Novel concepts for automated segmentation to facilitate MRI-guided radiotherapy in head and neck cancer

A thesis submitted for the degree of
Doctor of Philosophy

Jennifer Kieselmann

Joint Department of Physics

The Institute of Cancer Research and
The Royal Marsden NHS Foundation Trust

University of London

I, Jennifer Kieselmann, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Abstract

The dramatic increase of magnetic resonance imaging (MRI) in daily treatment planning and response assessment of radiotherapy (RT) requires the development of reliable auto-segmentation algorithms for organs-at-risk (OARs) and radiation targets. The current practice of manual segmentation is subjective and time-consuming, particularly for head and neck cancer (HNC) patients. New methodologies based on machine learning offer ample opportunities to solve this problem.

This thesis aimed to develop accurate and rapid auto-segmentation algorithms on MR images of HNC patients, employing established atlas-based algorithms and comparing the results with deep learning-based methods. The work is divided into design and implementation of auto-segmentation methods followed by extensive validation studies. For the latter, I developed a fully automated RT workflow enabling validation on purely geometric features of the automatically generated contours whose impact on key dosimetric features of a treatment plan was further analysed.

A common challenge for medical image segmentation is the limited availability of data due to the associated cost of obtaining expert contours. Moreover, frequent updates of imaging protocols or scanners may prevent algorithms, developed on existing databases, from working well on newly-acquired images. I designed domain adaptation methods which leverage large databases from related application domains to tackle this problem.

While both auto-segmentation strategies achieved clinically acceptable accuracy, atlas-based methods were slow and are, unlike deep learning-based models, difficult to share between hospitals due to data-confidentiality issues. Deep learning-based methods were able to alleviate the computational burden, generating contours within seconds. Moreover, when healthy tissue was infiltrated with irregular structures, deep learning was more accurate.

In conclusion, I demonstrated that auto-segmentation was feasible and can change clinical practice. Moreover, domain adaptation strategies hold promise in mitigating problems with small datasets in medical imaging and in eliminating the need to acquire new annotated datasets for each change in imaging protocols.

Acknowledgements

Research never happens behind closed doors! I want to take this opportunity to thank the numerous people that have contributed to this work in one way or the other. I am grateful to (in not necessarily a particular order)

- my supervisors **Uwe Oelfke** and **Simeon Nill** for enabling this project, their support throughout the last four years, reading through the numerous pages of my thesis. In particular, thank you for attributing a lot of trust to my skills. This allowed me to explore many, mostly exciting, roads on the way to this day and become independent in conducting research. Also, thank you for allowing me to go to basically any conference I ever suggested.
- my supervisor **Corijn Kamerling**, who gave me invaluable help in the first year of my PhD and who was always available to listen to ideas and to give advice, even after he had (unfortunately) left the institute.
- **Jorge Cardoso** and **Ninon Burgos** for many interesting discussions, a different perspective on my work outside the field of radiotherapy, as well as their generous hospitality at UCL.
- **Oliver Gurney-Champion** for setting up very useful, collaborative meetings, for numerous interesting discussions, the possibility to bounce back ideas, a lot of encouraging support, proofreading every single word of my thesis, many fun phBar shifts together and much more!
- **Peter Ziegenhein** for encouraging me to look into deep learning!
- **Merle Reinhart** and **Jenny Bertholet** for proofreading parts of my thesis - it was really helpful!
- my office mates **Martin Menten**, **Sarah Mason**, **Dave Edmunds** and **Jonathan Mohajer** for lots of fun, interesting discussions, support and making the thought of going to work much more enjoyable every day.

- the whole **RT-Phys-Modelling team** for lots of fun, as well as support, at uncountable lunches, coffee breaks, group socials, conferences and much more! I couldn't have wished for a better group!

- the **head and neck team** at the ICR/RMH for including me in their meetings and helping me out with the clinical perspective of my work.

- **Brian Hin**, **Arabella Hunt**, **Gemma McCormick** and **Imran Petkar** for the tedious manual delineation work. **Brian**, thank you for teaching me how to contour in the head and neck!

- my examiners **Coen Rasch** and **David Hawkes** for thoroughly reading my thesis, asking very challenging questions, interesting discussions and a surprisingly enjoyable experience at my PhD defense!

- all **volunteers** who suffered through an hour of having their head in an uncomfortable position in the MR scanner! Can you find yourself in my thesis?

- the **Oracle Cancer Trust** for their generous financial support, without which this project would not have been possible. Thank you for the many opportunities to talk about my work to your supporters! I will particularly remember the Gala evening at the Asian Achiever Awards!

- **NVIDIA** for their support of my work through the donation of a graphics card.

- the **MD Anderson Cancer Center**, in particular, **Dave Fuller**, **Yao Ding** and **Mona Kamal Jomaa** for helping me out with sharing their patient data!

- the Physics secretaries **Debbie Carrick**, **Annette Willmott**, **Caroline Saunders**, **Cheryl Taylor**, **Lesley Brotherston**, **Lesley Richards** and **Louise Sear** for their valuable and always reliable help and constant support throughout the years with many conference bookings and other organisational stuff!

- the great company at conference trips, including the road trip to AAPM in Washington D.C. (**Merle**, **Martin** and **Henry (El Tsango)**, as well as **Corijn**, **Dave** and **Martin Fast** during the conference), ESTRO in Vienna (**Dave**, **Oliver** and **Peter**), MR in RT in Sydney (**Filipa Costa** and **Joshua Freedman**), Machine learning workshop in California (**Oliver** and **Joshua**), train trip to MR in RT in Utrecht (**Filipa** and **Jenny**), deep learning summer school in Sicily (**Oliver**) and ICCR/MR in RT Canada (**Sarah Brueningk**, **Jenny** and **Oliver**).

- **Jenny** for the monthly fun with creating our newsletter!

- the **ICR girls and Martin** for lots of nice dinners, yearly Secret Santas, pub crawls, cake lunches, weddings and much more!
- the **board games group** for lots of delicious dinners and fun nights! Thanks for hosting me all the time in Sutton!
- **JFK** for being supportive fellow PhD students and for nice dinners, cakes, walks and so forth outside work - you are both almost there, too!
- **Amona** for being an excellent model to demonstrate some deep learning-basics in this thesis.
- my **library in Germany** for predicting early on in my childhood that I would be a Dr. (despite it probably being an honest mistake... well, who knew you were right in the end!)
- the **inventors of chocolate and Chai latte** for helping me through some decreased concentration capability in the afternoons.
- **numerous friends**, in London and all over the world for their constant support throughout my PhD with encouraging words, phone calls, visits and so forth (you know who you are)!
- **my family** for their encouragement, help and never-ending support. Also thanks for tolerating the fact that I moved "far" away and, especially towards the end of my PhD, did not find much time to visit home.

Contents

List of Figures

List of Tables

List of Abbreviations

2D two-dimensional.
3D three-dimensional.
ART adaptive radiation therapy.
CBCT cone-beam computed tomography.
CI Jaccard conformity index.
CNN convolutional neural network.
CPP control position point.
CT computed tomography.
CTV clinical target volume.
DSC Dice similarity coefficient.
GAN generative adversarial network.
GPU graphical processing unit.
GTV gross target volume.
Gy Gray.
HD Hausdorff distance.
HD95 95th percentile of the Hausdorff distance.
HNC head and neck cancer.
IMRT intensity-modulated radiation therapy.
linac linear accelerator.
MI mutual information.
MLC multi-leaf collimator.
MR magnetic resonance.
MRF Markov Random Field.
MRI magnetic resonance imaging.
MSD mean surface distance.
NCC (normalised) cross-correlation.
OAR organ at risk.
PET positron-emission tomography.

PRV planning risk volume.

PTV planning target volume.

ROI region of interest.

RT radiation therapy.

SD standard deviation.

SSM statistical shape and appearance model.

T1w T1-weighted.

T2w T2-weighted.

TPS treatment planning system.

VMAT volumetric modulated arc therapy.

Chapter 1

Introduction

1.1 Motivation

All models are wrong, but some are useful.
George E. P. Box

In radiation therapy (RT), tumours are irradiated with ionising radiation. While ionising radiation can kill tumour cells, it cannot be avoided that the radiation beams also traverse healthy tissues and deposit dose in these regions. Modern treatment planning systems can design highly conformal dose distributions, delivering a high radiation dose to the tumour with a sharp dose fall-off to minimize the irradiation of organs at risk (OARs). Full utilisation of this sharp dose fall-off requires accurate localisation of the target and the OARs.

An RT treatment is generally planned using the information on the patient's anatomy from an x-ray computed tomography (CT) image, which is typically acquired days or weeks before the actual treatment. The treatment itself can last several weeks, exploiting the finding that healthy tissues can recover better from radiation damage than tumorous tissues. As a first step in the clinical workflow of an RT treatment, a clinician conventionally outlines all regions of interest (ROIs) on the planning CT. This process is also called image segmentation. Image segmentation is especially tedious for the treatment of head and neck cancer (HNC) patients due to the complex anatomy, including many OARs and irradiation targets associated with HNC. Many of these ROIs are challenging to outline due to poor soft-tissue contrast provided by the CT images.

Image guidance in RT has seen a dramatic increase in magnetic resonance imaging (MRI), owing to its superior soft-tissue contrast [130] and the absence of ionising radiation compared to the conventionally used CT [31, 80, 95]. The information gained from high soft-tissue contrast magnetic resonance (MR) images can be used to improve the

contouring of ROIs on the CT for treatment planning [20, 38, 121, 122]. Also, integrated MRI and treatment delivery systems have become available in the past years [40, 91, 100, 120]. These systems allow MR scanning of the patient in treatment position directly before or during the treatment delivery. Classically, cone-beam computed tomography (CBCT) images are acquired. However, due to the poor soft-tissue contrast, only bony anatomy matches can be performed, leading to rigid shifts of the patient which do not fully account for changes in the patient's anatomy throughout the treatment. With the introduction of daily MRI, a daily adaptation to the current anatomy of the patient has become possible.

Moreover, in MR-only treatment workflows, MR images replace the conventionally used pre-treatment CT [76, 107]. In such workflows, treatment planning and dose calculation are solely based on the MR images. One of the challenges in an MRI-only workflow is that, contrary to a CT, the required electron density information for the dose calculated in treatment planning cannot be derived directly from the image intensities. Therefore, the creation of synthetic CTs is necessary to provide surrogates for electron densities [36].

To realise adaptive RT, a repeated delineation of all ROIs is necessary.The current practice of manual segmentation is a time-consuming and error-prone process [151]. Automating the outlining of ROIs would allow to alleviate the enormous workload of manual segmentation and reduce the inter- and intra-observer variabilities. Moreover, an adaptive RT workflow is only feasible with automated contouring tools as manual contouring can take up to hours, which is an unfeasible burden to the daily clinical workflow.

Automation of medical image segmentation is a challenging problem since the anatomy of each patient is different, varying in size, position and shape. Moreover, the image quality varies between successive acquisitions and different patients. To date, the most commonly used auto-segmentation methods are atlas-based (Fritscher et al. [43] and references therein), although recent developments in machine learning offer ample opportunities for further improving automated contouring [12]. Numerous studies have investigated CT-based automated delineation of critical structures in the head and neck region [26, 39, 43, 56, 63, 79, 115, 119, 134], yet only very few studies have been conducted on MR images [149, 154, 162].

A problem commonly encountered in medical image segmentation is the lack of delineated imaging data. To perform well, most developments of auto-segmentation methods need a large amount of example imaging data. Many machine learning-based applications originate from the field of natural images. Natural images denote photographs of people, landscapes, animals and other objects. There are typically

millions of natural images available to train algorithms in this field. However, obtaining annotated medical images is usually associated with a high cost because the collection of large medical imaging datasets is generally a time-consuming process. It involves the search for suitable data in large hospital systems with only moderately structured information, further processing (annotation) of the data by expert physicians and requires patient consent. Although this problem can be partly mitigated by increasing collection of large databases over time, for instance by collaborations between hospitals, the amount of data is still quite small when compared to natural images (typically hundreds of images compared to millions). Furthermore, MR images are likely to change in appearance due to changes in image acquisition parameters or updates of MR scanners. Hence, auto-segmentation approaches developed on an existing image database may not work well on newly acquired images.

To allow for MRI-guided adaptive RT treatments, it is, therefore, crucial to develop auto-segmentations methods which only require small amounts of data. This thesis aimed to design and develop auto-segmentation algorithms which can rapidly and accurately perform this task on MR images of HNC patients.

1.2 Thesis aim and outline

Chapter 2 provides a brief overview of the theoretical background of this thesis. Starting with the clinical motivation for the treatment of HNC, I introduce general concepts in RT with a focus on image-guided adaptive RT and, in particular, MRI-guided RT. Finally, I provide a general overview of various existing strategies for medical image segmentation, which was at the heart of this thesis.

Chapter 3 provides an overview of the imaging data used for this thesis and the employed preprocessing steps.

The central part of this thesis was the design and implementation of auto-segmentation algorithms, followed by extensive validation studies. Unfortunately, studies on auto-segmentation methods usually lack suitable evaluation metrics in the context of RT. For this reason, I developed a fully automated RT workflow, enabling validation on purely geometric features of the automatically generated contours whose impact on key dosimetric features of a treatment plan was further analysed. This workflow is described in chapter 4.

For the design of the segmentation algorithms, I employed atlas-based (chapter 5) and deep learning-based methods (chapter 6). Atlas-based methods are well-established and are integrated into some commercial treatment planning systems. However, they are generally slow and, due to data-confidentiality issues, hard to share between sites as

they require an image database. Deep learning-based approaches have recently shown promising results in computer vision problems, such as object detection or classification. Furthermore, their ability to make fast predictions and the easiness to share trained models with other users are desirable qualities for RT.

A common challenge for medical image segmentation is the limited availability of data due to the associated cost of obtaining expert contours. Moreover, frequent imaging protocol or scanner updates may prevent algorithms, developed on existing databases, from working well on newly-acquired images. To overcome these limitations, I designed domain adaptation methods which leverage large databases from related application domains. Chapter 6.5 describes a machine learning technique named transfer learning, where the gained knowledge from outlining ROIs on CT images initialised an algorithm to segment these ROIs on MR images. Chapter 6.6 describes an algorithm named cross-modality learning, developed for this thesis. It can leverage information learned from existing databases of one imaging modality (CT images or original MR sequence) to prevent the need for acquiring and annotating new datasets under the new protocol (MR images or updated MR sequence).

Chapter 7 summarises the quantitative results of the previous chapters and discusses the strengths and limitations of each method. It provides a qualitative analysis on the generalisability of the developed algorithms, employing a fully independent test dataset. Chapter 8 then summarises the main findings of this thesis, discusses the strengths and limitations of the presented work and concludes with suggestions for potential future research.

All methods in this thesis were developed explicitly for MRI-guided RT treatment of HNC. Nonetheless, many of the introduced concepts apply to other treatment sites or for MR-only treatment workflows.

Chapter 2

Theoretical background

This thesis introduces novel concepts for automated image segmentation to facilitate MRI-guided radiotherapy in head and neck cancer. This chapter provides the relevant theoretical background. First, the clinical problem, head and neck cancer, is introduced. I then give an overview of radiotherapy with state-of-the-art hardware and software developments. I conclude with a review of existing auto-segmentation methods in the literature, where I briefly introduce their methodologies and discuss their limitations.

2.1 Head and neck cancer

HNC is a heterogeneous group of cancers which arises in parts of the head and the neck, such as the pharynx, larynx and oral cavity [138]. Figure 2.1 indicates the location of these regions within the head and the neck. HNC is the 7th most common cancer type in the United Kingdom with an incidence rate of approximately 12,000 cases per year and a mortality rate of about 4,000 per year [11]. The most common causes are tobacco and alcohol consumption, as well as infection with the human papillomavirus [138]. The disease often spreads to the lymph nodes of the neck.

The three main types of treatment of HNC are RT, surgery and chemotherapy. The optimal combination of the three treatment modalities depends on the specific type and stage of HNC. Concurrent chemo-RT has become the standard of care for patients with locally advanced HNC (spread to the lymph nodes). While surgery is an invasive technique and, depending on the proximity of functional organs to the tumour, may not be able to to conserve their functionality completely, RT can rely on the ability of healthy tissues to recover from radiation effects, in particular by using appropriate fractionation schemes.

Many functional organs, including the salivary glands, the spinal cord, the brain stem and optical structures, are located in the head and the neck and often close to

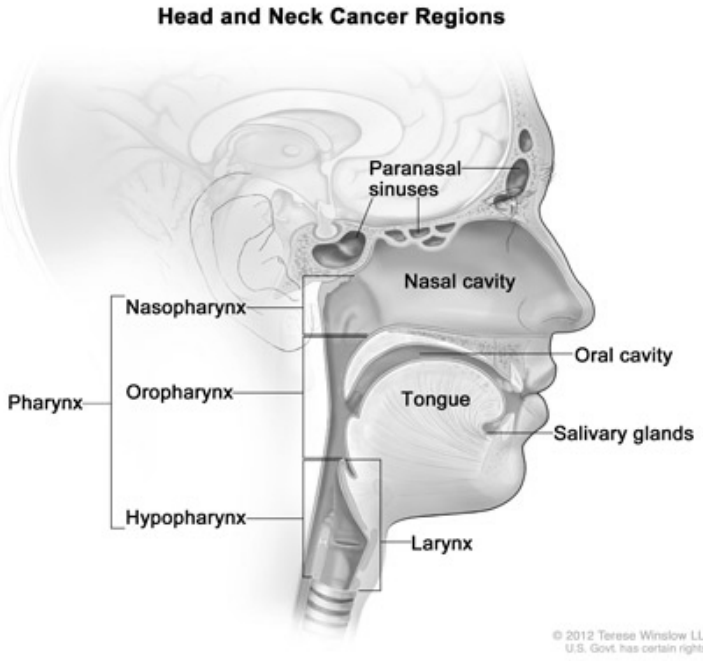


Figure 2.1: This figure highlights the heterogeneity of head and neck cancer, with labels of the various sites where cancer can occur. Figure courtesy: *https://cancer.gov*

the primary tumour. Due to this proximity, RT is known to lead to several side effects, such as a chronic dry mouth (xerostomia), oral mucositis or swallowing dysfunction (dysphagia) [71, 82, 103, 106]. These side effects can have a significant impact on the quality of life for these patients and render the treatment of HNC challenging. Advanced techniques in RT are developed to overcome these challenges.

2.2 Radiotherapy

RT uses ionising radiation to kill cancer cells by damaging their DNA. The energy given to a certain mass element of tissue, the dose, is measured in Gray (Gy), where $1\,\mathrm{Gy} = 1\,\mathrm{J/kg}$. There are two ways for the delivery of RT: injection of radioactive sources into or close to the tumour target (internal RT or brachytherapy) or application of external sources (external-beam RT, EBRT). In EBRT, the radiation can be delivered in the form of electron, particle (protons or heavy ions), or high-energy photon (i. e. x-ray) beams. Like most cancers, HNC is most commonly treated with high-energy photon beams, which are produced with a linear accelerator (linac). I hence focus on this type of RT in the remainder of this thesis.

Figure 2.2 illustrates a linac, where I annotated the most essential components:

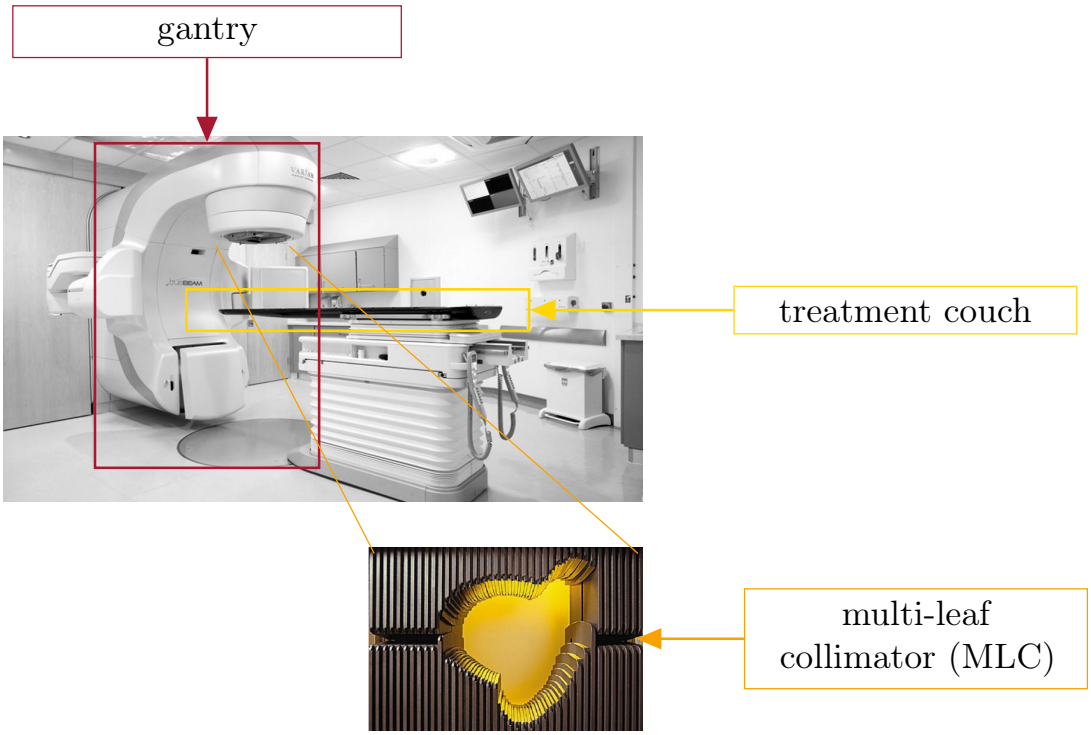


Figure 2.2: An image of a linear accelerator, with the most important components highlighted: The gantry (red) rotates around the treatment couch (yellow). The multi-leaf collimator is mounted at the head of the gantry and is used to shape the beams (orange).

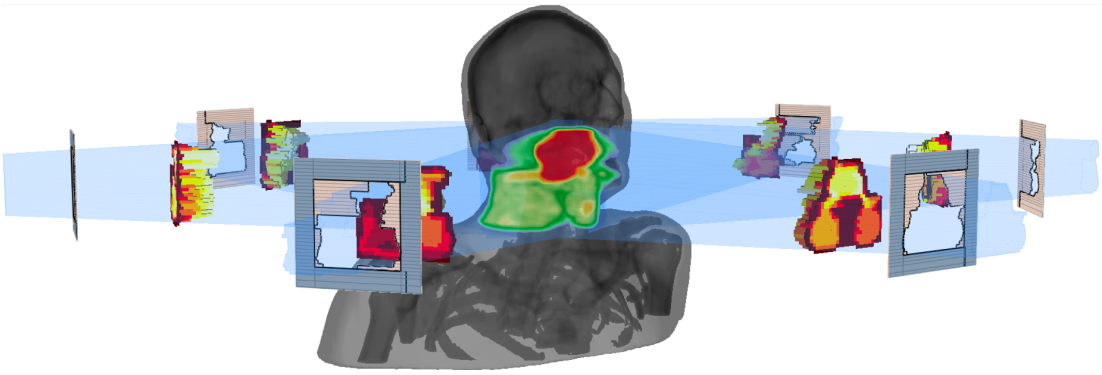


Figure 2.3: Step-and-shoot IMRT with 7 beams of an HNC patient: each of the beams uses a different MLC shape (indicated as grey squares) and contributes with a different photon fluence distribution (coloured distribution map) to the total dose distribution. The dose distribution is indicated as an overlay over the patient's anatomy, where the red region indicates a high dose, given to the target volume. Figure courtesy: Frederiksson [42].

during treatment, the patient lies on the treatment couch. The linac is mounted onto a gantry which can rotate around the patient. At the head of the gantry, a collimator can shape the treatment beams customised to the patient's anatomy.

While RT aims to irradiate the tumour, it cannot be avoided that the radiation beams also hit healthy tissue. It is particularly important to minimise the dose to essential radiation-sensitive tissues, so-called OARs, which might suffer damage from irradiation. Significant technical improvements towards achieving this aim have been accomplished in the delivery of RT in the last decades [8]. The balance between the probability of tumour control and the risk of normal tissue complications is a measure of the therapeutic ratio of the treatment. This therapeutic ratio denotes the relationship between the probability of tumour control and the likelihood of normal tissue damage and can be maximised in two main ways: by applying conformal dose distributions (see section 2.2.1) or fractionation schemes (see section 2.2.2).

2.2.1 Delivery techniques

Modern RT techniques, such as intensity-modulated radiation therapy (IMRT) [8, 105], employ multiple photon beams from various directions to achieve a cumulative irradiation effect to the tumour volume while minimising damage to healthy tissue. In IMRT, multi-leaf collimators (MLCs) are used to modulate the beam shapes and intensities to tailor the dose distribution conformally to the tumour with sharp dose gradients outside the target region. MLCs use thin bars of metal, called leaves, to block selective parts of the irradiation beam. Figure 2.2 illustrates an MLC. In static or step-and-shoot IMRT, the beams are split into a set of segments with differing MLC shapes and delivery is switched off while the MLC leaves move. Figure 2.3 illustrates the dose distribution

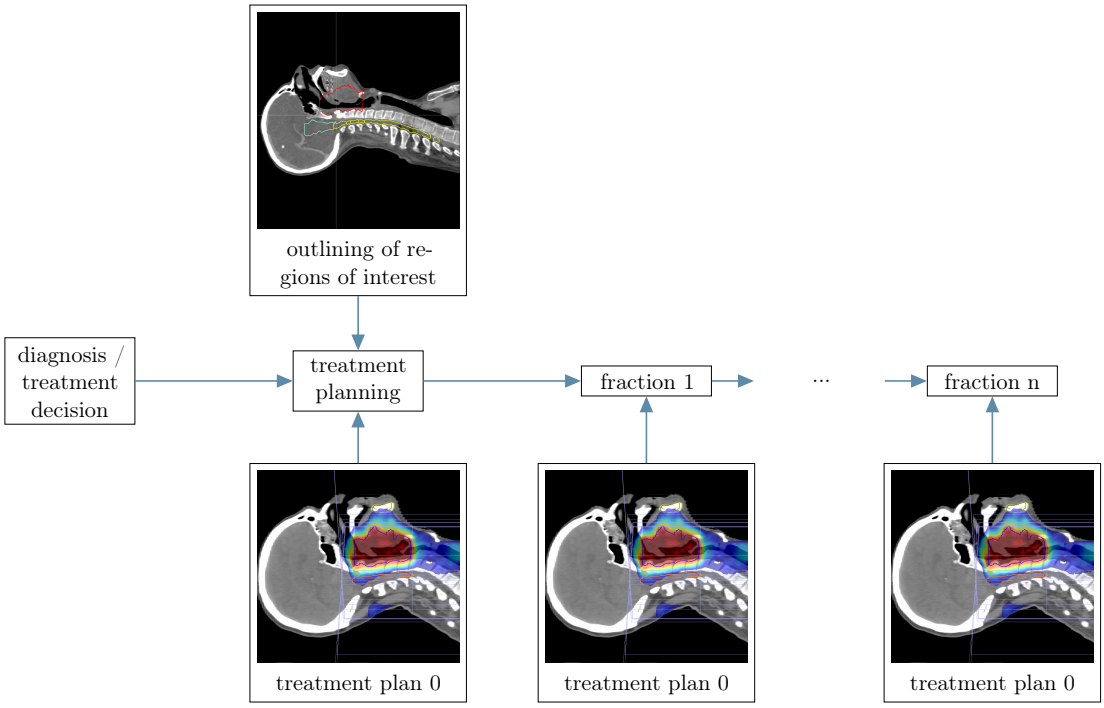


Figure 2.4: Conventional workflow in RT. Prior to the treatment, all regions of interest are outlined and a treatment plan is created. This treatment plan is then used for all remaining fractions of the treatment.

together with the MLC shapes for a step-and-shoot IMRT with seven beams of an HNC patient.

In dynamic IMRT or volumetric modulated arc therapy (VMAT), the modulation is achieved by continuously changing the beam's shape and intensity while the gantry rotates around the patient [109]. One or multiple arcs deliver the radiation as opposed to a fixed number of beams in static IMRT. IMRT can create highly conformal dose distributions.

2.2.2 Treatment schedule and planning

RT is typically administered on multiple days (fractions), taking advantage of the different radiobiological characteristics of tumourous and healthy tissue. An HNC RT treatment course usually consists of 30 to 40 fractions over six weeks. Figure 2.4 illustrates the conventional workflow of an RT treatment: before irradiation, a CT scan of the patient is acquired, and trained clinicians outline all ROIs, i. e. target regions and OARs, on the scan. A treatment plan which optimises the balance between sufficient dose coverage of the target regions and minimal doses to OARs is then created. This treatment plan is conventionally used over the whole course of the treatment. The following sections describe each of these steps in more detail.

2.2.2.1 Patient geometry

Before the first treatment fraction, the patient is scheduled for a CT scan. This CT scan provides three-dimensional (3D) information on the patient's anatomy. In a treatment planning system (TPS), the anatomy can be visualised and clinicians outline all ROIs: the gross target volume (GTV), clinical target volume (CTV), planning target volume (PTV) and OARs.

The GTV encompasses the visible tumour, as well as potentially involved lymph nodes where the tumour cells might have spread. The CTV is created by including a more extensive region around the GTV to account for a microscopic spread of the disease. In HNC there are typically two CTVs: the primary or boost CTV and the secondary or elective CTV. The primary CTV includes the primary, visible tumour, possible microscopic spread, as well as possible regional spread of tumour cells into lymph nodes. The secondary CTV includes regions of the presumed spread of tumour cells, distant from the primary tumour [67]. Finally, the purpose of the PTV is to ensure that the prescribed dose is delivered to the CTV with a clinically acceptable probability. An isotropic margin usually expands the CTV to the PTV, accounting for uncertainties in beam alignment, patient positioning, organ motion, and organ deformation [67].

Outlining the OARs is crucial to ensure that there is no unnecessary dose delivered to them. Analogous to the PTVs, margins can be added to account for uncertainties in patient positioning or day-to-day changes in the patient's anatomy. The resulting volume is called the planning risk volume (PRV). The physician then prescribes a dose to the target volumes and tolerance values for the OARs, guided by previous clinical experience. As these criteria are in general competing conditions, they are assigned priorities. In RT for HNC, there are typically two dose levels. The primary or boost PTV is treated to a higher dose, typically 65 Gy, whereas the secondary or elective CTV is treated at a lower dose, typically 54 Gy. Treatment planning aims to optimise beam settings to meet the pre-defined criteria in the best way possible.

2.2.2.2 Treatment plan creation

Modern RT employs inverse treatment planning [8]. Each of the pre-defined criteria is assigned a weight according to their rank in a priority list. A cost function, which incorporates the weighted conditions on target volumes and OARs, is then iteratively optimised with the help of a TPS. Calculating dose requires knowledge of the electron densities on the beam paths. This information is typically obtained from the voxel intensities of the acquired CT image, given in Hounsfield units (HU), which can be converted into electron densities.

Table 2.1: Example of clinical treatment planning goals, prescribing a mean dose of 65 and 54 Gy to the primary and secondary planning target volumes (PTVs), respectively. Depending on the clinical case, priorities 2a and 2b can change order. The index x in D_x refers to the type of dose volume parameter (e. g. "x =1cc": minimum dose to 1cc, "x =mean": mean dose.

priority	volume of interest	clinical goal
1	spinal cord	$D_{1cc} < 46$ Gy
1	spinal cord + 3mm PRV	$D_{1cc} < 48$ Gy
1	brainstem	$D_{1cc} < 54$ Gy
1	brainstem + 3mm PRV	$D_{1cc} < 56$ Gy
2a	primary PTV	$D_{99\%} > 90\%$ of 65 Gy
2a	primary PTV	$D_{95\%} > 95\%$ of 65 Gy
2a	primary PTV	$D_{50\%} = 65 \pm 1$ Gy
2a	secondary PTV	$D_{99\%} > 90\%$ of 54 Gy
2a	secondary PTV	$D_{95\%} > 95\%$ of 54 Gy
2a	secondary PTV	$D_{50\%} = 54 \pm 1$ Gy
2b	optical nerves	$D_{1cc} < 54$ Gy
2b	optical nerves + 1mm PRV	$D_{1cc} < 55$ Gy
2b	chiasm	$D_{1cc} < 55$ Gy
2b	chiasm + 1mm PRV	$D_{1cc} < 56$ Gy
2b	optical lenses	$D_{mean} < 6$ Gy
3	parotids	$D_{mean} < 26$ Gy

2.2.2.3 Dose-volume criteria

During creation and evaluation of a treatment plan, the planner ensures that the imposed criteria can be fulfilled. The criteria are typically expressed in the form of delivering a minimal, maximal or mean amount of dose to a particular region of the patient's anatomy. They are also known as dose-volume constraints, and their fulfilment can be verified by determining the cumulative dose-volume histograms (DVHs) of individual ROIs. A DVH relates the radiation dose to the volume of tissue. Standard terminology for dose-volume constraints is as follows:

- D_V, the dose D to a volume V. V is usually expressed as fractional (in percentage) or absolute volume (in cm^3). A typical example is "$D_{95} > 95\%\, D_{pres}$", which requires the dose to 95% of the PTVs to be larger than 95% of prescribed dose D_{pres}. Another example is "$D_{1cc} < 46\,Gy$", which requires the minimum dose to any $1\,cm^3$ of the spinal cord to be smaller than 46 Gy.

- D_{mean}, the mean dose to the ROI. A typical example for HNC is "$D_{mean} < 26\,Gy$" for the parotids.

- V_D, the fraction of a volume receiving a dose D or higher. A typical example is "$V_{95} > 95\% D_{pres}$", where more than 95% of the PTVs should receive a dose of at least 95% of the prescribed dose.

Table 2.1 provides an example of such a priority list of the planning constraints for HNC.

2.3 Imaging modalities for radiotherapy

With the invention of the CT in 1972 [64], treatment simulations in three dimensions became possible. CT images provide an excellent contrast between bone and soft tissues or air. Furthermore, information on the electron densities can be extracted, which are crucial for treatment planning. Despite these advantages, there are several drawbacks to this imaging modality, as further discussed in section 2.3.1. Other imaging modalities, e. g. positron-emission tomography (PET) and MRI, are, therefore, often integrated into the RT workflow [20, 27, 31, 38, 41, 54, 80, 95, 114, 121, 160]. These can provide additional information about the patient's anatomy and help in planning and delivering even more precise treatments. The following sections introduce the basics for these imaging techniques.

2.3.1 Computed tomography

CT is based on an x-ray tube and detectors which are rotating around the patient while continuously acquiring 2D projection images, a dataset denoted as a sinogram. A 3D volume is obtained from the sinogram via tomographic reconstruction. CT images represent an essential part of modern RT workflows. As described above, a CT scan is typically acquired before the treatment to outline the target volumes and OARs, and to determine a dose distribution tailored to these ROIs. CBCT scans are often used for patient positioning verification or treatment monitoring. For CBCT, a kilo-voltage x-ray tube is attached to the linac and CBCT images can be acquired just before treatment delivery [70]. However, due to the quality of the beam collimation and more substantial scattering effects compared to CT images, the image quality is inferior. Additionally, especially in the head and neck, artefacts can be large for both imaging modalities due to dental implants.

In comparison to PET and MRI, the acquisition times of CT and CBCT are very short. Furthermore, CT images are geometrically accurate and can be acquired at a high spatial resolution. As mentioned before, they provide information on the electron densities of the tissues, which are crucial for RT treatment planning. Drawbacks of x-ray based images are their poor soft-tissue contrast, as well as an additional dose to the patient.

2.3.2 Positron emission tomography

PET is based on tracers that are marked with radioactive substances. The most commonly used tracer is Fluoro-Deoxy-Glucose (FDG). PET yields the metabolic activity of tissue

and is hence valuable in imaging tumours. In current clinical practice, it is used to help define the target volume and particularly metabolically active regions to boost, or monitor treatment response [27, 41, 114]. Due to its low spatial resolution, combined PET-CT scanners are often used. While PET provides functional information on ROIs, the CT can provide anatomical and precise spatial information.

2.3.3 Magnetic resonance imaging

In contrast to x-ray based imaging techniques, MRI can provide an excellent soft tissue contrast. MRI exploits the difference in magnetic properties for different tissues to generate images. Thorough introductions to the physics of MRI have been published, e. g. by Bernstein et al. [6] and Brown et al. [10].The following paragraphs describe in a nutshell the basic concepts behind MRI.

A strong magnetic field (typically 1.5 to 3T in clinical MRI scanners), as well as a sequence of radio-frequency (RF) pulses and magnetic field gradients, are used to manipulate the spin of the protons of hydrogen atoms in tissues of the human body. MRI can be understood by elements of a classical and quantum-mechanical picture. For the basic concepts, it is sufficient to rely on the classical picture.

Any nucleus is composed of protons and neutrons. Each proton and neutron has an intrinsic angular momentum called spin. If there is an odd number of protons or neutrons in a nucleus, there is a net spin as not all protons or neutrons can couple to zero. This net spin leads to a moving, electrically charged particle and, therefore, creates a magnetic moment. Particularly relevant for MRI are hydrogen atoms as they are naturally abundant in water and fat tissue in the human body. The hydrogen atom is composed of a single proton. The spins usually are randomly oriented, precessing along their axes. Once placed in a static magnetic field B_0, they tend to align with the direction of the magnetic field and precess around that axis at the so-called Larmor frequency. Due to movements and nuclear interactions, this alignment only happens partially, leaving most of the spins still oriented randomly. All magnetic moments can be summarised as a net magnetisation $\mathbf{M}$. Due to a small excess of spins being aligned parallel to $\mathbf{B}_0$ (only 3 in a million protons in a magnetic field with strength 1 T), $\mathbf{M}$ is also aligned parallel to $\mathbf{B}_0$. Applying an RF pulse to this configuration disturbs this alignment by transmitting energy to the spin system through a rotating magnetic field $\mathbf{B}_1$ perpendicular to the stationary field. Only when the rotation frequency is very close to the Larmor frequency, the energy is large enough to tip $\mathbf{M}$ out of its parallel alignment with $\mathbf{B}_0$ at an angle proportional to the duration time of the RF pulse and the field strength $\mathbf{B}_1$. This phenomenon is called magnetic resonance.

Once the RF pulse is switched off again, the net magnetisation returns to its

equilibrium state, the parallel alignment with $\mathbf{B}_0$, in a spiral movement. This phenomenon is governed by two main underlying processes: T1 and T2 relaxation. The T1 relaxation time is the time after which the parallel component of $\mathbf{M}$ restores its maximum value. The T2 relaxation time is the time by which the transversal component of $\mathbf{M}$ is reduced to 0, as it was the case in the equilibrium state.

The T1 and T2 relaxation times are characteristic of different tissues. MRI exploits this fact to generate contrast between adjacent anatomical structures. The influence of T1 and T2 relaxation in the images can be controlled by varying parameters in RF pulse sequences, such as the echo time TE and the repetition time TR, governed by the Bloch equations [7]. MRI is a very flexible imaging technique were a multitude of different sequences can be used to obtain different contrasts in the images.

2.4 Image-guided adaptive radiotherapy

A treatment plan for RT is traditionally solely based on a single CT scan, obtained before the treatment, and the dose distribution is tailored to the anatomical structures delineated on this scan. This treatment plan is then used for the whole treatment course, which consists of multiple treatment days over several weeks.

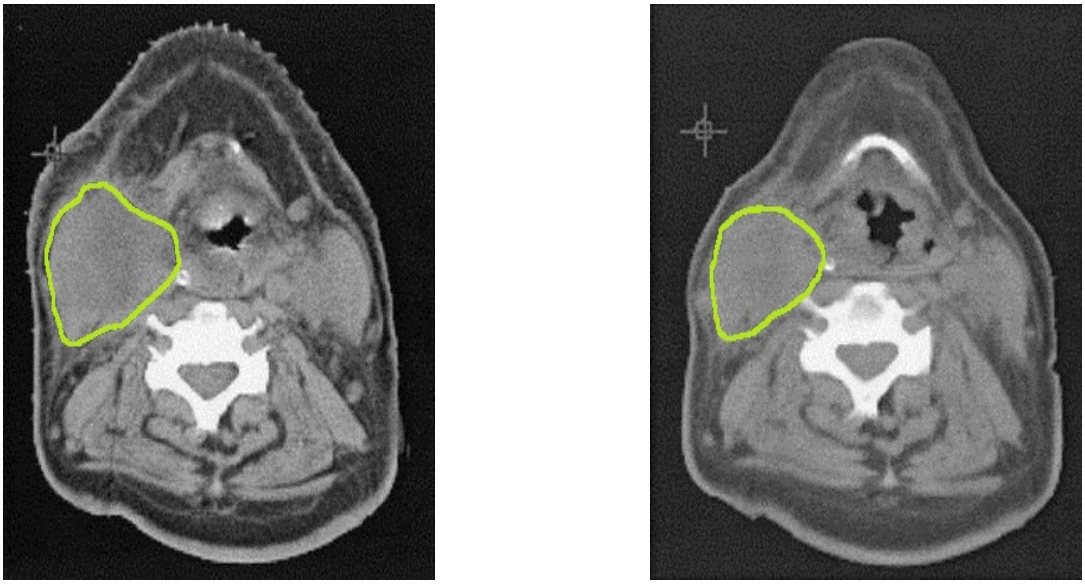
For each treatment fraction, the patient is placed on the treatment couch in the same position as during the planning CT scan. This process is usually guided with external markers placed on the patient during the planning phase. A thermoplastic mask is used to fit the head of HNC patients to reproduce the position from the planning CT at each fraction. As the dose distribution is ideally highly conformal to the target volume, errors in patient positioning can result in a suboptimal coverage of the PTVs, as well as overdosage of the OARs.

2.4.1 Image guidance in radiotherapy

Image guidance can be used to minimise these positioning errors. Commonly, CBCT images are acquired before treatment delivery for this purpose [70]. These images can then be used to compare the patient's current anatomy to the one on the treatment plan and adjust the patient's position such that the radiation can be delivered as planned. These adjustments are usually performed in a rigid manner, aligning the bony anatomy and accounting for translations and rotations of the patient.

2.4.2 Anatomical changes

Using the same treatment plan assumes that the anatomy of the patient stays the same throughout the treatment. However, HNC patients commonly undergo noticeable changes



CT slice
prior to treatment

corresponding CT slice
3 weeks into radiotherapy

Figure 2.5: This figure shows an example of drastic changes in the anatomy of a patient during the course of radiotherapy. The image on the left-hand side shows a CT slice before the treatment, the image on the right-hand side 3 weeks into treatment. The green outline encompasses the GTVs. Figure courtesy: Barker et al. [3].

in their anatomy throughout the treatment course [3, 87]. These include weight loss, changes in tumour size and shape, and normal tissue shrinkage or swelling. Figure 2.5 illustrates an example of a patient prior to treatment and three weeks into treatment. These changes cannot be accounted for with rigid translations or rotations of the patient.

2.4.3 Adaptive radiotherapy

Not accounting for anatomical changes often leads to discrepancies between the planned dose distribution and the actually delivered doses. In conventional RT, anatomical changes are taken into account by adding a safety margin to the CTV. The resulting large target volumes can significantly hinder the success of an RT treatment, as it might be necessary to either compromise on normal tissue sparing or target coverage. Therefore, HNC patients, in particular, would benefit from adapting the treatment to the observed changes in the anatomy [124].

Adaptive radiation therapy (ART) accounts for anatomical changes in an offline or online process. In theory, ART can be undertaken at three different time-scales: offline between fractions, online immediately before a fraction, or in real-time during a treatment fraction. Figure 2.6 illustrates an ART workflow.

