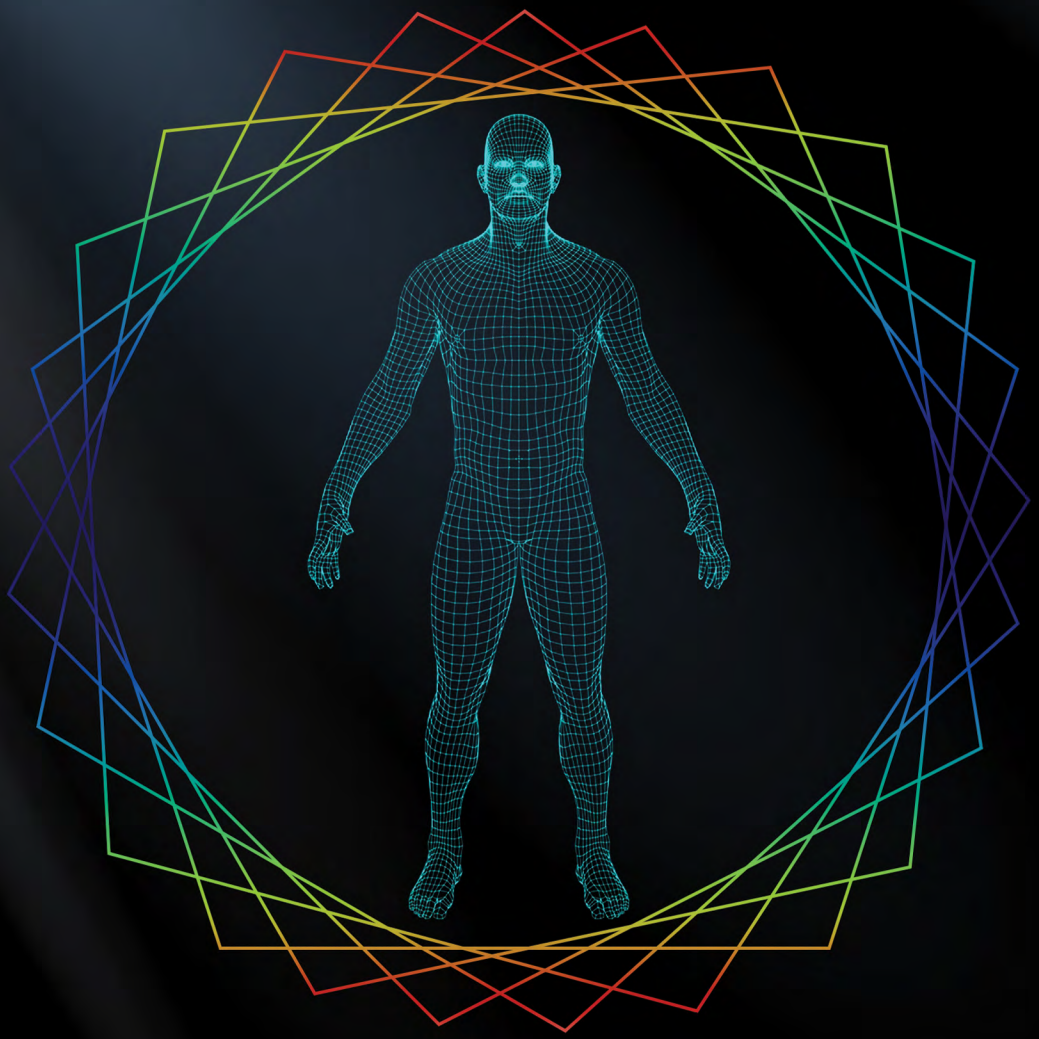
Deep learning for online adaptive radiotherapy

Elmahdy, M.S.E.

DEEP LEARNING

FOR ONLINE ADAPTIVE RADIOTHERAPY



MOHAMED S. ELMAHDY

Colophon

About the cover: The cover is a symbolic representation of a deformation grid overlaid on a coronal plane of a human body while being irradiated.

Deep Learning for Online Adaptive Radiotherapy
Mohamed S. Elmahdy

ISBN: 978-94-92597-93-9
Thesis layout & cover designed by Mohamed S. Elmahdy
Printed by Boekendeal

Deep Learning for Online Adaptive Radiotherapy

Proefschrift

ter verkrijging van
de graad van doctor aan de Universiteit Leiden,
op gezag van rector magnificus prof.dr.ir. H. Bijl,
volgens besluit van het college voor promoties
te verdedigen op dinsdag 15 maart 2022
klokke 15.00 uur.

door

Mohamed Said Elbially Elmahdy
geboren te Caïro, Egypte in 1991

Promotores:	Prof. dr. U. A. van der Heide Prof. dr. ir. B. P. F. Lelieveldt
Co-promotor:	Dr. ir. M. Staring
Leden promotiecommissie:	Prof. dr. C. Rasch Prof. dr. M. J. P. van Osch Prof. dr. J. P. W. Pluim *Technical University Eindhoven* Prof. dr. I. Isgum *University of Amsterdam*

The research in this thesis was performed at the Division of Image Processing (LKEB), Department of Radiology of Leiden University Medical Center, The Netherlands.

Financial support for the publication of this thesis was kindly provided by:
LKEB,
Library of Leiden University
Bonitus Stichting,

Contents

1

Introduction

1.1 Automatic delineation in adaptive radiotherapy

Cancer is a leading cause of mortality, accounting for about 10 million deaths worldwide annually. The National Cancer Institute (NCI) predicts that by 2040, the number of new cancer cases per year will rise to 29.5 million and the number of cancer-related deaths to 16.4 million [1]. Radiotherapy (RT) is one of the widely used treatment options for cancer diseases, where a high dose of ionizing radiation damages the DNA of cancerous cells [2]. RT is frequently combined with other treatment modalities such as chemotherapy, surgery, and recently immunotherapy [3]. RT dose is usually delivered by using a technique called Intensity Modulated Photon Therapy (IMRT) or Intensity Modulated Proton Therapy (IMPT). Prior to the radiation treatment of a patient, a personalized treatment plan is constructed based on a planning CT scan and sometimes augmented with MR and/or PET images. RT doses are usually fragmented over 4 to 8 weeks resulting in 20 to 40 daily fractions [4]. Since the dose is delivered over several sessions, variations in the size and shape of the target area and Organs-At-Risk (OARs) are bound to take place. These variations are caused by multiple factors such as patient motion, weight loss, breathing, organ filling, tumors and OARs shrinkage [5]. IMPT is more sensitive to these daily changes than IMRT, which may result in the dose being distorted or dose not match the anatomy of the day [4, 6]. The simplest strategy to compensate for these variations is to account for them beforehand by adding a margin to the Clinical Target Volume (CTV) to generate the Planning Target Volume (PTV), and then generate the treatment plan using the PTV. However, these margins result in extra dose to the OARs, thus increasing the risk of toxicity, making it a suboptimal strategy. A more sophisticated approach is called Adaptive Radiotherapy (ART), where we account for inter-fraction variations by adapting the treatment plan online to the daily anatomy. Figure 1.1 illustrates

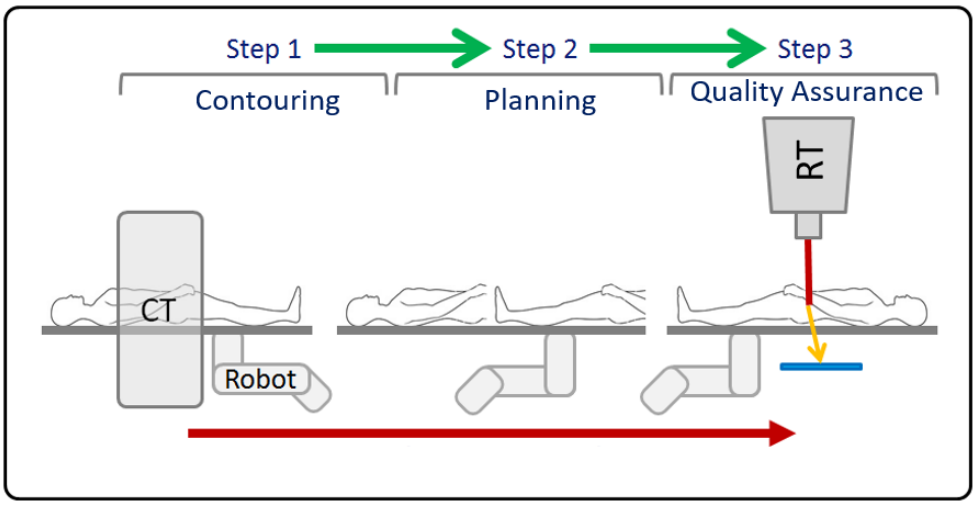


Figure 1.1: Illustration of the online adaptive radiotherapy workflow. A CT scan of the day is acquired using an in-room scanner, after which the patient table is moved to the treatment beam via a robotic arm. During the movement of the table, the daily CT scan is re-contoured and the treatment plan is adapted accordingly.

the workflow of online adaptive radiotherapy using an in-room CT scanner. However, these daily CT scans have to be delineated online 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 preference and experience differences 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 or correction strategies are needed. Automatic contouring can be done by direct segmentation of the daily scan, or by registration of the annotated planning scan with the daily scan followed by contour propagation. These two methods are discussed in detail in the following sections.

1.2 Image registration in radiotherapy

Image registration is the task of finding the geometrical correspondence between images that were acquired at different time steps or from different imaging modalities. In this dissertation we focus on deformable, non-linear, image registration (DIR) which accounts for local deformations such as stretching or shrinkage deformations. In the context of adaptive radiotherapy, the aim is to find the transformation that aligns the planning scan (moving image) and daily scan (fixed image) of the same

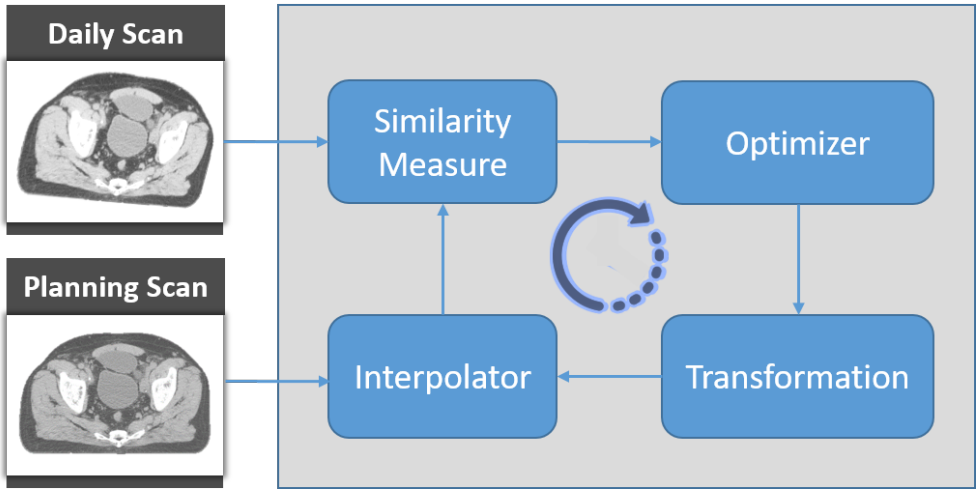


Figure 1.2: A diagram of iterative-based image registration algorithms.

patient (inter-fractional registration). DIR has been extensively integrated into RT applications such as dose planning, delivery and evaluation, since all these tasks can be improved by accounting for organ deformation. DIR is an ill-posed problem because there is no unique transformation between the fixed and moving images. Therefore it is formulated as an optimization problem (see Figure 1.2) that can be solved using iterative approaches such as Stochastic Gradient Descent (SGD):

$$\hat{\mu} = \arg\min_{\mu} C\left(I_f(\boldsymbol{x}),\, I_m(\boldsymbol{x});\, T_\mu(\boldsymbol{x})\right), \tag{1.1}$$

where $I_f(\boldsymbol{x})$, $I_m(\boldsymbol{x})$ are the fixed and moving images, C is a dissimilarity metric such as Normalized Cross Correlation (NCC), and $T(\boldsymbol{x})$ is the transformation that makes $I_m(T(\boldsymbol{x}))$ spatially aligned to $I_f(\boldsymbol{x})$. The transformation function can be modeled by a limited number of parameters (parametric transformation) or by a vector per voxel describing the displacement of this voxel in a continuous space using interpolation (non-parametric transformation). In this thesis we focus on non-parametric transformation. For contour propagation, the contour of the moving image is resampled in the fixed image domain by interpolation.

With the recent advance of deep learning, a variety of methods on learning-based registration were proposed to replace conventional iterative methods. This research either uses deep learning to model the transformation function via supervised training through synthesized deformations [12, 13, 14] or via unsupervised training by equipping the network with a spatial transformer [15] similar to the work presented in [16, 17, 18, 19]. For further details, the reader may refer to [20] where a detailed review on medical image registration using deep learning is provided.

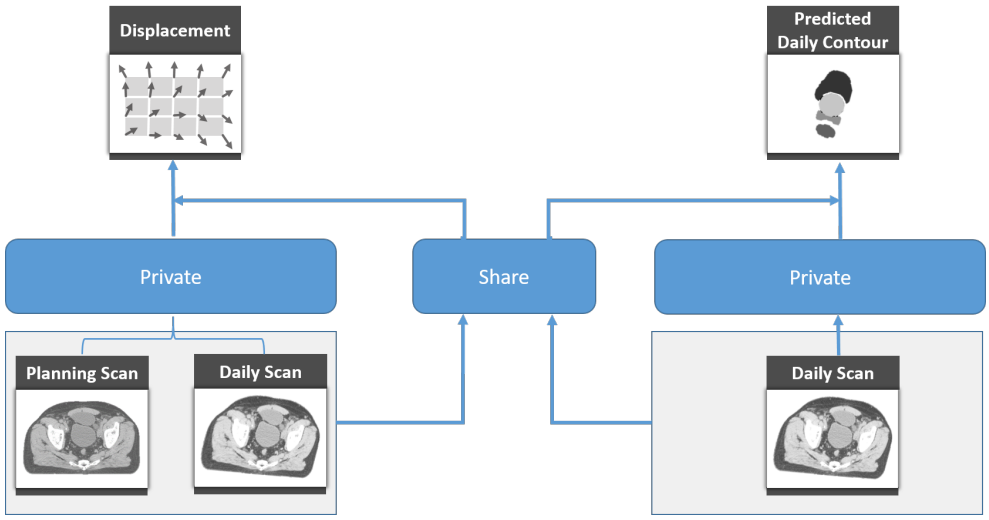


Figure 1.3: A diagram of joint registration and segmentation algorithms.

1.3 Joint registration and segmentation

Automatic contouring of the daily anatomy can be done by either a segmentation algorithm or via DIR as mentioned in Section 1.1. However, each of these tasks has its own strengths and weaknesses. For instance, image segmentation algorithms can directly delineate images based on texture and surrounding anatomy, and may therefore be robust to large organ deformations. However, it sometimes has difficulties with low contrast areas and irregularly shaped organs. On the other hand, image registration algorithms have the ability to encode prior knowledge of the patient's anatomy and therefore may perform better on low quality images. However, such methods sometimes have difficulty with large deformations. These two tasks are in fact complementary, as for example image atlases warped by image registration algorithms are often used for image segmentation [21, 22], while image contours can be used to guide the image registration method in addition to the intensity images [23, 17, 24]. Contours are also used for evaluating the quality of the registration [25, 26]. Therefore, coupling image registration and segmentation tasks and modeling them in a single framework could leverage their strengths and mitigate their weaknesses through the sharing of beneficial information. Figure 1.3 is a diagram showing how these tasks might be joined.

1.4 Accelerating MR acquisition for adaptive radiotherapy

CT-guided radiotherapy has long been the standard setup for radiation oncology despite its low contrast compared to MRI since it encodes crucial information about

tissue electron density needed for dose calculation and simulation. The introduction of MRI-guided radiation therapy (MRIgRT) has revolutionized the practice of radiation oncology. This is due to the superior performance of MR images over CT in terms of soft tissue contrast (see Figure 1.4), which facilitates the monitoring of the daily anatomical changes and subsequently adapt the treatment radiation in real-time. Any substantial latency in this MRI-guided workflow might, however, result in the introduction of new anatomical variations and subsequently the delivered dose would be different from the intended dose [27]. One of the potential causes for this latency is imaging latency, which represents the time delay between the anatomical change and its emergence in the reconstructed image.

Therefore, fast image acquisition and reconstruction are crucial to improve the performance of current MR scanners, which led in recent years to the development of techniques such as parallel reception, compressed sensing and multi-band accelerations. However, there is still a need for further scan acceleration. The long acquisition time is intrinsic to the scanner and physics properties of MRI. For the majority of scans performed in clinical practice, this acquisition is done through consecutive read-outs of single lines in k-space. The scanning time could be shortened by reducing the number of acquired lines in k-space, i.e. by undersampling the 2D or 3D k-space. However, this could violate the Nyquist criterion, resulting in aliasing and blurriness in the reconstructed images. These issues may result in a lag between the organ positions derived from the reconstructed image and the actual positions by the time the acquisition is finished. Compressed Sensing (CS) is one of the most common solutions for acceleration by undersampling, while maintaining image quality. CS was introduced by Donoho [28], Lustig [29] and Candes [30], where it leverages the fact that MR images can be compressed in some domain, restoring the missing k-space data through an iterative reconstruction algorithm [31].

Recently deep learning-based algorithms were introduced in order to reconstruct a high quality MR images at acceptable speed. These algorithms often focus on modeling iterative approaches similar to CS algorithm via deep learning models [32, 33, 34].

1.5 Outline of the thesis

The aim of the work described in this thesis is to develop a deep learning-based methodology for automatic contouring for real time adaptive radiotherapy either guided by CT or MR imaging modalities. Our proposed automatic contouring networks were trained and tested on CT images since it is the commonly used for treatment planning, while for MR we focused on the other bottleneck, i.e. the reconstruction time.

Chapter 2 presents a contour propagation pipeline that combines conventional iterative-based registration with a deep learning model. We propose a CNN network

Figure 1.4: Example of prostate images, where from left to right are CT, T2-weighted MR, and fat suppressed MR images [35].

that automatically segments the bladder, and then feeds it to the registration algorithm as prior knowledge of the underlying anatomy. We also introduce a GAN model to address the problem of gas pockets in the rectum to avoid the registration algorithm to be distracted.

Chapter 3 presents a novel transfer learning approach in order to leverage personalized anatomical knowledge accumulated over the treatment sessions. We adapt a baseline segmentation model as the patient goes through their RT treatment. Thus, instead of depending on a static deep learning segmentation model for all patients, we accumulate knowledge over successive sessions for a particular patient. This accumulated knowledge is then used to encourage the model into predicting a segmentation that has a higher quality for this specific patient.

Chapter 4 proposes to combine the registration and segmentation tasks in a deep learning setting using adversarial learning. The proposed framework consists of an unsupervised 3D end-to-end generator network that estimates the deformation vector field (DVF) between the input image pairs. Meanwhile, a discriminator network is trained to evaluate how well the registration is performing.

Chapter 5 proposes to formulate the registration and segmentation as a joint problem via a Multi-Task Learning (MTL) setting, allowing these tasks to leverage their strengths and mitigate their weaknesses through the sharing of beneficial information. We explored different joint network architectures as well as loss weighting methods for merging these tasks, thus pinpointing the best strategy to maximize the information flow between the two tasks.

Chapter 6 presents a fast MR reconstruction algorithm, which enables the application of the automatic contouring methods proposed in the previous chapters for online adaptive MR-guided radiotherapy. Starting from undersampled k-space data, an iterative learning-based reconstruction scheme inspired by compressed sensing theory is used to reconstruct the images. We developed a novel deep neural network to refine and correct prior reconstruction assumptions given the training data. The proposed

network was ranked #1, shared #1, and #3 on respectively the 8x accelerated multi-coil, the 4x multi-coil, and the 4x single-coil tracks in the fastMRI competition organized by Facebook and New York University (NYU).

Chapter 7 summarizes and discusses the ideas presented in this thesis.

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

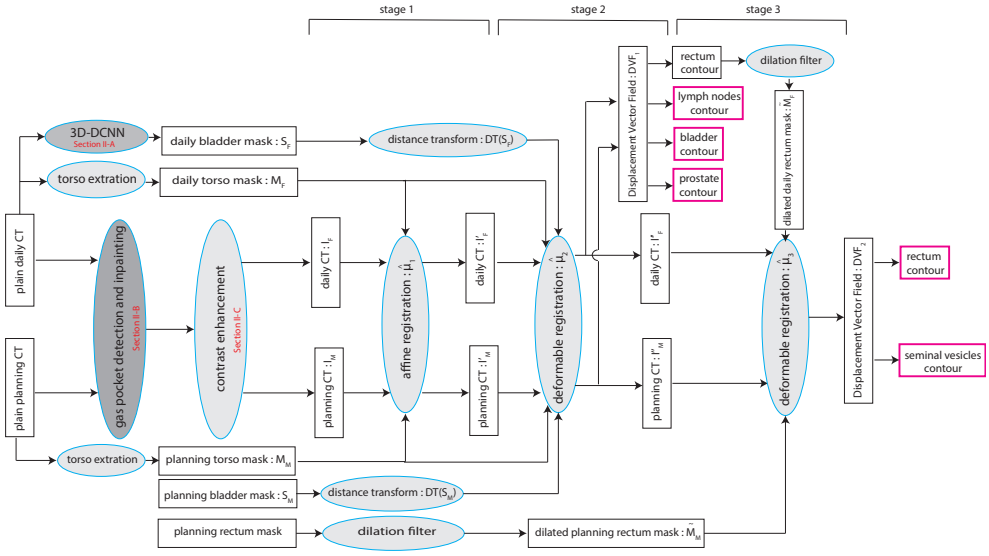


Figure 2.1: The proposed multi-stage registration process using `elastix` software and deep learning. The red boxes denote the contours finally used as output of the algorithm.

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

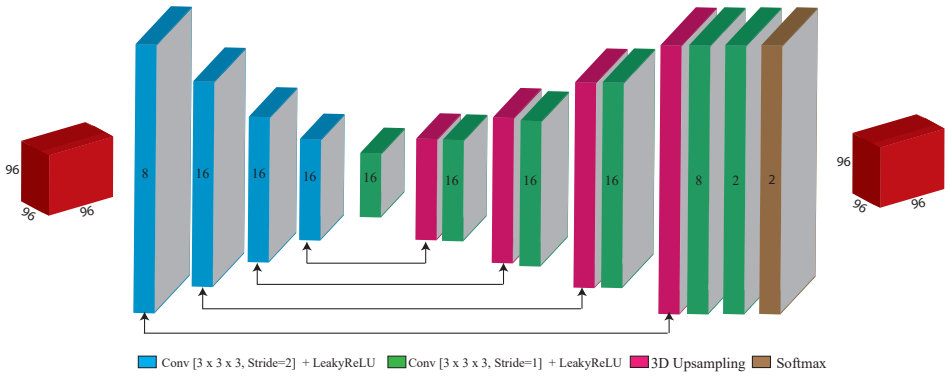


Figure 2.2: The architecture for the 3D-CNN network, where the numbers on the blocks denote the number of feature maps.

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.

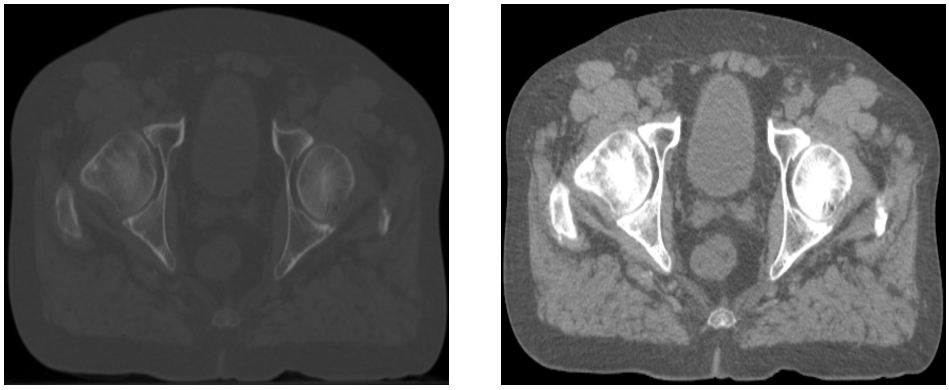
Alternatively, we also experimented with a simplified method for inpainting. Following the idea proposed by Rodriguez-Vila *et al.* [61] we fill the gas pockets with a fixed value and smooth the output to blend it with the surrounding tissues. A threshold of -200 is used to generate a binary mask of 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.

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

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

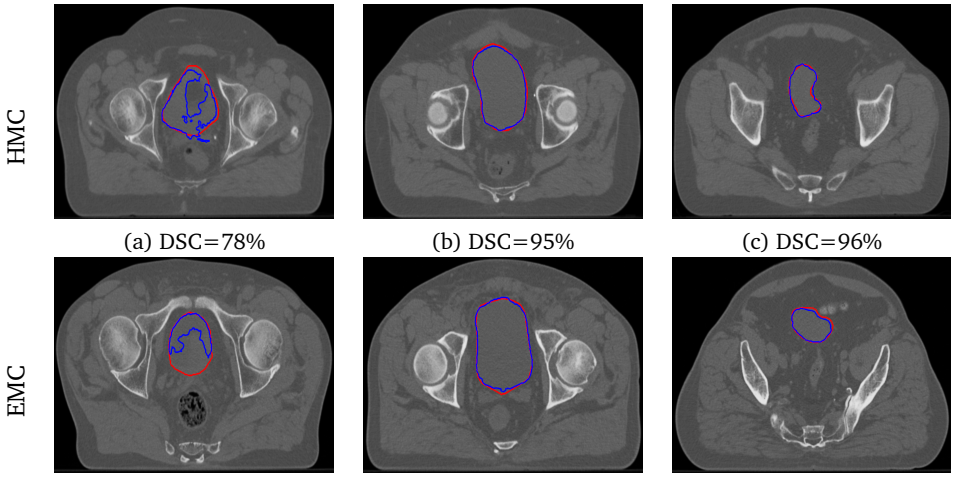
where $\{a_1, a_2, ..., a_n\}$ and $\{b_1, b_2, ..., b_m\}$ are the surface mesh points of the fixed and moving contours, respectively and $d\,(a_i, M) = \min_j ||\, b_j - a_i\,||$. The geometrical success rate, as a marker for geometric robustness, is defined as the percentage of registrations with 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

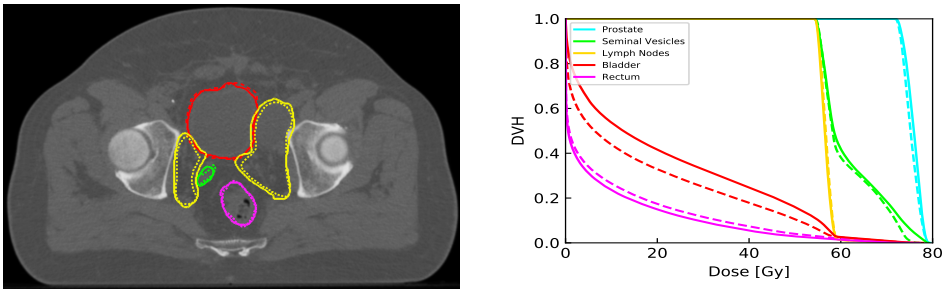
"Affine" refers to the affine registration defined in Eq. (2), which is considered a reference method. The weights 0.05 and 0.20 yielded very similar performance. We opted for a weight of 0.05 to avoid overfitting on the bladder. Since the target areas (prostate, lymph nodes, and seminal vesicles) obtained 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.

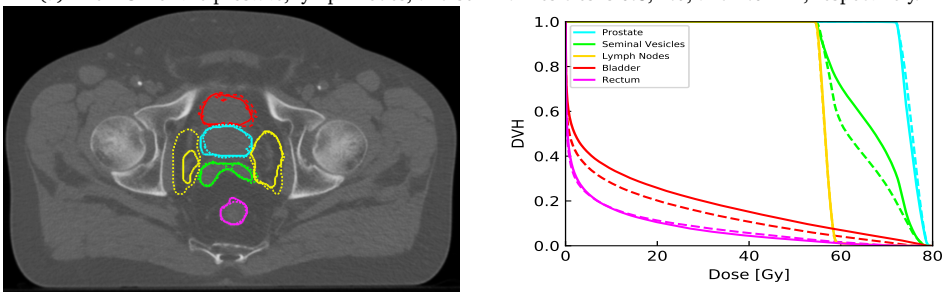
From the aforementioned experiments and analysis (Table 2.2 and 2.3), we noticed a similar performance between 100 and 500 iterations and in order to reduce the registration time, we considered only the results from 100 iterations for the final experiments.

2.3.5 Registration performance

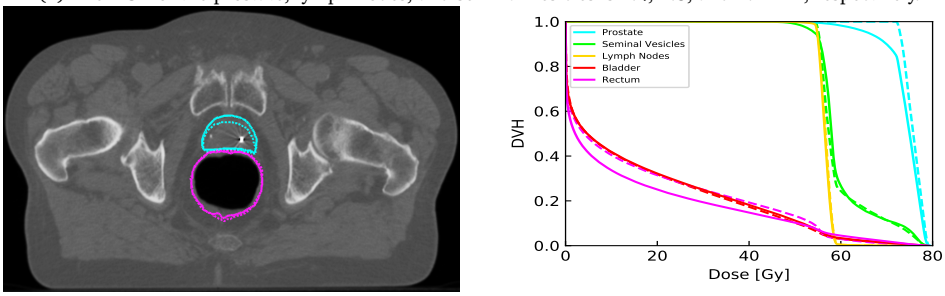
Since the LUMC dataset did not have any follow-up scans, we only consider the EMC and HMC datasets for evaluating the registration performance. Figure 2.6 shows example results of the automatically propagated contours. We compared the proposed method with the intensity-based registration approach of 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.



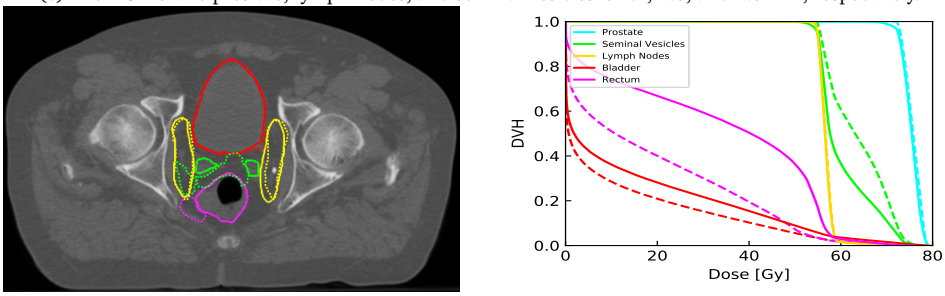
(a) The MSD of the prostate, lymph nodes, and seminal vesicles is 0.8, 1.6, and 1.0 mm, respectively.



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

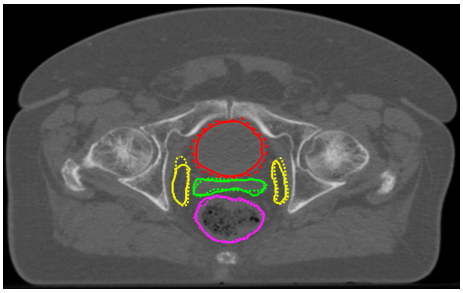
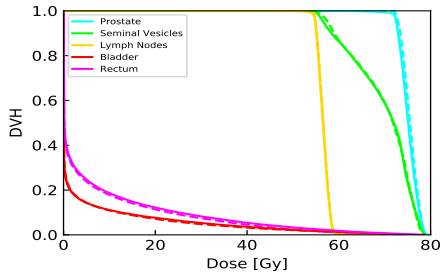


(c) The MSD of the prostate, lymph nodes, and seminal vesicles is 2.1, 1.6, and 1.5 mm, respectively.

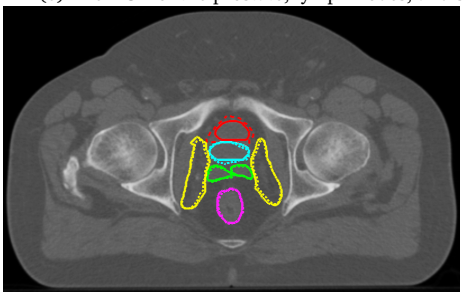
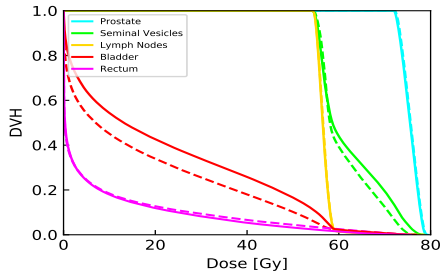


(d) The MSD of the prostate, lymph nodes, and seminal vesicles is 1.5, 1.6, and 11.0 mm, respectively.

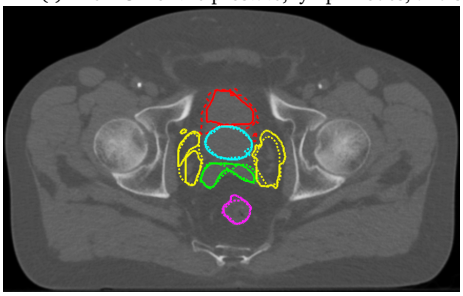
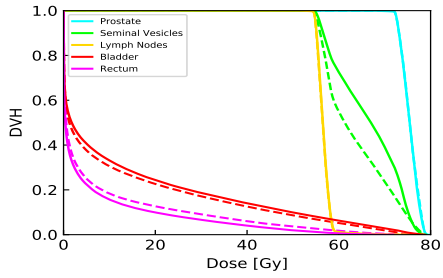
Figure 2.6: Examples from the automatic contours propagation of the HMC dataset and the corresponding dose volume histograms evaluated on the manual contours. The solid line represents the manual contouring results while the dotted line is the automatically propagated one.

(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.

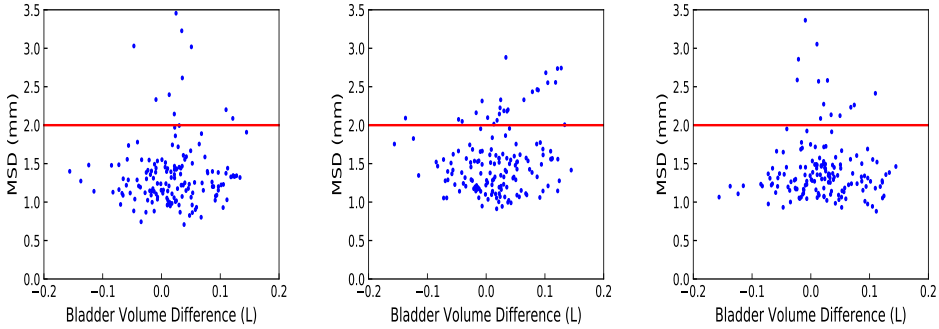



(g) The MSD of the prostate, lymph nodes, and seminal vesicles is 1.3, 1.1, and 1.1 mm, respectively.

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.

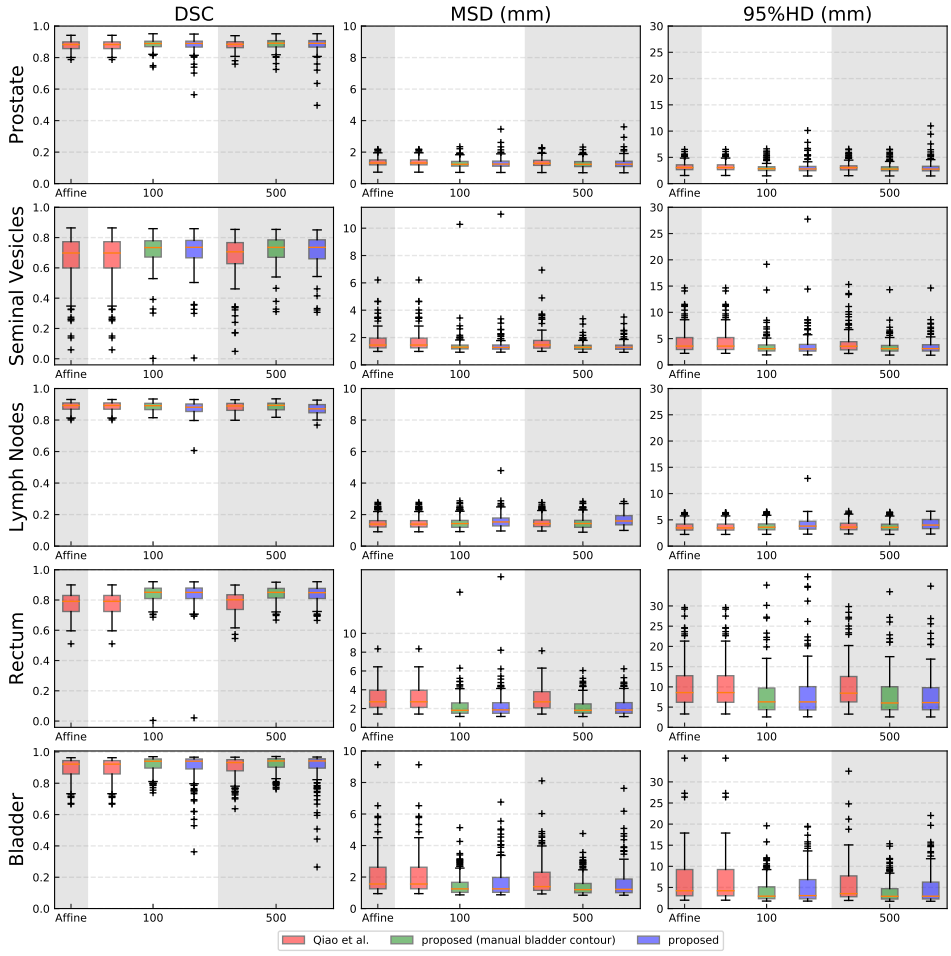


Figure 2.8: Boxplot comparison between Qiao *et al.* and the proposed algorithm for image registration on the HMC dataset versus the number of iterations. The columns show the DSC, MSD, and 95%HD from left to right. Prostate, seminal vesicles, lymph nodes, rectum, and bladder are shown from top to bottom 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

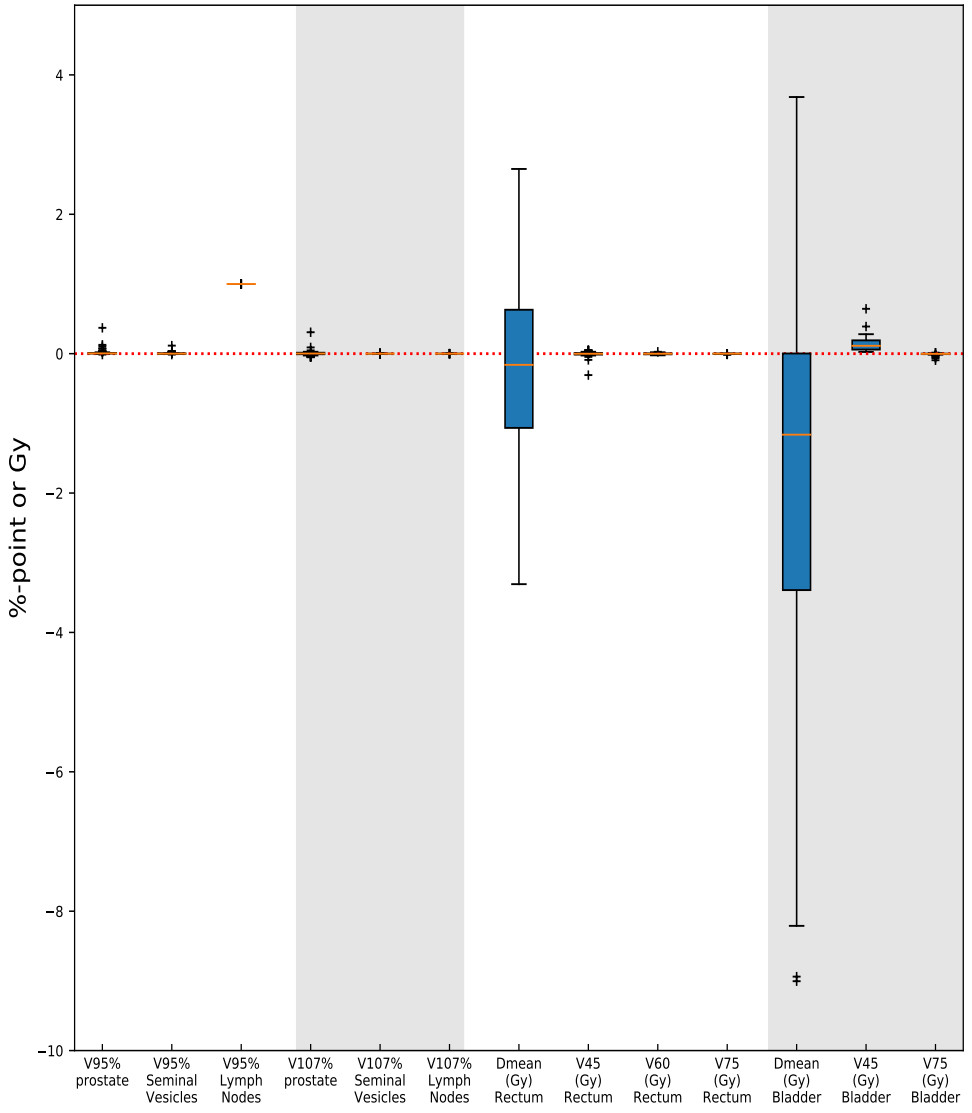


Figure 2.9: Boxplot depicting the difference in dosimetric parameters of the manual delineations, calculated by using either the treatment plan based on the automated delineations or the manual delineations for 99 scans of the HMC dataset.

