-U-net, so this network can be thought of as many U-nets inside a large U-net, as shown in Figure 2. Besides this, it makes use of dilated convolutions (8) and deep supervision (4; 7). This architecture was selected because it is at the state-of-the-art for image segmentation, with numerous applications from salient object detection (36), to human portrait drawing (44) and also

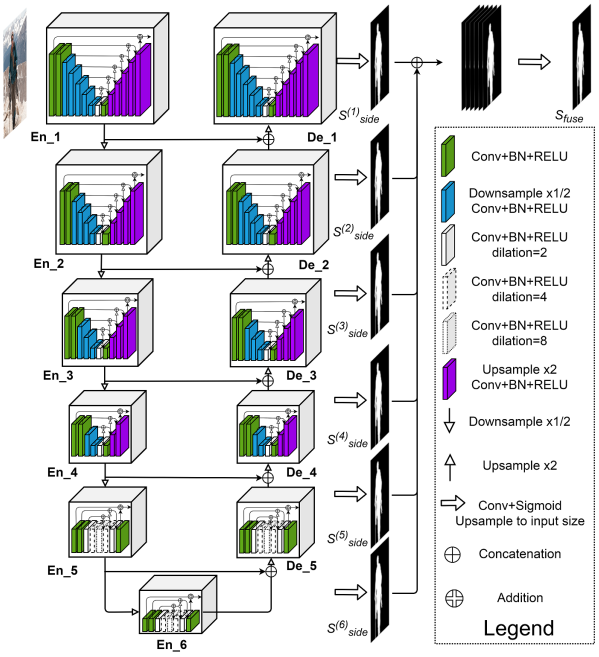

Fig. 2: Diagram of the baseline U^2-net (36).

medical imaging segmentation at the Thyroid Nodule Segmentation and Classification in Ultrasound Images challenge in MICCAI2020 (42). The developed implementation was adapted to a 3D architecture with 3D convolutional layers, as well as 3D MaxPooling and UpSampling layers.

In addition to the powerful features that the baseline network boasts, one found that the implementation of Attention Gates in the skip connections that bridge the transfer of information from the encoder to the decoder arms was beneficial. Figure 3 details the implemented attention gates. This adds a matricial sieve to focus the decoder arm's processing power on relevant regions of the imaging scan, as opposed to other less relevant regions. This is similar to having a weighted filter for the general location of the relevant organs, so that the network would assign more importance to those regions. The advantage of having the attention gate as designed is that it has learnable parameters that the network can tune during the training phase, by decreasing the weights of irrelevant background structures during the gradient backpropagation step (33).

In practical terms, as shown in Figure 3, the image encodings at each level of the left side of the U-net architecture are passed to an attention gate as x^l and is multiplied element-wise by the attention coefficients (α), to yield the output of the attention gate ($\hat{x}^l$). The attention coefficients are produced by an element-wise addition of the tensors x^l and g, the latter of which is the image encoding at the lowest level of the network, i.e. more processed and thus, with a better semantic representation of the image. Elements that are aligned will result in larger weights while the opposite results in a near cancelation of the weights. These are passed through a sigmoid activation layer producing the attention coefficients appropriately scaled between 0 (zero) and 1 (one). The Resampler layer ensures the sizes of the tensors match so that element-wise multiplication is possible.

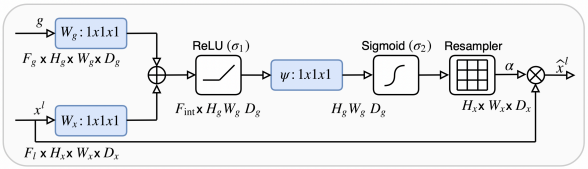

Fig. 3: Diagram of an Attention Gate as used in this work (33).

The network was implemented in Python using the Tensorflow package and trained for 100 epochs for a total time of 17.9 hours with a batch size of 2 and a learning rate of 3x10^{-4} using the Adam optimizer. The loss function was based on the soft volumetric Dice coefficient, which is differentiable (28), added to the categorical crossentropy loss. Training was performed on an NVIDIA DGX station with a V100 GPU with 32Gb of RAM and a 20-core Intel Xeon CPU. As stated above, the input samples were downscaled so the resulting segmentation masks were half the size at each axis. These had to be rescaled back to the original full-resolution in order to be used as the starting point for the level-set method.

2.4. Level-set method

In order to fine-tune the segmentation performed by the deep neural network on the coarse images to the fine detail of the full-resolution scans of each patient, a level-set method was implemented based on the Chan-Vese model, which is established as an energy minimization problem, applied to a level-set formulation. Being an active contour model, it relies on the evolution of a curve to define the segmentation, but does not depend on an edge-function to stop (3). Hence, it can segment objects whose edges do not have a clear gradient, and can have very smooth boundaries or a very noisy image, which are commonly seen in medical images, and for which the classical active contour models usually fail. Its stopping term is based on the Mumford-Shah functional (29). The basic idea is the minimization of a fitting term, which is positive for all points outside the target object and also for all points inside it, such that the only set of points to perfectly minimize the term is the correct boundary of the object of interest. The active curve will evolve towards that

	Prostate	Bladder	Rectum
μ	0.05	0.20	0.25
λ_1	160	140	135
λ_2	1.0	1.0	1.0
dt	1.5	0.5	1.5
iterations	6	8	3

Table 1
Best parameters found for the Chan-Vese method based on the validation set of 150 patients for each of the organs segmented. (μ corresponds to the curve length parameter; λ_1 is the difference from the average weight parameter for the values inside the region-of-interest, and λ_2 is the same for values outside the region-of-interest; dt is a multiplication factor applied at each iteration to the energy function.)

set of points in a sequence of iterations (45). There are also some regularizing terms, which correspond to the length of the curve and the area of the region inside the object, which help add stability to the method (3).