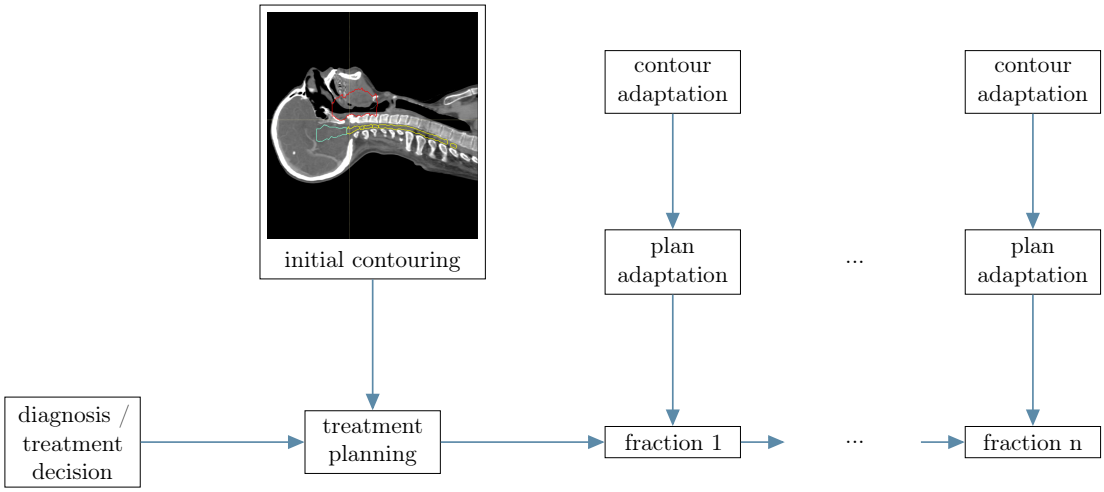


Figure 2.6: This figure illustrates the scenario of an adaptive treatment workflow, where at each or several fractions, an image of the patient is acquired and the treatment plan is adapted according to the observed anatomical changes.

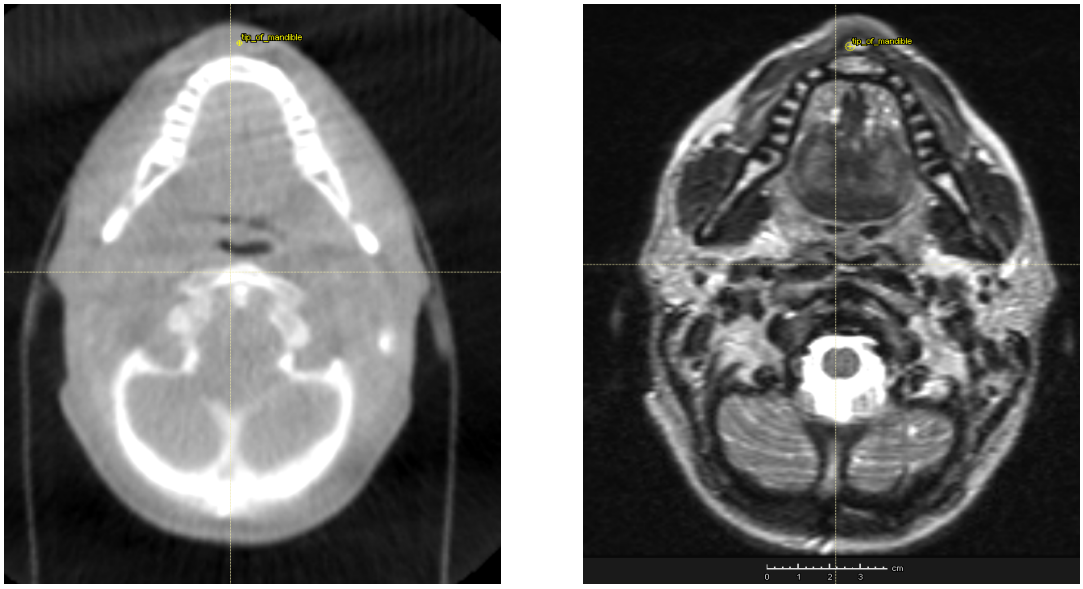
2.4.4 MRI-guided radiotherapy

Ideally, in-room or even on-board imaging devices would be used for ART. CBCTs could be used for this purpose [30, 81, 150, 164]. Recently, in-room image guidance with MRI has been introduced with combined MRI and treatment delivery systems [40, 91, 100, 120]. As with CBCTs, these systems help to inform treatment adaptation based on the current anatomy directly before or during the treatment.

As discussed in section 2.3, MRI can provide a much better soft-tissue contrast than CBCT , has a broad range of flexibility in choosing a specific contrast and does not add any additional dose to the patient. A direct comparison of an MRI and a CBCT scan of the same patient can be seen in figure 2.7.

Within the last few years, several varieties of combined MRI and treatment delivery systems have been installed in multiple centres across the world. Figure 2.8 provides an example of such a system, the Unity MR-linac by Elekta (Stockholm, Sweden), which was recently installed and is now in clinical use at our hospital.

With MRI-guidance, one could exploit the superior soft-tissue contrast of MRI for more accurate localisation of OARs and target volumes and adapting treatments. These adaptations could be offline or even in real-time, depending on the time-scale of anatomical changes. Functional imaging could provide information about the metabolism, diffusivity, perfusion or hypoxia of tumours, which could then be employed for an indication of response to and prognosis of treatments.



(a) axial slice of a CBCT

(b) corresponding axial slice of a (T2w) MRI

Figure 2.7: Direct comparison of (a) CBCT and (b) conventional MRI for an HNC patient on corresponding axial slices. The soft tissue contrast on the MRI is much better than on the CBCT.

Figure 2.8: MRI-guidance: Example of Elekta's Unity system at the Royal Marsden Hospital. Image courtesy: http://www.icr.ac.uk

2.5 Medical image segmentation

One of the key components of a successful RT treatment, in particular for ART, is the accurate localisation of all ROIs in the images acquired throughout the treatment. This process is also called contouring, delineation or segmentation. Manual segmentation is a time-consuming and error-prone process. As there is no access to the ground truth, it is subject to inter- and intra-observer variabilities [26, 46]. With daily or frequent treatment plan adaptations, segmentation poses a considerable burden to the clinical workflow. Automated segmentation can alleviate the burden of manual segmentation while providing consistent contours. However, there are challenges for automated segmentation of medical images due to the presence of noise, low contrast, inhomogeneity, partial volume effects and image artefacts.

A plethora of algorithms applied to the segmentation of ROIs in medical images can be found in the literature, including in-depth reviews [117, 131, 132]. The following paragraphs provide a brief overview of these methods, discuss potential drawbacks and include examples of published applications.

Auto-segmentation methods can be classified into various, not mutually exclusive, categories, such as intensity- and texture-based, supervised and unsupervised, pixel- and region-based, model-based, or based on prior knowledge.

In the early stages, due to a lack of computational power, auto-segmentation algorithms were purely based on image intensity values. These included thresholding and region growing methods.

Thresholding and region growing methods

Thresholding algorithms divide an image into one or multiple classes by defining one or multiple intervals of intensity values for each class, separated by threshold values. The threshold(s) can be set manually or determined automatically, based on intensity histograms [108]. The thresholding method can either be applied locally or globally. It is sensitive to noise and intensity inhomogeneities, which commonly occur in medical images.

Region growing methods start at so-called seed points and expand the region according to some pre-defined criterion, for instance, the homogeneity with respect to the intensities present in that region. If a pixel or voxel meets the criterion, it is included in the region. Similar to thresholding, region growing algorithms are sensitive to noise and inhomogeneities. Furthermore, a seed point is necessary to initialise the algorithm and region growing can therefore not be used in a fully automated manner, unless combined with other algorithms.

In recent publications, thresholding and region growing techniques were rarely used alone but in combination with other methods, with its main application in PET image segmentation [94, 116, 118]. The split-and merge-algorithm is an example of an automated version of region growing. Objects in medical images are often not homogeneous with respect to their intensity values, which limits an application of these algorithms to medical image segmentation.

Watershed algorithm

The watershed algorithm is an edge detection algorithm where the image is represented as topographic relief. Intuitively, flooding the relief with water, the lines dividing areas of water from different basins are known as the watersheds. These lines represent the outlines of the segmented structures. Lim et al. [90] segmented the liver in CT images using a watershed method.

Deformable models

Deformable models, such as active contours and level-set methods, use closed surfaces as initialisation of the algorithm. These closed surfaces can contract or expand to conform to structures within images. Zhuang et al. [170] applied an active contour approach to the segmentation of tumours on PET images of HNC patients. Tan et al. [142] used a combination of watershed and active contour algorithms to segment lung nodules on CT images. Lapeer et al. [84] highlighted the shortcomings of watershed algorithms in the application to the segmentation of abdominal organs in MR images and combined it with an active contour algorithm to overcome these shortcomings.

The auto-segmentation algorithms described so far do not include any prior knowledge but are purely based on the images themselves. The following paragraphs describe algorithms which include some form of prior knowledge.

Statistical shape and appearance models

Extensions to deformable models that incorporate prior knowledge are the statistical shape and appearance models (SSMs). These are frequently used in medical image segmentation. Heimann and Meinzer [58] provide a thorough review of SSM in medical imaging. These algorithms explore the fact that organs have a similar structure even among different individuals. Characteristic variations of shape and appearance are learned from a library of segmented images and build into a model. The segmentation of a new image can be constrained to an anatomically plausible shape using this model. SSM algorithms consist of two main elements: first, a shape model is built, where specific

methods on representing shapes and drawing correspondences between different shapes need to be employed; second, the model needs to be fitted to new images. For the latter, the appearance of shapes, in particular at their boundaries, need to be modelled and a search algorithm is used to find the corresponding shape in the new image. The statistical shape and appearance model is obtained by analysing the shapes of one or more structures in previously segmented template images. Landmarks are commonly used to represent shapes. Correspondences between landmarks in different shapes can be obtained manually, which is very time-consuming and subjective, or automatically. Most algorithms perform registration between the involved shapes. Dimensionality reduction techniques can yield a compact representation of the shape model. This reduction is usually accomplished by computing the mean shape and the most dominant modes of variations via Principal Component Analysis (PCA). Fitting the model to a new image can either be done similar to active contour models, where the deformations are constrained through the modelled shape variations, or by employing a statistical appearance model, learned from the training data. A search algorithm is then used to fit the model to the image. Due to their nature, SSM algorithms are limited to specific shapes and highly depend on the training data.

Classifier and clustering algorithms

Classifiers, such as the k-nearest neighbour algorithm (KNN), determine the pixel- or voxel-wise label by estimation from previously segmented images. In KNN, the pixel or voxel is assigned to the class that is most common amongst the k-nearest neighbours in feature space. The feature space can be derived from intensity values and is specific to the application. The challenge is to choose this feature space such that labels can be distinguished. KNN has, for instance, been applied to the segmentation of brain tissues [153]. An unsupervised variant of using classifiers are clustering algorithms, such as the k-means algorithm. It consists of two iterative steps: it first assigns a pixel or voxel to the closest cluster defined by its distance to the mean of the cluster and then calculates the mean of the cluster from all its elements. This process is repeated until it satisfies a pre-set condition, such as the variance within a cluster. This process has some drawbacks, one of them being the initialisation of the algorithm. If the initial clusters are poorly chosen, the algorithm might either only slowly converge or might not achieve an optimal solution.

As neighbouring voxels in an image are classified independently, classifier and clustering algorithms typically lack contextual information. To mitigate this drawback, these algorithms are often combined with Markov Random Field (MRF) models. MRF models describe interactions between neighbouring pixels or voxels. A difficulty associated

with MRF models is the proper selection of the parameters controlling the strength of spatial interactions. If the strength is set too high, resulting segmentation can be too smooth and details can be lost. If the strength is set too low, it can lead to isolated clusters.

Atlas-based methods

In atlas-based segmentation methods, prior knowledge is used in the form of images and their associated segmentations, so-called atlases. A new image is segmented by obtaining optimal transformations between the atlas images and the new image and by using this transformation to warp the corresponding atlas segmentation to the new image. Atlas-based algorithms consist of two major steps: first, all atlas images are registered to the new image and second, an atlas selection or fusion method is applied. A variety of image registration methods has been reported in the literature for this purpose [110, 129, 144], as well as atlas selection or fusion approaches [2, 24, 68, 83, 125, 161]. To date, atlas-based segmentation methods are the most commonly used auto-segmentation approaches in RT [132]. Chapter 5 describes atlas-based auto-segmentation algorithms in more detail.

Deep learning

Recently, deep learning-based methods have demonstrated great potential in computer vision tasks [12, 127]. Deep learning is a sub-discipline of machine learning, where data representations are learned from problem-specific example cases. More specifically, deep learning uses neural network architectures to learn a specific task from a dataset by representing the data through a hierarchy of non-linear functions. Complex representations of the data are learned by decomposing them into many simple concepts, such as edges and corners. Neural networks have been inspired by the structure of the human brain. Nowadays, the most commonly used deep learning architectures are convolutional neural networks (CNNs) with an increasing number of applications in the field of medical image segmentation. An in-depth introduction to the basic building blocks and relevant concepts of CNNs is given in chapter 6.

Chapter 3

Image acquisition and preparation

This chapter provides details on the acquisition and preparation of all imaging data, which built the basis of the development and design of all auto-segmentation methods of this PhD thesis. Processing of imaging data is an essential component of any deep learning-based method. This chapter mainly serves the purpose to ease the readability, as most of the images and the preprocessing steps were the same for all subsequent chapters and will, this way, not need to be repeated.

3.1 Database: images and annotations

A library of 27 patients, all with a tumour at the base of the tongue and treated with RT at the MD Anderson Cancer Center (Houston, Texas, USA), was available for this thesis. All patients had a baseline CT scan, as well as baseline T1-weighted (T1w) and T2-weighted (T2w) MR scans, typically a few weeks before the RT treatment. One clinician at the Royal Marsden Hospital (RMH, London, United Kingdom) manually delineated the left and parotid glands, the spinal cord and the mandible in all 27 T1w and T2w images. There were not enough axial slices in the MR images to cover the required anatomy for treatment planning in the superior-inferior direction. The focus of this thesis was therefore on the four mentioned OARs. Optical structures and the brainstem were outside the covered imaging field of view.

For the dosimetric evaluation in chapter 4, dose was calculated with the information on the electron densities of tissues from the CT images. To create a valid treatment plan, two clinicians at the RMH manually outlined the primary (including involved lymph nodes) and secondary (including more distant lymph nodes) CTVs, the optical nerves and lenses, the chiasm and the brainstem on the CT images.

All ROIs were manually delineated using the TPS Raystation (Raysearch, Stockholm, Sweden). The parotid glands, the spinal cord and the mandible were warped from the MR images to the CT images by employing the deformable image registration framework ADMIRE (research version 1.1, Elekta AB, Stockholm, Sweden). Figure 3.1 illustrates axial, sagittal and coronal slices of all imaging modalities for one example patient, together with the manually segmented ROIs. The last column guides the reader to the respective chapters in which these images were employed. Table 3.1 lists the relevant image acquisition parameters for each imaging modality.

202 CT images from the publicly available database of the Cancer Imaging Archive [53], as well as the MICCAI HNC segmentation challenge [123], together with the manual

Table 3.1: Imaging parameters of the main database (T1w and T2w MR, as well as CT images) for the design of all auto-segmentation algorithms of this thesis.

parameter	T2w MR	T1w MR	CT
FOV [#pixels]	512x512	512x512	512x512
#slices	30	30	[165, 235]
voxel size [mm^3]	0.5x0.5x4	0.5x0.5x4	0.98x0.98x2.5
TE [ms]	[96.72, 107.30]	[6.54, 7.85]	n.a.
TR [ms]	[3198, 4000]	[601, 800]	n.a.
flip angle [°]	90	90	n.a.
sequence type	2D T2w spin echo	2D T1w spin echo	n.a.
field strength/tube voltage	3 T	3 T	120 keV

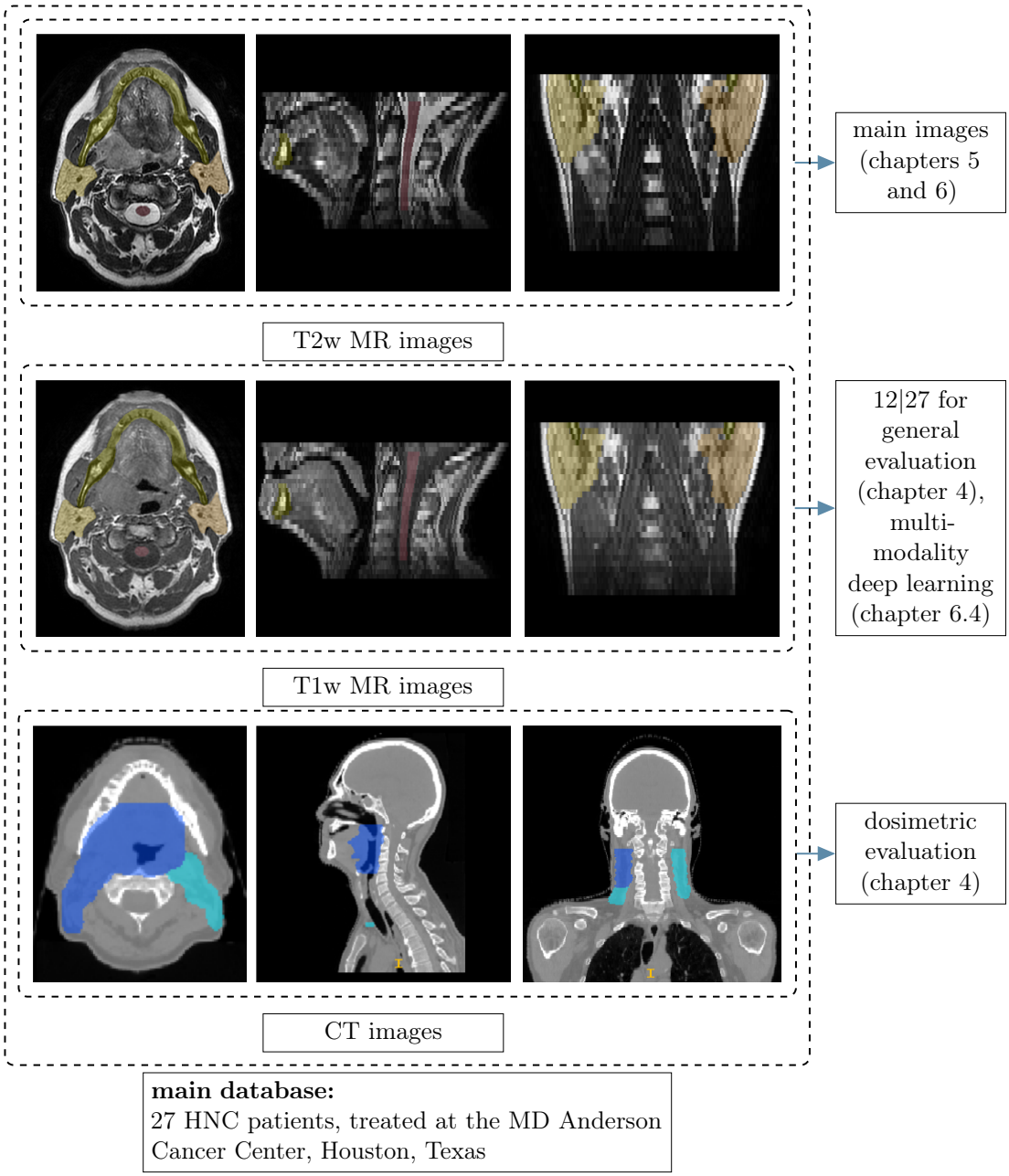


Figure 3.1: Guide to imaging data employed in this thesis: Each column represents axial, coronal and sagittal slices of the T2w and T1w MR images, as well as the CT images. The coloured regions represent the manually segmented ROIs of the primary PTV (blue) and the secondary PTV (turquoise) on the CT, as well as the left (orange) and right (yellow) parotids, the mandible (green) and the spinal cord (red) on the MR images. The last column provides references to the chapters in which they were used.

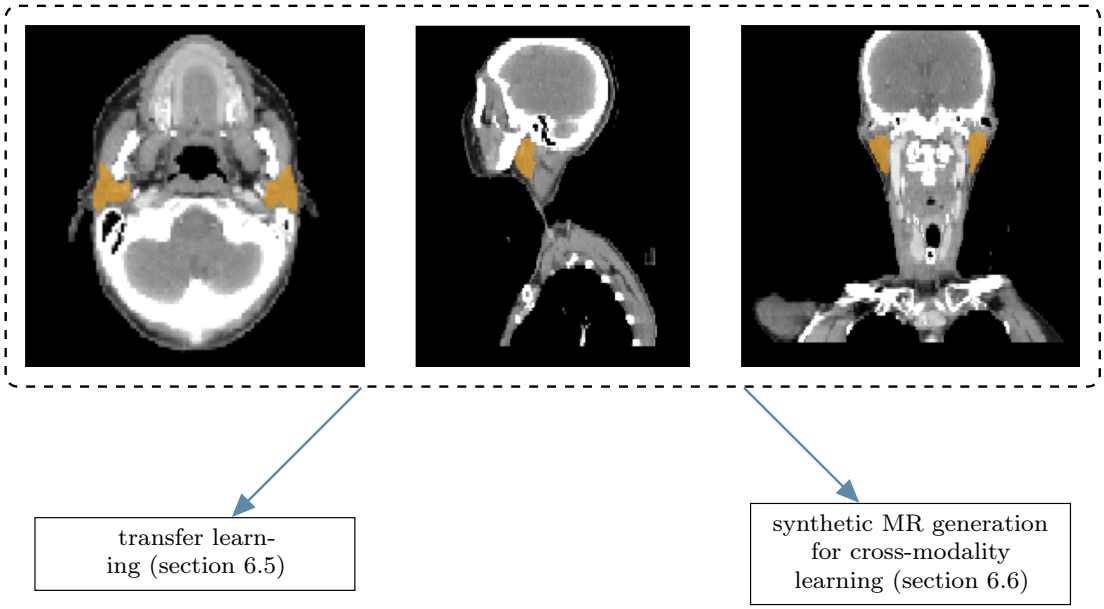


Figure 3.2: This figure illustrates an axial, sagittal and coronal slice of one patient from the database of 202 CT images (Cancer Imaging Archive [53], as well as the MICCAI HNC segmentation challenge [123]), together with the manual segmentation of the parotid glands (orange). These CT images served as additional data for the development of the domain adaptation methods in chapters 6.5 and 6.6.

segmentation of the parotid glands, served as additional data for the development of the domain adaptation methods in chapters 6.5 and 6.6.

3.2 Data processing for deep learning methods

3.2.1 Resolution and field of view

A bottleneck for deep learning-based methods is the graphical processing unit (GPU) memory. To handle this problem, I downsampled each axial MR slice by a factor of 2x2. To match the resolution of the MR images with the CT images for the transfer learning method (chapter 6.5) and cross-modality learning (chapter 6.6), I further resampled all CT images from the public database to an axial plane resolution of 1x1 mm^2. Compared to the MR images, the field of view of the CT images was larger in both, axial and sagittal planes. I, thus, cropped the CT images to a window of 256x256 voxels in the axial plane. This window was determined by finding the external outline of the head using Otsu thresholding [108] and a binary closing method [35] to fill potential holes.

3.2.2 Intensity scaling

As image intensities can vary between MR images, I standardised the contrast with an intensity histogram-based thresholding technique. With the histogram $h(i)$ of intensity values i, the minimum intensity $i_{\min}$ and maximum intensity $i_{\max}$, the histogram is normalised as:

$$h'(i) = \frac{N_{\text{voxels}} - \sum_{j=i_{\min}}^{i} h(j)}{N_{\text{voxels}}} = \frac{\sum_{j=i}^{i_{\max}} h(j)}{\sum_{j=i_{\min}}^{i_{\max}} h(j)}, \tag{3.1}$$

with the cumulative distribution function $\text{cdf}[h(i), h(i_{\min})] = \sum_{j=i_{\min}}^{i} h(j)$. All intensities i with $h'(i) < 10^{-4}$ were set to the maximum intensity and 100 bins were used in the histogram. To standardise voxel intensities and increase the visibility of the parotids, I rescaled the CT images to the recommended soft-tissue window (level 40, window 350 HU) [62]. Additionally, I mapped image intensities for both CT and MR images to a range of intensities between 0 and 255. This mapping is a standard procedure in deep learning (see also chapter 6.2.3).

Chapter 4

Validation of auto-segmentation methods in radiotherapy

The evaluation of auto-segmentation methods suffers from a lack of ground truth. Commonly, it is assumed that the manual segmentation by one or more experts approximates this ground truth. Geometric features are employed to compare the auto-segmented regions of interest to corresponding manually segmented ones and determine the algorithm's performance. This chapter introduces the fully automated validation workflow, designed to evaluate any auto-segmentation method in the context of RT.

4.1 Introduction

Before routinely using auto-segmentation methods in a clinical RT workflow, it is necessary to ensure that they achieve adequate accuracy for RT planning. It is, therefore, crucial to evaluate any auto-segmentation method and understand where errors might occur before implementing them into the clinical workflow. However, the evaluation of auto-segmentation methods is known to suffer from the lack of access to the ground truth. The ground truth would require surgery and is hence challenging to obtain. Commonly, one or more experts manually segment the images to define a gold standard. Despite being subjective, it is assumed that this gold standard approximates the ground truth well. Evaluation of auto-segmentation algorithms is then performed relative to that gold standard. If the auto-segmentation agrees well, one can ensure that treatment planning is at least as good as what a clinician could achieve. The inter- and intra-observer variability can provide an upper bound on the achievable accuracy.

Frequently, the performance of auto-segmentation algorithms is evaluated in terms of purely geometric criteria. However, in RT, the delineated ROIs are used to guide the optimisation process, which balances good dose coverage of the tumour with minimising the dose to OARs. Therefore, it is crucial to address the dosimetric impact of segmentation inaccuracies in the process of generating treatment plans. A few groups have addressed this need and looked at dosimetric differences on CT images [4, 25, 37, 101, 146, 152] with various methods. Nonetheless, to my knowledge, no single geometric measure has been observed to be suitable for the prediction of the dosimetric implications so far.

For this purpose, I developed a method to analyse the impact of geometric differences on dosimetric features of the planned dose distribution. I integrated this into a fully automated workflow to evaluate any auto-segmentation method within the context of RT. A dosimetric evaluation is time-consuming and cannot be done routinely at this stage. I, therefore, additionally investigated to what extent geometric measures are sufficient surrogates for the dosimetric impact.

This chapter describes the developed workflow in detail, which is mainly derived from the publication Kieselmann et al. [72]. It uses the example of atlas-based auto-segmentation algorithms for demonstration purposes and to find an answer to the question, whether purely geometric features can be used as surrogates for key dosimetric features.

4.2 Materials and methods

4.2.1 Data acquisition and preparation

12 T1w images, together with corresponding CT images, as introduced in figure 3.1 on page 39, served as imaging database for this study. I restricted this study to 12 out of the 27 patients, as at the time of this study only these had all necessary ROIs manually delineated and treatment planning was time-consuming. Unlike the MR images, the CT images provided information on the electron densities of tissues, as well as on the required anatomy for treatment planning. Another strategy to obtain electron densities would be to generate a synthetic CT from the MR image and calculate treatment plans optimising on these images. However, as of the time of this study, there was no approach available that could accurately predict CT from MR images and furthermore, the issue with the restricted coverage would not have been solved. I address strategies to generate synthetic CTs in chapter 6.6.

4.2.2 Fully automated evaluation workflow

To implement any auto-segmentation algorithm in a clinical RT workflow, one needs to perform a thorough validation. To facilitate this process, I established a fully automated workflow consisting of

(1) automated segmentation (see section 4.2.3 and more generally in chapters 5 and 6)

(2) automated treatment planning for any set of ROIs using a template approach (see section 4.2.4)

(3) automated geometric and dosimetric evaluation of auto-generated ROIs where manually drawn contours serve as the gold standard reference (see section 4.2.5)

(4) benchmarking the automated segmentation algorithm against the inter-observer variability (see section 4.2.6)

(5) a one-time correlation analysis between geometric and dosimetric evaluation measures to determine whether these are coherent (see section 4.2.7)

While the dosimetric impact of segmentation inaccuracies on treatment planning is an important quantity, thorough dosimetric evaluation is time-consuming and the geometric evaluation would be the preferred method as a surrogate estimation. To determine whether there was a sufficient correlation between geometric and dosimetric evaluation measures I, therefore, performed a correlation analysis in this study (step 5).

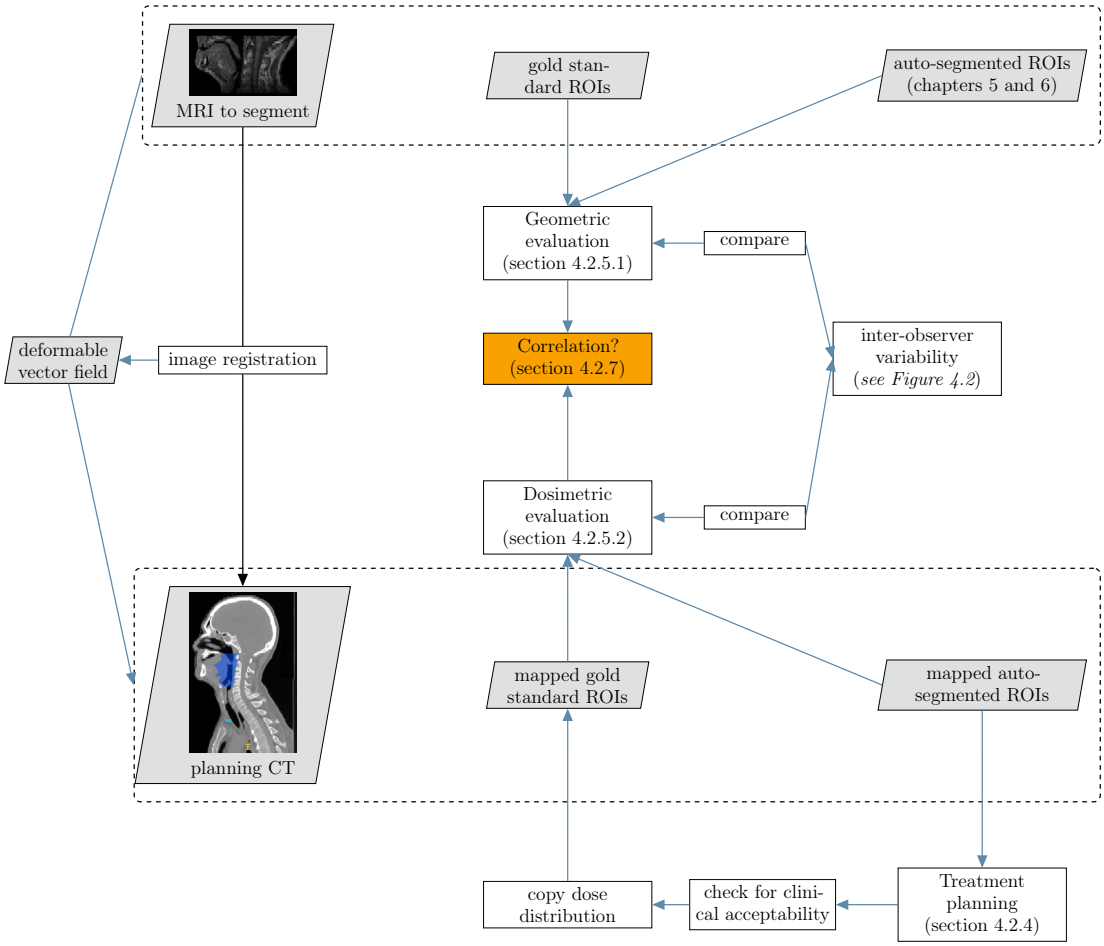


Figure 4.1: This figure illustrates the fully automated validation workflow to evaluate any auto-segmentation algorithm in the context of RT. The highlighted part in orange is the correlation analysis to determine whether a geometric evaluation suffices as a surrogate for key dosimetric features of a treatment plan. The top row illustrates an MR image, together with its gold standard and auto-segmented ROIs. To perform a dosimetric analysis, I registered the MR to its corresponding CT image via deformable image registration and used the resulting deformable vector field to warp the segmented ROIs from the MR to the CT. The central part of this figure shows the building blocks of the evaluation: the geometric and dosimetric evaluation, as well as the comparison to the inter-observer variability.

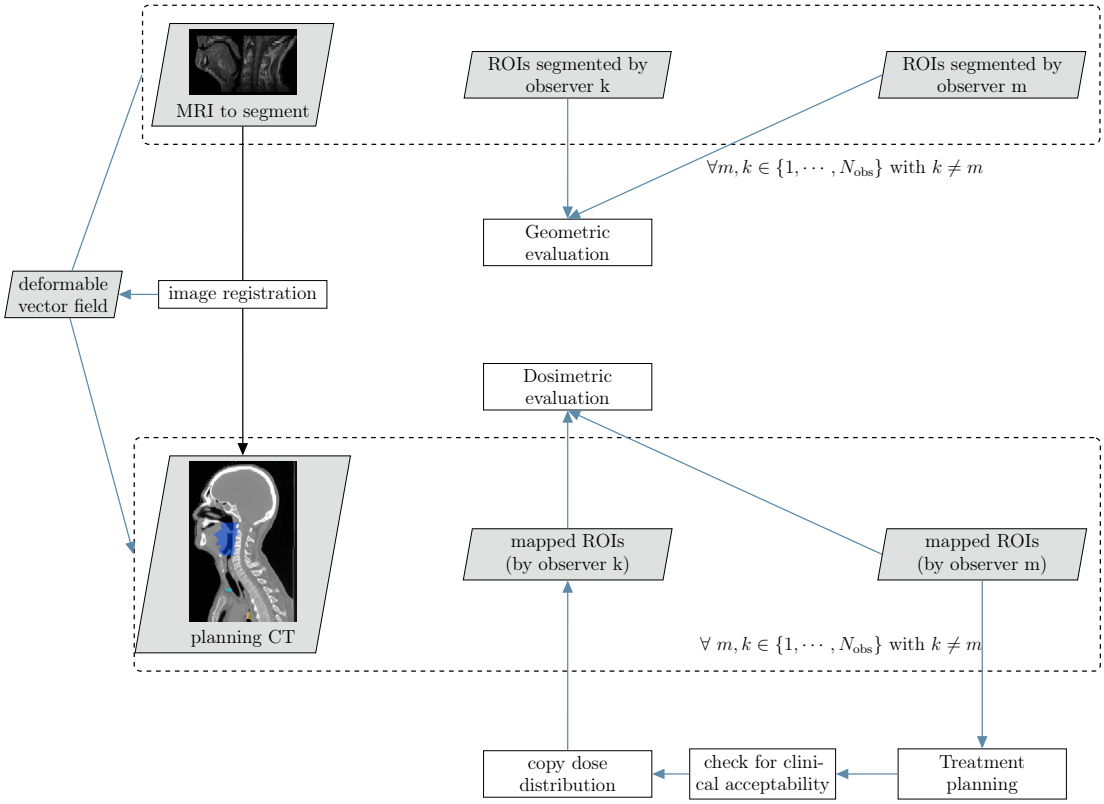


Figure 4.2: This figure illustrates how the inter-observer variability was determined for this study. For each observer pair (k,m) I determined the geometric and dosimetric differences in the same way as I did before with the gold standard and auto-segmented ROIs. The variability was then determined as the average of these pairwise differences.

We published this workflow, shown in detail in figure 4.1, using the example of atlas-based segmentation approaches [72]. It can easily be adapted to evaluate any auto-segmentation approach within the scope of RT.

Instead of manually creating treatment plans, I established an automated treatment planning process. Such an approach removes additional observer variation from the planning process and hence increasing treatment plan comparability. The inter-observer variability, determined with the workflow in figure 4.2, provided a benchmark for the segmentation algorithm. The following sections describe each of the steps of this workflow in detail.

4.2.3 Automated segmentation method (step 1)

For this study, atlas-based auto-segmentation methods were selected. These are described in detail in chapter 5. Three different fusion methods were employed:

(1) method A: best atlas

(2) method B: weighted majority voting (maWMV)

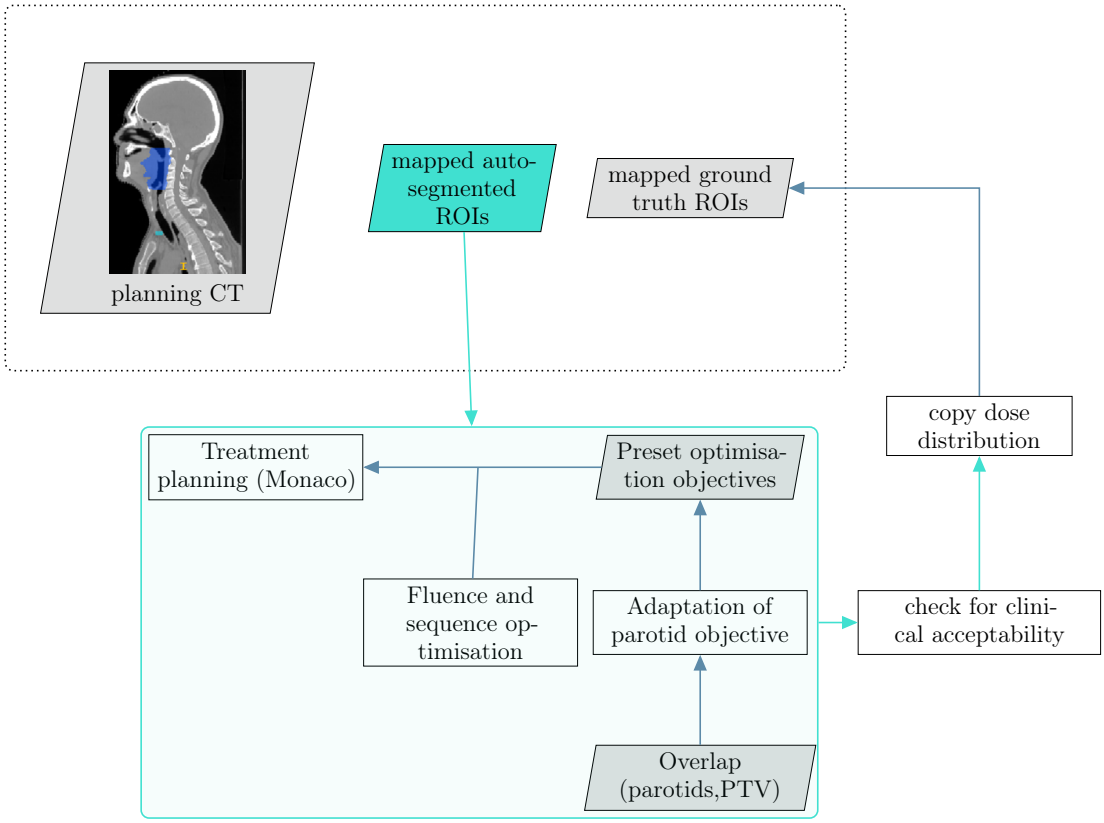


Figure 4.3: This figure illustrates the workflow to automatically generate a treatment plan for any set of auto-generated contours on the MR images. For each set of auto-segmented ROIs, a treatment plan is automatically generated, with an adaptation of the parotid constrant (see light turqouise box at the bottom). After ensuring clinical acceptability, the dose distribution is superimposed to the manually segmented "ground truth" ROIs.

(3) method C: STEPS (Similarity and Truth Estimation for Propagated. Segmentations; maSTEPS)

Method A employs a single atlas, the best in terms of similarity to the target image, to segment the target image. Methods B and C combine multiple atlases to predict the segmentation of the target image.

To evaluate the geometric and dosimetric accuracy of the auto-segmentation methods, I devised a planning study based on a leave-one-out cross-validation strategy: I applied the three auto-segmentation methods for each patient of the library described in section 4.2.1. The MR image of the respective patient was excluded from the library and used as the target, with the atlas library comprising the remaining MR images.

4.2.4 Automated treatment planning strategy (step 2)

To eliminate the uncertainties in the optimised plans which are introduced by the subjective and personal view of the treatment planner, I implemented an automated plan

generation approach making use of the research scripting interface of the TPS Monaco (research version 5.19.03, Elekta AB, Stockholm, Sweden [23]). Figure 4.3 illustrates the workflow of the automated plan generation. I generated treatment plans for a 9-beam step and shoot IMRT treatment on the Unity MR-Linac (Elekta AB, Stockholm, Sweden) prescribing mean doses of 65 Gy to the primary PTV and 54 Gy to the secondary PTV in 30 fractions, following the INSIGHT study protocol [158]. Details on the clinical goals are listed in table 2.1 on page 26.

Treatment plans were created for all auto-segmentation methods and the dose distributions superimposed on the gold standard ROIs. The treatment plans were generated on the CT images, which is why I warped the automatically and manually segmented OARs from the MR to the corresponding CT scans as described in chapter 3. I expanded the CTVs with a margin of 3 mm to obtain the PTVs . The brainstem and the spinal cord were expanded with a margin of 3 mm, the optical nerves and chiasm with a margin of 1 mm for the planning risk volumes (PRVs).

To calculate the dose, I used the GPU-based Monte Carlo dose engine (research version of GPUMCD, Elekta AB, Stockholm, Sweden [61]). As I was simulating treatments on the MR-Linac, I chose the MR-Linac beam model for a magnetic field of 1.5 T. I normalised each dose distribution such that 95% of the primary PTV was covered by 95% of the prescribed dose.

I defined a template cost function which incorporated optimisation objectives on the target volumes and OARs . For the sample of patients used for this study, there was a considerable overlap of the parotids with the target volumes. Therefore, the sparing of the parotids was challenging to achieve, and I chose to loosen the optimisation objective, as well as the clinical goal for the parotids. With the original condition being

$$D_{\text{mean}} < 26\,[\text{Gy}], \tag{4.1}$$

I determined the objective as a function of the overlap volume (OV) with the primary PTV :

$$D_{\text{mean}}(\text{OV}[\%]) < 24\,[\text{Gy}] + 0.6\,[\text{Gy}] \cdot \text{OV}[\%]. \tag{4.2}$$

This heuristic strategy has proven to be useful in clinical practice [65]. It represents the clinical guidelines at our hospital, where target coverage and the sparing of the brainstem, the spinal cord, as well as optical structures, are prioritised over a reduction of dose to the parotids.

The dose distribution, obtained through fluence and sequence optimisations in Monaco (research version 5.19.03, Elekta AB, Stockholm, Sweden), was then checked

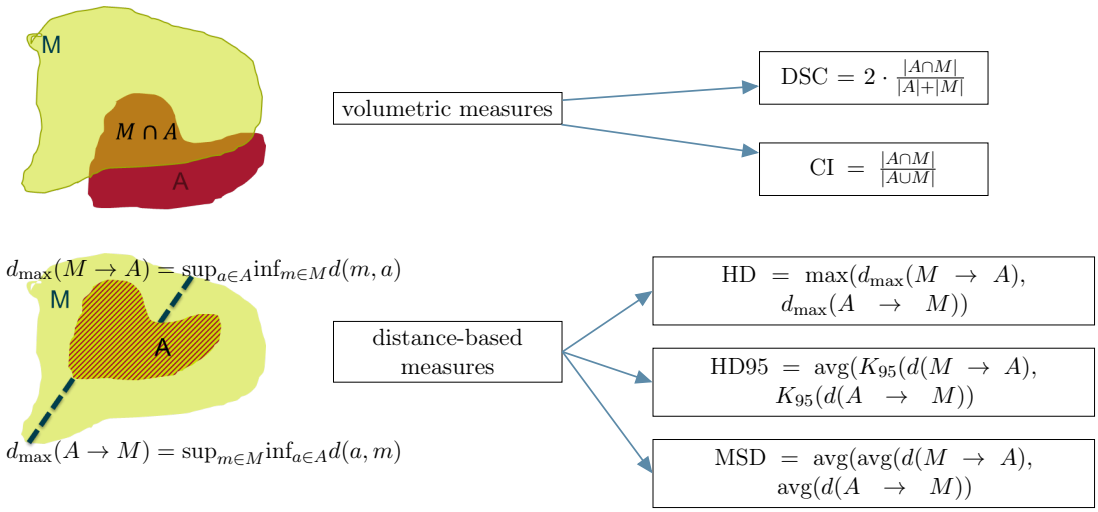


Figure 4.4: Geometric measures: This figure illustrates volumetric and distance-based measures with the example of two shapes A and M. The top part provides formula on the Dice similarity coefficient (DSC) and the Jaccard conformity index (CI), the bottom part the Hausdorff distance (HD), 95th percentile HD (HD95), as well as the mean surface distance (MSD). For reasons of simplicity, the illustration is given in 2D.

for clinical acceptability. I implemented an automated plan check algorithm to analyse whether all imposed clinical goals were fulfilled, using the research interface in Monaco (research version 5.19.03, Elekta AB, Stockholm, Sweden). At this stage, to test the performance of the automated plan algorithm, I additionally asked a clinician to visually inspect the dose distributions to ensure that the plans were clinically acceptable.

