



Universiteit  
Leiden  
The Netherlands

## Deep learning for online adaptive radiotherapy

Elmahdy, M.S.E.

### Citation

Elmahdy, M. S. E. (2022, March 15). *Deep learning for online adaptive radiotherapy*. Retrieved from <https://hdl.handle.net/1887/3278960>

Version: Publisher's Version

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

Downloaded from: <https://hdl.handle.net/1887/3278960>

**Note:** To cite this publication please use the final published version (if applicable).

# 3

## Patient Specific Automatic Segmentation of Prostate and Organs-at-Risk for Adaptive Radiotherapy

*This chapter was adapted from:*

M Elmahdy, T Ahuja, U van der Heide, and M Staring. **Patient Specific Automatic Segmentation of Prostate and Organs-at-Risk For Adaptive Radiotherapy**, *International Symposium on Biomedical Imaging (ISBI)*, Pages 577-580, 2020.

## Abstract

Contouring of the target volume and Organs-At-Risk (OARs) is a crucial step in radiotherapy treatment planning. In an adaptive radiotherapy setting, updated contours need to be generated based on daily imaging. In this work, we leverage personalized anatomical knowledge accumulated over the treatment sessions, to improve the segmentation accuracy of a pre-trained Convolution Neural Network (CNN), for a specific patient. We investigate a transfer learning approach, fine-tuning the baseline CNN model to a specific patient, based on imaging acquired in earlier treatment fractions. The baseline CNN model is trained on a prostate CT dataset from one hospital of 379 patients. This model is then fine-tuned and tested on an independent dataset of another hospital of 18 patients, each having 7 to 10 daily CT scans. For the prostate, seminal vesicles, bladder and rectum, the model fine-tuned on each specific patient achieved a Mean Surface Distance (MSD) of  $1.64 \pm 0.43$  mm,  $2.38 \pm 2.76$  mm,  $2.30 \pm 0.96$  mm, and  $1.24 \pm 0.89$  mm, respectively, which was significantly better than the baseline model. The proposed personalized model adaptation is therefore very promising for clinical implementation in the context of adaptive radiotherapy of prostate cancer.

### 3.1 Introduction

Prostate cancer is the second leading cause of cancer death in American men, behind lung cancer. About 1 man in 41 will die of prostate cancer. The American Cancer Society's estimates about 174,650 new cases of prostate cancer and about 31,620 deaths from prostate cancer in the United States for 2019 only [36]. Radiation Therapy (RT) is one of the treatment options for prostate cancer that uses a highly localized dose distribution to kill cancer cells. RT dose is usually fragmented over 4 to 8 weeks resulting in 20 to 40 daily fractions [4]. Since the dose is delivered in several sessions, variations in the size and shape of the target area and Organs-At-Risk (OARs) is bound to take place. Often these changes are due to organ deformation such as variations in the rectum and bladder filling [5]. Continuing RT using the initial planning CT scan and the corresponding dose distribution despite these changes may lead to under-dosing of the target area or over-dosing the OARs.

Adaptive radiation therapy aims to adapt the treatment dose distribution to the daily anatomy, in order to achieve safe dose escalation [78] or smaller margins. This adaptation can be done by re-imaging, re-contouring, and re-planning at every session. However, manual re-contouring takes a lot of time and consequently new variations in the size and the shape of the OARs could be introduced meanwhile. Therefore, there is a need to re-contour the CT images accurately and in a time-efficient manner, i.e. (semi-)automatically.

Literature of prostate segmentation can be broadly categorized into two approaches, non-learning and learning approaches. For non-learning approaches, Qiao *et al.* used the open source `elastix` software to apply deformable image registration and subsequently propagate the contours from the planning to the daily scans [65]. For learning-based approaches, Elmahdy *et al.* proposed a hybrid learning and iterative approach, where they used a CNN network to segment the bladder and explicitly feed it to the registration model as prior knowledge on the underlying anatomy [46]. A substantial improvement was observed compared to the results reported in [65]. Recently, deep learning and specifically convolutional neural networks (CNNs) are being used to automatically segment OARs. Deep learning has the power to extract information from the data rather than depending on hand-crafted features. Milletari *et al.* developed an end-to-end CNN network to segment 3D medical images and reported a Dice Similarity Coefficient (DSC) of  $0.869 \pm 0.033$  on the PROMISE 2012 challenge dataset [49] (MR data). Tong *et al.* proposed the use of a shape representation model, to learn highly representative shape characteristics of OARs and help the final segmentation of a Fully Convolution Neural Network (FCNN) [79].