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.

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 ± 0.43 mm, 2.38 ± 2.76 mm, 2.30 ± 0.96 mm, and 1.24 ± 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 ± 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 × 1.0 × 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 × 0.9 × 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 × 3 × 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 × 3 × 3 convolutions with a stride of one in each dimension, followed by a ReLu and an upsampling of 2 × 2 × 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 $\mathbf{base}^a$ and $\mathbf{base}^b$, 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, Φ the model adaptation process, (I_j, S_j) the image and segmentation pair at fraction j. Model adaptation is then performed as follows:

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

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 × 128 × 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 × 1 × 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†‡	SV†‡	Rectum†‡	Bladder†‡
		$\mu \pm \sigma$	$\mu \pm \sigma$	$\mu \pm \sigma$	$\mu \pm \sigma$
base^a		0.76 ± 0.06	0.55 ± 0.19	0.79 ± 0.08	0.86 ± 0.14
proposed^a	500	0.78 ± 0.06	0.61 ± 0.17	0.79 ± 0.07	0.89 ± 0.09
proposed^a	2000	0.80 ± 0.05	0.63 ± 0.17	0.80 ± 0.07	0.90 ± 0.07
proposed^a	5000	0.82 ± 0.05	$\mathbf{0.65 \pm 0.16}$	0.80 ± 0.07	$\mathbf{0.91 \pm 0.07}$
base^b		0.77 ± 0.07	0.46 ± 0.20	0.78 ± 0.08	0.87 ± 0.08
proposed^b	500	0.80 ± 0.05	0.50 ± 0.18	0.81 ± 0.06	$\mathbf{0.91 \pm 0.07}$
proposed^b	2000	$\mathbf{0.83 \pm 0.04}$	0.56 ± 0.15	$\mathbf{0.83 \pm 0.05}$	$\mathbf{0.91 \pm 0.06}$
proposed^b	5000	$\mathbf{0.83 \pm 0.04}$	0.55 ± 0.17	$\mathbf{0.83 \pm 0.05}$	$\mathbf{0.91 \pm 0.06}$

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

	#Itr	Prostate†‡	SV†‡	Rectum†‡	Bladder†‡
		$\mu \pm \sigma$	$\mu \pm \sigma$	$\mu \pm \sigma$	$\mu \pm \sigma$
base^a		2.70 ± 0.88	3.78 ± 4.76	2.77 ± 1.23	2.66 ± 9.46
proposed^a	500	2.38 ± 0.64	2.58 ± 2.67	2.76 ± 1.05	1.62 ± 1.28
proposed^a	2000	2.03 ± 0.55	$\mathbf{2.38 \pm 2.76}$	2.66 ± 1.03	1.48 ± 1.17
proposed^a	5000	1.78 ± 0.52	2.41 ± 3.17	2.59 ± 1.00	1.39 ± 1.12
base^a		2.29 ± 0.70	5.08 ± 5.33	2.78 ± 1.17	1.83 ± 0.96
proposed^b	500	1.95 ± 0.53	5.21 ± 16.71	2.46 ± 1.05	1.40 ± 1.03
proposed^b	2000	1.64 ± 0.43	3.18 ± 3.34	$\mathbf{2.30 \pm 0.96}$	$\mathbf{1.24 \pm 0.89}$
proposed^b	5000	$\mathbf{1.61 \pm 0.41}$	4.67 ± 16.96	$\mathbf{2.30 \pm 0.96}$	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^a and proposed^b. Here, † and ‡ represent a statistically significant difference between base^a vs proposed^a, and base^b vs proposed^b at 2000 iterations, respectively. In terms of DSC, MSD, and 95% HD, the larger base^b model is slightly better than base^a, 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†‡	SV†‡	Rectum	Bladder†‡
		$\mu \pm \sigma$	$\mu \pm \sigma$	$\mu \pm \sigma$	$\mu \pm \sigma$
base^a		9.1 ± 3.7	12.2 ± 10.7	14.4 ± 7.0	8.7 ± 12.9
proposed^a	500	7.9 ± 2.8	9.2 ± 8.4	14.7 ± 6.7	7.9 ± 6.7
proposed^a	2000	6.4 ± 2.1	$\mathbf{8.6 \pm 9.0}$	14.3 ± 6.7	7.2 ± 6.5
proposed^a	5000	5.9 ± 2.4	8.8 ± 10.6	14.0 ± 6.6	6.5 ± 6.1
base^b		6.9 ± 1.9	15.5 ± 9.7	13.9 ± 9.5	7.0 ± 4.5
proposed^b	500	6.1 ± 1.7	14.6 ± 20.4	13.5 ± 9.1	6.2 ± 5.8
proposed^b	2000	$\mathbf{5.2 \pm 1.4}$	11.9 ± 9.3	13.5 ± 8.4	$\mathbf{5.6 \pm 5.3}$
proposed^b	5000	$\mathbf{5.2 \pm 1.4}$	13.9 ± 20.9	$\mathbf{13.3 \pm 8.5}$	5.7 ± 5.6

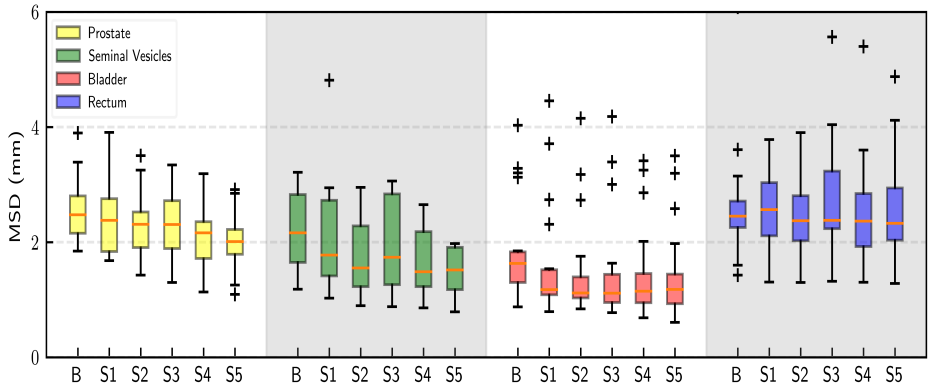


Figure 3.1: The effect of model adaptation over sessions in terms of MSD (mm). B is base model, S*i* after session i.

network for 5000 iterations seems as good as 2000 iterations as shown in the tables. Predicting a segmentation took ~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^b and proposed^b at 2000 iterations.

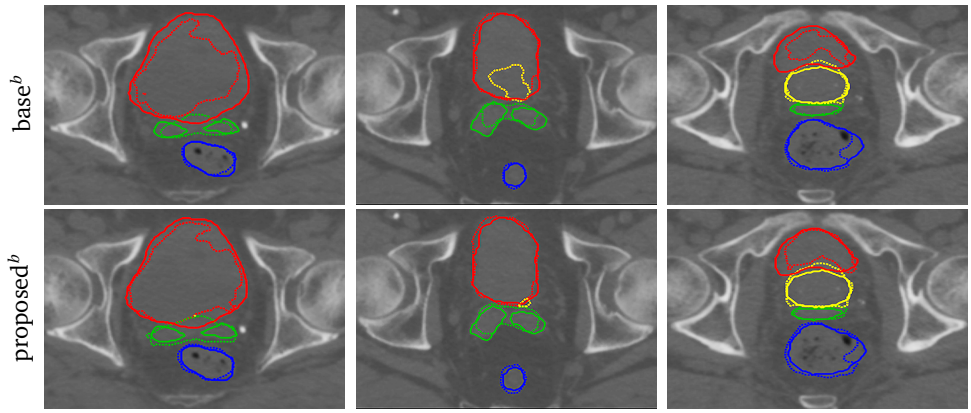


Figure 3.2: Three example results for the base^b model (top row) and the proposed^b at 2000 itr model (bottom row). The red, yellow, blue and green contours represent the bladder, prostate, rectum, and seminal vesicles, respectively. The solid line indicates manual contouring, while the dotted line the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

4

Adversarial Optimization for Joint Registration and Segmentation in Prostate CT Radiotherapy

This chapter was adapted from:

M Elmahdy, J Wolterink, H Sokooti, I Išgum, and M Staring. **Adversarial optimization for joint registration and segmentation in prostate CT radiotherapy**, *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, Pages 366-374, 2019.

Abstract

Joint image registration and segmentation has long been an active area of research in medical imaging. Here, we reformulate this problem in a deep learning setting using adversarial learning. We consider the case in which fixed and moving images as well as their segmentations are available for training, while segmentations are not available during testing; a common scenario in radiotherapy. The proposed framework consists of a 3D end-to-end generator network that estimates the deformation vector field (DVF) between fixed and moving images in an unsupervised fashion and applies this DVF to the moving image and its segmentation. A discriminator network is trained to evaluate how well the moving image and segmentation align with the fixed image and segmentation. The proposed network was trained and evaluated on follow-up prostate CT scans for image-guided radiotherapy, where the planning CT contours are propagated to the daily CT images using the estimated DVF. A quantitative comparison with conventional registration using `elastix` showed that the proposed method improved performance and substantially reduced computation time, thus enabling real-time contour propagation necessary for online-adaptive radiotherapy.

4.1 Introduction

Joint image registration and segmentation (JRS) has long been an active area of research in medical imaging. Image registration and segmentation are closely related and complimentary in applications such as contour propagation, disease monitoring, and data fusion from different modalities. Image registration could be enhanced and improved using an accurate segmentation, and vice versa registration algorithms could be used to improve image segmentation.

An important application in which coupling of image registration and segmentation is crucial, is online adaptive image-guided radiotherapy. In this application, clinically approved contours are propagated from an initial *planning* CT scan to *daily* inter-fraction CT scans of the same patient. Image registration can be used to correct for anatomical variations in shape and position of the underlying organs, as well as to compensate for any misalignment in patient setup. Ideally, contours should be propagated quickly to allow immediate computation of a new dose distribution. With these propagated contours, margins can be smaller and treatment-related complications may be reduced. Thus, it is important that the daily contours are of high quality, are consistent with the planning contours, and are generated in near real-time.

In the last decade, researchers have been working on fusing image registration and segmentation. Lu *et al.* [81] proposed a Bayesian framework for modelling segmentation and registration such that these could alternatingly constrain each other. Yezzi *et al.* [82] proposed using active contours to register and segment images. Unal *et al.* [83], generalizing on [82], proposed to use partial differential equations without any shape prior. Most of these methods require long computation times and complex parameter tuning. Recently, the widespread adoption of deep learning techniques has led to remarkable achievements in the field of medical imaging [53]. Among these techniques are generative adversarial networks (GANs), which are defined by joint optimization of a generator and discriminator network [54]. GANs have boosted the performance of traditional networks for image segmentation [84] as well as registration [85]. Recently, Mahapatra *et al.* [86] proposed a GAN for joint registration and segmentation of 2D chest X-ray images. However, this method requires reference deformation vector fields (DVFs) for training. In practice, these are often unavailable and it may be more practical to perform unsupervised registration [87], i.e. training without reference DVFs.

In this paper, we introduce a fast unsupervised 3D GAN to jointly perform deformable image registration and segmentation. A generator network estimates the DVF between two images, while a discriminator network is trained simultaneously to evaluate the quality of the registration and the segmentation and propagate the

feedback to the generator network. We consider the use-case in which fixed and moving images as well as their segmentations are available for training, which is a common scenario in radiation therapy. However, no segmentations are required for DVF estimation during testing. This paper has the following contributions. First, we propose an end-to-end 3D network architecture, which is trained in an adversarial manner for joint image registration and segmentation. Second, we propose a strategy to generate well-aligned pairs to train the discriminator network with. Third, we leverage PatchGAN as a local quality measure of image alignment. Fourth, the proposed network is much faster and more accurate than conventional registration methods. We quantitatively evaluate the proposed method on a prostate CT database, which shows that the method compares favorably to `elastix` software [62].

4.2 Methods

Image registration is the transformation of a moving image I_m to the coordinate system of a fixed image I_f. In this paper, we assume that all image pairs are affinely registered beforehand, and we focus on local non-linear deformations. In conventional contour propagation algorithms, registration and segmentation are disjoint. First, the DVF Φ is estimated using image registration, and then Φ is used to warp the contours S_m to the fixed coordinate space. Afterwards, during system evaluation, a similarity measure such as the Dice similarity coefficient (DSC) can be used to measure the quality of the propagated contours w.r.t. ground truth contours, but this information is not fed back to the registration algorithm. We call this an *open loop* system. In contrast, this paper proposes an end-to-end *closed loop* system to improve image registration based on feedback on the registration as well as the segmentation quality.

4.2.1 Adversarial training

We propose to train a GAN containing two CNNs: a generator network that predicts the DVF Φ given I_f and I_m, and a discriminator network that assesses the alignment of $I_f(\boldsymbol{x})$ and $I_m(\Phi(\boldsymbol{x}))$ as well as the overlap between $S_f(\boldsymbol{x})$ and $S_m(\Phi(\boldsymbol{x}))$. Hence, we assume that S_f and S_m are both available, but during training only. The GAN is trained using a Wasserstein objective [88], which has empirically been shown to improve training stability and convergence compared to the GAN objective in [54]. Equations (4.1) and (4.2) list the generator loss L_G^{GAN} and the discriminator loss L_D^{GAN} of WGAN:

$$L_G^{GAN} = E\left[D(I_f(\boldsymbol{x}), I_m(\Phi(\boldsymbol{x})), S_m(\Phi(\boldsymbol{x})))\right], \tag{4.1}$$

$$L_D^{GAN} = E\left[D(I_f(\boldsymbol{x}), I_m(\Phi(\boldsymbol{x})), S_m(\Phi(\boldsymbol{x})))\right] - \left[D(I_f, \Theta(I_f), S_f)\right], \tag{4.2}$$

where G and D denote the generator and discriminator networks with trainable parameters and Φ is the DVF provided by G. In a GAN, the discriminator is trained to distinguish between *real* and *fake* samples. In this case, fake samples are the triple

$(I_f, I_m(\Phi), S_m(\Phi))$, while real samples should be well-aligned images. As we perform unsupervised registration, and assume no knowledge about the ideal alignment of two images, we synthesize such image based on the fixed image and its segmentation alone: $(I_f, \Theta(I_f), S_f)$. Hence, Θ in Equation (4.2) is a random combination of disturbance functions, as follows. First, to mimic imaging noise, Gaussian noise and Gaussian smoothing are added with zero mean and a standard deviation of 0.04. Second, to mimic contrast variations, we apply gamma correction with a random gamma factor in the range $[-0.4, 0.4]$. Third, we mimic interpolation errors by applying a random deformation of less than 0.5 mm and resample the images using that deformation using linear interpolation.

In addition to these image-based quality measures, we include the segmentation of the deformed moving image as input to the discriminator in order to enforce DVFs that are consistent with the moving segmentation. We test two designs. The first design concatenates the segmentation as a third input channel in the discriminator, next to the fixed and moving image channels. The second design multiplies the fixed and moving image channel with the corresponding segmentation, so that the network learns to focus on the target structures and organs-at-risk instead of on the bowels and other less relevant soft tissue. These designs are named JRS-GAN^a and JRS-GAN^b, respectively.

We found that training the network using WGAN loss only, resulted in slow convergence and suboptimal registrations. Thus, a similarity loss L_{sim}, based on image similarity and segmentation overlap, was added to the generator:

$$L_{sim} = (1 - \mathrm{DSC}(S_m(\Phi(\boldsymbol{x})), S_f(\boldsymbol{x}))) + (1 - \mathrm{NCC}(I_m(\Phi(\boldsymbol{x})), I_f(\boldsymbol{x}))), \tag{4.3}$$

where DSC is the Dice similarity coefficient and NCC is normalized cross-correlation. Adding the DSC to L_{sim} ensures that the registration improves the segmentation and vice versa. Furthermore, to ensure smooth and continuous DVFs, the bending energy penalty of the DVF, L_{smooth}, was added as a regularization term to the overall generator loss, which was defined as:

$$L_G = L_{sim} + \lambda_1 L_{smooth} + \lambda_2 L_G^{GAN}, \tag{4.4}$$

where λ_1 and λ_2 are weights for the DVF smoothness and the generator loss.

During training of the network, for every iteration of the generator we used 100 iterations of the discriminator, for the first 25 iterations. After that we used the ratio 1:5. In each iteration, weights of the discriminator were clipped to the range $[-0.01, 0.01]$ [88].

4.2.2 Network architectures

Generator Network To estimate the parametric mapping function Φ between the fixed and moving images we use a 3D network similar to the U-net [89]. Figure 4.1 shows

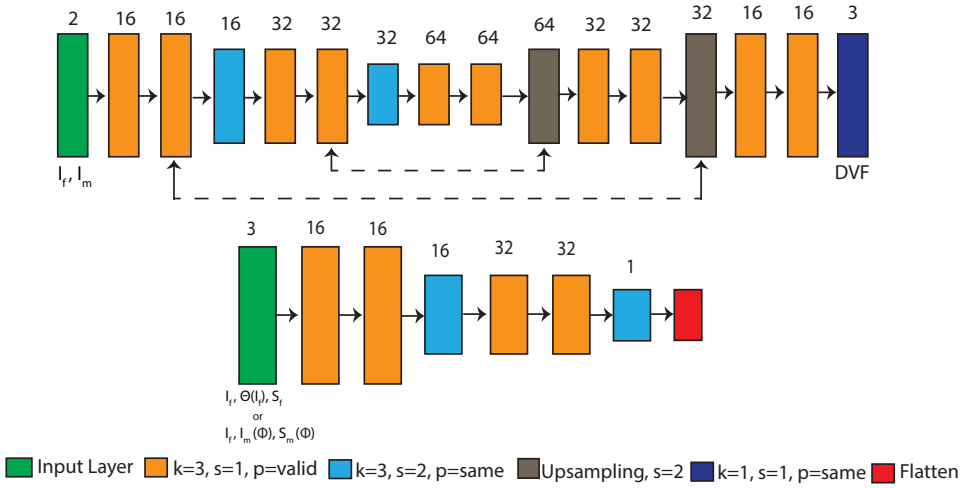


Figure 4.1: The proposed generator (top) and discriminator (bottom) networks, where k, s, and p represent the kernel size, stride size, and padding option, respectively. The numbers above the different layers represent the feature maps.

the network design in more detail. The input to the network is the concatenation of I_f and I_m. The network encodes the image pairs through a set of 3×3×3 convolution layers followed by LeakyReLU and batch normalization layers. Strided convolutions are used in the contractive path and upsampling layers are used in the expanding path. The output size of the network is smaller than the input size in order to consider a larger field of view. A resampling network adopted from NiftyNet [90] is used to warp the images using the estimated DVF during training time so that the network can be trained end-to-end.

Discriminator Network The discriminator is responsible for assessing whether the image pairs are well-aligned or not, as well as assessing whether the segmentations overlap. Figure 4.1 shows the network design, which is similar to the contracting path of the generator. The discriminator network was trained using PatchGAN [91]. Hence, instead of representing the quality of the whole patch with a single number, the network could quantify the sub-patch quality locally.

4.3 Experiments and results

4.3.1 Dataset, evaluation criteria and implementation details

This study includes eighteen patients who underwent intensity-modulated radiation therapy for prostate cancer in 2007 at Haukeland university hospital [67]. Each patient had a planning CT as well as 7 to 10 inter-fraction repeat CT scans. The prostate, lymph nodes, seminal vesicles, as well as the rectum and bladder were annotated.

Table 4.1: MSD (mm) values for different experiments, where † and ‡ represent a significant difference compared to `elastix`-MI and Reg-CNN, respectively.

	Prostate	Seminal vesicles	Lymph nodes	Rectum	Bladder
Evaluation	$\mu \pm \sigma$	$\mu \pm \sigma$	$\mu \pm \sigma$	$\mu \pm \sigma$	$\mu \pm \sigma$
`elastix`-NCC	1.81 ± 0.7	2.80 ± 1.6	1.19 ± 0.4	3.79 ± 1.2	5.31 ± 2.6
`elastix`-MI	1.73 ± 0.7	2.70 ± 1.6	1.18 ± 0.4	3.68 ± 1.2	5.26 ± 2.6
Reg-CNN	$1.44 \pm 0.5^{\dagger}$	$2.09 \pm 1.7^{\dagger}$	1.22 ± 0.3	$2.59 \pm 1.3^{\dagger}$	$4.18 \pm 2.6^{\dagger}$
JRS-CNN	$1.18 \pm 0.4^{\dagger\ddagger}$	$1.91 \pm 1.6^{\dagger\ddagger}$	$1.02 \pm 0.3^{\dagger\ddagger}$	$2.32 \pm 1.3^{\dagger\ddagger}$	$2.37 \pm 2.0^{\dagger\ddagger}$
Reg-GAN	$1.40 \pm 0.5^{\dagger}$	$2.14 \pm 1.7^{\dagger}$	$1.06 \pm 0.3^{\dagger\ddagger}$	$2.72 \pm 1.3^{\dagger}$	$4.31 \pm 2.8^{\dagger}$
JRS-GAN^{a}	$\mathbf{1.13 \pm 0.4}^{\dagger\ddagger}$	$\mathbf{1.81 \pm 1.6}^{\dagger\ddagger}$	$\mathbf{1.00 \pm 0.3}^{\dagger\ddagger}$	$\mathbf{2.21 \pm 1.3}^{\dagger\ddagger}$	$\mathbf{2.29 \pm 2.0}^{\dagger\ddagger}$
JRS-GAN^{b}	$1.17 \pm 0.4^{\dagger\ddagger}$	$1.90 \pm 1.5^{\dagger\ddagger}$	$1.01 \pm 0.3^{\dagger\ddagger}$	$2.34 \pm 1.3^{\dagger\ddagger}$	$2.41 \pm 2.1^{\dagger\ddagger}$

Each scan has 90 to 180 slices with a slice thickness of around 2 to 3 mm. All the slices were of size 512 × 512 with an in-plane resolution of around 0.9 mm. All the volumes were affinely registered using `elastix`. The volumes were resampled to isotropic voxel size of 1×1×1 mm. All volumes intensities were scaled to [-1, 1]. We split the dataset into 111 image pairs (from 12 patients) for training and validation and 50 image pairs (6 patients) for testing.

The quality of registration is quantified geometrically in 3D by comparing the manual delineations of the daily CT with the automatically propagated contours. We use the mean surface distance (MSD), and the 95% Hausdorff distance (HD). A Wilcoxon signed rank test at $p = 0.05$ is used to compare results.

The networks were implemented using TensorFlow (version 1.13) [73] with the RMSProp optimizer using a learning rate of 10^{-5}. The networks were trained and tested on an NVIDIA Tesla V100 GPU with 16 GB of memory. From each image pair, 1000 patches of size 96×96×96 voxels were sampled within the torso mask. To improve stability, the network was trained to warp the fixed patch to the moving patch and vice versa at the same training iteration. The magnitude of the three loss terms in Equation (4.3) was scaled by setting $\lambda_1 = 1$ and $\lambda_2 = 0.01$.

4.3.2 Experiments and results

Tables 4.1 and 4.2 provide quantitative results comparing the following methods. First, we include conventional iterative methods using `elastix` software [62] with NCC (`elastix`-NCC) and MI (`elastix`-MI) similarity measures, using the settings from [42]. Second, we evaluate two unsupervised deep learning-based methods without adversarial feedback: One uses the generator trained with the NCC loss (Reg-CNN), similar to [87]; the other uses the generator with both the NCC and DSC loss (JRS-CNN). Third, we evaluate several versions of our GAN-based approach. To study the effect of adversarial training without added segmentations, we perform an experiment named Reg-GAN. Finally, we evaluate the proposed JRS-GAN^{a} and

Table 4.2: %95HD (mm) values for different experiments, where † and ‡ represent a significant difference compared to `elastix`-MI and Reg-CNN, respectively.

	Prostate	Seminal vesicles	Lymph nodes	Rectum	Bladder
Evaluation	$\mu \pm \sigma$	$\mu \pm \sigma$	$\mu \pm \sigma$	$\mu \pm \sigma$	$\mu \pm \sigma$
`elastix`-NCC	4.2 ± 1.8	6.1 ± 3.3	$\mathbf{2.8 \pm 1.0}^{\ddagger}$	11.0 ± 5.2	$15.4 \pm 8.4^{\ddagger}$
`elastix`-MI	4.0 ± 1.7	6.0 ± 3.7	$2.8 \pm 1.0^{\ddagger}$	10.9 ± 5.2	$15.3 \pm 8.3^{\ddagger}$
Reg-CNN	5.3 ± 2.5	6.2 ± 3.5	4.4 ± 1.4	11.0 ± 6.5	16.6 ± 9.3
JRS-CNN	$3.6 \pm 1.5^{\dagger\ddagger}$	$5.4 \pm 3.4^{\dagger\ddagger}$	$3.1 \pm 0.9\ddagger$	$10.3 \pm 6.7^{\dagger\ddagger}$	$11.6 \pm 10.5^{\dagger\ddagger}$
Reg-GAN	$4.3 \pm 2.1\ddagger$	6.0 ± 3.6	$3.4 \pm 1.0\ddagger$	11.1 ± 6.4	$16.2 \pm 9.6\ddagger$
JRS-GAN^a	$\mathbf{3.4 \pm 1.4}^{\dagger\ddagger}$	$\mathbf{5.3 \pm 3.3}^{\dagger\ddagger}$	$3.1 \pm 0.9\ddagger$	$\mathbf{10.0 \pm 6.7}^{\dagger\ddagger}$	$\mathbf{11.0 \pm 10.3}^{\dagger\ddagger}$
JRS-GAN^b	$3.5 \pm 1.4^{\dagger\ddagger}$	$5.6 \pm 3.7\ddagger$	$3.0 \pm 1.0\ddagger$	$10.5 \pm 6.8^{\dagger\ddagger}$	$11.4 \pm 10.6^{\dagger\ddagger}$

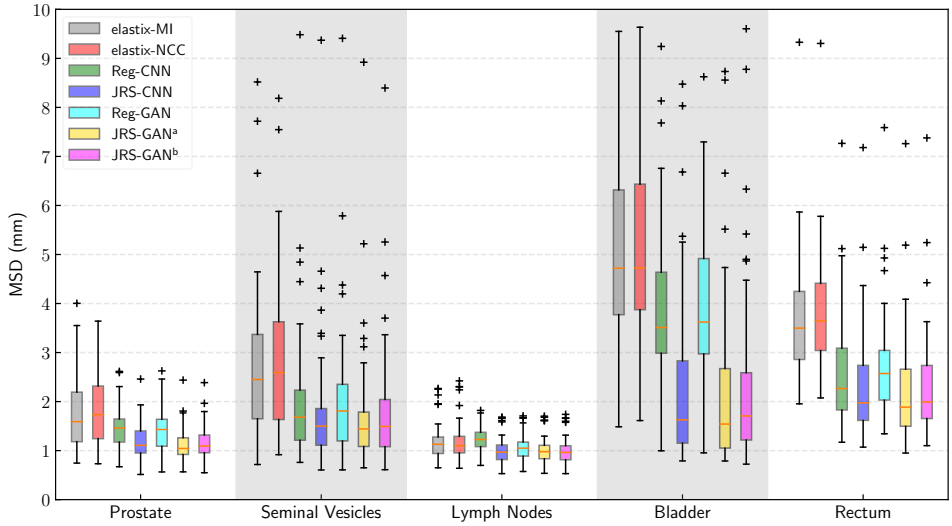


Figure 4.2: Boxplots for the evaluated methods in terms of MSD (mm).

JRS-GAN^b methods.

The MSD values in Table 2.5 show that for all organs, the GAN-based methods significantly improved over `elastix`. This is further shown in Figure 4.2. The results indicate a significant improvement when performing joint registration and segmentation instead of disjoint registration. Furthermore, the boxplot indicates that performance for JRS-GAN^a and JRS-GAN^b was very similar. Similarly, the 95% HD values in Table 2.6 show improvements in contour accuracy when the GAN-based method is used. Especially the organs-at-risk showed large improvements. The standard deviations of the Jacobian determinant of the estimated DVFs were 0.08 ± 0.01 and 0.17 ± 0.04 for `elastix`-MI and JRS-GAN^a, respectively. The average runtime for the proposed pipeline is 0.6 seconds on the GPU for a volume of size 256^3 voxels,

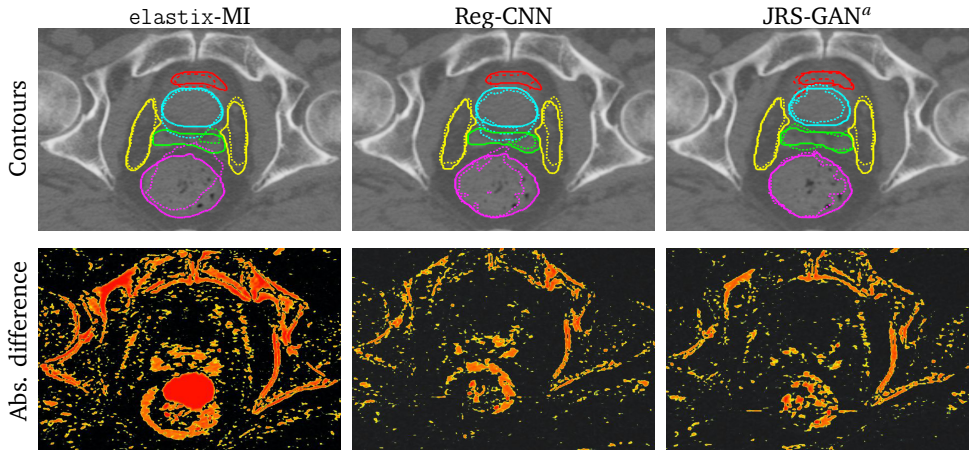


Figure 4.3: An example result for three of the methods. Top row shows the fixed image with propagated contours (solid line is manual; dotted is automatic result). The red, yellow, cyan, violet, and green contours represent the bladder, lymph nodes, prostate, rectum, and seminal vesicles, respectively. Bottom row shows heatmaps of absolute difference images between fixed and deformed moving image.

while the average runtime of `elastix` at 100 iterations is 13 seconds per volume on an Intel Xeon E51620 CPU using 4 cores. Figure 4.3 illustrates the segmentation and registration for an example case.

4.4 Discussion and conclusion

In this study, we investigated the performance of an end-to-end joint registration and segmentation network for adaptive image-guided radiotherapy. Unlike conventional registration methods, our network encodes and learns the most relevant features for joint image registration and segmentation, and exploits the combined knowledge on unseen images without segmentations.

We demonstrate that including the segmentation during training boosts the system's performance by a margin. Furthermore, adversarial feedback had a small benefit on performance, when comparing Reg-CNN with Reg-GAN. Results indicate a noticeable benefit of including segmentation masks as input to the discriminator during training. How exactly segmentation masks were embedded during training was less relevant, with only small differences observed for the seminal vesicles. This could be due to the small size and irregular nature of the seminal vesicles. A key advantage of the proposed deep learning-based contour propagation method is its runtime on new and unseen data, i.e. 0.6 s.

This work has shown that adversarial feedback can help improve registration, i.e. that a discriminator can learn a measure of image alignment. This is a promising

aspect that could be further explored in future work. This will include improved GAN objectives, such as the use of gradient penalty regularization.

To conclude, we have proposed a 3D adversarial network for joint image registration and segmentation with a focus on prostate CT radiotherapy. The proposed method demonstrated the effectiveness of training the registration and segmentation jointly. Moreover, it showed a substantial reduction in the computation time making it a strong candidate for online adaptive image-guided radiotherapy of prostate cancer. Since the proposed method did not only improve accuracy for the target areas, but substantially so for the organs-at-risk, this may aid reducing treatment-induced complications.

5

Joint Registration and Segmentation via Multi-Task Learning for Adaptive Radiotherapy of Prostate Cancer

This chapter was adapted from:

M Elmahdy, L Beljaards, S Yousefi, H Sokooti, F Verbeek, U van der Heide, and M Staring. **Joint Registration and Segmentation via Multi-Task Learning for Adaptive Radiotherapy of Prostate Cancer**, *IEEE Access*, 2021.

Abstract

Medical image registration and segmentation are two of the most frequent tasks in medical image analysis. As these tasks are complementary and correlated, it would be beneficial to apply them simultaneously in a joint manner. In this paper, we formulate registration and segmentation as a joint problem via a Multi-Task Learning (MTL) setting, allowing these tasks to leverage their strengths and mitigate their weaknesses through the sharing of beneficial information. We propose to merge these tasks not only on the loss level, but on the architectural level as well. We studied this approach in the context of adaptive image-guided radiotherapy for prostate cancer, where planning and follow-up CT images as well as their corresponding contours are available for training. At testing time the contours of the follow-up scans are not available, which is a common scenario in adaptive radiotherapy. The study involves two datasets from different manufacturers and institutes. The first dataset was divided into training (12 patients) and validation (6 patients), and was used to optimize and validate the methodology, while the second dataset (14 patients) was used as an independent test set. We carried out an extensive quantitative comparison between the quality of the automatically generated contours from different network architectures as well as loss weighting methods. Moreover, we evaluated the quality of the generated deformation vector field (DVF). We show that MTL algorithms outperform their Single-Task Learning (STL) counterparts and achieve better generalization on the independent test set. The best algorithm achieved a mean surface distance of 1.06 ± 0.3 mm, 1.27 ± 0.4 mm, 0.91 ± 0.4 mm, and 1.76 ± 0.8 mm on the validation set for the prostate, seminal vesicles, bladder, and rectum, respectively. The high accuracy of the proposed method combined with the fast inference speed, makes it a promising method for automatic re-contouring of follow-up scans for adaptive radiotherapy, potentially reducing treatment related complications and therefore improving patients quality-of-life after treatment. The source code is available at https://github.com/moelmahdy/JRS-MTL.

5.1 Introduction

Medical image analysis aims to extract clinically useful information that aids the diagnosis, prognosis, monitoring and treatment of diseases [92, 93]. Two of the most common tasks in such analyses are image registration and segmentation [94]. Image segmentation aims to identify and cluster objects that prevail similar characteristics into distinctive labels, where these labels can be used for diagnosis or treatment planning. Image registration is the task of finding the geometrical correspondence between images that were acquired at different time steps or from different imaging modalities. These two tasks are complementary, as for example image atlases warped by image registration algorithms are often used for image segmentation [21, 22], while image contours can be used to guide the image registration method in addition to the intensity images [23, 17, 24]. Contours are also used for evaluating the quality of the registration [25, 26]. However, each of these tasks has its own strengths and weaknesses. For instance, image segmentation algorithms can directly delineate images based on texture and surrounding anatomy, and may therefore be robust to large organ deformations. However it sometimes has difficulties with low contrast areas and irregularly shaped organs. On the other hand, image registration algorithms have the ability to encode prior knowledge of the patient's anatomy and therefore may perform better on low quality images. However, such methods sometimes have difficulty with large deformations. Therefore, coupling of image registration and segmentation tasks and modeling them in a single network could leverage their strengths and mitigate their weaknesses through the sharing of beneficial information.