4.2.5 Geometric and dosimetric evaluation (step 3)

4.2.5.1 Geometric metrics

Commonly, geometric measures are used to quantify the agreement between two segmented ROIs [132, 141]. These can be generalised into two categories: volumetric and distance-related measures. Figure 4.4 illustrates some examples of these measures. Denote A as the set of auto-segmented points and M as the set of manually segmented points. The most popular volumetric measure reported in the literature is the Dice similarity coefficient (DSC) [29], defined as

$$\mathrm{DSC} = 2 \cdot \frac{|A \cap M|}{|A| + |M|}, \tag{4.3}$$

the number of points common to both sets A and M, normalised to the average number of points in A and M. That means for 3D volumes, the DSC is given as the volume of the overlapping region divided by the sum of both volumes.

A similar measure is the Jaccard conformity index (CI), defined as

$$\mathrm{CI} = \frac{|A \cap M|}{|A \cup M|}, \tag{4.4}$$

the number of points common to both sets A and M, normalised to the union of points in A and M. DSC and CI both range from 0 to 1, where 1 indicates perfect overlap. As most studies in the literature report on the DSC, I chose this as the volumetric measure for comparison purposes.

Volumetric measures are useful in detecting mismatches in size and position of the ROI. However, they do not account for the shape of the structure. It is, therefore, essential to also quantify distance-related measures. The most commonly employed distance-related measures are the Hausdorff distance (HD) and the mean surface distance (MSD). The HD is defined as the supremum (sup), i. e. the least upper bound, of the distances $d(a, m)$ between each point a of one segmented ROI A to its infimum (inf), i. e. the greatest lower bound or the closest point, in the segmented ROI M.

$$\begin{aligned} \mathrm{HD} &= \max(d_{\max}(A \to M), d_{\max}(M \to A)) \\ \text{where } d_{\max}(A \to M) &= \sup_{a \in A} \inf_{m \in M} d(a, m). \end{aligned} \tag{4.5}$$

In other words, it is the largest of all distances from one point in A to the closest point in B and vice versa. The HD is sensitive to outliers because it uses the most mismatched points as a criterion. The 95 percentile (K_{95}) of the HD, on the other hand, refers to the distance which is larger or equal to 95% of all distances between the two segmented ROIs:

$$\mathrm{HD95} = \mathrm{avg}(K_{95}(d(M \to A), K_{95}(d(A \to M)). \tag{4.6}$$

The MSD is a measure of the average distances between surfaces and is defined as:

$$\mathrm{MSD} = \mathrm{avg}(\mathrm{avg}(d(M \to A), \mathrm{avg}(d(A \to M)) \tag{4.7}$$

The smaller the distance measures, the better is the agreement. For this study, I calculated the DSC, the MSD, the 95th percentile of the Hausdorff distance (HD95) and the HD between the automatically and manually segmented ROIs to determine the geometric accuracy.

4.2.5.2 Dosimetric metrics

To adequately address the dosimetric impact of segmentation inaccuracies on the process of generating treatment plans, I used the following steps, which best mimic the use of

such segmented ROIs in a clinical workflow:

(1) Calculate dose distributions, which are optimised for the automatically contoured ROIs. This procedure would be followed clinically if the treatment was based on the automatically proposed contours.

(2) Map the dose distributions from (1) to the respective gold standard ROIs.

(3) Calculate the dose differences between the dose to the automatically contoured and the gold standard ROIs. These differences represent the errors between what one expects to deliver to the ROIs (auto-contours) and what would be delivered (gold standard contours).

Since plan evaluation is based on dose-volume metrics, as introduced in chapter 2, I determined differences in these dose-volume metrics for the auto-segmented ROIs ($D_{x,auto}$) and manually segmented ROIs ($D_{x,man}$). To determine the relative impact, each of these differences was normalised to the respective clinical goal $D_{\mathrm{x,goal}}$:

$$\Delta D_{x,\mathrm{norm}} = \frac{D_{x,\mathrm{auto}}(Gy) - D_{x,\mathrm{manual}}(Gy)}{D_{x,\mathrm{goal}}(Gy)}. \tag{4.8}$$

The index x denotes the type of dose-volume parameter, e. g. "$x = 1cc$" represents the minimum dose to 1cc of the volume and "x =mean" refers to the mean dose. Negative $\Delta D_{x,\mathrm{norm}}$ mean that a larger dose would be delivered to the gold standard than what was planned for the auto-segmented ROIs.

Voet et al. [152] and Beasley et al. [4] used a similar approach to investigate the dosimetric differences and correlations between dosimetric and geometric measures on CT images of HNC patients. To eliminate subjectivity in treatment planning, I established an automated treatment planning strategy, further described in section 4.2.4.

I followed an adaptive approach for the parotids, as described in section 4.2.4, choosing the adaptive dose-volume constraint as defined in equation (4.8). I normalised the difference to the non-adapted clinical goal of 26 Gy. The spinal cord and the mandible were evaluated in terms of the minimum dose to 1 cm^3 with clinical goals of 46 and 67.25 Gy, respectively.

4.2.6 Inter-observer variability (step 4)

4.2.6.1 Overview of measures for inter-observer variability

Even though there are well-defined guidelines for contouring of HNC ROIs [52, 140], there is still a substantial inter- and intra-observer variability [46, 101, 156]. The inter-observer variability can provide an estimate of the upper bound on the necessary

auto-segmentation accuracy. To estimate the inter-observer variability, several observers are usually asked to (repeatedly) delineate the same ROIs on the same images, according to the same guidelines. The variability between the observers can then be determined according to predefined measures.

One algorithm to estimate the inter-observer variability between a set of observers is STAPLE [155]. It uses an Expectation-Maximization algorithm to iteratively estimate a reference standard out of the observers' segmentation, as well as the performance parameters (based on sensitivity and specificity) that quantify agreement of an individual observer with the estimated reference standard.

Another standard measure of the inter-observer variation is the conformity index, defined in equation (4.4) on page 50 for two observers. It can be defined analogously for multiple observers. A problem with this measure is that it strongly depends on the number of observers and can only decrease when more observers are added. While non-overlapping regions always increase or stay the same, overlapping regions can never increase with more observers, thus only allowing for a decrease of the measure. For this reason, Kouwenhoven et al. [77] introduced a generalised CI, defined as the ratio between the sum of all pairs of overlapping volumes and the sum of all pairs of unions of volumes.

A measure of inter-observer agreement, especially in classification problems, is Cohen's kappa or Fleiss' kappa (a generalisation of Cohen's kappa for more than two observers and categories). Cohen's kappa is defined as the observed agreement, normalised to the agreement occurring by chance. A further measure of inter-observer variability is the coefficient of variation, defined as the ratio between the standard deviation (SD) and the mean of all segmentation volumes.

For local observer agreement, one can calculate distances to a common surface map between observers and quantify the variation by, for instance, the standard deviation of the observers' distances [137].

4.2.6.2 Inter-observer variability in this study

To determine the inter-observer variability for the data used in this study, I asked two additional observers to outline the four OARs on the T1w MR images. Each of the observers followed the contouring guidelines defined in Sun et al. [140]. I estimated the inter-observer variability geometrically and dosimetrically. Since the inter-observer variability in this work served the purpose to determine a benchmark for the auto-segmentation algorithms, I employed the same pairwise measures, which I also used to evaluate the accuracy of the auto-segmentation algorithms.

Geometric inter-observer variability

To determine the geometric inter-observer variability between two observers, I first calculated the DSC, HD, HD95 and MSD (see section 4.2.5.1) between the respective observers' contours for each patient and defined the pairwise inter-observer variability as the average and SD of all patients. The overall inter-observer variability was then calculated as the average of the three pairwise inter-observer variabilities, with the overall SD being the root mean square (RMS) of the sum of the three individual SDs.

Dosimetric inter-observer variability

To determine the dosimetric inter-observer variability, I superimposed the dose distribution, which was optimised on the auto-segmented ROIs, on each of the three sets of manually segmented ROIs. I calculated the pairwise differences between the dose to each manually segmented and auto-segmented ROI according to equation (4.8) on page 51 for each patient and ROI. I approximated the dosimetric variability by the SD of these three dose difference values. From this dosimetric variability per patient and ROI, I estimated the overall variability for each ROI by calculating the mean and SD over all patients.

4.2.7 Geometric measures: suitable predictors for dosimetric accuracy? (step 5)

To determine whether geometric measures, such as the DSC and HD95, can reliably predict the dosimetric impact on planned dose-volume parameters, I investigated the correlation between the geometric and dosimetric quantities by calculating Spearman's correlation coefficients [136]. I calculated the correlation coefficients individually for the three different auto-segmentation approaches as these were determined for the same set of patients and could therefore not be treated as independent. Additionally, I performed a qualitative analysis by visual inspection of individual patient images to understand the dependency of the correlation on the shape and the size of the OAR, the dose metric, as well as the relative position to the target volume (i. e. location within large dose gradients).

4.2.8 Statistical evaluation

Tests for statistically significant differences were performed using Student's paired t-test [139] at a significance level of $p=0.05/3$ with a Bonferroni correction to account for multiple comparisons. As a condition of the paired t-test is the normal distribution of the data, I tested the results for normality by visual inspection of Q-Q-plots. All analyses

were performed using automated scripts I developed in Python. As for dosimetric differences the variance (standard deviation) is the most important quantity, I applied a Levene's test [89] to determine significance for the dosimetric evaluation.

4.3 Results

4.3.1 Geometric and dosimetric evaluation

Figure 4.5 shows boxplots of the DSC, MSD, HD95 and HD, as well as the dosimetric differences ΔD_{norm} for all ROIs and the three atlas fusion methods. Negative $\Delta D_{\text{x,norm}}$ mean that a larger dose would be delivered to the gold standard than what was planned for the auto-segmented ROIs. The stars indicate statistical significance, as defined in section 4.2.8. Tables 4.1 and 4.2 list the mean and standard deviations for all applied evaluation measures. The inter-observer variability was included as a reference value.

While there were statistically significant improvements when using one of the multi-atlas approaches B or C for all ROIs, no method was superior in terms of dosimetric differences. Dose differences took both positive and negative values but were close to a zero mean for all ROIs and segmentation approaches. Differences as large as 23 %

Table 4.1: Geometric evaluation for all ROIs and auto-segmentation approaches: mean values for DSC, MSD, HD and HD95. All mean values have been calculated by averaging over all 12 patients. For a reference, I also include the inter-observer variability, derived from the manual contours of three different observers.

ROI	method	$\overline{DSC}$	$\overline{HD}$ [mm]	$\overline{HD95}$ [mm]	$\overline{MSD}$ [mm]
right parotid	A	0.74±0.04	15.07±5.03	6.84±1.95	2.24±0.75
	B	0.80±0.03	16.51±6.96	5.65±1.41	1.61±0.43
	C	0.81±0.02	13.33±5.20	5.20±0.97	1.56±0.38
	IOV	0.84±0.04	10.76±4.35	4.97±1.66	1.40±0.45
left parotid	A	0.77±0.04	13.89±5.36	5.84±1.64	1.84±0.54
	B	0.82±0.03	15.00±4.62	5.17±1.62	1.47±0.41
	C	0.83±0.03	12.13±3.91	4.63±1.21	1.35±0.40
	IOV	0.83±0.04	10.94±3.75	5.27±1.76	1.59±0.63
spinal cord	A	0.71±0.08	12.72±3.91	7.68±3.56	2.26±1.10
	B	0.80±0.05	10.12±4.83	4.26±1.36	1.24±0.45
	C	0.80±0.05	10.35±3.75	4.39±1.33	1.21±0.44
	IOV	0.79±0.07	7.12±5.15	4.64±3.06	1.55±0.81
mandible	A	0.64±0.09	16.65±3.60	6.96±1.84	2.14±0.60
	B	0.80±0.04	13.33±4.06	4.31±1.05	1.10±0.28
	C	0.80±0.04	10.88±2.07	4.44±1.09	1.35±0.30
	IOV	0.85±0.04	8.94±3.16	3.85±1.56	0.92±0.45

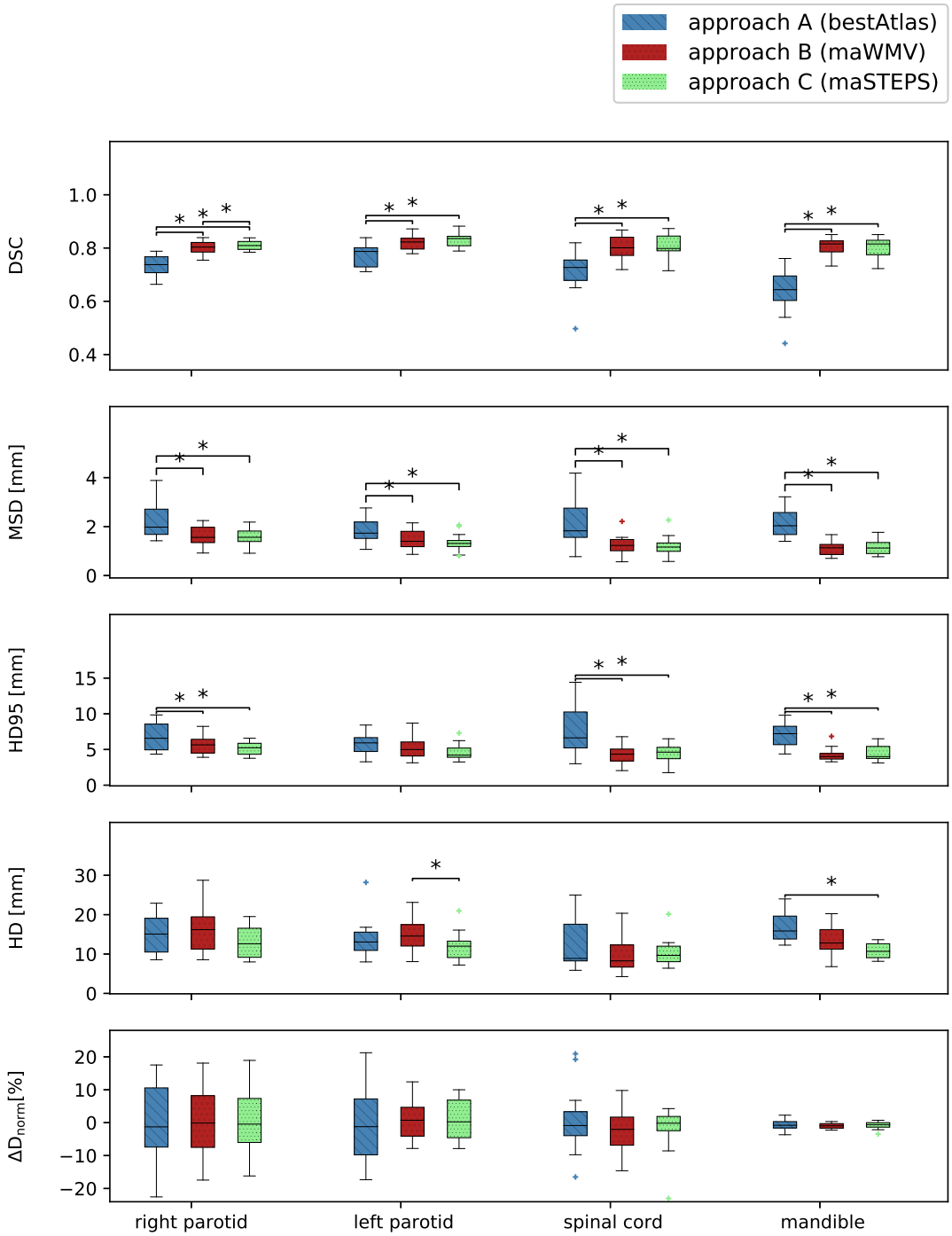


Figure 4.5: Boxplots of, from top to bottom, the DSC, MSD, HD95, HD and dosimetric difference ΔD_{norm} for all OAR (x-axis) and automated segmentation approaches (A in blue, B in red and C in green). The boxes indicate the interquartile range (IQR), the whiskers extend to the minimum and maximum values. Outliers are defined as data points beyond 1.5 IQRs from the IQR, denoted with a plus sign. Stars indicate statistical significance ($p<0.05/3$).

Table 4.2: Normalised dosimetric differences ΔD_{norm} (see equation (4.8)), as well as dosimetric variability (see section 4.2.6). Negative $\Delta D_{\text{x,norm}}$ mean that a larger dose would be delivered to the gold standard than what was planned for the auto-segmented ROIs. For a reference, the inter-observer variability (IOV) is included, where the values in column 4 need to be compared to the standard deviation of the dosimetric differences in column 3 (both in bold print).

ROI	method	$\overline{\Delta D_{\text{norm}}}$ [%]	IOV [%]
right parotid	A	0.06±**12.93**	
	B	-0.84±**10.82**	**5.56±4.78**
	C	0.02±**10.26**	
left parotid	A	-0.65±**11.39**	
	B	0.83±**6.51**	**6.00±3.93**
	C	0.68±**6.28**	
spinal cord	A	0.95±**10.68**	
	B	-2.77±**6.64**	**4.76±4.58**
	C	-2.17±**7.41**	
mandible	A	-0.66±**1.64**	
	B	-1.02±**0.85**	**0.46±0.26**
	C	-0.84±**1.18**	

of the clinical goal in either direction were observed for the parotids. Dose differences to the mandible were below 4 % of the clinical goal. The SDs of all mean dosimetric differences were within the range of the dosimetric variability (dosimetric inter-observer variability ± 1SD). However, the individual dosimetric difference in the parotids and the spinal cord was outside the range of the dosimetric variability for 50 % of the patients. In the mandible, this was the case in 75 %. Despite these substantial dose differences, all treatment planning objectives were still met for the manually segmented ROIs.

4.3.2 Geometric measures as predictors for dosimetric accuracy

Figure 4.6 depicts the absolute values of the dosimetric differences as a function of the three geometric measures (DSC, HD, HD95) for all ROIs and segmentation approaches. For a qualitative overall picture, I illustrate all approaches in the same subfigures. The correlation coefficients for each approach are included in each subfigure. Correlations between geometric and dosimetric measures were small with $R^2 < 0.5$ and did not have the expected sign in all cases, e. g. a negative correlation existed between the HD and $|\Delta D|$ for the left parotid, segmented using approach C.

Figure 4.7 highlights the pitfalls of performing a geometric-only evaluation. It illustrates three example pairs of cases with similar geometric accuracy yet large deviations between the dosimetric differences. The first two columns show a sagittal or axial image plane for two different patients. The coloured lines represent the isodose curves,

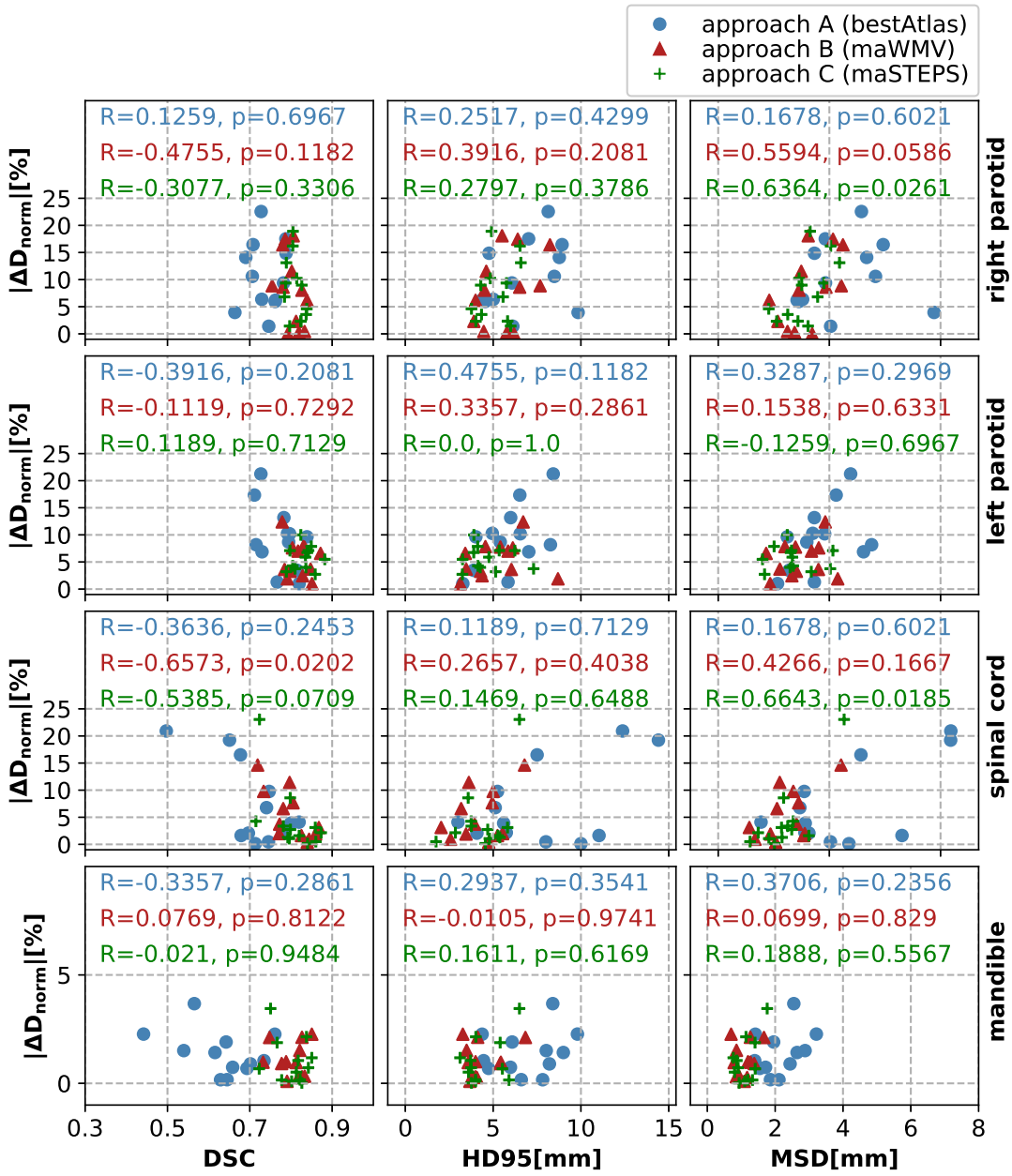
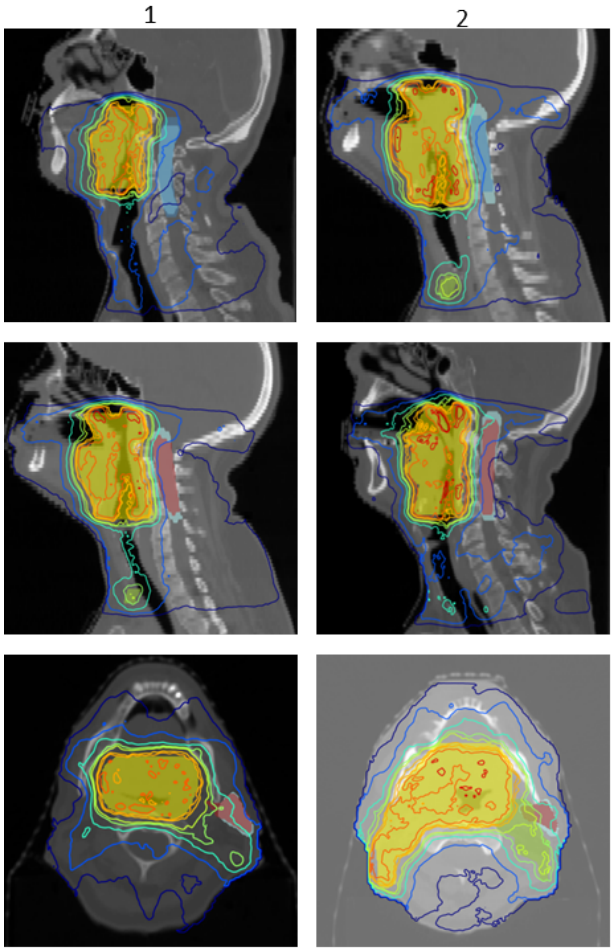


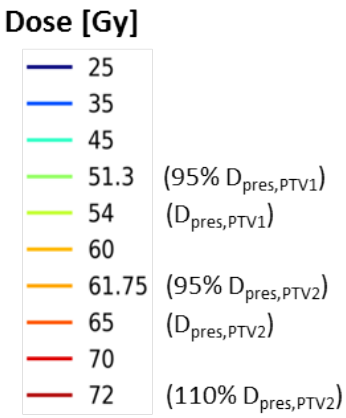
Figure 4.6: Scatter plots illustrating dose differences between manually and auto-segmented ROIs normalised to the clinical goal as a function of the respective geometric measures (from left to right: DSC, HD95 and MSD), separated according to the ROIs used in this study (from top to bottom: right parotid, left parotid, spinal cord and mandible). The different colours and symbols illustrate the three auto-segmentation methods of this study. The numbers in each subplot are the respective correlation coefficients R together with the p-values, calculated using Spearman's approach.



	1	2
DSC	0.74	0.75
HD [mm]	8.38	8.25
MSD [mm]	1.47	1.54
HD95 [mm]	5.11	5.25
ΔD [%]	6.77	-9.76

	1	2
DSC	0.80	0.80
HD [mm]	8.15	6.96
MSD [mm]	1.45	1.11
HD95 [mm]	4.95	3.62
ΔD [%]	-7.67	-11.42

	1	2
DSC	0.83	0.83
HD [mm]	12.5	12.5
MSD [mm]	1.20	1.39
HD95 [mm]	4.56	5.41
ΔD [%]	-7.84	7.75



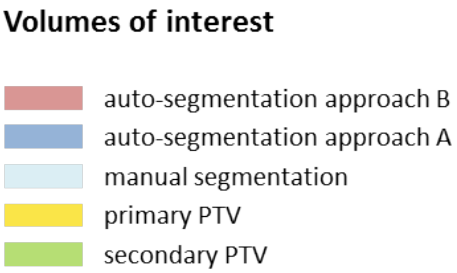


Figure 4.7: This figures illustrates three example cases where the geometric differences (DSC, HD, HD95 and MSD) were similar between the patients in columns 1 and 2 but the dosimetric impact differed. The first two rows illustrate examples for the spinal cord, the last row for the left parotid.

whereas the coloured areas show the manually and automatically segmented ROIs. The individual geometric and dosimetric differences between manually and automatically segmented ROIs are provided in the table in the third column. The first two rows illustrate examples for the spinal cord, where steep dose gradients have a tremendous influence due to the nature of the clinical goal (maximum dose). The last row shows an example for the parotid, where the relative position to the high dose region mainly impacts the dosimetric outcome. With the qualitative per-patient analysis, I found that more substantial dosimetric differences started to appear with the OAR being closer to the target volume. I additionally clustered the data as a function of the distance to the target volume and did not find any significant correlation.

4.4 Discussion

I developed a novel method to assess the dosimetric impact of segmentation errors automatically. Furthermore, I investigated the correlation between geometric and dosimetric differences. Dosimetric studies are essential in RT because segmentation errors can lead to an underdosage of target volumes, as well as an overdosage of OARs.

4.4.1 Geometric and dosimetric evaluation

A detailed geometric evaluation of atlas-based segmentation methods can be found in chapter 5. Several groups have studied the quantification of the impact of inaccurate localisations of ROIs on the planned dose distribution when using auto-generated contours in treatment planning. These can be summarised into three approaches.

The first approach is to use existing dose distributions on ground truth ROIs and superimpose these on the auto-segmented ROIs. The effect of contouring variations on dose parameters can then be determined by comparing dose differences to paired ground truth and auto-segmented ROIs. This method was applied by Eldesoky et al. [37] for the segmentation of breast tissues and by Conson et al. [25] for the segmentation of brain structures. A limitation of applying this method to the plan creation is that instead of generating new treatment plans for the automatically segmented ROIs, the original plans are used, thus ignoring the fact that different contours generate a different optimisation problem.

The second approach individually optimises the dose distributions for both, auto-segmented and ground truth ROIs, using the same beam parameters and planning constraints. Tsuji et al. [146] applied this approach for pairs of pre- and mid-treatment CTs of the head and neck region. A limitation of this method is that instead of comparing the direct dosimetric impact of contouring inaccuracies, two separately

generated treatment plans are compared. This means that rather than comparing the dosimetric impact, the feasibility of generating good quality treatment plans is compared.

The third approach is to create treatment plans for the auto-segmented sets of ROIs and superimpose the dose distributions to the ground truth ROIs. Nelms et al. [101] applied this approach to investigate the effects of inter-observer variabilities in manual OAR segmentations from 32 observers. A drawback of their study is that they only use the CT image of one patient for their evaluation. Voet et al. [152] applied the third approach to investigate whether geometric measures can predict the amount of underdosage in the PTV. Auto-segmented HNC ROIs edited by clinicians served as the ground truth. They included the neck levels and the parotids in their analysis. Beasley et al. [4] compared dosimetric differences and the geometric accuracy of auto-generated contours for the parotids and the larynx of 10 HNC patients, using the manually drawn contours of 5 observers as ground truth.

To properly account for the impact of segmentation inaccuracies on the planned dose distribution, the third approach has the smallest number of weaknesses. It solves the optimisation problem directly for the auto-segmented ROIs and, therefore, emulates the clinical reality in the case of an application to treatment plan generation. For this reason I chose this approach.

Both multi-atlas approaches outperformed the best-atlas approach in terms of the geometric accuracy (DSC, HD95 and MSD). This finding is in line with other published studies [26, 56, 143]. The HD was not a reliable measure for the geometric accuracy of the data used in this study. As this measure provides the maximum distance to the ground truth segmentations, it is susceptible to outliers and is hence not a good representative of the overall geometric accuracy.

In terms of the dosimetric accuracy, none of the three auto-segmentation approaches chosen in this work was superior to any other for any of the investigated OARs . Average absolute dose differences were below 3 % of the clinical goal for all OARs and segmentation approaches. However, dose differences for individual patients were widely spread with a standard deviation of up to 11 % of the mean. These broad ranges of dosimetric differences for individual patients are in line with published values. Beasley et al. [4] reported on an average difference in the mean dose to the parotids between auto-generated and ground truth ROIs, relative to the latter, of -4.8±3.4% with a range from -18% to 43%. They also compared mean doses for the larynx and found a difference of -8.4±2.3%, ranging from -20% to 3%. The inter-observer variability between 5 observers determined the uncertainty.

In contrast to these findings, there were no significant dose differences in the studies by Voet et al. [152] and Tsuji et al. [146]. Voet et al. [152] reported a small, statistically

non-significant dose difference for the parotids (-0.8±1.1 Gy, i. e. <3 %). For the target volume (CTV), they found that the mean reduction in dose to 99% of the volume (D_{99}) was considerable with 14.2 Gy (range of 1 to 54 Gy). Tsuji et al. [146] did not find any significant dose differences to the manually and automatically segmented OARs. However, instead of superimposing one treatment plan on both sets of ROIs for comparison, they generated individual treatment plans for each set of ROIs, therefore impairing a direct comparison.

4.4.2 Geometric measures as predictors for dosimetric accuracy

To understand whether the geometric measures used in this study (DSC, HD, HD95 and MSD) can be a reliable surrogate for dosimetric differences and treatment planning accuracy, I investigated the correlation between the geometric and dosimetric accuracy. If geometric measures were good predictors for the impact of segmentation inaccuracies on the dose distribution, one would expect large negative correlation coefficients R for the DSC and large positive R for the MSD, HD and HD95.

Voet et al. [152] showed that both DSC and mean contour distances did not have a large predictive value with respect to their influence on dose coverage of the target volume. They reported that an underdosage of 11 Gy might appear even for a decent geometric accuracy with DSC=0.8 and MSD<1 mm. Eldesoky et al. [37] investigated the relationship between geometric and dosimetric accuracy for four target volumes in breast cancer RT. They found a small significant correlation for only one of those target volumes between the DSC and dose-volume metrics.

In contrast to the studies mentioned above, I was focusing on OARs instead of target volumes. The results presented in Figure 4.6, illustrating the relation between geometric and dosimetric measures, did not imply a strong correlation between geometric and dosimetric measures. This finding was also reflected in the small correlation coefficients. All patients in this study had a tumour at the base of the tongue. For this reason, the relative positions of OARs and target volumes were similar. Despite this similarity, the relation between dose deposition and between the location of target volumes remained to be very complex. The inspection of individual patient images revealed that the impact of geometric inaccuracies on dosimetric outcome was influenced considerably by the shape of the structure, the type of clinical goal (maximum or mean dose) and the location of geometric differences (i. e. whether these lie within regions of high dose gradients or are far from those). Examples of high dose gradients influencing the correlation between geometric and dosimetric measures are shown in the first two example cases in figure 4.7.

These findings suggest that for the data used in this study, the investigated geometric measures are not reliable surrogates for the dosimetric outcome. The correlation values

for the DSC are in line with results reported by Beasley et al. [4]. Additionally, they found a substantial correlation (R=0.83) between the centroid distance and the differences in the mean dose to the parotids. However, evaluating this for the data in this study, There was not such a strong correlation. Furthermore, correlations with the distance-related measures were smaller compared to Beasley et al. [4].

While the SD of dosimetric differences for the full patient cohort was within the range of the dosimetric inter-observer variability, I found that for individual patients, the dosimetric difference was outside this variability despite a decent geometric accuracy. This finding highlights the need to investigate the dosimetric impact of contouring inaccuracies carefully. In this study, I developed an approach on how to accomplish this and when using an auto-segmentation approach in the clinic, it is crucial to ensure an adequate dosimetric accuracy.

4.4.3 Limitations and future work

One limitation of this study was the relatively small number of available training data. However, even with this small dataset, it could be shown that a decent geometric accuracy does not guarantee small dosimetric errors. Therefore, when using auto-segmentation algorithms in the clinic, a thorough dosimetric evaluation is crucial.

Furthermore, due to the small imaging coverage of the patients' anatomies in the superior-inferior direction, I could only include four OARs in the analysis. However, even though treatment planning of HNC requires the segmentation of more OARs such as the optical structures and the brainstem, the OARs in this analysis covered a variety of shapes and locations relative to the target.

Dose calculations in this study were performed simulating a 9-beam step and shoot IMRT treatment on an MR-linac in a magnetic field. While other radiation delivery techniques may lead to slightly different dosimetric results, the dosimetric evaluation method is independent of the treatment type and can be easily applied to more patient data. The template approach established in this study worked well for all included patients. One would anticipate some necessary changes to the template for very different anatomies compared to the patient data in this study.

The capability to estimate the dosimetric effect from the geometric evaluation directly would remove the need to optimise treatment plans for each set of auto-segmented ROIs. On the other hand, using geometric measures that do not reliably predict the impact on the dose distribution limits their applicability in a clinical validation for RT. Future work would investigate new measures than can more reliably predict the dosimetric effect of segmentation inaccuracies. These could, for instance, incorporate the distance to the target volumes, or, more generally, to high dose-gradient regions. Yang et al. [163],

for example, use the overlap volume histogram to quantify the distance between the rectum and prostate PTV to predict possible dose distributions. Furthermore, the first applications of machine learning approaches in RT seem promising and could be applied to this problem by, for example, modelling geometric uncertainties using neural networks and determining the effect on dose distributions. These approaches were outside the scope of this thesis.

4.5 Conclusion

To my knowledge, this was the first study to investigate the use of contours derived from atlas-based segmentation on HNC MR images in the context of treatment plan generation for RT with a complete analysis of the geometric and dosimetric accuracy. The inter-observer variability, determined for the imaging data used in this study, served as a benchmark on the achievable accuracy.

Since there appeared to be only a slight correlation between geometric (DSC, MSD and HD95) and dosimetric measures, the geometric measures alone were not sufficient to predict the dosimetric impact of segmentation inaccuracies on RT treatment plans. When performing exploratory research on auto-segmentation methods, geometric measures can provide an estimate of the performance in terms of accuracy to compare it to other auto-segmentation approaches. However, for a safe clinical implementation, it is crucial also to investigate the dosimetric impact of segmentation inaccuracies.

Chapter 5

Automated segmentation with atlas-based methods

Atlas-based segmentation methods are the most commonly used auto-segmentation approaches in radiotherapy. This chapter investigates different atlas-based segmentation methods of organs at risk in the head and neck region. I benchmark the developed approaches by comparing them to a commercially available atlas-based solution and the inter-observer variability.

This chapter is an extension of the following publication:

J P Kieselmann et al, Geometric and dosimetric evaluations of atlas-based segmentation methods of MR images in the head and neck region, Physics in Medicine and Biology (2018) 63 145007.

5.1 Introduction

To date, atlas-based methods are the most commonly used auto-segmentation approaches in RT [43]. The labelled anatomy of an image is used as a source of positional, topological and shape information about the ROIs in the form of an atlas. An atlas is an image together with its corresponding labels. The contours from a library of atlases are warped to the previously unseen image using image registration methods. One then needs to select an atlas or merge multiple atlases to obtain the final labels for the unseen image. Provided enough data is available, atlases can capture anatomical variations that naturally exist between patients.

Atlas-based segmentation methods in the literature differ by the type of image registration and atlas fusion or selection they apply. The auto-segmentation result can be derived by a single chosen atlas or by a combination of multiple atlases. Atlas-based segmentation has been implemented into many treatment planning systems, such as RayStation (Raysearch, Stockholm, Sweden) and Monaco (Elekta AB, Stockholm, Sweden).

This chapter describes a study on the performance of atlas-based segmentation in terms of its geometrical accuracy and computation time, applied to the segmentation of the parotids, the spinal cord and the mandible on MR images of 27 HNC patients. I chose the NiftyReg [98, 99], and NiftySeg [14, 147] software tools for this study, which were both developed at the University College London (UCL). I was kindly provided with the source code and extended the tools, automated parts of the segmentation process and developed a fully automated validation workflow (see chapter 4). Although I have shown in chapter 4 that calculating the dosimetric effect of auto-segmentation is a more accurate assessment of its quality than purely geometric measures, there was not enough time to do this within the scope of this thesis and it is therefore left for future work.

5.2 Materials and Methods

5.2.1 Data acquisition and preparation

The atlas database comprised 27 T2w pre-treatment MR images, introduced in figure 3.1 on page 39. The inter-observer variability was obtained from [72] (see the previous chapter).

5.2.2 Image registration: Basic concepts

Image registration is an essential component of atlas-based segmentation methods. This section thus introduces the basic concepts in the field of image registration, as well as an overview of commonly used algorithms. Extensive surveys on medical image registration can, for example, be found in [59, 93, 135].

Image registration is the process of finding a geometrical transformation $\mathcal{T}$ that spatially aligns points x in one image to points x' in another image:

$$\mathbf{x} \xrightarrow{\mathcal{T}} \mathbf{x}'. \tag{5.1}$$

Typically, one image is defined as the fixed image I_F, whereas the other image, the moving image I_M, is transformed to match the fixed image as closely as possible.

There are two main types of image registration: feature- or shape-based [9] and intensity-based methods [74, 129, 144]. Feature-based methods register images by identifying and matching features or objects that describe distinctive landmarks, edges or shapes. These features are often difficult to obtain. Intensity-based methods, on the other hand, are primarily based on a correlation between the intensity values of the pixels or voxels of two images. In this study, I discuss and apply only parametric intensity-based image registration methods.

Intensity-based image registration is typically formulated as an optimisation problem, aiming to minimise a cost function $C(\mathcal{T}, I_F, I_M)$ that measures the similarity between the two images. In an iterative process ,optimal transformation parameters $\hat{\Theta}$ are sought to maximise the similarity:

$$\hat{\Theta} = \arg\min_{\Theta} C(\mathcal{T}_\Theta, I_F, I_M). \tag{5.2}$$

Figure 5.1 illustrates the basic concepts of image registration for two example patients (two-dimensional (2D) illustration for ease of display). Intensity-based image registration is characterised by three main elements:

(1) the transformation type

(2) the cost function

(3) the optimisation method

which are described in the following sections.

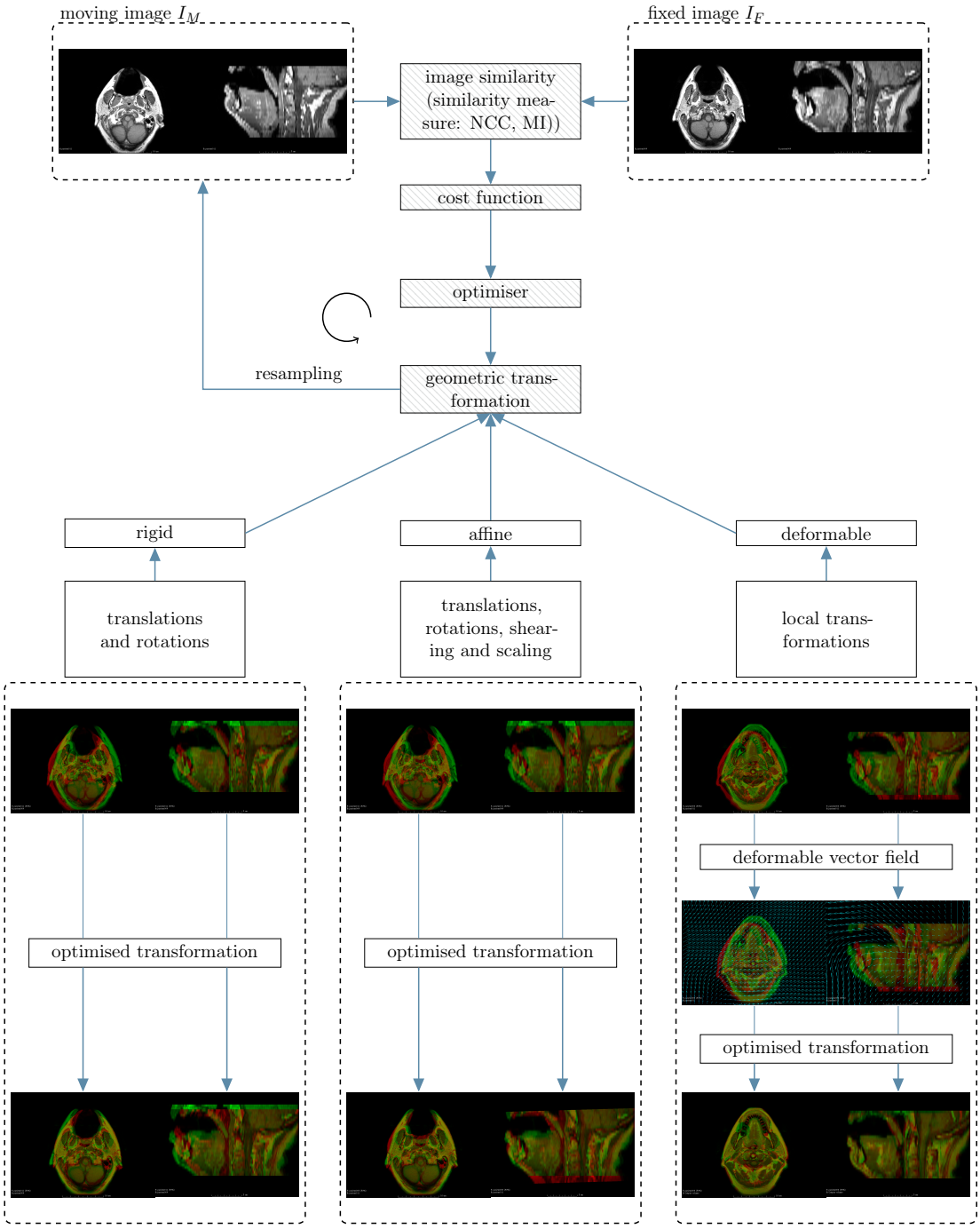


Figure 5.1: Image registration approaches: This figure illustrates the basics of image registration with the example of two MR images of two patients. The three main transformation types (rigid, affine and deformable) are illustrated in the bottom part of the figure. For illustration purposes, the images are displayed in 2 dimensions, whereas the transformations were performed in 3 dimensions.