To the best of our knowledge, no deep learning based approach has been used to adapt a CNN model to a particular patient at a certain time point. In this study, we

propose to adapt a baseline CNN model as the patient goes through their RT treatment. Thus, instead of depending on a static deep learning model, we accumulate the knowledge over successive sessions for the same patient. This accumulated knowledge is then used to encourage the model into predicting a more personalized segmentation. Moreover, we study the performance of the fine-tuned network when more and more imaging becomes available over the course of treatment.

## 3.2 Materials and methods

### 3.2.1 Dataset

This study includes two different datasets from two different institutes and scanners, for patients who underwent intensity-modulated RT for prostate cancer. First, a dataset from Leiden University Medical Center (LUMC), Netherlands, has a total of 379 patients with one CT scan each. The scans were acquired using a Toshiba scanner, having 68 to 240 slices with a voxel size of approximately  $1.0 \times 1.0 \times 3.0$  mm. The second dataset is from Haukeland University Hospital, Norway, and includes 18 patients with 8-11 CT scans each corresponding to multiple fractions. These scans were acquired using a GE scanner, having 90 to 180 slices with a voxel size of approximately  $0.9 \times 0.9 \times 3.0$  mm. The target structures (prostate and seminal vesicles) as well as OARs (bladder and rectum) were manually delineated by oncologists. For the LUMC data informed consent was waived by the local Medical Ethical Committee, while for the Haukeland data informed consent was given by all included patients.

### 3.2.2 Baseline segmentation CNN model

In this study, for the baseline CNN model we adopted the network introduced in Elmahdy *et al.* [46], which has a straightforward architecture, but increased the number of the output labels from 2 to 5 for multi-organ segmentation. Similar to the standard U-net, it has an encoder and decoder path with four resolutions. Each encoder block consists of  $3 \times 3 \times 3$  convolutions with a stride of two in each dimension, followed by a rectified linear unit (ReLU). In the synthesis path, each block of the decoder consists of  $3 \times 3 \times 3$  convolutions with a stride of one in each dimension, followed by a ReLU and an upsampling of  $2 \times 2 \times 2$ . For a better localization, high resolution features from the contracting path are combined with the upsampled output, using skip connections. This aids to recover fine-grained details that were lost during the compression phase. The upsampling path is then followed by one Fully Connected (FC) layer, similar to the original paper, or three FC layers. These models are denoted by **base<sup>a</sup>** and **base<sup>b</sup>**, respectively. A softmax layer finalizes the models.

### 3.2.3 Patient-specific daily model adaptation

In order to adapt the model to the patient-specific anatomy, we update the model with the anatomical knowledge from previous treatment sessions for the same patient.

We thus personalize the pre-trained baseline CNN and utilize the available imaging of the patient. Since the new data (daily scans) as well as the task (segmentation) are the same as the base models, we re-use the base model weights as initialization to the model adaptation, as in a transfer learning approach. The weights in the FC layers are then fine-tuned using the imaging from previous fractions, while all other weights remain fixed. For each patient, at the first fraction we fine-tune the base model based on the planning scan. At subsequent fractions we start with the previous fine-tuned model and continue fine-tuning based on the scan and segmentation pair of the previous fraction. Here we consider that the segmentation can be manually corrected to clinical quality before the start of the current fraction. To be precise, let  $M_0$  be the base model and  $M_j$  the fine-tuned model at fraction  $j$ ,  $\Phi$  the model adaptation process,  $(I_j, S_j)$  the image and segmentation pair at fraction  $j$ . Model adaptation is then performed as follows:

$$M_j = \Phi(M_{j-1}; I_{j-1}, S_{j-1}). \quad (3.1)$$

The segmentation prediction of the current fraction is then given by  $M_j(I_j)$ . The base models are trained on the LUMC dataset. The networks are adapted to the Haukeland dataset, which has multiple follow-up scans per patient, mimicking an adaptive RT scenario.

### 3.3 Experiments and results

#### 3.3.1 Evaluation measures and implementation

The Dice Similarity Coefficient (DSC), Mean Surface Distance (MSD), and 95% Hausdorff Distance (HD) are used to evaluate the error between the ground truth delineations and the segmentations predicted by the models. These are evaluated for the prostate, seminal vesicles, bladder, and rectum. A Wilcoxon signed rank test at  $p = 0.05$  is used to assess statistical significance. Evaluation results are shown for the Haukeland data only, as only this dataset has follow-up scans.

