Deep learning for online adaptive radiotherapy

Elmahdy, M.S.E.

2

Robust Contour Propagation Using Deep Learning and Image Registration for Online Adaptive Proton Therapy of Prostate Cancer

This chapter was adapted from:

M Elmahdy, T Jagt, Y Qiao, R Shahzad, H Sokooti, S Yousefi, L Incrocci, C Marijnen, M Hoogeman, and M Staring. **Robust contour propagation using deep learning and image registration for online adaptive proton therapy of prostate cancer**, *Medical Physics*, Pages 3329-3343, 2019.

Abstract

Purpose: To develop and validate a robust and accurate registration pipeline for automatic contour propagation for online adaptive Intensity-Modulated Proton Therapy (IMPT) of prostate cancer using `elastix` software and deep learning.

Methods: A 3D Convolutional Neural Network was trained for automatic bladder segmentation of the CT scans. The automatic bladder segmentation alongside the CT scan are jointly optimized to add explicit knowledge about the underlying anatomy to the registration algorithm. We included three datasets from different institutes and CT manufacturers. The first was used for training and testing the ConvNet, where the second and the third were used for evaluation of the proposed pipeline. The system performance was quantified geometrically using the Dice Similarity Coefficient (DSC), the Mean Surface Distance (MSD), and the 95% Hausdorff Distance (HD). The propagated contours were validated clinically through generating the associated IMPT plans and compare it with the IMPT plans based on the manual delineations. Propagated contours were considered clinically acceptable if their treatment plans met the dosimetric coverage constraints on the manual contours.

Results: The bladder segmentation network achieved a DSC of 88% and 82% on the test datasets. The proposed registration pipeline achieved a MSD of 1.29 ± 0.39, 1.48 ± 1.16, and 1.49 ± 0.44 mm for the prostate, seminal vesicles, and lymph nodes, respectively on the second dataset and a MSD of 2.31 ± 1.92 and 1.76 ± 1.39 mm for the prostate and seminal vesicles on the third dataset. The automatically propagated contours met the dose coverage constraints in 86%, 91%, and 99% of the cases for the prostate, seminal vesicles, and lymph nodes, respectively. A Conservative Success Rate (CSR) of 80% was obtained, compared to 65% when only using intensity-based registration.

Conclusion: The proposed registration pipeline obtained highly promising results for generating treatment plans adapted to the daily anatomy. With 80% of the automatically generated treatment plans directly usable without manual correction, a substantial improvement in system robustness was reached compared to a previous approach. The proposed method therefore facilitates more precise proton therapy of prostate cancer, potentially leading to fewer treatment related adverse side effects.

2.1 Introduction

Prostate cancer is one of the leading causes of mortality and the most common cancer among men. The National Cancer Society (NCS) estimates around 164,690 new cases and 24,430 deaths from prostate cancer in the United States only for 2018 [36]. Due to its slow progress, individuals could develop prostate cancer for many years without explicit signs. There are treatment options for prostate cancer including surgical removal of the prostate, hormone therapy, and radiotherapy. Intensity-Modulated Proton Therapy (IMPT) is able to deliver a highly localized dose distribution to the target volume, while minimizing collateral damage to the surrounding healthy tissues [37]. IMPT is however more sensitive to daily changes than photon therapy, which may result in distortion of the delivered dose distribution [4, 6]. These changes could arise from anatomical variations in the shape and position of both target volumes and Organs-At-Risk (OARs) or a misalignment in the patient setup. In order to compensate for these changes, a margin is added to the Clinical Target Volume (CTV) to generate the Planning Target Volume (PTV) in addition to robust treatment planning. These margins result in extra dose to the OARs, leading to an increase in the treatment-related toxicities that may prevent dose escalation. Traditionally, motion-induced variations are minimized by implanting fiducial markers in the prostate, subsequently compensating for the daily prostate motion using online imaging [38]. However, such correction strategies are invasive and only capable of correcting for translational motion and limited amount of rotational motion [39]. Online imaging and re-planning should be able to handle this problem without using fiducial markers [40]. These online CT scans have to be delineated first in order to update the treatment plan. Usually this task is done by radiation oncologists according to certain guidelines [7, 8]. However, intra and inter-observer inconsistency has been noted due to different preferences and experience among radiation oncologists [9, 10]. Typically, daily manual re-contouring is not performed because it is time consuming and new anatomical variations may be introduced in the time it takes to delineate the scan [11]. Automatic re-contouring algorithms can alleviate these issues, but robust methods are required, because otherwise still time consuming fallback strategies are needed.

Automatic re-contouring could be accomplished effectively using Deformable Image Registration (DIR) by deducing the correspondence between the daily CT and the planning CT. Using the generated Deformation Vector Field (DVF), manual contours can be propagated from the planning CT to the daily CT. The automatically generated contours together with fast re-optimization of the treatment plan [41] could compensate for the daily variation and ensure the delivery of the prescribed dose distribution at small margins and robust settings. DIR is a crucial step towards developing online adaptive IMPT alongside re-planning and personalized dose Quality

Assurance (QA). Currently, these steps are time consuming, thus severely limiting online procedures.

There are commercially available applications for automatic re-contouring such as Atlas Based Auto Segmentation (ABAS), Mirada, and RayStation. These applications are, however, considered a black box for the end-users and therefore limit the parameter choices and tuning. Open source DIR packages provide a high level of flexibility with a concrete scientific evidence and reproducibility. Qiao *et al.* [42] reported an MSD of 1.36 ± 0.30 mm, 1.75 ± 0.84 mm, 1.49 ± 0.44 mm for the prostate, seminal vesicles, and lymph nodes, respectively for 18 patients using the open source `elastix` software. A clinical success rate of 69% was achieved, which means that 31% of the delineations have to be corrected, leading to increased costs and a suboptimal patient workflow. In 2011, Thor *et al.* deployed DIR to propagate the contours of the prostate and OARs from CT to cone-beam CT [43]. The system achieved a mean DSC of 0.80 for the prostate, 0.77 for the rectum, and 0.73 for the bladder with a relatively high variance. Moreover, the system was not qualitatively evaluated in terms of dosimetric coverage. Recently, Woerner *et al.* [44] investigated the error between different radiologists and both DIR and rigid registration in different body regions. They only reported the results for the prostate, which were 0.90, 0.99 mm, and 8.12 mm for the DSC, MSD, and Hausdorff Distance (HD), respectively. Thörnqvist *et al.* [45] used two different demons-based registration algorithms, with one more conservative than the other. They achieved an average DSC of 0.88, 0.85, 0.89, 0.78 for the lymph nodes, prostate, bladder, and rectum, respectively.

In spite of the existence of quite accurate registration algorithms, they still suffer from a lack of robustness, which is a critical aspect for clinical application. Therefore, in this paper we focus on the robustness aspect of the registration pipeline. The main challenges in Qiao *et al.* were the presence of gas pockets and large deformations surrounding the seminal vesicles, bladder, and rectum. Hence, we propose to tackle these challenges by inpainting the rectum gas pockets as well as embedding the bladder segmentation in the registration pipeline using deep learning to enhance the system's robustness. The proposed registration pipeline was evaluated geometrically and dosimetrically for generating clinically acceptable IMPT plans. Compared to our conference paper [46], we made several improvements, such as the inclusion of more datasets, dealing with gas pockets, data normalization, and multi-stage registration. Moreover, we carried out an extensive dosimetric validation for the automatically generated contours to verify its clinical viability.