5.2.2.1 Transformation type

The transformation type can be either rigid (global translations and rotations), affine (rigid and shearing or scaling operations) or deformable (local transformations). Medical image registration typically combines multiple registration methods. In the following, without loss of generalisation, I use the example of 3D images.

Rigid transformation

A rigid transformation of 3D images has 6 degrees of freedom: 3 rotational and 3 translational ones. In homogeneous coordinates[1], the transformation can be parametrised by the transformation matrix $\mathcal{T} = T \circ R$ with translation matrix T and rotation matrix R defined as follows:

$$T = \begin{pmatrix} 1 & 0 & 0 & t_x \\ 0 & 1 & 0 & t_y \\ 0 & 0 & 1 & t_z \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

where $t_i \in \mathbb{R}$,

$$R_x = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & c_x & s_x & 0 \\ 0 & -s_x & c_x & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, R_y = \begin{pmatrix} c_y & 0 & s_y & 0 \\ 0 & 1 & 0 & 0 \\ -s_y & 0 & c_y & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, R_z = \begin{pmatrix} c_z & s_z & 0 & 0 \\ -s_z & c_z & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{5.3}$$

where $c_i = \cos(\vartheta_i)$ and $s_i = \sin(\vartheta_i)$

$t_i, c_i, \vartheta_i \in \mathbb{R}, i \in \{x, y, z\}$.

Affine transformation

In addition to translations T and rotations R, an affine transformation can consist of scaling S and shearing Q operations:

$$S = \begin{pmatrix} a_x & 0 & 0 & 0 \\ 0 & a_y & 0 & 0 \\ 0 & 0 & a_z & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

$$Q = \begin{pmatrix} 1 & a_{xy} & a_{xz} & 0 \\ a_{yx} & 1 & a_{yz} & 0 \\ a_{zx} & a_{zy} & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{5.4}$$

It has 6 additional degrees of freedom a_i, where $a_i \in \mathbb{R}$. The rigid transformation is a special case of an affine transformation.

[1] Homogeneous coordinates replace a 3D vector $(x, y, z)^T$ with the 4D vector $(x, y, z, 1)^T$.

Deformable transformation

A deformable transformation acts locally, whereas affine transformations act globally. Deformable image registration is usually described by a deformation vector field, where for each voxel in the image, a deformation vector is calculated, also named free-form deformation approach [129]. The degrees of freedom can be as large as three times the number of voxels in the image. As it is too computationally expensive to calculate deformation vectors at each voxel, a common approach is to overlay a grid of so-called control position points (CPPs) $p_{a,b,c}$, with $(a, b, c) \in (\{1, \cdots, n_x\}, \{1, \cdots, n_y\}, \{1, \cdots, n_z\})$, that are more sparsely spaced than the voxel positions. Cubic B-splines can then be used to interpolate the deformable transformation $T(\vec{x})$ at each position $\vec{x}$ from the CPPs [129].

To limit the deformations to physically plausible transformations, penalty terms can be incorporated into the cost function to apply regularisation. Examples of these are described in the following paragraphs.

5.2.2.2 Cost function: Similarity measures and penalty terms

To find the optimal correspondence between two images, several similarity measures can be applied. The two most popular ones in medical image registration are the (normalised) cross-correlation (NCC) for mono-modal images and the mutual information (MI) for multi-modal images. The NCC is similar to the convolution operation and is defined as follows:

$$\text{NCC} = \frac{1}{N} \sum_{i \in \mathbb{N}} \frac{(I_F(i) - \overline{I_F})(I_M \circ \mathcal{T}(i) - \overline{I_M})}{\sigma_{I_F} \sigma_{I_M}} \tag{5.5}$$

with

N : number of voxels

$I_F(i), I_M(i) \in \mathbb{R}$, intensity at voxel i in fixed and moving image

$\overline{I_F}, \overline{I_M}$: mean intensities

$\sigma_{I_F}, \sigma_{I_M}$: standard deviation of mean intensities

It multiplies the mean-subtracted intensities at each voxel of two images and averages them over all voxels, leading to large values when the mean-subtracted images are similar and to values close to zero for two random, non-similar images. It works well if the intensities in images I_F and I_M can be linearly related, i. e. for images with a similar

contrast. The MI is defined as

$$\mathrm{MI} = \sum_{i\in\mathbb{N}} \sum_{j\in\mathbb{N}} p_{FM}(I_F(i), I_M \circ \mathcal{T}(j)) \log \frac{p_{FM}(I_F(i), I_M \circ \mathcal{T}(j))}{p_F(I_F(i)) p_M(I_M \circ \mathcal{T}(j)} \tag{5.6}$$

with

p_{FM} : joint probably distribution

p_F, p_M : individual probability distributions.

The MI assumes no functional relationship between the images. It is maximised when there is a consistent relationship between voxels in one image and corresponding voxels in the other image. Mathematically, if the two images are completely independent, then

$$p_{FM}(I_F(i), I_M \circ \mathcal{T}(j)) = p_F(I_F(i)) \cdot p_M(I_M \circ \mathcal{T}(j)) \tag{5.7}$$

and MI=0. However, if the two images are fully dependent on each other (i. e. perfectly aligned), then

$$p_{FM}(I_F(i), I_M \circ \mathcal{T}(j)) = p_F(I_F(i)) = p_M(I_M \circ \mathcal{T}(j)). \tag{5.8}$$

The MI is well suited for applications to multi-modal images, where no linear relationship between voxels exists.

The cost function to be optimised consists of the similarity measure SM and the penalty terms P_i:

$$C = (1-\alpha) \cdot \mathrm{SM}(I_F, I_M \circ \mathcal{T}) + \alpha \sum_i w_i \cdot P_i(\mathcal{T}), \tag{5.9}$$

where α balances the influence of the similarity measure against the penalty terms and w_i are the individual weights for the penalty terms.

As image registration is generally an ill-posed problem, penalty terms are used for regularisation purposes. These terms allow for the inclusion of prior knowledge of the physical properties of the underlying deformations. Typical examples for penalty terms in deformable image registration are the following:

- the bending energy, which is the sum of the second order derivatives of the transformation field M, to ensure smoothness [129]

$$C_{\mathrm{smooth}} = \frac{1}{N} \sum_{(x,y,z)\in\mathbb{R}^3} \left[\left(\frac{\partial^2 \mathcal{T}(\vec{x})}{\partial x^2} \right)^2 + \left(\frac{\partial^2 \mathcal{T}(\vec{x})}{\partial y^2} \right)^2 + \left(\frac{\partial^2 \mathcal{T}(\vec{x})}{\partial z^2} \right)^2 +$$

$$2\left(\frac{\partial^2\mathcal{T}(\vec{x})}{\partial x\partial y}\right)^2 + 2\left(\frac{\partial^2\mathcal{T}(\vec{x})}{\partial x\partial z}\right)^2 + 2\left(\frac{\partial^2\mathcal{T}(\vec{x})}{\partial y\partial z}\right)^2\Bigg]. \tag{5.10}$$

By penalising large second derivatives, i. e. high curvature, this means for example that high compressions close to nearby high expansions are penalised.

- A Jacobian based term to ensure that no folding occurs

$$C_{\text{folding}} = \log(|J(\mathcal{T}(\vec{x}))|)$$

with the Jacobian matrix $\mathcal{J}$

$$\mathcal{J}(\mathcal{T}(\vec{x}) = \left(\begin{pmatrix} \frac{\partial\mathcal{T}(\vec{x})_x}{\partial x} & \frac{\partial\mathcal{T}(\vec{x})_y}{\partial x} & \frac{\partial\mathcal{T}(\vec{x})_z}{\partial x} \\ \frac{\partial\mathcal{T}(\vec{x})_x}{\partial y} & \frac{\partial\mathcal{T}(\vec{x})_y}{\partial y} & \frac{\partial\mathcal{T}(\vec{x})_z}{\partial y} \\ \frac{\partial\mathcal{T}(\vec{x})_x}{\partial z} & \frac{\partial\mathcal{T}(\vec{x})_y}{\partial z} & \frac{\partial\mathcal{T}(\vec{x})_z}{\partial z} \end{pmatrix}\right). \tag{5.11}$$

Folding means a cross-over between lines in the deformation grid, that means the penalty prevents foldings of structures onto themselves. A Jacobian determinant larger than zero ensures that the deformation vector field is invertible. Jacobian determinants larger than 1 mean a volume increase after registration and vice versa.

5.2.2.3 Optimisation methods

The optimisation problem in equation (5.2) on page 66 can be formulated as an iterative process. With the current parameters Θ_i, step size α_i (c. f. the learning rate in chapter 6) and an update or search direction u_i, the parameters in iteration $(i+1)$ are determined as follows:

$$\Theta_{i+1} = \Theta_i + \alpha_i \cdot u_i. \tag{5.12}$$

This process is repeated until a predefined convergence point.

The optimisation parameters α_i and u_i depend on the optimisation type. There is a range of optimisation strategies applied in medical image registration, the most prominent ones being gradient descent, conjugate gradient descent and Quasi-Newton methods [135]. The following paragraphs briefly introduce the basic concepts in optimisation.

Step size α_i

The optimisation parameter α_i is often set to a constant or a decaying function. The constant needs to be adapted to the underlying problem and is not always straightforward

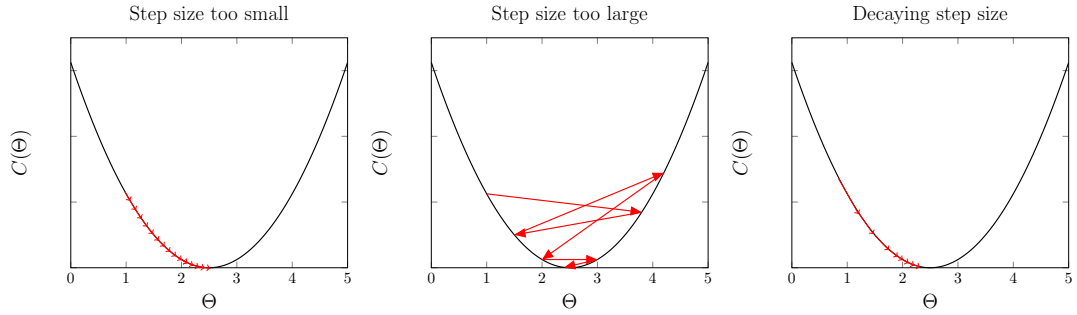


Figure 5.2: This figure illustrates the iterative optimisation process of minimising the cost function C as a function of the parameters Θ. For simplicity, I chose a convex, one-dimensional function. The right arrows depict the step taking from one iteration to the next. The left and the central graph illustrate the effect of too small (left) and too large (centre) step sizes on the optimisation process. Using too small step sizes can lead to long execution times as the optimiser only slowly converges to the optimum. Too large step sizes can lead to large fluctuations in the cost function and overshooting over the optimum. An example of a decaying step size over the number of iterations is illustrated in the right graph, where a larger step size is used in earlier iterations and smaller step sizes in later iterations when the optimum is approached.

to determine. If $\alpha_i = \alpha$ is too large then the optimiser might jump over minima and convergence may not be achieved. If α is set too small, convergence may be very slow and the optimiser may be stuck in local minima. A decaying function for α_i can be justified by the fact that it may be beneficial to reach close to an optimal solution at the beginning with larger steps, while slowly reaching the optimum towards the end with small steps. Figure 5.2 illustrates the effect of too small, too large and decaying step sizes on the optimisation process. For simplicity, I chose a convex, one-dimensional example for the cost function $C(\Theta)$ using a gradient descent method to obtain the search direction u_i. In practice, the cost function is typically highly non-convex.

Gradient descent

In the gradient descent method [75], u_i is the gradient of the cost function C with respect to its parameters Θ, evaluated using the current parameters Θ_i:

$$u_i = \frac{\partial C}{\partial \Theta}\bigg|_{\Theta_i}. \tag{5.13}$$

Gradient descent is usually combined with a constant or decaying step size α.

Quasi-Newton

As gradient descent methods can take long times to converge to an optimum, other methods can be used to speed up this process. For example, information on second-order derivatives of the cost function can be used (defined through the Hessian matrix, a

matrix of all second order derivatives with respect to its parameters). As this can be computationally expensive, the Quasi-Newton approach uses an approximation of the inverse Hessian matrix, L_i [75]. The parameters Θ_i are updated as follows:

$$\Theta_{i+1} = \Theta_i + \alpha_i \cdot L_i \cdot \frac{\partial C}{\partial \Theta}\bigg|_{\Theta_i}. \tag{5.14}$$

Conjugate gradient descent

The conjugate gradient does not require the calculation of second order derivatives but instead uses the previous search direction as an additional term to update in the next iteration:

$$u_{i+1} = \frac{\partial C}{\partial \Theta}\bigg|_{\Theta_i} + u_i. \tag{5.15}$$

This method has been shown to converge faster than the gradient descent method [75].

Multi-resolution approach

Due to image registration being an ill-posed problem with many possible solutions, a standard approach in optimisation is to use a multi-resolution approach [88]. For this purpose, the registration is split into several steps, starting with a coarse CPP grid to model coarse deformations and subsequently refining the deformation field, registering smaller structures, by using finer CPP grids. Multi-resolution can speed up convergence, reduce the number of examined transformations and avoid local minima. This is also called a hierarchical or pyramidal approach.

Block-matching approach

Commonly in rigid and affine registration approaches, one assumes a global relationship between the intensities in the fixed and the moving image, defines similarity as described above and finds the optimal transformation parameters according to equation (5.12). A problem with this approach is that a global optimisation is in general not straightforward due to numerous local minima of the cost function, as well as imaging artefacts that remove a global relationship between image intensities. To overcome these problems, several groups used a block-matching based approach [99, 111]. After dividing both images into equally sized blocks, this block-matching algorithm consists of two steps:

(1) finding correspondences between blocks of the two images to register

(2) extracting parameters from these correspondences for global transformation (rigid: rotation and translations; affine: rotation, translation, scaling and shearing).

For step (1) each block a_i is compared to blocks b_m in a pre-defined neighbourhood and the most similar block b_i in terms of some pre-defined similarity measure is chosen as the corresponding block. In step (2), the optimal global transformation parameters $\hat{\Theta}$ can be estimated by a least square fit of the residuals $r_i = a_i - T(b_i)$:

$$\hat{\Theta} = \arg\min_{\Theta} \sum_{i}^{N} ||r_i||^2, \tag{5.16}$$

with N being the total number of blocks. However, as this is sensitive to outliers, Modat et al. [99] propose to use a least-trimmed-square method instead:

$$\hat{\Theta} = \arg\min_{\Theta} \sum_{i}^{k} ||r_{i:N}||^2 = \arg\min_{\Theta} \sum_{i}^{p \cdot N} ||r_{i:N}||^2 \tag{5.17}$$

where $r_{i:N}$ are the residuals, sorted according to magnitude and $k = p \cdot N$ with p denoting the fraction of total blocks to be used.

Initialisation for deformable registration approaches:

Due to the ill-posed nature, in particular of deformable registration, it often needs a good initial position to converge to an optimum. For this purpose, a common approach is to initialise deformable registrations with a global registration (rigid or affine) and refine the transformation by a subsequent deformable registration [93].

5.2.3 Image registration software

5.2.3.1 Image registration in NiftyReg

I performed the registration using UCL's software tool NiftyReg. The choice of parameters in the following was justified by a combination of previous work in this area and an exploratory approach for the data used in this study. Table 5.1 lists the chosen parameters.

5.2.3.2 Image registration in RayStation

Since RayStation is a commercial TPS, I did not have access to the source code. There is no option to adapt parameters in RayStation as it provides a general-purpose solution. The registration is initialised with a rigid registration. The deformable registration is based on a hybrid solution between shape- and intensity-based registration called ANACONDA (ANAtomically CONstrained Deformation Algorithm). The algorithm is described in detail in Weistrand and Svensson [157]. The cost function incorporates

Table 5.1: Image registration parameters used for NiftyReg in this study.

	affine	**deformable**	**comments**
step	1	2	
algorithm	block-matching	free-form deformation	
# **resolution levels**	2	3	
resolution [mm^3]			registration with original resolution did not improve accuracy
(1)	2x2x4	4x4x4	
(2)	1x1x4	2x2x4	
(3)	-	1x1x4	
CPP/block size			changed default to real-world instead of voxel coordinates due to anisotropy
(1)	4x4x4 voxels	10x10x10 mm^3	
(2)	4x4x4 voxels	5x5x5 mm^3	
(3)	4x4x4 voxels	2.5x2.5x2.5 mm^3	
similarity measures	NCC	NCC	most suitable metric in previous work for registering images of the same modality [13]
penalty terms	-	bending energy (weight 0.005)	
	-	Jacobean determinant (weight 0.0001)	
optimiser	conjugate gradient descent	conjugate gradient descent	
other parameters	$p = 80\%$ (eq.(5.17))	$\alpha = 0.0051$ (eq.(5.9))	p as recommended in the original paper [98]

the NCC as similarity measure and a penalty function to ensure smoothness and avoid folding. Furthermore, it incorporates a constraint based on the shape of pre-defined controlling ROIs.

5.2.4 Automated segmentation

The automated segmentation is based on the software tool NiftySeg. To benchmark the auto-segmentation algorithms, I also looked at atlas-based segmentation implemented in the commercially available TPS RayStation.

In theory, an atlas-based approach only requires one reference image. However, due to substantial anatomical variations between individual patients, it has proven to be beneficial to use multiple reference images in the creation of the atlas database [56, 143].

In the following, I define an *atlas* as an MR image, paired with segmented ROIs. I call the previously unseen MR image the *target image*. Atlas-based segmentation consists of two major steps:

(1) image registration of all library images to the target image

(2) atlas selection or fusion of individual segmentation results from each atlas to a joint segmentation of the target image.

Image registration algorithms were discussed in the previous sections. The following section provides an overview of atlas selection and fusion methods.

5.2.4.1 Atlas selection and fusion

After the registration of all library images to the target image, one can use different approaches to merge the information obtained from each library image into one segmentation result. I used NiftySeg and developed automated scripts to compare three atlas selection and fusion approaches to obtain the final segmentation result. In all three approaches, I determined the similarity between two images by calculating the NCC coefficient. Figure 5.3 illustrates the roadmap to atlas-based segmentation and, in particular, shows the three different approaches used in this work, namely the best atlas approach (approach A), a weighted majority voting approach (approach B) and an approach called STEPS (approach C).

Approach A: best atlas

In the best atlas approach (approach A), the library image that was most similar to the target image was selected. The auto-segmented ROIs were obtained by warping the ROIs from the library image to the target image, using the DVF from the image registration.

Approach B: weighted majority voting

In this approach, the labels of the registered library images were combined into a single label with a weighted majority voting on a voxel-by-voxel basis. The weights were derived locally from the similarity between library and target image [13]. In this context, locally was defined as the application of a Gaussian kernel with a standard deviation (SD) of 2.5 voxels around each voxel. I call this the multi-atlas weighted majority voting (maWMV) approach.

Approach C: STEPS

Approach C was the multi-atlas Similarity and Truth Estimation for Propagated Segmentations (maSTEPS) [15], which is closely related to the well-established STAPLE method [155]. STEPS mainly consists of seven steps:

(1) All library images are registered to the target image.

(2) For each voxel, the n library images which locally are most similar to the target image are chosen, where local is defined as in approach B.

(3) An initial ground truth estimation of the ROIs is determined using a majority voting approach.

(4) The sensitivity and specificity with respect to the initial segmentation in (3) are determined for the chosen atlases.

(5) The ground truth estimation of the ROIs is updated with a maximum likelihood estimation using the sensitivity and specificity of the individual atlases as parameters.

(6) In the ground truth estimation, an MRF is used to enforce locally connected regions. If the strength of the MRF is too large, small details can be lost, if it is too small, isolated segmented regions can occur.

(7) If a pre-set fraction of atlases agrees on a label, this voxel is declared as solved and removed from the ground truth estimation.

(8) Steps (4) to (7) are repeated until convergence.

I chose n=15 for step (2), a fraction of 95 % for step (7) and set the strength of the MRF to 1, following the parameters recommended in [15].

Atlas-based segmentation in RayStation:

RayStation includes the option to create a library of images and perform an atlas-based segmentation on unsegmented images. After registering all atlas images to the target image, the N best atlases are used in an atlas fusion method to segment the target image. The user can set N. The details on the exact approach to atlas fusion are unclear, RayStation only states in their user manual that the fusion is done in an iterative process, initialised with a weighted majority voting. I selected N as the total number of atlases available.

5.2.5 Computation time

I determined computation times for programme execution on an Intel® Xeon® CPU E5-1660v3 (3GHz) processor by averaging the time over multiple runs. A programme execution included the image extraction time, image registrations between the target image and individual atlases from the library, as well as the atlas selection or fusion to obtain the final segmented ROIs for the target image.

5.2.6 Evaluation

To estimate and compare the performance of auto-segmentation algorithms, I followed a 9-fold cross-validation strategy. For each fold, I removed three distinct images from the library to be used as test images. The remaining 24 images compromised the library of atlases.

5.2.6.1 Geometrical evaluation

As the first indication of agreement, I calculated the volume of each auto-segmented ROI, averaged over all patients and compared to the volume of the manually segmented gold standard ROIs. Furthermore, I calculated four well-established geometric measures between the auto-segmented and the gold standard ROIs: the DSC [29] for volumetric differences, as well as the standard HD and the HD95 and the MSD [115] for distance related differences. Details on geometric measures as well as the evaluation of auto-segmentation methods can be found in chapter 4. I benchmarked the approaches by comparing it to the inter-observer variability, as well as the commercial algorithm in RayStation.

5.2.6.2 Statistical evaluation

The statistical analysis was performed using Student's t-test, as described in the previous chapter 4.

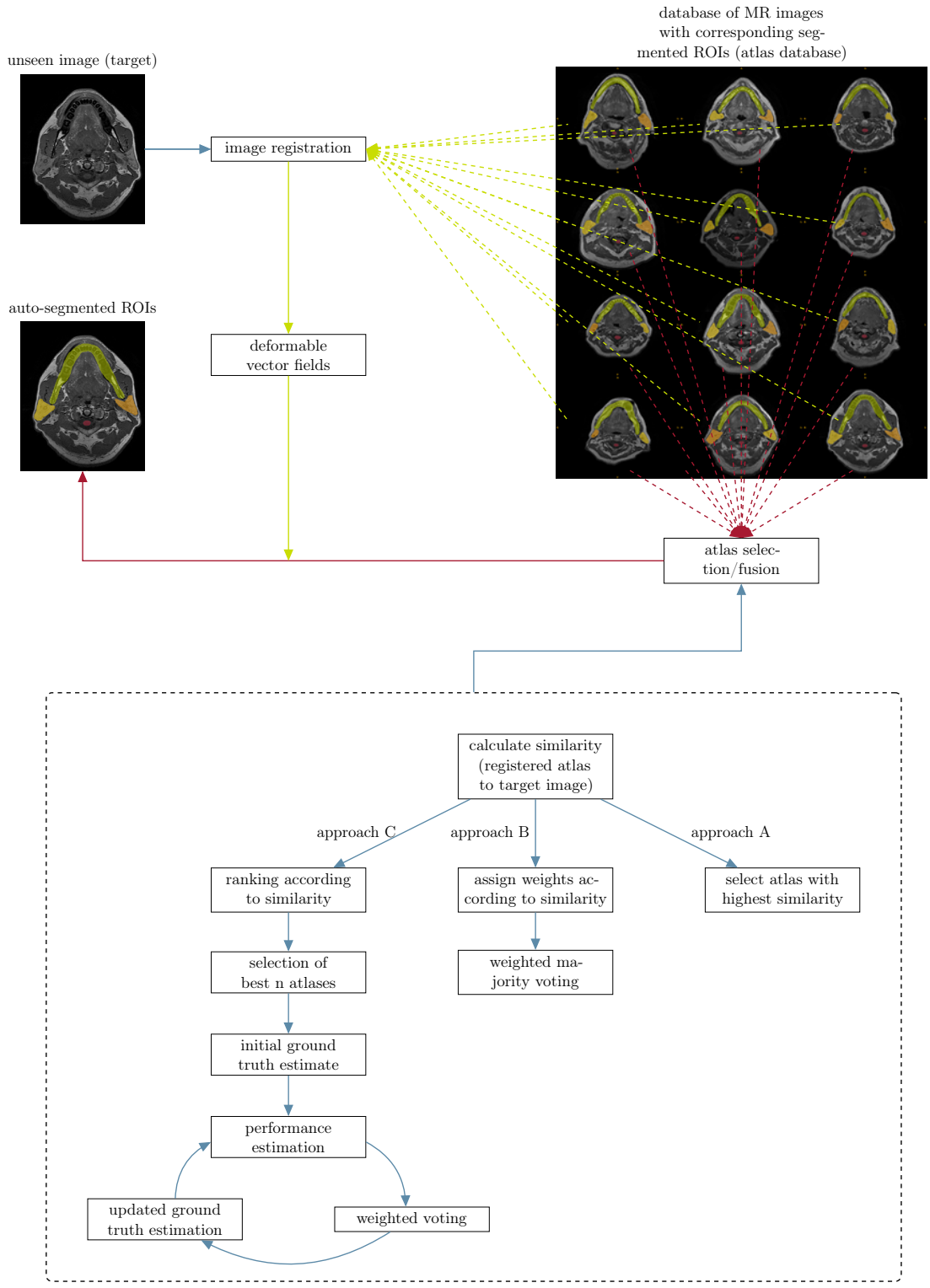


Figure 5.3: This figure shows a flowchart of the atlas-based segmentation approaches used in this work. The top part shows the general concept of auto-segmentation with the image registration of the unseen image to all atlas images, and, using the deformable vector fields from this process, the prediction of the segmented ROIs through atlas fusion or selection. The bottom part details the three atlas selection and fusion approaches used in this work.

5.3 Results

5.3.1 Computation time

The bottleneck of atlas-based segmentation concerning computation time is the image registration. One image registration took 5 minutes on average. As multiple image registrations were performed sequentially, this led to a total computation time of under 2 hours (23 registrations at 5 minutes each). The time attributed to selecting the best atlas in approach A did not add any significant time. The atlas fusion for approaches B and C added less than a minute in total. In RayStation, the full process took approximately 2 minutes.

5.3.2 Geometric evaluation

Figure 5.4 provides four typical examples from four different patients for a qualitative comparison of all three auto-segmentation approaches, as well as the commercial approach in RayStation, to the gold standard. The three multi-atlas approaches (rows 2, 3 and 4) clearly outperformed the best-atlas approach (first row) in all shown cases. Visually, RayStation leads to a comparable performance as the two multi-atlas approaches.

As the first indication of agreement, I calculated the volume of the automatically and manually segmented ROIs, averaged over all patients. Table 5.2 lists the mean volume, as well as the SD for all ROIs and segmentation approaches.

Table 5.2: Automatically segmented mean volumes with standard deviations for all approaches and ROIs with comparisons to manually segmented (gold standard) volumes.

ROI	manually segmented volume [cm^3]	approach	auto-segmented volume [cm^3]
right parotid	42.42±14.66	A (best atlas)	41.92±12.21
		B (maWMV)	39.92±9.29
		C (maSTEPS)	49.92±10.50
		D (RayStation)	41.91±7.41
left parotid	42.24±13.46	A (best atlas)	42.36±11.50
		B (maWMV)	40.21±9.23
		C (maSTEPS)	50.36±10.85
		D (RayStation)	43.39±7.78
spinal cord	6.21±1.54	A (best atlas)	6.62±1.46
		B (maWMV)	6.00±1.22
		C (maSTEPS)	10.33±1.83
		D (RayStation)	5.50±0.98
mandible	55.93±12.62	A (best atlas)	49.35±11.11
		B (maWMV)	50.80±11.98
		C (maSTEPS)	60.12±11.75
		D (RayStation)	55.95±9.08

Table 5.3: Geometric evaluation for all ROIs and auto-segmentation approaches: mean values for DSC, HD and MSD. All mean values have been calculated by averaging over all 27 patients. For a reference, I also include the inter-observer variability (IOV), derived from the manual contours of three different experts, and compare to the approach in the commercial system Raystation (RS).

ROI	method	$\overline{DSC}$	$\overline{HD}$ [mm]	$\overline{MSD}$ [mm]
right	A	0.79±0.04	28.17±15.38	2.54±1.13
parotid	B	0.85±0.04	17.83±9.95	1.65±1.08
	C	0.83±0.05	16.48±8.88	2.03±1.03
	RS	0.81±0.06	17.33±10.72	2.28±1.31
	IOV	0.84±0.04	10.76±4.35	1.40±0.45
left	A	0.80±0.04	18.05±4.86	2.01±0.56
parotid	B	0.85±0.03	14.98±6.88	1.39±0.54
	C	0.84±0.04	13.96±5.38	1.69±0.58
	RS	0.83±0.04	14.39±6.11	1.89±0.65
	IOV	0.83±0.04	10.94±3.75	1.59±0.63
spinal	A	0.73±0.10	15.74±10.54	2.50±2.31
cord	B	0.83±0.06	11.98±11.41	1.65±1.57
	C	0.73±0.12	15.03±11.92	2.33±2.01
	RS	0.74±0.10	14.76±10.56	2.02±1.49
	IOV	0.79±0.07	7.12±5.15	1.55±0.81
mandible	A	0.74±0.08	16.72±8.02	1.49±0.63
	B	0.84±0.04	12.16±5.68	0.83±0.24
	C	0.84±0.04	10.66±4.72	0.95±0.29
	RS	0.81±0.06	12.06±4.00	1.14±0.44
	IOV	0.85±0.04	8.94±3.16	0.92±0.45

The intervals of mean values ± 1 SD of manually and auto-segmented volumes overlapped for all ROIs and auto-segmentation methods, besides for the spinal cord with approach C (maSTEPS), where the auto-segmentation approach tended to over-segment in comparison to the manual approach. There was also a trend for the other ROIs towards larger segmented volumes for this approach.

Figure 5.5 illustrates boxplots of the DSC, HD and MSD for all ROIs and auto-segmentation methods The stars indicate statistical significance. Table 5.3 lists the mean and standard deviations for all applied geometric measures. The inter-observer variability was included as a reference value.

The mean DSC for approach A ranged from 0.73 to 0.80. I found statistically significant improvements when using the multi-atlas approaches B and C with a mean DSC larger than 0.83 for all ROIs, except for the spinal cord when using approach C. Differences between the mean DSC values ranged from 0.06 for the parotids to 0.10 for the mandible and the spinal cord. The approach in RayStation led to a mean DSC of

0.74 to 0.83. This finding was comparable to the multi-atlas approaches, where approach B outperformed RayStation by a small difference in terms of the DSC of 0.03 to 0.09.

The superior performance of the multi-atlas approaches also held for the mean MSD with 1.49 to 2.54 mm (approach A) compared to 0.83 to 2.03 mm (approaches B and C). With a mean MSD of 1.14 to 2.28 mm for the approach in RayStation, the performance of the multi-atlas approaches was again similar, with approach B slightly outperforming RayStation with a small difference in the mean MSD of 0.31 to 0.63 mm. Additionally, this was also the case for the mean HD with 15.75 to 28.17 mm (approach A) compared to 11.98 to 17.83 mm (approaches B and C), however, in this case, the differences were not as substantial as for the other measures. The same as for the DSC was also true for distance-related measures: the performance of approach C for the spinal was not significantly better than the other two approaches.

I found a trend towards smaller SDs for all quantitative measures and ROIs when applying multi-atlas approaches. When using the multi-atlas approaches (B and C), the mean values of all geometric measures for all ROIs were within one SD of the inter-observer variability. The best-atlas approach (A) had lower accuracy than the inter-observer variability in case of the right parotid and the mandible.

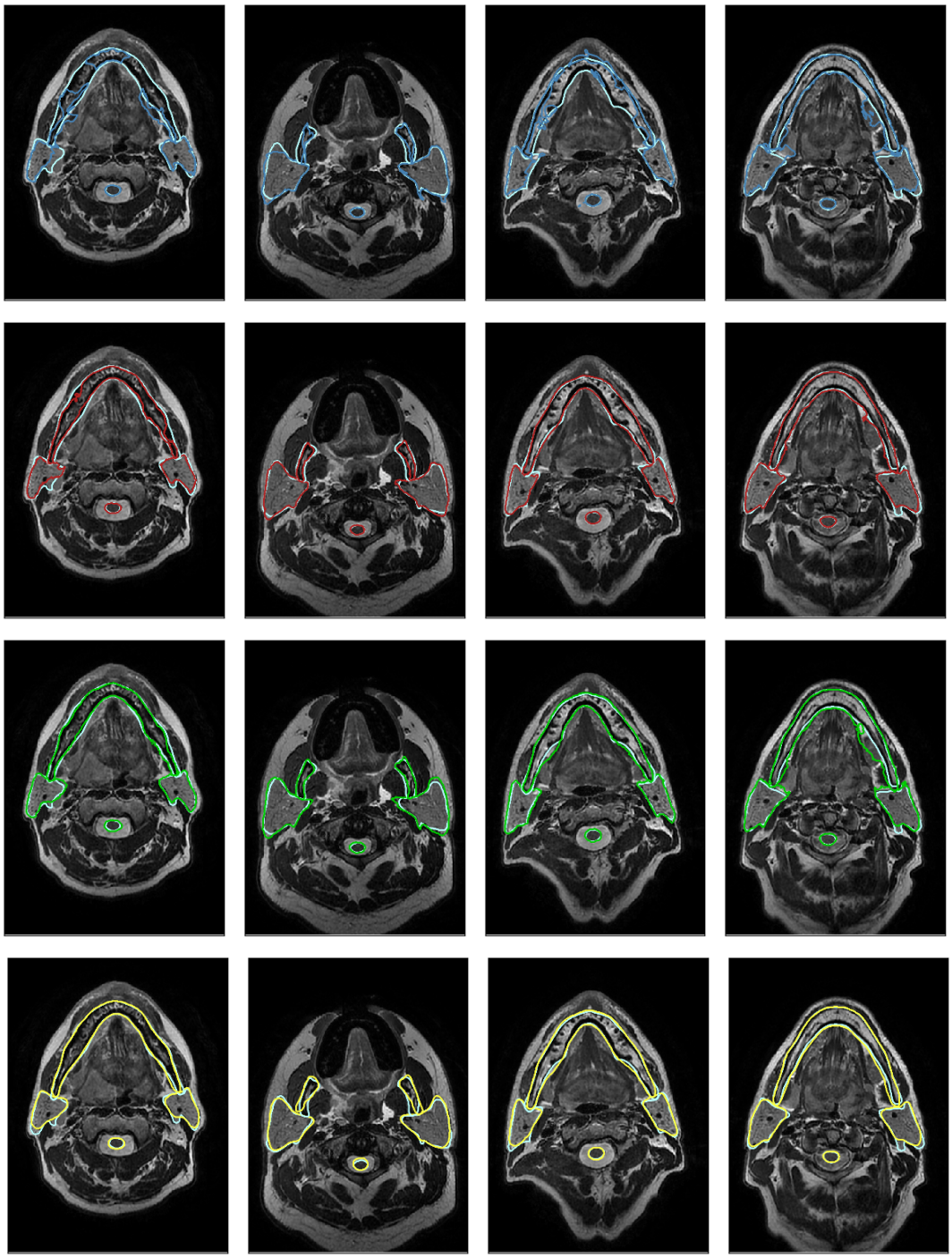


Figure 5.4: This figure shows in each column a typical example comparing the manually segmented ROIs (light blue) to approach A (dark blue, first row), approach B (red, second row), approach C (green, third row) and to the approach in Raystation, approach D (yellow, fourth row), respectively. Each example originates from a different patient image.

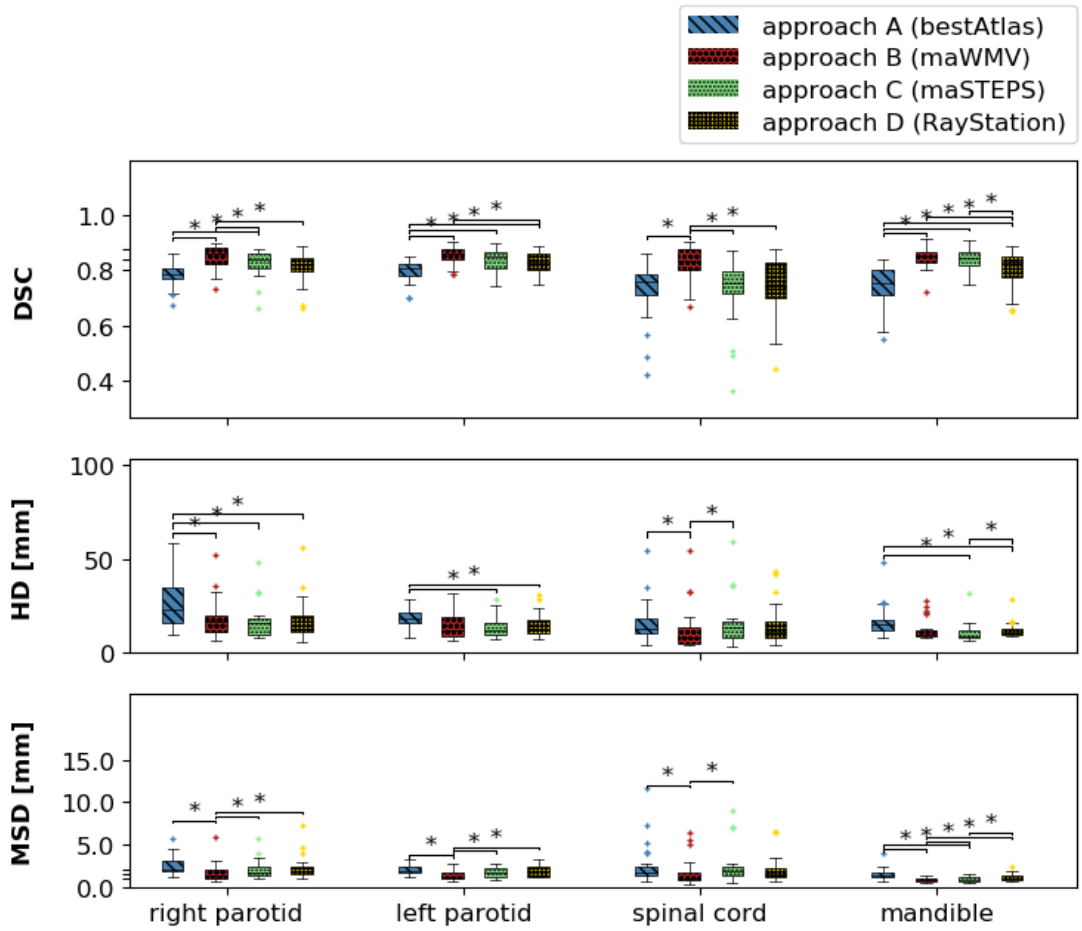


Figure 5.5: Boxplots of the DSC, HD and MSD for all OAR (x-axis) and automated segmentation approaches (A in blue, B in red and C in green). The boxes indicate the interquartile range (IQR), the whiskers extend to the minimum and maximum values. Outliers are defined as data points beyond 1.5 IQRs from the IQR, denoted with a plus sign. Stars indicate statistical significance ($p<0.05/3$).

5.4 Discussion

I investigated the application of different atlas-based methods to the segmentation of OARs and benchmarked them with a commercial atlas-based approach as well as the inter-observer variability.

5.4.1 Geometric evaluation

I compared three approaches, approach A (bestAtlas), approach B (maWMV) and approach C (maSTEPS). Both multi-atlas approaches B and C outperformed approach A in terms of the geometric accuracy (DSC, HD95 and MSD). This finding is in line with other published studies [26, 56, 143]. Comparing the two multi-atlas approaches B and C, there was no clear benefit of using one or the other, although approach B seems to have higher accuracy in most cases. As these two approaches only differ in the atlas fusion method, one can conclude that for the data utilised in this study, the performance of atlas-based approaches is mainly influenced by the quality of the image registration and choosing a local instead of a global approach (atlas fusion in the multi-atlas approaches versus global atlas selection in approach A).

The HD was not a reliable measure for the geometric accuracy of the data used in this study. As this measure provides the maximum distance between ROIs, it is susceptible to outliers and is hence not a good representative of the overall geometric accuracy (see chapter 4).

The multi-atlas approaches had similar performance compared to the commercial approach in RayStation. Approach B outperformed RayStation with a small difference in all geometric measures.

Table 5.4 lists mean reported geometric measures to compare my results with published auto-segmentation studies. The majority of the reported studies used CT scans. There were only three studies on auto-segmentation of H&N MR images [149, 154, 162]. Except for approach C for the spinal cord, the mean DSC was larger than 0.83 and the mean MSD smaller than 2 mm for the multi-atlas approaches. The developed approaches were, therefore, comparable to the reported values in table 5.4. Furthermore, all geometric measures were within one SD of the inter-observer variability that has been determined for the data in this study. Published results for the HD are sparse and have significant variations. This study here is the only one reporting on the HD for the mandible. For the parotids, my results are comparable to Daisne and Blumhofer [26] and Fritscher et al. [43]. For the spinal cord, I found a lower HD than Hoang Duc et al. [63]. The segmentation accuracy in terms of the DSC of the mandible was slightly worse in my approach compared to reported studies [56, 79, 119]. This may be attributed to

Table 5.4: This table lists geometric measures (mean DSC, mean HD and mean MSD) reported for the ROIs of this work. The mean values for the parotids are averaged between the left and right parotid.

ROI	DSC	HD[mm]	MSD[mm]	mod.	#pat.	study
paro-	0.80	23.11	2.28	MR	27	this study (A)
tids	0.85	16.41	1.52	MR	27	this study (B)
	0.84	15.22	1.86	MR	27	this study (C)
	0.82	15.86	2.09	MR	27	this study (RS)
	0.79	-	4.97	MR	14	Wardman et al. [154]
	0.77	-	-	CT	10	Beasley et al. [4]
	0.65	45	-	CT	100	Hoang Duc et al. [63]
	0.84	13	-	CT	18	Fritscher et al. [43]
	0.91	3.46	0.31	MR	15	Yang et al. [162]
	0.72	15	2.5	CT	20	Daisne et al. (2013)
	0.79	-	-	CT	5	La Macchia et al. [79]
	0.79	-	2.5	CT	10	Teguh et al. [143]
	0.83	5.8	-	CT	25	Qazi et al. [119]
	0.86	4.95	-	CT	25	Pekar et al. [115]
	0.68	-	-	CT	13	Sims et al. [134]
	0.85	-	-	CT	10	Han et al. [56]
spinal	0.73	15.74	2.50	MR	27	this study (A)
cord	0.83	11.98	1.65	MR	27	this study (B)
	0.73	15.03	2.33	MR	27	this study (C)
	0.74	14.76	2.02	MR	27	this study (RS)
	0.37	-	17.5	MR	14	Wardman et al. [154]
	0.75	40	-	CT	100	Hoang Duc et al. [63]
	0.81	-	-	CT	5	La Macchia et al. [79]
	0.78	-	2.3	CT	10	Teguh et al. [143]
	0.75	-	-	CT	10	Han et al. [56]
mandible	0.74	16.72	1.49	MR	27	this study (A)
	0.84	12.16	0.83	MR	27	this study (B)
	0.84	10.66	0.95	MR	27	this study (C)
	0.81	12.06	1.14	MR	27	this study (RS)
	0.86	-	-	CT	5	La Macchia et al. [79]
	0.93	-	2.64	CT	25	Qazi et al. [119]
	0.78	-	-	CT	13	Sims et al. [134]
	0.9	-	-	CT	10	Han et al. [56]

the fact that the published studies were conducted using CT images. As the mandible is a bony structure, it is more clearly visualised on CT images.