Adaptive image-guided radiotherapy is an exemplar application where the coupling of image registration and segmentation is vital. In radiotherapy, treatment radiation dose is delivered over a course of multiple inter-fraction sessions. In an adaptive setting, re-imaging of the daily anatomy and automatic re-contouring is crucial to compensate for patient misalignment, to compensate for anatomical variations in organ shape and position, and an enabler for the reduction of treatment margins or robustness settings [95, 96]. These have an important influence on the accuracy of the dose delivery, and improve the treatment quality, potentially reducing treatment related side-effects and increasing quality-of-life after treatment [97]. Automatic contouring can be done by direct segmentation of the daily scan, or by registration of the annotated planning scan with the daily scan followed by contour propagation. Image registration has the advantage of leveraging prior knowledge from the initial planning CT scan and the corresponding clinical-quality delineations, which may especially be helpful for challenging organs. On the other hand, image segmentation methods may better delineate organs that vary substantially in shape and volume between treatment fractions, which is often the case for the rectum and the bladder.

In this study, we propose to fuse these tasks at the network architecture level as well as via the loss function. Our key contributions in this paper are as follows:

1. We formulate image registration and segmentation as a multi-task learning problem, which we explore in the context of adaptive image-guided radiotherapy.

2. We explore different joint network architectures as well as loss weighting methods for merging these tasks.

3. We adopt the cross-stitch network architecture for segmentation and registration tasks and explore how these cross-stitch units facilitate information flow between these tasks.

4. Furthermore, we compare MTL algorithms against single-task networks. We demonstrate that MTL algorithms outperform STL networks for both segmentation and registration tasks. To the best of our knowledge this is the first study to investigate various MTL algorithms on an architectural level as well as on a loss weighing level for joint registration and segmentation tasks.

5. We thoroughly investigate the internals of the STL and MTL networks and pinpoint the best strategy to merge this information to maximize the information flow between the two tasks.

Initial results of this work were presented in [98], focusing on the cross-stitch unit in a proposed joint architecture. In the current paper we extend this study to the architectural fusion of these tasks as well as different loss weighting mechanisms. Moreover, an extensive analysis of the different methodologies was performed, detailing the effect of architectural choices, information flow between the two tasks, etc.

The remainder of this paper is organized as follows: Section 5.2 introduces single-task networks, multi-task networks, and loss weighting approaches. In Section 5.3 we introduce the datasets and details about the implementation as well as the experiments. In Sections 5.5 and 5.6, we discuss our results, provide future research directions, and present our conclusions.

5.1.1 Related work

In the last decade, researchers have been exploring the idea of fusing image segmentation and registration. Lu *et al.* [99] and Pohl *et al.* [100] proposed modeling these tasks using a Bayesian framework such that these tasks would constrain each other. Yezzi [101] proposed to fuse these tasks using active contours, while Unal *et al.* [102] proposed to generalize the previous approach by using partial differential equations without shape priors. Mahapatra *et al.* [24] proposed a Joint Registration and Segmentation (JRS) framework for cardiac perfusion images, where the temporal

intensity images are decomposed into sparse and low rank components corresponding to the intensity change from the contrast agent and the motion, respectively. They proposed to use the sparse component for segmentation and the low rank component for registration. However, most of the aforementioned methods require complex parameter tuning and yield long computation times.

Recently, deep learning-based networks have shown unprecedented success in many fields especially in the medical image analysis domain [20, 103, 104, 105, 106, 13], where deep learning models perform on par with medical experts or even surpassing them in some tasks [107, 108, 109, 110]. Several deep learning-based approaches have been proposed for joint registration and segmentation. The joining mechanisms in the literature can be classified in two categories, namely joining via the loss function and via the architecture as well as the loss function. Selected exemplar methods of the first approach are Hue *et al.* [111], who proposed to join segmentation and registration via a multi-resolution Dice loss function. Elmahdy *et al.* [23] proposed a framework that is a hybrid between learning and iterative approaches, where a CNN network segments the bladder and feeds it to an iterative-based registration algorithm. The authors integrated domain-specific knowledge such as air pocket inpainting as well as contrast clipping, moreover they added an extra registration step in order to focus on the seminal vesicles and rectum. Elmahdy *et al.* [17] and Mahapatra *et al.* [112] proposed a GAN-based (Generative Adversarial Network) approach, where a generative network predicts the correspondence between a pair of images and a discriminator network for giving feedback on the quality of the deformed contours. Exemplar methods of the second category are Xu *et al.* [113], who presented a framework that simultaneously trains a registration and a segmentation network. The authors proposed to jointly learn these tasks during training, however the networks can be used independently during test time. This enables prediction of only the registration output, when the labels are not available during test time. Estienne *et al.* [114] proposed to merge affine and deformable registration as well as segmentation in a 3D end-to-end CNN network. Recently Liu *et al.* [115] proposed an end-to-end framework called JSSR that registers and segments multi-modal images. This framework is composed of three networks: a generator network, that synthesizes the moving image to match the modality of the fixed image, a registration network that registers the synthesized image to the fixed image, and finally a segmentation network that segments the fixed, moving, and synthesized images.

All the previous methods explored the idea of joining segmentation and registration, where to the best of our knowledge none have explored how these tasks are best connected and how to optimize the information flow between them on both the loss and architectural levels.

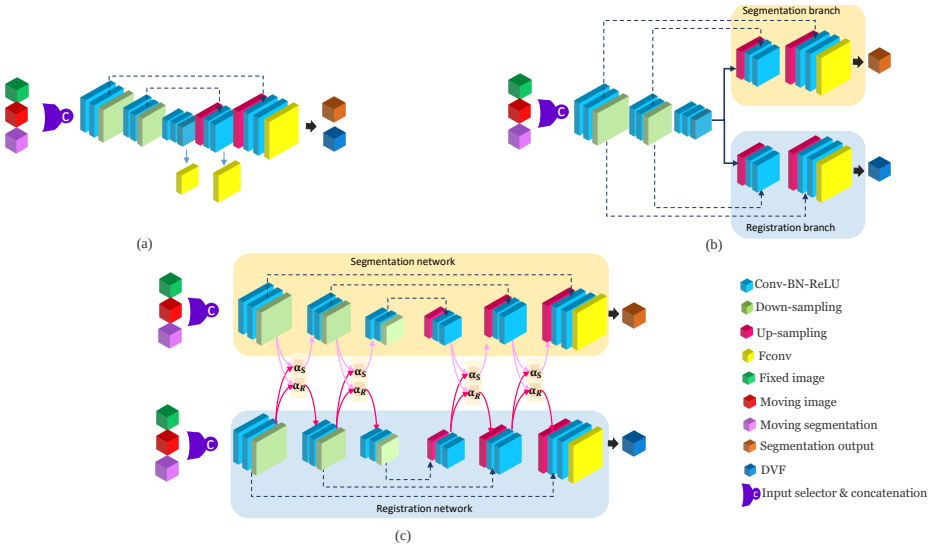


Figure 5.1: The proposed network architectures introduced in the paper. (a) is the base STL network architecture for either segmentation or registration, but also represents the dense parameter sharing MTL network architecture; (b) is the architecture with a shared encoder, while (c) is the Cross-stitch network architecture. Details about the number of feature maps are presented in Section 5.3.2.

5.2 Methods

5.2.1 Base network architecture

The base architecture for the networks in this paper is a 3D CNN network inspired by the U-Net and BIRNet architectures [116, 117]. Figure 5.1a shows the architecture of the base network. The network encodes the input through $3 \times 3 \times 3$ convolution layers with no padding. LeakyReLU [118] and batch normalization [119] are applied after each convolutional layer. We used strided convolutions in the down-sampling path and trilinear upsampling layers in the upsampling path. Through the upsampling path, the number of feature maps increases while the size of the feature maps decreases, and vice versa for the down-sampling path. The network has three output resolutions and is deeply supervised at each resolution. Each resolution is preceded by a $1 \times 1 \times 1$ fully convolution layer (Fconv) so that at coarse resolution, the network can focus on large organs as well as large deformations, while vice versa at fine resolution. In order to extract the groundtruth for different resolutions, we perform cropping of different sizes as well as strided sampling so that for every input patch of size n^3, the sizes of the coarse, mid, and fine resolution are $(\frac{n}{4} - 7)^3$, $(\frac{n}{2} - 18)^3$, and $(n - 40)^3$, respectively.

5.2.2 Single task learning

Single-task networks are designed to solve one task and therefore require a large amount of labeled training samples, which are scarce in the medical domain since it takes time and trained medical personnel to contour these images. The segmentation and registration networks have the same architecture as the base network depicted in Figure 5.1a, but differ in the input and output layers. Here, single-task networks are considered baseline networks for comparing with the performance of the proposed multi-task networks.

5.2.2.1 Segmentation network

The input to the segmentation network is the daily CT scan, referred to as the fixed image I_f, where the network predicts the corresponding segmentation S_f^{pred}. S_f^{pred} represents the probability maps for the background, target organs, and organs-at-risk. The network was trained using the Dice Similarity Coefficient (DSC) loss, which quantifies the overlap between the network prediction S_f^{pred} and the groundtruth S_f as follows:

$$\mathscr{L}_{\text{DSC}} = 1 - \frac{1}{K}\sum_{k=1}^{K}\frac{2 * \sum_x S_k^{\text{pred}}(x) \cdot S_k(x)}{\sum_x S_k^{\text{pred}}(x) + \sum_x S_k(x)}, \tag{5.1}$$

where K is the number of structures to be segmented, x is the voxel coordinate, S_k is the ground truth segmentation, and S_k^{pred} the predicted probabilities. The network has 779,436 trainable parameters.

5.2.2.2 Registration network

The input to the registration network is the concatenation of the planning scan, referred to as the moving image $I_m$and the daily scan I_f. The network predicts the geometrical correspondence between the input images. This correspondence is represented by the displacement vector field (DVF), referred to as ϕ^{pred}. This DVF is then used to warp I_m. In an ideal scenario, the warped moving image I_m^{warped} would be identical to I_f. The network is trained using Normalized Cross Correlation (NCC) in order to quantify the dissimilarity between I_m^{warped} and I_f. Since the images are from a single imaging modality (CT) with a similar intensity distribution, NCC is an obvious choice abundantly used in the registration literature. Moreover, the implementation is straightforward and efficient when using plain convolution operations. NCC is defined by the following equation:

$$\mathscr{L}_{\text{NCC}} = 1 - \frac{\sum_x [(I_f(x) - \overline{I_f}) \cdot (I_m^{\text{warped}}(x) - \overline{I_m^{\text{warped}}})]}{\sigma_{I_f}\sigma_{I_m^{\text{warped}}}}, \tag{5.2}$$

where x is the voxel coordinate, and σ_{I_f} and $\sigma_{I_m^{\text{warped}}}$ are the standard deviation of the fixed and warped images, respectively. In order to encourage the network to predict a

smooth DVF, a bending energy penalty term is added for regularization:

$$\mathscr{L}_{\text{BE}} = \frac{1}{N} \sum_{x} \| H(\phi^{pred}(x)) \|_2^2, \tag{5.3}$$

where H is the Hessian matrix. Now the total registration loss becomes:

$$\mathscr{L}_{\text{Registration}} = \mathscr{L}_{\text{NCC}} + w \cdot \mathscr{L}_{\text{BE}}, \tag{5.4}$$

where w is the bending energy weight. For more details on the selection of w, see Section 5.4.1. The network has 779,733 trainable parameters.

5.2.3 Multi task learning

In Multi-Task Learning (MTL), related tasks regularize each other by introducing an inductive bias, thus making the model agnostic to overfitting compared to its STL counterparts [120]. MTL can also be considered as an implicit data augmentation strategy, since it effectively increases the training sample size while encouraging the model to ignore data-dependent noise. Because different tasks have different noise patterns, modeling these tasks simultaneously enables the model to generalize well [121]. Moreover, in MTL models, some features can be more easily learned by one task than another, thus encouraging information cross-talk between tasks [122].

Also, in real-world scenarios, physicians usually incorporate knowledge from different imaging modalities or previous tasks in order to come up with a diagnosis or better understanding of the underlying problem. This illustrates that the knowledge embedded in one task can be leveraged by other tasks and hence it is beneficial to jointly learn related tasks.

Choosing the architecture of an MTL network is based on the following two factors [123]: *what to share* and *how to share*. *What to share* defines the form in which knowledge is shared between tasks. This knowledge sharing can be done through hand-crafted features, input images, and model parameters. *How to share* determines the optimal manner in which this knowledge is shared. In this paper, we focus on parameter-based sharing.

In the following sections, we investigate different MTL network architectures in order to best understand how segmentation and registration tasks share information on the architectural level. The investigated networks predict two sets of contours, one set resulting from the segmentation task and one from the registration task. In this paper, we select the best set of contours as the final output, based on the validation results. More sophisticated strategies are discussed in Section 5.5.

5.2.3.1 Joint registration and segmentation via the registration network

The network in this method, dubbed JRS-reg, has the same architecture as the STL registration network from Section 5.2.2.2, except that this network is optimized using a joint loss as presented in Eq. 5.6.

5.2.3.2 Dense parameter sharing

In this architecture both segmentation and registration tasks are modeled using a single network, where both tasks share all parameters except for the task-specific parameters in the output layer, see Figure 5.1a. The network architecture is the same as the base network (see Section 5.2.1) except for the input and output layers. This dense sharing eliminates overfitting issues since it enforces the parameters to model all the tasks at once, however it does not guarantee the best representation for individual tasks [123]. The input to the network is the concatenation of I_m, I_f, and S_m. The network predicts the ϕ^{pred} between input images as well as S_f^{pred}. The network has 781,164 trainable parameters.

5.2.3.3 Encoder parameter sharing

Since the input to the segmentation and registration tasks are both CT scans, this means they both encode similar features in the down-sampling path of the network. Therefore in this network both tasks share the encoding path and then splits into two upsampling task specific decoder paths. We call this network the Shared Encoder Double Decoder (SEDD) network. Figure 5.1b shows the architecture of the network. The input to the network is the concatenation of I_m, I_f, and S_m. The network predicts ϕ^{pred} between the input images from the registration path while predicting S_f^{pred} from the segmentation path. The network has 722,936 trainable parameters.

5.2.3.4 Cross-stitch network

A flexible approach to share parameters is via a Cross-Stitch (CS) network [124]. In contrast to the heuristic approach of manually choosing which layers are shared and which are task-specific, the CS network introduces a learning-based unit to determine the amount of feature sharing between tasks. The CS units learn to linearly combine feature maps from the two networks, one for segmentation and one for registration, as shown in Figure 5.1c. The unit itself is defined as:

$$\begin{bmatrix} \bar{X}_S^{\ell,k} \\ \bar{X}_R^{\ell,k} \end{bmatrix} = \begin{bmatrix} \alpha_{SS}^{\ell,k} & \alpha_{SR}^{\ell,k} \\ \alpha_{RS}^{\ell,k} & \alpha_{RR}^{\ell,k} \end{bmatrix} \begin{bmatrix} X_S^{\ell,k} \\ X_R^{\ell,k} \end{bmatrix}, \tag{5.5}$$

where $X_S^{\ell,k}$ and $X_R^{\ell,k}$ represent the feature maps k at layer l for the segmentation and registration networks, respectively. $\alpha_{SS}^{\ell,k}$, $\alpha_{SR}^{\ell,k}$, $\alpha_{RS}^{\ell,k}$, and $\alpha_{RR}^{\ell,k}$ represent the learnable parameters of the CS unit. $\bar{X}_S^{\ell,k}$ and $\bar{X}_R^{\ell,k}$ are the output feature maps for the segmentation and registration networks, respectively. The advantage of CS units is that the network can dynamically learn to share the feature maps in case this is beneficial in terms of the final loss value. In case there is no benefit, an identity matrix can be learned, so that the feature maps become task-specific. This allows the network to learn a smooth sharing between the tasks at a negligible increase in the

number of parameters. As suggested by the original paper, we placed the CS units after the downsampling and upsampling layers resulting in a total of 4 CS units. The CS network has 779,000 trainable parameters.

5.2.4 Loss weighting

The loss function for the MTL networks is defined by:

$$\mathscr{L} = w_0 \cdot \mathscr{L}_{\text{NCC}} + w_1 \cdot \mathscr{L}_{\text{DSC-R}} + w_2 \cdot \mathscr{L}_{\text{DSC-S}} + w_3 \cdot \mathscr{L}_{\text{BE}}, \tag{5.6}$$

where w_i are the loss weights. They are chosen based on the relative contribution of their corresponding tasks, so that different tasks would learn at the same pace. These weights can be chosen manually based on empirical knowledge, or automatically. A simple choice would be to weigh the losses equally with a fixed weight of 1. Following are some exemplar algorithms for choosing the loss weights automatically. Chen *et al.* proposed GradNorm [125] to weigh different tasks by dynamic tuning of the gradient magnitudes of the tasks. This tuning is achieved by dynamically changing the learning rate for each task so that all tasks would be learning at the same speed. The drawback of this approach is that it requires access to the internal gradients of the shared layers which could be cumbersome. Moreover, one needs to choose which shared layer to back propagate to in case of multiple shared layers. Kendall *et al.* [126] proposed to weigh each task by considering the homoscedastic uncertainty of that task, so that tasks with high output variance will be weighted less than tasks with low variance. This approach only adds few trainable parameters, namely equal to the number of loss functions. Inspired by GradNorm, Liu *et al.* proposed Dynamic Weight Averaging (DWA) [127], where each task is weighted over time by considering the rate of change of the relative loss weights. Contrary to GradNorm, DWA only requires the numerical values of the loss functions rather than their derivatives. In this paper, we compared equal weights versus homoscedastic uncertainty and DWA. For all the experiments, we set the weight of the bending energy to a fixed value of 0.5 (for more details see Section 5.4.1) instead of a trainable one. This is to prevent the network to set it too low in order to improve the DSC of the deformed contours on the account of the smoothness of the predicted DVF.

5.2.4.1 Homoscedastic uncertainty

Homoscedastic uncertainty was proposed as a loss weighting method by Kendall *et al.* [126]. This is a task-dependant uncertainty which is not dependant on the input data but rather varies between tasks. The authors derived their finding by maximizing the Gaussian likelihood while considering the observational noise scalar σ that represents the homoscedastic uncertainty term related to each task. The following equation describes the weight loss using homoscedastic uncertainty, where σ is a trainable

parameter:

$$\mathscr{L}_{\text{homoscedastic}} = \sum_{i=1}^{T} \frac{1}{\sigma_i^2} \mathscr{L}_i + \log \sigma_i, \tag{5.7}$$

where T is the number of tasks. The higher the uncertainty of task i, the lower the contribution of its associated loss $\mathscr{L}_i$ to the overall loss. The log term can be viewed as a regularization term, so that the network would not learn a trivial solution by setting the uncertainty of all tasks to extreme values.

5.2.4.2 Dynamic weight averaging

Dynamic Weight Averaging (DWA) was proposed by Liu *et al.* [127]. Similar to GradNorm [125], DWA weights the losses via the rate of change of the loss of each task over the training iterations t. In contrast to GradNorm, DWA does not require access to the internal gradients of the network, but only requires the numerical loss values. According to DWA, the weight w of the loss $\mathscr{L}$ associated with the task k is defined as:

$$w_k(t) = \frac{K \exp(r_k(t-1)/tmp)}{\sum_i \exp(r_i(t-1)/tmp)}, \; r_k(t-1) = \frac{\mathscr{L}_k(t-1)}{\mathscr{L}_k(t-2)}, \tag{5.8}$$

where r_k is the relative loss ratio and tmp is the temperature that controls the smoothness of the the task weighting. Here, we set $tmp = 1$ as suggested by the original paper. For the initial two iterations, $r_k(t)$ is set to 1.

5.3 Datasets, implementation, and evaluation

5.3.1 Datasets

This study involves two datasets from two different institutes and scanners for patients who underwent intensity-modulated radiotherapy for prostate cancer. The first dataset is from Haukeland Medical Center (HMC), Norway. The dataset has 18 patients with 8-11 daily CT scans, each corresponding to a treatment fraction. These scans were acquired using a GE scanner and have 90 to 180 slices with a voxel size of approximately 0.9 × 0.9 × 2.0 mm. The second dataset is from Erasmus Medical Center (EMC), The Netherlands. This dataset consists of 14 patients with 3 daily CT scans each. The scans were acquired using a Siemens scanner, and have 91 to 218 slices with a voxel size of approximately 0.9 × 0.9 × 1.5 mm. The target structures (prostate and seminal vesicles) as well as organs-at-risk (bladder and rectum) were manually delineated by radiation oncologists. All datasets were resampled to an isotropic voxel size of 1 × 1 × 1 mm. All scans and corresponding contours were affinely registered beforehand using `elastix` [128], so that corresponding anatomical structures would fit in the network's field of view. The scan intensities were clipped to [-1000, 1000] .

5.3.2 Implementation and training details

All experiments were developed using Tensorflow (version 1.14) [129]. The convolutional layers were initialized with a random normal distribution ($\mu = 0.0$, $\sigma = 0.02$). All parameters of the Cross-stitch units were initialized using a truncated normal distribution ($\mu = 0.5$, $\sigma = 0.25$) in order to encourage the network to share information at the beginning of the training. In order to ensure fairness regarding the number of parameters in all the networks, the number of filters for the Cross-stitch network were set to [16, 32, 64, 32, 16], while for the other networks the numbers were scaled by $\sqrt{2}$ resulting in [23, 45, 91, 45, 23] filtermaps. This results in approximately 7.8×10^5 trainable parameters for each network. The networks were trained using the RAdam optimizer [130] with a fixed learning rate of 10^{-4}. Patches were sampled equally from the target organs, organs-at-risk and torso. All networks were trained for 200K iterations using an initial batch size of 2. The batch size is then doubled by switching the fixed and moving patches so that the network would warp the fixed patch to the moving patch and vice versa at the same training iteration.

The networks were trained and optimized on the HMC dataset, while the EMC dataset was used as an independent test set. Training was performed on a subset of 111 image pairs from 12 patients, while validation and optimization was carried out on the remaining 50 image pairs from 6 patients.

From each image, 1,000 patches of size 96 × 96 × 96 voxels were sampled. The size of the patch was chosen so that it would fit in the GPU memory, while still producing a patch size of 17^3 at the lowest resolution, which is a reasonable size to encode the deformation from the surrounding region. Losses from the deeply supervised resolutions were weighted equally, $\frac{1}{3}$ each. Training was performed on a cluster equipped with NVIDIA RTX6000, Tesla V100, and GTX1080 Ti GPUs with 24, 16 and 11 GB of memory, respectively. The source code is available at https://github.com/moelmahdy/JRS-MTL.

5.3.3 Evaluation metrics

The automatically generated contours are evaluated geometrically by comparing them against the manual contours for the prostate, seminal vesicle, rectum, and bladder. The Dice similarity coefficient (DSC) measures the overlap between contours:

$$\mathrm{DSC} = \sum \frac{2 \mid S_f \cap S_g \mid}{\mid S_f \mid + \mid S_g \mid}, \tag{5.9}$$

where S_g is the generated contour from either the segmentation or the registration network. The distance between the contours is measured by the Mean Surface Distance

Table 5.1: The effect of network input for the different architectures on the validation set (HMC) in terms of MSD (mm). Lower values are better. Here, $\oplus$ is the concatenation operation, and $\cdot\|\cdot$ represents the inputs to the segmentation network (left of $\|$) and the inputs to the registration network (right of $\|$). Stars denote one-way ANOVA statistical significance with respect to the Cross-stitch network with $I_f \,||\, I_f \oplus I_m \oplus S_m$ as inputs.

			Prostate		Seminal vesicles		Rectum		Bladder	
Network	Input	Output path	$\mu \pm \sigma$	median	$\mu \pm \sigma$	median	$\mu \pm \sigma$	median	$\mu \pm \sigma$	median
Seg	I_f		1.49 ± 0.3*	1.49	2.50 ± 2.6	2.09	3.39 ± 2.2	2.73	1.60 ± 1.1*	1.13
	$I_f \oplus S_m$		1.31 ± 0.4	1.23	1.63 ± 0.9	1.26	2.88 ± 3.4	2.06	1.12 ± 0.5	0.97
	$I_f \oplus I_m$		3.06 ± 0.6*	3.01	5.36 ± 4.4	3.71	14.57 ± 9.4*	11.58	1.46 ± 1.3	1.12
	$I_f \oplus I_m \oplus S_m$		1.26 ± 0.4	1.20	2.08 ± 2.2	1.27	2.79 ± 1.6	2.45	1.05 ± 0.4	0.97
Reg	$I_f \oplus I_m$		1.43 ± 0.8*	1.29	1.71 ± 1.4*	1.37	2.44 ± 1.1*	2.17	3.40 ± 2.3*	2.71
	$I_f \oplus I_m \oplus S_m$		1.91 ± 1.3	1.59	1.92 ± 1.5	1.44	2.58 ± 1.1	2.33	3.88 ± 2.5	3.16
JRS-reg	$I_f \oplus I_m$		1.16 ± 0.3	1.16	1.32 ± 0.6	**1.11**	2.08 ± 1.0	1.82	2.57 ± 2.0	2.04
	$I_f \oplus I_m \oplus S_m$		1.20 ± 0.4	1.13	1.35 ± 0.7	1.16	2.08 ± 1.0	1.82	2.63 ± 2.3	1.90
Cross-stitch	$I_f \,\|\|\, I_f \oplus I_m$	Segmentation	1.47 ± 0.3*	1.48	2.93 ± 3.0*	2.08	2.93 ± 2.0*	2.25	1.19 ± 1.0	0.89
		Registration	1.10 ± 0.3	1.07	1.38 ± 0.7	1.17	2.12 ± 1.0	1.89	2.55 ± 2.1	1.89
	$I_f \,\|\|\, I_f \oplus I_m \oplus S_m$	Segmentation	**1.06 ± 0.3**	**0.99**	**1.27 ± 0.4**	1.15	**1.76 ± 0.8**	**1.47**	**0.91 ± 0.4**	**0.82**
		Registration	1.10 ± 0.3	1.06	1.30 ± 0.6	1.13	2.00 ± 1.0	1.75	2.45 ± 2.1	1.81
	$I_f \oplus S_m \,\|\|\, I_f \oplus I_m \oplus S_m$	Segmentation	2.05 ± 0.7*	2.00	3.66 ± 4.4*	2.19	2.44 ± 1.0*	2.35	1.09 ± 0.5*	0.93
		Registration	1.40 ± 0.4	1.35	1.31 ± 0.6	1.17	2.27 ± 1.0	2.02	2.56 ± 1.9	1.96
	$I_f \oplus I_m \oplus S_m \,\|\|\, I_f \oplus I_m \oplus S_m$	Segmentation	1.08 ± 0.3	1.05	1.54 ± 0.9*	1.28	1.88 ± 1.0	1.61	1.01 ± 0.7	0.82
		Registration	1.20 ± 0.3	1.18	1.35 ± 0.7	1.16	2.12 ± 1.1	1.87	2.54 ± 2.2	1.80

(MSD) and Hausdorff Distance (HD) defined as follows:

$$\mathrm{MSD} = \frac{1}{2}\left(\frac{1}{N}\sum_{i=1}^{n} d\left(a_i, S_g\right) + \frac{1}{M}\sum_{i=1}^{m} d\left(b_i, S_f\right)\right), \tag{5.10}$$

$$\mathrm{HD} = \max\left\{\max_i\left\{d\left(a_i, S_g\right)\right\}, \max_j\left\{d\left(b_i, S_f\right)\right\}\right\}, \tag{5.11}$$

where $\{a_1;\ a_2;\ \ldots;\ a_n\}$ and $\{b_1;\ b_2;\ \ldots;\ b_m\}$ are the surface mesh points of the manual and generated contours, respectively, and $d\left(a_i, S_g\right) = \min_j \|b_j - a_i\|$. For all the experiments, we apply the largest connected component operation on the network prediction.

In order to evaluate the quality of the deformations, we calculate the determinant of the Jacobian matrix. A Jacobian of 1 indicates that no volume change has occurred; a Jacobian > 1 indicates expansion, a Jacobian between 0 and 1 indicates shrinkage, and a Jacobian ≤ 0 indicates a singularity, i.e. a place where folding has occurred. We can quantify the smoothness and quality of the DVF by indicating the fraction of foldings per image and by calculating the standard deviation of the Jacobian alongside the MSD of the segmentation.

A repeated one-way ANOVA test was performed using a significance level of $p = 0.05$. P-values are only stated for the comparisons between the best network with the other networks.

5.4 Experiments and results

In the paper we present two single-task networks dubbed *Seg* and *Reg* networks (see Sections 5.2.2.1 and 5.2.2.2 for more details). Moreover, we investigated multiple multi-task networks, namely JRS-reg, dense, SEDD, and Cross-stitch (see Sections 5.2.3.1, 5.2.3.2, 5.2.3.3, and 5.2.3.4 for more details). We compared our proposed methods against three state-of-the-art methods that were developed for prostate CT contouring. These methods represent three approaches, namely an iterative conventional registration method, a deep learning-based registration method, and a hybrid method. For the iterative method, we used `elastix` software [128] with the NCC similarity loss using the settings proposed by Qiao *et. al.* [131]. In the deep learning method proposed by Elmahdy *et. al.* [17], a generative network is trained for contour propagation by registration, while a discrimination network evaluates the quality of the propagated contours. Finally, we compare our methods against the hybrid method proposed by Elmahdy *et. al.* [23], where a CNN network segments the bladder and then feeds it to the iterative registration method as prior knowledge.

Following, we optimize some of the network settings on the validation set (HMC), in order to investigate the influence of the bending energy weight, network inputs, weighting strategy and network architecture on the results. Then, on the independent test set, we present the final results comparing with methods from the literature.

5.4.1 Bending energy weight

We compared the single-task registration, the JRS-reg method and the Cross-stitch network for a set of bending energy weights, see Equations (5.4) and (5.6), while the weights of the other loss functions are set to 1. Figure 5.2 shows the performance of the aforementioned methods using different bending energy weights. The optimal performance of the registration network occurs at a bending weight of 0.5, while the optimal bending weight for both JRS-reg and Cross-stitch network is much lower but with higher standard deviation of the Jacobian. Therefore, for the remainder of the paper we set the weight of the bending energy to 0.5 since it achieves the best compromise between the contour performance in terms of MSD and the registration performance in terms of the std. of the Jacobian determinant.

5.4.2 Optimization of the networks inputs

During training, validation, and testing, we have access to the fixed image I_f, the moving image I_m, and the moving segmentation S_m. In Table 5.1 we compared different sets of inputs on the validation dataset. This experiment helps to better understand how these network interpret and utilize these inputs and how this would reflect on the network outcome represented by the MSD metric. For this experiment we used equal loss weights for the MTL networks.

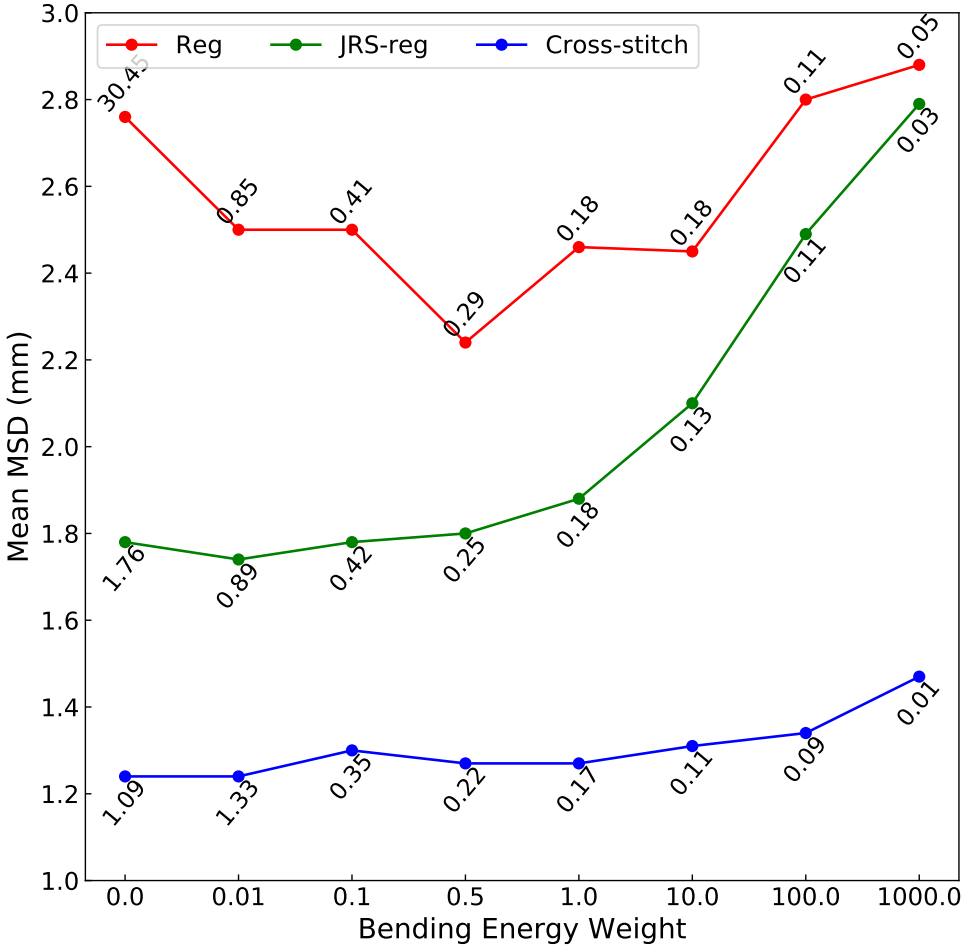


Figure 5.2: The performance of the registration, JRS-registration and Cross-stitch networks with different bending energy weights on the validation set (HMC), in terms of mean MSD averaged over the four organs. The annotation at each point represents the standard deviation of the determinant of the Jacobian.

Feeding S_m to the segmentation network improves the results substantially compared to only feeding I_f, especially for the seminal vesicles, while feeding I_m deteriorates the results. For the registration and JRS-reg networks, feeding S_m alongside I_f and I_m resulted in a similar performance compared to not feeding it. Since the Cross-stitch network is composed of two networks, one for segmentation and the other for registration, we experimented with various combinations of inputs. The results are very consistent with our previous findings on the single-task networks on the effect of

Table 5.2: MSD (mm) values for the different networks and loss weighting methods for the HMC dataset. Lower values are better. Stars and daggers denote one-way ANOVA statistical significance for inter-network experiments with respect to Homoscedastic weights and intra-network experiments with respect to Cross-stitch with Equal weights, respectively. Grey numbers represent the values of the worst path between the segmentation and registration paths, while bold numbers represent the best results.

			Prostate		Seminal vesicles		Rectum		Bladder	
Network	Weight	Output path	$\mu \pm \sigma$	median	$\mu \pm \sigma$	median	$\mu \pm \sigma$	median	$\mu \pm \sigma$	median
JRS-reg	Equal	Registration	1.20 ± 0.4	1.13	1.35 ± 0.7	1.16	2.08 ± 1.0	1.82	$2.63 \pm 2.3^{*}$	1.90
	Homoscedastic	Registration	1.20 ± 0.3	1.20	1.22 ± 0.5	1.07	2.05 ± 1.0	1.81	2.34 ± 2.2	1.60
	DWA	Registration	1.22 ± 0.3	1.18	$1.37 \pm 0.7^{*}$	1.20	$2.29 \pm 1.1^{*}$	2.04	$3.18 \pm 2.4^{*}$	2.43
Dense	Equal	Segmentation	1.14 ± 0.4	1.06	1.73 ± 2.1	1.12	1.91 ± 0.9	1.64	1.04 ± 0.7	0.87
		Registration	1.20 ± 0.3	1.11	$1.33 \pm 0.7^{*}$	1.10	2.16 ± 1.1	1.85	2.56 ± 1.9	1.90
	Homoscedastic	Segmentation	1.09 ± 0.3	1.04	1.51 ± 1.2	1.13	1.86 ± 0.8	1.69	0.99 ± 0.4	0.91
		Registration	1.17 ± 0.3	1.15	1.31 ± 0.6	1.13	2.17 ± 1.0	1.96	$2.63 \pm 2.0^{*}$	1.95
	DWA	Segmentation	$1.12 \pm 0.3^{*\dagger}$	1.04	1.74 ± 2.0	1.13	$1.99 \pm 0.9^{*}$	1.77	1.00 ± 0.4	0.85
		Registration	1.14 ± 0.3	1.14	1.27 ± 0.6	**1.07**	$2.24 \pm 1.1^{*}$	1.97	2.72 ± 1.9	2.13
SEDD	Equal	Segmentation	$1.47 \pm 0.6^{*\dagger}$	1.31	2.81 ± 4.6	1.34	1.97 ± 1.0	1.59	1.21 ± 1.0	0.94
		Registration	$1.28 \pm 0.4^{*}$	1.19	$1.50 \pm 0.9^{*}$	1.26	$2.26 \pm 1.1^{*}$	1.94	$2.61 \pm 2.1^{*}$	1.83
	Homoscedastic	Segmentation	$1.15 \pm 0.3^{\dagger}$	1.14	1.47 ± 1.0	1.22	2.12 ± 1.1	1.91	0.99 ± 0.2	0.94
		Registration	1.19 ± 0.3	1.21	1.23 ± 0.5	1.13	2.15 ± 1.0	1.92	2.31 ± 2.0	1.64
	DWA	Segmentation	$1.22 \pm 0.3^{*\dagger}$	1.18	1.44 ± 0.8	1.21	2.12 ± 1.4	1.73	1.10 ± 0.6	0.93
		Registration	1.22 ± 0.3	1.22	$1.32 \pm 0.6^{*}$	1.10	$2.30 \pm 1.1^{*}$	2.01	$2.86 \pm 1.9^{*}$	2.41
Cross-stitch	Equal	Segmentation	$\mathbf{1.06 \pm 0.3}$	**0.99**	$\mathbf{1.27 \pm 0.4}$	1.15	$\mathbf{1.76 \pm 0.8}$	**1.47**	$\mathbf{0.91 \pm 0.4}$	**0.82**
		Registration	$1.10 \pm 0.3^{*}$	1.06	1.30 ± 0.6	1.13	$2.00 \pm 1.0^{*}$	1.75	2.45 ± 2.1	1.81
	Homoscedastic	Segmentation	$1.23 \pm 0.3^{\dagger}$	1.16	1.51 ± 1.2	1.17	2.37 ± 1.0	2.09	0.92 ± 0.2	0.89
		Registration	1.24 ± 0.3	1.24	1.32 ± 0.6	1.13	2.12 ± 1.0	1.89	2.45 ± 1.9	1.97
	DWA	Segmentation	$1.34 \pm 0.4^{*\dagger}$	1.27	1.75 ± 1.7	1.29	$2.32 \pm 0.9^{\dagger}$	2.11	$1.17 \pm 0.8^{*}$	0.91
		Registration	1.22 ± 0.3	1.19	1.27 ± 0.6	1.09	$2.21 \pm 1.0^{*}$	2.00	$2.93 \pm 2.3^{*}$	2.27

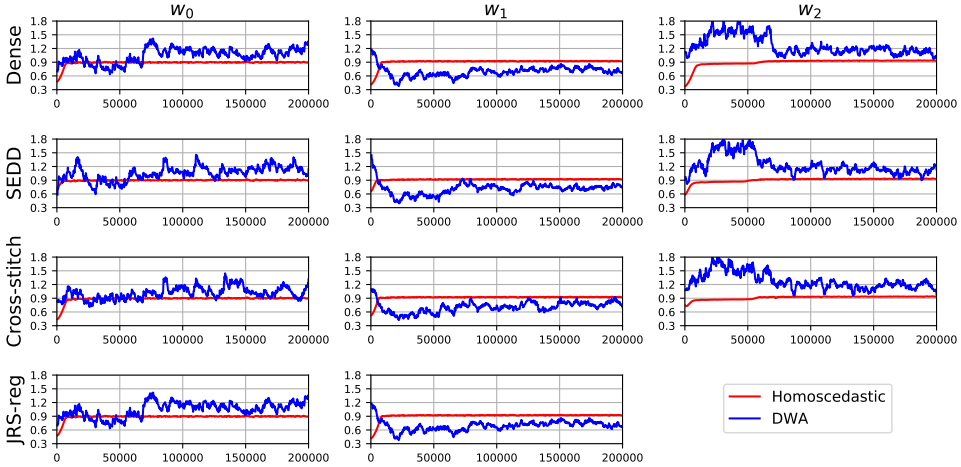


Figure 5.3: The evolution of the loss weights during training for different multi-task networks on the validation dataset (HMC).

using S_m as an input.