2.2 Methods

The prostate and seminal vesicles are positioned between the bladder and the rectum, therefore prostate motion is mainly influenced by the filling and motion of both the

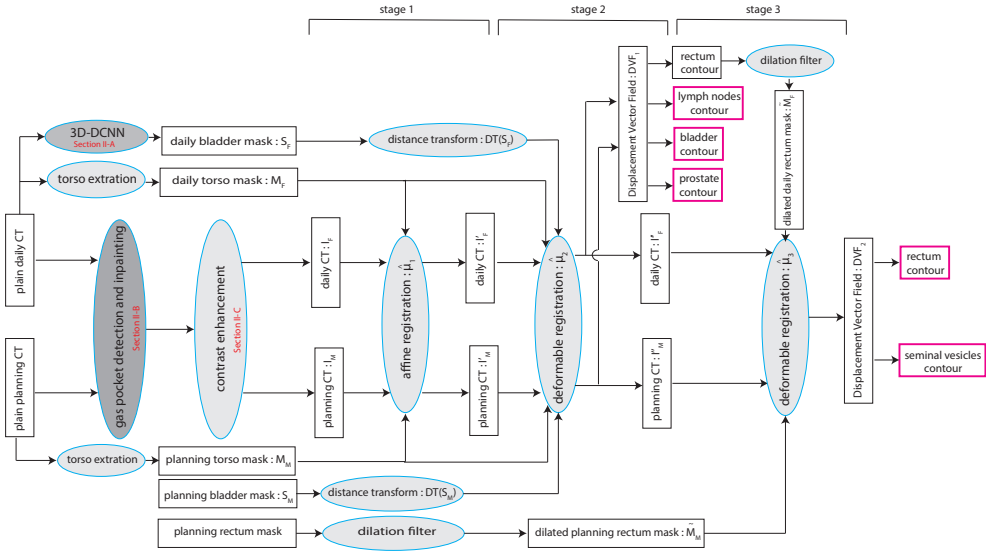


Figure 2.1: The proposed multi-stage registration process using `elastix` software and deep learning. The red boxes denote the contours finally used as output of the algorithm.

bladder and the rectum [47]. Hence, we hypothesize that embedding an explicit prior knowledge about the deformation of either organs to the intensity-based DIR method may improve the accuracy and robustness of the registration. Here, we considered the bladder because it has a well-defined shape that could be more easily delineated in a fully automatic manner than the rectum. Since the registration is intensity-based, the quality of the registration process is correlated to the quality of the input images. Hence, we introduced multiple data preprocessing steps to enhance the quality of the input images. These steps include rectum gas pocket detection and inpainting and contrast clipping as shown in Figure 2.1.

2.2.1 Bladder segmentation using deep learning

In this study, we automatically segment the bladder using a 3D U-net Convolutional Neural Network (3D-CNN) similar to the architecture introduced in [48]. The network consists of encoding and decoding branches connected with skip connections as shown in Figure 2.2. In order to represent the volumetric information and tissue homogeneity of the CT volume, 3D convolution layers followed by non-linear leaky rectified linear units were used. The original maxpooling layers were replaced by strided convolution in both encoder and decoder branches. Negative Dice Similarity Coefficient (DSC) [49] is deployed as a cost function and the network is trained using the Adam optimizer [50] with a fixed learning rate of 10^{-4}. The network has 64,320 trainable parameters

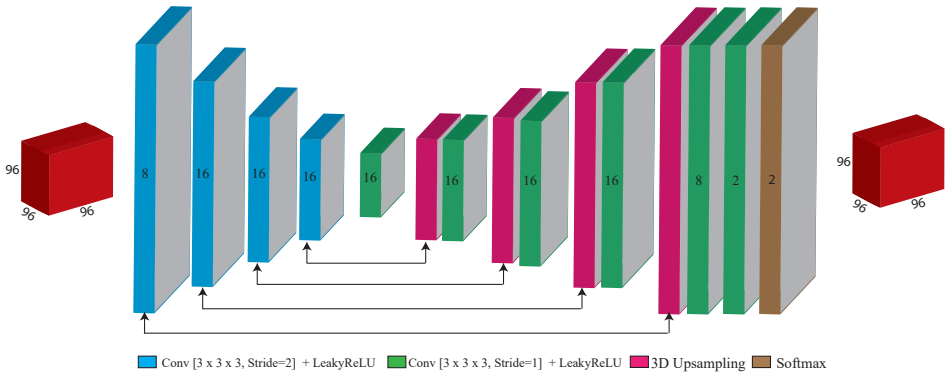


Figure 2.2: The architecture for the 3D-CNN network, where the numbers on the blocks denote the number of feature maps.

which enables network inference of the entire CT image in approximately 2 seconds. The network was designed to output the same size as input, however the input size should be divisible by 16. Largest connected component analysis was applied as a post-processing step to eliminate irrelevant activations.

2.2.2 Gas pocket detection and inpainting

A problem that usually arises for intensity-based DIR of the pelvic region is the presence of gas pockets in the bowel and rectum. These pockets appear as dark areas surrounded by soft tissue. Usually the size and position of these pockets are not the same in the planning and the daily CT. In such situations, physical correspondence between images at different sessions does not exist because of the insertion or occlusion of image content. Only few studies addressed this issue in the literature. Gao *et al.* [51] proposed introducing a virtual gas pocket to the planning CT that follows the pocket in the daily CT. They tested it on 15 prostate cancer patients with distended rectum. Foskey *et al.* [52] proposed to deflate the pocket to a virtual point. In both papers, the authors assumed no gas pockets in the planning CT, which is not usually the case. Recently, deep learning based algorithms have revolutionized the medical image analysis field [53]. One category of deep learning architectures is Generative Adversarial Networks (GANs) introduced by Goodfellow *et al.* [54] in 2014. GANs have been growing since then in generating realistic natural and synthetic images. As for medical images, GANs have been used in image segmentation [55], synthesis [56], registration [57], and denoising [58]. Recently Yu *et al.* [59] proposed a 2D GAN network with a contextual attention model to restore and inpaint occluded regions in natural images. The network also blends the restored region with the surrounding texture to make it look more realistic. The proposed model has two successive networks

(a) (b) (c)

Figure 2.3: Different inpainting algorithms, where (a), (b), and (c) represent the original CT, the result from simple-inpainting, and the result from GAN-inpainting, respectively.

for image generation in order to generate patches with fine quality. The first 'generator' network generates a coarse result through a dilated convolution network. This result is then fed to the second network. The second 'discriminator' network has two routes, one goes to a dilated convolution network while the other goes through a contextual attention model. Finally, the results from these two routes are concatenated and fed to a prediction network. This network has shown an improvement over a similar network proposed by Iizuka *et al.* [60]. In this paper, we retrained this network so that it can inpaint (fill) gas pockets of different shapes and sizes with a more sophisticated and realistic content rather than a fixed value. The same implementation and hyper parameters were used as in the original paper.

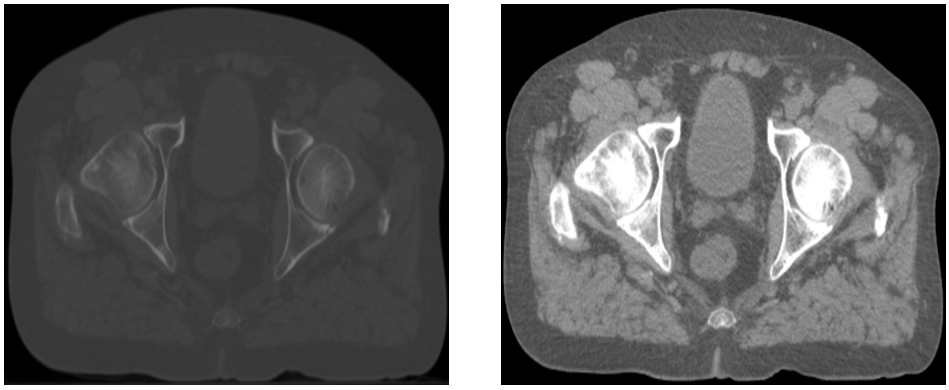
Alternatively, we also experimented with a simplified method for inpainting. Following the idea proposed by Rodriguez-Vila *et al.* [61] we fill the gas pockets with a fixed value and smooth the output to blend it with the surrounding tissues. A threshold of -200 is used to generate a binary mask of the gas pockets. This mask is then dilated with a kernel of size 7x7x1 voxels (M) while the CT image is filled with a fixed HU number of 60 (the average HU number for faeces), and smoothed with a sigma of 4mm ($I_{smoothed}$). Equation (2.1) shows the simple inpainting process:

$$I_{out} = I_{input} \times (1 - M) + I_{smoothed} \times M \tag{2.1}$$

Figure 2.3 shows a comparison between gas pocket inpainting using the GAN network and simple inpainting.

2.2.3 Contrast enhancement

To enhance the soft tissue contrast, the CT intensity was clipped to the range of $[-300, 300]$. This clipping is similar to viewing the soft tissue with an appropriate window level. Moreover, such enhancement improves the registration convergence. Figure 2.4 shows the effect of intensity clipping.



(a) (b)

Figure 2.4: The effect of contrast clipping, where (a) and (b) represent the image before and after intensity clipping, respectively.