The results published by Yang et al. [162] demonstrate a superior performance of their algorithm. They used an atlas-based approach, refined by a machine learning post-processing step. However, in contrast to my study, they applied their approach to the auto-segmentation of post-RT MRIs using pre-RT MRIs from the same patient.

This resulted in a smaller expected variance between atlas and target images.

The performance of approach C for the spinal cord was worse than for the other multi-atlas approaches. In a closer analysis, I found that it tends to segment the spinal cord on more slices than given in the manually segmented ROIs. The brainstem is a continuation of the spinal cord towards the top of the head. With including information on the location of the brainstem in the segmentation process, I expect to solve this over-segmentation problem.

From this study, one can conclude that both multi-atlas approaches and RayStation can segment the investigated ROIs accurately enough. The approach in RayStation was faster than our current implementation of the multi-atlas methods.

5.4.2 Limitations and future work

Although I have shown in chapter 4 that calculating the dosimetric effect of auto-segmentation is a more accurate assessment of its quality than purely geometric measures, there was not enough time to do this within the scope of this thesis and it is therefore left for future work.

One limitation of this study was the relatively small number of available training data. Considering the substantial appearance variations between different patients' anatomies, a larger database would be necessary to account for these variations. However, a larger database would not invalidate the conclusions on the accuracy of the atlas-based segmentation. Instead, one would expect a higher geometrical accuracy, as more variation in the library will also more likely include images similar to the target image.

Furthermore, due to the small imaging coverage of the patients' anatomies in the superior-inferior direction, I could only include four OARs in the analysis. Treatment planning of HNC requires the segmentation of more organs at risk such as the optical structures and the brainstem. However, other ROIs can easily be included in this algorithm without any adjustments.

In this work, I looked at the segmentation of OARs which are similar in terms of shape and location for different patients. Segmentation of target volumes, however, is challenging with atlas-based segmentation due to substantial variations in shape, size and location. Due to these reasons, atlas-based segmentation is likely to fail for the segmentation of target volumes and hence, other methods, such as described in chapter 6 need to be employed.

It is a known problem that the evaluation of auto-segmentation suffers from the lack of ground truth. While I determined the inter-observer variability to provide an estimate of the upper bound on the desired auto-segmentation accuracy, I compared to the contours of one expert. This was the expert whose contours were used to create

the atlas for the auto-segmentation. Previous publications suggested combining the contours of several experts into one joint contour, for example, by using an approach called Simultaneous Truth and Performance Level Estimation (STAPLE) [155]. With STAPLE one could obtain a gold standard that might be closer to the unknown ground truth by considering the agreement between different experts on the absence or presence of the ROI at a particular location within the image. In future work, one could consider using the STAPLE of several observers as the gold standard ROIs, both, as input for the atlas-based segmentation, as well as a reference for comparison purposes.

A limitation of the atlas-based segmentation approach is the computation time. With computation times of several minutes, this approach scales with the number of atlases used in the database and would not be suitable for an online workflow. However, the use of a multi-atlas approach for the offline segmentation of pre-treatment images would already represent a significant time-gain compared to manual segmentations which can take up to several hours. In an adaptive RT workflow, one could then use previous, already segmented, images of the same patient in a single-atlas approach, which would necessitate the registration of only one image to the target image and reduce time significantly to a few minutes. I furthermore expect that one could significantly reduce the registration time by changes in the algorithm itself, e. g. by parallelising image registrations for different library images and cutting down the time for the affine registration.

5.5 Conclusion

This study showed that multi-atlas approaches could achieve a geometric accuracy in the range of the inter-expert variability for the imaging data used. I additionally benchmarked the accuracy with a commercial atlas-based approach in the treatment planning system RayStation and achieved a comparable, if not better, accuracy. For example, the segmentation of the parotid glands achieved an average DSC of 0.85 ± 0.05 for the multi-atlas approach of this study compared to 0.82 ± 0.07 using the approach in RayStation. All multi-atlas approaches outperformed the simple best-atlas approach.

This study showed that atlas-based auto-segmentation approaches could achieve clinically acceptable results. While the computation time is currently too large for an application in an online workflow, it can be used for an offline treatment planning while reducing the time burden to the clinician significantly.

Chapter 6

Deep learning-based algorithms

The work presented in this chapter aimed to investigate the feasibility of, as well as to design and develop deep learning-based methods for the segmentation of head and neck MR images. The nature of this work was exploratory, where I applied established approaches which have been shown to work well in other applications and made adjustments accordingly. I furthermore explored the potential of deep learning approaches to handle small annotated datasets, which is a very commonly encountered problem in medical imaging and in particular radiotherapy. For this purpose, I designed dedicated methods, able to overcome these limitations.

6.1 Introduction

Deep learning is a group of machine learning approaches that learn to abstract data hierarchically, composing multiple non-linear operations. The underlying aim is to learn a function which maps input data to outputs by presenting many example cases to the algorithm. For instance, one might want to develop an algorithm that can recognise the type of animal shown in an image (see figure 6.1). For this purpose, one could train an algorithm, represented by the black box in this figure, that can recognise that the image shows a cat.

In traditional machine learning approaches, such as support vector machines or random forests, features that are used to guide this process need to be hand-crafted and require extensive expert knowledge. For instance, in the cat classification example in figure 6.1, the black box would be fully designed to detect hand-crafted features which one associates with a cat, such as a fur, the eyes, the nose or the ears. In deep learning, these features are learned directly from the data. By feeding many images into the "black box" and telling it whether it was an image of a cat, the algorithm learns to detect cats in new images. This process can be compared to how a child would learn to recognise a cat: its parents would teach the child how a cat looks like by "labelling" many examples of cats in real life or images. In medical imaging, a radiologist has seen many MR images and learned how a specific ROI looked like before being able to detect these regions of interest in new images.

In this work, I focused on CNNs, which are learning complex functions through convolutional operations, designed to detect specific patterns or features in the input data. CNNs have shown great success in solving computer vision tasks, such as object detection and segmentation. Applications of CNNs to the segmentation of ROIs on MR images are still in their infancy. Litjens et al. [92] provide a detailed review of recent

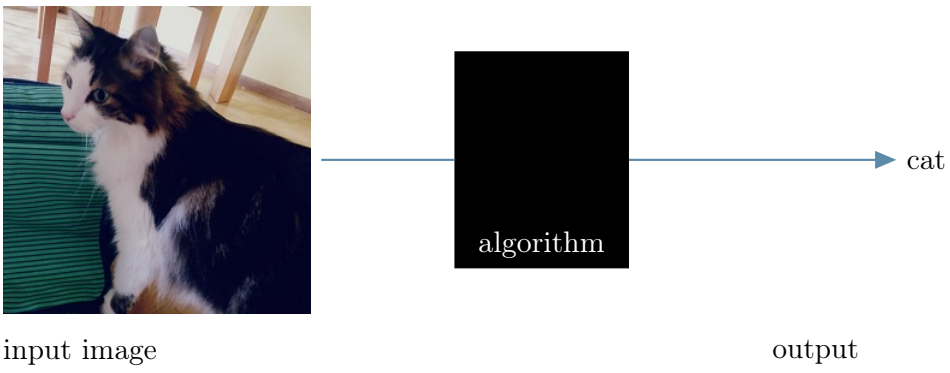


Figure 6.1: An example of a classification problem: a cat image is fed into the algorithm, whereas the aim of the algorithm is to predict that a cat can be seen in the image.

publications in the field of medical image analysis.

Despite first developments of CNNs already in the 1980s [44] and first applications in the early 1990s by [86], they only became popular recently. The main breakthrough was in 2012 when Krizhevsky et al. [78] won the popular ImageNet challenge [28] by a large margin. This breakthrough can mainly be attributed to the lack of high quality annotated training data previously and the only recent universal availability of hardware improvements in terms of GPUs and computational power.

Deep learning is a rapidly growing field with many new publications each day. The field has seen a wide range of applications, from outperforming humans at computer games to automatically driving cars and medical image processing. Therefore, I focus on introducing the main concepts in deep learning and mention only publications that are closely related to my work. I refer the interested reader to Goodfellow et al. [50], providing an in-depth introduction to the concepts behind deep learning, and Litjens et al. [92], giving a broad overview of recent publications of applications in the field of medical imaging.

Although deep learning has shown promising results for many tasks, the interpretability, i. e. identifying how a model makes a prediction, is still challenging. Deep learning is often described as a black box: data goes in (in our example, the cat image) and a decision comes out. However, the processes happening between the input and output are difficult to grasp and often very complex. This "black box"-phenomenon renders research in this field still empirical.

The aim of the presented work in this chapter was to develop and investigate deep learning-based methods applied to the segmentation of the parotid glands in HNC MR images. Section 6.2 introduces the basic building blocks and concepts behind CNNs, which are essential to understand the following sections. This introduction is mainly derived from Goodfellow et al. [50] and Litjens et al. [92]. I then present results on auto-segmentation with CNNs, applied to the exemplary case of the parotid glands, in section 6.4. Afterwards, I address the problem of a lack of annotated training data by implementing two approaches: transfer learning (see section 6.5), as well as synthetic data generation (see section 6.6). Most of this work has been submitted as abstracts to peer-reviewed conferences and was accepted as oral presentations (see appendix A.2).

Although I have shown in chapter 4 that calculating the dosimetric effect of auto-segmentation is a more accurate assessment of its quality than purely geometric measures, there was not enough time to do this within the scope of this thesis and all presented segmentation methods are therefore only evaluated in terms of their geometric accuracy.

6.2 A short guide to convolutional neural networks

The following sections 6.2.1 to 6.2.4 provide an overview of the basic concepts in deep learning, necessary to understand sections 6.4, 6.5 and 6.6. I first introduce the basic building blocks of CNNs in 6.2.1. Section 6.2.2 describes how CNNs are trained and how they can be used to infer predictions on unseen data.

An often encountered problem of the application of CNNs is overfitting. Overfitting describes the situation where an algorithm is able to describe the training data well but fails to infer good predictions on unseen data. This is particularly true when only using small training datasets. Section 6.2.3 describes concepts on how to mitigate this problem and introduces methods of regularisation. Finally, section 6.2.4 discusses a challenging property of CNNs, the presence of many external parameters, also known as hyperparameters, which need to be adjusted according to the specific application.

6.2.1 The basic building blocks of convolutional neural networks

A CNN is composed of multiple layers of non-linear operations. In contrast to traditional machine learning approaches, features are not hand-crafted but learned by the algorithm itself. Artificial neural networks are roughly modelled according to our understanding of the human brain. In the top row of figure 6.2, the basic elements of an artificial neural network are illustrated. A neural network is composed of multiple neurons, also knowns as nodes, which are arranged in layers. A neuron is the basic unit of a neural network. Like a neuron in the brain, it comprises a number of weighted inputs (dendrites), a processing unit (cell nucleus) and an output (axon), which is the weighted sum of the inputs and a bias unit. The connection of neurons are illustrated on the right of figure 6.2. Conventionally, each input node is connected to each node in the consecutive layer. These types of networks are also known as fully connected layers. The first layer is called the input layer, whereas the last layer is called the output layer. In image classification or segmentation problems, the input layer is a function of the voxel intensities. The output layer is either an integer or a vector of integers for classification problems, a label map for segmentation problems, or a number for regression problems. Any layer (one or multiple) in between the input and the output layer is known as a hidden layer as, unlike the input and output layers, these layers do not have a direct connection with the outside world. Each hidden node is determined as a function of the weighted sum of all nodes of the previous layer and a constant bias node (illustration on the top right of figure 6.2. These linear terms are then passed on to a non-linear function σ. With input

nodes x_i, bias b_i and weight matrix W, a node in the hidden layer is determined as

$$a_i = \sigma\left(\sum_{j\in\mathbb{N}} w_{ij}x_j + b_i\right). \tag{6.1}$$

The function σ is the so-called activation function. See more details in section 6.2.1.2 on page 96. Weights and biases are usually randomly initialised and learned throughout the training process by iteratively updating them (see section 6.2.2 on page 98).

The basic building blocks of CNNs are convolutional layers. Instead of connecting each node in one layer to each node in the consecutive layer, connections are only local (so-called sparse connectivity). The weights are composed of convolutional kernels that are shared for the full layer. This property is different from fully connected layers, where each element of the weight matrix is only used once in the network and llustrated in the bottom part of figure 6.2. Objects that share similar properties but occur at different locations in an image can, therefore, be detected by the same convolutional kernel, which is shared in the layer. Furthermore, in this way, the number of parameters that need to be learned is reduced considerably.

I introduce the terminology commonly used in applications of deep learning in the following sections, guided with an example of the classification of a cat image.

Architectures of CNNs consist of three primary mathematical operations: convolutional, (non-linear) activation and pooling. I briefly discuss each operation.

6.2.1.1 Convolutional operation

The purpose of a convolutional operation is to extract features from the input images. In each convolutional layer, a series of so-called filters or feature detectors are defined and convolved with the images. The resulting object is called a feature map. Figure 6.3 illustrates an example of the first convolutional layer for the classification of cat images.

Intuitively, a convolutional filter is a filter that leads to large values in the feature map (input image convolved with filter) when a particular type of structure is present in the image, such as an edge or corner. In the example of the classification of the cat image in figure 6.3, earlier layers would detect edges and lines in the images through multiple convolutional filters, proceeding to higher-order features such as the fur, eyes, ear and nose in later layers and finally efficiently describing a cat. These convolutional filters are learnt by the network, unlike in traditional machine learning approaches where they are hand-crafted.

The following parameters define each convolutional operation:

- **(spatial) filter size F:** This is the size of the convolutional filters or kernels.

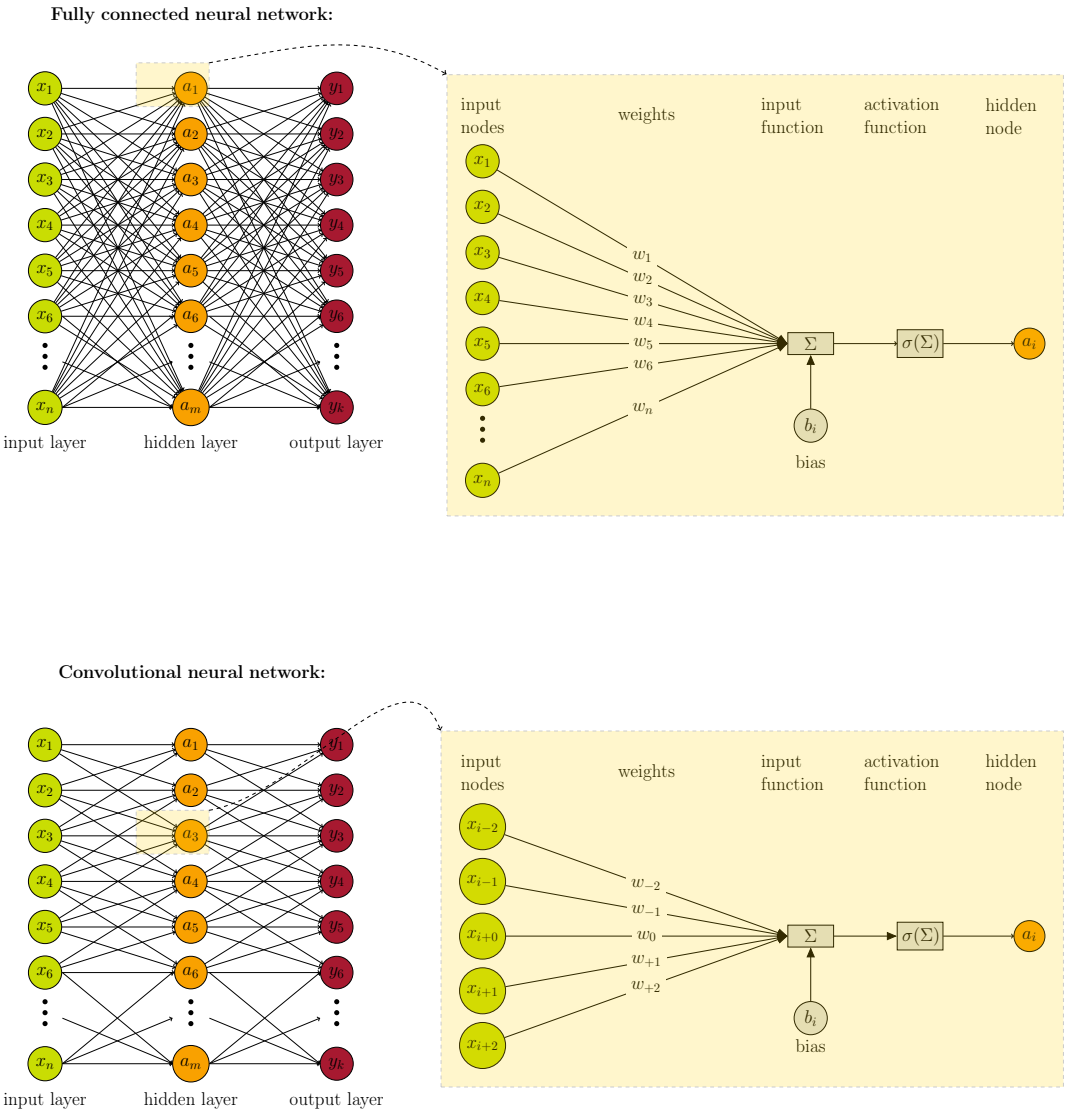


Figure 6.2: Artificial neural networks: This figure illustrates the composition of a fully connected neural network (top), compared to a convolutional neural network (bottom). On the right, an illustration of how a hidden node is determined from the previous nodes is provided.

Commonly, squared or cubic filters with side lengths of 3 or 5 pixels or voxels in each dimension, are used. In the example, the first layer uses a filter size of 3x3. The depth of the filters is always equal to the number of channels of the previous layer (e. g. 3 for coloured images), and a sum is performed over the depth elements. The filter depth in hidden layers is equal to the number of applied filters in the previous layer.

- **number of filters:** For each convolutional layer, not only one but multiple filters are used to extract different features. The cat classification example applied 64 filters in the first convolutional layer.

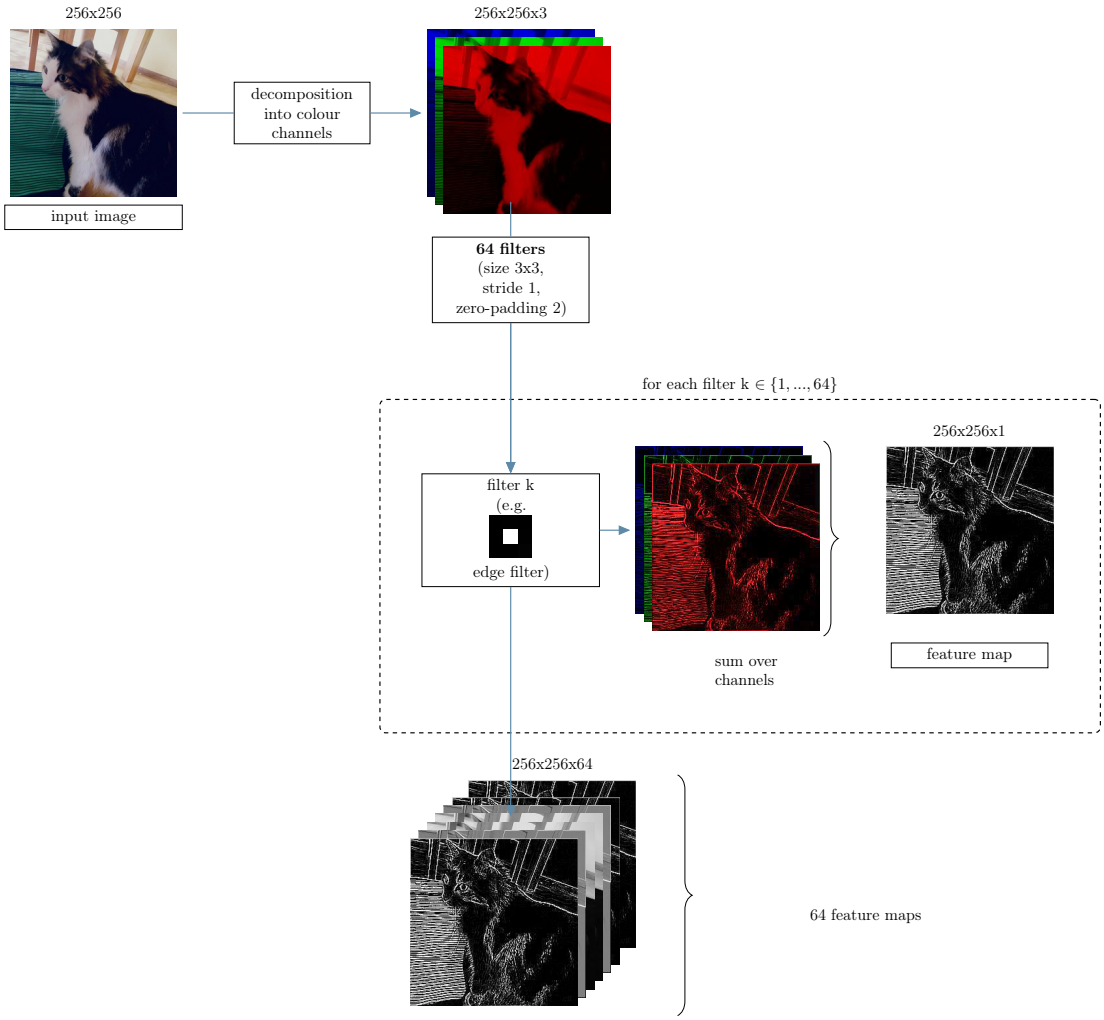


Figure 6.3: Convolutional operation for the example of the classification of a coloured cat image. After decomposing the coloured image into its three colour channels, 64 filters are applied to each of the colour channels. One example of such a filter is illustrated in the central part of this figure. The feature map for each filter is derived from the sum over the channels, resulting in 64 feature maps for this particular example.

- **stride:** The stride is the number of pixels/voxels that a filter moves at a time. The example used a stride of 1.
- **zero-padding:** Due to the nature of convolutions, rudely applying the convolution would reduce the size of the output in each operation by the "filter size minus 1". To avoid this reduction, the input image can be padded with zeros around the border, such that the filter is also applied to the bordering elements of the input image. Commonly, a zero-padding equal to the "filter size minus 1" is applied. The presented example employed a zero-padding of 2 to counter the size reduction due to the convolutions.

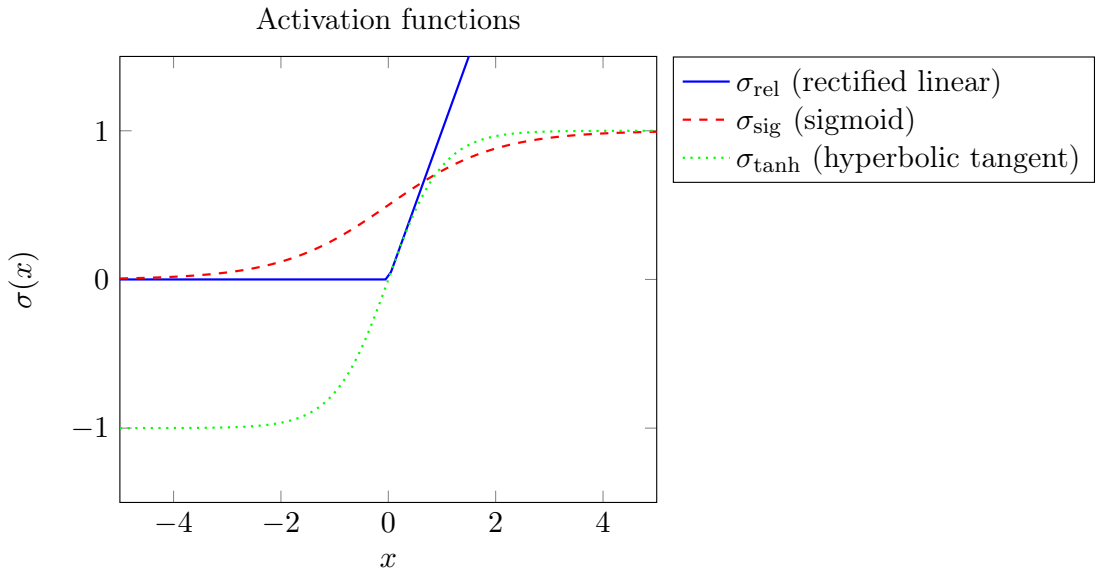


Figure 6.4: Activation functions commonly used in CNNs: the rectified linear function (solid blue), the sigmoid function (dashed red) and the hyperbolic tangent function (dotted green).

6.2.1.2 Activation

A network of subsequent linear operations could mathematically be simplified by a single linear layer and would not be able to learn anything non-linear. To introduce non-linearity, a so-called activation function is applied. Commonly used activation functions are the rectified linear function,

$$\sigma_{\text{rel}}(x) = \max(0, x), \tag{6.2}$$

the sigmoid function,

$$\sigma_{\text{sig}}(x) = \frac{1}{1 + e^{-x}} \tag{6.3}$$

and the hyperbolic tangent function,

$$\sigma_{\text{tanh}}(x) = \tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}. \tag{6.4}$$

Figure 6.4 illustrates the activation functions. The rectified linear function has shown to be most effective in recent works [50].

The task of the last layer is to turn any values produced in the cascade of the neural network into values that can be interpreted by humans. In a classification problem, this could the either "yes" or "no", a class label, or the probability of belonging to a particular class. A sigmoid function yields a value between 0 and 1 which can be

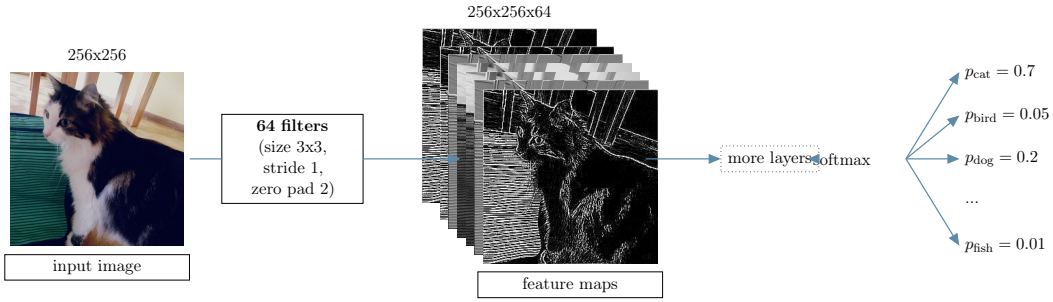


Figure 6.5: Cascade of layers leading to the prediction layer for the example of a classification of a cat image. A softmax activation function (see equation (6.5)) is applied to obtain the probability of the image belonging to a particular class (dog, cat, bird,...).

interpreted as a probability. For multiple classes, one can use the softmax activation function:

$$\sigma_{\text{softmax},i,k} = \frac{e^{-x_i}}{\sum_{j=1}^{k} e^{-x_j}} \tag{6.5}$$

with i being the class of interest, x_i the feature vector from the previous layer for class i and k the total number of classes.

In the cat classification example, the last layer would yield the probability of this image belonging to a particular class, e. g. a cat, dog or bird as shown in figure 6.5.

6.2.1.3 Pooling

The purpose of pooling is to keep the essential information of each feature map while discarding information, which is not relevant. Pooling uses some function to summarise subregions. There are different types of pooling operations, for example, max-pooling or average-pooling. A pooling kernel is slid over the image, where in each region, summary statistics replace the individual pixels or voxels, that is, for max-pooling the maximum value and for average-pooling the average value. The pooling kernel moves by one stride size at a time. Figure 6.6 illustrates a max-pooling operation for the classification of a cat image.

6.2.1.4 Receptive field

A receptive field is the size of a region that affects a particular feature. It is an important quantity to determine how much of the surrounding voxels are taken into account in the decision chain. For simplicity, I assume square sizes for the input, the convolutional and the pooling filters in the following. With a kernel of size F and stride s, the receptive

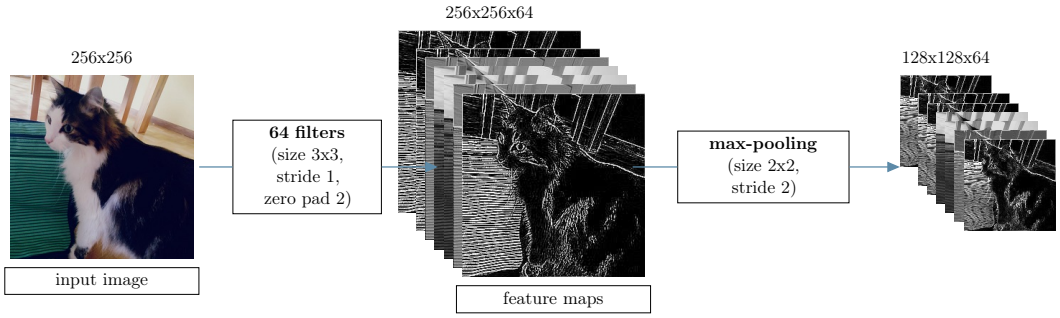


Figure 6.6: Pooling operation for the example of a classification of a cat image.

(square) field size r_m of each layer m can be calculated as follows:

$$r_m = r_{m-1} + (F-1) \cdot j_{m-1} \tag{6.6}$$

with

$$\begin{aligned} j_m &= s \cdot j_{m-1} \\ r_0 &= 1 \\ j_0 &= 1. \end{aligned}$$

The variable j_m can be interpreted as a jump between neighbouring features in terms of input space coordinates. For the example of the classification of a cat image, the receptive field after the first convolution has size $r_{\text{conv}} = 2$ and after the pooling operation $r_{\text{pool}} = 4$. That means that any voxel in the 64 feature maps has "seen" 2x2 voxels of the original image, whereas after the first pooling operation, any voxel in the pooled feature maps has "seen" 4x4 voxels of the original image.

6.2.2 Training and prediction phases of neural networks

Neural networks need to be trained before predictions on new data can be inferred. Training means to determine the best set of weights to maximise some performance parameters of a neural network. After the training phase ends, the trained model, i. e. the network with its fixed weights and biases, can be applied to unseen data to make predictions. Figure 6.7 illustrates the primary two phases in CNNs with the example of image segmentation.

Learning which weights optimise the desired performance measure is infeasible with a brute-force approach as there are too many parameters in neural networks. The AlexNet, the CNN introduced by Krizhevsky et al. [78] in 2012, which won the ImageNet challenge by a large margin, has 60 million of these parameters, for example. Therefore, one needs

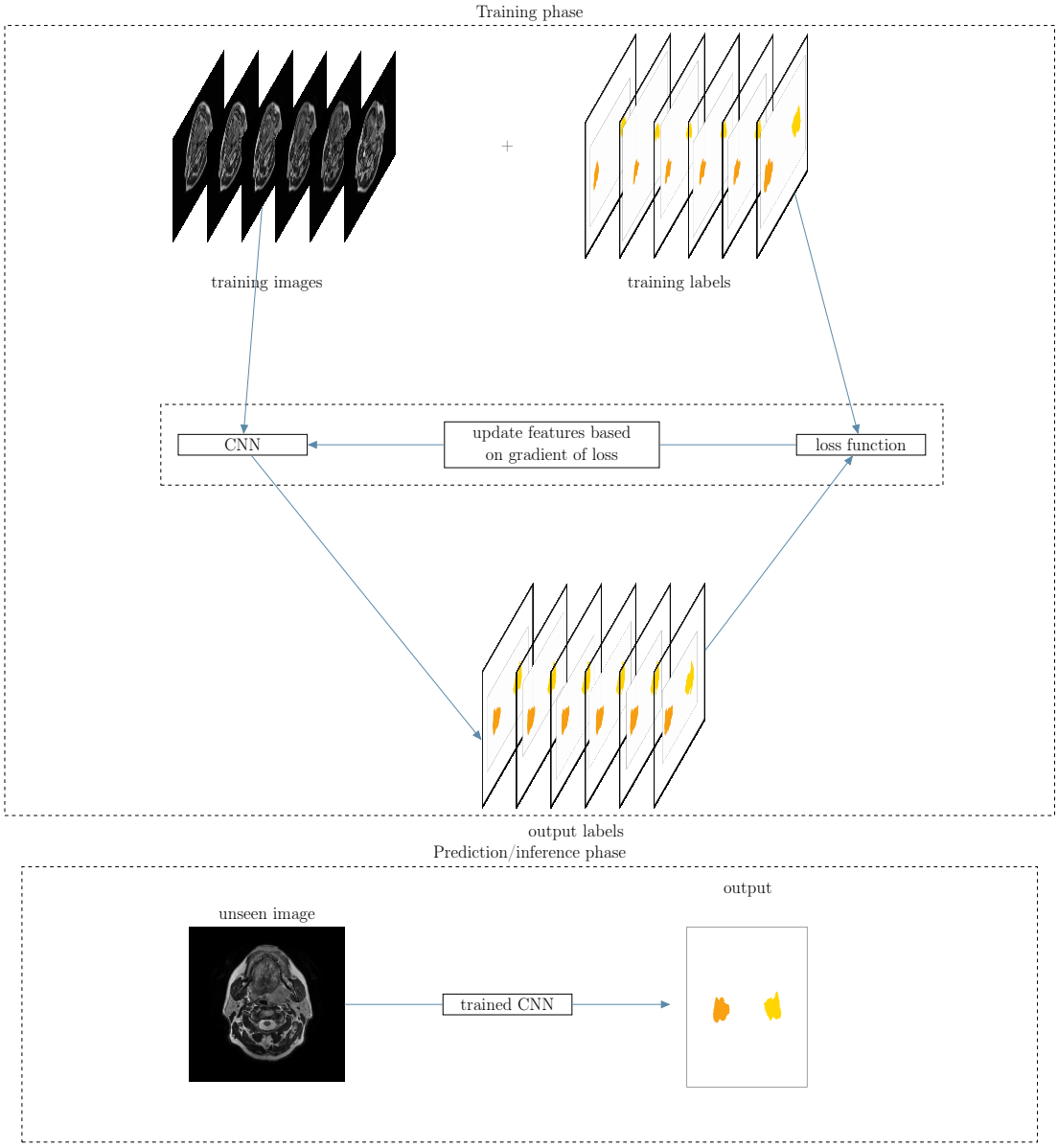


Figure 6.7: This figure illustrates the two main phases for developing CNN-based approaches: First, one needs to train the network to learn an optimal set of the weights by using many example images (top part of figure). This is an iterative process. After training, the weights are fixed and the testing images are fed into the network to generate an output (here: labels).

to find a more elegant solution to this problem. In supervised learning, one defines a so-called cost or loss function $\mathcal{L}$ that quantifies the error between the desired output and the network's output. The aim is to minimise this error throughout the training process.

6.2.2.1 Data splitting

A general approach for the design and validation of a deep learning model is to split the input data into training, validation and testing. The training data is used to learn the

optimal weights of the network, which minimise the loss function. As this could lead to overfitting to this particular set of data (read more in section 6.2.3), one generally uses an independent validation set to test the performance of the model on unseen data. If the model performance does not meet the desired accuracy, one can change the model, i. e. adjust the hyperparameters, train the network again and test it again on the validation data. This process is iterated until the model performance meets the desired accuracy. However, this process can still lead to overfitting as the validation data was used to find an optimal solution. For this purpose, one can use a further dataset, the test dataset, which has neither been used in the training, nor the validation phase. The developed model is fixed and the only purpose of the test dataset is to check the model performance.

A problem with this approach is that, if the available dataset is small, the split into training and testing is highly biased and furthermore, the training dataset becomes even smaller as not all the data is used to find the best model. K-fold cross-validation mitigates this dilemma by splitting the dataset into k equally sized groups. In each fold, one group is used as the validation data and the remaining k-1 groups are used to train the model. This process is repeated k times and performance is induced by averaging over all folds.

6.2.2.2 Loss functions

A suitable loss function is highly dependent on the method and data at hand. In general, these can be classified into regression and classification loss functions. With Θ being the network's parameters (weights and biases), N the number of training examples, $x^{(i)}$ the input values and $y^{(i)}$ the ground truth of the ith training example, the prediction of the ith training example is defined as:

$$\hat{y}^{(i)} := \sigma(\Theta, x^{(i)}). \tag{6.7}$$

Typical examples of loss functions in regression problems are the mean square error (MSE):

$$\mathcal{L}_{\text{MSE}} = \sum_{i=1}^{N} \frac{\left(y^{(i)} - \hat{y}^{(i)}\right)^2}{N} \tag{6.8}$$

and the L1 loss:

$$\mathcal{L}_{\text{L1}} = \sum_{i=1}^{N} \left| y^{(i)} - \hat{y}^{(i)} \right|. \tag{6.9}$$

I restrict myself to binary classification problems for simplicity. All loss functions can be easily extended to multi-class problems. Typical classification losses are the binary cross-entropy:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}\left(y^{(i)} \cdot \log\left(\hat{y}^{(i)}\right) + \left(1 - y^{(i)}\right) \cdot \log\left(1 - \hat{y}^{(i)}\right)\right) \tag{6.10}$$

or (for segmentation problems) the Dice loss:

$$\mathcal{L}_{\text{Dice}} = \sum_{i=1}^{N}\left(1 - \frac{2 \cdot \sum_{k=1}^{N_{\text{vox}}}(y_k^{(i)} \cdot \hat{y}_k^{(i)})}{\sum_{k=1}^{N_{\text{vox}}} y_k^{(i)} \cdot \sum_{k=1}^{N_{\text{vox}}} \hat{y}_k^{(i)}}\right) \tag{6.11}$$

with the number of voxels N_{vox}. The Dice loss function has been introduced by Milletari et al. [96] to account for highly unbalanced problems where the foreground in the image (region of interest) only contributes to a small percentage of the full image. Other studies use a weighted cross-entropy to account for this issue:

$$\mathcal{L}_{\text{wBCE}} = -\frac{1}{N}\sum_{i=1}^{N} w_{\text{fg}}\left(y^{(i)} \cdot \log\left(\hat{y}^{(i)}\right) + w_{\text{bg}}\left(1 - y^{(i)}\right) \cdot \log\left(1 - \hat{y}^{(i)}\right)\right), \tag{6.12}$$

with w_{fg} being the weight for foreground voxels, and w_{bg} the weight for background voxels. In general, loss functions need to be differentiable with respect to the weights and biases in the optimisation process.

6.2.2.3 Optimisation algorithms

Most optimisation algorithms are based on calculating gradients of the loss function with respect to the network's parameters. A standard approach is the gradient descent algorithm. With $\Theta = \left((w_l)_{l\in[1,N_{\text{w}}]}, (b_k)_{k\in[1,N_{\text{b}}]}\right)$ being the network's parameters and $\mathcal{L}(\Theta)$ the loss function, one would like to solve the following problem:

$$\arg\min_{\Theta} \mathcal{L}(\Theta). \tag{6.13}$$

As this is not solvable in an analytical way, one gradually decreases the loss function. This is done by iteratively calculating the current slope of the loss function and moving in the direction that decreases its value. The slope of the loss function is obtained through its gradient with respect to its parameters:

$$\nabla_{\Theta}\mathcal{L} = \left(\frac{\partial\mathcal{L}}{\partial w_1}, \frac{\partial\mathcal{L}}{\partial w_2}, \cdots, \frac{\partial\mathcal{L}}{\partial w_{N_w}}, \frac{\partial\mathcal{L}}{\partial b_1}, \frac{\partial\mathcal{L}}{\partial b_2}, \cdots, \frac{\partial\mathcal{L}}{\partial b_{N_b}}\right). \tag{6.14}$$

The parameters are then updated as follows:

$$\Theta \mapsto \Theta - \alpha \cdot \nabla_{\Theta} \mathcal{L} \tag{6.15}$$

where α is the so-called learning rate, the amount by which the loss function is updated in each iteration. This algorithm is referred to as gradient descent and a standard approach in many optimisation problems. In practice, it is infeasible to follow this simple approach. For each training example, one would need to calculate equation (6.15), which can be computationally expensive. For this reason, one usually uses a variant of this approach. Instead of computing the gradient over the entire training set, one can calculate the gradient only for one training example, or a small number of training examples. The number of training examples shown before updating weights is known as the batch size. Following this approach is a stochastic approximation of the gradient over the full training dataset and, therefore, the approach is called stochastic gradient descent.

This iterative process of calculating gradients and updating the parameters is repeated before a pre-defined stopping criterion is met. The deep learning term for one iteration over the whole training dataset is an epoch.

Backpropagation

As the loss function is not a direct function of the network parameters, one cannot directly calculate derivatives with respect to the network parameters. Instead, the derivatives have to be calculated via chain rules. This process is called backpropagation, as one calculates the error to a specific weight by starting from the output with the error in the loss function and then subsequently "propagates" this error through each layer of the network until one reaches the weight in question. This is done for every weight in the network.

Advanced optimisation algorithms

The optimisation procedure above requires to carefully choose a learning rate α that describes the step size made to update the parameters. This can be a challenge. As illustrated for image registration problem earlier in figure 5.2 on page 72, if the learning rate is chosen too small, it can take very long to reach a minimum and there is a high risk to be stuck in suboptimal local minima. However, if the learning rate is too large, convergence might not be achieved and one might get large fluctuations in the loss function (overshooting over the minima). Furthermore, different parameters in the network might require different individual learning rates. In the algorithms described

above, the parameters are all updated at the same rate.

Several approaches address these problems. In the following paragraphs, I introduce basic concepts and describe a popular algorithm called Adam [73] in detail, as this is the approach I used in this work. More background information can be found in Ruder [128].

Momentum Close to local minima, the surface of a loss function is generally much steeper in some dimensions compared to others. One can visualise this by imagining a long ravine, where the minimum is along the long axis of the ravine. This leads to fluctuations in the loss function, up and down the edges of the ravine, and a slow progression towards the minimum along the ravine. To prevent this, one can define an additional term in the parameter update to help the optimiser move in the relevant direction. This additional term is proportional to the past parameter update. With the update u_t at step t, the parameters are updated as follows:

$$\Theta_t = \Theta_{t-1} - u_t \tag{6.16}$$

with

$$u_t = \gamma \cdot u_{t-1} + \alpha \nabla_\Theta \mathcal{L}(\Theta) \tag{6.17}$$
$$\gamma : \text{momentum term, usually } \gamma \approx 0.9$$

This moves the parameter updates in the same direction as in a previous step and decreases updates in another direction. Therefore, fluctuations in the loss function can be reduced.

Adam: Adaptive moment estimation algorithm Adam was suggested by Kingma and Ba [73] to address two main issues: biasing the update of the parameters in the direction of the current gradient and individualising the size of the updates according to individual parameters. The algorithm updates the network's parameters by the first and second moments of the gradient (i. e. the mean and the standard deviation): Let m_t be the first moment of the gradient at time t, v_t the second moment and β_1 an β_2 the decay rates for the moments, respectively,

$$m_t = \frac{\beta_1 \cdot m_{t-1} + (1-\beta_1) \cdot \nabla_\Theta \mathcal{L}(\Theta)}{1-\beta_1^t} \tag{6.18}$$

$$v_t = \frac{\beta_2 \cdot v_{t-1} + (1-\beta_2) \cdot (\nabla_\Theta \mathcal{L}(\Theta))^2}{1-\beta_2^t}. \tag{6.19}$$

The parameters are then updated as follows:

$$\Theta_t = \Theta_{t-1} - u_t$$

with

$$u_t = \alpha \cdot \frac{m_t}{v_t + \epsilon}. \tag{6.20}$$

The term ϵ prevents from dividing by zero when $v_t = 0$. The moments are initialised with zero: $m_0 = 0 = v_0$. The authors recommend the settings $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-8}$ and a global learning rate $\alpha = 0.001$. The first moment ensures moving with the momentum, the second moment scales parameter updates.