We used the Tensorflow library for the implementation of the 3D CNNs. Out of the 379 LUMC patients, 70% (259 cases) were used for training the baseline models and the remaining 30% (111 cases) were used for validation. A total of 1000 patches of size  $128 \times 128 \times 128$  were extracted from each CT volume, sampling the classes with a uniform distribution to handle class imbalance. All scans were resampled to a fixed voxel size of  $1 \times 1 \times 2$  mm in order to handle variations in voxel size. The Dice Similarity Coefficient (DSC) is deployed as a cost function and the network was trained using the state-of-the-art Rectified Adam (RAdam) optimizer introduced in [80], with a fixed learning rate of  $10^{-4}$ . The network was trained for 1,000,000 iterations with a batch size of 4 using an NVIDIA Titan Xp GPU with 12 GB of memory. For model adaptation, a total of 2000 patches were extracted from a single CT volume with the

Table 3.1: DSC values of the target volumes and OARs.

|                       | #Itr | Prostate <sup>†‡</sup> | SV <sup>†‡</sup>   | Rectum <sup>†‡</sup> | Bladder <sup>†‡</sup> |
|-----------------------|------|------------------------|--------------------|----------------------|-----------------------|
|                       |      | $\mu \pm \sigma$       | $\mu \pm \sigma$   | $\mu \pm \sigma$     | $\mu \pm \sigma$      |
| base <sup>a</sup>     |      | 0.76 ± 0.06            | 0.55 ± 0.19        | 0.79 ± 0.08          | 0.86 ± 0.14           |
| proposed <sup>a</sup> | 500  | 0.78 ± 0.06            | 0.61 ± 0.17        | 0.79 ± 0.07          | 0.89 ± 0.09           |
| proposed <sup>a</sup> | 2000 | 0.80 ± 0.05            | 0.63 ± 0.17        | 0.80 ± 0.07          | 0.90 ± 0.07           |
| proposed <sup>a</sup> | 5000 | 0.82 ± 0.05            | <b>0.65 ± 0.16</b> | 0.80 ± 0.07          | <b>0.91 ± 0.07</b>    |
| base <sup>b</sup>     |      | 0.77 ± 0.07            | 0.46 ± 0.20        | 0.78 ± 0.08          | 0.87 ± 0.08           |
| proposed <sup>b</sup> | 500  | 0.80 ± 0.05            | 0.50 ± 0.18        | 0.81 ± 0.06          | <b>0.91 ± 0.07</b>    |
| proposed <sup>b</sup> | 2000 | <b>0.83 ± 0.04</b>     | 0.56 ± 0.15        | <b>0.83 ± 0.05</b>   | <b>0.91 ± 0.06</b>    |
| proposed <sup>b</sup> | 5000 | <b>0.83 ± 0.04</b>     | 0.55 ± 0.17        | <b>0.83 ± 0.05</b>   | <b>0.91 ± 0.06</b>    |

Table 3.2: MSD values of the target volumes and OARs.

|                       | #Itr | Prostate <sup>†‡</sup> | SV <sup>†‡</sup>   | Rectum <sup>†‡</sup> | Bladder <sup>†‡</sup> |
|-----------------------|------|------------------------|--------------------|----------------------|-----------------------|
|                       |      | $\mu \pm \sigma$       | $\mu \pm \sigma$   | $\mu \pm \sigma$     | $\mu \pm \sigma$      |
| base <sup>a</sup>     |      | 2.70 ± 0.88            | 3.78 ± 4.76        | 2.77 ± 1.23          | 2.66 ± 9.46           |
| proposed <sup>a</sup> | 500  | 2.38 ± 0.64            | 2.58 ± 2.67        | 2.76 ± 1.05          | 1.62 ± 1.28           |
| proposed <sup>a</sup> | 2000 | 2.03 ± 0.55            | <b>2.38 ± 2.76</b> | 2.66 ± 1.03          | 1.48 ± 1.17           |
| proposed <sup>a</sup> | 5000 | 1.78 ± 0.52            | 2.41 ± 3.17        | 2.59 ± 1.00          | 1.39 ± 1.12           |
| base <sup>a</sup>     |      | 2.29 ± 0.70            | 5.08 ± 5.33        | 2.78 ± 1.17          | 1.83 ± 0.96           |
| proposed <sup>b</sup> | 500  | 1.95 ± 0.53            | 5.21 ± 16.71       | 2.46 ± 1.05          | 1.40 ± 1.03           |
| proposed <sup>b</sup> | 2000 | 1.64 ± 0.43            | 3.18 ± 3.34        | <b>2.30 ± 0.96</b>   | <b>1.24 ± 0.89</b>    |
| proposed <sup>b</sup> | 5000 | <b>1.61 ± 0.41</b>     | 4.67 ± 16.96       | <b>2.30 ± 0.96</b>   | 1.25 ± 0.91           |