2.2.4 Image registration

For carrying out the DIR experiments, we used the open software package `elastix` [62]. For more details, see the website `http://elastix.isi.uu.nl`. All the experiments were performed on a cluster of workstations operated on the Oracle Grid Engine (OGE), which has 500 nodes with a total of 800 cores. Testing time is reported using a PC with 16 GB memory, Windows 7 Professional 64 bit operation system and an Intel Xeon E51620 CPU with 4 cores at 3.6 GHz, utilizing only the CPU.

In this study, the planning CT scan (moving image) was aligned with the daily CT scan (fixed image) of each patient. The registrations were initialized based on the center-of-gravity of the bony anatomy defined by a Hounsfield number larger than 200. A mask of the body torso was generated using Pulmo software [63] to remove the effect of the CT table. The registration process is done in three stages. First, the moving and fixed images are registered using a single resolution affine transformation using 200 iterations as defined in Eq. (2.2):

$$\hat{\mu_1} = \arg\min_{\mu} C_1\left(I_F, I_M, M_F, M_M; T_{\mu_1}\right), \tag{2.2}$$

where I_F is the daily scan, I_M is the planning scan, M_F is the torso mask of the daily scan, M_M is the torso mask of the planning scan, and C_1 is the mutual information cost function. The affine transformation aligns the bones and large structures. Second, a deformable registration is applied to tackle the local deformations of the organs. In this stage, the planning CT of each patient combined with the manual delineation of the bladder are considered the moving images, while the repeat CT of the same patient accompanied with the bladder segmentation resulting from the proposed 3D-CNN are

the fixed images. Equation (2.3) defines the optimization problem for this stage:

$$\hat{\mu_2} = \arg\min_{\mu} \{C_1\left(I_F, I_M, M_F, M_M, T_{\mu_1}; T_{\mu_2}\right) + \alpha\, C_2\left(DT\left(S_F\right), DT\left(S_M\right), T_{\mu_1}; T_{\mu_2}\right)\}, \quad (2.3)$$

where C_2 is the Mean Squared Difference (MSD) cost function, α is a weight for balancing these two cost functions, $DT(S_F)$ is the distance transform of the 3D-CNN bladder segmentation, and $DT(S_M)$ is the distance transform of the manual annotation of the planning scan. The Distance Transform (DT) of the bladder segmentations is used instead of the binary segmentations themselves, to ensure a smooth and stable optimization process. The generated Deformation Vector Field (DVF) from this step is then used to propagate the contours of the prostate, lymph nodes, bladder, and rectum from the planning CT to the repeat CT. Because the seminal vesicle is a small irregular structure, which is highly affected by the deformation in the rectum, we introduce a third stage to focus the registration on the rectum and seminal vesicle region. In this stage, the rectum contour of the planning CT and the rectum contour of the daily CT (from the previous stage) are dilated with a kernel of 45x45x1 voxels and used as a registration mask together with the fixed and moving CT scans. The contours of the rectum and seminal vesicles are then propagated using the generated DVF from the final stage. Equation (2.4) defines the optimization problem for this stage:

$$\hat{\mu_3} = \arg\min_{\mu} C_1\left(I_F, I_M, \tilde{M}_F, \tilde{M}_M, T_{\mu_1}, T_{\mu_2}; T_{\mu_3}\right), \quad (2.4)$$

where $\tilde{M}_M$ is the dilated rectum mask of the planning CT and $\tilde{M}_F$ is the dilated rectum mask of the daily CT. A fast recursive implementation of the B-spline transformation was employed for DIR [64] in stage 2 and 3. Adaptive stochastic gradient descent was used for optimization [65] in all three stages. For the DIR stage we used a three level Gaussian pyramid with smoothing factors of 4, 2, and 1 mm. Figure 2.1 illustrates the proposed registration pipeline in detail.

2.3 Experiments and results

2.3.1 Dataset

This study includes three datasets representing three different institutes and CT scanners from three different vendors for patients who underwent intensity-modulated radiation therapy for prostate cancer. Table 2.1 shows detailed information about these datasets. The LUMC dataset was used to train and validate the neural network for segmenting the bladder (Section 2.2.1) as well as the inpainting network (Section 2.2.2), while the EMC and HMC dataset were used as independent test sets for the complete registration pipeline. Geometric evaluation was performed on both the EMC and HMC dataset. Eleven out of the eighteen HMC patients were considered for dosimetric evaluation due to the availability of not only the manual delineations

Table 2.1: Details of the datasets reported in this study. LUMC, EMC, and HMC are abbreviations for Leiden University Medical Center (Netherlands), Erasmus Medical Center (Netherlands), and Haukeland Medical Center (Norway), respectively. SV and LN denote Seminal Vesicles, and Lymph Nodes, respectively.

Institute	Scanner	#Patients	#Scans/ patient	Image size	Voxel spacing (mm)	Manual delineations
LUMC	Toshiba	418	1	512x512x(68-240)	~1.0x1.0x3.0	bladder, rectum
EMC [66]	Siemens	14	4	512x512x(91-218)	~0.9x0.9x1.5	prostate, SV bladder, rectum
HMC [67]	GE	18	8-11	512x512x(90-180)	~0.9x0.9x2.0	prostate, SV, LN bladder, rectum

for the target organs (prostate, seminal vesicles, lymph nodes) and OARs (bladder, rectum), but moreover the manual delineations of the bowels and femoral heads needed for planning.

2.3.2 Evaluation measures

The quality of the registration is quantified in terms of geometric aspects and dosimetric coverage. The geometric quality is measured by comparing the manual contours and the automatically propagated contours of the daily CT for the prostate, lymph nodes, seminal vesicles, rectum, and bladder. The Dice Similarity Coefficient (DSC) measures the overlap between the segmentations, while the Mean Surface Distance (MSD), and the 95% Hausdorff Distance (HD) measure the residual distance between the contours in 3D space.

$$DSC = \sum \frac{2\,|\,F \cap M\,|}{|\,F\,| + |\,M\,|}, \tag{2.5}$$

where F and M are the propagated contour and the ground truth contour, respectively.

$$MSD = \frac{1}{2}\left(\frac{1}{n}\sum_{i=1}^{n} d\,(a_i, M) + \frac{1}{m}\sum_{i=1}^{m} d\,(b_i, F)\right), \tag{2.6}$$

$$HD = \max\left\{\max_i\{d\,(a_i, M)\}, \max_j\{d\,(b_i, F)\}\right\}, \tag{2.7}$$

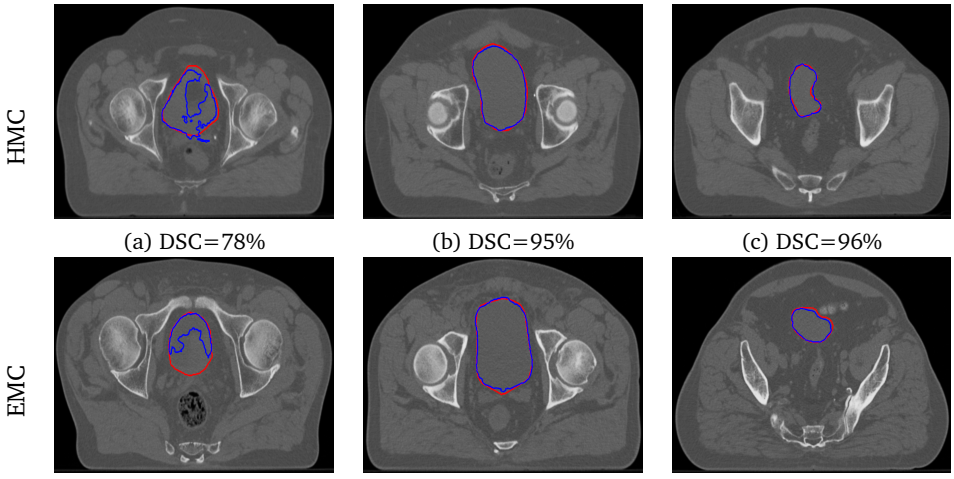
where $\{a_1, a_2, ..., a_n\}$ and $\{b_1, b_2, ..., b_m\}$ are the surface mesh points of the fixed and moving contours, respectively and $d\,(a_i, M) = \min_j \,||\, b_j - a_i \,||$. The geometrical success rate, as a marker for geometric robustness, is defined as the percentage of registrations with MSD < 2 mm (slice thickness): $\gamma = \frac{n}{N}\{\text{MSD} < 2\,mm\}$, where (N) is the total number of registrations performed.