This model requires an initial segmentation, which typically consists of a basic geometric form, such as a circle on the center of the input image, and greatly influences the end result (45). In our case, by using the segmentation resultant of the DL model, one of the problems of implementing a level-set method in practice is overcome. This initialization also helps with the problems of stability and convergence, as all that is required of the level-set method is a small fine-tuning step, adjusting the already high quality DL segmentation to the organ boundaries evident in the full-resolution scans.

The Chan-Vese model was applied on a slice-by-slice basis, to each organ separately, using the neural network segmentation as starting point. The slices were then joined to rebuild the finished volumetric segmentation. The Chan-Vese method takes several parameters, which were found by successive trial experimentation, using the validation set composed of 150 samples. For each organ, in order to maximized the target metrics, a different set of parameters was found, which are shown in Table 1.

2.5. Evaluation metrics

For a comparison between segmentation masks, i.e. segmentation results, to be possible, useful and representative metrics that can be easily measured are needed. In this study, the segmentation quality was evaluated using four metrics, to capture a robust picture of the performance of the proposed method. However, for validation purposes, in order to select the best parameters, only one metric can be used, because some small parameter changes can benefit one metric, but worsen others. For this, the Surface Dice-Sørensen Coefficient (DSC) suggested in (30) was chosen.

This surface DSC metric is tailored to the boundary of the segmented object, ignoring the voxels that are clearly inside or outside the volume. We found this to be much more aligned with our own qualitative evaluation of the segmentations for this particular problem, since the edges are the most important feature, but also because all organs are filled (as opposed to hollow) and there is no influence by the overall size and shape of the organ, unlike other metrics; e.g., the volumetric Dice-Sørensen Coefficient favors large round objects such as the bladder over long and thin ones such as the rectum. Furthermore, the surface DSC allows for the use of a tolerance factor, which determines an acceptable deviation from the ground-truth, only penalizing deviations larger than that tolerance, as shown in Figure 4. Often a small deviation in the boundary of an organ is not clinically significant, although it can have a large impact on volumetric DSC if the organ is small; and the opposite can also happen where a relatively large deviation in the boundary can have little effect on volumetric DSC if the organ is large, but it can have a significant impact in clinical terms. In this study, a tolerance of 2 mm was used, as this corresponds to the voxel resolution of the downsampled images used for the neural network training.

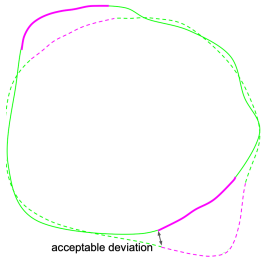

Fig. 4: Detail of the surface DSC metric focused on the surface of the segmentation, with an 'acceptable deviation' parameter. The metric only penalizes contour points which are far from the ground-truth more than the acceptable tolerance. (Continuous line: ground-truth; Dashed line: model segmentation; adapted from (30).)

Nevertheless, for final test purposes and results reporting, the standard volumetric DSC was also calculated (6), as is the case in most organ segmentation studies (39; 25; 24; 17; 4; 14):

$$Dice\ Sorensen\ Coefficient = \frac{2|A \cap B|}{|A| + |B|}, \tag{1}$$

where A and B are the segmented volumes to be compared.

The other metrics calculated are based on linear point distances. The Average Boundary Distance, ABD, and Hausdorff distance, HD, are computed as:

$$ABD = \frac{1}{|A_s| + |B_s|} \cdot \left(\sum_{a \in A_s} min_{b \in B_s} ||a - b|| + \sum_{b \in B_s} min_{a \in A_s} ||a - b|| \right), \tag{2}$$

$$HD = max_{a \in A_s}(min_{b \in B_s} ||a - b||), \tag{3}$$

where A_s and B_s are the surfaces of the segmented volumes to be compared, and $||a - b||$ is the Euclidean distance between two points on A and B.

All metrics were calculated with the full-resolution scans, i.e., upscaling the segmentation masks produced by the models as needed.

3. Experimental results and discussion

Three segmentation methods were compared for their performance on the test set: (A) the baseline U^2-net as implemented by its authors, available in (35); (B) the modified neural network, with the addition of attention gates; and (C) the overall framework, as described in Section 2. The same preprocessing steps were performed for the three methods.

The achieved quantitative results are summarized in Table 2. Both on average as well as for each organ individually, the proposed combined framework produced the segmentations that are closest to those of human experts. On average, the performance gain from the modified network architecture was not as large as that offered by the fine-tuning step. Overall, the highest result was achieved for the bladder and the prostate proved to be the most difficult. This can be explained by the fact that a large region of the bladder is usually surrounded by fatty tissue, which appears darker than the organs, while the prostate is often touching the bladder anteriorly and rectum posteriorly, which have approximately the same intensity values.

The runtime for making segmentations on the downsized scans with the trained DL model was 0.373 seconds per patient (with GPU acceleration), while the level-set step took 1.399 seconds per patient. Although the level-set step adds considerable time in relative terms to the overall framework, it is negligible in terms of the typical time taken by medical experts to perform segmentation of a radiotherapy case. Even if it were to take much longer, it could be parallelized with other tasks such as patient consultation.