There are various other algorithms, such as Adagrad [33], Adadelta [167] or RMSProp [60]. The Adam algorithm has generally shown to be fairly robust and to perform well for many applications.

6.2.2.4 Initialisation

Before training a neural network, the weights and biases have to be initialised. It is crucial not to initialise the weights to the same value, as otherwise the gradients are identical. Hence, it is important to initialise them to random values to circumvent this problem. Another possibility of initialisation the weights is through pre-training: in this case, the network has been trained using data from a different application. Instead of random initialisation, these weights are then used to train the network for a new application. This is a widespread technique called transfer learning [18]. I address this in more details in section 6.5.

6.2.2.5 Summary: Basic training steps

In summary, training a neural network is composed of the following steps:

(1) initialise the network's parameters (weights and biases)

(2) feedforward the network with training data in packages of a pre-defined batch size to predict output

(3) calculate the error using a pre-defined loss function between the desired and the predicted output

(4) backpropagate the error through the network for each network parameter

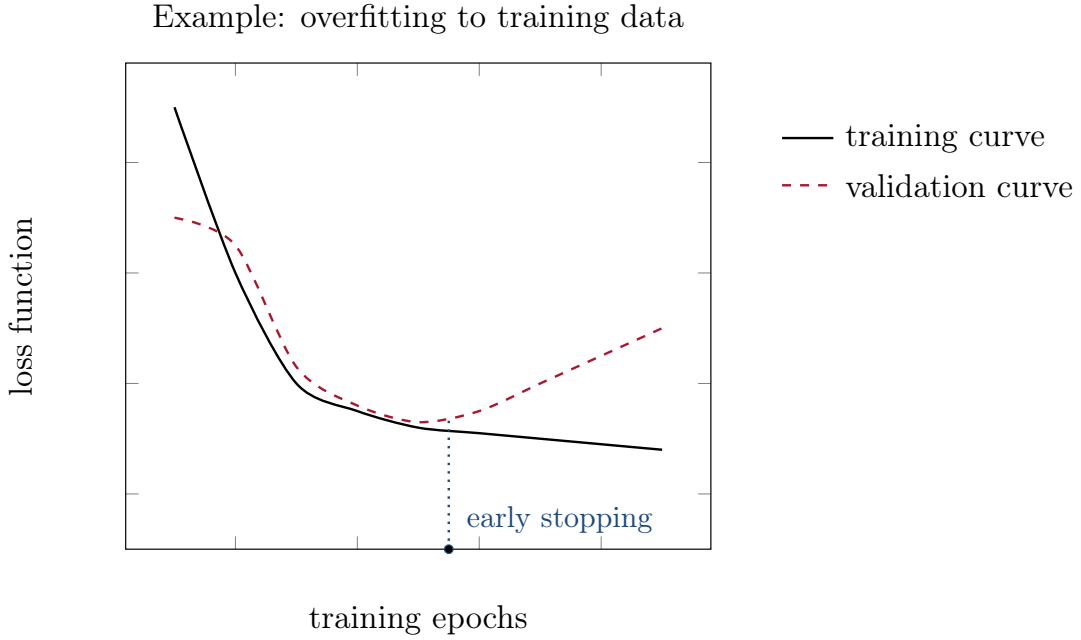


Figure 6.8: This figure depicts the loss function as a function of the number of training epochs. When developing CNN-based models, the training data (used to adjust weights) are generally well represented by the data but may fail for unseen data (validation data). After a certain number of epochs, the validation curve increases, while the training curve still decreases. A method to mitigate overfitting is early stopping, as indicated in blue.

(5) update each network parameter with a step size defined by the learning rate using an optimisation algorithm, e. g. stochastic gradient descent

(6) repeat steps (2) to (5) for a pre-defined number of iterations (epochs) or until a stopping criterion is met

6.2.3 Generalisability and regularisation

Training neural networks can easily lead to overfitting to the training set, especially if the training set is small or does not reflect the variance in the population data adequately. Figure 6.8 illustrates the training and validation loss as a function of the number of training epochs. After a certain number of epochs, the model still performs well for the training data but deteriorates for the validation data. Regularisation can help reduce overfitting and make the trained model generalise better. There are several techniques for this, some of which I introduce in the following sections.

6.2.3.1 Input and batch normalisation

The input to a CNN is typically standardised to a common scale to help with a stable training process. In the training, the inputs are multiplied with weights and biases are

added. As typically all weights are updated with the same step size (learning rate) in the backpropagation process, input standardisation helps the updates to the weights to be of similar sizes.

As the training progresses, this normalisation may be lost, owing to different magnitudes of features. This can slow down the training progress. Introducing batch normalisation avoids this and furthermore leads to a regularisation, as features are brought to a common range. Batch normalisation usually calculates the mean and variance of features for each training batch and brings the features to a common mean and variance.

6.2.3.2 Dropout

Another strategy to reduce overfitting is the so-called dropout method. Dropout randomly drops nodes and their connections from the network with a pre-defined probability. In each training iteration (epoch), different nodes may be dropped. This prevents nodes from co-adapting too much: The value of the loss function decreases in each iteration. The update of an individual node depends on the update of its connected nodes. Therefore, nodes may change in a way to fix the mistakes of other, connected nodes. This would lead to overfitting to the training dataset as this would not generalise well to unseen data. By dropping nodes, this can be avoided as nodes cannot "rely" on the presence of other nodes, and consequently, co-adaptation might be prevented.

6.2.3.3 Early Stopping

As the training loss can always be reduced but does not necessarily improve the performance of the prediction on unseen data, one needs to choose the number of epochs to train a network carefully. To avoid overfitting one can stop the training once the performance on the validation data deteriorates. This is known as early stopping. The point at which the training is stopped ideally is indicated in blue in figure 6.8. However, in reality, the validation loss is generally not a smooth function and it is not straightforward to set a stopping criterion for when to stop the training. The validation error might well decrease again after it has increased for a certain number of epochs. The choice of a stopping criterion aims to balance training time versus generalisation error.

6.2.3.4 Data augmentation

A larger and representative training dataset helps the CNN model to generalise well to unseen data. However, it is not always possible to obtain more training data. To address this problem, there are several ways to increase the training data size by augmenting

the data. Typical examples of data augmentations are to rotate, translate, mirror, scale or deform the image, or scale the intensities values. Choosing the augmentation approach is not a straightforward problem as data augmentation needs to cover realistic transformations of the original training data and also cover the variability seen in the full population data. I address this problem and introduce a novel approach to data augmentation using generative adversarial networks in section 6.6.

6.2.4 Hyperparameters

Hyperparameters are parameters that need to be set before training and predicting with a deep learning algorithm. The hyperparameters determine the architecture of the network (e. g. number of layers, type of layers and size of filter) and the learning process (e. g. optimiser, learning rate and stopping criteria). The number of hyperparameters is generally large and depends on the application. For this reason, it is challenging to fine-tune them carefully in order to obtain the best performance in terms of accuracy and training time.

Most commonly, a manual or grid search is performed to explore the space of hyperparameters, while another suggestion was to perform a random search [5].

6.2.5 Network architectures

All CNN architectures follow similar general design principles of a sequence of convolutional layers and decreasing the spatial dimensions while increasing the number of feature maps. Many innovative ways of constructing layers allow for efficient learning. In this section, I introduce some key networks and key building blocks, commonly used and relevant for this work.

The pioneering network by LeCun et al. [86], the LeNet, was composed of two convolutional layers, intervened by average-pooling operations, and followed by a fully connected layer, as illustrated in figure 6.9. The LeNet was developed to recognise handwritten digits.

CNNs did not show many promising results until the breakthrough of Krizhevsky et al. [78] in 2012 with their AlexNet, as illustrated in figure 6.10. AlexNet comprises eight layers, with the first five being convolutional layers, intervened by max-pooling layers, and finally three fully-connected layers.

After the introduction of AlexNet, new network architectures have mostly increased in depth, i. e. by adding more layers. However, deeper networks are hard to train due to the problem of vanishing or exploding gradients as gradients may increase or decrease exponentially when backpropagated (see section 6.2.2.3) through the network

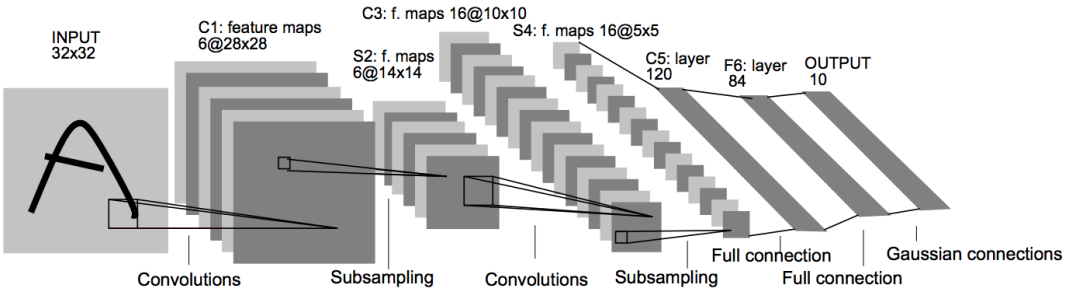


Figure 6.9: LeNet architecture: The LeNet is composed of two convolutional layers, intervened by average-pooling operations, and finished with a fully connected layer. The network was develop to recognise handwritten digits, here an example of recognising the letter "A". Figure courtesy: LeCun et al. [86].

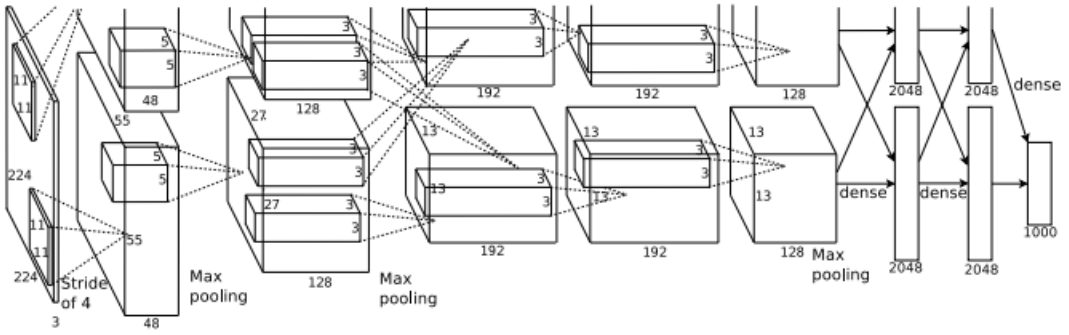


Figure 6.10: AlexNet architecture: The AlexNet consists of eight layers - 5 convolutional layers, intervened by max-pooling layers, and three fully connected layers. Figure courtesy: Krizhevsky et al. [78].

[48]. Furthermore, the degradation of the network's performance has been demonstrated for deeper networks. Degradation describes the observation that the training error deteriorates beyond a certain depth. While degradation was a counter-intuitive finding, as one would expect a performance not worse than the shallower counterpart, approaches to solving this problem have been suggested and led to the introduction of residual networks [57].

Residual networks utilize short-cuts or skip connections, as illustrated in figure 6.11. Instead of learning the function $\mathcal{F}(x)$, the output from a previous layer is added to the output of a deeper layer. In the extreme case of vanishing gradients or redundant functions, an identity mapping is learned and the network is equivalent to its shallower counterpart.

The networks presented so far are compressing the information contained in images to classify them as belonging to a particular class, or to predict some number in regression problems. Image segmentation aims to label each pixel of an image with a corresponding class. Stacking convolutional and pooling operations help the network to learn what can

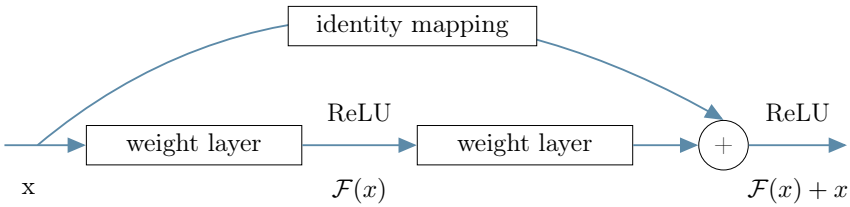


Figure 6.11: Residual block: this is an important building block of residual neural networks. The output of an earlier layer is fed directly into a later layer via a sum. This allows for reducing vanishing gradients, as well as developing deeper networks, without degradation of the training performance (see text).

be seen in the image but loose information on where it can be seen. One way to segment an image is to classify each pixel in the network separately [21].

Another way is to recover the information on where something is present in the image. Modern segmentation networks typically consist of encoding and decoding parts, where the encoding part learns about the "what" and the decoding part recovers the "where" information. A widespread network for segmentation in biomedical imaging has been introduced by Ronneberger et al. [126], which comprises an encoding or contracting part and a decoding or expanding part. As I used this network in my work, I introduce its components in detail. Figure 6.12 illustrates the architecture of the U-Net network. It has approximately 15 million parameters.

The expanding part mirrors the contracting part, which gives it the u-shaped architecture. The contracting path consists of repeated applications of convolutions, (each followed by a ReLU activation) and max-pooling operations. Each pooling layer in the contracting part is replaced by an upsampling or transposed convolution part in the expanding part. There are various methods to conduct an upsampling operation: nearest neighbour interpolation, bilinear or bicubic interpolation. Instead of manually choosing the interpolation method, a U-Net learns how to upsample optimally by using an operation known as transposed convolution. Dumoulin and Visin [34] provide more details on how transposed convolutions work. Feature maps from the contracting part are copied to the expanding part via skip connections. This feeds contextual information into the location information branch, retaining the lost spatial information from pooling operations. A 1×1 convolution in the last layer reduces the number of output channels to the number of labels.

6.2.6 Software tools and libraries

Several open-source deep learning libraries have been developed that provide an efficient GPU-implementation of essential operations in neural networks, allowing users to

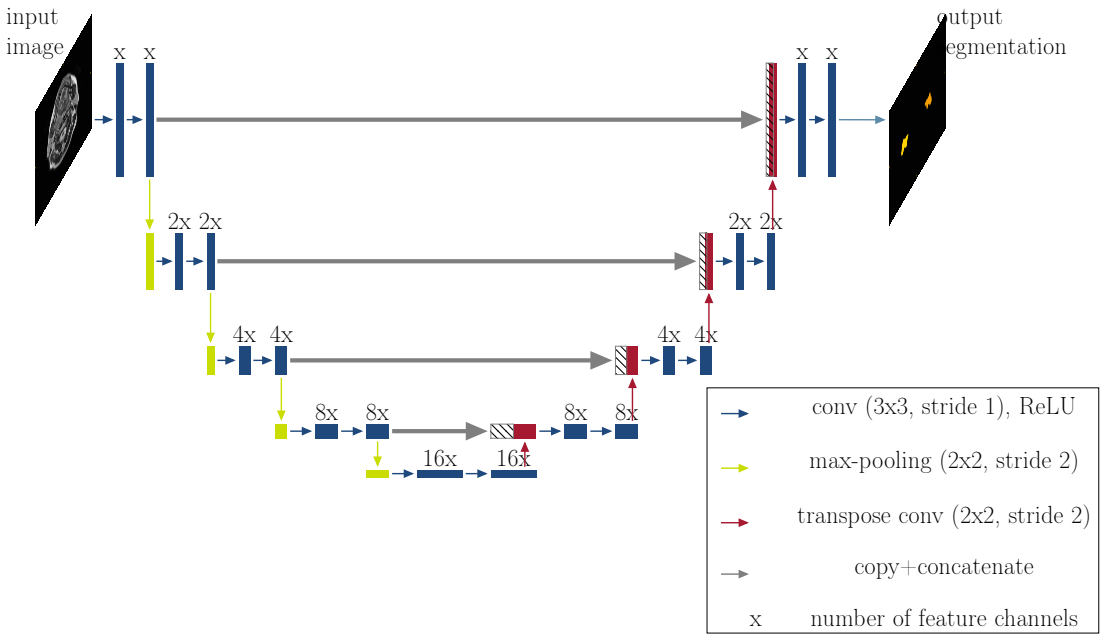


Figure 6.12: U-Net architecture: example for 5 different resolution levels, starting at x feature channels and doubling the feature channels at each level (ending at 16x feature channels at the lowest resolution). Each rectangle corresponds to a feature map. The feature channels are denoted at the top of the rectangles. Striped boxes represent copied feature maps. The coloured arrows denote different operations, as indicated in the legend. Due to the fully convolutional nature of the network, the image input size is flexible.

implement ideas and focus on how to set up a particular network without needing to reimplement the basic operations. The most popular ones include Tensorflow [1], Keras [19] and Pytorch [113]. They all provide application programming interfaces, mostly in the programming language Python.

6.3 General infrastructure

The aim of the presented work in this chapter was to investigate the feasibility of deep learning-based approaches applied to the segmentation of OARs in HNC MR images. We restricted this investigation to the segmentation of the parotid glands. The most prevalent side effect for HNC patients is xerostomia [71, 82, 106], which is caused by radiation damage to the salivary glands. Sparing the parotid glands in an RT treatment could therefore significantly reduce complications [106, 156] and renders an accurate segmentation is crucial. Moreover, parotid glands can vary significantly in shape and size between patients.

I implemented all deep learning approaches in Python (version 3.6) using the open-source libraries Tensorflow (version 1.10.0) [1] and Keras (version 2.2.2) [19]. For pre- and postprocessing of input and output, I further used the python modules numpy

(version 1.14.5), nibabel (version 2.3.0), scikit-learn (version 0.20.0) and scipy (version 1.1.0).

As discussed in section 6.2.1, many hyperparameters can be tuned, rendering it impractical to determine the best ones in a grid search. For this reason, I started from a deep learning-based approach that has been proven to work well in similar applications and restricted myself to adapt only a few of the hyperparameters. Due to the small amount of MR imaging data, I used an 80/20 split into training/validation data for hyperparameter optimisation and validated the developed approaches in a 9-fold cross-validation setting, where in each fold, 8/9 of the data was used for training and 1/9 for testing purposes. This led to the training of 9 networks with a different combination of training and testing data. This was different for the CT imaging data, as more data was available, and hence, I used a 70/10/20 split into training/validation/testing for training these networks. In contrast to the MR training, where multiple networks were trained for each fold in the cross-validation, only one network was trained for the CT network.

For evaluation purposes, I used the geometrical evaluation tools, as described in chapter 4. I did not perform any dosimetric evaluation, as the nature of this work was exploratory. For clinical validation, one would need to follow the workflow as described in chapter 4 and generate treatment plans for each set of auto-generated contours.

6.4 Automated segmentation using a convolutional neural network

6.4.1 Introduction

In the last two years, after I started looking into deep learning approaches, a few studies were investigating the segmentation of ROIs in HNC, all of these applied to CT images [16, 66, 97, 127, 145]. Table 6.1 lists details on these studies.

This work aimed to investigate the feasibility of using CNN-based approaches to segment OARs on HNC MR images, and to determine the drawbacks and necessary consequent steps to counter these drawbacks. For this purpose, I started from a widespread network architecture [126] and investigated different techniques, such as varying the dimensionality of the input (2D, 2.5D, 3D) and following a multi-modality approach. Additionally, I studied the impact of standard deep learning techniques such as the application of dropout, data augmentation and different loss functions. To the best of my knowledge, I was the first to study CNN-based approaches for the segmentation of OARs on HNC MR images.

6.4.2 Materials and Methods

6.4.2.1 Data acquisition and preparation

The 27 T1w and T2w MR images from MD Anderson, together with the manual segmentation of the parotid glands served as input to the networks. For more details on image acquisition parameters and image preprocessing steps, see chapter 3.

6.4.2.2 Neural network specifications

General settings

I chose the U-Net architecture [126] (5 resolution levels, starting at 64 features and ending at 1024 features at the lowest resolution in the bottleneck), as it has proven to be successful in many recent applications for medical imaging. The architecture of a 2D U-Net is illustrated in figure 6.12 on page 110. I used the Adam method [73] to optimise a Dice loss function as defined in equation (6.11) on page 101.

I explored the usage of dropout layers with various probabilities and placements within the architecture of the U-Net. Furthermore, I introduced random data augmentations (translations, rotations and mirroring). Both methods were described in section 6.2.3 as a common strategy to mitigate generalisation problems. As I did not find any increase in the model performance, I decided not to pursue these methods any further for this study.

Table 6.1: An overview of published studies on the application of CNNs to the segmentation of ROIs in HNC.

study	#	modality	input data	approach	regions of interest
[66]	50	CT	2.5D (tri-planar orthogonal patches)	separate CNN per ROI	spinal cord, mandible, parotid glands, submandibular glands, larynx, pharynx, eye globes, optic nerves, optic chiasm
[145]	32	CT	3D (cropped images to patient contour)	multi-organ, shape-constrained U-Net with residual blocks	brainstem, optic chiasm, mandible, optical nerves, parotids, and submandibular glands
[97]	43	CT & MR	2D (orthogonal patches)	6 2D CNNs, then stacked	parotid glands
[16]	200	CT	3D patches	multi-step, adapted 3D U-Net	spinal cord, mandible, parotid glands, oral cavity, brainstem, larynx, oesophagus, submandibular glands and the temporomandibular joints
[127]	157	CT	3D patches	3D U-Net	submandibular glands, parotid glands, larynx, cricopharynx, pharyngeal constrictor muscle, upper oesophageal sphincter, brain stem, oral cavity and oesophagus

In addition, I employed two different loss function: Dice loss and weighted cross-entropy. I did not see any difference between using a Dice or a weighted cross-entropy loss function and therefore decided to use the Dice loss function as opposed to the weighted cross-entropy, this loss function does not include any further hyperparameters.

The output of the U-Net is a segmentation map with values between 0 and 1, providing the likelihood of a structure being present or not (in this case the parotid glands). I chose to set everything above 0.5 to 1 and 0 otherwise.

Applying deep learning to 2D slices only considers image information in 2D planes. 3D volumes consume a large computational memory, and 3D convolutions have high computational costs on the other hand, and therefore need to compromise at other points. As a compromise, I first implemented a 2.5D network, where instead of using 2D slices as inputs, I used three adjacent slices to take into account information from neighbouring slices. As another approach, I chose 3D volumes around the regions of interest, where instead of feeding the full 3D volumes, I restricted the field of view in the axial plane to the region of interest.

In MRI, different sequence parameters are used to manipulate the contrast of the resulting images and highlight regions of interest. To exploit this, I implemented a multi-modality approach, where I used the information from T1w images to guide the segmentation of the T2w images. All approaches are detailed in the following paragraphs and illustrated in figure 6.13.

2D approach

The input data to the 2D U-Net comprised the 2D axial slices of each 3D volumetric imaging dataset. Each 2D slice was fed into the network, leading to 24 times 30 slices for the training dataset in each fold. I trained the network for 60 epochs with a learning rate of 10^{-4}. The network is illustrated in the top part of figure 6.13.

Adjacent slices: 2.5D approach

In the 2.5D network, I fed the two adjacent slices as two additional input channels into the network, as illustrated in the central part of figure 6.13. The network was trained for 60 epochs and a learning rate of 10^{-4}.

3D approach as a two-step process

Due to limitations in GPU memory and to focus the network on the relevant regions of interest, I used 3D patches of 128x128x16 voxels (1x1x4 mm^3), centred at the centre of mass of each parotid gland. To obtain the centre of mass for the testing data, I

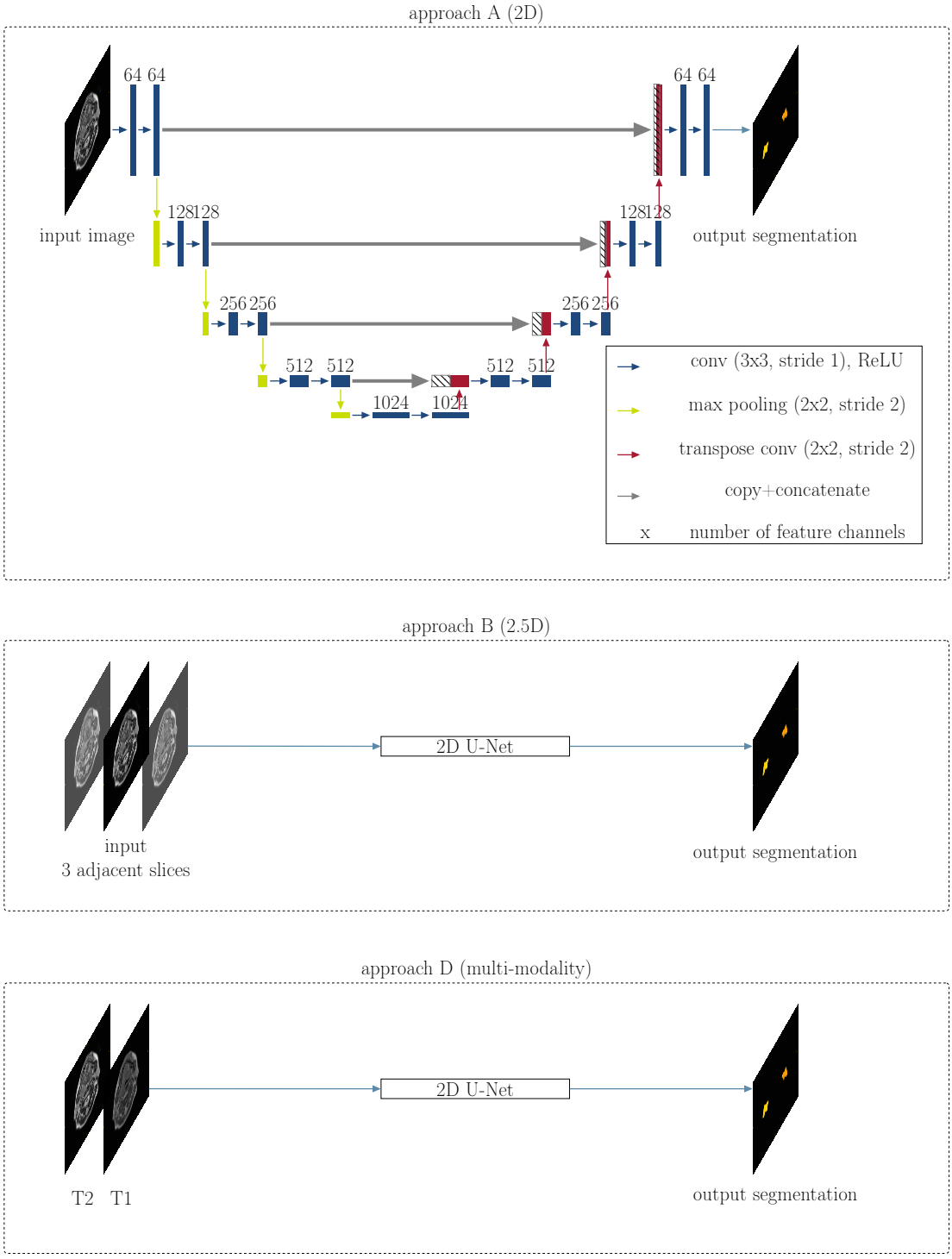


Figure 6.13: This figure illustrates methods A, B and D, all using the same network architecture (2D U-Net with 5 resolution levels, starting at 64 features and ending at 1024 features at the lowest resolution in the bottleneck). Each rectangle corresponds to a feature map. The feature channels are denoted at the top of the rectangles. Striped boxes represent copied feature maps. The coloured arrows denote the different operations as indicated in the legend. The output for all three approaches is a 2D segmentation map.

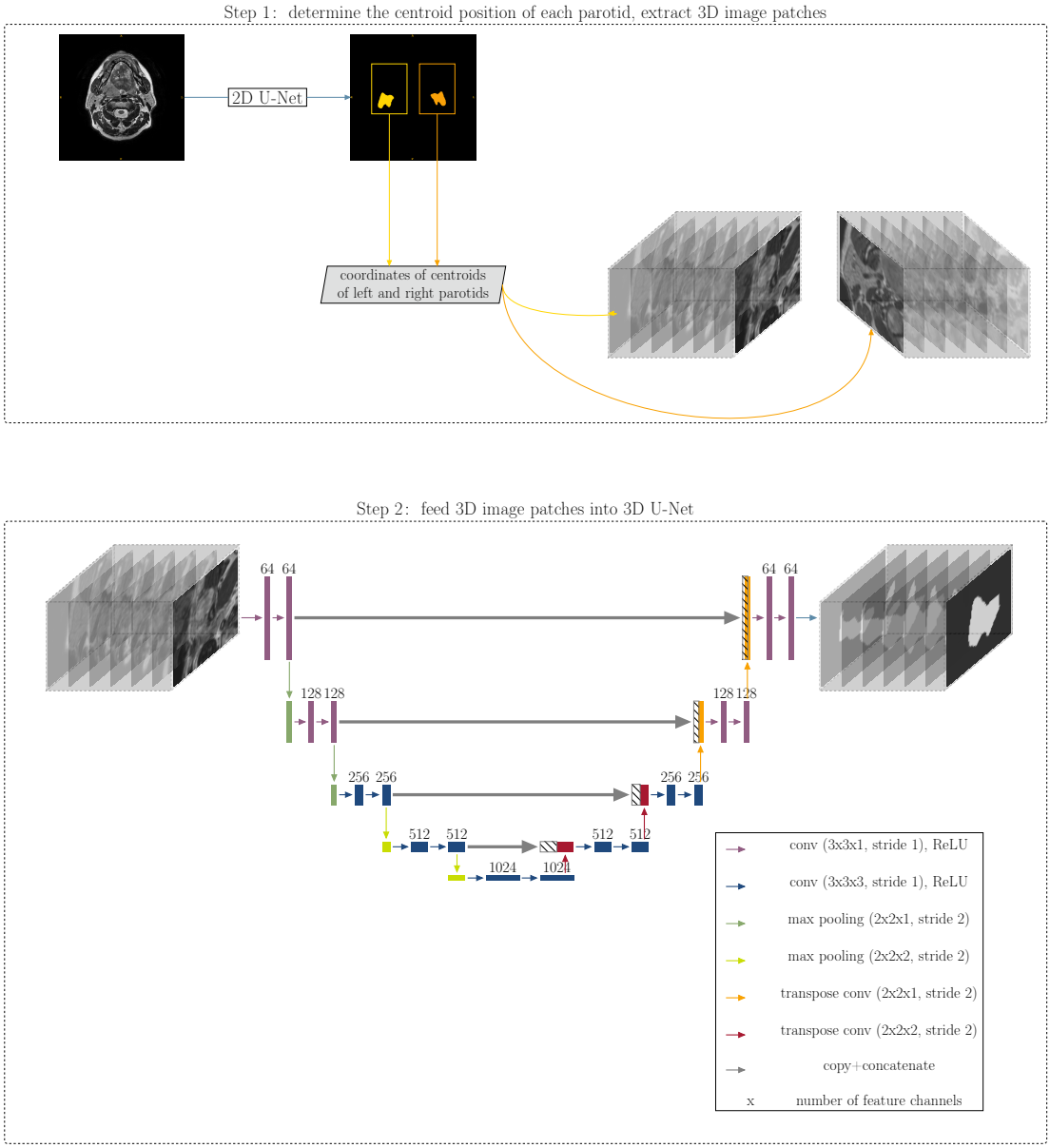


Figure 6.14: 3D U-Net architecture: In the first step, a rough segmentation is performed (using a 2D U-Net) to determine the location of the left and right parotid glands. From this, 3D patches centred at the centroid coordinate of each parotid are extracted and fed into a 3D U-Net, as shown below in step 2.

performed a rough segmentation using a 2D U-Net as a first step. Figure 6.14 illustrates the workflow of this approach. I used 3D convolutions instead of the 2D convolutions in the previous approaches. Due to highly anisotropic voxels, I performed the convolution in the first two levels with anisotropic convolutional kernels (3x3x1 voxels), as well as max-pooling kernels (2x2x1 voxels). In all other levels, I used isotropic kernels (convolutional: 3x3x3 voxels, max-pooling: 2x2x2 voxels).

I trained the network for 60 epochs with a learning rate of 10^{-4} in the first step,

using downsampled images (128x128x30 voxels, 2x2x4 mm^3 voxel size). The 3D network was trained for 40 epochs with a learning rate of 10^{-5}.

Multi-modality

To use complementary information on the regions of interest, I explored the usage of other MRI contrasts to guide the segmentation. For this, I registered corresponding T1w images to the T2w images and fed them into the network as two input channels. The network is illustrated in the bottom part of figure 6.13. I trained the network for 60 epochs with a learning rate of 10^{-4}.

6.4.2.3 Computation time

Run time was determined for programme execution on a single Tesla V100 with 16 GB VRAM. I calculated the mean and average values from the individual run times of the 9-fold cross-validation. The inference time is stated per patient.

6.4.2.4 Geometric evaluation

To evaluate the geometrical accuracy of the U-Net approach, I chose a 9-fold cross-validation, where for each fold, 3 MR images comprised the test data and the remaining 24 MR images the training data. Geometric differences between the ground truth and the CNN-derived segmentations were evaluated by calculating the 3D DSC, HD and MSD, as described in chapter 4. Due to their symmetry, I simultaneously segmented both parotid glands and divided them into the left and right part in a post-processing step.

6.4.3 Results

Figure 6.15 provides examples of the auto-generated contours overlayed with the manual contours, comparing all four approaches (2D, 2.5D, 3D and multi-modality). The rows represent the four approaches, whereas the columns are four different examples of patients. The contours generally follow the manual contours well. The auto-segmentation of the patient in the last column is a negative example for approaches C (3D) and D (multi-modality), where for approach C, most of the left parotid is left out, being entirely missed out for approach D. The patient in the second column is an example where it is difficult to tell whether the auto-generated or manual contours are the more accurate ones: the part included additionally for the right parotid in the auto-generated contours in comparison to the manual ones might as well be part of the parotid gland which might have been missed out in the manual procedure.

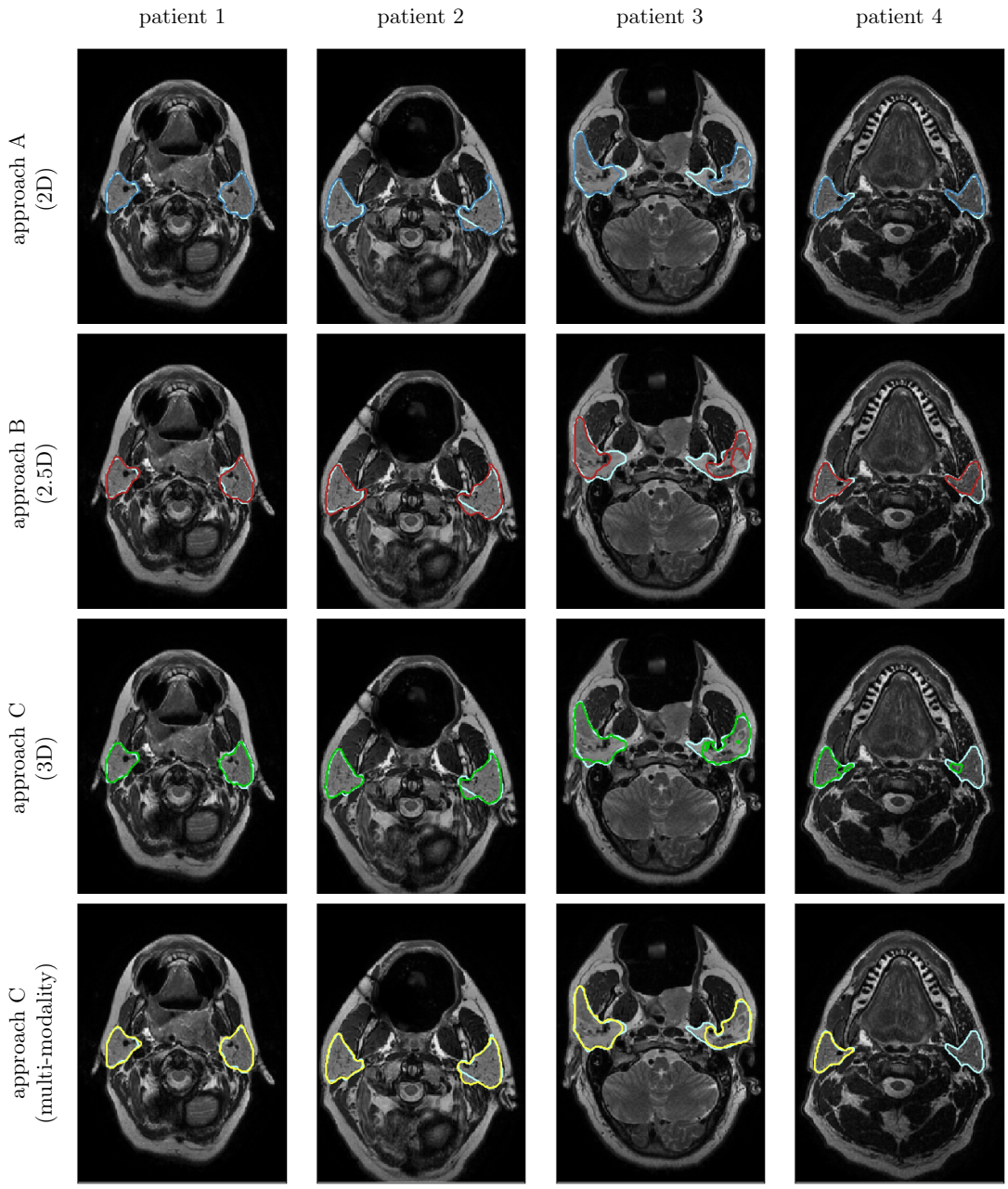


Figure 6.15: This figure shows in each column a typical example comparing the manually segmented parotids (light blue) to approach A (2D, dark blue, first row), approach B (2.5D, red, second row), approach C (3D, green, third row) and approach D (multi-modality, yellow, fourth row), respectively. Each example originates from a different patient image.

Table 6.2: Training and inference times for the four CNN-based approaches (2D, 2.5D, 3D and multi-modality). Training times are the average of 9 folds, whereas inference times are the average of 27 patients.

method	t_{train} [min]	t_{inf} [s]
2D (A)	24.65 ± 0.22	0.90 ± 0.06
2.5D (B)	25.35 ± 0.25	0.88 ± 0.04
3D (C)	376.28 ± 75.85	1.54 ± 0.44
multi-modality (D)	25.56 ± 0.21	0.90 ± 0.04

Table 6.3: Evaluation of geometric accuracy of auto-segmenting the left and right parotid gland, comparing different methods using the U-Net.

ROI	method	$\overline{DSC}$	$\overline{HD}$ [mm]	$\overline{MSD}$ [mm]
right parotid	approach A (2D)	0.84 ± 0.07	16.75 ± 10.37	2.00 ± 1.46
	approach B (2.5D)	0.81 ± 0.07	19.15 ± 10.08	2.27 ± 1.38
	approach C (3D)	0.83 ± 0.06	16.15 ± 11.64	1.85 ± 1.27
	approach D (multi-modality)	0.82 ± 0.09	15.69 ± 9.96	2.04 ± 1.45
	inter-observer variability	0.84 ± 0.04	10.76 ± 4.35	1.40 ± 0.45
left parotid	approach A (2D)	0.85 ± 0.08	15.19 ± 8.09	1.63 ± 1.27
	approach B (2.5D)	0.83 ± 0.05	15.89 ± 5.91	1.83 ± 0.77
	approach C (3D)	0.80 ± 0.12	16.36 ± 6.78	1.90 ± 1.08
	approach D (multi-modality)	0.79 ± 0.17	16.85 ± 9.97	2.38 ± 2.19
	inter-observer variability	0.83 ± 0.04	10.94 ± 3.75	1.59 ± 0.63

6.4.3.1 Computation time

Training and inference times can be found in table 6.2. Approaches A, B and D all had a training time of approximately 25 minutes and an inference time of less than 1 second. In comparison, approach C (3D) took on average 376 minutes to train with an average inference time of 1.54 seconds.

6.4.3.2 Geometric evaluation

Figure 6.16 shows the boxplots of all methods and ROIs employed in this study. Table 6.3 lists mean values and standard deviations for all approaches and ROIs. For comparison, I included the inter-observer variability.

With a mean DSC larger than 0.8 and a mean MSD smaller than 2.5 mm, all approaches achieved a geometric accuracy which was in the same ballpark as the inter-observer variability. There were no significant differences in the average performance of any of the approaches. While the 3D approach tended to have a smaller variance of accuracies and less severe outliers, the multi-modality approach added some uncertainty in comparison to the plain 2D approach. Adding adjacent slices did not improve the segmentation accuracy.

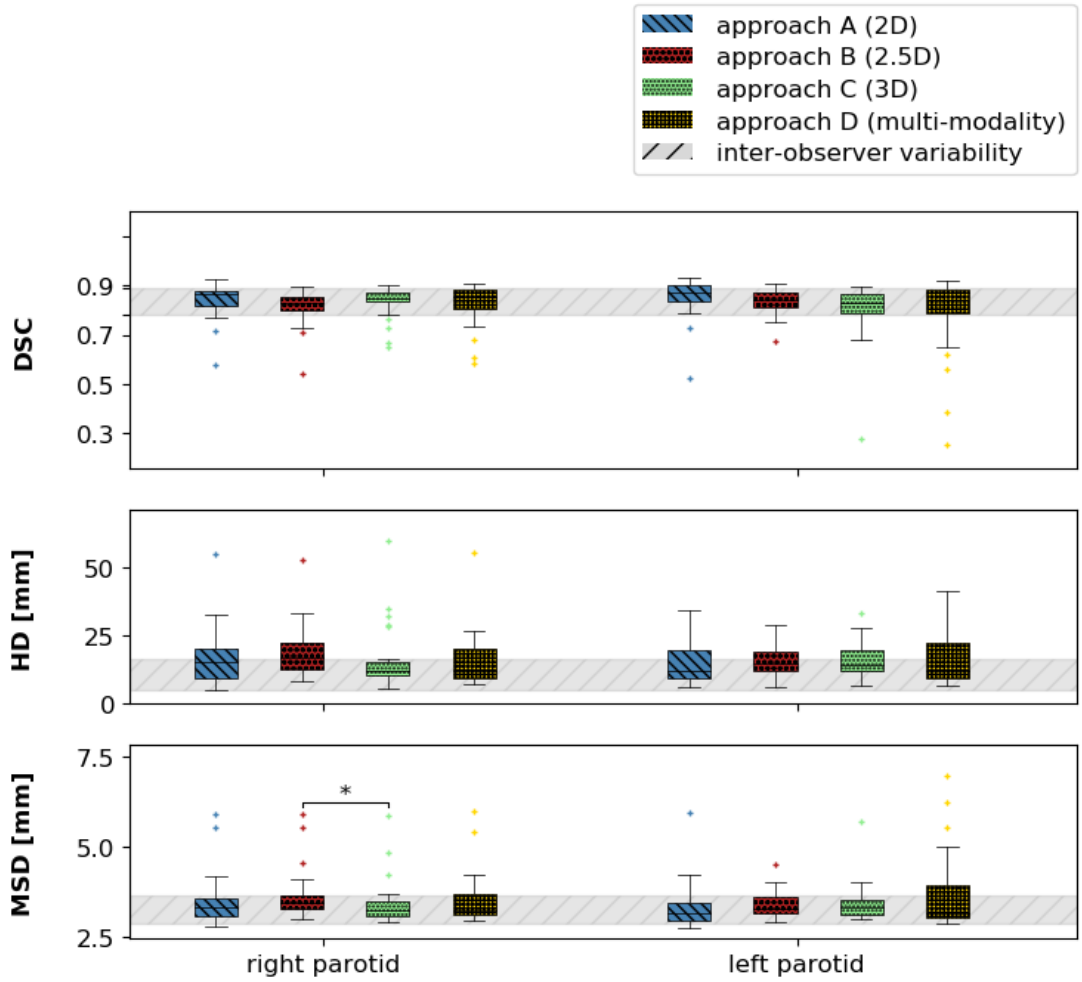


Figure 6.16: Comparison of different CNN-based approaches: Boxplots of, from top from bottom, the DSC, the MSD and the HD for both parotid glands (x-axis) and all automated segmentation approaches (A in blue, B in red, C in green, D in yellow). I also included the inter-observer variability (grey). Stars indicate statistical significance ($p<0.05/3$).