IMPT plans were generated for 11 patients from the HMC dataset using both the manual and the automatic delineations. The plans were then evaluated on the manual delineations to investigate the clinical effect of the error between these two delineations. Erasmus-iCycle, an in-house developed treatment planning optimization

system, [68, 69, 70, 71, 72] together with the Astroid dose engine were used to generate the IMPT plans. Erasmus-iCycle uses a multi-criteria optimization to generate a clinically desirable Pareto optimal treatment plan on the basis of a wish list consisting of hard constraints and objectives. A small margin of 2 mm around the prostate and 3.5 mm around the lymph nodes and seminal vesicles is used to compensate for the marginal error of the propagated contours and to account for intra-observer variations in the manual contouring. These margins alone can not account for variations in shape and location of the target volumes. Dose was prescribed according to a simultaneously integrated boost scheme in which the high-dose PTV (prostate + 2 mm margin) was assigned 74 Gy and the low-dose PTV (seminal vesicles and lymph nodes + 3.5 mm margin) 55 Gy, to be delivered using two laterally opposed beams. In order to avoid under-dose, the optimization ensures that at least 98% of the target volumes receive at least 95% of the prescribed dose ($V_{95\%} \geq 98\%$). To avoid overdose the optimization ensures that less than 2% of the target volumes receive more than 107% of the highest prescribed dose ($V_{107\%} \leq 2\%$). To achieve a clinically acceptable result, automatically generated treatment plans from the propagated contours should still fulfill these goals. Hence, IMPT plans from the propagated contours are evaluated based on the manual contours. The clinical success rate, as a marker for geometric robustness, is defined as the percentage of registrations for which the prostate directly meets the dose treatment criteria: $\eta = \frac{n}{N}\{V_{95\%} \geq 98\%\}$. Conservative Success Rate (CSR) is a more conservative measure of clinical success when all target volumes (the prostate, seminal vesicles and lymph nodes) meet this dosimetric criterion. For dosimetric coverage calculation $N = 99$.

2.3.3 Network training and performance

We implemented the 3D-CNN and GAN-inpainting networks using Tensorflow [73]. For training these networks, we used the LUMC dataset. This dataset was a sufficiently large dataset to be able to train the neural networks. Since the LUMC dataset only had one CT scan per patient, it was not used for registration evaluation. From the 418 LUMC patients, 350 patients were used for network training, and 68 patients for validation. The trained network was then applied without modification to the CT scans in the EMC and HMC datasets. In order to account for the variations in voxel size between datasets and scans, all scans were resampled to a fixed voxel size of 1.0×1.0×2.0 mm. For the 3D-CNN, 100,000 patches of size 96×96×96 voxels were randomly extracted from the training volumes, making sure they are equally distributed between foreground and background. For the GAN-inpainting network, all the slices with gas pockets were eliminated from training. Moreover, all slices were resampled to a pixel size of 1.0 × 1.0 mm and centrally cropped to 256x256 pixels so that more patches could fit into memory as well as it would be beneficial for the



(a) DSC=78% (b) DSC=95% (c) DSC=96%

(d) DSC=71% (e) DSC=93% (f) DSC=91%

Figure 2.5: Examples of the automatic bladder segmentation using the 3D-CNN alongside the DSC of the volume. First and second rows represent samples from HMC and EMC, respectively. (a) and (d) are suboptimal results and the rest are good results. The red line represents the ground truth and the blue line is the network output.

network to learn the most relevant contextual information to the rectum. Randomly selected windows of size 64x64 pixels were occluded in order to train the network to inpaint these regions with a realistic content. Both the 3D-CNN and the 2D-GAN-inpainting network were trained for 100,000 iterations on the raw CT patches without any preprocessing except for resampling. All the experiments were carried out using an NVIDIA GTX1080 Ti with 11 GB of GPU memory. The 3D-CNN bladder segmentation network obtained a DSC of $85.4\% \pm 1.4\%$ on the validation scans. Moreover, the network was tested on the EMC and HMC datasets and achieved an average DSC of $82.3\% \pm 1.5\%$ and $87.9\% \pm 1.2\%$, respectively. Using a single GPU, the average inference time of the segmentation and inpainting networks were approximately 2 seconds and 3 seconds per volume depending on the number of slices per volume. Figure 2.5 shows examples of the network output.

2.3.4 Parameter optimization and preprocessing analysis

For a fair comparison, the same registration parameters as in [42] were used. For the weight α that balances the contribution of the bladder segmentation in the cost function (2.3), we investigated multiple settings based on initial experiments on EMC and HMC datasets. The weight was set for the coarse (first) resolution only and was set to zero for the other two resolutions, in order to avoid overfitting issues. Here we compared four settings for α: 0.2, 0.1, 0.05, and 0.01. For this experiment we did not use inpainting. The results are shown in Table 2.2 for the HMC dataset where

Table 2.2: MSD (mm) of the target volumes and OARs of the HMC dataset for different registration and weight settings after the third stage of registration. Registrations using 100 and 500 iterations were both tested.

		Prostate	Seminal vesicles	Lymph nodes	Rectum	Bladder
Method	α	$\mu \pm \sigma$	$\mu \pm \sigma$	$\mu \pm \sigma$	$\mu \pm \sigma$	$\mu \pm \sigma$
Affine, 200		1.63 ± 0.74	2.92 ± 1.74	1.23 ± 0.49	3.89 ± 1.62	4.37 ± 2.11
B-spline, 100	0.20	1.55 ± 0.90	1.70 ± 0.74	1.63 ± 0.58	2.70 ± 1.12	1.85 ± 1.85
	0.10	1.53 ± 0.82	1.72 ± 0.73	1.58 ± 0.50	2.72 ± 1.11	1.85 ± 1.71
	0.05	1.50 ± 0.75	1.74 ± 0.79	1.55 ± 0.46	2.75 ± 1.16	1.86 ± 1.56
	0.01	1.41 ± 0.36	1.75 ± 0.86	1.57 ± 0.38	2.76 ± 1.15	1.98 ± 1.19
B-spline, 500	0.20	1.49 ± 0.90	1.76 ± 0.80	1.65 ± 0.64	2.87 ± 1.39	1.74 ± 1.63
	0.10	1.45 ± 0.77	1.77 ± 0.93	1.59 ± 0.52	2.78 ± 1.19	1.77 ± 1.58
	0.05	1.43 ± 0.77	1.78 ± 0.90	1.55 ± 0.47	2.79 ± 1.19	1.81 ± 1.57
	0.01	1.36 ± 0.47	1.76 ± 0.82	1.56 ± 0.48	2.81 ± 1.18	1.84 ± 1.24

Table 2.3: MSD (mm) of the target volumes and OARs for different registration settings and inpainting methods at $\alpha = 0.05$. Registrations using 100 and 500 iterations were both tested.

		Prostate	Seminal vesicles	Lymph nodes	Rectum	Bladder
# It.	Inpainting Method	$\mu \pm \sigma$	$\mu \pm \sigma$	$\mu \pm \sigma$	$\mu \pm \sigma$	$\mu \pm \sigma$
100	Simple	1.29 ± 0.39	1.48 ± 1.16	1.49 ± 0.44	2.39 ± 1.92	1.72 ± 1.17
	GAN	1.29 ± 0.41	1.70 ± 2.12	1.49 ± 0.44	2.65 ± 2.17	1.71 ± 1.16
500	Simple	1.28 ± 0.42	1.36 ± 0.40	1.49 ± 0.44	2.19 ± 1.03	1.67 ± 1.22
	GAN	1.28 ± 0.42	1.36 ± 0.38	1.48 ± 0.45	2.33 ± 0.95	1.67 ± 1.22

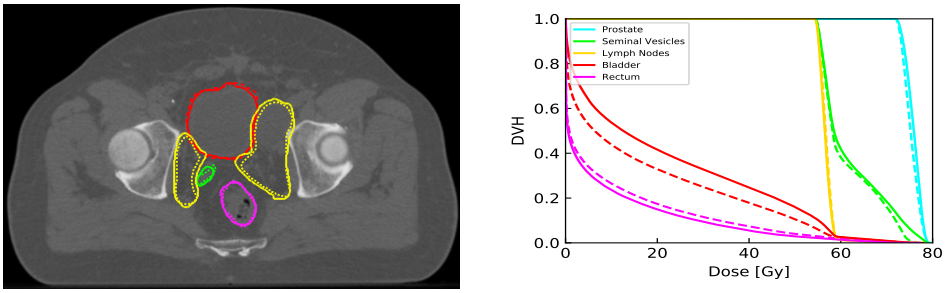
"Affine" refers to the affine registration defined in Eq. (2), which is considered a reference method. The weights 0.05 and 0.20 yielded very similar performance. We opted for a weight of 0.05 to avoid overfitting on the bladder. Since the target areas (prostate, lymph nodes, and seminal vesicles) obtained slightly better accuracy for a lower weight and these are important for radiotherapy planning, we selected 0.05. For the EMC dataset a similar experiment gave a weight of 0.01 (not reported). Therefore, for the remainder of the paper these weights have been used.