The overlap metrics clearly show that the level-set step improved the quality of the segmentation for each of the three organs. The gains were slightly better in the case of Surface DSC, which is explained by the fact that small changes in the boundary position can have a larger effect on this metric as opposed to volumetric DSC, which relies on the size of the whole sets. This is also the reason why volumetric DSC was slightly higher for any method and for any organ, due to the large amount of overlapping voxels in the middle of the segmentations. The Hausdorff distance did not show a significant improvement, in fact it becomes slightly worse for the prostate and rectum. This metric gives the maximum distance of one set to the nearest point of the other set, which is very influenced by the distances in the top-down axis. While the scans were resampled to 1x1x1 mm^3 resolution, the ground-truth segmentation is capped at a specific slice level, often causing the Hausdorff distance to be on the top or bottom of the organ, overwhelming the comparably smaller changes in the segmentations produced by the level-set method in the anterior-posterior and lateral axes. This is less noticeable in the case of the bladder, which has more clearly defined top and bottom. To overcome this limitation, the average boundary distance is used, which is the average of the distances of all points in a surface to the closest point in the reference surface. For this reason, this metric better reflected a modest improvement for all three organs with the use of the level-set method.

A statistical analysis was performed based on the paired samples T-test which assesses the difference between the means of two groups with measurements of the same subject, in this case organ segmentations of the same patient. For most of the metrics evaluated, statistical significance was found, as shown in Table 2, which also indicates a significant advantage with the level-set fine-tuning step compared to the method that relies only on a neural network.

Another important quantitative result is the percentage of patients in the test set whose segmentations improved after the level-set method, which were 78%, 94% and 62% for the prostate, bladder and rectum, respectively, based on the surface DSC. Furthermore, in all but one of the 50 test set patients at least one of the organ segmentations improved with the level-set method, compared to the U^2-net

		Prostate	Bladder	Rectum	*Average*
Surface Dice Coefficient (%)	(A)	79.41 ± 1.23	94.03 ± 0.50	82.17 ± 0.78	85.20 ± 0.65
	(B)	79.67 ± 1.06	92.96 ± 0.52	83.63 ± 0.80	85.42 ± 0.60
	(C)	**81.00 ± 1.02***	**95.36 ± 0.36***	**83.68 ± 0.74**	**86.68 ± 0.53***
Volumetric Dice Coefficient (%)	(A)	85.91 ± 0.45	94.63 ± 0.22	83.02 ± 0.59	87.85 ± 0.31
	(B)	85.98 ± 0.37	94.54 ± 0.20	83.86 ± 0.60	88.13 ± 0.28
	(C)	**86.21 ± 0.36***	**95.55 ± 0.19***	**84.44 ± 0.59***	**88.73 ± 0.28***
95% Hausdorff Distance (mm)	(A)	4.60 ± 2.64	2.41 ± 0.85	13.02 ± 9.00	6.68 ± 3.30
	(B)	**4.59 ± 2.41**	2.61 ± 0.96	**10.61 ± 8.68**	5.94 ± 3.11
	(C)	4.63 ± 2.43	**2.39 ± 1.10***	10.66 ± 8.61	**5.89 ± 3.11**
Average Boundary Distance (mm)	(A)	1.40 ± 0.51	0.85 ± 0.23	1.92 ± 0.96	1.39 ± 0.42
	(B)	1.38 ± 0.40	0.87 ± 0.23	1.73 ± 0.96	1.33 ± 0.39
	(C)	**1.31 ± 0.40***	**0.66 ± 0.22***	**1.66 ± 0.94**	**1.21 ± 0.38***

Table 2
Comparative results of the frameworks under study using the test set: (A) Baseline U^2-net; (B) U^2-net with attention gates; and (C) U^2-net with attention gates followed by level-set fine-tuning at full resolution - the proposed method. (Mean and standard deviation are presented; best values are in bold. * indicates statistical significance in the paired samples T-test: $p<0.05$)

with attention gates alone; likewise, the segmentation was improved for all three organs in 23 of the patients.

As for qualitative results, in Figure 5 it is depicted that the level-set method can improve the segmentation particularly in areas close to the bone where the DL model does not respect bone surface limits (visible on the right column images). One other important detail to note is as to the anterior part of the bladder (on the top of the images): because the level-set method uses the DL output as starting point, it does not add the missing segmentation on the first set of images (on the left column), adjusting only the part closest to the center and, actually, worsening the simple 2D Dice coefficient on this particular slice. However, in the second set (on the centre column), the level-set improves the segmentation quality by reducing the size of the part of the bladder that was oversegmented by the neural network. Overall, in these images, the fine-tuning step decreased the segmentation quality in the first slice, but improved in the others, 60.46 to 57.21%, 76.45 to 83.68% and 86.56 to 92.22%, respectively.

Similarly, on the images of the left column of Figure 6, the DL segmentation had a general shape typical of the prostate shape for a slice at this level, but the level-set adjustment was able to better adapt the prostate contour to this particular patient, improving the evaluation metrics, from 82.14 to 92.45%. On the images of the centre and right columns, examples of slight oversegmentation by the DL model on the rectum can be observed, which were corrected by the fine-tuning step. The 2D Dice coefficient improved from 80.12 to 85.12% and 78.56 to 95.53%, on the images of the centre and right columns, respectively.

For a better representation of the achieved three-dimensional segmentations, Figure 7 shows sagittal and coronal views of the same cases shown in the axial slices for each of the three organs.

The DL model was able to accurately determine the average shape of the organs required and exclude individual particularities to learn the main characteristics of prostate, bladder and rectum shapes, not too dependent on specific image details, but taking into account the general location compared to other anatomical landmarks independent of imaging artifacts. However, it lacked in fine detail, often overflowing into nearby bone structures (high intensity, lighter gray in CT) or soft fatty tissue (low intensity, darker gray on CT). The strength in the proposed combined approach stems from the adjustment of the level-set method as a means of post-processing, such that the final segmentation is better adapted to the anatomy and fine details of each patient.