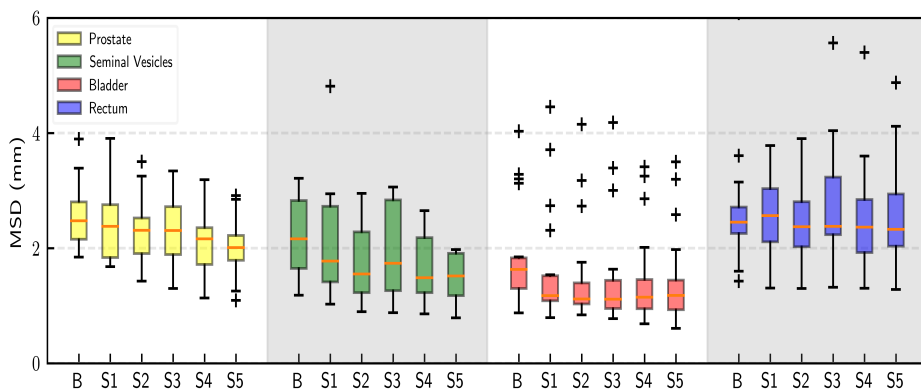
same patch size, cost function, and learning rate as the base model. For investigating the effect of the number of adaptation iterations on the network performance, we varied this parameter between 500, 2000, and 5000 iterations. This adaptation can be performed offline, and took less than an hour per fraction at 2000 iterations.

### 3.3.2 Results

Tables 3.1, 3.2 and 3.3 show the quantitative results of all the models, averaged over all the treatment fractions. Adapted models are denoted by proposed<sup>a</sup> and proposed<sup>b</sup>. Here, † and ‡ represent a statistically significant difference between base<sup>a</sup> vs proposed<sup>a</sup>, and base<sup>b</sup> vs proposed<sup>b</sup> at 2000 iterations, respectively. In terms of DSC, MSD, and 95% HD, the larger base<sup>b</sup> model is slightly better than base<sup>a</sup>, except for the seminal vesicles. This pattern is also visible for the proposed models. All proposed models outperform the baseline models, sometimes by a margin. Furthermore, adapting the

Table 3.3: 95% HD value of the target volumes and OARs.

|                       | #Itr | Prostate <sup>†‡</sup>          | SV <sup>†‡</sup>                | Rectum                           | Bladder <sup>†‡</sup>           |
|-----------------------|------|---------------------------------|---------------------------------|----------------------------------|---------------------------------|
|                       |      | $\mu \pm \sigma$                | $\mu \pm \sigma$                | $\mu \pm \sigma$                 | $\mu \pm \sigma$                |
| base <sup>a</sup>     |      | $9.1 \pm 3.7$                   | $12.2 \pm 10.7$                 | $14.4 \pm 7.0$                   | $8.7 \pm 12.9$                  |
| proposed <sup>a</sup> | 500  | $7.9 \pm 2.8$                   | $9.2 \pm 8.4$                   | $14.7 \pm 6.7$                   | $7.9 \pm 6.7$                   |
| proposed <sup>a</sup> | 2000 | $6.4 \pm 2.1$                   | <b><math>8.6 \pm 9.0</math></b> | $14.3 \pm 6.7$                   | $7.2 \pm 6.5$                   |
| proposed <sup>a</sup> | 5000 | $5.9 \pm 2.4$                   | $8.8 \pm 10.6$                  | $14.0 \pm 6.6$                   | $6.5 \pm 6.1$                   |
| base <sup>b</sup>     |      | $6.9 \pm 1.9$                   | $15.5 \pm 9.7$                  | $13.9 \pm 9.5$                   | $7.0 \pm 4.5$                   |
| proposed <sup>b</sup> | 500  | $6.1 \pm 1.7$                   | $14.6 \pm 20.4$                 | $13.5 \pm 9.1$                   | $6.2 \pm 5.8$                   |
| proposed <sup>b</sup> | 2000 | <b><math>5.2 \pm 1.4</math></b> | $11.9 \pm 9.3$                  | $13.5 \pm 8.4$                   | <b><math>5.6 \pm 5.3</math></b> |
| proposed <sup>b</sup> | 5000 | <b><math>5.2 \pm 1.4</math></b> | $13.9 \pm 20.9$                 | <b><math>13.3 \pm 8.5</math></b> | $5.7 \pm 5.6$                   |