In order to investigate the difference between simple-inpainting and GAN-inpainting, we run the registration on HMC dataset using both techniques as shown in Table 2.3. The results shows a very similar performance for simple-inpainting and GAN-inpainting. Hence, the simple-inpainting is used for gas pocket inpainting for the remainder of the paper.

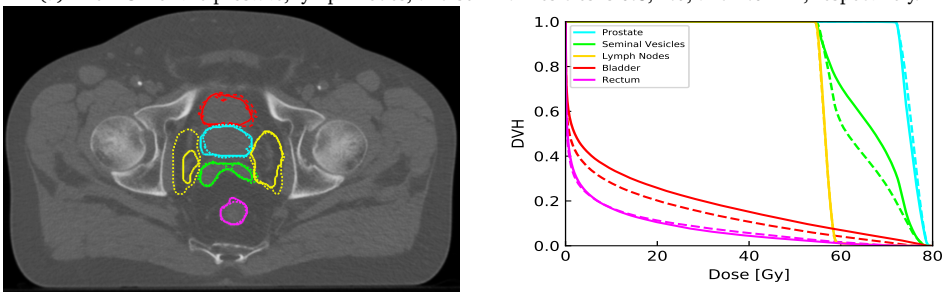
From the aforementioned experiments and analysis (Table 2.2 and 2.3), we noticed a similar performance between 100 and 500 iterations and in order to reduce the registration time, we considered only the results from 100 iterations for the final experiments.

2.3.5 Registration performance

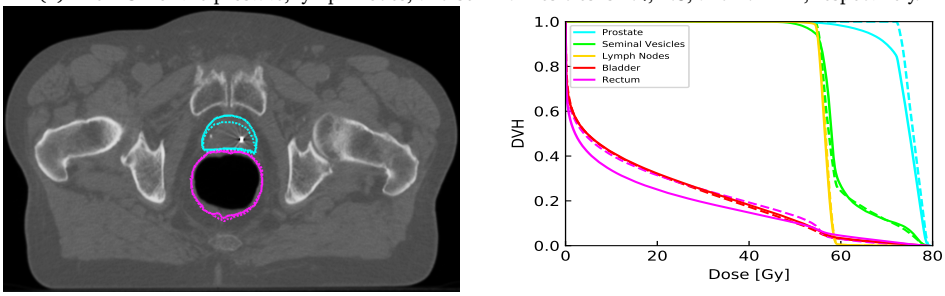
Since the LUMC dataset did not have any follow-up scans, we only consider the EMC and HMC datasets for evaluating the registration performance. Figure 2.6 shows example results of the automatically propagated contours. We compared the proposed method with the intensity-based registration approach of Qiao *et al.* [42]. For the HMC data we directly compare with the results reported in [42], as the same dataset was used. For the EMC data we applied their algorithm, and compare with our results. The DSC overlap of the proposed algorithm is presented in Table 2.4. For the HMC dataset, the prostate, lymph nodes, and bladder performed similarly for the proposed method and Qiao *et al.*, while the seminal vesicles and rectum showed substantial improvements. The median DSC values of the prostate, seminal vesicles, lymph nodes, rectum, and bladder were 0.88, 0.70, 0.89, 0.78, and 0.91 ,respectively for Qiao *et al.*, while they were 0.89, 0.73, 0.89, 0.85, and 0.94, respectively for the proposed method. For the EMC dataset, the proposed algorithm showed consistent improvement for the seminal vesicles, rectum, and bladder. The median DSC values of the prostate, seminal vesicles, rectum, and bladder were 0.91, 0.80, 0.76, and 0.86, respectively for Qiao *et al.* and 0.89, 0.81, 0.81, and 0.90, respectively for the proposed method. For the MSD results shown in Table 2.5, the proposed method outperformed Qiao *et al.* for all the target areas and OARs. The MSD of most of the targets and the OARs was less than one voxel (2 mm). The geometrical success rate was 97%, 93%, and 87% for the prostate, seminal vesicles, and lymph nodes, respectively for the HMC dataset and 67% and 71% for the prostate and seminal vesicles for the EMC dataset. Table 2.6 shows the 95% HD, yielding a significant improvement for the proposed method over Qiao *et al.* on the HMC dataset, but less improvement for the EMC dataset. Moreover, Qiao *et al.* and the proposed method show a significant improvement from the affine method except for the lymph nodes. Figure 2.7 shows a scatter plot depicting the effect of the bladder distension (volume difference between planning and daily CT) on the Mean Surface Distance (MSD) of different target organs of the HMC dataset. The figure shows that the MSD of the proposed method is less than the slice thickness (2 mm) for most of the cases, and that there is little correlation between registration performance and bladder distensibility. Figure 2.8 shows the comparison of the registration performance between Qiao *et al.* (intensity only) and the proposed method (intensity and bladder segmentation), both using 100 iterations for the HMC dataset. The comparison illustrates the performance in terms of DSC, MSD, and 95%HD for the target volumes and OARs. The figure shows a similar pattern between the proposed method using the manually annotated contours of the bladder and the contours from the 3D-CNN network. This pattern emphasizes that the proposed method achieved the upper limit of the system. The average runtime for the proposed pipeline is 98.3 seconds for each registration at 100 iterations.



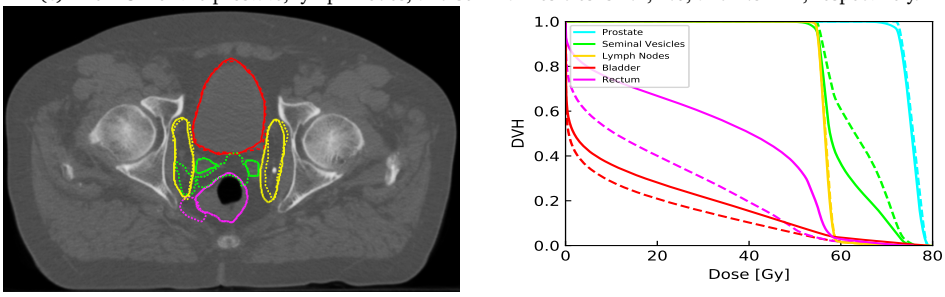
(a) The MSD of the prostate, lymph nodes, and seminal vesicles is 0.8, 1.6, and 1.0 mm, respectively.



(b) The MSD of the prostate, lymph nodes, and seminal vesicles is 1.4, 1.8, and 1.2 mm, respectively.

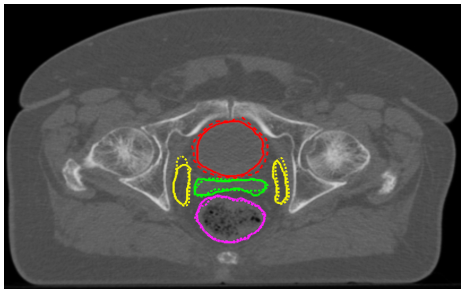
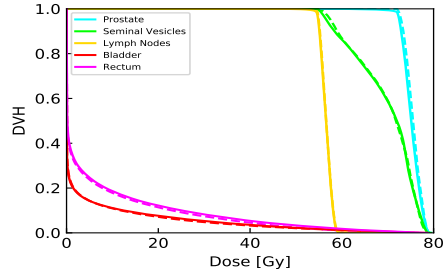


(c) The MSD of the prostate, lymph nodes, and seminal vesicles is 2.1, 1.6, and 1.5 mm, respectively.

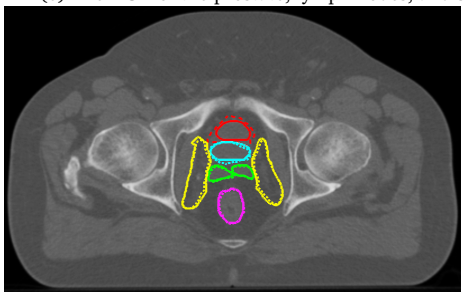
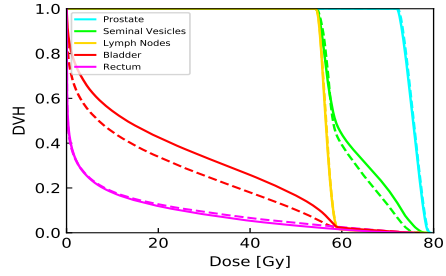


(d) The MSD of the prostate, lymph nodes, and seminal vesicles is 1.5, 1.6, and 11.0 mm, respectively.