There is reason to believe that a fully end-to-end DL framework on the full-resolution scans could surpass the accuracy of the proposed hybrid model. However, it is more computationally intensive, and more time-consuming to fine-tune a neural network model and perform hyper parameter search than to iterate on the level-set method used in the proposed method. Likewise, a coarse DL network followed by a fine DL network trained on a volumetric patch of the full-size images could provide accurate results and presents as an interesting avenue of research to pursue.

By performing the fine-tuning step on each organ individually, the resulting segmentations have a few superpositions between the organs, even if very limited because of the non-overlapping starting points. While not impactful in the

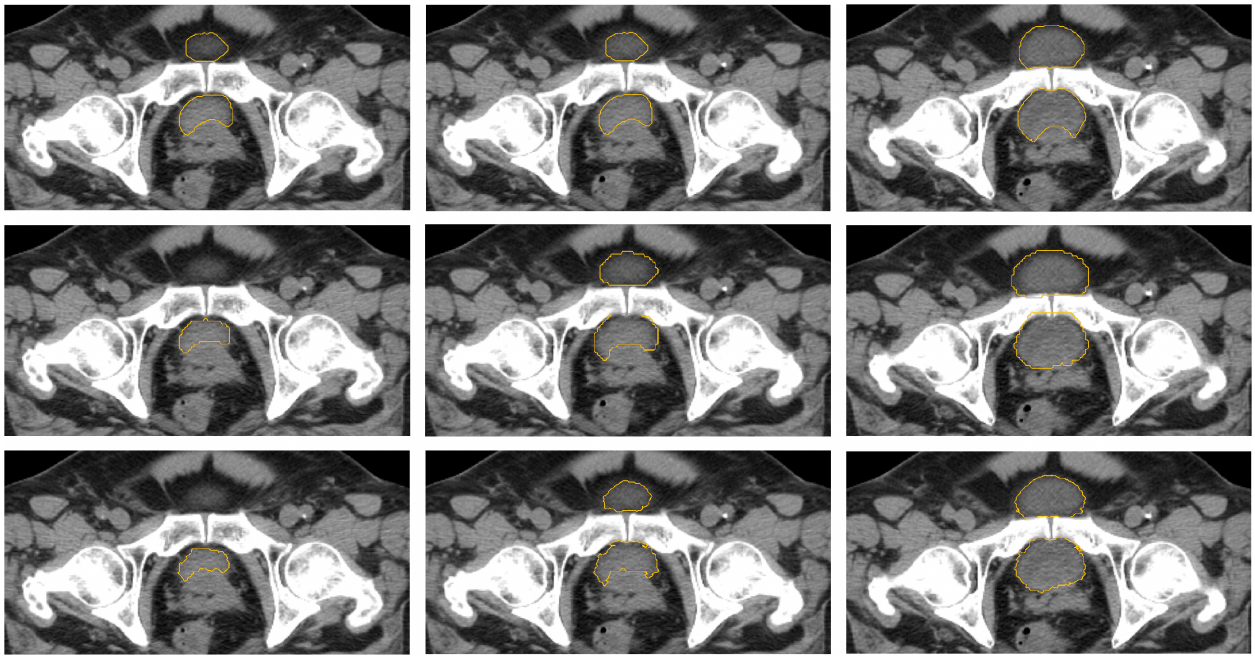

Fig. 5: Segmentation of the bladder in three slices of the same patient. Top row: Manual ground-truth; Middle row: U^2-net with attention gates; and Bottom row: full proposed framework.

quantitative metrics, this is not ideal in the clinical setting as in this particular case there should be no overlaps between these organs. There may be computational algorithms, namely adversary curve evolutions, to help overcome this problem, which we plan to explore in future work.

The use of a specific set of parameters for the level-set method can impact the generalizability of the overall workflow, and it is why we performed the parameter search in a randomly picked validation set, but finally assessing the overall method on a different test set. This way we show that these parameters are not only tuned to the specific images for which they were chosen. Still, all CT scans come from the same dataset, but even these come from two different centers, and the ground-truths were contoured by more than a dozen experts, with the inherent variation thereof. Besides, the CT scans used for radiotherapy planning are specially calibrated so that the Hausdorff units have some correlation with the mass number of the cellular tissues, in order for radiation absorption estimations to be made; this results in more consistent images than typical CT scans for other uses (5).

A limitation of the 2D approach implemented for the level-set method is that it only adjusts the segmentation performed by the DL model on a given slice, such that it never decides to eliminate a segmentation from a given slice, nor add a new one to an adjacent slice that was not segmented by the DL model. This has an impact on the calculated metrics, because any slice that has no ground-truth segmentation, for the reason that the organ of interest is not displayed on that slice, will have a DSC of 0 (zero), both for the DL segmentation and after the level-set refinement. This affected the rectum and was particularly noticeable in the Hausdorff distance metric. Furthermore, in some scans there was a lack of smooth continuity from slice to slice after the final step of the proposed framework. This is not sufficient to negate its benefits, but shows a pathway for further improvements.

4. Conclusion

This article presented a combined framework to perform organ segmentation on three-dimensional CT scans of the male pelvis, where a DL model is compounded by a region-based active contour method, which is able to fine-tune the segmentation to the patient's specific anatomy and improve key metrics.