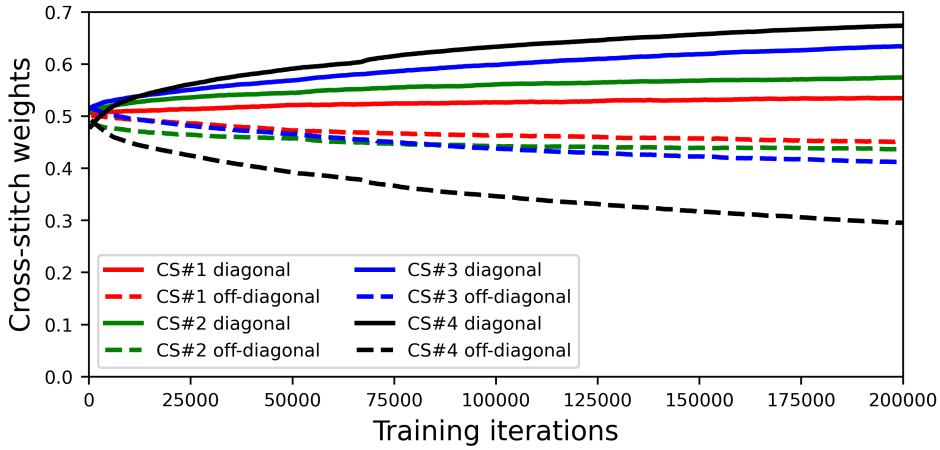


Figure 5.4: The evolution of the Cross-stitch units weights during training using equal weights. CS#1 and CS#2 are placed in the down-sampling path, while CS#3 and CS#4 are placed in the upsampling path. The solid lines represent the mean of the weights across the diagonal of the CS unit, while the dashed lines represent the mean of the off-diagonal weights.

For the remainder of this paper, we chose to use I_f as input for the segmentation network, and I_f and I_m as inputs for the registration network. Although adding S_m proved to be better especially for the segmentation network, here we exclude it, since these two methods act as a baseline and this is the standard setting in single-task networks. For dense, SEDD, and JRS-reg networks, we select a concatenation of I_m, I_f, and S_m for the final network. For the Cross-stitch network, we select I_f for the segmentation network and the concatenation of I_m, I_f, and S_m for the registration network.

5.4.3 Optimization of loss weighting strategy

In this experiment we investigate the performance of the various loss weighting strategies introduced in Section 5.2.4 in order to select the best weighting method for the underlying tasks.

Table 5.2 shows the results of the different weighting strategies for the MTL networks in terms of MSD. For the JRS-reg network architecture, weighting the losses with homoscedastic uncertainty achieved comparable results to using equal weights, while DWA scored somewhat less. For the dense and SEDD architectures, homoscedastic weighting achieved a slightly better performance, while equal weights was best for the Cross-stitch network. For these architectures (dense, SEDD, and

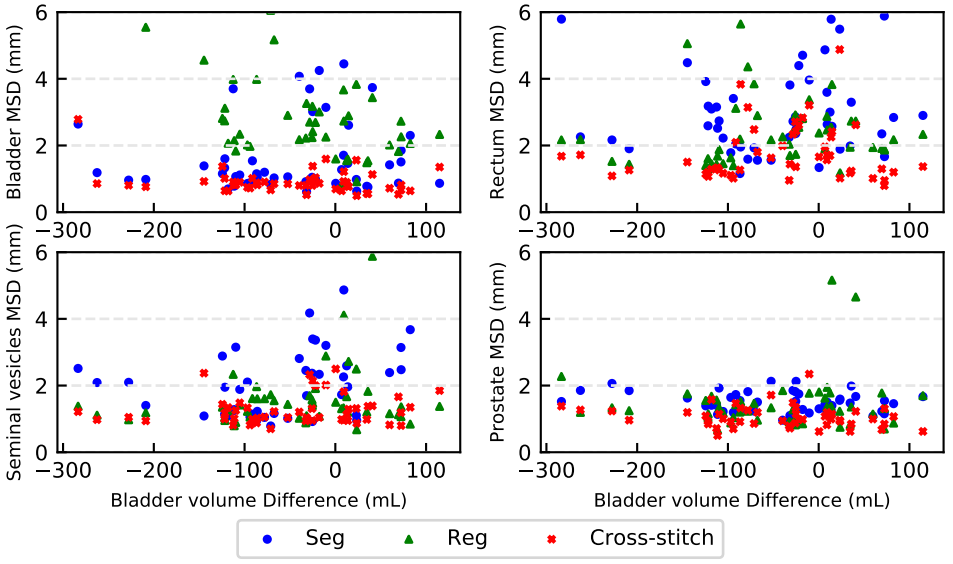


Figure 5.5: The effect of the bladder volume deviation from the planning volume on the performance of the Seg, Reg, and Cross-stitch networks for the validation set (HMC).

Cross-stitch), the segmentation output path showed improvement over the registration output path.

Figure 5.3 illustrates the evolution of the loss weights w_i during training, for different multi-task network architectures and weighting strategies.

For the remainder of this paper and based on the previous findings, we chose the homoscedastic uncertainty weighting strategy for the JRS-reg, dense and SEDD networks, while using equal weights for the Cross-stitch network.

5.4.4 Analysis of cross-stitch units

Analysis of the behavior of the Cross-stitch units during training facilitates the understanding of how the segmentation and registration networks interacts in the MTL settings. Figure 5.4 shows the mean of the CS units across the diagonal and off-diagonal (See Equation (5.5)). Higher weights on the diagonal means that the network tends to separate the task-specific feature maps, while higher weights off-diagonal means that the network tends to share the corresponding feature maps.

5.4.5 Effect of the bladder filling

For the HMC dataset, which was used for training and validation, a bladder filling protocol was in place, meaning that the deformation of the bladder between daily and

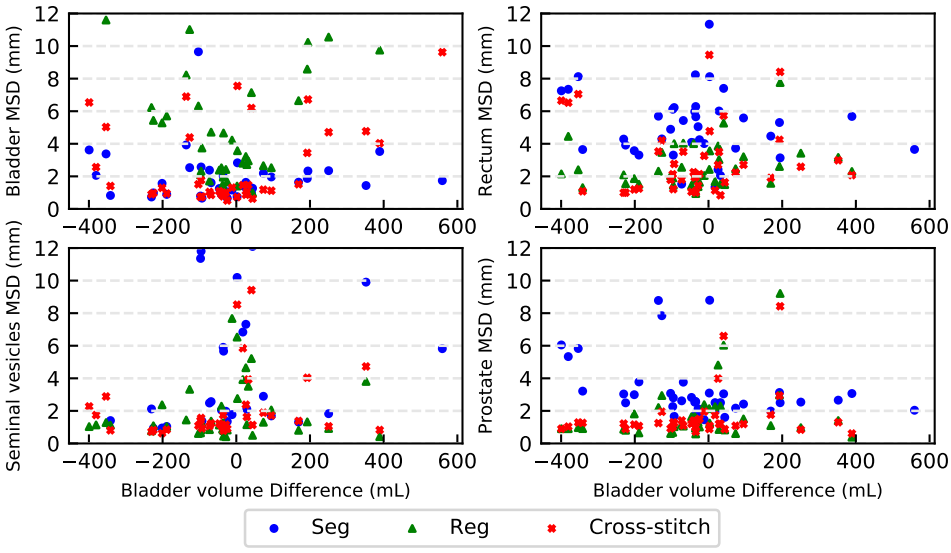


Figure 5.6: The effect of the bladder volume deviation from the planning volume on the performance of the STL and the Seg, Reg, and Cross-stitch networks for the independent test set (EMC).

planning scans is not large. However, this is not the scenario for the EMC dataset, the test set.

Figure 5.5 and 5.6 illustrates the effect of the bladder volume variation from the planning scan on the performance of the Seg, Reg, and Cross-stitch networks. The Cross-stitch network is resilient to bladder filling for both the HMC and EMC datasets.

5.4.6 Evaluation of the quality of the DVF

The smoothness of the predicted DVF is an important parameter to evaluate the predicted deformation field. Table 5.5 shows a detailed analysis of the DVF in terms of the standard deviation of the determinant of the Jacobian as well as the folding fraction for the registration path of the different networks.

5.4.7 Comparison against the state-of-the-art

Table 5.3 and 5.4 show the results for the validation set (HMC) and test set (EMC), respectively. The first two networks in each table are single-task networks. For both sets, the registration network outperformed the segmentation network for all organs except the bladder. The mean MSD for the independent test set is higher than the corresponding numbers in the validation set for most organs. However, the median values are on par. For the MTL networks, the segmentation path of the networks

Table 5.3: MSD (mm) values for the different networks on the validation set (HMC). Lower values are better.

		Prostate		Seminal vesicles		Rectum		Bladder	
Network	Output path	$\mu \pm \sigma$	median	$\mu \pm \sigma$	median	$\mu \pm \sigma$	median	$\mu \pm \sigma$	median
Seg	Segmentation	1.49 ± 0.3	1.49	2.50 ± 2.6	2.09	3.39 ± 2.2	2.73	1.60 ± 1.1	1.13
Reg	Registration	1.43 ± 0.8	1.29	1.71 ± 1.4	1.37	2.44 ± 1.1	2.17	3.40 ± 2.3	2.71
JRS-reg	Registration	1.20 ± 0.3	1.20	1.22 ± 0.5	**1.07**	2.05 ± 1.0	1.81	2.34 ± 2.2	1.60
Dense	Segmentation	1.09 ± 0.3	1.04	1.51 ± 1.2	1.13	1.86 ± 0.8	1.69	0.99 ± 0.4	0.91
	Registration	1.17 ± 0.3	1.15	1.31 ± 0.6	1.13	2.17 ± 1.0	1.96	2.63 ± 2.0	1.95
SEDD	Segmentation	1.15 ± 0.3	1.14	1.47 ± 1.0	1.22	2.12 ± 1.1	1.91	0.99 ± 0.2	0.94
	Registration	1.19 ± 0.3	1.21	1.23 ± 0.5	1.13	2.15 ± 1.0	1.92	2.31 ± 2.0	1.64
Cross-stitch	Segmentation	$\mathbf{1.06 \pm 0.3}$	**0.99**	$\mathbf{1.27 \pm 0.4}$	1.15	$\mathbf{1.76 \pm 0.8}$	**1.47**	$\mathbf{0.91 \pm 0.4}$	**0.82**
	Registration	1.10 ± 0.3	1.06	1.30 ± 0.6	1.13	2.00 ± 1.0	1.75	2.45 ± 2.1	1.81
Elastix [131]	Registration	1.73 ± 0.7	1.59	2.71 ± 1.6	2.45	3.69 ± 1.2	3.50	5.26 ± 2.6	4.72
Hybrid [23]	Registration	1.27 ± 0.3	1.25	1.47 ± 0.5	1.32	2.03 ± 0.6	1.85	1.75 ± 1.0	1.26
JRS-GAN [17]	Registration	1.14 ± 0.3	1.04	1.75 ± 1.3	1.44	2.17 ± 1.1	1.89	2.25 ± 1.9	1.54

Table 5.4: MSD (mm) values for the different networks on the independent test set (EMC). Lower values are better. Results for JRS-GAN are not available for this dataset.

		Prostate		Seminal vesicles		Rectum		Bladder	
Network	Output path	$\mu \pm \sigma$	median	$\mu \pm \sigma$	median	$\mu \pm \sigma$	median	$\mu \pm \sigma$	median
Seg	Segmentation	3.18 ± 1.8	2.57	9.33 ± 10.1	5.82	5.79 ± 3.4	5.18	1.88 ± 1.5	1.50
Reg	Registration	2.01 ± 2.5	1.18	2.86 ± 5.2	1.18	2.89 ± 2.5	2.23	5.98 ± 4.7	4.44
JRS-reg	Registration	1.94 ± 2.6	1.16	2.48 ± 4.8	1.01	2.67 ± 2.4	2.05	4.80 ± 4.6	2.12
Dense	Segmentation	2.01 ± 2.6	**1.15**	4.08 ± 7.2	1.23	3.70 ± 5.4	**2.03**	2.75 ± 3.1	1.23
	Registration	1.93 ± 2.5	1.15	2.53 ± 4.7	1.01	2.67 ± 2.3	2.13	5.08 ± 4.4	3.01
SEDD	Segmentation	1.99 ± 2.4	1.24	6.26 ± 8.9	3.01	4.21 ± 4.9	2.12	2.43 ± 2.9	1.04
	Registration	1.92 ± 2.5	1.19	2.43 ± 4.5	1.07	2.72 ± 2.4	2.17	4.86 ± 4.4	2.22
Cross-stitch	Segmentation	1.88 ± 1.9	1.30	2.76 ± 3.5	1.28	4.87 ± 6.8	2.49	$\mathbf{1.66 \pm 1.7}$	**0.85**
	Registration	1.91 ± 2.3	1.23	2.41 ± 4.5	**0.95**	2.78 ± 2.4	2.16	4.90 ± 4.0	2.84
Elastix [131]	Registration	$\mathbf{1.42 \pm 0.7}$	1.17	2.07 ± 2.6	1.24	3.20 ± 1.6	3.07	5.30 ± 5.1	3.27
Hybrid [23]	Registration	1.55 ± 0.6	1.36	$\mathbf{1.65 \pm 1.3}$	1.22	$\mathbf{2.65 \pm 1.6}$	2.36	3.81 ± 3.6	2.26

Table 5.5: Analysis of the determinant of the Jacobian for the validation and the independent test sets. Lower values are better.

	Validation set (HMC)		Independent test set (EMC)	
Network	Std. Jacobian	Folding fraction	Std. Jacobian	Folding fraction
Reg	0.2935 ± 0.1022	0.0049 ± 0.0039	0.4129 ± 0.2258	0.0112 ± 0.0115
JRS-reg	0.2543 ± 0.0505	0.0030 ± 0.0014	0.3148 ± 0.1106	0.0066 ± 0.0062
Dense	0.2062 ± 0.0431	0.0018 ± 0.0012	0.2558 ± 0.0899	0.0036 ± 0.0027
SEDD	0.2626 ± 0.1167	0.0019 ± 0.0016	0.4287 ± 0.3000	0.0066 ± 0.0074
Cross-stitch	0.2241 ± 0.0784	0.0024 ± 0.0018	0.3301 ± 0.1869	0.0071 ± 0.0070

achieved better performance than the registration path on both datasets except for the seminal vesicles. The Cross-stitch network achieved the best results compared to the other MTL networks.

The proposed STL and MTL networks were compared against other state-of-the-art methods that were evaluated using the HMC dataset. For the validation set, the STL network achieved comparable results, while the Cross-stitch network outperformed

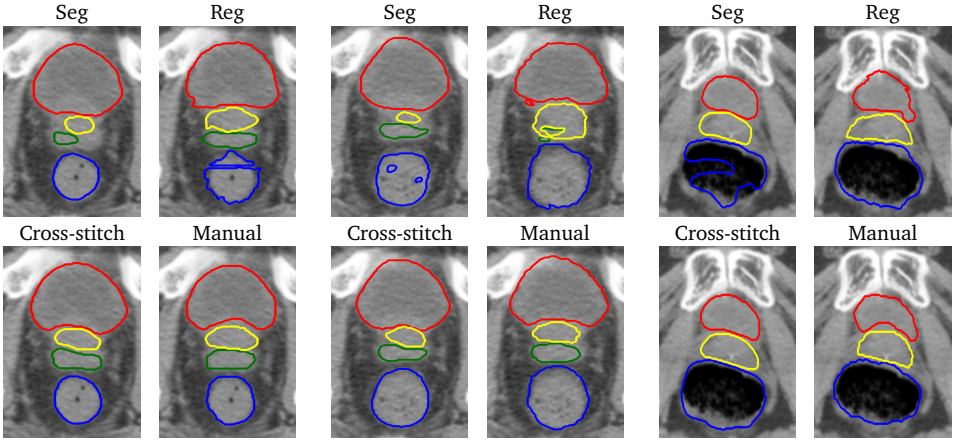


Figure 5.7: Example contours from the validation dataset (HMC) generated by the proposed STL and MTL networks. From left to right, the selected cases are the first, second, and third quartile in terms of the prostate MSD of the Cross-stitch network. The contours of the bladder, prostate, seminal vesicles, and rectum are colored in red, yellow, green, and blue, respectively.

these methods for both output paths. On the test set, *elastix* [131] and the Hybrid method [23] performed better except for the bladder, although the median values of the MTL networks were better.

For the quality of the predicted contours, Figure 5.7 and 5.8 show example contours from the HMC and EMC datasets for the Seg, Reg, and Cross-stitch networks. The examples show that the Cross-stitch network achieves better results compared to the Seg and Reg networks especially for the seminal vesicles and rectum with large gas pockets.

5.5 Discussion

In this study, we proposed to merge image registration and segmentation on the architectural level as well as the loss, via a multi-task learning setting in order to leverage their strengths and mitigate their weaknesses through the sharing of beneficial information. We studied different network architectures and loss weighting methods in order to explore how these tasks interact, and thereby leverage the shared knowledge between them. Moreover, we carried out extensive quantitative analysis in the context of adaptive radiotherapy, and compared the proposed multi-task methods to their single-task counterparts. In this paper, a substantial number of experiments were executed, where we explored the following methodological choices: the bending energy weight, the input to the STL and MTL networks, and the loss weighting method. We also performed a thorough analysis on how Cross-stitch units and loss

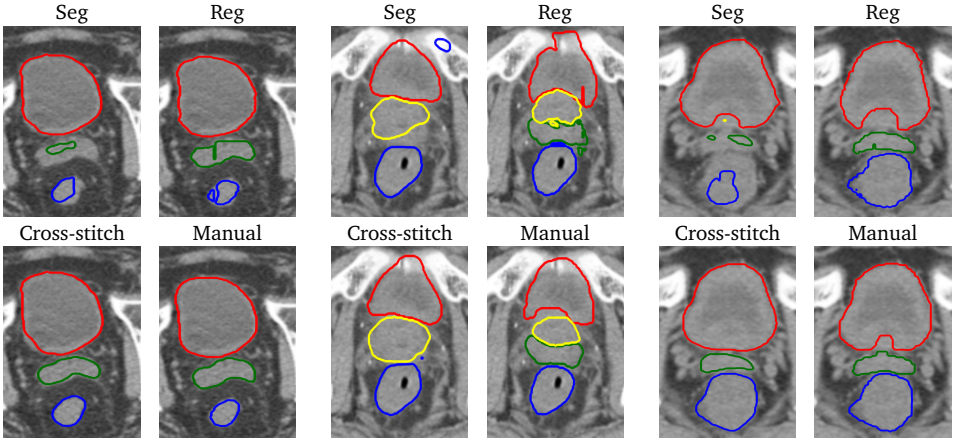


Figure 5.8: Example contours from the independent test set (EMC) generated by the proposed STL and MTL networks. From left to right, the selected cases are the first, second, and third quartile in terms of the prostate MSD of the Cross-stitch network.

weights evolve during training. Finally, we compared our proposed methods against state-of-the-art methods.

In all the experiments we fixed the weight of the bending energy weight so that the network would not set it too low in order to improve the DSC of the deformed contours on the account of the smoothness of the predicted DVF. As shown in Figure 5.2 low bending energy weights result in better contour quality on the account of the smoothness of the predicted DVF.

For the inputs to the STL networks, additionally feeding S_m to the segmentation network resulted in a statistically significant improvement especially for the seminal vesicles. Apparently the network considers S_m as an initial estimation for S_f and subsequently uses it as a guidance for its final prediction. When feeding $I_m$the results deteriorated; this may confuse the network as I_f and I_m have the same anatomy but with different shapes and local positions. The addition of both I_m and S_m performed similar to the addition of only S_m, which indicates that the networks learned to ignore I_m. For the registration network, the addition of S_m resulted in a sub-optimal result, since the S_m contours on its own does not represent the underlying deformation well.

For the inputs to the MTL networks, in the JRS-reg network, feeding S_m alongside I_f and I_m resulted in a similar performance compared to not feeding it. This indicates that the incorporation of S_m via the DSC loss, already enables the JRS-reg network to exploit this extra information, and that additionally adding S_m as a network input does not provide further benefits. In the Cross-stitch network, we found that adding S_m to the registration network results in a statistically significant improvement. Furthermore, feeding S_m to one of the networks is sufficient, proving that segmentation

and registration networks communicate their knowledge efficiently through the Cross-stitch units.

We selected the STL networks with I_f (for segmentation) and I_f alongside I_m (for registration) as input to our baseline methods. Between these two networks, the registration network performed better overall, since the registration network leverages prior knowledge from the organs in the moving image. For the bladder, the segmentation network achieved better results; Apparently the registration network had difficulties finding the correspondence between the bladder in the fixed and moving images, since it tends to deform considerably between visits. However, the segmentation network failed to segment the seminal vesicles for five cases. That is explained by the fact that the seminal vesicles is a difficult structure to segment, due to its relatively small size, undefined borders, and poor contrast with its surroundings. The registration network on the other hand is able to employ the surrounding anatomy as context, to accurately warp the seminal vesicles.

For the multi-task networks, we demonstrated that fusing segmentation and registration tasks is performing better than its single-task counterparts. Merging these tasks using Cross-stitch network achieved the best results on both the validation and testing datasets.

Different loss weighting methods achieved comparable results as shown in Table 5.2. In Figure 5.3, homoscedastic uncertainty tended to weigh all losses equally, using almost a fixed weight of 0.9 during most of the training iterations. On the contrary, DWA tended to fluctuate during training as the weights are updated based on the ratio of the loss from previous iterations, which fluctuates due to the batch-based training. Since the fixed and moving images are affinely registered beforehand, DWA tended to down-weigh the registration loss and the associated DSC at the beginning of the training, while weighting the segmentation network loss more in order to improve its prediction. Later during training, all the weights stabilized around 0.9 similar to homoscedastic uncertainty. Although both methods stabilized by the end of the training around the same value (0.9), the homoscedastic uncertainty achieved slightly better results compared to DWA and equal weighting methods, except for the Cross-stitch network. Our reasoning behind this is that homoscedastic uncertainty, unlike other methods, is learnable during the training and highly dependent on the underlying task uncertainty.

By analyzing the performance of the Cross-stitch units as demonstrated in Figure 5.4, we found that the Cross-stitch units tended to average feature maps for the down-sampling path, while preferring to be more task-specific for the upsampling path. This somewhat mimics the shared encoder double decoder (SEDD) network, but in contrast to this network, the Cross-stitch network does not completely split the decoder paths. This finding confirms that the segmentation and registration tasks are

correlated and thereby encode similar features.

We carried out an experiment to study the effect of the bladder filling protocol between the HMC and EMC datasets. As shown in Figure 5.5, the HMC dataset has a bladder filling protocol so the volume of the bladder changes slightly around 100 mL between different sessions, which is not the case for the EMC dataset as shown in Figure 5.6. Since the registration-based networks and joint networks were trained on small bladder deformations, they failed on large deformations, however the segmentation network was not affected since it does not depend on the deformation but rather the underlying texture to segment the bladder.

In terms of the smoothness of the predicted DVF shown in Table 5.5, MTL networks achieved lower numbers for the standard deviation of the Jacobian as well as for the folding fraction, compared to the STL network (Reg), on both the test and validation set. Our reasoning is that joining the segmentation task to the registration task works as an additional regularization to the registration network. Due to the fact that the higher the quality of the predicted DVF, the higher the quality of the propagated contours and subsequently the lower the DSC loss. The numbers on the test set are slightly higher than the validation set, but this is due to the variance between the deformations between both sets and the fact that the network has not seen the test set before. This can be addressed using transfer learning as suggested by Elmahdy *et al.* [106] or by using synthetic deformations that mimic the one presented in the EMC dataset.

In the paper, we compared our algorithm against different algorithms from various categories: non-learning (`elastix` [128], a popular conventional tool); hybrid [23], and GAN-based [17]. The presented multi-task networks outperformed these approaches on the validation set and performed on par to these methods for the test set. However, the test time for the hybrid and `elastix` methods are in the order of minutes, while the presented methods have the advantage of fast prediction in less than a second. This enables online automatic re-contouring of daily scans for adaptive radiotherapy. Moreover, in our hybrid study [23] we carried out an extensive dosimetric evaluation alongside the geometric evaluation. The predicted contours from that study met the dose coverage constraints in 86%, 91%, and 99% of the cases for the prostate, seminal vesicles, and lymph nodes, respectively. Since our multi-task networks outperformed the geometrical results in that study, we expect that our contours would achieve a higher success rate in terms of the dose coverage. This could potentially reduce treatment related complications and therefore improve patient quality-of-life after treatment.

A promising direction for future research is the addition of a third task, potentially radiotherapy dose plan estimation. Hence, we can generate contours that are consistent with an optimal dose planning. Further studies could also focus on sophisticated

MTL network architectures similar to sluice networks [132] or routing networks [133]. Moreover, we can study how to fuse the contours from the segmentation and registration paths in a smarter way rather than simply selecting one of them based on the validation set.

5.6 Conclusion

In this paper, we propose to formulate the registration and segmentation tasks as a multi-task learning problem. We presented various approaches in order to do so, both on an architectural level and via the loss function. We experimented with different network architectures in order to investigate the best setting that maximizes the information flow between these tasks. Moreover, we compared different loss weighting methods in order to optimally combine the losses from these tasks.

We proved that multi-task learning approaches outperform their single-task counterparts. Using an adaptive parameter sharing mechanism via Cross-stitch units gives the networks freedom to share information between these two tasks, which resulted in the best performance. An equal loss weighting approach had similar performance to more sophisticated methods.

The cross stitch network with equal loss weights achieved a median MSD of 0.99 mm, 0.82 mm, 1.13 mm and 1.47 mm on the validation set and 1.09 mm, 1.24 mm, 1.02 mm, and 2.10 mm on the independent test set for the prostate, bladder, seminal vesicles, and rectum, respectively. That is equal or less than slice thickness (2 mm). Due to the fast inference of the methods, the proposed method is highly promising for automatic re-contouring of follow-up scans for adaptive radiotherapy, potentially reducing treatment related complications and therefore improving patient quality-of-life after treatment.

5.7 Acknowledgment

The HMC dataset with contours was collected at Haukeland University Hospital, Bergen, Norway, and was provided to us by responsible oncologist Svein Inge Helle and physicist Liv Bolstad Hysing. The EMC dataset with contours was collected at Erasmus University Medical Center, Rotterdam, The Netherlands, and was provided to us by radiation therapist Luca Incrocci and physicist Mischa Hoogeman. They are gratefully acknowledged.

6

An Adaptive Intelligence Algorithm for Undersampled Knee MRI Reconstruction

This chapter was adapted from:

N Pezzotti, S Yousefi, M Elmahdy, C Schülke, M Doneva, T Nielsen, S Kastryulin, B Lelieveldt, M van Osch, E de Weerdt, and M Staring. **An Adaptive Intelligence Algorithm for Undersampled Knee MRI Reconstruction**, *IEEE Access*, Pages 204825-204838, 2020.

Abstract

Adaptive intelligence aims at empowering machine learning techniques with the additional use of domain knowledge. In this work, we present the application of adaptive intelligence to accelerate MR acquisition. Starting from undersampled k-space data, an iterative learning-based reconstruction scheme inspired by compressed sensing theory is used to reconstruct the images. We developed a novel deep neural network to refine and correct prior reconstruction assumptions given the training data. The network was trained and tested on a knee MRI dataset from the 2019 fastMRI challenge organized by Facebook AI Research and NYU Langone Health. All submissions to the challenge were initially ranked based on similarity with a known groundtruth, after which the top 4 submissions were evaluated radiologically. Our method was evaluated by the fastMRI organizers on an independent challenge dataset. It ranked #1, shared #1, and #3 on respectively the 8x accelerated multi-coil, the 4x multi-coil, and the 4x single-coil tracks. This demonstrates the superior performance and wide applicability of the method.

6.1 Introduction

Magnetic Resonance Imaging (MRI) is a widely applied non-invasive imaging modality, with excellent soft tissue contrast and high spatial resolution. Unlike Computed Tomography (CT) scanning, MRI does not expose patients to any ionizing radiation, making it a compelling alternative. MR images are essential for clinical assessment of soft tissue as well as functional and structural measurements, which leads to early detection and diagnosis of many diseases. However, MRI is relatively slow compared to other imaging modalities. The total examination time can vary from 15 minutes for knee imaging to an hour or more for cardiac imaging. Remaining still for this long in a confined space is challenging for any patient, being especially difficult for children, elderly and patients under pain. Motion artifacts are not only difficult to correct, which may require a complete re-scan [134]. Furthermore, the acquisition time affects the temporal resolution and subsequently limits the potential of MRI for dynamic imaging, where high temporal resolution and robustness against motion are critical for diagnosis. Moreover, the relatively long scan times lead to high costs that limit the availability of MRI scanners [135]. Therefore, fast acquisition and reconstruction are crucial to improve the performance of current MR scanners, which led in recent years to the development of techniques such as parallel reception, compressed sensing and multi-band accelerations. However, there is still a need for further scan acceleration.

The long acquisition time is intrinsic to the scanner and physics properties of MRI. For the majority of scans performed in clinical practice, this acquisition is done through consecutive reading-out of single lines in k-space. These readouts are constrained by physical limitations of the hardware, the contrast generating principle, and human physiology. The scanning time could be shortened by reducing the number of acquired lines in k-space, i.e. by undersampling the 2D or 3D k-space. However, this could violate the Nyquist criterion, resulting in aliasing and blurriness in the reconstructed images, rendering them unqualified for clinical purpose. Compressed Sensing (CS) and Parallel Imaging are the most common solutions for acceleration by undersampling, while maintaining image quality. Compressed Sensing, the focus of this paper, introduced by Donoho [28], Lustig [29] and Candes [30], leverages the fact that MR images can be compressed in some domain, restoring the missing k-space data through an iterative reconstruction algorithm [31]. Parallel Imaging uses multiple receive coils that provide an additional signal encoding mechanism, allowing to reduce the number of necessary k-space lines to reconstruct an image, thus partially parallelizing the data acquisition [136].

When CS is used to accelerate MR acquisitions, the k-spaces is sampled pseudo-randomly and the image is subsequently reconstructed by promoting a sparse solution. In the optimal setting, the reconstructed image will be identical to the Fourier

transform of the full k-space and have a limited number of large coefficients when transformed to the sparse domain. Equation (6.1) shows the optimization function that describes the CS algorithm:

$$\min_{\mathbf{x}} \{ \| M\mathcal{F}\mathbf{x} - M\mathbf{y} \|_2^2 + \lambda \| \Psi\mathbf{x} \|_1 \}, \tag{6.1}$$

where $\mathbf{x}$ is the reconstructed image, $\mathbf{y}$ is the fully measured k-space data, $\mathcal{F}$ is the Fourier transform, M (mask) is the undersampling operation, $\Psi\mathbf{x}$ represents the sparsity transform coefficients, and λ is the regularization parameter. The ℓ_1 norm is used to enforce sparsity of the solution in a domain specified by the transformation Ψ. The ℓ_2 norm is used as a similarity measure between the measured k-space data $M\mathbf{y}$ and the reconstructed k-space $M\mathcal{F}\mathbf{x}$, called the "data consistency" term. Note that, in case of multi-coil acquisitions, the data consistency term is given by:

$$\sum_q \| M\mathcal{F}\left(\mathbf{S}_q \cdot \mathbf{x}\right) - M\mathbf{y}_q \|_2^2, \tag{6.2}$$

where q denotes the coil element and $\mathbf{S_q}$ the corresponding coil sensitivity map. The coil sensitivity maps $\mathbf{S}$ are computed using the fully centered region of k-space. A low-passed version of the coil images $\mathbf{x}_q^{\text{lpf}}$ is obtained by cropping the available region of k-space. The sensitivity map $\mathbf{S}_q$, for the individual coil element is computed as follows:

$$\mathbf{S}_q = \frac{\mathbf{x}_q^{\text{lpf}}}{\sqrt{\sum_j \left(\mathbf{x}_j^{\text{lpf}}\right)^2}} \tag{6.3}$$

To simplify notation, without loss of generality, the single-coil data consistency term will be used throughout this paper.

Recently, deep learning has shown promising results for speeding up MR acquisition by adopting Convolutional Neural Networks (CNN) and Generative Adversarial Networks (GAN). In contrast to iteratively solving optimization problems, deep learning offers a solution for reconstructing highly-accelerated scans by adopting learnable reconstruction schemes.

The literature of deep learning-based reconstruction algorithms can be divided into two categories [137]. First, data-driven approaches, where a neural network is trained to find the optimal transformation from the zero-filled k-space to the desired reconstruction. Here, the network is completely dependent on the underlying training dataset without any task-specific prior knowledge on the domain; following are selected exemplar algorithms of this approach. Quan *et al.* [138] developed a GAN network for MR reconstruction starting from undersampled data. Their network consists of two consecutive networks, one for reconstruction and one for refining the results. They used a cyclic data consistency term alongside the WGAN loss.

Mardani *et al.* [139] developed a GAN network for CS. The proposed network corrects aliasing artifacts of MR images. Guo et al. [140] proposed a WGAN with recurrent context-awareness to reconstruct MRI images from highly undersampled k-space data. Schlemper et al. propose a cascaded CNN-based compressive sensing (CS) technique for the reconstruction of diffusion tensor cardiac MRI [141]. Yang et al. proposed a conditional GAN-based architecture for de-aliasing and fast CS-MRI [142, 143]. Putzky *et al.* [144] treated the MR reconstruction problem as an inverse problem. They applied the previously introduced invertible Recurrent Inference Machine (i-RIM) model [145], which iteratively updates its current state based on the output of the forward model. The model was trained and evaluated on the single- and multi-coil data at 4x and 8x accelerations from the fastMRI challenge (see Section 6.2 for more details). AUTOMAP [146] reports good reconstruction results with an architecture that learns to directly transform k-space into image data. Lee *et al.* [147] introduced two separate deep residual networks for magnitude and phase. The proposed networks successfully reconstructed images even when obtained with high undersampling factors.

Second, hybrid approaches are presented in the literature. This class of algorithms builds on top of existing reconstruction solutions and integrate learning-based approaches to substitute part of the original computations, often by adopting an unrolled implementation of an iterative algorithm [32]. A notable example is the Variational network presented by Hammernik et al. [33] utilizing learned filters in an existing iterative optimization scheme, while Yang et al. presented the Deep ADMM-Net [34], which extends the Alternating Direction Method of Multipliers (ADMM) [148] approach by integrating learnable operators.

Aggarwal *et al.* [149] introduced a model based deep learning architecture named MoDL to solve the inverse problem, including MR reconstruction. The proposed model consists of a series of recursive linear CNN networks. These networks share weights for regularization and reduction in the number of parameters. The proposed network imitates the CS algorithm and for numerical optimization, the authors introduced a data consistency term using a conjugate gradient (CG) optimization scheme at every iteration. The model was trained on multi-coil brain MR slices from 4 patients and tested on one patient. Ramzi *et al.* [150] provided a reproducible benchmark of deep learning based reconstruction methods on the single-coil part of the fastMRI dataset [151]. The benchmark consists of a U-net [116], cascade net [152], KIKI-net [153], and PD-net [154]. Cascade net has been inspired by a dictionary learning approach [155]. This approach is composed of residual convolutional blocks applied in image space followed by data consistency layers. The data consistency layers enforce the k-space values be close to the original k-space measurements. KIKI-net is a cascaded network where a non-residual convolutional block has been added to perform k-space

completion, while PD-net provides a learnable and unrolled version of the Primal Dual Hybrid Gradient optimization algorithm [156]. Seitzer et al. discussed the inadequacy of loss function for training a CS-MRI reconstruction CNN [157]. In that study they proposed a refinement method which incorporates both loss functions in a harmonious way to improve the training stability.

Recently, Zhang and Ghanem [158] developed a deep learning approach called ISTA-Net that mimics the conventional ISTA algorithm, but enriches it by replacing the sparsifying transform and the thresholding with learned operations. The resulting network does not implement a fully iterative algorithm, but it simulates it by adopting a fixed number of iterations, effectively enabling the implementation of a deep neural network that can be trained by the backpropagation algorithm. Inspired by the work of Zhang and Ghanem [158], in this paper we propose a deep-learning based solution, Adaptive-CS-Network, that mimics the ISTA algorithm, but introduces strong prior information, i.e., inductive biases, to better constrain the reconstruction problem. The main contributions of this work are: i) we propose a novel CNN network that integrates and enhances the conventional CS approach; ii) it integrates multiscale sparsification, inspired by wavelet transforms, but in a learnable manner; iii) we adopt domain-specific knowledge, such as data consistency, a prior on known phase behavior, and the location of the background: these computations cannot be easily learned by a CNN; iv) the proposed model exploits the correlation between neighbouring slices by adopting a 2.5D learning approach. In addition, we propose a hierarchical training strategy that leverages the available data. We conducted extensive experiments to investigate the performance of the network, and show that domain specific information is crucial for reconstructing high-quality MR images. The proposed network showed superior performance by winning one, and co-winning a second track out of the three tracks of the fastMRI challenge [151].