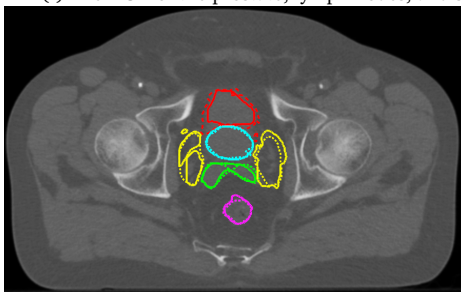
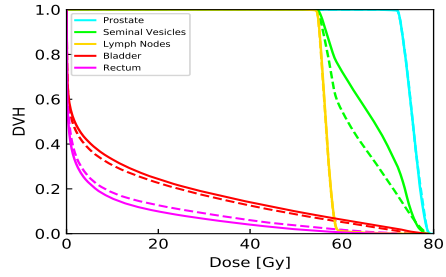
Figure 2.6: Examples from the automatic contours propagation of the HMC dataset and the corresponding dose volume histograms evaluated on the manual contours. The solid line represents the manual contouring results while the dotted line is the automatically propagated one.

(e) The MSD of the prostate, lymph nodes, and seminal vesicles is 1.3, 1.3, and 1.1 mm, respectively.

(f) The MSD of the prostate, lymph nodes, and seminal vesicles is 0.9, 1.4, and 0.9 mm, respectively.

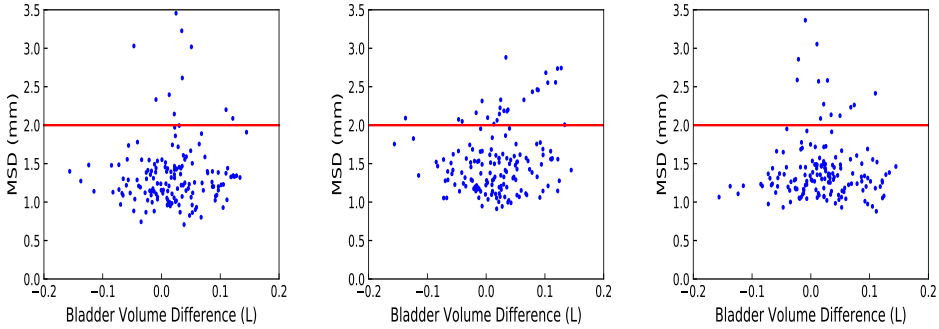



(g) The MSD of the prostate, lymph nodes, and seminal vesicles is 1.3, 1.1, and 1.1 mm, respectively.

Figure 2.6: Continued.

2.3.6 Dosimetric performance

Figure 2.6 shows the Dose Volume Histogram (DVH) of the target organs and OARs for some examples. The clinical constraints in terms of $V_{95\%}$ and $V_{107\%}$ were calculated for the prostate, seminal vesicles, and lymph nodes based on the manual contours. In order to monitor the accumulated dose for the OARs, we calculated $V_{45Gy\%}$, $V_{60Gy\%}$, $V_{75Gy\%}$, and D_{mean} for the rectum, as well as $V_{45\%}$, $V_{65Gy\%}$, and D_{mean} for the bladder. Here D_{mean} is the structure's average dose and $V_{xxGy\%}$ is the percentage of volume receiving a dose of xx Gy. Table 2.7 shows a comparison between the propagated contours from Qiao *et al.* and the proposed algorithm in terms of the percentage of scans that achieved the clinical criteria of $V_{95\%} \geq 98\%$ and $V_{107\%} \leq 2\%$. The Table shows a significant improvement for the seminal vesicles, which is a small and difficult



(a) Prostate (b) Lymph nodes (c) Seminal vesicles

Figure 2.7: Scatter plot showing the effect of the bladder volume change between planning and daily scans of the HMC dataset on the performance of the proposed method in terms of MSD. Red line represents the slice thickness.

Table 2.4: DSC value of the target volumes and the OARs of the HMC and EMC datasets for different registration methods. † represents a significant difference (at $p = 0.05$) between Qiao *et al.* and the proposed algorithm.

			Prostate	Seminal vesicles	Lymph nodes	Rectum	Bladder
	Method	# It.	$\mu \pm \sigma$	$\mu \pm \sigma$	$\mu \pm \sigma$	$\mu \pm \sigma$	$\mu \pm \sigma$
HMC	Affine	200	0.84 ± 0.11	0.46 ± 0.26	0.90 ± 0.08	0.71 ± 0.10	0.77 ± 0.11
	Qiao *et al.*	100	0.87 ± 0.08	0.65 ± 0.18	0.88 ± 0.07	0.77 ± 0.09	0.88 ± 0.11
	Proposed	100	0.87 ± 0.08	$0.70 \pm 0.13^{\dagger}$	0.87 ± 0.07	$0.82 \pm 0.12^{\dagger}$	0.89 ± 0.12
EMC	Affine	200	0.78 ± 0.20	0.49 ± 0.32	-	0.62 ± 0.18	0.66 ± 0.25
	Qiao *et al.*	100	0.87 ± 0.13	0.70 ± 0.26	-	0.72 ± 0.16	0.78 ± 0.22
	Proposed	100	0.87 ± 0.12	$0.75 \pm 0.18^{\dagger}$	-	$0.78 \pm 0.15^{\dagger}$	$0.83 \pm 0.17^{\dagger}$

target organ, while the performance of the prostate and lymph nodes was very similar. The boxplot in Figure 2.9 illustrates the difference between the dosimetric parameter values of the manual delineations, calculated by using either the treatment plan based on the automated delineations or the manual delineations. We can see that the difference for all dosimetric parameters of all the target organs and OARs is almost 0 % or Gy except for the lymph nodes, which is approximately 1%.

2.4 Discussion

In this study, we developed and evaluated an automatic contour propagation pipeline using DIR, while considering the robustness, accuracy, and clinical acceptance rate for the target organs and the OARs of prostate cancer. Online adaptive IMPT is a crucial step towards treatment with small margins for target organs. In this study we used margins of 2 mm for the prostate and 3.5 mm for the seminal vesicles and lymph nodes, respectively. Such small margins are only viable when online and daily re-planning is

Table 2.5: MSD (mm) of the target volumes and the OARs of the HMC and EMC datasets for different registration methods. † represents a significant difference (at $p = 0.05$) between Qiao *et al.* and the proposed algorithm.

			Prostate	Seminal vesicles	Lymph nodes	Rectum	Bladder
	Method	# It.	$\mu \pm \sigma$	$\mu \pm \sigma$	$\mu \pm \sigma$	$\mu \pm \sigma$	$\mu \pm \sigma$
HMC	Affine	200	1.70 ± 0.96	3.02 ± 1.96	1.26 ± 0.51	3.92 ± 1.59	4.47 ± 2.27
	Qiao *et al.*	100	1.40 ± 0.47	1.85 ± 1.26	1.51 ± 0.44	3.13 ± 1.38	2.38 ± 1.79
	Proposed	100	1.29 ± 0.39	1.48 ± 1.16	1.49 ± 0.44	$2.39 \pm 1.92^{\dagger}$	$1.72 \pm 1.17^{\dagger}$
EMC	Affine	200	2.82 ± 3.18	4.42 ± 6.03	-	4.63 ± 3.01	8.03 ± 6.46
	Qiao *et al.*	100	1.41 ± 0.76	2.24 ± 3.14	-	3.21 ± 1.85	5.42 ± 5.84
	Proposed	100	1.54 ± 0.67	$1.67 \pm 1.38^{\dagger}$	-	$2.67 \pm 1.76^{\dagger}$	$3.89 \pm 4.00^{\dagger}$

Table 2.6: %95HD (mm) of the target volumes and the OARs of the HMC and EMC datasets for different registration methods. † represents a significant difference (at $p = 0.05$) between Qiao *et al.* and the proposed algorithm.