The framework was applied to a large dataset of radiation therapy planning CTs whose breadth of anatomical variation helps validate the achieved results. The application of attention gates to the U^2-net architecture was successful for this specific task, and improved the baseline by adding a processing layer at the bridge between encoder and decoder

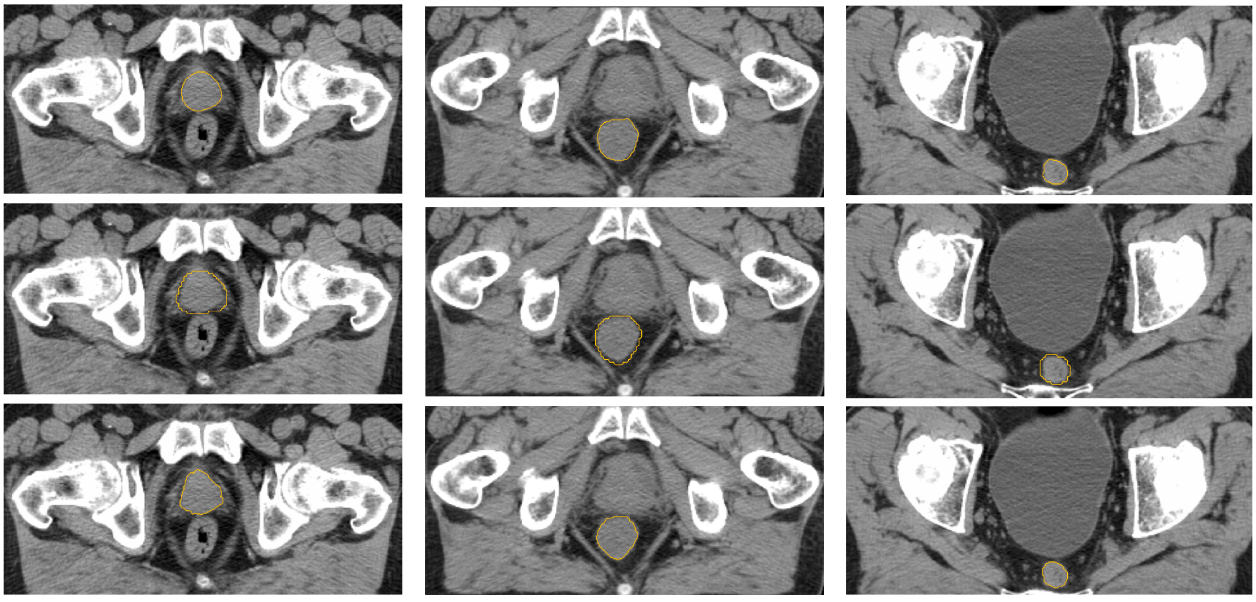

Fig. 6: Segmentation of prostate (left column) and rectum (centre and right columns). Top row: Manual ground-truth; Middle row: U^2-net with attention gates; Bottom row: full proposed framework.

arms of the network. The Chan-Vese level-set method was shown to perform an appropriate final adjustment to each patient's anatomy allowing for the DL model to be applied to a lower resolution scan avoiding computational memory issues and decreasing computational needs.

In essence, this study shows there is a pathway for traditional image segmentation algorithms to help improve the immense gains that DL neural networks have brought to the medical imaging segmentation domain. As future work, we plan on improving this framework by adding a method to avoid superpositions of the different organs, and also applying it to CT scans of patients who had the prostate surgically removed before undergoing radiation therapy, where the surgical bed is the target region.

Acknowledgments

The authors would like to acknowledge and thank *Fundação para a Ciência e Tecnologia* (FCT) for the PhD grant (reference SFRH/BD/146887/2019) awarded to the first author, which this work is a part of.

References

[1] Ruba Alkadi, Ayman El-Baz, Fatma Taher, and Naoufel Werghi. A 2.5D deep learning-based approach for prostate cancer detection on T2-weighted magnetic resonance imaging. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11132 LNCS:734–739, 2019. doi:10.1007/978-3-030-11018-5_66.

[2] David J. Brenner and Eric J. Hall. Computed tomography: An increasing source of radiation exposure. *New England Journal of Medicine*, 357(22):2277–2284, 2007. PMID: 18046031. URL: https://doi.org/10.1056/NEJMra072149, arXiv:https://doi.org/10.1056/NEJMra072149, doi:10.1056/NEJMra072149.

[3] T.F. Chan and L.A. Vese. Active contours without edges. *IEEE Transactions on Image Processing*, 10(2):266–277, 2001. doi:10.1109/83.902291.

[4] Ruida Cheng, Holger R. Roth, Nathan Lay, Le Lu, Baris Turkbey, William Gandler, Evan S. McCreedy, Peter Choyke, Ronald M. Summers, and Matthew J. McAuliffe. Automatic MR prostate segmentation by deep learning with holistically-nested networks. *Medical Imaging 2017: Image Processing*, 10133(4):101332H, 2017. doi:10.1117/12.2254558.

[5] Anne T Davis, Antony L Palmer, and Andrew Nisbet. Can CT scan protocols used for radiotherapy treatment planning be adjusted to optimize image quality and patient dose? A systematic review. *The British journal of radiology*, 90(1076):20160406, aug 2017. doi:10.1259/bjr.20160406.

[6] L. R. Dice. Measures of the amount of ecologic association between species, 1945. URL: https://doi.wiley.com/10.2307/1932409, doi:10.2307/1932409.

[7] Xue Dong, Yang Lei, Sibo Tian, Tonghe Wang, Pretesh Patel, Walter J. Curran, Ashesh B. Jani, Tian Liu, and Xiaofeng Yang. Synthetic MRI-aided multi-organ segmentation on male pelvic CT using cycle consistent deep attention network. *Radiother Oncol*, pages 1–8, 2019. URL: https://doi.org/10.1016/j.radonc.2019.09.028, doi:10.1016/j.radonc.2019.09.028.

[8] Vincent Dumoulin and Francesco Visin. A guide to convolution arithmetic for deep learning, 2016. arXiv:1603.07285.