Figure 3.1: The effect of model adaptation over sessions in terms of MSD (mm). B is base model,  $S_i$  after session  $i$ .

network for 5000 iterations seems as good as 2000 iterations as shown in the tables. Predicting a segmentation took  $\sim 0.6$  seconds.

To investigate the relation between the amount of patient-specific data available for fine-tuning and segmentation performance, we applied the models  $M_j$ ,  $j \in \{0, 5\}$  on scan  $I_6$  from each patient. Boxplots of the results are shown in Figure 3.1 for the MSD. Segmentation of most structures tend to improve when more data is available, while such a trend is not visible for the rectum. Some example segmentation results are given in Figure 3.2 for the models base<sup>b</sup> and proposed<sup>b</sup> at 2000 iterations.



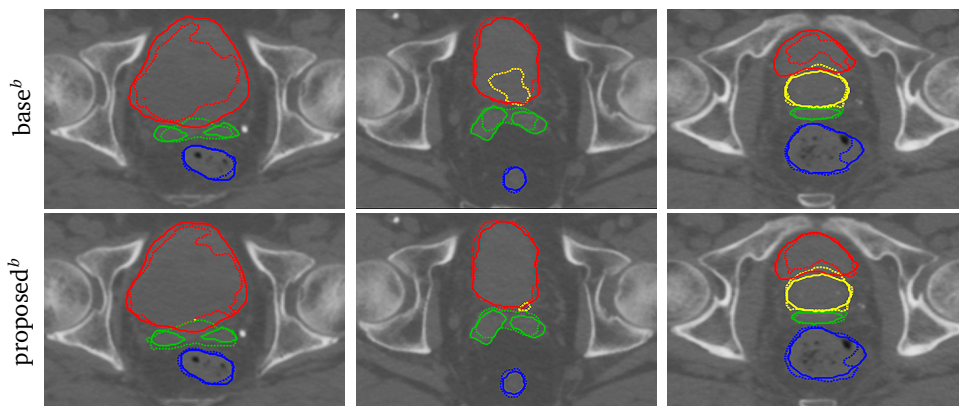


Figure 3.2: Three example results for the base<sup>b</sup> model (top row) and the proposed<sup>b</sup> at 2000 itr model (bottom row). The red, yellow, blue and green contours represent the bladder, prostate, rectum, and seminal vesicles, respectively. The solid line indicates manual contouring, while the dotted line the automatic prediction.

### 3.4 Discussion and conclusion

In this study we investigated the hypothesis of personalizing the automatic contouring by deep learning, of the target organs and OARs for adaptive radiotherapy of prostate cancer. Unlike traditional CNN networks, the proposed model adaptation strategy predicts the segmentation of the daily CT based on the accumulated anatomical knowledge from the previous scans of the same patient. We demonstrate that adapting the model to a specific patient anatomy boosted the performance of the network. Furthermore, increasing the network capacity by adding more fully connected layers was beneficial. Moreover, from Fig. 3.1 we observe that the rectum delineation did not benefit from fine-tuning, which may be explained by the normal variation in rectum filling over treatment fractions. Such variations in the rectum and seminal vesicles could be further investigated in future work by more sophisticated networks and model generalization methods.

To conclude, we proposed an adaptive training mechanism for personalized automatic contour segmentation for prostate cancer. This adaptation showed potential for improving the prediction of the daily anatomy based on personalized imaging accumulated over fractions. Since the segmentation time is less than a second, the adaptation mechanism is therefore very promising for clinical implementation in the context of adaptive radiotherapy of prostate cancer.

**Acknowledgements.** This study was financially supported by Varian Medical Systems and ZonMw, grant number 104003012. The Haukeland dataset was provided by oncologist Svein Inge and physicist Liv Bolstad; they are gratefully acknowledged.