6.2 FastMRI challenge

The fastMRI challenge is a challenge organized by Facebook AI Research and NYU Langone Health [151]. The aim of the challenge is to advance and encourage AI-based research in MR reconstruction in order to allow acceleration of the acquisition and, subsequently, to reduce the examination time. The challenge is divided in three tracks: 4x single-coil, 4x multi-coil, and 8x multi-coil accelerations. Eight teams participated in the multi-coil track and 17 teams in the single-coil track [159].

6.2.1 Dataset

The challenge organizers released a large-scale dataset of raw MR data of the knee [160]. The data was acquired with a 2D protocol in the coronal direction with a 15 channel knee coil array using Siemens MR machines at two different field strengths: 1.5T and 3T [151]. The data was acquired using two pulse sequences: a *proton*

density weighting with (PDFS) and without (PD) fat suppression. The data is divided approximately equally between these pulse sequences. The pixel size is 0.5 mm × 0.5 mm with a slice thickness of 3 mm.

The dataset is divided in 4 categories: training (973 volumes, 34,742 slices), validation (199 volumes, 7,135 slices), test (118 volumes, 4,092 slices), and challenge (104 volumes, 3,810 slices). These numbers are the same for multi-coil and single-coil data, with the exception of the test and challenge categories, where single-coil data has respectively 10 and 12 volumes less than the multi-coil data. The training, validation and test sets were publicly available since late November, 2018, while the challenge set was available since September 2019. The full k-space was available for all the datasets except for the test and challenge sets. Training and validation sets were considered for training and optimizing our model, while the test set was used for evaluating model performance on a public leaderboard. The final model was evaluated by the organizers on the independent challenge set.

The k-space data provided in the challenge were undersampled using a Cartesian mask, where k-space lines are set to zero in the phase encoding direction. The sampling density is dependent on the acceleration rate (4x or 8x), where the sampled lines are randomly selected. All masks, however, are fully sampled in the central area of k-space which corresponds to the low frequencies of the image. For the 4x accelerated scans, this percentage is 8% while it is 4% for 8x acceleration. Besides making the reconstruction problem easier to solve, such lines allow for obtaining a low-pass filtered version of the image that is used to compute the coil sensitivity maps $\mathbf{S}_q$ as presented in Equation (6.3) using a root sum of square approach [151].

6.2.2 Quantitative evaluation

In order to measure the accuracy of the reconstructed volumes **r** compared to the target volumes **t**, the following metrics were considered:

6.2.2.1 Normalized mean square error (NMSE)

measures the square of the Euclidean norm between a pair of images:

$$\text{NMSE} = \frac{||\mathbf{r}-\mathbf{t}||_2^2}{||\mathbf{t}||_2^2} \tag{6.4}$$

6.2.2.2 Peak signal-to-noise ratio (PSNR)

the ratio between the maximum intensity and the underlying distortion noise:

$$\text{PSNR} = 10\log_{10}\frac{\max(\mathbf{t})^2}{\frac{1}{N}||\mathbf{r}-\mathbf{t}||_2^2} \tag{6.5}$$

6.2.2.3 Structural similarity index metric (SSIM)

measures image similarity using human perception aspects [161]. SSIM is calculated by measuring three image distortions including luminance $l(\cdot)$, contrast $c(\cdot)$ and structure $s(\cdot)$:

$$\text{SSIM} = l(\mathbf{r},\mathbf{t})^{\alpha} c(\mathbf{r},\mathbf{t})^{\beta} s(\mathbf{r},\mathbf{t})^{\gamma}, \tag{6.6}$$

where α, β, γ are the distortion weights, here chosen as 1. In this study, similar to the fastMRI challenge, the SSIM score is computed on the magnitude version of the 2D MR scans, leading to grayscale images.

6.2.3 Radiological evaluation on the challenge dataset

We submitted the reconstructions on the challenge dataset via an online form, which were then evaluated independently by the fastMRI organizers, described in detail by Knoll *et al.* [159]. All submissions were ranked by the SSIM metric, after which only the 4 highest ranking submissions were evaluated by a panel of 7 radiologists. The panel was asked to evaluate the reconstructions on a scale from 1 to 5 on four different categories, where 1 is the best and 5 is the worst. The 4 categories were the rating of artifacts, reconstruction sharpness, perceived contrast-to-noise ratio and diagnostic confidence. The radiological scores were subsequently averaged and translated to a final ranking.

6.3 Methods

In this section we present the background of our solution, first by introducing the Iterative Shrinkage-Thresholding Algorithm (ISTA) [162] and, second, by introducing its deep learning-based variant, ISTA-Net [158]. Then, we present our solution, the Adaptive-CS-Network, that builds on top of the ISTA-Net framework by introducing several improvements, including strong inductive biases derived from domain knowledge on the reconstruction problem.

6.3.1 ISTA background

ISTA is an optimization algorithm to solve (6.1) in an iterative fashion, starting from the reconstruction $\mathbf{x}_0$, which is often obtained by reconstructing the zero-filled undersampled k-space. The initial estimate is refined using the following update rules:

$$\mathbf{r}_{i+1} = \mathbf{x}_i - \rho \mathscr{F}^T (M\mathscr{F}\mathbf{x}_i - M\mathbf{y}), \tag{6.7}$$

$$\mathbf{x}_{i+1} = \arg\min_{\mathbf{x}} \frac{1}{2} \| \mathbf{x} - \mathbf{r}_{i+1} \|_2^2 + \lambda \| \Psi\mathbf{x} \|_1, \tag{6.8}$$

where F^T denotes inverse Fourier transform, $\mathbf{r}_{i+1}$ is an update of the estimate $\mathbf{x}_i$, where the error in the measured data $M\mathbf{y}$ is corrected by a step ρ. Equation (6.8) is a special case of the proximal mapping, with a regularization weight λ, and a crucial

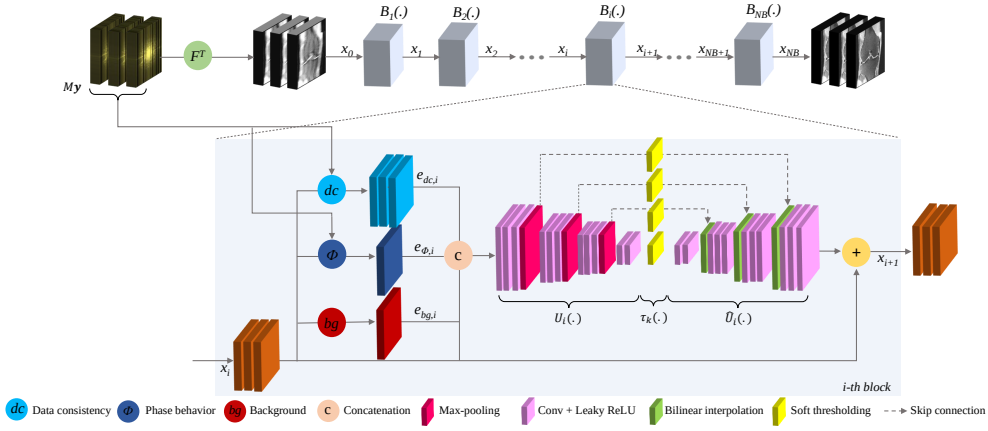


Figure 6.1: Proposed adaptive Adaptive-CS-Net architecture. The input and output of the network are stacks of three consequent knee MR images.

step for optimization algorithms such as ISTA, ADMM [148] and AMP [163]. When Ψ is a wavelet transform $\mathbf{W}$, it can be proven that

$$\mathbf{x}_{i+1} = \mathbf{W}^{-1} soft(\mathbf{W}\mathbf{r}_{i+1}, \lambda), \tag{6.9}$$

where $soft$ is the soft-tresholding operator defined as $soft(\mathbf{u}, \lambda) = \max(|\mathbf{u}| - \lambda, 0) \cdot \frac{\mathbf{u}}{|\mathbf{u}|}$. In general, solving (6.8) is not straightforward for non-linear operators Ψ, limiting the applicability of the ISTA framework to simple transforms. Another problem of this family of algorithms, is the difficulty of tuning the hyperparameters λ and ρ in addition to its slow convergence, hence requiring a lot of iterations to achieve the optimal solution of (6.1).

6.3.2 ISTA-Net

Recently, Zhang and Ghanem introduced a deep-learning approach to overcome the limitations of the ISTA framework for image-to-image reconstruction. Their solution, called ISTA-Net [158], replaces the handcrafted transform Ψ with a learned operator $\mathscr{S}(\cdot)$, which consists of a 2D learnable convolution followed by a rectified linear unit (ReLU) and a second convolution. By replacing Ψ with $\mathscr{S}(\cdot)$ in (6.8), we can rewrite the update rule as

$$\mathbf{x}_{i+1} = \arg\min_{\mathbf{x}} \frac{1}{2} \| \mathbf{x} - \mathbf{r}_{i+1} \|_2^2 + \lambda \| \mathscr{S}(\mathbf{x}) \|_1, \tag{6.10}$$

and, by defining $\hat{\mathscr{S}}$ as the inverse of $\mathscr{S}$, i.e., $\hat{\mathscr{S}} \circ \mathscr{S} = I$, Zhang and Ghanem propose to update (6.9) as follows:

$$\mathbf{x}_{i+1} = \hat{\mathscr{S}}(soft(\mathscr{S}(\mathbf{r}_{i+1}), \lambda)), \tag{6.11}$$

where $\hat{\mathscr{S}}$ has a similar architecture as $\mathscr{S}$.

The model is trained end-to-end, where the iterations of the ISTA algorithm are "unrolled", i.e., a number b of identical reconstruction blocks are created. Note that in the ISTA-Net approach, the learnable parameters are shared among all the blocks in the unrolled network, unlike our solution. The training loss is defined as a combination of the reconstruction and discrepancy loss:

$$\mathscr{L} = \mathscr{L}_{reconstruction} + \sigma \mathscr{L}_{discrepancy} \tag{6.12}$$

$$\mathscr{L}_{reconstruction} = \| \mathbf{x}_b - \mathscr{F}^T \mathbf{y} \|_2^2 \tag{6.13}$$

$$\mathscr{L}_{discrepancy} = \frac{1}{b} \sum_{i=1}^{b} \| \hat{\mathscr{S}}(\mathscr{S}(\mathbf{x}_i)) - \mathbf{x}_i \|_2^2 \tag{6.14}$$

The reconstruction loss encodes the need for the final reconstruction, defined as $\mathbf{x}_b$, to be as close as possible in the least squares sense to the ground-truth image, i.e., $F^T\mathbf{y}$. The discrepancy loss stimulates that $\hat{\mathscr{S}} \circ \mathscr{S} = I$. The σ parameter allows to control the weight given to the discrepancy loss, and it is chosen to be arbitrarily small, e.g., $\sigma = 0.01$. An extension, called ISTA-Net$^+$ is also presented by the authors, where residual computations are adopted.

6.3.3 Adaptive-CS-Network

Starting from the network developed by Zhang and Ghanem, we developed the Adaptive-CS-Network approach. Our solution builds on top of the ISTA-Net solution based on three key innovations, here ordered by importance to the final network performance: i) the use of multi-scale and ii) multi-slice computations, together with iii) the introduction of soft MRI priors. We present them independently, building towards the update rule of the Adaptive-CS-Network model as presented in (6.16). Fig. 6.1 illustrates the proposed network.

First, many non-learned CS algorithms make use of **multi-scale** transforms to sparsify the signal. An example is given in (6.9), where $\mathbf{W}$ is a wavelet transform; a decomposition of the signal into a set of basis functions at different scales. We include this inductive bias in our design, and adopt a multi-scale transform $\mathscr{U}$, and its inverse $\hat{\mathscr{U}}$. As an additional design choice, we decide to sparsify and learn only the residual, therefore our update rule is written as follows:

$$\mathbf{x}_{i+1} = \hat{\mathscr{U}}(soft(\mathscr{U}(\mathbf{r}_{i+1}), \lambda_{s,f_s})) + \mathbf{r}_{i+1}, \tag{6.15}$$

where $\mathscr{U}$ comprises of 2D convolutions and non-linearities in the form of Leaky-ReLU to counteract the problem of dying neurons. To generate a multiscale representation, a max-pooling layer is used and the resulting features are then processed again by convolutional blocks and non-linearities. The exact design of $\mathscr{U}$ is presented in Fig. 6.1. The feature maps produced at the different scales are then thresholded using the soft-max function. Differently from ISTA-Net$^+$, we learn a lambda parameter and feature

channel f_s for each scale s. This approach gives the network the flexibility of tuning the thresholds independently, hence reducing the complexity of the transforms learned by the convolutional operators. Finally, the filtered channels are transformed back into the image domain by the inverse $\hat{\mathcal{U}}$, consisting of interpolation, 2D convolutions and Leaky-ReLU operators. Note that, contrary to the latest literature in deep learning networks, we decided not to adopt strided convolutions for sub- and up-sampling, which would increase the risk of creating checkerboard artifacts [164]; instead we took the more conservative approach of adopting pooling and interpolation layers for achieving better image quality. Overall, the computation represented by $\hat{\mathcal{U}}(soft(\mathcal{U}(\mathbf{r}_{i+1}), \lambda_{s,f_s}))$ is implemented with a UNet-like architecture [116], where the feature maps before the skip connections are filtered according to the parameter λ_{s,f_s}.

Second, it is important to note that the slice thickness of the dataset is much higher than the in-plane resolution. This indicates that inter-slice correlations are less useful for finer scales, and potentially damaging as they will become a confounder for the network. However, such information becomes beneficial at coarser scales, e.g., to facilitate the delineation of the bone in several slices. Since our transform $\mathcal{U}$ is multi-scale by nature, we found it beneficial to inject **neighboring slices** into the model, while leaving it to the network to identify at which scale the information will be used. To reduce the memory footprint of the model, we adopted a 2.5D convolution approach by concatenating neighbouring slices into the input tensor along the channel dimension, enabling to "reinvest" the saved GPU memory as compared to a truly 3D convolution approach, into more unrolled iterations. More details on the number of slices used and the definition of the loss function are given in Section 6.4.3.

Finally, we adopted a hybrid- or nudge- approach to incorporate additional **prior knowledge** into the reconstruction algorithm. We therefore computed additional information derived from the current estimate $\mathbf{x}_i$ together with k-space $M\mathbf{y}$. These soft priors, which are presented in the next section, capture some properties of an MR image that cannot be easily learned by a deep neural network due to the limited size of the receptive field. The priors come in the form of images, and are provided as extra input channel to the transform $\mathcal{U}$. In this way, they are integrated in the computations performed by $\mathcal{U}$ whenever this is beneficial for the optimization of the loss function.

6.3.4 Final design

The overall update for a block B_{i+1} in the Adaptive-CS-Network model is defined as follows:

$$\begin{aligned}\mathbf{x}_{i+1} = B_{i+1}(\mathbf{x}_i) = \\ \mathbf{x}_i + \hat{\mathcal{U}}_i\left(soft\left(\mathcal{U}_i\left(\mathbf{x}_i, \mathbf{e}_{\mathrm{dc},i}, \mathbf{e}_{\phi,i}, \mathbf{e}_{\mathrm{bg},i}\right), \lambda_{s,f_s}\right)\right).\end{aligned} \tag{6.16}$$

Each block in the network learns different transforms $\mathcal{U}_i$ and $\hat{\mathcal{U}}_i$, enabling each block to focus on different properties of the image and effectively increasing the network

capacity. Note that $\mathcal{U}_i$ and $\hat{\mathcal{U}}_i$ are different for every reconstruction block i.

In our final design, the transform $\mathcal{U}_i$ does not receive the data consistent image $\mathbf{r}_i$, as defined in (6.7), but rather the current estimate $\mathbf{x}_i$ together with the data consistency prior $\mathbf{e}_{\mathrm{dc},i}$ computed as follows:

$$\mathbf{e}_{\mathrm{dc},i} = \mathcal{F}^T(M\mathcal{F}\mathbf{x}_i - M\mathbf{y}). \tag{6.17}$$

This "soft data-consistency" update allows the network to evaluate the reliability of the acquired data and potentially compensate errors in the coil combination defined by $\mathcal{F}$ in (6.1).

The second prior we provided to the network, $\mathbf{e}_{\phi,i}$, represents the known phase response for spin-echo MR sequences. Theoretically, spin-echo sequences have zero phase everywhere in the image. In practice, however, slowly varying phase will occur, i.e. nonzero phase only in the low frequencies, due to hardware and acquisition imperfections. Taking this into account, it is noted that the final reconstructed image should be a real valued image after removal of the slowly varying phase. This information is captured in the following prior:

$$\mathbf{e}_{\phi,i} = \left\{ \mathbf{x}_i \cdot \frac{\mathbf{x}^*_{i,\mathrm{lpf}}}{\|\mathbf{x}_{i,\mathrm{lpf}}\|_2} \right\}_{imag}, \tag{6.18}$$

where * denotes the complex conjugate, and lpf refers to low pass filtering. The low pass filter is chosen such that it corresponds to the center part of k-space which is fully sampled. By doing so, the low pass filtered image $\mathbf{x}_{i,\mathrm{lpf}}$ can be derived beforehand only once, hence $\mathbf{x}_{i,\mathrm{lpf}}$ is replaced by $\mathbf{x}_{0,\mathrm{lpf}}$.

Finally, we adopt a simple approach to estimate the location in $\mathbf{x}_i$ where the background is found, which is common in parallel imaging techniques. The following prior is applied:

$$\mathbf{e}_{\mathrm{bg},i} = \frac{\mathbf{x}_i}{\|\mathbf{x}_{i,\mathrm{lpf}}\|_2}. \tag{6.19}$$

This prior will penalize estimated signal content where $\|\mathbf{x}_{i,\mathrm{lpf}}\|$ is low, i.e., within the background. Again, $\mathbf{x}_{i,\mathrm{lpf}}$ is replaced by $\mathbf{x}_{0,\mathrm{lpf}}$. Because $\mathbf{x}_{0,\mathrm{lpf}}$ is based on the fully measured central part of k-space, the image is artefact free albeit at low spatial resolution, leading to a reliable background identification.

In Fig. 6.1 the design of the Adaptive-CS-Network is shown, including the multi-scale transforms, the multi-slice computation and the priors provided as input. Note how the spin-echo and background priors are computed only for the central slice, in order to save GPU memory.

6.3.5 Network training and implementation details

We implemented our models in PyTorch [165]. All the optimization experiments were performed on an NVIDIA V100 GPU with 16 GB RAM and the final network was trained

on two NVIDIA V100 GPU with 16 GB RAM. In order to run as many experiments as possible given the challenge deadline, model optimization (see Section 6.4) was done with a relatively small model (≤ 10 blocks), which we trained for 20 epochs. All the optimization networks were trained and validated on the highest acceleration rate of the challenge, i.e. 8x and for single-coil data, except for the number of the blocks which was performed for both 4x and 8x, and for the priors which are more relevant for the multi-coil data. Since the ground truth for the test set was not available, all the quantitative comparisons were only done on the validation set.

For the challenge, we trained the final model using the training and validation datasets for 25 epochs and accelerations randomly selected from 2x to 10x. The residual connections designed on a per-iteration basis, facilitates the learning and prevents the degradation of the error gradient throughout the architecture. The model was subsequently fine-tuned on eight data sub-populations identified by the acceleration (4x and 8x), the protocol (PD and PDFS) and the scanner field strength (1.5T and 3T). Fine-tuning was then performed for 10 epochs on the sub-populations. This procedure was performed independently for the single- and multi-coil datasets, resulting in a total of 8 models. All models were trained using an exponentially decaying learning rate of 10^{-4}. The final models have 33M trainable parameters each; for the single-coil data this leads to an inference time of approximately 327 ms, while it takes approximately 518 ms to compute the reconstruction of a multi-coil dataset on an NVIDIA V100 GPU.

6.4 Experiments and results: Model optimization

In this section we present how we optimized the network configuration, on a smaller model with 10 reconstruction blocks, using the quantitative measures reported in Section 6.2.2 for validation. We performed experiments on the number of the blocks, the loss functions, the influence of using adjacent slices, the optimizer, and the soft priors. A repeated measure one-way ANOVA test was performed on the SSIM values using a significance level of $p = 0.05$. P-values are only stated for the comparisons between the best method and the other methods. In all the experiments a learning rate of 0.0001 was used.

6.4.1 Number of blocks

The proposed model consists of multiple blocks, related to the number of unrolled iterations of the ISTA scheme. Increasing the number of blocks leads to an increase in the number of parameters of the model, and subsequently training time and GPU memory usage as well as an increase in risk of overfitting. In this experiment we investigated the effect of the number of the blocks on the quality of reconstructed images. Tests were ran with the 2D network for 4x and 8x acceleration rates without neighboring slices, MSE as loss function, RMSprop as optimizer, and with the Unet-like

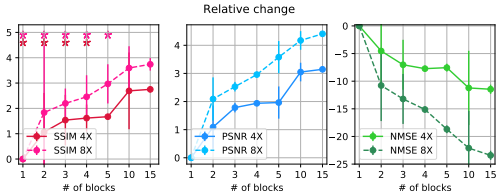


Figure 6.2: The effect of the number of blocks on performance, using the 4x and 8x single-coil validation data. The variance values are shown by the bars. The stars in the first plot show one-way ANOVA statistical significance.

Table 6.1: The effect of the loss function on performance, using the 8x single-coil validation data. Stars denote one-way ANOVA statistical significance.

Loss function	SSIM $\mu \pm \sigma$	NMSE $\mu \pm \sigma$	PSNR $\mu \pm \sigma$	p-value
MSE	0.657±0.149*	0.046±0.029	30.2±2.8	$\ll$0.001
Perceptual loss	0.664±0.157*	0.061±0.044	29,2±3.2	$\ll$0.001
Huber	0.664±0.148*	0.062±0.041	29.1±3.0	$\ll$0.001
ℓ_1	0.664±0.148*	0.062±0.041	29.1±3.0	$\ll$0.001
SSIM	0.662±0.145*	0.065±0.041	28.9±2.8	$\ll$0.001
MSSIM [168]	0.671±0.143*	0.050±0.034	30.1±3.1	$\ll$0.001
Eq. (6.21)	**0.673±0.143**	**0.048±0.033**	**30.3±3.1**	

architecture of 16 filter maps for each convolutional layer. Fig. 6.2 reports the relative changes to a single block of our quantitative metrics. Based on the experiments, increasing the number of the blocks will improve the performance of the network. Therefore, the final network was configured with the maximum number of blocks that could be fitted into GPU memory: 25 blocks. However, for the optimization experiments below only 10 blocks were employed to limit the duration of the training.

6.4.2 Loss functions

In this experiment we investigated the effect of a wide range of differentiable loss functions on the performance of our network. Here, we used the single slice reconstruction network with only 10 blocks, RMSprop as the optimizer, and 16 filter maps for each convolutional layer. The models were trained for 20 epochs to ensure convergence of the model. The evaluated loss functions included MSE, perceptual loss (PL) [166], ℓ_1, Huber [167] and multi-scale structural similarity index (MSSIM) [168]. The PL loss function was calculated using a pre-trained VGG-16 at layers relu1_2, relu2_2, and relu3_3.

MSSIM [168] builds upon SSIM (see Section 6.2.2.3) by incorporating structural similarity at multiple image resolutions, thereby supplying more flexibility compared

to SSIM, and is defined as follows:

$$\text{MSSIM} = \left[l_M(\mathbf{r^c}, \mathbf{t^c})\right]^{\alpha_M} \prod_{i=1}^{M} \left[c_i(\mathbf{r^c}, \mathbf{t^c})\right]^{\beta_i} \left[s_i(\mathbf{r^c}, \mathbf{t^c})\right]^{\gamma_i}, \tag{6.20}$$

where $\mathbf{r}^c, \mathbf{t}^c$, denote the reconstructed and target images respectively, M is the number of scales used, l_M, c_i and s_i are the luminance, contrast, and structure as defined in [161], α_M, β_i, and γ_i are the weights of the distortion factors at different resolution levels. We adopted the same weights as reported in [168].

Zhao et al. [169] reported that a linear combination of SSIM and ℓ_1 preserves the different properties of an image better than each separately: SSIM encourages the network to preserve structural and contrast information, while ℓ_1 enforces sensitivity to uniform biases to preserve luminance [170]. Since MSSIM reached higher metric values than SSIM (see Table 6.1), we deployed a weighted summation of MSSIM [168] and ℓ_1:

$$L = \alpha \text{MSSIM}(\mathbf{r^c}, \mathbf{t^c}) + (1 - \alpha)\|\mathbf{r^c} - \mathbf{t^c}\|_1, \tag{6.21}$$

where $\alpha = 0.84$ was chosen, following Zhao et al. [169]. Note that, compared to the ISTA-Net approach, we found it beneficial not to adopt the discrepancy loss as presented in Eq. (6.12) for two reasons. First, we empirically found that tuning the loss multiplier θ is not straightforward, leading to sub-optimal results in terms of the reconstruction loss. Secondly, computing the discrepancy loss is very demanding in terms of GPU memory, requiring to perform a second forward pass where only the thresholding operation is ignored. While feasible, it requires to make the model significantly smaller in terms of learnable parameters, hence reducing model performance significantly.

Table 6.1 reports the quantitative results for the different loss functions. The weighted linear combination of MSSIM and ℓ_1 yielded the best results, where the p-values indicate that the improvement achieved thanks to our modifications is highly consistent across all scans, despite the small improvements on SSIM-values. Fig. 6.3 shows two example results for the different loss functions, confirming the favorable results for the model trained using a combination of MSSIM and ℓ_1. Therefore, this loss function was selected for training the final model. For the remainder of the experiments, MSSIM is used as loss function.

6.4.3 Multi-slice network

The resolution of the images in the dataset is anisotropic with a voxel size of $0.5 \times 0.5 \times 3$ mm^3. Due to the correlation between adjacent slices with respect to anatomical structures in MRI images, we performed an experiment to assess whether inclusion of neighbouring slices into the reconstruction might improve the performance. We compared the 2D scheme using only the center slice with three alternative 2.5D

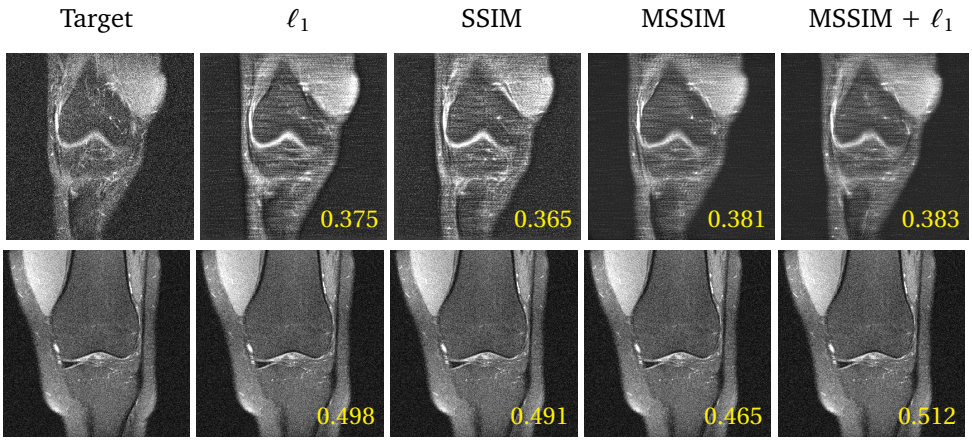


Figure 6.3: Two examples of single-coil 4x for the different loss functions. A small network is used to test several losses. SSIM values are shown in yellow.

Table 6.2: The effect of adopting a 2.5D approach on the 8x single-coil data using the small model. W denotes the loss weight applied to the neighboring slices. Stars denote one-way ANOVA statistical significance.

Network	SSIM $\mu \pm \sigma$	NMSE $\mu \pm \sigma$	PSNR $\mu \pm \sigma$	p-value
2D	0.671±0.143*	0.050±0.034	30.1±3.1	$\ll$0.001
2.5D W0.1	0.549±0.128*	0.089±0.034	26.8±2.1	$\ll$0.001
2.5D W0.2	0.548±0.128*	0.090±0.033	26.8±2.1	$\ll$0.001
2.5D	**0.674±0.143**	**0.048±0.033**	**30.3±3.1**	

schemes: i) the neighboring slices were used together with the center slice as input, but only the center slice was used in the loss function (network 2.5D); ii) and iii) the neighboring slices are also used in the loss, with different weights (0.1 vs 0.2 for the neighbors; 1.0 for the center slice). To compute the first and last slice, we pad the volume with replicas of the edge slices. MSSIM was used for the loss function, 10 blocks, RMSprop as the optimizer, and 16 feature maps.

Table 6.2 shows the results of this experiment, showing that the 2.5D schema very consistently improves over the 2D scheme, and that the loss should only be defined on the center slice. For the final model, this scheme was selected.

6.4.4 Optimizer

We experimented with different optimizers including RMSprop, rectified Adam (RAdam) [130], LookAhead [171] and Ranger [172]. RAdam exploits a dynamic rectifier to adjust the adaptive momentum of Adam [173]. LookAhead not only uses an adaptive learning rate and accelerated schemes but also iteratively updates two sets of weights,

Table 6.3: The effect of the optimizer on performance, using the 8x single-coil validation data. Stars denote one-way ANOVA statistical significance.

Optimizer	SSIM $\mu \pm \sigma$	NMSE $\mu \pm \sigma$	PSNR $\mu \pm \sigma$	p-value
RMSprop	0.673±0.143*	0.048±0.033	30.3±3.1	$\ll$0.001
LookAhead	0.668±0.140*	0.050±0.032	30.0±2.9	$\ll$0.001
Ranger	0.668±0.140*	0.050±0.032	30.0±2.9	$\ll$0.001
RAdam	**0.674±0.141**	**0.048±0.032**	**30.3±3.0**	

Table 6.4: Adaptive-CS-Net vs ISTA-Net$^+$ on the 8x single-coil dataset. ISTA-Net$^+$ has 0.75M trainable parameters, while ISTA-Net-L$^+$ and A-CS-Net have 2.12M trainable parameters. Stars denote one-way ANOVA statistical significance.

Model	SSIM $\mu \pm \sigma$	NMSE $\mu \pm \sigma$	PSNR $\mu \pm \sigma$	p-value
ISTA-Net$^+$	0.547±0.117*	0.169±0.022	23.8±1.9	$\ll$0.001
ISTA-Net-L$^+$	0.543±0.119*	0.103±0.038	26.2±2.0	$\ll$0.001
A-CS-Net	**0.671±0.143**	**0.050±0.034**	**30.1±3.1**	

i.e. fast and slow weights. Ranger combines Radam and LookAhead optimizers into a single one. We used the 2D network with 10 blocks and 16 feature maps for each layer, and MSSIM the loss function.

Table 6.3 tabulates the results for the different optimizers. Since the best results were obtained for the RAdam optimizer, very consistently improving over the other optimizers, this was used for the final network.

6.4.5 Adaptive-CS-Net vs ISTA-Net$^+$

In this experiment, we compare the proposed model to ISTA-Net$^+$ [158]. For this experiment, a 2D network with 10 blocks and 16 feature maps per layer was used, SSIM as loss function, and RAdam as the optimizer. Since ISTA-Net$^+$ uses a much smaller single scale architecture with much fewer network parameters, we added an experiment increasing the feature maps for ISTA-Net$^+$ such that the number of parameters was the same as for our architecture. According to the results reported in Table 6.4, the proposed model outperforms ISTA-Net$^+$ significantly. Figure 6.4 shows a qualitative comparison between ISTA-Net$^+$ and Adaptive-CS-Net on the single-coil 4x dataset. Although for the first image Adaptive-CS-Net reconstructed a better output in terms of the anatomical structure, the output of ISTA-Net-L$^+$ has a higher SSIM value. This implies that the radiological evaluation is a complementary step to judge the quality of the results.

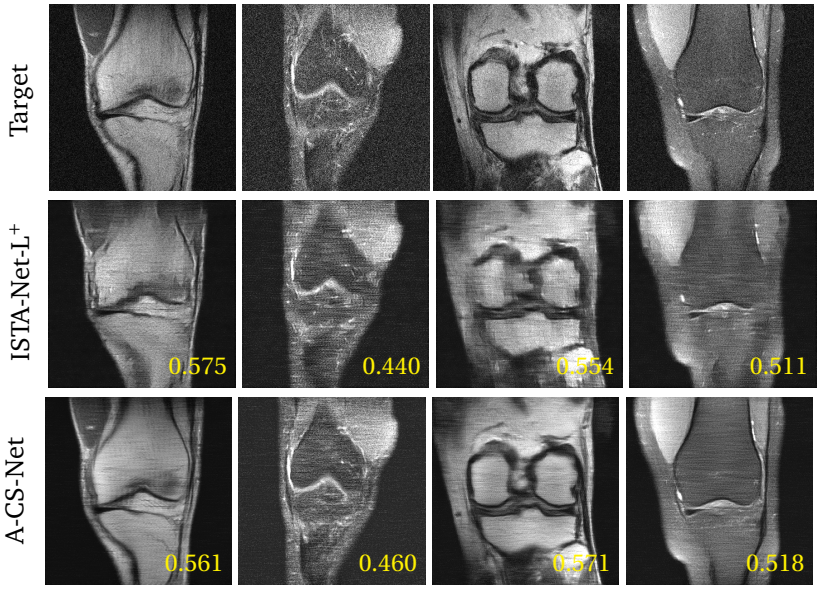


Figure 6.4: Qualitative comparison of Adaptive-CS-Net vs ISTA-Net$^+$ on the single-coil 4x dataset. The SSIM values are shown in yellow.

Table 6.5: The effect of adding priors to the final network on performance, using the multi-coil test data.

Acceleration	prior	SSIM	NMSE	PSNR
4x	–	0.772	**0.025**	30.98
	+	**0.773**	0.028	**33.49**
8x	–	0.674	**0.038**	**30.90**
	+	**0.675**	0.044	30.27

6.4.6 Soft priors

To assess the contribution of the additional soft priors, we compared the full model against a version without known phase behaviour $\mathbf{e}_{\phi,i}$ and without background information $\mathbf{e}_{\text{bg},i}$. Visually, we observed only small differences. To verify the differences in a realistic setting, we submitted the results to the public leaderboard of the fastMRI challenge. As shown in Table 6.5, the network with all priors performed better in terms of the SSIM metric, although the results worsened in terms of NMSE and PSNR. Despite the fact that the improvement was minimal, we decided to adopt all priors for the final model to ensure our participation in the last challenge phase, since the selection was based on SSIM.

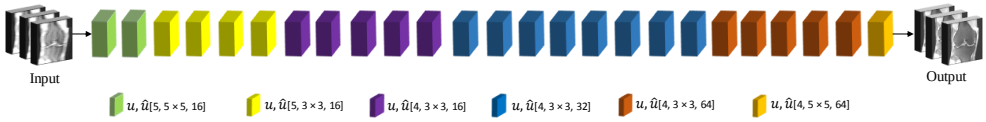


Figure 6.5: Final network, each block has the same structure as shown in Fig. 6.1 and is defined by $\mathscr{U}, \hat{\mathscr{U}}$[number of scales, kernel size, number of feature maps in the first scale]. For all layers Leaky ReLU was used as the activation function.

6.5 Adaptive-CS-NET: Submitted model

In this section, we describe the configuration of the submitted model [174] and analyze the resulting reconstructions. The final performance is evaluated with the quantitative metrics on the test and challenge datasets, and by presenting the radiological scores for the challenge dataset as performed by the fastMRI challenge organizers.

Following our model optimization study, the configuration of the final model was determined as follows. The linear combination of MSSIM and ℓ_1 (6.21) was chosen as the loss function. The 2.5D scheme was chosen with two neighboring slices, with the loss applied only on the central slice. For training the model, the RAdam optimizer was deployed. Fig. 6.5 shows the structure of the final network. Each block is determined by three parameters for the denoiser: 1) the number of scales for the denoiser $\mathscr{U}, \hat{\mathscr{U}}$, 2) the kernel size used in the convolutions and, 3) the number of feature maps in the first convolutional layer, which is then doubled at each scale. According to the experiments presented in Fig. 6.2, the number of reconstruction blocks greatly affects the reconstruction performance, empirically observing that performance still improves when 15 blocks are used. The available GPU memory is a limiting factor when designing a deep neural network. To allow for a large number of blocks, we chose a different design in each block, mixing a less powerful design (16 filters) with more powerful ones (64 filters). By adopting this strategy, our final design contained 25 reconstruction blocks and has 33M parameters.

Fig. 6.6 shows example results of the final network for the multi-coil track from the validation dataset. Fig. 6.7 shows examples from the test and challenge datasets. Table 6.6 shows the SSIM, NMSE, and PSNR values for the test and challenge set (as described in Section 6.2.1), for the images with and without fat suppression and both combined, for both single- and multi-coil MRI scans. For the radiological evaluation, our method scored 2.285, 1.286, and 2.714 for multi-coil 4x, multi-coil 8x, and single-coil 4x, respectively (the closer to 1, the better). The average runtimes for the model are 518 and 327 milliseconds for the multi-coil and the single-coil data, respectively. More details on the results for the challenge were presented in [159].

Table 6.6: Results for the final model for single- and multi-coil data on the test and challenge dataset.

Dataset	Coil		Detail	SSIM	NMSE	PSNR
Test	multi	4x	ALL	0.928	0.005	39.9
			PD	0.961	0.002	41.7
			PDFS	0.891	0.009	37.9
		8x	ALL	0.888	0.009	36.8
			PD	0.937	0.005	38.5
			PDFS	0.843	0.013	35.3
	single	4x	ALL	0.777	0.027	33.7
			PD	0.877	0.010	36.9
			PDFS	0.685	0.043	30.7
		8x	ALL	0.680	0.042	30.5
			PD	0.777	0.019	32.4
			PDFS	0.575	0.067	28.5
Challenge	multi	4x	ALL	0.927	0.005	39.9
		8x	ALL	0.901	0.009	37.4
	single	4x	ALL	0.751	0.030	32.7

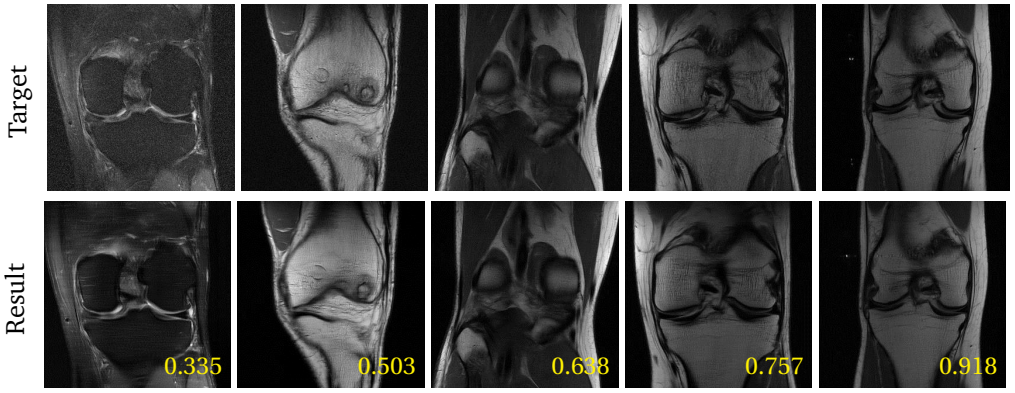


Figure 6.6: Example results of the final model for the multi-coil track accelerated by 8x on the validation dataset. Top row depicts the target image, bottom row the reconstructed images with the SSIM value in yellow.