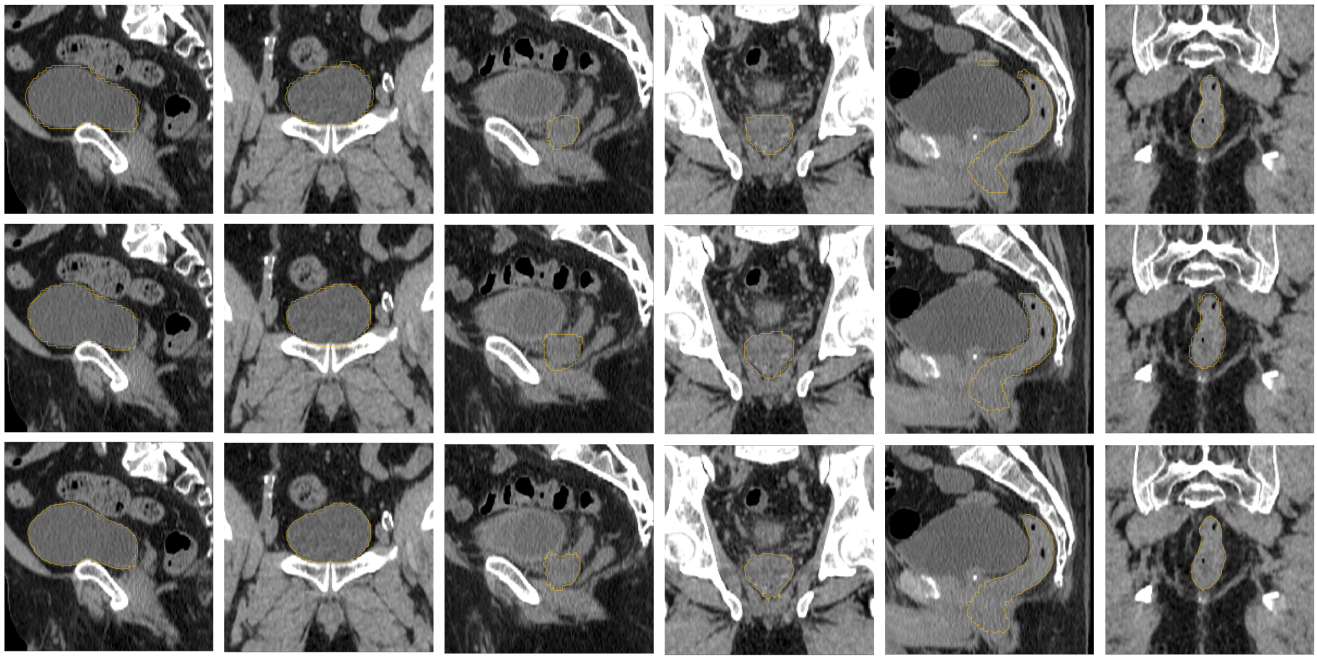

Fig. 7: Sagittal and coronal views of segmentation of the bladder (two leftmost columns), prostate (two centre columns) and rectum (two rightmost columns). Top row: Manual ground-truth; Middle row: U^2-net with attention gates; Bottom row: full proposed framework.

[9] Claudio Fiorino, Michele Reni, Angelo Bolognesi, Giovanni Mauro Cattaneo, and Riccardo Calandrino. Intra- and inter-observer variability in contouring prostate and seminal vesicles: Implications for conformal treatment planning. *Radiother Oncol*, 47(3):285–292, 1998. doi:10.1016/S0167-8140(98)00021-8.

[10] Lisa J. Forrest. Computed tomography imaging in oncology. *Veterinary Clinics of North America: Small Animal Practice*, 46(3):499–513, 2016. Diagnostic Radiology. URL: https://www.sciencedirect.com/science/article/pii/S0195561615001849, doi:https://doi.org/10.1016/j.cvsm.2015.12.007.

[11] I. Fotina, C. Lütgendorf-Caucig, M. Stock, R. Pötter, and D. Georg. Critical discussion of evaluation parameters for inter-observer variability in target definition for radiation therapyKritische Diskussion von Evaluierungsparametern der Inter-Beobachter-Variabilität bei der Konturierung von Zielvolumina in der Strahlenth. *Strahlentherapie und Onkol*, 188(2):160–167, 2012. doi:10.1007/s00066-011-0027-6.

[12] Zhanrong Gao, David Wilkins, Libni Eapen, Christopher Morash, Youssef Wassef, and Lee Gerig. A study of prostate delineation referenced against a gold standard created from the visible human data. *Radiother Oncol*, 85(2):239–246, 2007. doi:10.1016/j.radonc.2007.08.001.

[13] HA Gay, HJ Barthold, E O'Meara, WR Bosch, IE Naga, R Al-Lozi, SA Rosenthal, C Lawton, WR Lee, H Sandler, A Zietman, R Myerson, LA Dawson, C Willett, LA Kachnic, A Jhingran, L Portelance, J Ryu, W Small, D Gaffney, AN Viswanathan, and JF Michalski. Male Pelvis Normal Tissue - RTOG Consensus Contouring Guidelines. Technical report, 2007. URL: https://www.rtog.org/LinkClick.aspx?fileticket=054g99vNGps{%}3D{&}tabid=354.

[14] Lei Geng, Jia Wang, Zhitao Xiao, Jun Tong, Fang Zhang, and Jun Wu. Encoder-decoder with dense dilated spatial pyramid pooling for prostate MR images segmentation. *Comput Assist Surg*, 24(sup2):13–19, 2019. URL: https://doi.org/10.1080/24699322.2019.1649069, doi:10.1080/24699322.2019.1649069.