6.4.4 Discussion

I investigated the application of CNN-based approaches to the segmentation of the parotid glands and benchmarked them with the inter-observer variability, as well as the multi-atlas based approach, introduced in chapter 5. I furthermore compared to published studies in similar applications. To my knowledge, this was the first study to demonstrate that auto-contouring parotid glands on MR images using CNN-based approaches can achieve an accuracy comparable to the inter-observer variability.

6.4.4.1 Computation time

Compared to a conventional commercial method, for instance, atlas-based segmentation, the computation time, especially the inference time, was much faster for the deep learning

approaches. While the computation time for atlas-based approaches can be further reduced, by, e. g. parallel computations, they still require minutes in comparison to a second for deep learning-based approaches. This renders deep learning-based approaches feasible in an adaptive clinical workflow, such as envisioned with the MR-Linac.

6.4.4.2 Geometric evaluation

I compared four different approaches: approach A (2D), approach B (2.5D), approach C (3D) and approach D (multi-modality). All approaches achieved an accuracy that was of the same size as the inter-observer variability with a DSC larger than 0.8 and an MSD smaller than 2.5 mm. There were no clear benefits of using one or the other for the data in this study. Since all approaches achieved an accuracy that was already in the same range as the inter-observer variability, it was challenging to detect differences in the approaches. As found before in the atlas-based approach, the HD was sensitive to local errors and did not reflect the overall accuracy well.

Adding further information, such as done in the multi-modality approach (approach D), did not increase the accuracy but instead led to more outliers. This may be attributed to the finding that the T2w and corresponding T1w images may not have been correctly aligned and therefore led to inconsistent information for the CNN. While one might have expected an improved performance by using additional information from the adjacent slices, this approach also did not lead to an increase in performance. The 3D approach decreased the general spread of values, but the overall accuracy did not improve in comparison to the 2D approach. Due to memory-related restrictions of the hardware, I followed a two-step approach by first finding an approximated bounding box close to the parotids and then using a 3D patch with the centroid at this bounding box. This requires the first step to being sufficiently accurate as otherwise the parotid might be missed in the extracted 3D patch. Another approach would be to use overlapping 3D patches at random locations of the images without the need for the first step. I did not explore this approach in this study due to time limitations.

The accuracy of the deep learning-based approaches was comparable to the atlas-based segmentation (DSC: 0.85±0.04, MSD: 1.65±1.08 mm, HD: 16.41±12.10 mm, see chapter 5).

Table 6.4 lists mean reported geometric measures for a comparison of my results to published studies on CNN-based approaches applied to the segmentation of the parotid glands. All of the reported studies used CT images (with one study adding MR images in a multi-modality approach). In particular, the 2D approach ranked highest compared to published studies. The training data of published studies were of similar size as mine except for two studies [16, 127]. I expect that with more training data, one can increase

Table 6.4: Comparison of geometric evaluation for the approaches developed in this study to published studies

ROI	DSC	MSD[mm]	mod	#	study
right	0.84 ± 0.07	2.00 ± 1.46	MR	27	approach A (2D)
parotid	0.81 ± 0.07	2.27 ± 1.38	MR	27	approach B (2.5D)
	0.83 ± 0.06	1.85 ± 1.27	MR	27	approach C (3D)
	0.82 ± 0.09	2.04 ± 1.45	MR	27	approach D (multi)
	0.85 ± 0.04	1.65 ± 1.08	MR	27	atlas-based (chapter 5)
	0.84 ± 0.02	1.12 ± 0.56	CT	32	Tong et al. [145]
	0.79	1.57	CT+MR	43	Močnik et al. [97]
	0.78 ± 0.05	-	CT	50	Ibragimov and Xing [66]
	0.86 ± 0.05	-	CT	200	Chan et al. [16]
	0.83 ± 0.02	-	CT	157	Rooij et al. [127]
left	0.85 ± 0.08	1.63 ± 1.27	MR	27	approach A (2D)
parotid	0.83 ± 0.05	1.83 ± 0.77	MR	27	approach B (2.5D)
	0.80 ± 0.12	1.90 ± 1.08	MR	27	approach C (3D)
	0.79 ± 0.17	2.38 ± 2.19	MR	27	approach D (multi)
	0.83 ± 0.06	1.65 ± 1.57	MR	27	atlas-based (chapter 5)
	0.84 ± 0.03	0.96 ± 0.34	CT	32	Tong et al. [145]
	0.79	1.57	CT+MR	43	Močnik et al. [97]
	0.77 ± 0.06	-	CT	50	Ibragimov and Xing [66]
	0.85 ± 0.03	-	CT	200	Chan et al. [16]
	0.83 ± 0.03	-	CT	157	Rooij et al. [127]

the generalisability of the trained model and reduce the number of outliers.

6.4.4.3 Limitations and future work

In this study, I focused on the segmentation of the parotid glands. While the segmentation of these OARs is crucial, the methodology from this study is not limited to this specific organ and I expect that one can easily transfer the approach to the segmentation of other ROIs. Out of the many OARs relevant for the treatment planning of RT in HNC, the parotid glands are the most challenging ones to contour. Furthermore, including multiple OARs as different labels in one network could improve the geometric accuracy as more information would be available to the CNN and outliers may be reduced. Therefore I believe that automatic contouring of other OARs should perform similar or better than of the parotids. However, including these data was beyond the scope of this thesis.

A further limitation of this study was the small number of available training data. To account for variations found between different patients and even of the same patients on different days, one would need a larger database that incorporates these substantial variations. Moreover, the algorithm may fail when applied to images with different

settings compared to the ones trained on, for instance, different contrast settings or resolutions. See a more thorough evaluation of this issue in chapter 7. Potential solutions to this problem using deep learning based methods are addressed in the following sections.

6.4.5 Conclusion

This study demonstrated the enormous potential for the application of CNNs to segment ROIs for RT treatment planning purposes with the accuracy comparable to conventional approaches such as atlas-based segmentation. In comparison to atlas-based segmentation methods, the computation time was much shorter (sub-second compared to minutes or hours) with a simple 2D approach. These short computation times render deep learning-based approaches suitable for online treatment planning workflows. A limiting factor of this study was the small amount of training data. I address potential solutions to this problem in the following sections.

6.5 Transfer learning from CT images

6.5.1 Motivation

Deep learning faces a bottleneck with the lack of the availability of sufficient annotated training data. While it is still unclear how many training examples these algorithms will require, it is evident that the generalisability increases with an increasing variety in the training data. However, as discussed before, a sufficient amount of representative annotated training data is usually not available. Images are acquired with different scanners or scan protocols. Manual annotation of data usually requires expert knowledge and is a time-consuming and error prone process. Moreover, some labels needed for the training of such algorithms may not be necessary for the clinical routine and therefore often need to be created only for the underlying research studies.

Recently, it has been shown that a technique called transfer learning can significantly improve the performance of deep learning models suffering from limited training data by leveraging existing data of related problems [18]. In transfer learning, a model, which was constructed to learn a task A from a large dataset, is applied to boost the performance of a model to learn a task B on a typically much smaller dataset. The model designed for task A is called a pre-trained model. Transfer learning originates from the finding that many deep learning tasks share common elements which can be recycled in a new task [166]. An analogy in the real world is the finding that it is generally easier for someone to learn a third language compared to the second language, due to shared vocabulary and general transferable logic of languages. In a clinical scenario, a clinician may find it easier to annotate an ROI for a new imaging modality, e. g. MRI, with his or her experience on annotating another imaging modality, e. g. CT.

Litjens et al. [92] and Cheplygina [18] provide reviews on the application of transfer learning techniques to medical image processing. Published studies have used both pre-training on natural images and related medical images. However, due to the availability of more training data in the field of natural image processing, the former has been used more frequently until now. A drawback of using pre-trained classification models on natural images is that they are often restricted to fixed image sizes and comprise unnecessary model complexities, rendering a direct application to medical imaging challenging.

With MRI only starting to be routinely used in RT, as well as the existing variety in scanners and scanning protocols, the wealth of training data as seen for natural images is generally not available for consistent RT-specific MR images. On the other hand, there is a wealth of publicly available, annotated CT images, for instance, in the

Cancer Imaging Archive [22]. In this section, I exploited this finding to train a 2D CNN using annotated CT data and subsequently fine-tuned the learned model by training the network on MR images. To the best of my knowledge, the only two published studies using transfer learning for medical image segmentation are [17, 148], where only the latter applied transfer learning to a CNN. To my knowledge, this study was the first to retrain CT networks for automatically contouring the parotid glands on MR images in HNC patients.

6.5.2 Materials and Methods

6.5.2.1 Data preparation

The 27 T2w MR images, as well as the public database, including 202 CT images, served as imaging database. For more details on acquisition and preprocessing techniques, see chapter 3.

6.5.2.2 Segmentation approach

I chose a 2D U-Net [126], as introduced in the previous section 6.4, with 5 levels, as well as 64 feature channels in the first level and 1024 channels in the last level. For the baseline approach (training with the CT data), as well as the transfer learning approaches, I optimised a Dice loss function with the Adam optimiser.

Pre-training using CT data

I trained the network with the CT data for 60 epochs, with an initial learning rate of 10^{-4}. I gradually reduced the learning rate throughout the training process by monitoring the validation loss: if the validation loss did not improve after 10 epochs, the learning rate was automatically reduced to half its size, down to a minimum of 10^{-8} .

Transfer learning

There are two general approaches to transfer learning:

(1) training a CNN with source data and extracting specific features from this network to train a classifier for the target data.

(2) training a CNN with source data and training the same CNN with the target data using the pre-trained network as initialisation.

In this work, I focused on the second approach. During the training of the CNN using the target data, it is common to keep some of the weights fixed, also known as 'frozen' and

only allow for updates of some weights. Typically in transfer learning, the pre-trained weights from the shallower levels (earlier in the network chain) are kept unchanged and only deeper levels (later in the network chain) are adapted to the problem at hand. This procedure stems from the widely accepted intuition that shallower levels refer to low-level features, such as edges, and develop into more complex shapes in the deeper levels [85]. Another approach is to only initialise the network with the pre-trained weights and perform a fine-tuning of all the layers on the new dataset [18].

I aimed to transfer information about the variety of shapes, as well as the locations of the parotid glands within the head, from the CT to the MR segmentation problem. For the U-Net architecture, it was non-trivial to know where this information was exactly stored and hence which weights to fix. According to the widely accepted intuition of how more complex structures are learned in deeper layers, and the architecture of the U-Net, the desired information on shape and appearance is likely contained in the encoding part with more complex structures towards the bottleneck. The "where" information is restored from the condensed information of the bottleneck in the decoding part.

In this work, I initialised the network's weights with the pre-trained weights from the CT network. As it is unclear where precisely a U-net stores the relevant information for transfer learning, I implemented three educated guesses for this purpose:

(1) train all layers (all trainable, approach A)

(2) freeze the encoding path (encoder fixed, approach B)

(3) freeze layers around the bottleneck (bottleneck fixed, approach C)

By initialising the weights with the CT network in approach A (all trainable), one would have a "warm" start in the optimisation process, which should be superior to a random initialisation. Retraining all layers enabled the possibility to adjust all weights according to the new images.

Approach B (encoder fixed) is typically done in transfer learning. The intuition is that the concept of edges and simple shapes, which are thought to be described by the encoder, is common to all images and can hence be transferred from one to another application.

Approach C (bottleneck fixed) built on the likelihood that the complexity of shapes for the parotid glands is "stored" in these layers. As I wanted to transfer this knowledge to the MR application, I kept the weights in the bottleneck fixed in this approach.

Figure 6.17 illustrates all of the approaches. For all three approaches, I trained the network for 30 epochs with a learning rate of 10^{-6}.

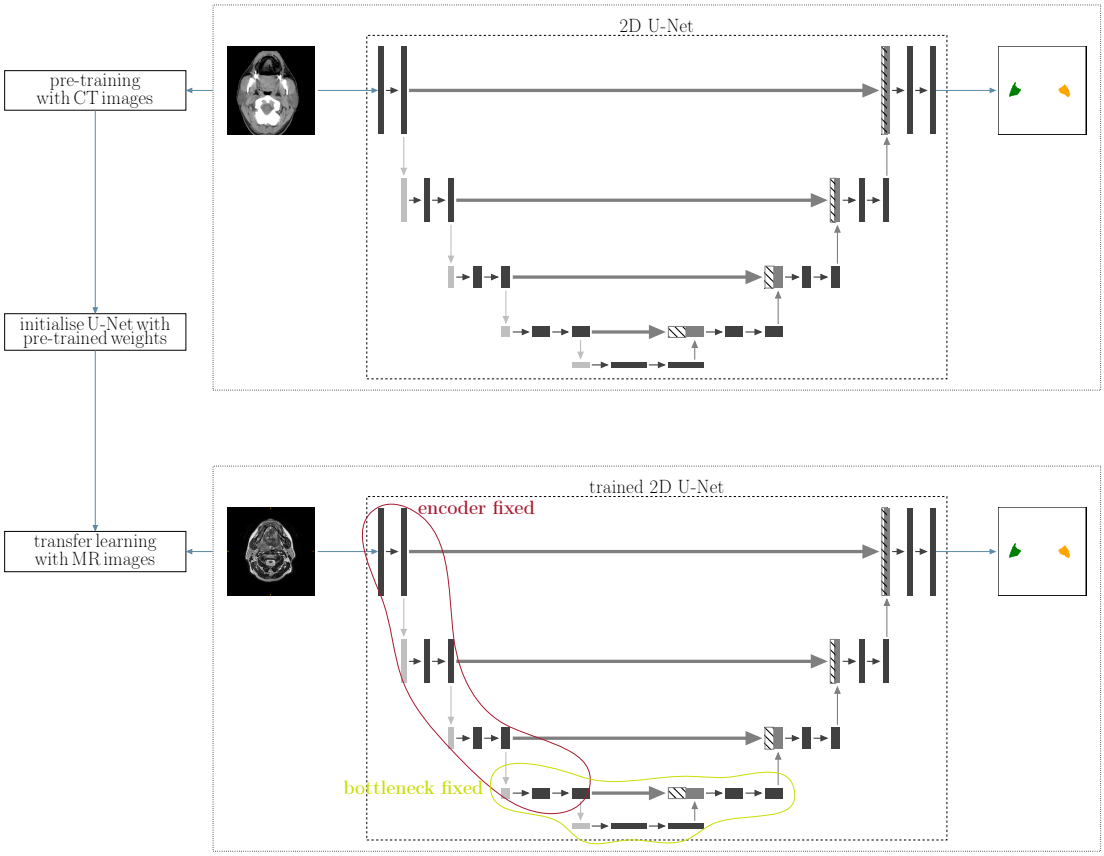


Figure 6.17: Transfer learning workflow: First the network was trained using the CT data in (1). Then this network was used with its pre-trained weights (2) as initialisation and trained on the MR data in (3). I employed three different strategies: re-training all layers, freezing the encoding part (red line) and freezing the bottleneck (green line).

6.5.2.3 Computation time

The run time was determined for programme execution on a single Tesla V100 with 16 GB VRAM. I calculated the mean and standard deviations of multiple runs (9-fold cross-validation). All inference times are stated per patient.

6.5.2.4 Geometric evaluation

I randomly split the 202 CT scans into 70% training, 10% validation and 20% testing data. The validation data were only used to choose the best hyperparameters, whereas the testing data were never seen by the network during the training phase and only used to evaluate the final performance.

To evaluate the accuracy of the transfer learning approach for the segmentation of the 27 MR images, I performed a 9-fold cross-validation (for each fold 24 patients for training, 3 for testing the network). I evaluated the geometric performance by calculating geometric differences between the manual and the CNN-derived segmentations with the

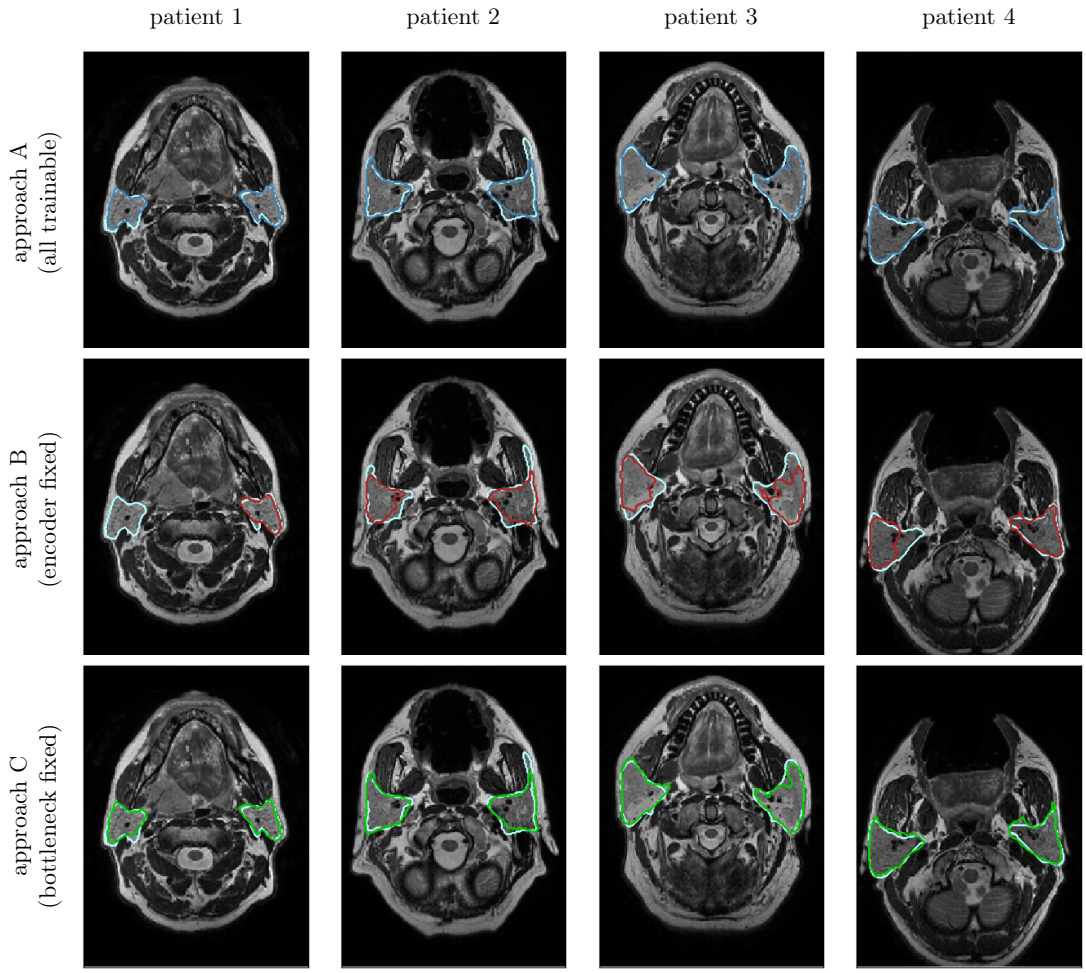


Figure 6.18: This figure shows in each column a typical example comparing the manually segmented parotids (light blue) to approach A (all trainable, dark blue, first row), approach B (encoder fixed, red, second row) and approach C (bottleneck fixed, green, third row), respectively. Each example originates from a different patient image.

original image resolution, using the DSC, the HD and the MSD as geometric metrics. I furthermore compared its performance to the inter-observer variability, determined by comparing manual segmentations from three different experts, as described in chapter 4 and published in [72].

6.5.3 Results

Figure 6.18 provides examples of 4 different patients, comparing all three transfer learning methods. While approach A and approach C (top and bottom row) generally follow the manual contours well, approach B (encoder fixed) misses out on parts of the parotids and in some cases even misses out on the full parotid (see the first column in the second row).

Table 6.5: This table lists training and inference times for all transfer learning approaches. Training times were averaged over the 9-fold cross-validation. Inference times were averaged over all 27 patients.

method	t_{train} [min]	t_{inf} [s]
all trainable (A)	12.40 ± 0.12	0.97 ± 0.16
encoder fixed (B)	10.21 ± 0.02	0.93 ± 0.13
bottleneck fixed (C)	11.40 ± 0.12	0.97 ± 0.15

Table 6.6: Evaluation of geometric accuracy of auto-segmenting the left and right parotid gland, comparing three different transfer learning approaches: all trainable (approach A), encoder fixed (approach B) and bottleneck fixed (approach C). As a benchmark, I also include the geometric accuracy of the CT-trained network, the inter-observer variability (see chapter 4), as well as the training from scratch of the MR network from section 6.4.

ROI	method	$\overline{DSC}$	$\overline{HD}$ [mm]	$\overline{MSD}$ [mm]
right parotid	all trainable (A)	0.79 ± 0.06	18.88 ± 8.96	2.45 ± 1.08
	encoder fixed (B)	0.52 ± 0.15	28.08 ± 14.47	5.44 ± 2.87
	bottleneck fixed (C)	0.80 ± 0.05	19.05 ± 9.20	2.38 ± 1.06
	CT only	0.81 ± 0.07	13.01 ± 5.61	1.87 ± 0.84
	from scratch (MR)	0.84 ± 0.07	16.75 ± 10.37	2.00 ± 1.46
	inter-observer variability	0.84 ± 0.04	10.76 ± 4.35	1.40 ± 0.45
left parotid	all trainable (A)	0.80 ± 0.06	17.79 ± 7.01	2.17 ± 0.98
	encoder fixed (B)	0.70 ± 0.12	19.14 ± 7.65	2.96 ± 1.51
	bottleneck fixed (C)	0.80 ± 0.07	17.93 ± 7.50	2.19 ± 1.01
	CT only	0.82 ± 0.05	12.98 ± 5.15	1.74 ± 0.53
	from scratch (MR)	0.85 ± 0.08	15.19 ± 8.09	1.63 ± 1.27
	inter-observer variability	0.83 ± 0.04	10.94 ± 3.75	1.59 ± 0.63

6.5.3.1 Computation time

Table 6.5 lists training and inference times for the three different transfer learning approaches. While the training time increased slightly with a smaller number of fixed layers, differences were small, in the order of a minute, for all approaches. The overall training time was in the order of 10-12 minutes. The inference time did not differ substantially between the three approaches and was approximately 1 second.

6.5.3.2 Geometric evaluation

Figure 6.19 illustrates the boxplots of the DSC, HD and MSD of the three transfer learning approaches for both parotid glands. Mean and standard deviations are provided in table 6.6. As a reference, I also included the accuracy of the trained CT network, the MR training from scratch, as well as the inter-observer variability.

With a mean DSC of 0.8, mean HD of 19 mm and mean MSD of 2.4 mm, approaches A (retraining all layers) and C (keeping the weights of the bottleneck fixed) achieved a similar accuracy. While the accuracy stayed below the inter-observer variability, as

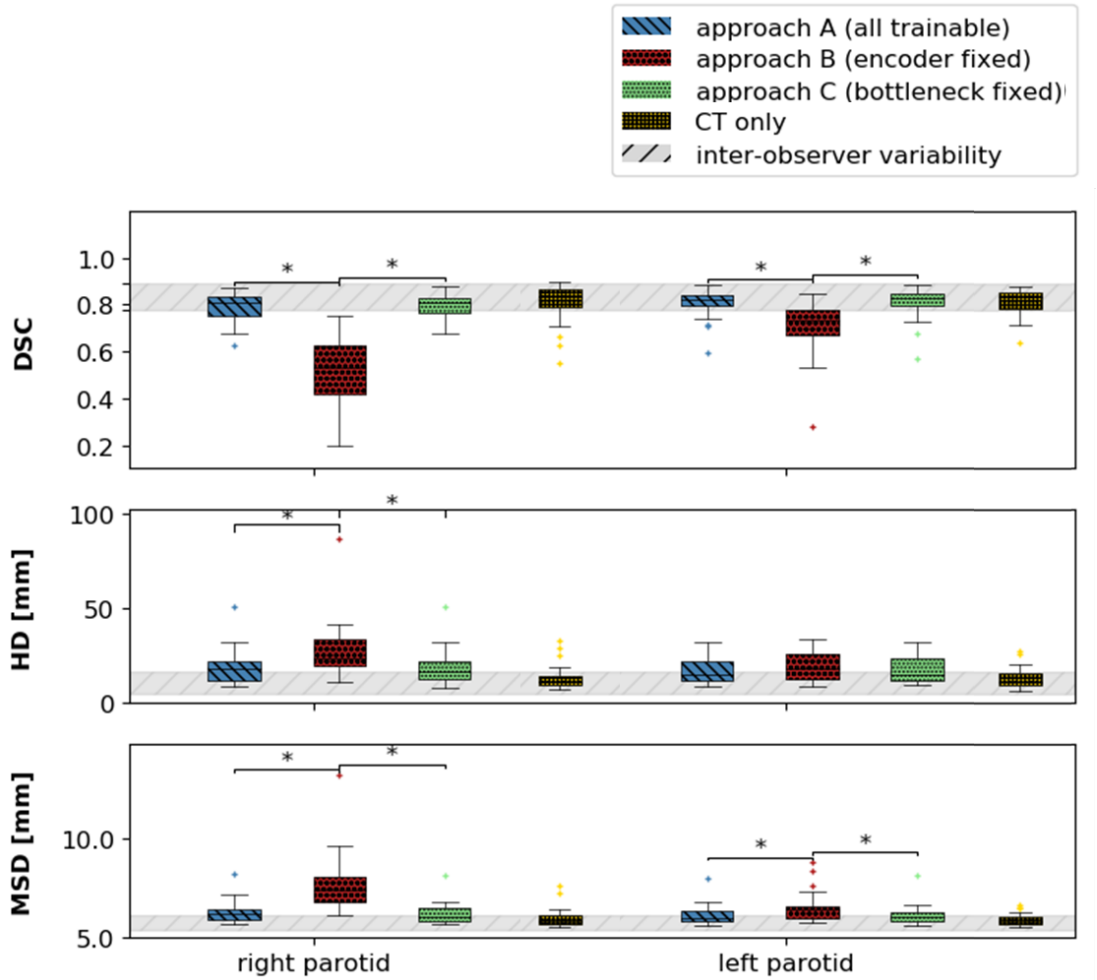


Figure 6.19: Boxplots of the DSC, the MSD and the HD for both parotid glands (x-axis) and all automated segmentation approaches (A in blue, B in red, C in green). As a benchmark, I included the accuracy of the source network (CT only, in yellow). I also show the inter-observer variability (grey). Stars indicate statistical significance ($p<0.05/3$).

well as training the network from scratch, these values are included within one SD of all approaches. Approach B (keeping the encoder fixed) had a worse performance, in particular for the right parotid (mean DSC=0.5, mean HD=28.1mm and mean MSD=5.4 mm).

6.5.4 Discussion

I investigated the application of transfer learning approaches for a 2D U-Net, where I pre-trained with CT images (202 in total) and re-trained with MR images (27 in total). Unlike in typical transfer learning applications, I did not merely want to transfer the ability to detect edges and simple shapes. Instead, I aimed to transfer the gained knowledge about the variety of shapes and locations of the parotid glands from the network trained

(a) Texture image
81.4% **Indian elephant**
10.3% indri
8.2% black swan

(b) Content image
71.1% **tabby cat**
17.3% grey fox
3.3% Siamese cat

(c) Texture-shape cue conflict
63.9% **Indian elephant**
26.4% indri
9.6% black swan

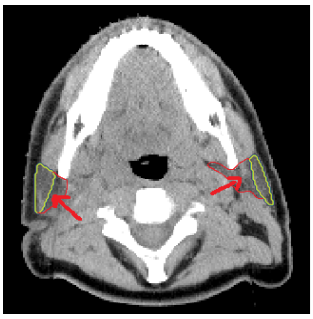
Figure 6.20: This figure demonstrates a case where deep learning went wrong and was fooled by an elephant-skin-like texture overlayed over a cat image to be classified as elephant. This is an example supporting the texture-based picture of how networks learn. Figure extracted from Geirhos et al. [47].

on CT images to MR images. For this purpose, I explored different transfer learning strategies, varying the number and location of "frozen" layers (weights fixed). To the best of my knowledge, I was the first to re-train CT networks to automatically contour the parotid glands on MR images of HNC patients. I showed that it was indeed possible to apply transfer learning to this application, however, it was challenging to determine where the desired information was stored in the networks.

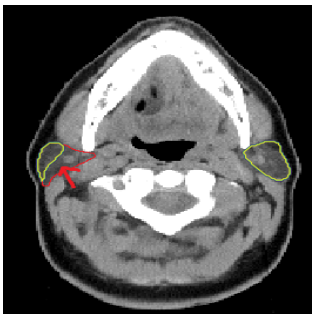
Deep learning research is currently still mostly a grey box where it is not entirely clear what the network learns in which layer. As mentioned at the beginning of this section, it is a commonly accepted intuition that the shallower levels refer to low-level features, such as edges, and develop into more complex shapes in the deeper levels [85].

This shape-based intuition has been contradicted by recent research on CNN features referring to texture based learning. This was, for instance, evident in a study where a network would classify an object with the shape of a cat but an elephant-skin-like texture as an elephant [47]. An illustration of this failure is provided in figure 6.20.

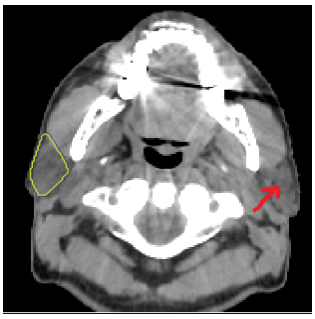
This finding supports the hypothesis that texture-based features are more important than shape-based features. I hypothesised that information about the contrast is most likely learnt in the encoding part of the network. This hypothesis is supported by the significantly worse performance of approach B, in particular for the right parotid. There was a large difference in the geometric measures compared to the other two approaches (Δ(DSC): 0.10-0.28, Δ(HD): 2-10 mm, Δ(MSD): 0.80-3.06 mm). In this approach, I kept the pre-trained weights of the encoder fixed and only re-trained the decoder. I believe that the poor accuracy was due to the difference in contrast between CT and



parotid contours only circumvent half the parotids



right parotid only encloses half the parotid



contour of left parotid fully missing in this slice

Figure 6.21: This figure illustrates three typical example of poor quality CT contours, where the contours do not enclose the full parotids (left and right parotid in the first column, right parotid in the second column) or are fully missing (left parotid in the last column).

MR images, which was most likely learnt in the encoding part of the network. Hence, keeping those weights fixed to the CT-initialised ones may not allow for the network to detect MR-specific features in the transfer learning. Furthermore, the MR images provide more details on soft-tissue contrast. This information was not present in the CT images and hence was not learned in the pre-trained network.

There was no substantial difference between keeping the weights of the bottleneck fixed (approach C) or re-training all layers (approach A). This finding supports the hypothesis that the variety of shapes is stored in the bottleneck of the network and therefore does not need to be adapted to the MR images. Additionally, the network might "forget" about the different shapes of the structure of interest that it has learned in the pre-training when there are no examples reflected in the target dataset.

All transfer learning approaches performed slightly worse compared to training the network from scratch with the MR data (transfer learning: DSC$\geq$0.80, HD$\leq$17.79 and MSD$\leq$2.17, compared to DSC=0.85$\pm$0.11, HD=15.97$\pm$13.14 and MSD=1.82$\pm$1.93) The quality of the manual contours for the CT images was not as high as for the MR images. The CT images have been contoured by a range of different clinicians from various hospitals and are not as consistent as the ones created for this dataset. Three typical poor examples of these contours are illustrated in figure 6.21. The quality of the CT contours certainly had an impact on the initialised weights and could therefore have contributed to the worse accuracy I obtained via transfer learning.

Due to all of the mentioned arguments, it remains challenging to transfer specific knowledge gained from pre-trained networks to a new application. In the following section, I investigate a potential solution to this problem.

A limitation of this study was that the dataset which I used as a source database was small in comparison to other studies of transfer learning (hundreds of CT images compared to millions of natural images). However, this will often be the case in medical imaging. Furthermore, while I explored on a suitable setting of the hyperparameters, I did not perform a systematic investigation of the optimal parameters. This could have an impact on the final performance of the networks. However, such grid searches are time-consuming and could lead to an overfitting to the data, in particular for small datasets.

Additionally, I could not exclude the possibility that the capacity of the network was large enough for approach C (bottleneck fixed) to be only relying on the non-frozen weights to fit the MR training data, without a significant benefit of having the weights initialised based on the training with the CT data.

6.5.5 Conclusion

In this study, I investigated the use of transfer learning from one larger imaging dataset (202 CT images) to a smaller dataset (27 MR images). While I could achieve an accuracy close to the inter-observer variability, it remained unclear what information was transferred through these approaches. I propose potential solutions to this problem in the next section.

6.6 Cross-modality learning

6.6.1 Introduction

A general approach to tackle the lack of training data is to augment them with random rotations, translations, geometric scaling, mirroring or contrast stretching. While these methods try to increase the variety in the training data, they generally are not able to mimic the large variabilities existing in the full population of patients' anatomies.

Instead of only using rotations and translations, one could employ elastic deformations of the training data to increase the variations between images. However, a major drawback of elastic deformations is that arbitrary deformations may not reflect typical anatomy seen in patients, whereas designing representative deformations is a time-consuming process and requires expert-knowledge on variations seen in patients. Moreover, data augmentations introduce additional hyperparameters into the deep learning setup, e. g. translation range, rotation angles, deformation settings, which need to be tuned in addition to the already existing hyperparameters.

Another approach is to use pre-trained networks on related problems via transfer learning. However, I have shown in the previous section (6.5) that conventional transfer learning approaches did not increase the performance of the final network and I believe that the variety in anatomical shapes from the larger (CT) dataset may be lost in the re-training process. Moreover, both approaches, i. e. training a network from scratch and transfer learning, require that both, annotation and training, need to be repeated for every novel MR contrast setting.

Recently, image generation methods with deep learning have been introduced. The so-called generative adversarial networks (GANs) [51] can learn to mimic any distribution of data. A GAN consists of two neural networks competing with each other: the generator and the discriminator network. The basic principle is illustrated in figure 6.22. The discriminator's task is to classify presented examples as real or fake, while the generator needs to fool the discriminator by generating real-looking examples from random noise. Competition in this adversarial "game" drives both, the discriminator and the generator network, to improve through joint optimisation until the fake data are indistinguishable from the real data. GANs have been applied to image-to-image translation problems, such as described in [69], where input images from one domain were mapped to output images of another domain. Instead of a pure GAN, they used a conditional GAN, where the input images were used as a condition on the generated data distribution. Instead of random noise, images from the source domain were fed into the generator and real data was represented by the target imaging domain. A drawback of the approach

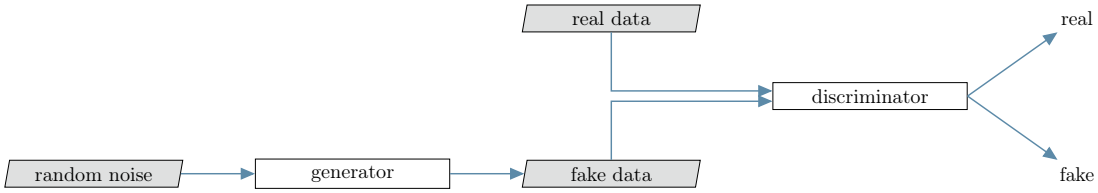


Figure 6.22: Generative adversarial network (GAN): The discriminator tries to classify input images as real or fake ones, while the generator tries to generate fake data which closely matches the data distribution of the real data to fool the discriminator.

by Isola et al. [69] is that paired input and output data are needed. Zhu et al. [169] extended this approach to unpaired datasets in their CycleGAN. The focus was to learn a mapping between two different image collections through cycle-consistent GANs, e. g. by transforming photographs into paintings of a certain artist and vice versa. Wolterink et al. [159] applied a CycleGAN to the creating of synthetic CT images from brain MR data.

In this study, I used a CycleGAN to generate synthetic MR images from CT images. Instead of using the generated synthetic images for data augmentation, I took one step further and trained a 2D CNN solely based on the synthetic MR images to segment the parotid glands. This resembled the situation where one would like to reuse annotated data from a different imaging domain (here CT images) to a new imaging domain (here MR images) without the necessity to employ the time-consuming and expensive annotation process. To the best of my knowledge, this was the first study to generate synthetic MR images from CT images for the purpose of training a network to segment MR images.

6.6.2 Materials and Methods

Figure 6.23 provides an overview of the method employed in this study. It consisted of three steps:

(1) For each axial slice of the CT images, a corresponding synthetic MR axial slice was generated using the CycleGAN.

(2) A 2D U-Net was trained using the synthetic MR images and corresponding manual contours from CT images as input.

(3) The trained 2D U-Net was used to propose contours on unseen real MR images.

6.6.2.1 Data preparation

Imaging data comprised the 27 T2w MR images from the MD Anderson Cancer Center, as well as the 202 CT images from public databases. Further details on image acquisition

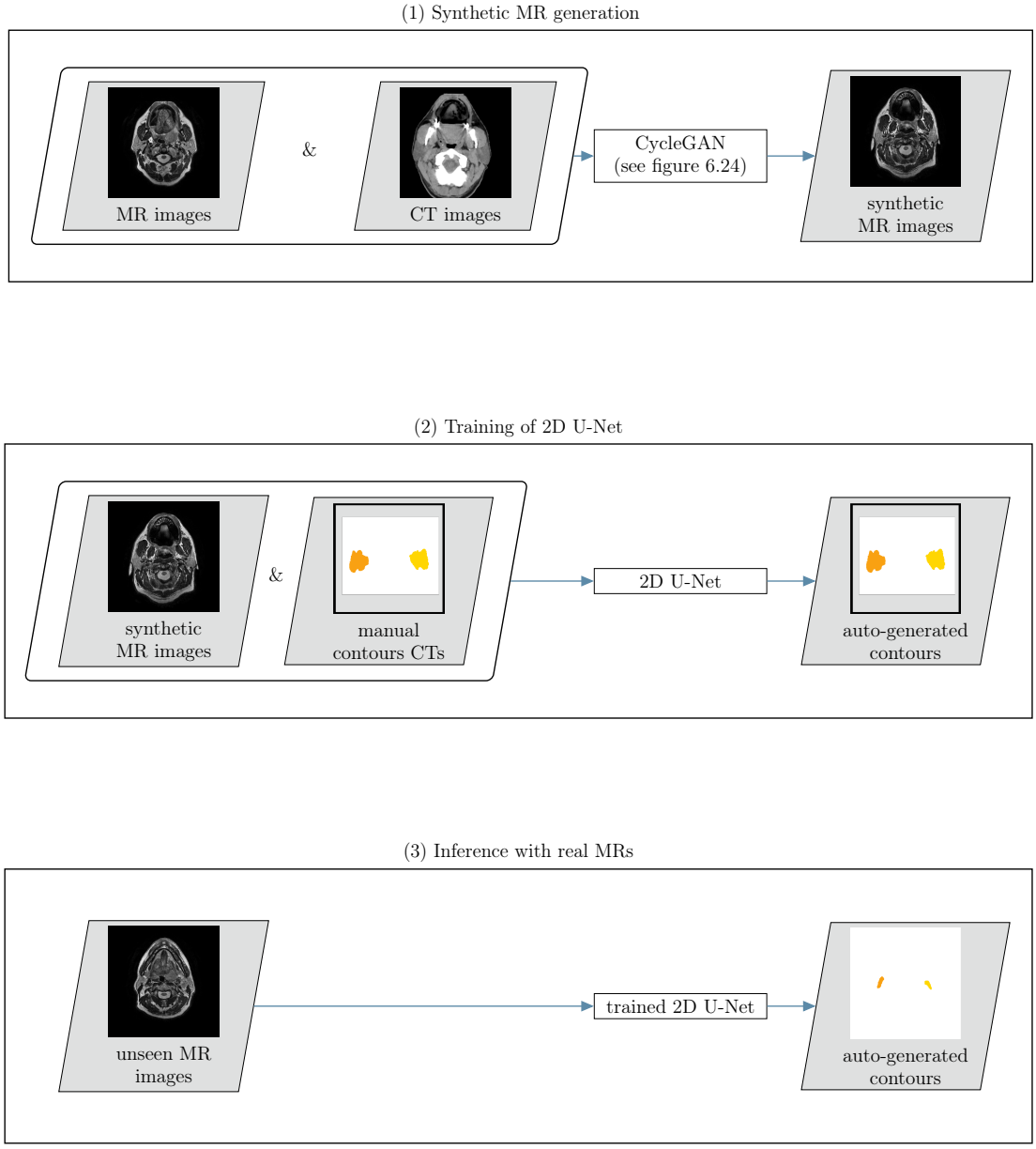


Figure 6.23: This figure provides an overview on the novel cross-modality learning method: in the first step (top row), synthetic MR images are generated through the CycleGAN method. The synthetic MRs are then fed into a 2D U-Net, together with the annotations from the CT images (second row). In a third step, the trained network is applied to unseen real MR images (bottom row).

parameters or image processing steps are provided in chapter 3.

6.6.2.2 Synthetic MR generation

Step (1) of the workflow consisted of the synthetic MR generation. The unpaired 2D slices from the CT and MR images were fed into a 2D CycleGAN network to generate synthetic MR images for each of the 202 CT images. I used the PyTorch [113] implementation provided by Zhu et al. [169], available on Github[1]. In the following paragraphs, I shortly describe the CycleGAN and the adjustments I made to this implementation. Further details on the original implementation are provided on Github[1] and in [169].

General workflow and objectives

The aim was to generate a "corresponding"[2] MR image I_{MR} for each CT image I_{CT}. For this purpose, the generator network G_{MR} was trained. To ensure that the generated MR images were indistinguishable from real MR images, the discriminator network D_{MR} was introduced. It aimed to distinguish between real and fake MR images. As described before, these two networks compete with each other in an "adversarial game". The adversarial loss was defined as:

$$\mathcal{L}_{\mathrm{adv}}(G_{\mathrm{MR}}, D_{\mathrm{MR}}, I_{\mathrm{MR}}, I_{\mathrm{CT}}) = (D_{\mathrm{MR}}(I_{\mathrm{MR}}))^2 + (1 - D_{\mathrm{MR}}(G_{\mathrm{MR}}(I_{\mathrm{CT}})))^2 \tag{6.21}$$

for an unpaired input $(I_{\mathrm{CT}}, I_{\mathrm{MR}})$.

The "game" of the two networks can be understood as a "min-max game", where the discriminator aims to maximise the objective and the generator tries to minimise it:

$$\min_{G_{\mathrm{MR}}} \max_{D_{\mathrm{MR}}} \mathcal{L}_{\mathrm{adv}}(G_{\mathrm{MR}}, D_{\mathrm{MR}}, I_{\mathrm{CT}}, I_{\mathrm{MR}}) \tag{6.22}$$

In theory, this loss function does not guarantee that an individual input I_{CT} is mapped to the desired I_{MR}. Moreover, it could learn to map any CT image to the same, unique, MR image. To reduce the space of possible mappings, Zhu et al. [169], therefore, introduced a cycle-consistency loss. For this, they included two further networks, generator G_{CT}, which maps real MR to synthetic CT images, and discriminator D_{CT}, discriminating between real and fake CT images. The adversarial loss for these two networks can be obtained by replacing the MR with the CT in equation (6.21).