			Prostate	Seminal vesicles	Lymph nodes	Rectum	Bladder
	Method	# It.	$\mu \pm \sigma$	$\mu \pm \sigma$	$\mu \pm \sigma$	$\mu \pm \sigma$	$\mu \pm \sigma$
HMC	Affine	200	3.97 ± 1.96	6.61 ± 3.70	3.12 ± 1.27	11.8 ± 5.98	12.5 ± 7.06
	Qiao *et al.*	100	3.31 ± 1.16	4.59 ± 2.95	3.73 ± 1.02	10.4 ± 5.99	7.41 ± 6.85
	Proposed	100	3.07 ± 1.30	$3.82 \pm 3.19^{\dagger}$	3.74 ± 1.02	$8.66 \pm 6.92^{\dagger}$	$5.11 \pm 4.38^{\dagger}$
EMC	Affine	200	5.98 ± 6.19	8.11 ± 7.66	-	13.2 ± 6.88	21.3 ± 16.3
	Qiao *et al.*	100	3.65 ± 2.31	4.80 ± 5.09	-	11.3 ± 6.77	16.5 ± 17.2
	Proposed	100	3.93 ± 2.24	4.92 ± 5.13	-	10.4 ± 7.77	$11.5 \pm 12.5^{\dagger}$

Table 2.7: Percentage of registrations that meets the dose constraints for different registration iterations. Conservative Success Rate (CSR) refers to the percentage of registrations for which all target volumes (the prostate, seminal vesicles and lymph nodes) meet the dose constraints.

	$V_{95\%} \geq 98\%$				$V_{107\%} \leq 2\%$		
	Prostate	SV	LN	CSR	Prostate	SV	LN
Qiao *et al.*	83.8%	75.7%	97.9%	65%	100%	100%	100%
Proposed	85.8%	90.9%	98.9%	80%	100%	100%	100%

performed. This re-planning procedure should be accurate as well as robust to avoid any subsequent adverse side effects. The automatically propagated contours were validated geometrically on the EMC and HMC datasets as well as dosimetrically on the HMC dataset in order to investigate whether or not the propagated contours meet the clinical acceptance criteria for dose coverage. DSC, MSD, and 95%HD were chosen for geometric validation while $V_{95\%} \geq 98\%$ and $V_{107\%} \leq 2\%$ were used for dosimetric coverage validation. Here, $V_{95\%} \geq 98\%$ ensures that at least 98% of the target volumes receive at least 95% of the prescribed dose and $V_{107\%} \leq 2\%$ ensures that less than 2% of the target volumes receive more than 107% of the highest prescribed dose.

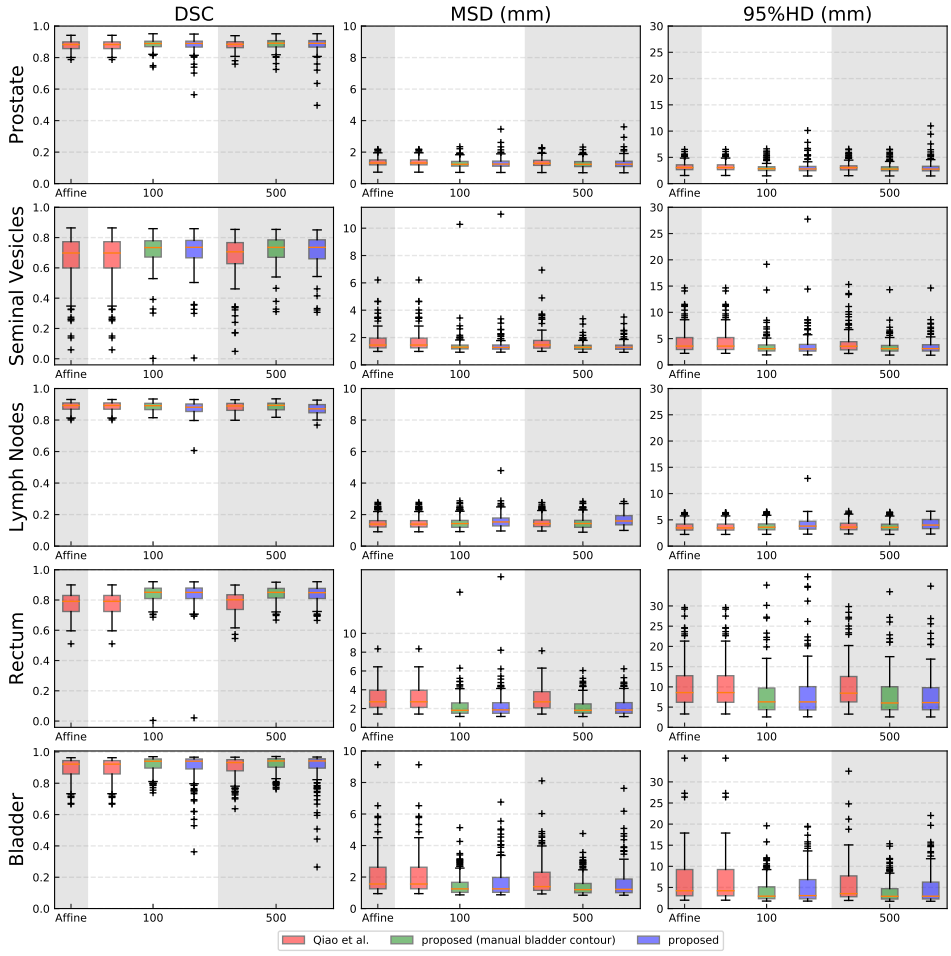


Figure 2.8: Boxplot comparison between Qiao *et al.* and the proposed algorithm for image registration on the HMC dataset versus the number of iterations. The columns show the DSC, MSD, and 95%HD from left to right. Prostate, seminal vesicles, lymph nodes, rectum, and bladder are shown from top to bottom rows, respectively. The red box is the method from Qiao *et al.*, the blue box is the proposed method, while the green box is an upper bound of the proposed method using manual daily contours.

In order to enhance the registration robustness, the segmentation of the bladder was introduced to steer the optimization. Since the registration process is partially driven by the bladder segmentation, this segmentation should be as accurate and robust as possible. Hence, we chose a 3D-CNN for bladder segmentation, and obtained a DSC of 87.9% and a Jaccard index of 80.2%, which is very comparable to the reported Jaccard index of 81.9% in [74], where the authors developed a CNN network alongside