[15] Colin Jacobs, Anton Schreuder, Sarah J. van Riel, Ernst Th. Scholten, Rianne Wittenberg, Mathilde M. Winkler Wille, Bartjan de Hoop, Ralf Sprengers, Onno M. Mets, Bram Geurts, Mathias Prokop, Cornelia Schaefer-Prokop, and Bram van Ginneken. Assisted versus manual interpretation of low-dose ct scans for lung cancer screening: Impact on lung-rads agreement. *Radiology: Imaging Cancer*, 3(5):e200160, 2021. PMID: 34559005. URL: https://doi.org/10.1148/rycan.2021200160, arXiv:https://doi.org/10.1148/rycan.2021200160, doi:10.1148/rycan.2021200160.

[16] Paul Keall. 4-dimensional computed tomography imaging and treatment planning. *Seminars in Radiation Oncology*, 14(1):81–90, 2004. High-Precision Radiation Therapy of Moving Targets. URL: https://www.sciencedirect.com/science/article/pii/S1053429603000870, doi:https://doi.org/10.1053/j.semradonc.2003.10.006.

[17] Vasant Kearney, Jason W Chan, Tianqi Wang, Alan Perry, Sue S Yom, and Timothy D Solberg. Attention-enabled 3D boosted convolutional neural networks for semantic CT segmentation using deep supervision. *Phys Med Biol*, 64(13):135001, 2019. doi:10.1088/1361-6560/ab2818.

[18] Marie Kloenne, Sebastian Niehaus, Leonie Lampe, Alberto Merola, Janis Reinelt, Ingo Roeder, and Nico Scherf. Domain-specific cues improve robustness of deep learning-based segmentation of CT volumes. *Scientific Reports*, 10(1):10712, 2020. URL: https://doi.org/10.1038/s41598-020-67544-y, doi:10.1038/s41598-020-67544-y.

[19] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015. doi:10.1038/nature14539.

[20] Michael M Lell and Marc Kachelrieß. Recent and Upcoming Technological Developments in Computed Tomography: High Speed, Low Dose, Deep Learning, Multienergy. *Investigative Radiology*, 55(1), 2020. URL: https://journals.lww.com/investigativeradiology/Fulltext/2020/01000/Recent_and_Upcoming_Technological_Developments_in.2.aspx, doi:10.1097/RLI.0000000000000601.

[21] Anke M. Leufkens, Maurice A.A.J. van den Bosch, Maarten S. van Leeuwen, and Peter D. Siersema. Diagnostic accuracy of computed tomography for colon cancer staging: A systematic review. *Scandinavian Journal of Gastroenterology*, 46(7-8):887–894, 2011. URL: https://doi.org/10.3109/00365521.2011.574732, arXiv:https://doi.org/10.3109/00365521.2011.574732, doi:10.3109/00365521.2011.574732.

[22] Geert Litjens, Oscar Debats, Jelle Barentsz, Nico Karssemeijer, and Henkjan Huisman. ProstateX Challenge data, 2017. doi:10.7937/K9TCIA.2017.MURS5CL.

[23] Geert Litjens, Robert Toth, Wendy van de Ven, Caroline Hoeks, Sjoerd Kerkstra, Bram van Ginneken, Graham Vincent, Gwenael Guillard, Neil Birbeck, Jindang Zhang, Robin Strand, Filip Malmberg, Yangming Ou, Christos Davatzikos, Matthias Kirschner, Florian Jung, Jing Yuan, Wu Qiu, Qinquan Gao, Philip Eddie Edwards, Bianca Maan, Ferdinand van der Heijden, Soumya Ghose, Jhimli Mitra, Jason Dowling, Dean Barratt, Henkjan Huisman, and Anant Madabhushi. Evaluation of prostate segmentation algorithms for MRI: The PROMISE12 challenge. *Med Image Anal*, 18(2):359–373, 2014. URL: http://dx.doi.org/10.1016/j.media.2013.12.002, doi:10.1016/j.media.2013.12.002.

[24] Chang Liu, Stephen J. Gardner, Ning Wen, Mohamed A. Elshaikh, Farzan Siddiqui, Benjamin Movsas, and Indrin J. Chetty. Automatic Segmentation of the Prostate on CT Images Using Deep Neural Networks (DNN). *International Journal of Radiation Oncology Biology Physics*, 104(4):924–932, 2019. URL: https://doi.org/10.1016/j.ijrobp.2019.03.017, doi:10.1016/j.ijrobp.2019.03.017.

[25] Ling Ma, Rongrong Guo, Guoyi Zhang, Funmilayo Tade, David M. Schuster, Peter Nieh, Viraj Master, and Baowei Fei. Automatic segmentation of the prostate on CT images using deep learning and multi-atlas fusion. *Medical Imaging 2017: Image Processing*, 10133:101332O, 2017. doi:10.1117/12.2255755.

[26] Meghan W. Macomber, Mark Phillips, Ivan Tarapov, Rajesh Jena, Aditya Nori, David Carter, Loic Le Folgoc, Antonio Criminisi, and Matthew J. Nyflot. Autosegmentation of prostate anatomy for radiation treatment planning using deep decision forests of radiomic features. *Physics in Medicine and Biology*, 63(23), 2018. doi:10.1088/1361-6560/aaeaa4.

[27] Philippe Meyer, Vincent Noblet, Christophe Mazzara, and Alex Lallement. Survey on deep learning for radiotherapy. *Comput Biol Med*, 98(May):126–146, 2018. URL: https://doi.org/10.1016/j.compbiomed.2018.05.018, doi:10.1016/j.compbiomed.2018.05.018.

[28] Fausto Milletari, Nassir Navab, and Seyed Ahmad Ahmadi. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. *Proc - 2016 4th Int Conf 3D Vision, 3DV 2016*, pages 565–571, 2016. arXiv:arXiv:1606.04797v1, doi:10.1109/3DV.2016.79.