6.6 Discussion

In this paper we propose a general method, named Adaptive-CS-Net, for reconstructing undersampled MRI data, combining ideas from compressed sensing theory with ideas from MR physics and deep learning. The method was developed in the context of the 2019 fastMRI challenge, which focused on accelerating knee MR imaging. The proposed network is an unrolled iterative learning-based reconstruction scheme, in which a large number of reconstruction blocks refine the MR image by denoising the

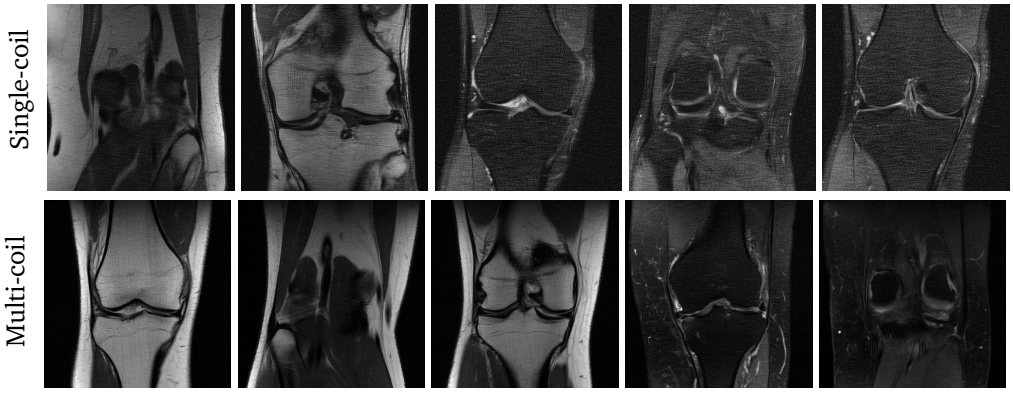


Figure 6.7: Example results of the final network from the test and challenge datasets, for which no ground truth reconstructions are available.

signal in a learned and multi-scale fashion. Moreover, we added neighboring slices as input to the sparsifying transform, as well as a number of soft priors that encode MRI domain knowledge.

The main driver of the performance of our network is the multi-scale architecture, as demonstrated in a direct comparison with ISTA-Net$^+$ that is corrected by the number of trainable parameters. According to the experimental results on the number of blocks for 4x and 8x accelerations of both single- and multi-coil data, we showed that the number of blocks has a large impact on model performance. Therefore, it was decided to use the maximum number of blocks that we could fit into the GPU memory, where we adopted different model designs for the different blocks to save memory. It might be expected that beyond a certain number of blocks, overfitting of the data might occur. However, signs of overfitting were not observed during training and the final number of blocks was only marginally larger than tested in the optimization experiments. Whether further increase in the number of blocks could result in even better performance could be the topic of further experiments. This would, however, need better hardware, as the current design is memory- and time-bound during training. With the current configuration, final model training took approximately 7 days on two V100 GPUs.

We experimented with a large variety of loss functions. Results showed that the linear summation of MSSIM and ℓ_1 performed best. Figure 6.6 showed that poor SNR data yield very low SSIM scores. Surprisingly, within high SNR data, a large variance of SSIM scores is also found. This highlights the fact that further research is required in order to develop better quality metrics. Moreover, we defined a 2.5D scheme to train the network in which three adjacent slices were reconstructed while the loss function was calculated only for the central slice. The proposed scheme outperformed

the 2D network as well as 2.5D networks in which the loss was calculated over all slices. By incorporating the neighbouring slices, the network can exploit existing correlations into the reconstruction of the target slice, which is our main target as defined by our loss. It can be expected that for MRI acquisition with less asymmetric voxel sizes, the inclusion of information of neighbouring slices would become more important. However, weighing the loss of the neighbouring slices resulted in less optimal results since it forces the network to solve a more difficult problem: the network has to reconstruct multiple slices instead of a single one. This reduces the effective network capacity per slice, leading to a degradation of the reconstruction performance. We tested different optimizers, where the newly introduced RAdam outperformed the others and we used it for training the final network. We also incorporated prior knowledge, including data consistency, known phase behaviour and background discrimination to support the network in the reconstruction process. We observed that these priors provided only limited extra performance to the network, resulting in visually similar images and minimal difference in the metrics.

We can conclude that the Adaptive-CS-Net is sufficiently powerful to learn directly from the data how to reconstruct the undersampled k-space, being the multi-scale structure and the use of many reconstruction blocks the main driver of our performance. As a future work, we want to better understand how much the network is relying on the priors by adopting interpretable AI techniques such as differentiable image parameterizations for feature visualization [175]. Stronger use of the priors via the loss function is an additional option.

As mentioned before, the radiologist scores were based on the visual quality of the reconstructed images and not on diagnostic interchangeability. Therefore, designing a network based on the diagnosis can be considered a point for further research. We furthermore observed that optimizing for SSIM was needed for reaching the final stage of the challenge, but is not necessarily an ideal representative of radiological image quality. This observation was very recently confirmed in a comparative study by others [176]. The proposed method outperforms the benchmark networks, including U-net [116], cascade net [152], KIKI-net [153], and PD-net [154], on the single-coil track as reported in [150]. It outperforms as well the i-RIM model [145] on the Multi-coil track but not the single coil track [159].

6.7 Conclusion

In this paper we propose an adaptive intelligence algorithm called Adaptive-CS-Net, which was developed in the context of the 2019 fastMRI challenge. In the two clinically relevant tracks of the challenge, using multi-coil MRI acquisitions, the proposed method was leading, while on a simulated single-coil track the method ranked 3rd.

6.8 Acknowledgements

Facebook AI Research and NYU Langone Health are acknowledged for organizing the 2019 fastMRI challenge.

7

Summary and Future Work

Adapting a radiotherapy treatment plan to the daily anatomy is a crucial task to ensure adequate irradiation of the target without unnecessary exposure of healthy tissue. This adaptation can be performed by automatically generating contours of the daily anatomy together with fast re-optimization of the treatment plan. These measures can compensate for the daily variation and ensure the delivery of the prescribed dose distribution at small margins and high robustness settings. In this thesis, we focused on developing a deep learning-based methodology for automatic contouring for real-time adaptive radiotherapy either guided by CT or MR imaging modalities. In this chapter, we summarize the previous chapters, discuss overall results found in this thesis, and give potential directions of future research.

7.1 Summary

In the first chapter, we introduced adaptive radiotherapy and gave background about automatic contouring methods as well as MR image reconstruction. In Chapter 2, we proposed a contour propagation pipeline that is a hybrid between iterative-based registration and deep learning-based segmentation. The pipeline was trained and evaluated using three datasets and performed better than only using iterative-based methods. In Chapter 3, we proposed an adaptive training mechanism for personalized automatic contour segmentation. This adaptation showed potential for improving the prediction of the daily anatomy based on personalized imaging accumulated over fractions. In Chapter 4, We proposed a 3D adversarial network for joint image registration and segmentation. This work showed that that a discriminator can learn a measure of image alignment to improve registration. In Chapter 5, we further studied joining registration and segmentation in a multi-task learning setting. We showed that training these tasks jointly can substantially improve the network outcome. All the networks in the previous chapters were trained and tested on CT images for adaptive

CT-guided radiotherapy, however this can be generalized for MR-guided radiotherapy by retraining these networks. The remaining problem would be developing a fast MR reconstruction algorithm, which we proposed in Chapter 6.

Chapter 2 In this chapter, we developed and validated a robust registration pipeline for automatic contour propagation for online adaptive IMPT of prostate cancer using elastix software and deep learning. A 3D CNN 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. The first was used for training and testing the ConvNet, where the second and the third were used for the evaluation of the proposed pipeline. The propagated contours were validated clinically as well. The segmentation network achieved a DSC of 88% and 82% on the test datasets. The proposed 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. 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.

Chapter 3 In this chapter, we leveraged personalized anatomical knowledge accumulated over the treatment sessions, to improve the segmentation accuracy of a pre-trained CNN network, for a specific patient. We investigated a transfer learning approach, where we fine-tuned 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 ± 0.43 mm, 2.38 ± 2.76 mm, 2.30 ± 0.96 mm, and 1.24 ± 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.

Chapter 4 proposed to combine the registration and segmentation tasks in a deep learning setting using adversarial learning. We considered the case in which fixed and

moving images as well as their segmentations are available for training, while segmentations are not available during testing; a common scenario in radiotherapy. The proposed framework consists of a 3D end-to-end generator network that estimates the deformation vector field (DVF) between fixed and moving images in an unsupervised fashion and applies this DVF to the moving image and its segmentation. A discriminator network is trained to evaluate how well the moving image and segmentation align with the fixed image and segmentation. The proposed network was trained and evaluated on follow-up prostate CT scans for image-guided radiotherapy, where the planning CT contours are propagated to the daily CT images using the estimated DVF. The proposed GAN network achieved a MSD of 1.13 ± 0.4 mm, 1.81 ± 1.6 mm, 1.00 ± 0.3 mm, 2.21 ± 1.3 mm, and 2.29 ± 2.0 mm, respectively, which was significantly better than the deep learning baseline model as well as the conventional algorithm. The inference time of the proposed model is 0.6 sec., thus enabling real-time contour propagation necessary for online-adaptive radiotherapy.

In **Chapter 5**, we formulated registration and segmentation jointly via a deep learning-based Multi-Task Learning setting, allowing these tasks to leverage their strengths and mitigate their weaknesses through the sharing of beneficial information. We propose to merge these tasks not only on the loss level, but on the architectural level as well. The study involves two datasets from different manufacturers and institutes. We carried out an extensive quantitative comparison between the quality of the automatically generated contours from different network architectures as well as loss weighting methods. Moreover, we evaluated the quality of the generated deformation vector field (DVF). We show that MTL algorithms outperform their Single-Task Learning (STL) counterparts and achieve better generalization on the independent test set. The best algorithm achieved a mean surface distance of 1.06 ± 0.3 mm, 1.27 ± 0.4 mm, 0.91 ± 0.4 mm, and 1.76 ± 0.8 mm on the validation set for the prostate, seminal vesicles, bladder, and rectum, respectively. The high accuracy of the proposed method combined with the fast inference speed, makes it a promising method for automatic re-contouring of follow-up scans for online adaptive radiotherapy.

Chapter 6 presented a fast MR reconstruction algorithm, which enables the application of the automatic contouring methods proposed in the previous chapters for online adaptive MR-guided radiotherapy. Starting from undersampled k-space data, an iterative learning-based reconstruction scheme inspired by compressed sensing theory is used to reconstruct the images. We developed a novel deep neural network to refine and correct prior reconstruction assumptions given the training data. The proposed network was ranked #1, shared #1, and #3 on respectively the 8x accelerated multi-coil, the 4x multi-coil, and the 4x single-coil tracks in the 2019 fastMRI competition organized by Facebook and New York University (NYU). The average runtimes for the model are 0.5 and 0.3 seconds for the multi-coil and the single-coil data, thus enabling

fast reconstruction for MR guided adaptive radiotherapy. This superior performance in terms of reconstruction quality and time makes the model a good candidate for MR guided adaptive radiotherapy.

7.2 Discussion and future work

The aim of the work presented in this thesis was to develop and investigate various methods of automatic contouring for adaptive radiotherapy. Despite the fact that all experiments presented in this thesis were validated on prostate CT scans, all proposed methods are generic and can potentially be applied to MR images as well, thus enabling adaptive CT as well as MR guided radiotherapy. So far the remaining technical bottleneck for applying these methods on MRIgRT would be the fast reconstruction of the MR images, which we addressed in the last chapter of the thesis.

In Chapters 4 and 5, we showed that joining segmentation and registration tasks is very beneficial and clinically important, especially for applications where the output from both tasks is required such as for dose accumulation and dose planning. We concluded that the best mechanism to combine both tasks is by modeling these tasks using a cross-stitch network that shares internal parameters across the tasks. Using a cross-stitch network, we were able to achieve superior performance in the order of 1 mm for the mean surface distance of the target organs and OARs. However, on the independent test set, the deep joint method achieved a result in the order of 2 mm. Compared to the performance of the deep joint network in chapter 5, the hybrid method presented in Chapter 2 had a better generalization since it is a non-learnable method and registers image pairs in an iterative manner. However, the hybrid method running time is in the order of minutes while the joint method takes less than a second, which makes it a better candidate for online adaptive radiotherapy. The generalization of the deep joint method could be further improved, however, by deploying one of the following strategies. First, we could re-train the model and include patients similar to the test data distribution. This way one could adapt the general model to, for example, a specific hospital. Second, we could augment the existing training data with realistic deformations instead of random ones, since such random deformations did not show improvement in our earlier experiments. This realistic deformation can be drawn from a generative mechanical model such as the ones presented in [177, 178]. Using such realistic augmentation would help the network to learn various types and scales of deformations that might not be available in the original training dataset. At last, we can adopt a transfer learning strategy similar to the one proposed in Chapter 3, where we proposed to leverage personalized anatomical knowledge accumulated over the treatment sessions. In that chapter, we demonstrated that adapting the model to a specific patient anatomy can improve the performance of the network especially for the organs that do not deform much from the planning scan as shown in Figure 3.1.

For the organs that deform between sessions such as the rectum and the bladder, the performance of the adapted network did not improve significantly as shown in Table 3.2. This adaptation strategy is effective for any existing model, therefore it can be used to adapt or personalize a pre-shipped model on in-house data without the need to re-train the model from scratch. This would improve model performance over time at a minimal cost while maintaining data privacy. Moreover, by adjusting the transfer learning strategy it can be used for continuously improving model prediction after being corrected by clinical radiologists.

The networks developed in Chapters 2, 3, 4, and 5 focused primarily on automatic contour propagation for prostate CT. These networks, however, can be generalized to MR images for MRIgRT. Various CNN networks with a U-Net variant architecture have already been proposed in literature for MR prostate segmentation [179, 180, 181, 182] as well as other anatomies such as the brain [183, 184, 185] and the heart [186, 187]. Furthermore, similar networks were successfully adopted for MR registration at different anatomical sites [16, 117, 188]. The success of the aforementioned methods that use similar network architecture, makes it very promising our proposed deep joint networks to generalize to MR images of the prostate and potentially for other anatomical sites. In order to validate that hypothesis, we need to conduct a study that involves multiple datasets from different anatomical sites as well as imaging modalities.

For MRIgRT application, the acquisition and reconstruction time of the MR images themselves are still a bottleneck for online MRIgRT application. In Chapter 6, we addressed this problem, where we developed a fast reconstruction algorithm that works for both single-coil and multi-coil MR images. The output from the network was clinically acceptable based on the clinical evaluation performed by radiologists [159]. The outcome of this clinical evaluation makes it very promising and encouraging to be deployed in clinical trials after doing further clinical assessment on different anatomical sites.

In terms of the clinical readiness of the proposed deep joint registration and segmentation network, in Chapter 2 we performed an extensive dosimetric evaluation on the automatically generated contours from the hybrid method. We found that improving the quality of the generated contours in terms of MSD, resulted in a boost of the dosimetric measures in terms of V_{95} and the Conservative Success Rate (CSR) compared to when only using intensity-based registration. Since the cross-stitch network achieved an even better geometric performance, we hypothesize that this improvement would also result in a boost in the corresponding dosimetric measures and we are currently validating this hypothesis.

A promising direction for future research for the joint network is the addition of a third task, potentially being radiotherapy dose plan estimation. Hence, we can

generate contours that are consistent with an optimal dose planning. Further studies could also focus on sophisticated MTL network architectures like sluice networks [132] or routing networks [133]. Moreover, we can study how to fuse the contours from the segmentation and registration paths in a smarter way rather than simply selecting one of them based on the validation set. It also worth investigating semi-supervised training techniques [189] in order to improve the generalizability of the proposed networks. For the fast MR reconstruction network, as a future direction, it will be beneficial to better understand how much the network is relying on the priors by adopting interpretable AI techniques such as differentiable image parameterizations for feature visualization [175]. Stronger use of the priors via the loss function is an additional option. Considering the end goal of MRIgRT which is to extract image contours and to subsequently generate a dose planning, it would be interesting to investigate if we can sacrifice image quality without losing segmentation and registration performance, which may even further accelerate the MR imaging, going from minutes to seconds. This could then in the future lead to an almost realtime steering and control modality, which has benefits for more rapidly moving organs like the lungs and heart.

7.3 General conclusions

In conclusion, this thesis proposes a deep learning based automatic contouring methodology for real time adaptive radiotherapy. The proposed networks were evaluated on prostate CT images, a commonly used modality for treatment planning, but may be generalized for MR images. Moreover, we proposed a fast MR reconstruction algorithm in order to accelerate MR acquisition so that our models can be used potentially for MR guided adaptive radiotherapy as well. All deep learning methods proposed in this thesis have a runtime of less than a second, thus enabling real-time automatic contouring necessary for online-adaptive radiotherapy.

Samenvatting en toekomstig werk

Het aanpassen van een behandelplan voor radiotherapie aan de dagelijkse anatomie is een cruciale taak, om te zorgen voor een adequate bestraling van het doelwit zonder onnodige blootstelling van gezond weefsel. Deze aanpassing kan worden uitgevoerd door automatisch contouren van de dagelijkse anatomie te genereren, samen met een snelle heroptimalisatie van het behandelplan. Deze maatregelen kunnen de dagelijkse variatie compenseren en zorgen voor de aflevering van de voorgeschreven dosisverdeling met kleine marges en hoge robuustheidsinstellingen. In dit proefschrift hebben we ons gericht op het ontwikkelen van een op deep learning gebaseerde methodologie voor automatische contouren voor realtime adaptieve radiotherapie, ofwel geleid door CT- ofwel MR-beeldvormingsmodaliteiten. In dit hoofdstuk vatten we de voorgaande hoofdstukken samen, bespreken we de algemene resultaten die in dit proefschrift zijn gevonden en geven we mogelijke richtingen voor toekomstig onderzoek.

Samenvatting

In het eerste hoofdstuk hebben we adaptieve radiotherapie geïntroduceerd en achtergrondinformatie gegeven over automatische contouring methoden en MR beeldreconstructie. In hoofdstuk 2 stelden we een contour propagatie pijplijn voor die een hybride is tussen iteratieve registratie en segmentatie op basis van deep learning. De pijplijn werd getraind en geëvalueerd met behulp van drie datasets en presteerde beter dan alleen met behulp van iteratieve-gebaseerde methoden. In Hoofdstuk 3 stelden we een adaptief trainingsmechanisme voor gepersonaliseerde automatische contoursegmentatie voor. Deze aanpassing toonde potentieel voor het verbeteren van de voorspelling van de dagelijkse anatomie op basis van gepersonaliseerde beeldvorming geaccumuleerd over fracties. In Hoofdstuk 4 stelden we een 3D adversarial netwerk voor voor gezamenlijke beeldregistratie en segmentatie. Dit werk toonde aan dat een discriminator een maat voor beelduitlijning kan leren om de registratie te verbeteren. In Hoofdstuk 5 bestudeerden we verder het samenvoegen van registratie en segmentatie in een multi-task learning setting. We toonden aan dat het gezamenlijk trainen van deze taken het resultaat van het netwerk aanzienlijk kan verbeteren. Alle netwerken in de voorgaande hoofdstukken zijn getraind en

getest op CT-beelden voor adaptieve CT-geleide radiotherapie, maar dit kan worden gegeneraliseerd voor MR-geleide radiotherapie door deze netwerken te hertrainen. Het resterende probleem zou de ontwikkeling van een snel MR-reconstructiealgoritme zijn, dat we in hoofdstuk 6 hebben voorgesteld.

Hoofdstuk 2 In dit hoofdstuk hebben we een robuuste registratie pipeline ontwikkeld en gevalideerd voor automatische contour propagatie voor online adaptieve IMPT van prostaatkanker met behulp van elastix software en deep learning. Een 3D CNN netwerk werd getraind voor automatische blaassegmentatie van de CT scans. De automatische blaassegmentatie naast de CT scan zijn gezamenlijk geoptimaliseerd om expliciete kennis over de onderliggende anatomie toe te voegen aan het registratie algoritme. We hebben drie datasets in de studie opgenomen. De eerste werd gebruikt voor het trainen en testen van het ConvNet, terwijl de tweede en de derde werden gebruikt voor de evaluatie van de voorgestelde pijplijn. De gepropageerde contouren werden ook klinisch gevalideerd. Het segmentatie netwerk behaalde een DSC van 88% en 82% op de test datasets. De voorgestelde pijplijn bereikte een MSD van $1,29 \pm 0,39$, $1,48 \pm 1,16$, en $1,49 \pm 0,44$ mm voor de prostaat, zaadblaasjes, en lymfeklieren, respectievelijk op de tweede dataset en een MSD van $2,31 \pm 1,92$ en $1,76 \pm 1,39$ mm voor de prostaat en de zaadblaasjes op de derde dataset. De automatisch gepropageerde contouren voldeden in 86%, 91% en 99% van de gevallen aan de dosisvoorwaarden voor respectievelijk de prostaat, de zaadblaasjes en de lymfeknopen. Een conservatief succespercentage (CSR) van 80% werd verkregen, vergeleken met 65% wanneer alleen gebruik werd gemaakt van registratie op basis van intensiteit. Met 80% van de automatisch gegenereerde behandelplannen die direct bruikbaar zijn zonder handmatige correctie, werd een aanzienlijke verbetering van de robuustheid van het systeem bereikt in vergelijking met een eerdere aanpak. De voorgestelde methode maakt dus een nauwkeurigere protonentherapie van prostaatkanker mogelijk, wat mogelijk leidt tot minder behandelingsgerelateerde bijwerkingen.

Hoofdstuk 3 In dit hoofdstuk maken we gebruik van gepersonaliseerde anatomische kennis, opgebouwd tijdens de behandelingssessies, om de segmentatienauwkeurigheid van een voorgetraind CNN netwerk te verbeteren, voor een specifieke patiënt. We hebben een transfer-learning aanpak onderzocht, waarbij we het baseline CNN model hebben afgestemd op een specifieke patiënt, gebaseerd op beeldvorming verkregen in eerdere behandelingsfracties. Het baseline CNN model is getraind op een prostaat CT dataset van 379 patiënten van één ziekenhuis. Dit model wordt vervolgens verfijnd en getest op een onafhankelijke dataset van een ander ziekenhuis van 18 patiënten, die elk 7 tot 10 CT-scans per dag hebben. Voor de prostaat, de zaadblaasjes, de blaas en het rectum bereikte het op elke specifieke patiënt afgestemde model een gemiddelde oppervlakte-afstand (MSD) van respectievelijk $1,64 \pm 0,43$ mm, $2,38 \pm 2,76$ mm, $2,30 \pm 0,96$ mm, en $1,24 \pm 0,89$ mm, wat aanzienlijk beter was dan het basismodel.

De voorgestelde gepersonaliseerde modelaanpassing is daarom veelbelovend voor klinische toepassing in het kader van adaptieve radiotherapie van prostaatkanker.

Hoofdstuk 4 stelde voor om de registratie en segmentatie taken te combineren in een deep learning setting met behulp van adversarial learning. We beschouwden het geval waarin vaste en bewegende beelden en hun segmentaties beschikbaar zijn voor training, terwijl segmentaties niet beschikbaar zijn tijdens het testen; een veel voorkomend scenario in radiotherapie. Het voorgestelde raamwerk bestaat uit een 3D end-to-end generator netwerk dat zonder toezicht het vervormingsvectorveld (DVF) tussen vaste en bewegende beelden schat en dit DVF toepast op het bewegende beeld en de segmentatie daarvan. Een discriminator-netwerk wordt getraind om te evalueren hoe goed het bewegende beeld en de segmentatie overeenkomen met het vaste beeld en de segmentatie. Het voorgestelde netwerk werd getraind en geëvalueerd op follow-up prostaat CT-scans voor beeldgestuurde radiotherapie, waarbij de planning CT-contouren worden doorgegeven aan de dagelijkse CT-beelden met behulp van de geschatte DVF. Het voorgestelde GAN-netwerk bereikte een MSD van respectievelijk $1,13 \pm 0,4$ mm, $1,81 \pm 1,6$ mm, $1,00 \pm 0,3$ mm, $2,21 \pm 1,3$ mm, en $2,29 \pm 2,0$ mm, wat significant beter was dan het deep learning basismodel en het conventionele algoritme. De inferentietijd van het voorgestelde model is 0,6 sec., waardoor real-time contourpropagatie mogelijk is, noodzakelijk voor online-adaptieve radiotherapie.

In **Hoofdstuk 5** hebben we registratie en segmentatie gezamenlijk geformuleerd via een deep learning-gebaseerde Multi-Task Learning setting, waardoor deze taken hun sterke punten kunnen benutten en hun zwakke punten kunnen verzachten door het delen van nuttige informatie. We stellen voor om deze taken niet alleen op loss niveau samen te voegen, maar ook op architecturaal niveau. De studie omvat twee datasets van verschillende fabrikanten en instituten. We hebben een uitgebreide kwantitatieve vergelijking uitgevoerd tussen de kwaliteit van de automatisch gegenereerde contouren van verschillende netwerkarchitecturen en loss wegingsmethoden. Bovendien hebben we de kwaliteit van het gegenereerde deformatievectorveld (DVF) geëvalueerd. We tonen aan dat MTL algoritmen beter presteren dan hun Single-Task Learning (STL) tegenhangers en een betere generalisatie bereiken op de onafhankelijke testset. Het beste algoritme bereikte een gemiddelde oppervlakte-afstand van $1,06 \pm 0,3$ mm, $1,27 \pm 0,4$ mm, $0,91 \pm 0,4$ mm, en $1,76 \pm 0,8$ mm op de validatie set voor de prostaat, zaadblaasjes, blaas, en rectum, respectievelijk. De hoge nauwkeurigheid van de voorgestelde methode gecombineerd met de snelle inferentiesnelheid, maakt het een veelbelovende methode voor automatische re-contouring van vervolgscans voor online adaptieve radiotherapie.

Hoofdstuk 6 presenteerde een snel MR-reconstructie-algoritme, dat de toepassing mogelijk maakt van de automatische contourmethodes die in de vorige hoofdstukken zijn voorgesteld voor online adaptieve MR-geleide radiotherapie. Uitgaande van

onderbemonsterde k-ruimtegegevens, wordt een iteratief, op leren gebaseerd reconstructieschema, geïnspireerd door compressed sensing theorie, gebruikt om de afbeeldingen te reconstrueren. We hebben een nieuw diep neuraal netwerk ontwikkeld om eerdere reconstructie-aannames te verfijnen en te corrigeren op basis van de trainingsgegevens. Het voorgestelde netwerk werd gerangschikt als #1, gedeeld #1 en #3 op respectievelijk de 8x versnelde multi-coil, de 4x multi-coil en de 4x single-coil tracks in de fastMRI-wedstrijd van 2019, georganiseerd door Facebook en New York University (NYU). De gemiddelde looptijden voor het model zijn 0,5 en 0,3 seconden voor de multi-coil en de single-coil data, waardoor een snelle reconstructie voor MR-geleide adaptieve radiotherapie mogelijk wordt. Deze superieure prestatie in termen van reconstructiekwaliteit en tijd maakt het model een goede kandidaat voor MR-geleide adaptieve radiotherapie.

Discussie en toekomstig werk

Het doel van het werk dat in dit proefschrift wordt gepresenteerd was het ontwikkelen en onderzoeken van verschillende methoden van automatische contouring voor adaptieve radiotherapie. Ondanks het feit dat alle experimenten in dit proefschrift gevalideerd zijn op CT-scans van de prostaat, zijn alle voorgestelde methoden generiek en kunnen ze mogelijk ook worden toegepast op MR-beelden, waardoor zowel adaptieve CT als MR-geleide radiotherapie mogelijk is. Tot nu toe zou het resterende technische knelpunt voor het toepassen van deze methoden op MRIgRT de snelle reconstructie van de MR-beelden zijn, die we in het laatste hoofdstuk van het proefschrift hebben besproken.

In de hoofdstukken 4 en 5, hebben wij aangetoond dat het combineren van segmentatie- en registratietaken zeer gunstig en klinisch belangrijk is, vooral voor toepassingen waarbij de output van beide taken vereist is, zoals voor dosisaccumulatie en dosisplanning. Wij hebben geconcludeerd dat het beste mechanisme om beide taken te combineren is door deze taken te modelleren met behulp van een kruissteeknetwerk dat interne parameters over de taken deelt. Met behulp van een kruissteeknetwerk waren we in staat om superieure prestaties te bereiken in de orde van grootte van 1 mm voor de gemiddelde oppervlakteafstand van de doelorganen en OAR's. Op de onafhankelijke testreeks behaalde de diepe gezamenlijke methode echter een resultaat in de orde van grootte van 2 mm. Vergeleken met de prestatie van de hybride methode, gepresenteerd in hoofdstuk 2, had de hybride methode een betere generalisatie, omdat het een niet-leerbare methode is en beeldparen op een iteratieve manier registreert. De looptijd van de hybride methode ligt echter in de orde van minuten, terwijl de gezamenlijke methode er minder dan een seconde over doet, wat het een betere kandidaat maakt voor online adaptieve radiotherapie. De generalisatie van de diepe gezamenlijke methode kan echter verder worden

verbeterd door een van de volgende strategieën toe te passen. Ten eerste kunnen we het model opnieuw trainen en patiënten includeren die vergelijkbaar zijn met de verdeling van de testgegevens. Op die manier zou men het algemene model kunnen aanpassen aan, bijvoorbeeld, een specifiek ziekenhuis. Ten tweede kunnen we de bestaande trainingsgegevens uitbreiden met realistische vervormingen in plaats van willekeurige, aangezien dergelijke willekeurige vervormingen geen verbetering vertoonden in onze eerdere experimenten. Deze realistische vervormingen kunnen worden ontleend aan een generatief mechanisch model zoals gepresenteerd in [177, 178]. Het gebruik van een dergelijke realistische vergroting zou het netwerk helpen om verschillende soorten en schalen van vervormingen te leren die misschien niet beschikbaar zijn in de oorspronkelijke trainingsdataset. Tenslotte kunnen we een transfer-leerstrategie toepassen die vergelijkbaar is met die voorgesteld in Hoofdstuk 3, waar we voorstelden om gebruik te maken van gepersonaliseerde anatomische kennis die tijdens de behandelingssessies werd opgebouwd. In dat hoofdstuk toonden we aan dat het aanpassen van het model aan een specifieke anatomie van de patiënt de prestaties van het netwerk kan verbeteren, vooral voor de organen die niet veel vervormen ten opzichte van de planningscan, zoals te zien is in Figuur 3.1. Voor de organen die tussen de sessies vervormen, zoals het rectum en de blaas, is de prestatie van het aangepaste netwerk niet significant verbeterd, zoals blijkt uit Tabel 3.2. Deze aanpassingsstrategie is doeltreffend voor elk bestaand model, en kan dus worden gebruikt om een vooraf geleverd model op interne gegevens aan te passen of te personaliseren zonder dat het nodig is het model van nul af aan opnieuw te trainen. Dit zou de prestaties van het model in de loop van de tijd verbeteren tegen minimale kosten en met behoud van de privacy van de gegevens. Door de transfer-leerstrategie aan te passen kan het bovendien worden gebruikt om de voorspelling van het model voortdurend te verbeteren nadat het door klinische radiologen is gecorrigeerd.

De netwerken die in de hoofdstukken 2, 3, 4 en 5 zijn ontwikkeld, waren in de eerste plaats gericht op automatische contourpropagatie voor prostaat CT. Deze netwerken kunnen echter worden gegeneraliseerd naar MR-beelden voor MRIgRT. Verschillende CNN netwerken met een U-Net variant architectuur zijn reeds voorgesteld in de literatuur voor MR prostaat segmentatie [179, 180, 181, 182] evenals andere anatomieën zoals de hersenen [183, 184, 185] en het hart [186, 187]. Bovendien werden soortgelijke netwerken met succes gebruikt voor MR-registratie op verschillende anatomische plaatsen [16, 117, 188]. Het succes van de eerder genoemde methoden die gebruikmaken van een vergelijkbare netwerkarchitectuur, maakt het veelbelovend dat de door ons voorgestelde diepe gezamenlijke netwerken kunnen worden gegeneraliseerd naar MR-beelden van de prostaat en mogelijk ook voor andere anatomische locaties. Om die hypothese te valideren, zouden we een studie moeten uitvoeren met meerdere datasets van verschillende anatomische sites en

beeldvormingsmodaliteiten.

Voor MRIgRT toepassingen vormen de acquisitie en reconstructietijd van de MR-beelden zelf nog steeds een knelpunt. In hoofdstuk 6 hebben we dit probleem aangepakt, waarbij we een snel reconstructie-algoritme hebben ontwikkeld dat zowel voor single-coil als multi-coil MR-beelden werkt. De output van het netwerk was klinisch acceptabel op basis van de klinische evaluatie uitgevoerd door radiologen. Het resultaat van deze klinische evaluatie maakt het veelbelovend en bemoedigend om te worden ingezet in klinische proeven na het doen van verdere klinische beoordeling op verschillende anatomische plaatsen.

In termen van de klinische gereedheid van het voorgestelde diepe gezamenlijke registratie- en segmentatienetwerk, hebben we in Hoofdstuk 2 een uitgebreide dosimetrische evaluatie uitgevoerd op de automatisch gegenereerde contouren van de hybride methode. We ontdekten dat het verbeteren van de kwaliteit van de gegenereerde contouren in termen van MSD, resulteerde in een verhoging van de dosimetrische maatregelen in termen van V_{95} en de Conservative Success Rate (CSR) vergeleken met wanneer alleen op intensiteit gebaseerde registratie werd gebruikt. Aangezien het kruissteeknetwerk een nog betere geometrische prestatie behaalde, veronderstellen we dat deze verbetering ook zou resulteren in een boost in de overeenkomstige dosimetrische metingen en we valideren momenteel deze hypothese.

Een veelbelovende richting voor toekomstig onderzoek voor het gezamenlijke netwerk is de toevoeging van een derde taak, mogelijk het schatten van het dosisplan voor radiotherapie. Zo kunnen we contouren genereren die consistent zijn met een optimale dosisplanning. Verdere studies kunnen zich ook richten op geavanceerde MTL-netwerkarchitecturen zoals sluisnetwerken [132] of routeringsnetwerken [133]. Bovendien kunnen we onderzoeken hoe we de contouren uit de segmentatie- en registratiepaden slimmer kunnen samenvoegen in plaats van er simpelweg één te selecteren op basis van de validatieset. Het is ook de moeite waard om semi-gesuperviseerde trainingstechnieken [189] te onderzoeken om de generaliseerbaarheid van de voorgestelde netwerken te verbeteren. Voor het snelle MR-reconstructienetwerk, als toekomstige richting, zal het gunstig zijn om beter te begrijpen in hoeverre het netwerk afhankelijk is van de priors door interpreteerbare AI-technieken toe te passen, zoals differentieerbare beeldparameterisaties voor featurevisualisatie [175]. Sterker gebruik van de priors via de loss-functie is een extra optie. Gezien het einddoel van MRIgRT, namelijk het extraheren van beeldcontouren en het vervolgens genereren van een dosisplanning, zou het interessant zijn om te onderzoeken of we de beeldkwaliteit kunnen opofferen zonder verlies van segmentatie- en registratieprestaties, wat de MR-beeldvorming zelfs nog verder kan versnellen, gaande van minuten tot seconden. Dit zou in de toekomst kunnen leiden tot een bijna realtime stuur- en controlemodaliteit, wat voordelen heeft voor sneller bewegende organen zoals de longen en het hart.

Algemene conclusies

Samengevat stelt dit proefschrift een op deep learning gebaseerde automatische contourmethode voor real-time adaptieve radiotherapie voor. De voorgestelde netwerken werden geëvalueerd op CT-beelden van de prostaat, een veelgebruikte modaliteit voor behandelplanning, maar kan worden gegeneraliseerd voor MR-beelden. Daarnaast hebben we een snel MR-reconstructie-algoritme voorgesteld om de MR-acquisitie te versnellen, zodat onze modellen mogelijk ook kunnen worden gebruikt voor MR-geleide adaptieve radiotherapie. Alle deep learning-methoden die in dit proefschrift worden voorgesteld, hebben een looptijd van minder dan een seconde, waardoor real-time automatische contouren mogelijk zijn die nodig zijn voor online adaptieve radiotherapie.

Bibliography

[1] *National Cancer Institute. Cancer Statistics*. URL: https://www.cancer.gov/about-cancer/understanding/statistics.

[2] *National Cancer Institute. Radiation Therapy to Treat Cancer*. URL: https://www.cancer.gov/about-cancer/treatment/types/radiation-therapy.

[3] J. C. Jagodinsky, P. M. Harari, and Z. S. Morris. "The promise of combining radiation therapy with immunotherapy". In: *International Journal of Radiation Oncology* Biology* Physics* 108.1 (2020), pages 6–16.

[4] M Zhang, D. Westerly, and T. Mackie. "Introducing an on-line adaptive procedure for prostate image guided intensity modulate proton therapy". In: *Physics in Medicine & Biology* 56.15 (2011), page 4947.

[5] J.-J. Sonke, M. Aznar, et al. "Adaptive radiotherapy for anatomical changes". In: *Seminars in radiation oncology*. Volume 29. 3. Elsevier. 2019, pages 245–257.

[6] A. Lomax. "Intensity modulated proton therapy and its sensitivity to treatment uncertainties 1: the potential effects of calculational uncertainties". In: *Physics in Medicine & Biology* 53.4 (2008), page 1027.

[7] D. Boehmer, P. Maingon, P. Poortmans, M.-H. Baron, R. Miralbell, V. Remouchamps, C. Scrase, A. Bossi, and M. Bolla. "Guidelines for primary radiotherapy of patients with prostate cancer". In: *Radiotherapy and Oncology* 79.3 (2006), pages 259–269.