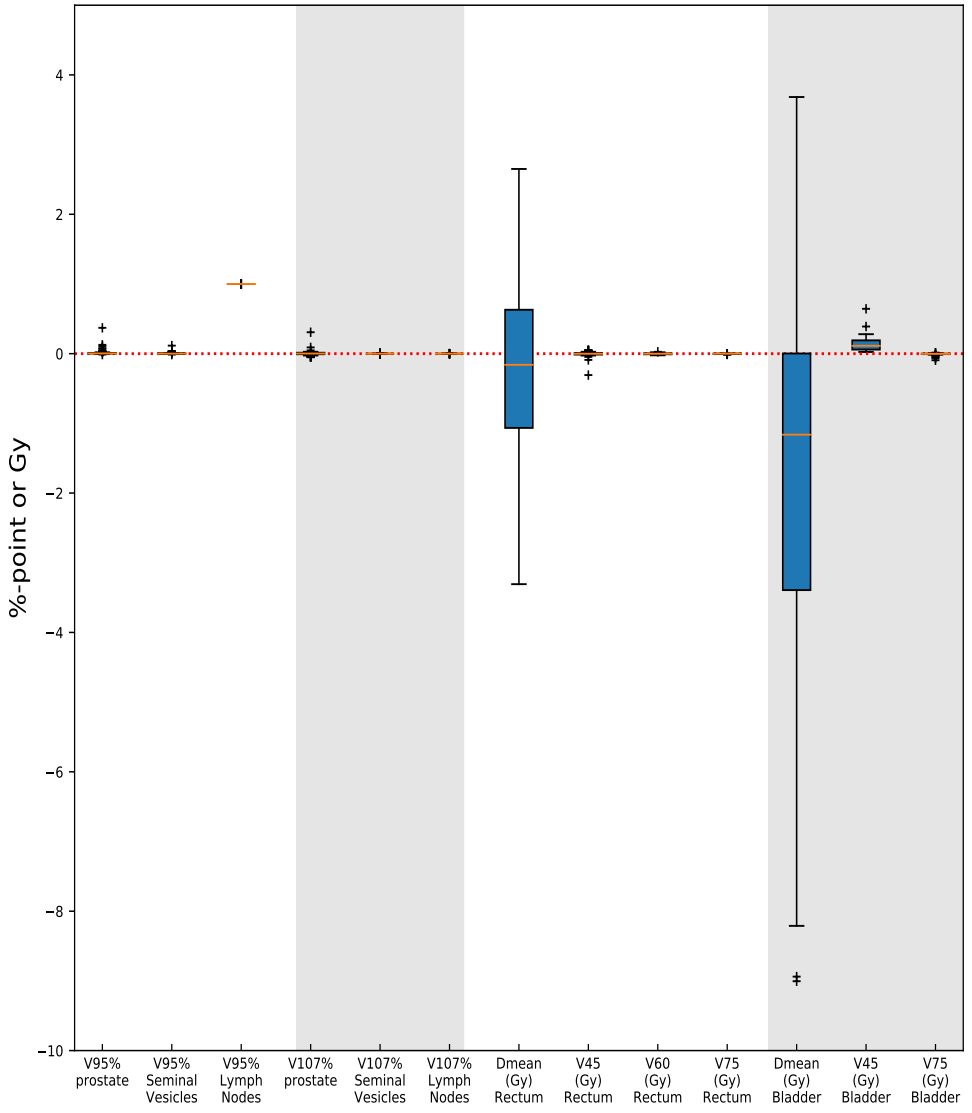


Figure 2.9: Boxplot depicting the difference in dosimetric parameters of the manual delineations, calculated by using either the treatment plan based on the automated delineations or the manual delineations for 99 scans of the HMC dataset.

level-sets to segment the bladder in CT urography. Moreover, our proposed network outperformed the 2D CNN network developed by Zhou *et al.* [75], where the authors reported a DSC of 72%. The high performance of the proposed network may be attributed to the use of a large receptive field as well as replacing the 2D convolutions with 3D convolutions, which helps the network to embed depth information.

Applying contrast clipping to the CT scans before registration was beneficial to the registration process, since the registration is intensity-based, which is consistent with the findings in [76]. Inpainting gas pockets in the rectum enhanced the registration of the rectum as well as the seminal vesicles. The presence of these pockets were challenging for the registration due to the physical non-correspondence between the daily and planning CT scans. Although the inpainting results from the GAN-inpainting network were more realistic than the simple-inpainting procedure, a similar performance with respect to the registration was obtained. Our explanation for this finding is that the mutual information similarity metric pays more attention to the overall intensity distribution and since the results from the simple-inpainting were blended and smoothed with respect to its neighbours, it produces a similar histogram distribution to the GAN-inpainting and subsequently gives a similar registration performance.

The initialization of the registration algorithm on the bony structures is a crucial step for optimal performance, which is consistent with the reported results in [42]. Moreover, masking out the couch using a torso mask removed its disrupting effect on the registration. Increasing the number of iterations had a minimal effect on the registration performance while increasing the registration time. We found that the effect of adding a third registration step focussing on the rectal area, boosted the performance regarding the rectum and seminal vesicles while there was no detrimental effect for the prostate, lymph nodes, and bladder.

In this study, we focused on the generalizability and robustness of the registration represented by performance on different datasets and the number of failed registrations according to geometrical and dosimetric criteria. This target is achieved through several steps. First, inpainting the rectum gas pockets. Second, enhancing the CT image contrast by contrast clipping. Third, introducing the bladder segmentation with an optimized weights ($\alpha = 0.05$ and 0.01) to steer the optimization problem to a better local minimum while avoiding overfitting to the bladder. Fourth, using a third stage for registration to focus on the rectum and consequently the seminal vesicles by using a dilated mask for the rectum. Overall, these steps yielded a more robust registration and substantially decreased the number of registrations with insufficient quality, especially for the seminal vesicles, rectum, and bladder. Improving the MSD for the seminal vesicles, which is an important target volume, resulted in a more precise targeting with potential benefits in terms of local control (lower probability of recurrences). Moreover, both the rectum and the bladder improved in terms of MSD and 95% HD, thereby avoiding treatment-induced complications after the therapy, so a higher probability of better quality-of-life after treatment. For the bladder, 11 of the 18 registrations with an MSD larger than the top whisker in Fig. 2.8, were belonging to two patients. For these two patients the 3D-CNN achieved an average

DSC of 0.65, explaining the suboptimal performance of the proposed method on these cases. From the CT images no apparent reason for this was found. In terms of the geometric success rate defined by the number of registrations that achieved an MSD lower than 2 mm (slice thickness), the system achieved 97%, 93%, and 87% for the prostate, seminal vesicles, and lymph nodes, respectively. This compares to a success rate of 95%, 78%, and 86% for Qiao *et al.*, i.e. especially improving the performance for the seminal vesicles. Moreover, the proposed system showed robustness to the change in bladder distension between planning and daily CT as shown in Figure 2.7. The proposed registration method achieved quite similar results on the EMC and HMC datasets, except for the bladder. We suspect this is partially due to the difference in bladder segmentation performance of the neural network, which was 82% on the EMC data and 88% on the HMC data. It could also be related to the affine registration results for the EMC dataset (Table V) being slightly less than HMC dataset. We visually checked the affine results and noticed that the field of view for some cases were cropped or zoomed. The average runtime for the proposed pipeline is 98.3 seconds for each registration at 100 iterations, comparing to 13.5 seconds reported by Qiao *et al.* However, the pipeline could be further optimized and adapted for GPU acceleration. For validating the clinical acceptance of the proposed algorithm, we considered $V_{95\%} \geq 98\%$, $V_{107\%} \leq 2\%$, and CSR for dosimetric coverage for 99 registrations. All the scans meet the $V_{107\%} \leq 2\%$ constraint. Fourteen out of the 99 registrations (14.1%) did not directly meet the $V_{95\%} \geq 98\%$ constraint for the prostate. After visual inspection of these failure cases, we found inconsistencies between the manual delineations for the planning and daily CT scans for 7 cases. These cases had a $V_{95\%}$ of $92.5\% \pm 0.1\%$, meaning that these cases were still close to be dosimetrically acceptable. The proposed algorithm improved the contouring quality and robustness especially for the seminal vesicles, which directly increased the percentage of acceptable scans from 75.5% to 90.9% for this important target organ. These success rates imply that the automatically generated contours have the potential to be employed for online adaptive IMPT. Moreover, the typical 7 mm margins [77] may be replaced with smaller daily margins, which means delivering an effective dose with potentially less adverse effects.

The reported performance of the proposed pipeline could be further improved by correcting the inconsistency present in the manual contouring. Also, the weighting parameter α could be selected automatically by introducing it as a trainable parameter. Moreover, the current 3D-CNN was trained using CT scans without contrast material, and therefore is unlikely to perform well on scans acquired with contrast. In case the clinical protocol dictates contrast-enhanced CT acquisitions, the network could be easily retrained. We may further investigate the effect on segmentation performance of CT clipping as a preprocessing step for the 3D-CNN for bladder segmentation. We also consider developing an end-to-end neural network to jointly optimize the registration

and segmentation tasks to further improve the system robustness and accuracy.

2.5 Conclusion

In this study we proposed a registration pipeline for automatic contour propagation for online adaptive IMPT of prostate cancer using the open source package `elastix` software in combination with deep learning. The proposed pipeline achieved a geometrical success rate of 97%, 93%, and 87% for the prostate, seminal vesicles, and lymph nodes, respectively for HMC dataset as well as 67% and 71% for the prostate and seminal vesicles, respectively for ECM dataset. The HMC automatically propagated contours meet the dose coverage constraints in 86%, 91%, and 99% of cases for these targets. A Conservative Success Rate (CSR) of 80% was achieved, meaning that 80% of the automatically generated treatment plans can be directly used without manual correction. This re-contouring showed a promise for generating daily treatment plans. Moreover, it showed a substantial improvement in the system robustness compared to a previous open source method, which means that more treatment plans can be directly used without manual correction, which is a crucial factor for enabling online daily adaptation and thus the use of relatively small treatment margins. Therefore, the proposed method could facilitate online adaptive proton therapy of prostate cancer. The authors have no relevant conflicts of interest to disclose.

2.6 Acknowledgements

This study was financially supported by Varian Medical Systems and ZonMw, the Netherlands Organization for Health Research and Development, grant number 104003012. The HMC dataset with contours were collected at Haukeland University Hospital, Bergen, Norway and were provided to us by responsible oncologist Svein Inge Helle and physicist Liv Bolstad Hysing; they are gratefully acknowledged.