[29] D. Mumford and J. Shah. Optimal approximations by piecewise smooth functions and associated variational problems. *Communications on Pure and Applied Mathematics*, 42:577–685, 1989. doi:10.1002/cpa.3160420503.

[30] Stanislav Nikolov, Sam Blackwell, Alexei Zverovitch, Ruheena Mendes, Michelle Livne, Jeffrey De Fauw, Yojan Patel, Clemens Meyer, Harry Askham, Bernardino Romera-Paredes, Christopher Kelly, Alan Karthikesalingam, Carlton Chu, Dawn Carnell, Cheng Boon, Derek D'Souza, Syed Ali Moinuddin, Bethany Garie, Yasmin McQuinlan, Sarah Ireland, Kiarna Hampton, Krystle Fuller, Hugh Montgomery, Geraint Rees, Mustafa Suleyman, Trevor Back, CÃŋan Hughes, Joseph R. Ledsam, and Olaf Ronneberger. Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy, 2021. arXiv:1809.04430.

[31] AM Noone, N Howlader, M Krapcho, D Miller, A Brest, M Yu, J Ruhl, Z Tatalovich, A Mariotto, DR Lewis, HS Chen, EJ Feuer, and KA Cronin. SEER cancer statistics review, National Cancer Institute. Technical report, 2017.

[32] The Royal College of Radiologists, Society of Radiographers, College, Institute of Physics in Medicine, and Engineering. On target: ensuring geometric accuracy in radiotherapy. Technical report, The Royal College of Radiologists, 2008.

[33] Ozan Oktay, Jo Schlemper, Loic Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Hammerla, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention U-Net: Learning Where to Look for the Pancreas. *1st Conf Med Imaging with Deep Learn (MIDL 2018)*, 2018.

[34] Gisele C. Pereira, Melanie Traughber, and Raymond F. Muzic. The role of imaging in radiation therapy planning: Past, present, and future. *Biomed Res Int*, 2014(2), 2014. doi:10.1155/2014/231090.

[35] Xuebin Qin and Zichen Zhang et al. Github repository for U2-Net, found at https://github.com/xuebinqin/U-2-Net, 2020.

[36] Xuebin Qin, Zichen Zhang, Chenyang Huang, Masood Dehghan, Osmar R. Zaiane, and Martin Jagersand. U2-net: Going deeper with nested u-structure for salient object detection. *Pattern Recognition*, 106:107404, Oct 2020. URL: http://dx.doi.org/10.1016/j.patcog.2020.107404, doi:10.1016/j.patcog.2020.107404.

[37] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Nassir Navab, Joachim Hornegger, William M Wells, and Alejandro F Frangi, editors, *Med Image Comput Comput Interv – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing.

[38] Maysam Shahedi, Martin Halicek, James D Dormer, David M Schuster, and Baowei Fei. Deep learning-based three-dimensional segmentation of the prostate on computed tomography images. *Journal of medical imaging (Bellingham, Wash.)*, 6(2):25003, apr 2019. doi:10.1117/1.JMI.6.2.025003.

[39] Yinghuan Shi, Wanqi Yang, Yang Gao, and Dinggang Shen. Does Manual Delineation only Provide the Side Information in CT Prostate Segmentation? In M. Descoteaux, editor, *MICCAI 2017, Part III, LNCS 10435*, volume 10435, pages 692–700. Springer International Publishing, 2017. doi:10.1007/978-3-319-66179-7_79.

[40] American Cancer Society. Facts & Figures 2019. Technical report, 2019.

[41] Simon John Christoph Soerensen, Richard Fan, Arun Seetharaman, Leo Chen, Wei Shao, Indrani Bhattacharya, Michael Borre, Benjamin Chung, Katherine To'o, Geoffrey Sonn, and Mirabela Rusu. ProGNet: prostate gland segmentation on MRI with deep learning. In Ivana Išgum and Bennett A Landman, editors, *Medical Imaging 2021: Image Processing*, volume 11596, pages 743–750. International Society for Optics and Photonics, SPIE, 2021. URL: https://doi.org/10.1117/12.2580448, doi:10.1117/12.2580448.

[42] Manh The Van, JianQiao Zhou, XiaoHong Jia, Yijie Dong, Dong Ni, Alison Noble, RuoBing Huang, Tao Tan, Xing Tao, and Rui Li. Tn-scui2020 thyroid nodule segmentation and classification in ultrasound images, 2020. URL: https://tn-scui2020.grand-challenge.org/Home/.

[43] Wanqi Yang, Yinghuan Shi, Sang Hyun Park, Ming Yang, Yang Gao, and Dinggang Shen. An Effective MR-Guided CT Network Training

for Segmenting Prostate in CT Images. *IEEE journal of biomedical and health informatics*, 24(8):2278–2291, aug 2020. doi:10.1109/JBHI.2019.2960153.

[44] Ran Yi, Yong-Jin Liu, Yu-Kun Lai, and Paul L. Rosin. Apdrawinggan: Generating artistic portrait drawings from face photos with hierarchical gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[45] Haiping Yu, Fazhi He, and Yiteng Pan. A survey of level set method for image segmentation with intensity inhomogeneity. *Multimedia Tools and Applications*, 79(39):28525–28549, 2020. URL: https://doi.org/10.1007/s11042-020-09311-9, doi:10.1007/s11042-020-09311-9.

[46] Sihang Zhou, Dong Nie, Ehsan Adeli, Jianping Yin, Jun Lian, and Dinggang Shen. High-Resolution Encoder-Decoder Networks for Low-Contrast Medical Image Segmentation. *IEEE Transactions on Image Processing*, 29(X):461–475, 2019. doi:10.1109/tip.2019.2919937.