[8] C. Salembier, G. Villeirs, B. De Bari, P. Hoskin, B. R. Pieters, M. Van Vulpen, V. Khoo, A. Henry, A. Bossi, G. De Meerleer, et al. "ESTRO ACROP consensus guideline on CT-and MRI-based target volume delineation for primary radiation therapy of localized prostate cancer". In: *Radiotherapy and Oncology* 127.1 (2018), pages 49–61.

[9] N. K. Jensen, D. Mulder, M. Lock, B. Fisher, R. Zener, B. Beech, R. Kozak, J. Chen, T.-Y. Lee, and E. Wong. "Dynamic contrast enhanced CT aiding gross tumor volume delineation of liver tumors: an interobserver variability study". In: *Radiotherapy and Oncology* 111.1 (2014), pages 153–157.

[10] A. C. Riegel, J. G. Antone, H. Zhang, P. Jain, J. Raince, A. Rea, A. M. Bergamo, A. Kapur, and L. Potters. "Deformable image registration and interobserver variation in contour propagation for radiation therapy planning". In: *Journal of Applied Clinical Medical Physics* 17.3 (2016), pages 347–357.

[11] P. Kupelian, T. Willoughby, A. Mahadevan, T. Djemil, G. Weinstein, S. Jani, C. Enke, T. Solberg, N. Flores, D. Liu, et al. “Multi-institutional clinical experience with the Calypso System in localization and continuous, real-time monitoring of the prostate gland during external radiotherapy”. In: *International Journal of Radiation Oncology* Biology* Physics* 67.4 (2007), pages 1088–1098.

[12] J. Krebs, H. Delingette, B. Mailhé, N. Ayache, and T. Mansi. “Learning a probabilistic model for diffeomorphic registration”. In: *IEEE Transactions on Medical Imaging* 38.9 (2019), pages 2165–2176.

[13] H. Sokooti, B. de Vos, F. Berendsen, M. Ghafoorian, S. Yousefi, B. P. Lelieveldt, I. Isgum, and M. Staring. “3D convolutional neural networks image registration based on efficient supervised learning from artificial deformations”. In: *arXiv preprint arXiv:1908.10235* (2019).

[14] M.-M. Rohé, M. Datar, T. Heimann, M. Sermesant, and X. Pennec. “SVF-Net: Learning deformable image registration using shape matching”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2017, pages 266–274.

[15] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu. “Spatial transformer networks”. In: *arXiv preprint arXiv:1506.02025* (2015).

[16] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, and A. V. Dalca. “VoxelMorph: a learning framework for deformable medical image registration”. In: *IEEE Transactions on Medical Imaging* 38.8 (2019), pages 1788–1800.

[18] B. D. de Vos, F. F. Berendsen, M. A. Viergever, H. Sokooti, M. Staring, and I. Išgum. “A deep learning framework for unsupervised affine and deformable image registration”. In: *Medical Image Analysis* 52 (2019), pages 128–143.

[19] B. D. de Vos, F. F. Berendsen, M. A. Viergever, M. Staring, and I. Išgum. “End-to-end unsupervised deformable image registration with a convolutional neural network”. In: *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer, 2017, pages 204–212.

[20] Y. Fu, Y. Lei, T. Wang, W. J. Curran, T. Liu, and X. Yang. “Deep learning in medical image registration: a review”. In: *Physics in Medicine & Biology* 65.20 (2020), 20TR01.

[21] Y. Huo, Z. Xu, Y. Xiong, K. Aboud, P. Parvathaneni, S. Bao, C. Bermudez, S. M. Resnick, L. E. Cutting, and B. A. Landman. “3D whole brain segmentation using spatially localized atlas network tiles”. In: *NeuroImage* 194 (2019), pages 105–119.

[22] H. Wang, Y. Cao, and T. Syeda-Mahmood. “Multi-atlas segmentation with learning-based label fusion”. In: *International Workshop on Machine Learning in Medical Imaging*. Springer. 2014, pages 256–263.

[24] D. Mahapatra, Z. Li, F. Vos, and J. Buhmann. “Joint segmentation and groupwise registration of cardiac dce mri using sparse data representations”. In: *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2015, pages 1312–1315.

[25] A. J. Woerner, M. Choi, M. M. Harkenrider, J. C. Roeske, and M. Surucu. “Evaluation of deformable image registration-based contour propagation from planning CT to cone-beam CT”. In: *Technology in Cancer Research & Treatment* 16.6 (2017), pages 801–810.

[26] X. Gu, B. Dong, J. Wang, J. Yordy, L. Mell, X. Jia, and S. B. Jiang. “A contour-guided deformable image registration algorithm for adaptive radiotherapy”. In: *Physics in Medicine & Biology* 58.6 (2013), page 1889.

[27] J. L. Bedford, M. F. Fast, S. Nill, F. M. McDonald, M. Ahmed, V. N. Hansen, and U. Oelfke. “Effect of MLC tracking latency on conformal volumetric modulated arc therapy (VMAT) plans in 4D stereotactic lung treatment”. In: *Radiotherapy and Oncology* 117.3 (2015), pages 491–495.

[28] D. L. Donoho. “Compressed sensing”. In: *IEEE Transactions on Information Theory* 52.4 (2006), pages 1289–1306.

[29] M. Lustig, D. Donoho, and J. M. Pauly. “Sparse MRI: The application of compressed sensing for rapid MR imaging”. In: *Magnetic Resonance in Medicine* 58.6 (2007), pages 1182–1195.

[30] E. J. Candes, Y. C. Eldar, D. Needell, and P. Randall. “Compressed sensing with coherent and redundant dictionaries”. In: *Applied and Computational Harmonic Analysis* 31.1 (2011), pages 59–73.

[31] Z.-P. Liang, F. Boada, R. Constable, E. Haacke, P. Lauterbur, and M. Smith. “Constrained reconstruction methods in MR imaging”. In: *Reviews in Magnetic Resonance in Medicine* 4.2 (1992), pages 67–185.

[32] G. Wang, J. C. Ye, K. Mueller, and J. A. Fessler. “Image reconstruction is a new frontier of machine learning”. In: *IEEE Transactions on Medical Imaging* 37.6 (2018), pages 1289–1296.

[33] K. Hammernik, T. Klatzer, E. Kobler, M. P. Recht, D. K. Sodickson, T. Pock, and F. Knoll. “Learning a variational network for reconstruction of accelerated MRI data”. In: *Magnetic Resonance in Medicine* 79.6 (2018), pages 3055–3071.

[34] J. Sun, H. Li, et al. “Deep ADMM-Net for compressive sensing MRI”. In: *Advances in Neural Information Processing Systems*. 2016, pages 10–18.

[35] L. Kerkmeijer, M Maspero, G. Meijer, J. v. d. V. van Zyp, H. De Boer, and C. van den Berg. “Magnetic resonance imaging only workflow for radiotherapy simulation and planning in prostate cancer”. In: *Clinical Oncology* 30.11 (2018), pages 692–701.

[36] *National Cancer Society. Cancer Stat Facts: Prostate Cancer*. URL: `https://seer.cancer.gov/statfacts/html/prost.html`.

[37] H. Kooy and C Grassberger. “Intensity modulated proton therapy”. In: *The British Journal of Radiology* 88.1051 (2015), page 20150195.

[38] F. Van den Heuvel, J. Fugazzi, E. Seppi, and J. D. Forman. “Clinical application of a repositioning scheme, using gold markers and electronic portal imaging”. In: *Radiotherapy and Oncology* 79.1 (2006), pages 94–100.

[39] M. S. Hoogeman, M. van Herk, J. de Bois, and J. V. Lebesque. “Strategies to reduce the systematic error due to tumor and rectum motion in radiotherapy of prostate cancer”. In: *Radiotherapy and Oncology* 74.2 (2005), pages 177–185.

[40] E. K. Hansen, M. K. Bucci, J. M. Quivey, V. Weinberg, and P. Xia. “Repeat CT imaging and replanning during the course of IMRT for head-and-neck cancer”. In: *International Journal of Radiation Oncology* Biology* Physics* 64.2 (2006), pages 355–362.

[41] T. Jagt, S. Breedveld, R. Van Haveren, B. Heijmen, and M. Hoogeman. “An automated planning strategy for near real-time adaptive proton therapy in prostate cancer”. In: *Physics in Medicine & Biology* 63.13 (2018), page 135017.

[42] Y. Qiao. “Fast optimization methods for image registration in adaptive radiation therapy”. PhD thesis. Ph. D. thesis, Leiden University Medical Center, 2017.

[43] M. Thor, J. B. Petersen, L. Bentzen, M. Høyer, and L. P. Muren. “Deformable image registration for contour propagation from CT to cone-beam CT scans in radiotherapy of prostate cancer”. In: *Acta Oncologica* 50.6 (2011), pages 918–925.

[44] A. J. Woerner, M. Choi, M. M. Harkenrider, J. C. Roeske, and M. Surucu. “Evaluation of deformable image registration-based contour propagation from planning CT to cone-beam CT”. In: *Technology in Cancer Research & Treatment* 16.6 (2017), pages 801–810.

[45] S. Thörnqvist, J. B. Petersen, M. Høyer, L. N. Bentzen, and L. P. Muren. “Propagation of target and organ at risk contours in radiotherapy of prostate cancer using deformable image registration”. In: *Acta Oncologica* 49.7 (2010), pages 1023–1032.

[46] **Elmahdy, Mohamed S**, T. Jagt, S. Yousefi, H. Sokooti, R. Zinkstok, M. Hoogeman, and M. Staring. “Evaluation of multi-metric registration for online adaptive proton therapy of prostate cancer”. In: *International Workshop on Biomedical Image Registration*. Springer. 2018, pages 94–104.

[47] S. Mangar, J. Coffey, H. McNair, V. Hansen, S Sohaib, R Huddart, C Parker, A Horwich, and D Dearnelay. “Prostate radiotherapy: evaluating the effect of bladder and rectal changes on prostate movement–a CT study”. In: *Trends Med Res* 1.1 (2006), pages 55–65.

[48] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger. “3D U-Net: learning dense volumetric segmentation from sparse annotation”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2016, pages 424–432.

[49] F. Milletari, N. Navab, and S.-A. Ahmadi. “V-net: Fully convolutional neural networks for volumetric medical image segmentation”. In: *2016 fourth international conference on 3D vision (3DV)*. IEEE. 2016, pages 565–571.

[50] D. P. Kingma and J. Ba. “Adam: A method for stochastic optimization”. In: *arXiv preprint arXiv:1412.6980* (2014).

[51] S. Gao, L. Zhang, H. Wang, R. De Crevoisier, D. D. Kuban, R. Mohan, and L. Dong. “A deformable image registration method to handle distended rectums in prostate cancer radiotherapy”. In: *Medical Physics* 33.9 (2006), pages 3304–3312.

[52] M. Foskey, B. Davis, L. Goyal, S. Chang, E. Chaney, N. Strehl, S. Tomei, J. Rosenman, and S. Joshi. “Large deformation three-dimensional image registration in image-guided radiation therapy”. In: *Physics in Medicine & Biology* 50.24 (2005), page 5869.

[53] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez. “A survey on deep learning in medical image analysis”. In: *Medical Image Analysis* 42 (2017), pages 60–88.

[54] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. “Generative adversarial networks”. In: *arXiv preprint arXiv:1406.2661* (2014).

[55] Y. Xue, T. Xu, H. Zhang, L. R. Long, and X. Huang. “Segan: Adversarial network with multi-scale l 1 loss for medical image segmentation”. In: *Neuroinformatics* 16.3 (2018), pages 383–392.

[56] D. Nie, R. Trullo, J. Lian, C. Petitjean, S. Ruan, Q. Wang, and D. Shen. “Medical image synthesis with context-aware generative adversarial networks”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2017, pages 417–425.

[57] Y. Hu, E. Gibson, N. Ghavami, E. Bonmati, C. M. Moore, M. Emberton, T. Vercauteren, J. A. Noble, and D. C. Barratt. “Adversarial deformation regularization for training image registration neural networks”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2018, pages 774–782.

[58] Q. Yang, P. Yan, Y. Zhang, H. Yu, Y. Shi, X. Mou, M. K. Kalra, Y. Zhang, L. Sun, and G. Wang. “Low-dose CT image denoising using a generative adversarial network with Wasserstein distance and perceptual loss”. In: *IEEE Transactions on Medical Imaging* 37.6 (2018), pages 1348–1357.

[59] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang. “Generative image inpainting with contextual attention”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pages 5505–5514.

[60] S. Iizuka, E. Simo-Serra, and H. Ishikawa. “Globally and locally consistent image completion”. In: *ACM Transactions on Graphics (ToG)* 36.4 (2017), pages 1–14.

[61] B Rodriguez-Vila, F Garcia-Vicente, and E. Gomez. “Methodology for registration of distended rectums in pelvic CT studies”. In: *Medical Physics* 39.10 (2012), pages 6351–6359.

[62] S. Klein, M. Staring, K. Murphy, M. A. Viergever, and J. P. Pluim. “Elastix: a toolbox for intensity-based medical image registration”. In: *IEEE Transactions on Medical Imaging* 29.1 (2009), pages 196–205.

[63] M Staring, M. Bakker, J Stolk, D. Shamonin, J. Reiber, and B. Stoel. “Towards local progression estimation of pulmonary emphysema using CT”. In: *Medical Physics* 41.2 (2014), page 021905.

[64] W. Huizinga, S. Klein, and D. H. Poot. “Fast multidimensional B-spline interpolation using template metaprogramming”. In: *International Workshop on Biomedical Image Registration*. Springer. 2014, pages 11–20.

[65] Y. Qiao, B. van Lew, B. P. Lelieveldt, and M. Staring. “Fast automatic step size estimation for gradient descent optimization of image registration”. In: *IEEE Transactions on Medical Imaging* 35.2 (2015), pages 391–403.

[66] G. J. van der Wielen, T. F. Mutanga, L. Incrocci, W. J. Kirkels, E. M. V. Osorio, M. S. Hoogeman, B. J. Heijmen, and H. C. de Boer. “Deformation of prostate and seminal vesicles relative to intraprostatic fiducial markers”. In: *International Journal of Radiation Oncology* Biology* Physics* 72.5 (2008), pages 1604–1611.

[67] L. P. Muren, E. Wasbø, S. I. Helle, L. B. Hysing, À. Karlsdottir, O. H. Odland, H. Valen, R. Ekerold, and D. C. Johannessen. “Intensity-modulated radiotherapy of pelvic lymph nodes in locally advanced prostate cancer: planning procedures and early experiences”. In: *International Journal of Radiation Oncology* Biology* Physics* 71.4 (2008), pages 1034–1041.

[68] S. Breedveld, P. R. Storchi, P. W. Voet, and B. J. Heijmen. “iCycle: Integrated, multicriterial beam angle, and profile optimization for generation of coplanar and noncoplanar IMRT plans”. In: *Medical Physics* 39.2 (2012), pages 951–963.

[69] S. van de Water, A. Kraan, S. Breedveld, W. Schillemans, D. Teguh, H. Kooy, T. Madden, B. Heijmen, and M. Hoogeman. “Improved efficiency of multi-criteria IMPT treatment planning using iterative resampling of randomly placed pencil beams”. In: *Physics in Medicine & Biology* 58.19 (2013), page 6969.

[70] S. van de Water, H. M. Kooy, B. J. Heijmen, and M. S. Hoogeman. “Shortening delivery times of intensity modulated proton therapy by reducing proton energy layers during treatment plan optimization”. In: *International Journal of Radiation Oncology* Biology* Physics* 92.2 (2015), pages 460–468.

[71] S. Breedveld, P. R. Storchi, and B. J. Heijmen. “The equivalence of multi-criteria methods for radiotherapy plan optimization”. In: *Physics in Medicine & Biology* 54.23 (2009), page 7199.

[72] P. W. Voet, M. L. Dirkx, S. Breedveld, D. Fransen, P. C. Levendag, and B. J. Heijmen. “Toward fully automated multicriterial plan generation: a prospective clinical study”. In: *International Journal of Radiation Oncology* Biology* Physics* 85.3 (2013), pages 866–872.

[73] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, et al. “Tensorflow: Large-scale machine learning on heterogeneous distributed systems”. In: *arXiv preprint arXiv:1603.04467* (2016).

[74] K. H. Cha, L. Hadjiiski, R. K. Samala, H.-P. Chan, E. M. Caoili, and R. H. Cohan. “Urinary bladder segmentation in CT urography using deep-learning convolutional neural network and level sets”. In: *Medical Physics* 43.4 (2016), pages 1882–1896.

[75] X. Zhou, T. Ito, R. Takayama, S. Wang, T. Hara, and H. Fujita. “Three-dimensional CT image segmentation by combining 2D fully convolutional network with 3D majority voting”. In: *Deep Learning and Data Labeling for Medical Applications*. Springer, 2016, pages 111–120.

[76] K. Men, J. Dai, and Y. Li. “Automatic segmentation of the clinical target volume and organs at risk in the planning CT for rectal cancer using deep dilated convolutional neural networks”. In: *Medical Physics* 44.12 (2017), pages 6377–6389.

[77] S. Thörnqvist, L. Bentzen, J. B. Petersen, L. B. Hysing, and L. P. Muren. “Plan robustness of simultaneous integrated boost radiotherapy of prostate and lymph nodes for different image-guidance and delivery techniques”. In: *Acta Oncologica* 50.6 (2011), pages 926–934.

[78] D. Yan, F. Vicini, et al. “Adaptive radiation therapy”. In: *Physics in Medicine & Biology* 42.1 (1997), page 123.

[79] N. Tong, S. Gou, et al. “Fully automatic multi-organ segmentation for head and neck cancer radiotherapy using shape representation model constrained fully convolutional neural networks”. In: *Medical Physics* 45.10 (2018), pages 4558–4567.

[80] S. J. Reddi, S. Kale, et al. “On the convergence of adam and beyond”. In: *arXiv preprint arXiv:1904.09237* (2019).

[81] C. Lu, S. Chelikani, X. Papademetris, J. P. Knisely, M. F. Milosevic, Z. Chen, D. A. Jaffray, L. H. Staib, and J. S. Duncan. “An integrated approach to segmentation and nonrigid registration for application in image-guided pelvic radiotherapy”. In: *Medical Image Analysis* 15.5 (2011), pages 772–785.

[82] A. Yezzi, L. Zöllei, and T. Kapur. “A variational framework for integrating segmentation and registration through active contours”. In: *Medical Image Analysis* 7.2 (2003), pages 171–185.

[83] G. Unal and G. Slabaugh. “Coupled PDEs for non-rigid registration and segmentation”. In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05)*. Volume 1. IEEE. 2005, pages 168–175.

[84] S. Kazeminia, C. Baur, A. Kuijper, B. van Ginneken, N. Navab, S. Albarqouni, and A. Mukhopadhyay. “GANs for medical image analysis”. In: *Artificial Intelligence in Medicine* (2020), page 101938.

[85] G. Haskins, U. Kruger, and P. Yan. “Deep learning in medical image registration: a survey”. In: *Machine Vision and Applications* 31.1 (2020), pages 1–18.

[86] D. Mahapatra, Z. Ge, S. Sedai, and R. Chakravorty. “Joint registration and segmentation of xray images using generative adversarial networks”. In: *International Workshop on Machine Learning in Medical Imaging*. Springer. 2018, pages 73–80.

[87] B. D. de Vos, F. F. Berendsen, M. A. Viergever, H. Sokooti, M. Staring, and I. Išgum. “A deep learning framework for unsupervised affine and deformable image registration”. In: *Medical Image Analysis* 52 (2019), pages 128–143.

[88] M. Arjovsky, S. Chintala, and L. Bottou. “Wasserstein generative adversarial networks”. In: *International Conference on Machine Learning*. PMLR. 2017, pages 214–223.

[89] O. Ronneberger, P. Fischer, and T. Brox. “U-net: Convolutional networks for biomedical image segmentation”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2015, pages 234–241.

[90] E. Gibson, W. Li, C. Sudre, L. Fidon, D. I. Shakir, G. Wang, Z. Eaton-Rosen, R. Gray, T. Doel, Y. Hu, et al. “NiftyNet: a deep-learning platform for medical imaging”. In: *Computer methods and programs in biomedicine* 158 (2018), pages 113–122.

[91] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. “Image-to-image translation with conditional adversarial networks”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pages 1125–1134.

[92] M. Nilashi, N. Ahmadi, S. Samad, L. Shahmoradi, H. Ahmadi, O. Ibrahim, S. Asadi, R. Abdullah, R. A. Abumalloh, and E. Yadegaridehkordi. “Disease diagnosis using machine learning techniques: A review and classification”. In: *Journal of Soft Computing and Decision Support Systems* 7.1 (), pages 19–30.

[93] D. Shen, G. Wu, and H.-I. Suk. “Deep learning in medical image analysis”. In: *Annual review of biomedical engineering* 19 (2017), pages 221–248.

[94] D. Rueckert and J. A. Schnabel. “Registration and segmentation in medical imaging”. In: *Registration and Recognition in Images and Videos*. Springer, 2014, pages 137–156.

[95] E. K. Hansen, M. K. Bucci, J. M. Quivey, V. Weinberg, and P. Xia. “Repeat CT imaging and replanning during the course of IMRT for head-and-neck cancer”. In: *International Journal of Radiation Oncology* Biology* Physics* 64.2 (2006), pages 355–362.

[96] K. K. Brock. “Adaptive radiotherapy: moving into the future”. In: *Seminars in radiation oncology*. Volume 29. 3. NIH Public Access. 2019, page 181.

[97] J.-J. Sonke, M. Aznar, and C. Rasch. “Adaptive radiotherapy for anatomical changes”. In: *Seminars in radiation oncology*. Volume 29. 3. Elsevier. 2019, pages 245–257.

[99] C. Lu, S. Chelikani, X. Papademetris, J. P. Knisely, M. F. Milosevic, Z. Chen, D. A. Jaffray, L. H. Staib, and J. S. Duncan. “An integrated approach to segmentation and nonrigid registration for application in image-guided pelvic radiotherapy”. In: *Medical Image Analysis* 15.5 (2011), pages 772–785.

[100] K. M. Pohl, J. Fisher, W. E. L. Grimson, R. Kikinis, and W. M. Wells. “A Bayesian model for joint segmentation and registration”. In: *NeuroImage* 31.1 (2006), pages 228–239.

[101] A. Yezzi, L. Zöllei, and T. Kapur. “A variational framework for integrating segmentation and registration through active contours”. In: *Medical Image Analysis* 7.2 (2003), pages 171–185.

[102] G. Unal and G. Slabaugh. “Coupled PDEs for non-rigid registration and segmentation”. In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05)*. Volume 1. IEEE. 2005, pages 168–175.

[103] S. Yousefi, H. Sokooti, **Elmahdy, Mohamed S**, I. M. Lips, M. T. M. Shalmani, R. T. Zinkstok, F. J. Dankers, and M. Staring. “Esophageal Tumor Segmentation in CT Images using a 3D Convolutional Neural Network”. In: *arXiv preprint arXiv:2012.03242* (2020).

[104] T. Kiljunen, S. Akram, J. Niemelä, E. Löyttyniemi, J. Seppälä, J. Heikkilä, K. Vuolukka, O.-S. Kääriäinen, V.-P. Heikkilä, K. Lehtiö, et al. “A Deep Learning-Based Automated CT Segmentation of Prostate Cancer Anatomy for Radiation Therapy Planning-A Retrospective Multicenter Study”. In: *Diagnostics* 10.11 (2020), page 959.

[105] X. Cao, J. Yang, L. Wang, Z. Xue, Q. Wang, and D. Shen. “Deep learning based inter-modality image registration supervised by intra-modality similarity”. In: *International Workshop on Machine Learning in Medical Imaging*. Springer. 2018, pages 55–63.

[107] P. Tschandl, N. Codella, B. N. Akay, G. Argenziano, R. P. Braun, H. Cabo, D. Gutman, A. Halpern, B. Helba, R. Hofmann-Wellenhof, et al. “Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: an open, web-based, international, diagnostic study”. In: *The Lancet Oncology* 20.7 (2019), pages 938–947.

[108] D. Ardila, A. P. Kiraly, S. Bharadwaj, B. Choi, J. J. Reicher, L. Peng, D. Tse, M. Etemadi, W. Ye, G. Corrado, et al. “End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography”. In: *Nature Medicine* 25.6 (2019), pages 954–961.

[109] L. Hu, D. Bell, S. Antani, Z. Xue, K. Yu, M. P. Horning, N. Gachuhi, B. Wilson, M. S. Jaiswal, B. Befano, et al. “An observational study of deep learning and automated evaluation of cervical images for cancer screening”. In: *JNCI: Journal of the National Cancer Institute* 111.9 (2019), pages 923–932.

[110] R. H. Mak, M. G. Endres, J. H. Paik, R. A. Sergeev, H. Aerts, C. L. Williams, K. R. Lakhani, and E. C. Guinan. “Use of crowd innovation to develop an artificial intelligence–based solution for radiation therapy targeting”. In: *JAMA oncology* 5.5 (2019), pages 654–661.

[111] Y. Hu, M. Modat, E. Gibson, N. Ghavami, E. Bonmati, C. M. Moore, M. Emberton, J. A. Noble, D. C. Barratt, and T. Vercauteren. “Label-driven weakly-supervised learning for multimodal deformable image registration”. In: *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE. 2018, pages 1070–1074.

[112] D. Mahapatra, Z. Ge, S. Sedai, and R. Chakravorty. “Joint registration and segmentation of xray images using generative adversarial networks”. In: *International Workshop on Machine Learning in Medical Imaging*. Springer. 2018, pages 73–80.

[113] Z. Xu and M. Niethammer. “DeepAtlas: Joint semi-supervised learning of image registration and segmentation”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2019, pages 420–429.

[114] T. Estienne, M. Vakalopoulou, S. Christodoulidis, E. Battistela, M. Lerousseau, A. Carre, G. Klausner, R. Sun, C. Robert, S. Mougiakakou, et al. “U-ReSNet: Ultimate coupling of registration and segmentation with deep nets”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2019, pages 310–319.

[115] F. Liu, J. Cai, Y. Huo, L. Lu, and A. P. Harrison. “JSSR: A Joint Synthesis, Segmentation, and Registration System for 3D Multi-Modal Image Alignment of Large-scale Pathological CT Scans”. In: *arXiv preprint arXiv:2005.12209* (2020).

[116] O. Ronneberger, P. Fischer, and T. Brox. “U-net: Convolutional networks for biomedical image segmentation”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2015, pages 234–241.

[117] J. Fan, X. Cao, P.-T. Yap, and D. Shen. “BIRNet: Brain image registration using dual-supervised fully convolutional networks”. In: *Medical Image Analysis* 54 (2019), pages 193–206.

[118] V. Nair and G. E. Hinton. “Rectified linear units improve restricted boltzmann machines”. In: *ICML*. 2010.

[119] S. Ioffe and C. Szegedy. “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift”. In: *Proceedings of the 32nd International Conference on Machine Learning*. Edited by F. Bach and D. Blei. Volume 37. Proceedings of Machine Learning Research. Lille, France: PMLR, 2015, pages 448–456.

[120] J. Baxter. “A model of inductive bias learning”. In: *Journal of Artificial Intelligence Research* 12 (2000), pages 149–198.

[121] E. Meyerson and R. Miikkulainen. “Pseudo-task augmentation: From deep multitask learning to intratask sharing—and back”. In: *International Conference on Machine Learning*. PMLR. 2018, pages 3511–3520.

[122] Y. S. Abu-Mostafa. “Learning from hints in neural networks”. In: *Journal of Complexity* 6.2 (1990), pages 192–198.

[123] Y. Zhang and Q. Yang. “A survey on multi-task learning”. In: *arXiv preprint arXiv:1707.08114* (2017).

[124] I. Misra, A. Shrivastava, A. Gupta, and M. Hebert. “Cross-stitch networks for multi-task learning”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pages 3994–4003.

[125] Z. Chen, V. Badrinarayanan, C.-Y. Lee, and A. Rabinovich. “Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks”. In: *International Conference on Machine Learning*. 2018, pages 794–803.

[126] A. Kendall, Y. Gal, and R. Cipolla. “Multi-task learning using uncertainty to weigh losses for scene geometry and semantics”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pages 7482–7491.

[127] S. Liu, E. Johns, and A. J. Davison. “End-to-end multi-task learning with attention”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019, pages 1871–1880.

[128] S. Klein, M. Staring, K. Murphy, M. A. Viergever, and J. P. Pluim. “Elastix: a toolbox for intensity-based medical image registration”. In: *IEEE Transactions on Medical Imaging* 29.1 (2009), pages 196–205.

[129] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, et al. “Tensorflow: Large-scale machine learning on heterogeneous distributed systems”. In: *arXiv preprint arXiv:1603.04467* (2016).

[130] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han. “On the variance of the adaptive learning rate and beyond”. In: *arXiv preprint arXiv:1908.03265* (2019).

[131] Y. Qiao. “Fast optimization methods for image registration in adaptive radiation therapy”. PhD thesis. Ph. D. thesis, Leiden University Medical Center, 2017.

[132] S. Ruder, J. Bingel, I. Augenstein, and A. Søgaard. “Sluice networks: Learning what to share between loosely related tasks”. In: *arXiv preprint arXiv:1705.08142* 2 (2017).

[133] C. Rosenbaum, T. Klinger, and M. Riemer. “Routing networks: Adaptive selection of non-linear functions for multi-task learning”. In: *arXiv preprint arXiv:1711.01239* (2017).

[134] M. Zaitsev, J. Maclaren, and M. Herbst. “Motion artifacts in MRI: A complex problem with many partial solutions”. In: *Journal of Magnetic Resonance Imaging* 42.4 (2015), pages 887–901.

[135] M. S. Cohen and R. M. Weisskoff. “Ultra-fast imaging”. In: *Magnetic Resonance Imaging* 9.1 (1991), pages 1–37.

[136] K. P. Pruessmann, M. Weiger, M. B. Scheidegger, and P. Boesiger. “SENSE: sensitivity encoding for fast MRI”. In: *Magnetic Resonance in Medicine* 42.5 (1999), pages 952–962.

[137] D. Liang, J. Cheng, Z. Ke, and L. Ying. “Deep MRI Reconstruction: Unrolled optimization algorithms meet neural networks”. In: *arXiv preprint arXiv:1907.11711* (2019).

[138] T. M. Quan, T. Nguyen-Duc, and W.-K. Jeong. “Compressed sensing MRI reconstruction using a generative adversarial network with a cyclic loss”. In: *IEEE Transactions on Medical Imaging* 37 (2018), pages 1488–1497.

[139] M. Mardani, E. Gong, J. Y. Cheng, S. S. Vasanawala, G. Zaharchuk, L. Xing, and et al. “Deep generative adversarial neural networks for compressive sensing MRI”. In: *IEEE Transactions on Medical Imaging* 38.1 (2018), pages 167–179.

[140] Y. Guo, C. Wang, H. Zhang, and G. Yang. “Deep Attentive Wasserstein Generative Adversarial Networks for MRI Reconstruction with Recurrent Context-Awareness”. In: *arXiv preprint arXiv:2006.12915* (2020).

[141] J. Schlemper, G. Yang, P. Ferreira, A. Scott, L.-A. McGill, Z. Khalique, M. Gorodezky, M. Roehl, J. Keegan, D. Pennell, et al. “Stochastic deep compressive sensing for the reconstruction of diffusion tensor cardiac MRI”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2018, pages 295–303.

[142] G. Yang, S. Yu, H. Dong, G. Slabaugh, P. L. Dragotti, X. Ye, F. Liu, S. Arridge, J. Keegan, Y. Guo, et al. “DAGAN: Deep de-aliasing generative adversarial networks for fast compressed sensing MRI reconstruction”. In: *IEEE Transactions on Medical Imaging* 37.6 (2017), pages 1310–1321.

[143] S. Yu, H. Dong, G. Yang, G. Slabaugh, P. L. Dragotti, X. Ye, F. Liu, S. Arridge, J. Keegan, D. Firmin, et al. “Deep de-aliasing for fast compressive sensing MRI”. In: *arXiv preprint arXiv:1705.07137* (2017).

[144] P. Putzky, D. Karkalousos, J. Teuwen, N. Miriakov, B. Bakker, M. Caan, and M. Welling. “i-RIM applied to the fastMRI challenge”. In: *arXiv preprint arXiv:1910.08952* (2019).

[145] P. Putzky and M. Welling. “Invert to learn to invert”. In: *Advances in Neural Information Processing Systems*. 2019, pages 446–456.

[146] B. Zhu, J. Z. Liu, S. F. Cauley, B. R. Rosen, and M. S. “Image reconstruction by domain-transform manifold learning”. In: *Nature* 555.7697 (2018), pages 487–492.

[147] D. Lee, J. Yoo, S. Tak, and J. C. Ye. “Deep residual learning for accelerated MRI using magnitude and phase networks”. In: *IEEE Transactions on Biomedical Engineering* 65.9 (2018), pages 1985–1995.

[148] J. Yang, Y. Zhang, and W. Yin. “A fast alternating direction method for TVL1-L2 signal reconstruction from partial Fourier data”. In: *IEEE Journal of Selected Topics in Signal Processing* 4.2 (2010), pages 288–297.

[149] H. K. Aggarwal, M. P. Mani, and M. Jacob. “MoDL: Model-based deep learning architecture for inverse problems”. In: *IEEE Transactions on Medical Imaging* 38.2 (2018), pages 394–405.

[150] Z. Ramzi, P. Ciuciu, and J.-L. Starck. “Benchmarking deep nets MRI reconstruction models on the fastMRI publicly available dataset”. In: *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2020, pages 1441–1445.

[151] J. Zbontar, F. Knoll, A. Sriram, M. J. Muckley, M. Bruno, A. Defazio, M. Parente, K. J. Geras, J. Katsnelson, H. Chandarana, et al. “fastMRI: An open dataset and benchmarks for accelerated MRI”. In: *arXiv preprint arXiv:1811.08839* (2018).

[152] J. Schlemper, J. Caballero, J. V. Hajnal, A. Price, and D. Rueckert. “A deep cascade of convolutional neural networks for MR image reconstruction”. In: *International Conference on Information Processing in Medical Imaging*. Springer. 2017, pages 647–658.

[153] T. Eo, Y. Jun, T. Kim, J. Jang, H.-J. Lee, and D. Hwang. “KIKI-net: cross-domain convolutional neural networks for reconstructing undersampled magnetic resonance images”. In: *Magnetic Resonance in Medicine* 80.5 (2018), pages 2188–2201.

[154] J. Adler and O. Öktem. “Learned primal-dual reconstruction”. In: *IEEE Transactions on Medical Imaging* 37.6 (2018), pages 1322–1332.

[155] J. Caballero, A. N. Price, D. Rueckert, and J. V. Hajnal. “Dictionary learning and time sparsity for dynamic MR data reconstruction”. In: *IEEE Transactions on Medical Imaging* 33.4 (2014), pages 979–994.

[156] A. Chambolle and T. Pock. “A first-order primal-dual algorithm for convex problems with applications to imaging”. In: *Journal of Mathematical Imaging and Vision* 40.1 (2011), pages 120–145.

[157] M. Seitzer, G. Yang, J. Schlemper, O. Oktay, T. Würfl, V. Christlein, T. Wong, R. Mohiaddin, D. Firmin, J. Keegan, et al. "Adversarial and perceptual refinement for compressed sensing MRI reconstruction". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2018, pages 232–240.

[158] J. Zhang and B. Ghanem. "ISTA-Net: Interpretable optimization-inspired deep network for image compressive sensing". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pages 1828–1837.

[159] F. Knoll, T. Murrell, A. Sriram, N. Yakubova, J. Zbontar, M. Rabbat, A. Defazio, M. J. Muckley, D. K. Sodickson, et al. "Advancing machine learning for MR image reconstruction with an open competition: Overview of the 2019 fastMRI challenge". In: *Magnetic Resonance in Medicine* (2020).

[160] F. Knoll, J. Zbontar, A. Sriram, M. J. Muckley, M. Bruno, A. Defazio, and et al. "fastMRI: A Publicly Available Raw k-Space and DICOM Dataset of Knee Images for Accelerated MR Image Reconstruction Using Machine Learning". In: *Radiology: Artificial Intelligence* 2.1 (2020).

[161] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. "Image quality assessment: from error visibility to structural similarity". In: *IEEE Transactions on Image Processing* 13.4 (2004), pages 600–612.

[162] A. Beck and M. Teboulle. "A fast iterative shrinkage-thresholding algorithm for linear inverse problems". In: *SIAM Journal on Imaging Sciences* 2.1 (2009), pages 183–202.

[163] D. L. Donoho, A. Maleki, and A. Montanari. "Message-passing algorithms for compressed sensing". In: *Proceedings of the National Academy of Sciences* 106.45 (2009), pages 18914–18919.

[164] A. Araujo, W. Norris, and J. Sim. "Computing Receptive Fields of Convolutional Neural Networks". In: *Distill* 4.11 (2019), e21.

[165] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. "Automatic differentiation in pytorch". In: (2017).

[166] J. Johnson, A. Alahi, and L. Fei-Fei. "Perceptual losses for real-time style transfer and super-resolution". In: *European Conference on Computer Vision*. Springer. 2016, pages 694–711.

[167] P. J. Huber. "Robust estimation of a location parameter". In: *Breakthroughs in statistics*. Springer, New York, NY, 1992, pages 492–518.

[168] Z. Wang, E. P. Simoncelli, and A. C. Bovik. "Multiscale structural similarity for image quality assessment". In: *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*. Volume 2. IEEE. 2003, pages 1398–1402.

[169] H. Zhao, O. Gallo, I. Frosio, and J. Kautz. "Loss functions for image restoration with neural networks". In: *IEEE Transactions on Computational Imaging* 3.1 (2016), pages 47–57.

[170] H. Zhao, O. Gallo, I. Frosio, and J. Kautz. “Loss functions for neural networks for image processing”. In: *arXiv preprint arXiv:1511.08861* (2015).

[171] M. Zhang, J. Lucas, J. Ba, and G. E. Hinton. “Lookahead Optimizer: k steps forward, 1 step back”. In: *Advances in Neural Information Processing Systems*. 2019, pages 9593–9604.

[172] Q. Tong, G. Liang, and J. Bi. “Calibrating the Learning Rate for Adaptive Gradient Methods to Improve Generalization Performance”. In: *arXiv preprint arXiv:1908.00700* (2019).

[173] D. P. Kingma and J. Ba. “Adam: A method for stochastic optimization”. In: *International Conference for Learning Representations* (2015).

[174] N. Pezzotti, E. de Weerdt, S. Yousefi, **Elmahdy, Mohamed S**, J. van Gemert, C. Schülke, and et al. “Adaptive-CS-Net: fastMRI with Adaptive Intelligence”. In: *arXiv preprint arXiv:1912.12259* (2019).

[175] A. Mordvintsev, N. Pezzotti, L. Schubert, and C. Olah. “Differentiable image parameterizations”. In: *Distill* 3.7 (2018), e12.

[176] A. Mason, J. Rioux, S. E. Clarke, A. Costa, M. Schmidt, V. Keough, T. Huynh, and S. Beyea. “Comparison of Objective Image Quality Metrics to Expert Radiologists’ Scoring of Diagnostic Quality of MR Images”. In: *IEEE Transactions on Medical Imaging* 39.4 (2020), pages 1064–1072.

[177] V. Sandfort, K. Yan, P. J. Pickhardt, and R. M. Summers. “Data augmentation using generative adversarial networks (CycleGAN) to improve generalizability in CT segmentation tasks”. In: *Scientific reports* 9.1 (2019), pages 1–9.

[178] M. Nakao, M. Nakamura, T. Mizowaki, and T. Matsuda. “Statistical deformation reconstruction using multi-organ shape features for pancreatic cancer localization”. In: *Medical Image Analysis* 67 (2021), page 101829.

[179] N. Ghavami, Y. Hu, E. Gibson, E. Bonmati, M. Emberton, C. M. Moore, and D. C. Barratt. “Automatic segmentation of prostate MRI using convolutional neural networks: Investigating the impact of network architecture on the accuracy of volume measurement and MRI-ultrasound registration”. In: *Medical Image Analysis* 58 (2019), page 101558.

[180] Z. Tian, L. Liu, Z. Zhang, and B. Fei. “PSNet: prostate segmentation on MRI based on a convolutional neural network”. In: *Journal of Medical Imaging* 5.2 (2018), page 021208.

[181] A. Ushinsky, M. Bardis, J. Glavis-Bloom, E. Uchio, C. Chantaduly, M. Nguyentat, D. Chow, P. D. Chang, and R. Houshyar. “A 3D-2D hybrid U-net convolutional neural network approach to prostate organ segmentation of multiparametric MRI”. In: *American Journal of Roentgenology* 216.1 (2021), pages 111–116.

[182] Q. Xiangxiang, Z. Yu, and Z. Bingbing. “Automated Segmentation Based on Residual U-Net Model for MR Prostate Images”. In: *2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*. IEEE. 2018, pages 1–6.

[183] D. Daimary, M. B. Bora, K. Amitab, and D. Kandar. “Brain tumor segmentation from MRI images using hybrid convolutional neural networks”. In: *Procedia Computer Science* 167 (2020), pages 2419–2428.

[184] M. Bateriwala and P. Bourgeat. “Enforcing temporal consistency in Deep Learning segmentation of brain MR images”. In: *arXiv preprint arXiv:1906.07160* (2019).

[185] K. Munir, F. Frezza, and A. Rizzi. “Brain Tumor Segmentation Using 2D-UNET Convolutional Neural Network”. In: *Deep Learning for Cancer Diagnosis*. Springer, 2021, pages 239–248.

[186] C. Liu, S. J. Gardner, N. Wen, M. A. Elshaikh, F. Siddiqui, B. Movsas, and I. J. Chetty. “Automatic segmentation of the prostate on CT images using deep neural networks (DNN)”. In: *International Journal of Radiation Oncology* Biology* Physics* 104.4 (2019), pages 924–932.

[187] F. Kong and S. C. Shadden. “A Generalizable Deep-Learning Approach for Cardiac Magnetic Resonance Image Segmentation Using Image Augmentation and Attention U-Net”. In: *International Workshop on Statistical Atlases and Computational Models of the Heart*. Springer. 2020, pages 287–296.

[188] S. Shan, W. Yan, X. Guo, E. I. Chang, Y. Fan, Y. Xu, et al. “Unsupervised end-to-end learning for deformable medical image registration”. In: *arXiv preprint arXiv:1711.08608* (2017).

[189] G. Bortsova, F. Dubost, L. Hogeweg, I. Katramados, and M. de Bruijne. “Semi-supervised medical image segmentation via learning consistency under transformations”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2019, pages 810–818.

In this appendix we provide a detailed results for the proposed methods and associated experiments in terms of DSC and %95 HD.

Table 1: The effect of network input for the different architectures on the validation set (HMC) in terms of DSC. Higher values are better. Here, $\oplus$ is the concatenation operation, and $\cdot\|\cdot$ represents the inputs to the segmentation network (left of $\|$) and the inputs to the registration network (right of $\|$).

			Prostate		Seminal vesicles		Rectum		Bladder	
Network	Input	Output path	$\mu \pm \sigma$	median	$\mu \pm \sigma$	median	$\mu \pm \sigma$	median	$\mu \pm \sigma$	median
Seg	I_f		0.84 ± 0.03	0.84	0.60 ± 0.14	0.62	0.75 ± 0.10	0.77	0.90 ± 0.07	0.93
	$I_f \oplus S_m$		0.85 ± 0.05	0.86	0.66 ± 0.16	0.72	0.79 ± 0.12	0.82	0.93 ± 0.03	0.94
	$I_f \oplus I_m$		0.66 ± 0.08	0.67	0.39 ± 0.21	0.40	0.39 ± 0.21	0.41	0.91 ± 0.08	0.93
	$I_f \oplus I_m \oplus S_m$		0.86 ± 0.04	0.87	0.64 ± 0.16	0.70	0.78 ± 0.08	0.78	0.93 ± 0.03	0.94
Reg	$I_f \oplus I_m$		0.85 ± 0.06	0.86	0.62 ± 0.18	0.68	0.79 ± 0.08	0.81	0.82 ± 0.10	0.84
	$I_f \oplus I_m \oplus S_m$		0.82 ± 0.08	0.83	0.60 ± 0.17	0.65	0.77 ± 0.08	0.80	0.79 ± 0.13	0.83
JRS-reg	$I_f \oplus I_m$		0.87 ± 0.04	0.87	0.68 ± 0.14	0.72	0.82 ± 0.06	0.84	0.87 ± 0.08	0.91
	$I_f \oplus I_m \oplus S_m$		0.87 ± 0.04	0.87	0.67 ± 0.15	0.72	0.83 ± 0.06	0.84	0.87 ± 0.08	0.91
Cross-stitch	$I_f \| \| I_f \oplus I_m$	Segmentation	0.85 ± 0.03	0.85	0.57 ± 0.19	0.60	0.81 ± 0.08	0.83	0.93 ± 0.05	0.94
		Registration	0.87 ± 0.03	0.88	0.67 ± 0.15	0.70	0.82 ± 0.06	0.84	0.87 ± 0.08	0.91
	$I_f \| \| I_f \oplus I_m \oplus S_m$	Segmentation	**0.88 ± 0.04**	0.88	**0.70 ± 0.11**	**0.74**	**0.86 ± 0.05**	**0.88**	**0.94 ± 0.02**	**0.95**
		Registration	0.87 ± 0.03	0.88	0.68 ± 0.15	0.73	0.84 ± 0.05	0.85	0.88 ± 0.08	0.91
	$I_f \oplus S_m \| \| I_f \oplus I_m \oplus S_m$	Segmentation	0.77 ± 0.11	0.79	0.52 ± 0.19	0.57	0.80 ± 0.05	0.80	0.93 ± 0.03	0.94
		Registration	0.85 ± 0.04	0.85	0.66 ± 0.14	0.72	0.80 ± 0.06	0.82	0.87 ± 0.08	0.90
	$I_f \oplus I_m \oplus S_m \| \| I_f \oplus I_m \oplus S_m$	Segmentation	0.88 ± 0.04	**0.89**	0.67 ± 0.15	0.72	0.85 ± 0.05	0.86	0.94 ± 0.03	0.95
		Registration	0.86 ± 0.04	0.87	0.67 ± 0.16	0.72	0.83 ± 0.06	0.84	0.88 ± 0.08	0.91

Table 2: The effect of network input for the different architectures on the validation set (HMC) in terms of %95 HD (mm). Lower values are better. Here, $\oplus$ is the concatenation operation, and $\cdot\|\cdot$ represents the inputs to the segmentation network (left of $\|$) and the inputs to the registration network (right of $\|$).

			Prostate		Seminal vesicles		Rectum		Bladder	
Network	Input	Output path	$\mu \pm \sigma$	median	$\mu \pm \sigma$	median	$\mu \pm \sigma$	median	$\mu \pm \sigma$	median
Seg	I_f		4.4 ± 1.0	4.4	8.6 ± 8.6	7.3	16.5 ± 11.0	13.3	6.9 ± 6.6	4.0
	$I_f \oplus S_m$		3.9 ± 1.4	3.6	5.9 ± 5.9	4.1	12.1 ± 9.7	8.9	4.3 ± 3.2	3.0
	$I_f \oplus I_m$		9.1 ± 2.3	8.7	14.9 ± 10.5	11.7	45.1 ± 17.3	41.8	5.3 ± 5.6	3.6
	$I_f \oplus I_m \oplus S_m$		3.8 ± 1.1	3.6	7.3 ± 9.2	4.2	11.5 ± 6.7	9.6	3.3 ± 1.5	3.0
Reg	$I_f \oplus I_m$		5.5 ± 4.5	4.0	5.6 ± 4.1	4.3	11.0 ± 6.4	9.4	15.7 ± 9.6	12.1
	$I_f \oplus I_m \oplus S_m$		7.7 ± 6.3	5.5	6.2 ± 4.2	4.8	11.6 ± 6.8	9.2	17.0 ± 9.5	14.7
JRS-reg	$I_f \oplus I_m$		3.6 ± 1.3	3.0	4.5 ± 3.0	3.3	9.6 ± 5.7	8.2	13.1 ± 10.1	9.4
	$I_f \oplus I_m \oplus S_m$		3.6 ± 1.9	3.1	4.4 ± 2.8	3.7	9.8 ± 5.9	8.1	13.4 ± 10.7	10.6
Cross-stitch	$I_f \| \| I_f \oplus I_m$	Segmentation	5.1 ± 2.3	4.4	9.5 ± 9.6	6.1	17.2 ± 14.0	12.6	5.0 ± 6.6	3.0
		Registration	3.3 ± 0.9	3.0	4.7 ± 3.0	3.7	10.1 ± 6.3	9.0	12.6 ± 10.0	9.4
	$I_f \| \| I_f \oplus I_m \oplus S_m$	Segmentation	**3.0 ± 1.0**	**3.0**	**4.3 ± 1.7**	3.9	**9.5 ± 6.2**	**7.2**	**3.3 ± 2.9**	**2.3**
		Registration	3.2 ± 0.9	3.0	4.5 ± 3.3	3.6	9.8 ± 6.3	8.6	12.2 ± 10.1	9.7
	$I_f \oplus S_m \| \| I_f \oplus I_m \oplus S_m$	Segmentation	5.8 ± 2.0	5.9	11.0 ± 13.4	5.8	10.2 ± 4.9	8.5	4.5 ± 4.3	3.0
		Registration	4.4 ± 1.6	4.1	4.5 ± 3.3	3.6	10.2 ± 5.7	9.3	12.9 ± 9.3	11.1
	$I_f \oplus I_m \oplus S_m \| \| I_f \oplus I_m \oplus S_m$	Segmentation	3.1 ± 1.0	3.0	5.4 ± 5.4	4.4	9.7 ± 5.6	8.9	4.2 ± 5.6	2.6
		Registration	3.5 ± 1.2	3.2	4.4 ± 3.1	**3.4**	10.2 ± 6.3	9.1	12.5 ± 10.6	8.7

Table 3: DSC values for the different networks and loss weighting methods for the HMC dataset. Higher values are better.

			Prostate		Seminal vesicles		Rectum		Bladder	
Network	Weight	Output path	$\mu \pm \sigma$	median	$\mu \pm \sigma$	median	$\mu \pm \sigma$	median	$\mu \pm \sigma$	median
JRS-reg	Equal	Registration	0.84 ± 0.16	0.89	0.67 ± 0.25	0.79	0.76 ± 0.14	0.79	0.79 ± 0.17	0.88
	Homoscedastic	Registration	0.84 ± 0.16	0.89	0.68 ± 0.25	0.77	0.76 ± 0.15	0.80	0.80 ± 0.18	0.89
	DWA	Registration	0.83 ± 0.16	0.88	0.66 ± 0.25	0.78	0.74 ± 0.15	0.79	0.76 ± 0.18	0.84
Dense	Equal	Segmentation	0.83 ± 0.15	0.88	0.55 ± 0.29	0.65	0.78 ± 0.16	0.81	0.88 ± 0.11	0.93
		Registration	0.83 ± 0.16	0.88	0.66 ± 0.25	0.75	0.76 ± 0.15	0.80	0.79 ± 0.16	0.87
	Homoscedastic	Segmentation	0.84 ± 0.16	0.89	0.63 ± 0.27	0.75	0.79 ± 0.16	0.82	0.87 ± 0.13	0.93
		Registration	0.84 ± 0.16	0.88	0.68 ± 0.25	0.78	0.77 ± 0.14	0.80	0.78 ± 0.17	0.86
	DWA	Segmentation	0.84 ± 0.15	0.89	0.58 ± 0.28	0.67	**0.79 ± 0.15**	**0.83**	0.88 ± 0.12	0.93
		Registration	0.84 ± 0.16	0.89	0.67 ± 0.24	0.76	0.76 ± 0.15	0.79	0.79 ± 0.16	0.87
SEDD	Equal	Segmentation	0.79 ± 0.16	0.85	0.46 ± 0.28	0.53	0.77 ± 0.14	0.80	0.85 ± 0.12	0.91
		Registration	0.82 ± 0.16	0.87	0.66 ± 0.26	0.78	0.75 ± 0.15	0.79	0.78 ± 0.16	0.86
	Homoscedastic	Segmentation	0.84 ± 0.15	0.89	0.50 ± 0.28	0.58	0.76 ± 0.18	0.82	0.88 ± 0.13	0.94
		Registration	0.84 ± 0.16	0.88	0.68 ± 0.24	0.78	0.76 ± 0.15	0.80	0.79 ± 0.17	0.88
	DWA	Segmentation	0.83 ± 0.14	0.88	0.62 ± 0.27	0.74	0.78 ± 0.16	0.83	0.87 ± 0.14	0.94
		Registration	0.84 ± 0.15	0.88	0.67 ± 0.24	0.78	0.75 ± 0.15	0.79	0.78 ± 0.18	0.86
Cross-stitch	Equal	Segmentation	**0.84 ± 0.14**	**0.89**	0.61 ± 0.27	0.73	0.78 ± 0.14	0.81	0.88 ± 0.10	0.93
		Registration	0.84 ± 0.15	0.89	**0.68 ± 0.24**	**0.80**	0.77 ± 0.15	0.80	0.80 ± 0.16	0.87
	Homoscedastic	Segmentation	0.84 ± 0.13	0.87	0.65 ± 0.24	0.76	0.74 ± 0.18	0.80	**0.92 ± 0.08**	**0.95**
		Registration	0.84 ± 0.15	0.89	0.68 ± 0.24	0.79	0.75 ± 0.15	0.79	0.80 ± 0.17	0.87
	DWA	Segmentation	0.82 ± 0.14	0.86	0.66 ± 0.24	0.76	0.75 ± 0.18	0.79	0.92 ± 0.08	0.95
		Registration	0.84 ± 0.15	0.89	0.68 ± 0.23	0.79	0.75 ± 0.15	0.78	0.77 ± 0.17	0.83

Table 4: %95 HD (mm) values for the different networks and loss weighting methods for the HMC dataset. Lower values are better.

			Prostate		Seminal vesicles		Rectum		Bladder	
Network	Weight	Output path	$\mu \pm \sigma$	median	$\mu \pm \sigma$	median	$\mu \pm \sigma$	median	$\mu \pm \sigma$	median
JRS-reg	Equal	Registration	5.2 ± 5.7	3.2	6.5 ± 7.1	4.0	12.6 ± 6.7	12.0	20.3 ± 14.0	18.6
	Homoscedastic	Registration	5.7 ± 5.9	3.7	6.2 ± 7.1	3.6	13.0 ± 7.3	**11.5**	18.5 ± 14.0	13.0
	DWA	Registration	5.7 ± 5.9	3.5	6.4 ± 6.8	3.7	13.2 ± 7.3	12.2	20.0 ± 13.2	17.6
Dense	Equal	Segmentation	5.7 ± 5.4	4.1	14.4 ± 17.2	6.8	16.8 ± 12.6	13.6	10.9 ± 10.9	5.5
		Registration	5.6 ± 5.6	4.0	6.6 ± 7.8	4.0	13.1 ± 6.7	13.0	19.6 ± 12.0	17.4
	Homoscedastic	Segmentation	5.8 ± 5.9	3.3	10.0 ± 11.6	5.1	17.1 ± 16.6	13.8	11.4 ± 11.3	5.9
		Registration	5.3 ± 5.7	3.0	6.4 ± 6.8	**3.2**	13.0 ± 6.5	12.6	19.2 ± 13.7	14.2
	DWA	Segmentation	5.4 ± 5.5	3.6	12.7 ± 17.0	5.9	16.2 ± 12.5	14.4	10.8 ± 10.7	6.2
		Registration	5.3 ± 5.6	3.5	**6.0 ± 6.6**	3.3	13.1 ± 7.2	13.0	19.4 ± 11.9	17.4
SEDD	Equal	Segmentation	8.5 ± 7.1	6.0	18.9 ± 19.5	8.6	16.7 ± 11.9	14.7	12.7 ± 11.0	8.5
		Registration	5.6 ± 5.8	3.6	6.7 ± 7.2	4.1	13.3 ± 7.0	12.0	19.0 ± 12.7	15.2
	Homoscedastic	Segmentation	5.7 ± 5.5	3.9	16.0 ± 16.3	10.6	18.8 ± 16.5	15.3	9.4 ± 9.9	4.1
		Registration	5.5 ± 5.6	3.3	6.3 ± 6.7	3.6	13.3 ± 7.3	13.0	18.8 ± 13.5	14.6
	DWA	Segmentation	6.2 ± 5.4	4.4	11.5 ± 14.0	5.0	16.8 ± 14.4	13.0	9.5 ± 10.8	4.4
		Registration	5.8 ± 5.7	4.0	6.4 ± 7.4	3.6	13.4 ± 7.5	12.5	21.9 ± 11.5	19.0
Cross-stitch	Equal	Segmentation	5.8 ± 5.4	4.0	12.2 ± 15.8	5.0	17.0 ± 14.7	14.0	10.8 ± 11.3	4.4
		Registration	**5.1 ± 5.5**	**3.2**	6.2 ± 8.6	3.3	**12.6 ± 6.7**	12.0	19.1 ± 12.5	16.2
	Homoscedastic	Segmentation	5.9 ± 5.4	4.1	7.8 ± 7.4	4.6	20.5 ± 18.9	14.7	7.8 ± 8.7	**3.1**
		Registration	6.2 ± 5.6	4.5	6.1 ± 7.2	3.2	13.5 ± 7.3	13.5	19.4 ± 12.3	16.3
	DWA	Segmentation	6.7 ± 5.8	4.2	7.6 ± 9.1	4.1	20.7 ± 18.6	14.9	**7.5 ± 8.8**	3.5
		Registration	6.0 ± 5.7	4.1	6.1 ± 6.8	3.4	13.5 ± 7.5	13.6	21.5 ± 11.6	20.1

Table 5: DSC values for the different networks on the validation set (HMC). Higher values are better.

		Prostate		Seminal vesicles		Rectum		Bladder	
Network	Output path	$\mu \pm \sigma$	median	$\mu \pm \sigma$	median	$\mu \pm \sigma$	median	$\mu \pm \sigma$	median
Seg	Segmentation	0.84 ± 0.03	0.84	0.60 ± 0.14	0.62	0.75 ± 0.10	0.77	0.90 ± 0.07	0.93
Reg	Registration	0.85 ± 0.06	0.86	0.62 ± 0.18	0.68	0.79 ± 0.08	0.81	0.82 ± 0.10	0.84
JRS-reg	Registration	0.86 ± 0.03	0.87	0.69 ± 0.13	0.73	0.83 ± 0.06	0.84	0.88 ± 0.08	0.92
Dense	Segmentation	0.88 ± 0.04	**0.89**	0.70 ± 0.12	0.73	0.85 ± 0.04	0.86	0.94 ± 0.02	0.94
	Registration	0.87 ± 0.04	0.88	0.68 ± 0.15	0.73	0.82 ± 0.06	0.83	0.87 ± 0.08	0.90
SEDD	Segmentation	0.87 ± 0.04	0.88	0.69 ± 0.12	0.72	0.83 ± 0.07	0.84	0.93 ± 0.02	0.94
	Registration	0.86 ± 0.04	0.87	0.69 ± 0.13	0.74	0.82 ± 0.06	0.83	0.88 ± 0.08	0.92
Cross-stitch	Segmentation	$\mathbf{0.88 \pm 0.04}$	0.88	$\mathbf{0.70 \pm 0.11}$	**0.74**	$\mathbf{0.86 \pm 0.05}$	**0.88**	$\mathbf{0.94 \pm 0.02}$	**0.95**
	Registration	0.87 ± 0.03	0.88	0.68 ± 0.15	0.73	0.84 ± 0.05	0.85	0.88 ± 0.08	0.91
Elastix [131]	Registration	0.84 ± 0.07	0.86	0.50 ± 0.25	0.53	0.74 ± 0.06	0.74	0.75 ± 0.10	0.76
Hybrid [23]	Registration	0.88 ± 0.04	0.89	0.70 ± 0.14	0.72	0.85 ± 0.06	0.87	0.91 ± 0.08	0.95
JRS-GAN [17]	Registration	0.86 ± 0.04	0.87	0.61 ± 0.20	0.67	0.82 ± 0.06	0.83	0.88 ± 0.08	0.92

Table 6: % 95 HD (mm) values for the different networks on the validation set (HMC). Lower values are better.

		Prostate		Seminal vesicles		Rectum		Bladder	
Network	Output path	$\mu \pm \sigma$	median	$\mu \pm \sigma$	median	$\mu \pm \sigma$	median	$\mu \pm \sigma$	median
Seg	Segmentation	4.4 ± 1.0	4.4	8.6 ± 8.6	7.3	16.5 ± 11.0	13.3	6.9 ± 6.6	4.0
Reg	Registration	5.5 ± 4.5	4.0	5.6 ± 4.1	4.3	11.0 ± 6.4	9.4	15.7 ± 9.6	12.1
JRS-reg	Registration	3.8 ± 1.3	3.2	4.1 ± 2.8	3.2	9.9 ± 6.2	8.4	11.7 ± 10.3	9.2
Dense	Segmentation	3.2 ± 1.0	3.0	5.8 ± 7.6	3.9	9.6 ± 5.8	8.0	3.8 ± 3.9	**2.8**
	Registration	3.4 ± 1.1	3.2	4.4 ± 3.0	3.2	10.5 ± 6.0	9.0	12.6 ± 9.2	10.2
SEDD	Segmentation	3.5 ± 1.1	3.3	5.2 ± 5.2	4.0	10.5 ± 5.5	9.7	$\mathbf{3.3 \pm 1.3}$	3.0
	Registration	3.6 ± 1.2	3.2	4.1 ± 2.6	3.1	10.4 ± 6.3	9.5	11.7 ± 9.9	8.7
Cross-stitch	Segmentation	3.0 ± 1.0	3.0	4.3 ± 1.7	3.9	9.5 ± 6.2	7.2	3.3 ± 2.9	2.3
	Registration	3.2 ± 0.9	3.0	4.5 ± 3.3	3.6	9.8 ± 6.3	8.6	12.2 ± 10.1	9.7
Elastix [131]	Registration	4.0 ± 1.7	3.7	6.0 ± 3.4	5.6	10.9 ± 5.2	9.8	15.3 ± 8.3	13.6
Hybrid [23]	Registration	$\mathbf{2.9 \pm 0.9}$	**2.8**	$\mathbf{3.8 \pm 2.2}$	**3.1**	$\mathbf{7.7 \pm 4.5}$	**6.1**	5.7 ± 4.6	3.3
JRS-GAN [17]	Registration	3.4 ± 1.2	3.0	5.3 ± 3.0	4.6	10.1 ± 6.1	8.4	11.0 ± 9.6	7.6

Table 7: DSC values for the different networks on the independent test set (EMC). Higher values are better.

		Prostate		Seminal vesicles		Rectum		Bladder	
Network	Output path	$\mu \pm \sigma$	median	$\mu \pm \sigma$	median	$\mu \pm \sigma$	median	$\mu \pm \sigma$	median
Seg	Segmentation	0.73 ± 0.11	0.77	0.37 ± 0.30	0.28	0.67 ± 0.10	0.68	$\mathbf{0.91 \pm 0.07}$	0.93
Reg	Registration	0.83 ± 0.16	0.88	0.64 ± 0.26	0.74	0.72 ± 0.16	0.77	0.75 ± 0.19	0.82
JRS-reg	Registration	0.84 ± 0.16	0.89	0.68 ± 0.25	0.77	0.76 ± 0.15	0.80	0.80 ± 0.18	0.89
Dense	Segmentation	0.84 ± 0.16	0.89	0.63 ± 0.27	0.75	0.79 ± 0.16	0.82	0.87 ± 0.13	0.93
	Registration	0.84 ± 0.16	0.88	0.68 ± 0.25	0.78	0.77 ± 0.14	0.80	0.78 ± 0.17	0.86
SEDD	Segmentation	0.84 ± 0.15	0.89	0.50 ± 0.28	0.58	0.76 ± 0.18	0.82	0.88 ± 0.13	**0.94**
	Registration	0.84 ± 0.16	0.88	0.68 ± 0.24	0.78	0.76 ± 0.15	0.80	0.79 ± 0.17	0.88
Cross-stitch	Segmentation	0.84 ± 0.14	0.89	0.61 ± 0.27	0.73	0.78 ± 0.14	0.81	0.88 ± 0.10	0.93
	Registration	0.84 ± 0.15	0.89	0.68 ± 0.24	0.80	0.77 ± 0.15	0.80	0.80 ± 0.16	0.87
Elastix [131]	Registration	$\mathbf{0.89 \pm 0.05}$	**0.91**	0.72 ± 0.24	**0.82**	0.75 ± 0.12	0.76	0.79 ± 0.18	0.87
Hybrid [23]	Registration	0.88 ± 0.04	0.89	$\mathbf{0.77 \pm 0.15}$	0.81	$\mathbf{0.80 \pm 0.10}$	**0.82**	0.85 ± 0.13	0.90

Table 8: %95 HD (mm) values for the different networks on the independent test set (EMC). Lower values are better.

		Prostate		Seminal vesicles		Rectum		Bladder	
Network	Output path	$\mu \pm \sigma$	median	$\mu \pm \sigma$	median	$\mu \pm \sigma$	median	$\mu \pm \sigma$	median
Seg	Segmentation	10.7 ± 5.4	9.3	21.4 ± 17.9	15.4	30.5 ± 12.9	29.0	11.2 ± 8.5	10.0
Reg	Registration	6.7 ± 5.9	4.2	7.5 ± 8.6	4.3	13.1 ± 6.9	12.0	22.7 ± 14.0	20.2
JRS-reg	Registration	5.7 ± 5.9	3.7	6.2 ± 7.1	3.6	13.0 ± 7.3	11.5	18.5 ± 14.0	13.0
Dense	Segmentation	5.8 ± 5.9	3.3	10.0 ± 11.6	5.1	17.1 ± 16.6	13.8	11.4 ± 11.3	5.9
	Registration	5.3 ± 5.7	3.0	6.4 ± 6.8	3.2	13.0 ± 6.5	12.6	19.2 ± 13.7	14.2
SEDD	Segmentation	5.7 ± 5.5	3.9	16.0 ± 16.3	10.6	18.8 ± 16.5	15.3	$\mathbf{9.4 \pm 9.9}$	**4.1**
	Registration	5.5 ± 5.6	3.3	6.3 ± 6.7	3.6	13.3 ± 7.3	13.0	18.8 ± 13.5	14.6
Cross-stitch	Segmentation	5.8 ± 5.4	4.0	12.2 ± 15.8	5.0	17.0 ± 14.7	14.0	10.8 ± 11.3	4.4
	Registration	5.1 ± 5.5	3.2	6.2 ± 8.6	3.3	12.6 ± 6.7	12.0	19.1 ± 12.5	16.2
Elastix [131]	Registration	$\mathbf{3.6 \pm 2.0}$	**2.9**	$\mathbf{4.6 \pm 4.4}$	3.2	11.3 ± 6.0	11.3	16.1 ± 14.8	10.4
Hybrid [23]	Registration	3.9 ± 1.9	3.4	4.8 ± 4.7	**3.1**	$\mathbf{10.3 \pm 6.8}$	**8.6**	11.1 ± 10.6	6.6

Publications

Journal articles

Elmahdy, Mohamed S, T. Jagt, R. T. Zinkstok, Y. Qiao, R. Shahzad, H. Sokooti, S. Yousefi, L. Incrocci, C. Marijnen, M. Hoogeman, et al. "Robust contour propagation using deep learning and image registration for online adaptive proton therapy of prostate cancer". In: *Medical Physics* 46.8 (2019), pages 3329–3343

N. Pezzotti, S. Yousefi, **Elmahdy, Mohamed S**, J. H. F. Van Gemert, C. Schuelke, M. Doneva, T. Nielsen, S. Kastryulin, B. P. Lelieveldt, M. J. Van Osch, et al. "An Adaptive Intelligence Algorithm for Undersampled Knee MRI Reconstruction". In: *IEEE Access* 8 (2020), pages 204825–204838

H. Sokooti, S. Yousefi, **Elmahdy, Mohamed S**, B. P. Lelieveldt, and M. Staring. "Hierarchical Prediction of Registration Misalignment Using a Convolutional LSTM: Application to Chest CT Scans". In: *IEEE Access* 9 (2021), pages 62008–62020

Elmahdy, Mohamed S, L. Beljaards, S. Yousefi, H. Sokooti, F. Verbeek, U. Van Der Heide, and M. Staring. "Joint Registration and Segmentation via Multi-Task Learning for Adaptive Radiotherapy of Prostate Cancer". In: *IEEE Access* (2021)

S. Yousefi, H. Sokooti, **Elmahdy, Mohamed S**, I. M. Lips, M. T. M. Shalmani, R. T. Zinkstok, F. J. Dankers, and M. Staring. "Esophageal Tumor Segmentation in CT Images Using a Dilated Dense Attention Unet (DDAUnet)". In: *IEEE Access* 9 (2021), pages 99235–99248

Conference proceedings

Elmahdy, Mohamed S, T. Jagt, S. Yousefi, H. Sokooti, R. Zinkstok, M. Hoogeman, and M. Staring. "Evaluation of multi-metric registration for online adaptive proton therapy of prostate cancer". In: *International Workshop on Biomedical Image Registration*. Springer. 2018, pages 94–104

S. Yousefi, H. Sokooti, **Elmahdy, Mohamed S**, F. P. Peters, M. T. M. Shalmani, R. T. Zinkstok, and M. Staring. "Esophageal gross tumor volume segmentation using a 3D convolutional neural network". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2018, pages 343–351

Elmahdy, Mohamed S, J. M. Wolterink, H. Sokooti, I. Išgum, and M. Staring. "Adversarial optimization for joint registration and segmentation in prostate CT radiotherapy". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2019, pages 366–374

S. Yousefi, L. Hirschler, M. van der Plas, **Elmahdy, Mohamed S**, H. Sokooti, M. Van Osch, and M. Staring. "Fast Dynamic Perfusion and Angiography Reconstruction Using an End-to-End 3D Convolutional Neural Network". In: *International Workshop on Machine Learning for Medical Image Reconstruction*. Springer. 2019, pages 25–35

Elmahdy, Mohamed S, T. Ahuja, U. A. van der Heide, and M. Staring. "Patient-Specific Finetuning of Deep Learning Models for Adaptive Radiotherapy in Prostate CT". in: *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2020, pages 577–580

L. Beljaards, **Elmahdy, Mohamed S**, F. Verbeek, and M. Staring. "A Cross-Stitch Architecture for Joint Registration and Segmentation in Adaptive Radiotherapy". In: *arXiv preprint arXiv:2004.08122* (2020)

P. M. Johnson, K. Hammernik, J. Schlemper, J. Duan, D. Rueckert, N. Pezzotti, E. D. Weerdt, S. Yousefi, **Elmahdy, Mohamed S**, et al. "Evaluation of the Robustness of Learned MR Image Reconstruction to Systematic Deviations Between Training and Test Data for the Models from the fastMRI Challenge". In: *International Workshop on Machine Learning for Medical Image Reconstruction*. Springer. 2021, pages 25–34

Y. Li, **Elmahdy, Mohamed S**, M. S. Lew, and M. Staring. "Transformation-Consistent Semi-Supervised Learning for Prostate CT Radiotherapy". In: *SPIE Medical Imaging: Image Processing*. Proceedings of SPIE. San Diego, CA, USA, 2022

Code Repository

Mohamed S. Elmahdy, JRS-GAN, *GitHub*, github.com/moelmahdy/JRS-GAN
Mohamed S. Elmahdy, JRS-MTL, *GitHub*, github.com/moelmahdy/JRS-MTL

Acknowledgements

ٱلْحَمْدُ لِلَّهِ رَبِّ ٱلْعَٰلَمِينَ

The success of my PhD journey would not have been possible without the support and inspiration of the wonderful colleagues who made me enthusiast for science. Therefore, I am profoundly grateful to all those who contributed, supported, and helped me during this journey.

First and foremost, I would like to thank my PhD supervisors Dr.ir. Marius Staring, Prof.dr.ir. Uulke Van der Heide, and Prof.dr.ir. Boudewijn Lelieveldt for giving me the opportunity to work in a diversified and prestigious lab such as LKEB. Your immense knowledge and plentiful experience have encouraged me by all means through my academic research and daily life. To Marius, thank you for being the best supervisor any one could have. You have always been friendly, considerable, and more importantly always available for a walk in discussions. I learned a lot from your expertise in image registration and deep learning. I will always be grateful for your willingness to adopt my ideas and for helping me answering my posed research questions as well as for your continuous encouragement. To Uulke, it has been a privilege to work with you, I have always admired your impressive radiotherapy knowledge and how you can be concise. I usually come to you with questions and leave with lots of ideas that enrich my research and horizon. I'm grateful for Boudewijn for easing every obstacle I had during my PhD and for giving me the opportunity to attend multiple conferences across the world that definitely broaden my horizon.

I would like to offer my sincere gratitude to my colleagues at LKEB for their friendship and support and for making LKEB a work friendly environment. Thanks to Hessam, Kilany, Xiaowu, Qing, Qian, Mody, Irene, Sahar, Zhewei, Antonios, Oleh, Denis, Laurens, Viktor, Li-Hsin, Jingnan, Patrick, Alexander, Jeroen, Tahereh, and Kirsten, I learned a lot from you folks and enjoyed our discussions over lunch and through the Monday morning talks and Deep learning meetups. I am deeply grateful to Rob for offering me a postdoc position at LKEB and also for the nice boat trip in Leiden canals and for the nice BBQ at his lovely house. Thanks Berend, Els, and Jouke for being so friendly and supportive. Thanks Michele for all the IT support and for never letting us down. Special Thanks to Hessam, Sahar, Tanuj, and Laurens for the

great discussions we had during our weekly meetings and for the great papers that we collaborated on.

I would like to thank Thyrza Jagt for all her amazing work on the dose calculation and evaluation, I also would like to thank Mischa Hoogeman for our fruitful discussions at Holland PTC. During the thesis, I had the opportunity to work with brilliant colleagues from Philips research and radiology dept. at LUMC on the fastMRI competition, special thanks to Nicola Pezzotti, Elwin de Weerdt, Thijs van Osch, and Sahar Yousefi. I also would like to extend my sincere thanks to Jelmer Wolterink who helped me a lot while training the GAN networks and Ivana Išgum for her insights and feedback.

Thank Patrick and Tamim for being my paranymphs and for doing your best in helping me organize the defense ceremony.

I'm deeply grateful for my friends Kilany, Tamim, Walid, Mahfouz, Eltager, Dr. Ryad, and Dr. Elsify who made me always feel home every time we get together here in Netherlands. I would like to express my sincere gratitude to my life time friends and professors from Egypt, Hisham, Tarek, Salah, Yassin, Dr. Ayman, Dr. Inas, Dr. Tamer, Dr. Noha, Dr. Roshdi, Dr. Morsy, Dr. Elbialy, Dr. Hesham, and Dr. Sahar.

My sincere gratitude to my family for their continuous and undivided love, help, and support. I am forever indebted to my parents, Said and Ahlam for giving me the opportunities and experiences that have made me who I am. They selflessly encouraged me to explore new directions in life and seek my own destiny. This journey would not have been possible if not for them, and I dedicate this milestone to them. I am grateful to my sister Dina and my brother Mahmoud for always being there for me as my best friends.

إلي والدي الحبيبين، أنا مدينٌ لكما إلي الأبد لإعطائي الفرص و الخبرات التي جعلتني ما أنا عليه الآن.إلي أمي الغالية أحلام، أطال الله لي عمرك و كتب لكي دوام الصحة و العافية. إلي أبي العزيز سعيد، حفظك الله لي و أدامك تاج علي رأسي دائما و أبداً.

Finally, to my beloved wife Sarah, this journey would not have been possible without you by my side with your tremendous understanding and encouragement in the past few years. I would forever be grateful for your countless sacrifices to help me get to this point. Your unlimited and unconditional love was the light in my darkest and toughest days. To my son, Adam, watching you growing and bonding with you every day was the best thing that ever happened while working from home during COVID time. You filled my life with so much love and precious memories. I hope one day you will be proud of me.

Curriculum Vitae

Mohamed S. Elmahdy was born in Cairo, Egypt. He received his BSc in biomedical engineering from Cairo University in 2013. In his BSc project, he designed an early prototype of a low-cost high-performance hand-held CAD system for early diagnosis of abnormalities in mammography. He obtained his MSc degree in biomedical engineering from the same university in 2017, where he developed an end-to-end deep learning framework for subvocal speech recognition via close-talk microphone and surface electromyogram.

From August 2017, he started his PhD study in the Division of Image Processing (LKEB) under the Department of Radiology at Leiden University Medical Center in the Netherlands. His PhD project mainly focuses on developing deep learning algorithms for online adaptive radiotherapy.

Starting from March 2021, he works as a post-doctoral researcher in LKEB, on a project of cardiac MR analysis using deep learning.