To restrict the mapping of the CT image to an MR image that resembles features of this CT image, cycle-consistency losses are introduced. These guarantee that the

[1] https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix

[2] There is no one-to-one mapping for this case, so the aim is to map to a "plausible" MR image.

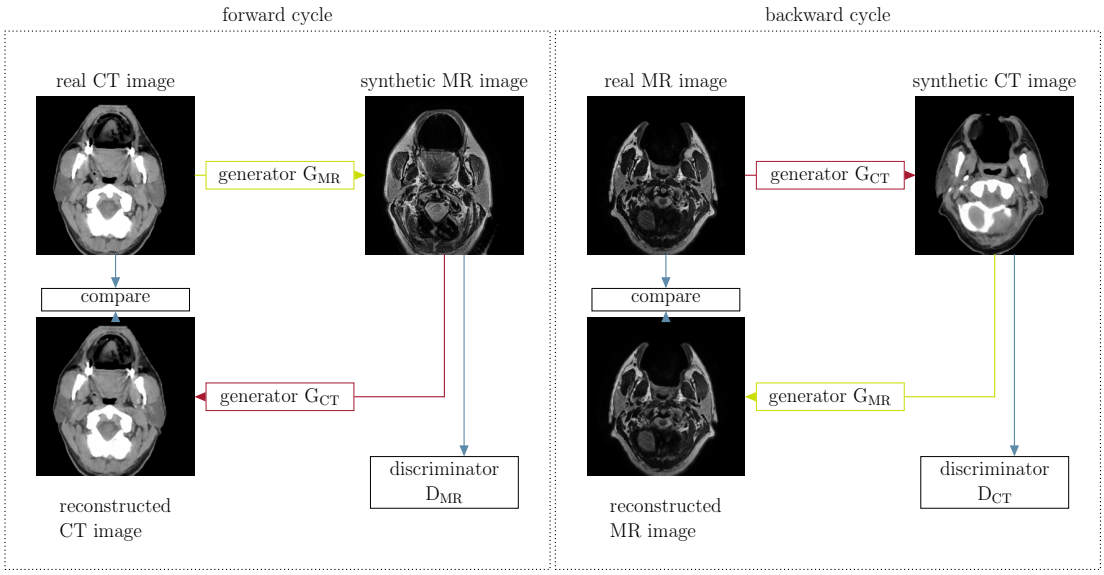


Figure 6.24: Illustration of the CycleGAN method: two cycles are introduced such that the generated synthetic images resemble the input images (Cycle for synthetic MR images at the top and for synthetic CT images at the bottom).

generated CT image that has gone through the full cycle (CT->MR->CT) is similar to the original CT. With the L1 norm $||\cdot||_1$, the cycle-consistency loss yields

$$\mathcal{L}_{\text{cycle,CT}}(G_{\text{MR}}, G_{\text{CT}}, I_{\text{CT}}) = ||G_{\text{CT}}(G_{\text{MR}}(I_{\text{CT}})) - I_{\text{CT}}||_1 \tag{6.23}$$

for a real CT image I_{CT} and vice versa for I_{MR}. Figure 6.24 illustrates these forward (CT→MR→CT) and backward cycles (MR→CT→MR).

To constrain the generated synthetic MR images to ones that geometrically match the source CT images, I introduced a further geometric consistency loss. For this purpose, I determined the external masks through Otsu thresholding and binary closing operations (see also chapter 3) of both, the source CT and the synthetic MR and calculated the binary cross-entropy between these masks. I introduced the same loss for the mapping in the opposite direction (source MR to synthetic CT). With $M(I)$ denoting the external mask of an image I, the geometric loss term yields as

$$\begin{aligned}\mathcal{L}_{\text{geo,CT}}(G_{\text{MR}}, I_{\text{CT}}) =& M(G_{\text{MR}}(I_{\text{CT}})) \cdot \log(M(I_{\text{CT}})) &(6.24)\\ &+ (1 - M(G_{\text{MR}}(I_{\text{CT}}))) \cdot \log(1 - M(I_{\text{CT}})) &(6.25)\end{aligned}$$

for a real CT image I_{CT} and vice versa for I_{MR}. This loss function is different from the

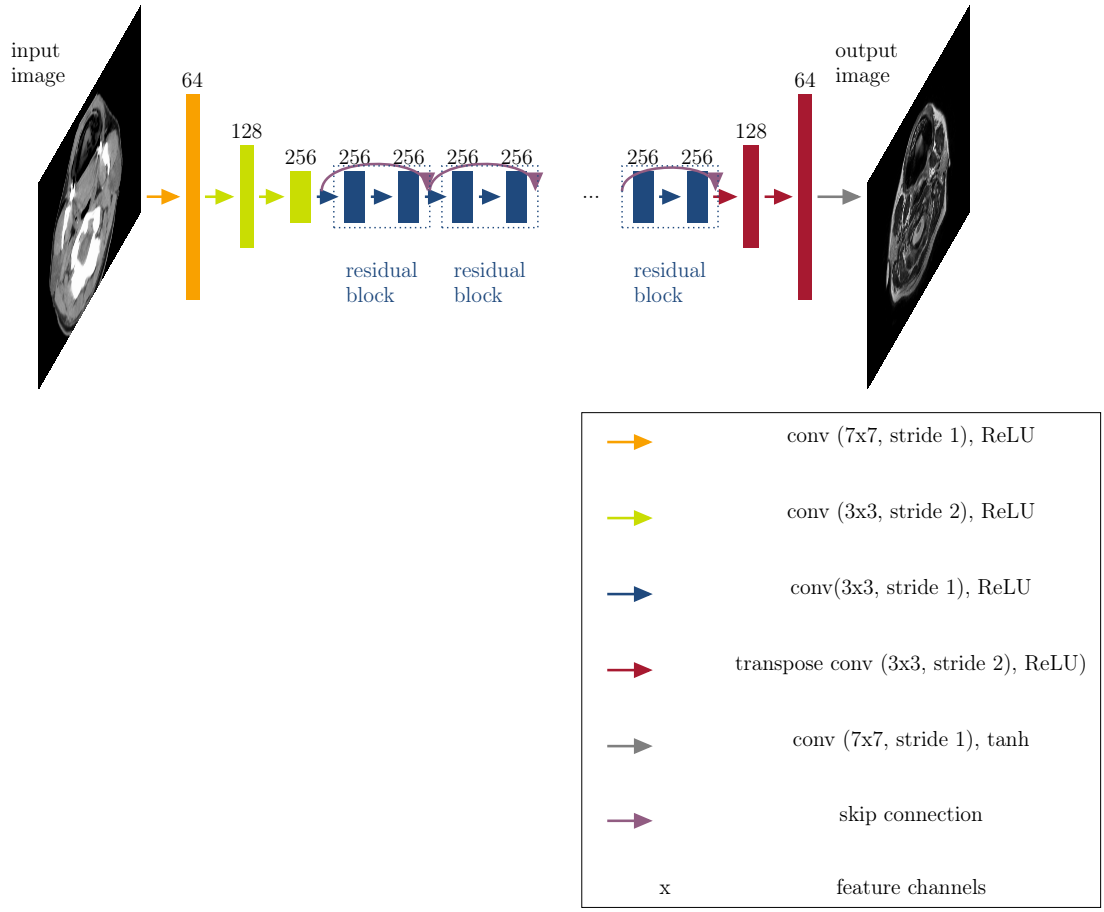


Figure 6.25: Generator network: This figure illustrates the network used for the generator in this study. It consists of 3 convolutional layers, followed by 9 residual blocks, 2 transpose convolutional layers and a final convolutional layer with a tanh activation function.

default network from Github[3]. The full training objective is given as:

$$\begin{aligned}\mathcal{L}_{\text{cycleGAN}} =& \mathcal{L}_{\text{adv}}(G_{\text{MR}}, D_{\text{MR}}, I_{\text{CT}}, I_{\text{MR}}) + \mathcal{L}_{\text{adv}}(G_{\text{CT}}, D_{\text{CT}}, I_{\text{MR}}, I_{\text{CT}}) \\ &+ \lambda_{\text{CT}} \cdot \mathcal{L}_{\text{cycle,CT}} + \lambda_{\text{MR}} \cdot \mathcal{L}_{\text{cycle,MR}} \\ &+ \lambda_{\text{geo,CT}} \cdot \mathcal{L}_{\text{geo,CT}} + \lambda_{\text{geo,MR}} \cdot \mathcal{L}_{\text{geo,MR}}. \end{aligned} \tag{6.26}$$

with relative weights λ_i for each of the individual contributions.

Generator network

The generator network takes as input a 2D image of the source domain and generates a 2D image of the target domain. It is illustrated in figure 6.25. The network consists of

[3] https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix

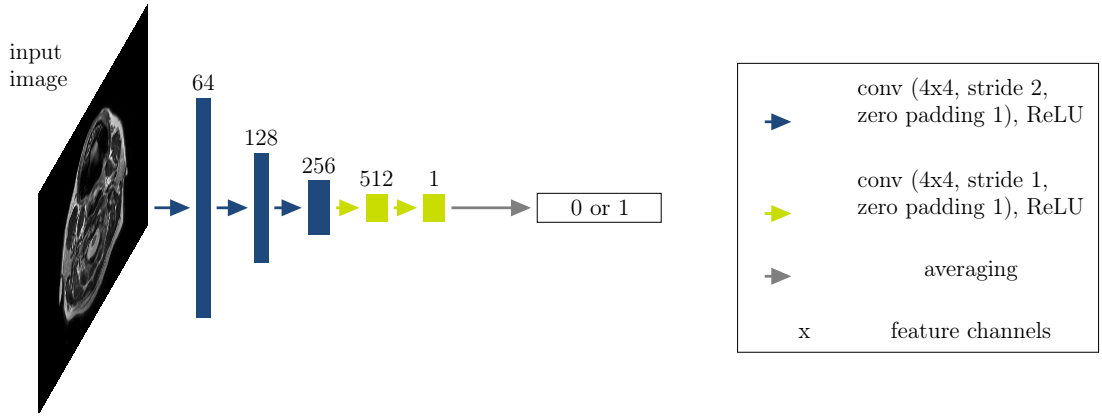


Figure 6.26: Discriminator network: This network consists of 5 convolutional layers and classifies images into two categories: real or fake. It was introduced in [69] as PatchGAN.

- 3 convolutional layers, each followed by batch normalisation and ReLU activation
- 9 residual blocks, as introduced in chapter 6.11
- 2 transposed convolutional layers
- a 1x1 convolutional layer with a tanh activation function (see section 6.2.1.2).

Residual networks utilise so-called short-cuts or skip connections, as illustrated in figure 6.11 on page 109. For more details on the generator network, I refer to [169].

Discriminator network

The discriminator network takes as input a 2D image and classifies the image as real (label=1) or fake (label=0). It consists of 5 convolutional layers. This discriminator network was introduced by Isola et al. [69]. They called it a 70x70 PatchGAN. The patch, in this case, does not mean that patches are used as input to the network but that the prediction is formed from overlapping regions with a receptive field size of 70x70. This can be calculated from equation (6.6). The PatchGAN is essentially a fully convolutional neural network. Figure 6.26 illustrates this network. Further details can be found in [69].

Training parameters

I employed the recommended settings, as described in [169]. I used the Adam optimiser [73] with a batch size of 1 and an initial learning rate of 0.0002. I kept this learning rate fixed for the first 100 epochs and linearly decayed the learning rate to zero for the next 100 epochs. I furthermore found the best settings for the weights of the respective contributions to the loss function to be $\lambda_{\mathrm{CT}} = \lambda_{\mathrm{MR}} = 10$ and $\lambda_{\mathrm{geo,CT}} = \lambda_{\mathrm{geo,MR}} = 10$.

Data cleaning as input for segmentation network

Since not all synthetic MR images perfectly matched the input CTs, I performed a data cleaning where I only selected slices that were suitable for the segmentation of the parotid glands. I explored constraints on the external outline of the head and decided to perform a refinement 2D registration to map synthetic MR images to the original CTs. I performed the registration using the Elastix toolkit [74] in two steps: first determining any possible translations and finally a deformable registration with a B-spline transformation, as introduced in chapter 5. I used a CPP grid spacing of 8 mm in the deformable registration. For both steps, I employed the mutual information as similarity measure between CT and synthetic MR images and used the gradient descent method as the optimiser. I chose parameter settings for the deformable registration as recommended in the user manual (SP_α=1, SP_A=20, SP_a=1000 and control point grid spacing of 8 mm). As the synthetic MR images were already generated in the same geometrical space as the CTs, the segmentation of the CTs formed the gold standard MR segmentation for the segmentation network.

6.6.2.3 Segmentation network

After data cleaning, I fed all remaining 2D synthetic MR images (approximately 1500) into a 2D U-Net as training data. I split the data into 80% training and 20% validation to choose suitable hyperparameters. I used the same network architecture as introduced in figure 6.13. The inference was performed on the 27 real MR images. I trained the network for 100 epochs and with an initial learning rate of 5×10^{-5}. I gradually reduced the learning rate by monitoring the validation loss, down to a minimum of 10^{-7}.

6.6.2.4 Computation time

The run times were determined for programme execution on a single Tesla V100 with 16 GB VRAM. Inference times are stated per patient, where I calculated the average over all 27 patients, as well as the standard deviation.

6.6.2.5 Geometric evaluation

I evaluated the performance of the segmentation network by calculating the DSC, HD and MSD between manual and auto-generated contours and compared to the accuracy of the CT network as a benchmark.

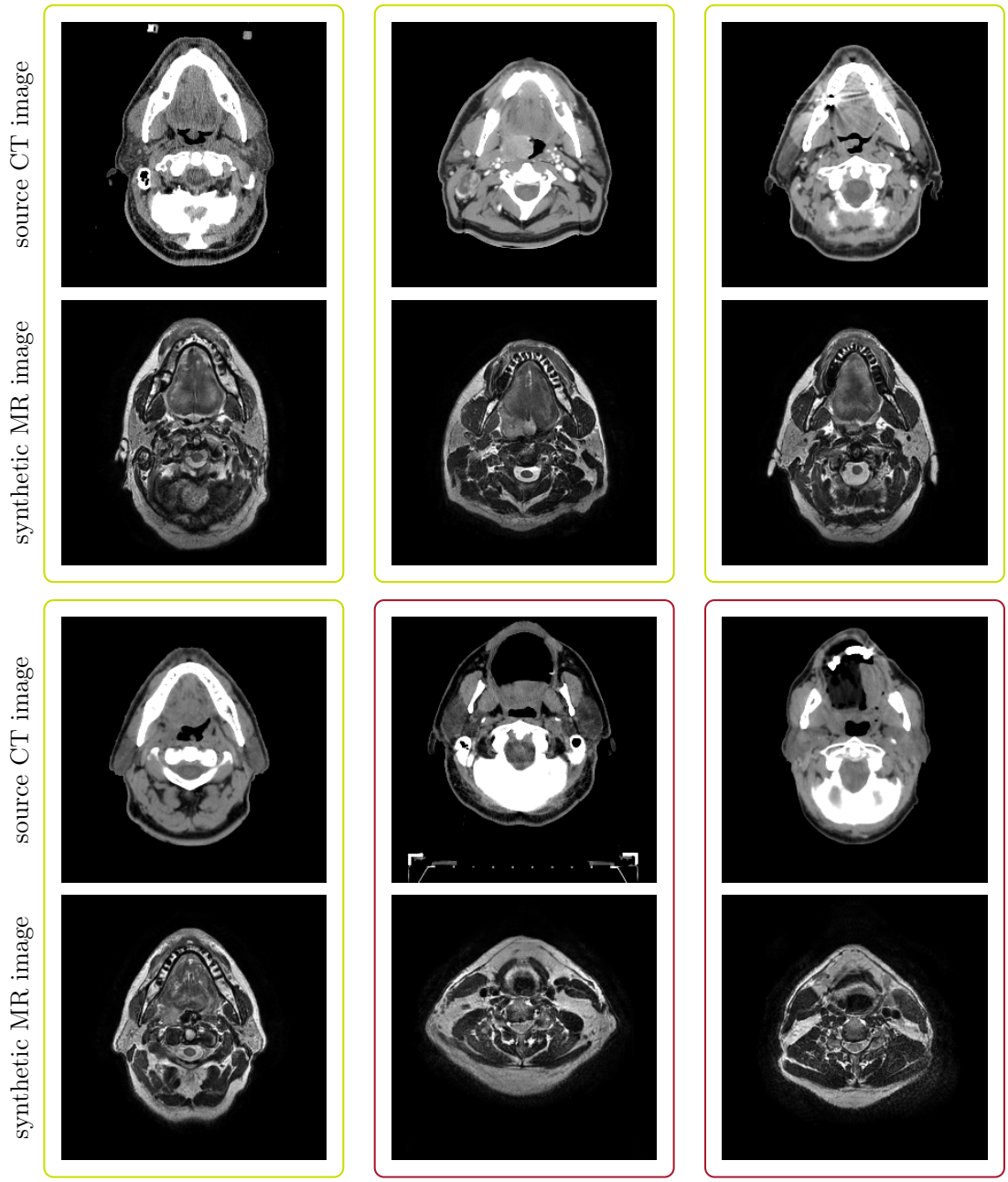


Figure 6.27: Typical examples of synthetic MRIs and their corresponding source CTs: The green boxes highlight example cases that were selected for further learning. The red boxes highlight cases where the CycleGAN failed to produce anatomically corresponding MR images for the corresponding CT images and hence were rejected for further analysis.

6.6.3 Results

6.6.3.1 Synthetic MR generation

Figure 6.27 illustrates selected (green box) and rejected example cases (red boxes) of synthetic MR images together with their corresponding source CT images. In some

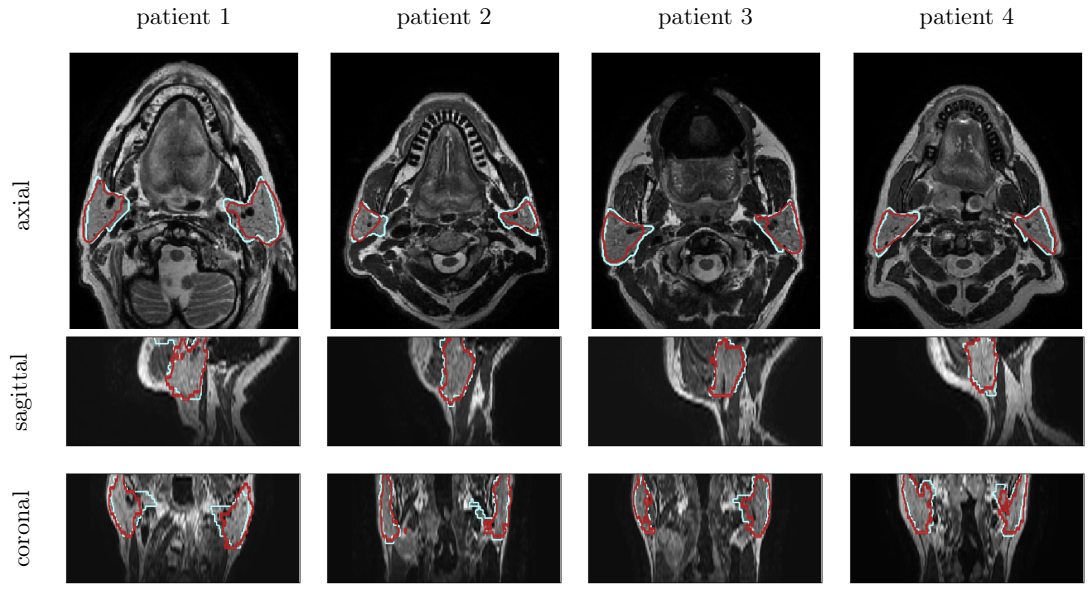


Figure 6.28: This figure shows in each column a typical example case of the cross-modality learning approach (in red). The manual contours are shown in blue. The rows correspond to an axial, sagittal and coronal cross section, respectively. Each example originates from a different patient image.

rejected cases, the synthetic MR images appeared as if they could be real MR images. However, they did not reflect the anatomy visible in the source CT images.

6.6.3.2 Computation time

Training of the CycleGAN took approximately 72 hours. The training of the 2D U-Net took approximately 150 minutes, whereas inference was made within 0.86 ± 0.02 seconds.

6.6.3.3 Qualitative segmentation results

Figure 6.28 illustrates four typical example cases for auto-generated contours using the cross-modality approach, comparing to the manual contours. I selected an axial, sagittal and coronal view for each of the patients. The auto-generated contours followed the manual ones closely.

6.6.3.4 Geometric evaluation

Table 6.7 lists mean and standard deviations for the DSC, HD and MSD, comparing the cross-modality learning to the accuracy of the trained CT network. The cross-modality learning accuracy (DSC: 0.77±0.07, HD: 18.32±10.12 mm, MSD:2.51±1.47 mm) stayed below, but was close to the inter-observer variability (0.84±0.04, 10.76±4.35 mm,

Table 6.7: Evaluation of the geometric accuracy of auto-segmenting the left and right parotid gland of the cross-modality learning approach. As a benchmark, I also include the geometric accuracy of the CT-trained network.

ROI	method	$\overline{DSC}$	$\overline{HD}$ [mm]	$\overline{MSD}$ [mm]
right parotid	**cross-modality learning**	**0.76 ± 0.06**	**18.32 ± 10.12**	**2.66 ± 1.26**
	CT only	0.81 ± 0.07	13.01 ± 5.61	1.87 ± 0.84
	MR only	0.84 ± 0.07	16.75 ± 10.37	2.00 ± 1.46
	inter-observer variability	0.84 ± 0.04	10.76 ± 4.35	1.40 ± 0.45
left parotid	**cross-modality learning**	**0.77 ± 0.04**	**17.75 ± 7.49**	**2.36 ± 0.75**
	CT only	0.82 ± 0.05	12.98 ± 5.15	1.74 ± 0.53
	MR only	0.85 ± 0.08	15.19 ± 8.09	1.63 ± 1.27
	inter-observer variability	0.83 ± 0.04	10.94 ± 3.75	1.59 ± 0.63

1.40±0.45 mm), as well as the CT-trained (DSC: 0.82±0.09, HD: 13.01±5.61 mm, MSD: 1.81 ± 0.99 mm) and MR-trained networks (DSC: 0.84±0.07, HD: 16.75±10.37 mm, MSD: 2.00±1.46 mm).

6.6.4 Discussion

In this study, I employed a new technique, cross-modality learning, to reuse knowledge gained from one application (annotated CT images) in a new application (non-annotated MR images). To the best of my knowledge, I was the first to generate synthetic MR images from annotated CT images to train an MR segmentation network. I found that it was indeed possible to obtain decent quality annotations of MR images with only annotated CT data.

6.6.4.1 Synthetic MR generation

The CycleGAN was generally able to generate synthetic MR image from the input CT images. In some cases, it failed, however, most of these failed synthetic MR images often still looked like an MR image, albeit not corresponding to the anatomy of the source CT image. Depending on the application, such images still could be useful. However, for my purpose, where I assume that the contours are still correct, one requires a satisfactory agreement between the represented anatomies. The failed generation could be because I only had a small number of real MR images from which the CycleGAN could perform a style transfer. As the CycleGAN learns to map features from the source data (here: CT) to the target data (here: MR), it might focus on irrelevant features, such as smaller heads in the target data. Failure to generate an MR that corresponded well to the input CT especially happened at the superior and inferior boundary slices. Due to the limited field of view of the training MR images in that direction, there were not a lot of samples available for the CycleGAN to learn.

I furthermore detected a systematically narrower external outline of the head for the synthetic MR images compared to the source CTs. In theory, no penalty in the CycleGAN prevents it from learning this narrowing function, as it could learn to generate more "narrow" MR images in the forward generator and go back to "broader" CT images in the backwards generator. This issue could be related to the skin outline being visible in the CT images but not in the MR images. While I tried to enforce a better overlay between these outlines by incorporating a geometric consistency penalty in the loss function, I was not able to entirely remove this issue. As GANs have been shown to be susceptible to hyperparameter settings, for instance, the weights of the individual loss contributions, this issue may be improved by a better optimisation of these parameters. In this study, I performed a 2D registration between the CT and the corresponding synthetic MR image to mitigate these detected "narrowing" transformations.

Recent research has shown that GANs are generally challenging to train and face problems with non-convergence, mode collapse (producing limited varieties of samples) and diminishing gradients of the generator when the discriminator becomes too powerful [49]. As they have been shown to be highly susceptible to hyperparameter selections [49], I expect that one could improve the synthetic MR generation further by tuning more hyperparameters. However, this would require more training such that overfitting can be avoided. While this study was a proof-of-concept study, in future research one could optimise these parameters further, with more data, to overcome the addressed limitations.

6.6.4.2 Geometric evaluation

There are several points where the achieved accuracy can be further improved. The quality of the ground truth contours for the CTs was not as high as for the MR images. This was also evident from the accuracy of the CT-trained network. With more high-quality CT contours, I am confident that the accuracy of the cross-modality learning will improve.

Compared to the accuracy of training a network with MR images from scratch, as described in section 6.4 (DSC: 0.85 ± 0.11, MSD: 1.81 ± 1.94 mm), the cross-modality learning performed worse. However, obtaining a high-quality segmentation is a time-consuming process and would need to be repeated for every new contrast setting. Furthermore, the comparison may be deemed unfair as the manual CT segmentation was less consistent than the MR segmentation (see figure 6.21 on page 132 in the previous chapter).

In comparison to the transfer learning approach of the previous section 6.5, I could directly incorporate the varieties found in a larger patient database to the small subset

of MR images. Furthermore, unlike the transfer learning approach, no additional manual segmentation was necessary.

I anticipate that this method could be used to generally adapt a trained network of one imaging modality to another imaging modality. Auto-segmentation approaches are usually trained on a very particular subset of imaging data. These approaches might work well when the target images are similar to the ones that have been used in the development phase. However, in the clinical routine, there are frequent changes, especially in MR image settings. While in a conventional approach this could mean that a new database with annotations of the new images would need to be created, the cross-modality learning would be able to reuse the already existing annotations on existing data and transfer it to the new imaging settings.

6.6.4.3 Limitations and future work

A limitation of this approach was that 2D slices were predicted instead of directly generating 3D volumes. This led to inconsistencies between some slices and only allowed for a 2D segmentation network. Employing a fully 3D approach may reduce the number of falsely predicted synthetic MR images. However, the 2D CycleGAN is already consuming a large amount of memory and a 3D approach would further increase this.

At this point, 2D image registration between CT and synthetic MR slices was still necessary. I am confident that in future work, this need could be removed with larger datasets as this would enable the CycleGAN to capture the important features in both imaging modalities and lead to better-quality synthetic MR images. Moreover, the CT data I used in this study originated from different hospitals with various imaging protocols and patient specifics. Despite this challenge, I found promising results in this study and think that the quality of the synthetic MR images will improve further with more consistent data.

6.6.5 Conclusion

This technique of cross-modality learning can be of great value for segmentation problems with sparse annotated training data. I anticipate using this method with any non-annotated MR dataset to generate synthetic MR images of the same type via image style transfer from CT images. To properly learn the appearance of these MR images, this would ideally be a large database. Furthermore, as this technique allows for fast adaptation of annotated datasets from one imaging modality to another, it could prove useful for translating between large varieties of MRI contrasts due to differences in imaging protocols within and between institutions.

6.7 Main findings and Conclusion

In this chapter, I explored the potential of deep learning-based approaches to the segmentation of MR images of HNC patients for RT treatment planning. In the following, the main findings and implications for future research are discussed.

Although I have shown in chapter 4 that calculating the dosimetric effect of auto-segmentation is a more accurate assessment of its quality than purely geometric measures, there was not enough time to do this within the scope of this thesis and it is therefore left for future work.

Overall, the results of this chapter indicate that deep learning-based approaches are capable of achieving accuracies comparable to state-of-the-art methods, such as atlas-based segmentation approaches. Moreover, deep learning-based approaches are much faster in segmenting images than atlas-based approaches (seconds compared to minutes or hours). Additionally, atlas-based approaches scale linearly with the number of training images, at least in their conventional and presented form. While the training phase of deep learning-based approaches may increase, more training data do not impact the inference time on new, unseen data. These findings suggest that deep learning-based approaches may be excellent candidates for adaptive treatment workflows.

A limitation of this study was the limited number of training data. With this drawback, it was not possible to entirely avoid overfitting to the data-at-hand, and it is crucial to bear in mind a possible bias and a decreased performance in generalisability to new data. Nonetheless, the methods developed in this study can be generalised for other data and it is to be expected that the generalisability will improve with more variety in an increasing collection of new data.

There has been little quantitative analysis of the amount of training data needed for the development of generalisable and accurate deep learning-based methods in medical imaging. While in computer vision, typically thousands and millions of image are available, this is not the case for medical image. However, medical images have unique characteristics which alleviate the strong need for massive datasets. First of all, they are subject to more standardisation than natural images. Additionally, there is typically less variance in the data distribution than found for natural images. These findings, combined with the findings of this study, suggest that deep learning-based approaches are feasible with smaller datasets than generally found in applications to natural image processing.

An important finding was that increasing quantity does not necessarily imply increasing segmentation quality. As one could see in the transfer learning approach, the accuracy of training a network to segment CT images was below the accuracy of the

trained MR network. While the variance in the CT database was more substantial and may have contributed to lower overall accuracy, a likely reason for this finding is the inferior quality of ground truth segmentations of the CT images.

I demonstrated promising applications that can alleviate the data problem at the heart of medical image segmentation further. In particular, I introduced a synthetic image generation method, cross-modality learning, by employing a generative adversarial network approach. Despite some challenges in creating geometrically accurate synthetic images, I am confident that this method can have an impact on using auto-segmentation algorithms clinically. A problem with applying auto-segmentation methods in a clinical routine is that the data which were used to develop the algorithm can differ significantly from the real data encountered in the clinic, which can be messy and subject to frequent changes. Cross-modality learning is able to overcome these pitfalls of developed segmentation methods as it can bridge the gap between the data used for developing segmentation algorithms and the data used in the clinical workflow.

Chapter 7

Final comparison and exploration of limitations

In this chapter, I summarise the findings from the previous chapters and compare qualitatively the different auto-segmentation methods developed in this thesis, as well as discuss their respective strengths and limitations.

In this thesis, I designed, developed and implemented different auto-segmentation methods. The first part of this chapter provides a summary and quantitative comparison of the findings from the previous chapters. Afterwards, potential limitations are explored, employing an independent test dataset. A quantitative assessment of auto-segmentation accuracy requires observer delineations of ROIs. Unfortunately these were not available for the independent test dataset so this chapter provides a purely qualitative assessment of this dataset.

7.1 Summary and comparison of auto-segmentation methods

Table 7.1 summarises the quantitative results of all auto-segmentation methods analysed in this thesis and the strengths and limitations for each method, which are further explored in the second part of this chapter. A 2D CNN could achieve an accuracy comparable to a multi-atlas-based method (DSC: 0.85±0.11 vs. 0.85±0.05, MSD: 1.82±1.94 vs. 1.67±1.21 mm) but within a considerably shorter computation time (<1 s vs. 1800 s). No significant differences between any of the CNN-based algorithms could be detected for the data utilised in this study.

While the overall accuracy was similar between atlas- and deep learning-based methods, the deep learning-based methods were better able to detect the boundaries of parotid glands which were infiltrated by involved lymph nodes or primary tumours and did not include them within the segmented ROI. Figure 7.1 illustrates typical examples. The last column represents a rare example where the atlas-based outperformed the deep learning-based method. The lymph nodes appear to have a similar texture compared to the parotid gland itself, which was likely the reason that the network annotated this part as parotid gland. However, in the majority of cases, the deep learning algorithm more often excluded regions which were not part of the ROI. Since the shape and location of tumours and involved lymph nodes varies widely among patients, it is difficult for an atlas-based approach to recognize infiltrated normal structures. Deep-learning strategies, on the other hand, are recognizing local shapes and edges and are, therefore, less affected by global geometrical changes.

The 2D CNN was ignorant about any context information along the axial direction. One would expect more consistent outlines in that direction by feeding information on neighbouring axial slices into the network. I, therefore, explored two approaches: 2.5D (feeding three adjacent slices as input) and fully 3D (feeding 3D patches). While in theory, both methods should provide more consistent contours in that direction, no significant differences to the 2D method could be detected for the data in this study,

Table 7.1: Summary of quantitative results of all auto-segmentation methods analysed in this thesis, as well as their respective strengths and limitations

method	modality	# scans	DSC	MSD [mm]	t_{train} [min]	$t_{\text{inference}}$ [s]	strengths	limitations
atlas-based	T2w	27	0.85 ± 0.05	1.67 ± 1.21	-	1800	accurate, independent of modality changes	slow, sensitive to head posture and field of view
deep learning							fast, flexible, better handling of abnormalities	sensitive to sequence settings
2D	T2w	27	0.85 ± 0.11	1.82 ± 1.94	24.65 ± 0.22	0.90 ± 0.06	accurate	neglects 3D context
2.5D	T2w	27	0.82 ± 0.09	2.05 ± 1.58	25.35 ± 0.25	0.88 ± 0.04	information on adjacent slices	more network parameters
3D	T2w	27	0.82 ± 0.13	1.88 ± 1.67	376.28 ± 75.85	1.54 ± 0.44	full 3D information	more network parameters, (currently) 2-step method
multi-modality	T1w+T2w	27+27	0.81 ± 0.19	2.21 ± 2.63	25.56 ± 0.21	0.90 ± 0.04	complementary information	accuracy limited by registration accuracy
transfer learning	CT+T2w	202+27	0.80 ± 0.08	2.31 ± 1.46	12.40 ± 0.12	0.97 ± 0.16	information from larger database when little training data available	not clear in which layers desired information is stored
cross-modality learning	CT+T2w	202+27	0.77 ± 0.07	2.51 ± 2.15	4320 + 150	0.86 ± 0.02	deals with image sequence/scanner updates	manual selection of synthetic images, 2D

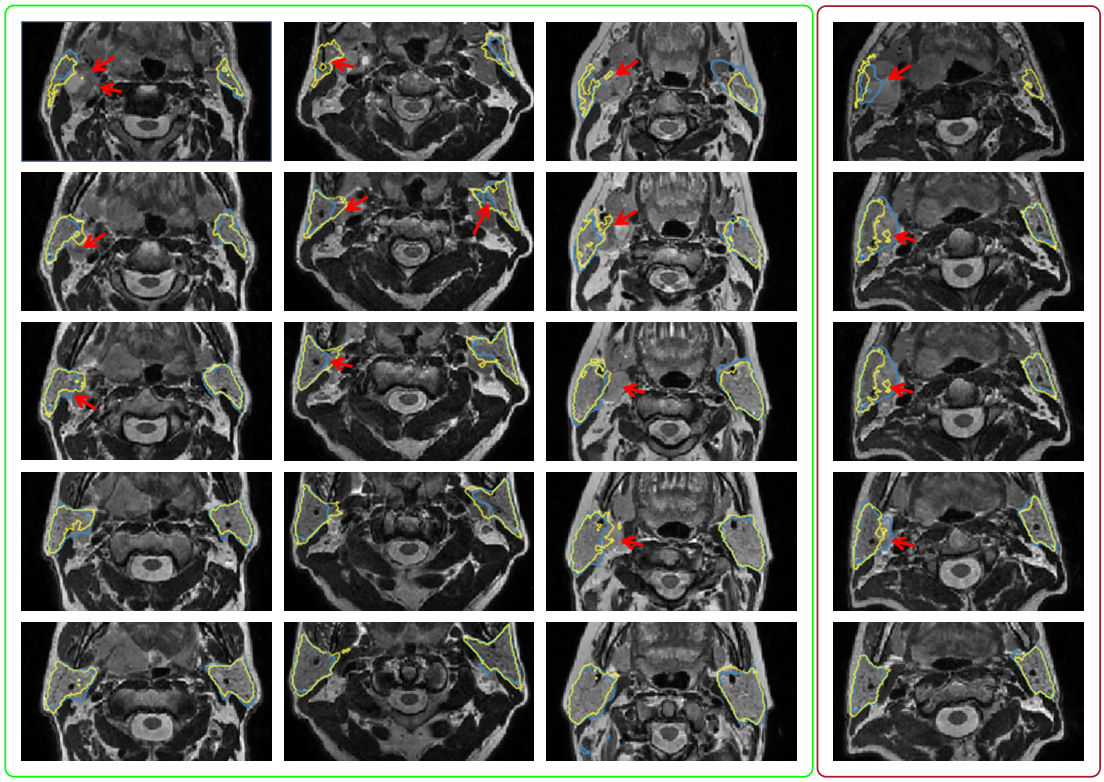


Figure 7.1: Segmentation examples: this figure provides 4 cases where the parotid glands were infiltrated by involved lymph nodes. The blue line denotes the delineation using a 2D deep learning-based approach, whereas the yellow line was created following an atlas-based approach. Each column shows an example case with consecutive axial slices shown in each row. The red arrows point towards involved lymph nodes. The first three columns highlight examples where deep learning outperformed an atlas-based approach (green box) whereas the last column provides a less common counter-example (red box).

which may be attributed to the small amount of data.

Combining multiple modalities could provide the network with complementary information about tissues. I, therefore, explored how the combination of T1w and T2w contrast could improve auto-segmentation. For the data in this study, there was no improvement following this strategy. As the information is used complementary, a very accurate registration is crucial between the two images, such that the network can relate the contrast in two corresponding voxels accordingly. Despite registering the images before feeding them into the network, there may have been residual errors in the registration, and therefore, this multi-modality strategy did not help to improve the auto-segmentation.

A limitation of the studies discussed so far was the small dataset. Deep learning would greatly benefit from more input data to generalise well on new data. I explored a technique named transfer learning, where knowledge gained in one imaging domain with a large number of available annotated images is transferred to an imaging domain

with smaller datasets. Transfer learning was used to transfer knowledge of parotid segmentation on CT imaging to MR imaging. While certain information, such as size, shape and location are domain-independent and can be transferred to a new imaging modality, it remains challenging to determine which layers of the network store the desired transferable information.

With frequent sequence and scanner updates, algorithms trained using the pre-update annotated images may not work well for post-update newly-acquired images. Cross-modality learning can leverage the information learned from existing data to prevent the need to acquire and annotate new datasets under the new protocol.

7.2 Exploring potential limitations using an independent test dataset

As summarised in the previous part of this chapter, I could show in this thesis that all developed auto-segmentation approaches were achieving an accuracy close to the inter-observer variability. This was determined in a well-defined dataset (database 1). In clinical reality, there may be changes in how images are acquired, leading to differences, for instance, in contrast settings, image quality or field of view. To be aware of potential limitations and solutions to them, I applied atlas- and deep learning-based methods to a fully independent dataset.

7.2.1 Materials and methods

7.2.1.1 Data acquisition and preparation

The training data (database 1) comprised the library of 27 T2w images together with the manual delineation of a clinician, introduced in section 3.1 on page 39. For testing purposes, a library of T2w images of 13 healthy volunteers and 7 HNC patients (database 2) was acquired in the RT treatment position (following Schmidt and Payne [130]). Figure 7.2 illustrates the set-up of a volunteer on the MR scanner. Three coils were utilised: an 8-channel anterior coil, fixed to a bridge and folded around the volunteer or patient's head and neck, in conjunction with a 32-channel posterior coil, embedded into the table, as well as a 32-channel anterior coil to cover the lower part of the neck and the shoulders.

A T2w 3D turbo-spin-echo sequence was optimised with a varying flip angle for radiotherapy planning purposes, balancing scanning time against a high signal-to-noise-ratio. The final scanning time was approximately 7 minutes. The image acquisition parameters of both training and testing data are provided in table 7.2. For the testing

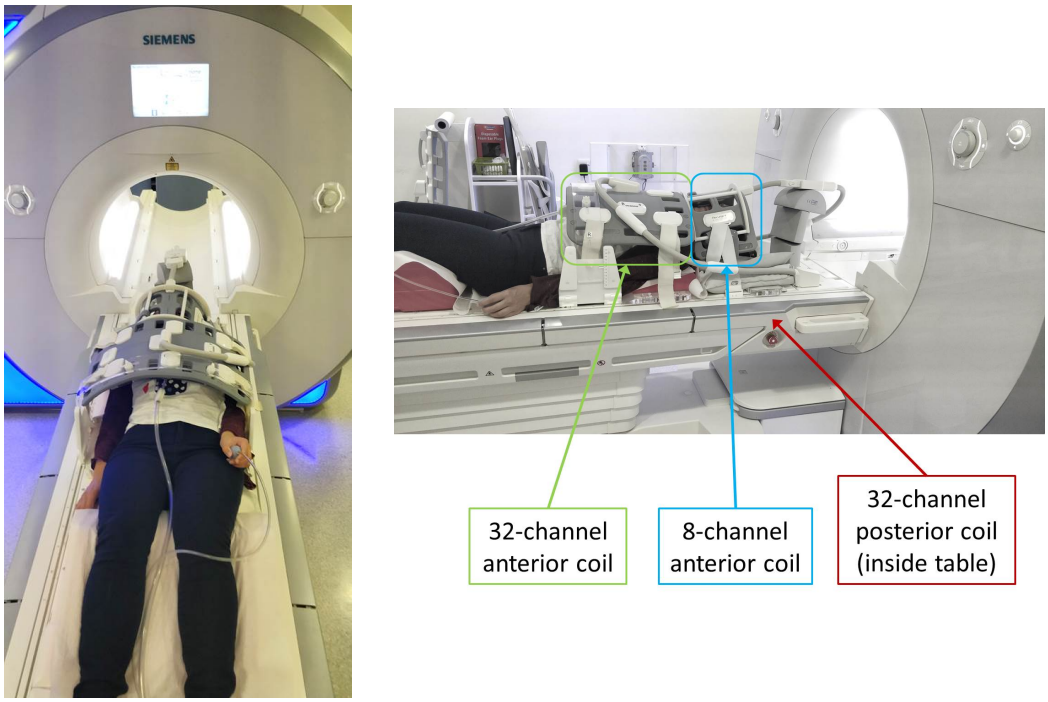


Figure 7.2: Scanning setup: This figure illustrates a top and side view of a volunteer, set up on the Siemens MR scanner. The respective coils are indicated in the picture in the second column.

Table 7.2: Image acquisition parameters of the two databases (database 1: training, database 2: testing).

parameter	database 1 (training)	database 2 (testing)
vendor	GE Healthcare	Siemens
sequence type	2D T2w spin-echo	3D T2w turbo-spin-echo
field strength	3 T	1.5 T
FOV [#pixels]	512 x 512	384 x 288
#slices	30	208
voxel size [mm^3]	0.5 x 0.5 x 4	0.78 x 0.78 x 0.80
phase encoding direction	unknown	anterior -> posterior
orientation	axial	sagittal
TE [ms]	[96.72, 107.30]	184
TR [ms]	[3198, 4000]	3500
varying flip angle [°]	90 (fixed)	120
number of excitations (averaging)	1	1.6
acceleration	unknown	GRAPPA (factor 3)

data, no gold standard contours were available. Figure 7.3 illustrates axial, sagittal and coronal views of example cases from the database. There were substantial differences between the two datasets, as images were acquired on different scanners from different vendors, as well as different sequences and algorithms were employed. These factors contributed to differences in signal-to-noise ratio, resolution or image quality.

To match the resolution of the testing and the training data, I resampled the test

images to a voxel size of 1x1x1 mm^3 and the training images to a voxel size of 1x1x4 mm^3 with a bilinear interpolation scheme.

7.2.1.2 Atlas-based approach

To auto-segment each image from database 2 (13 healthy volunteers, 7 patients), it was registered to each image of database 1 (27 patients), respectively, employing an affine registration as initialisation and refining it with a deformable registration. Details on the registration can be found in chapter 5. The test images (database 2) covered a much larger field of view along the axial direction (head to foot). Therefore, a preprocessing step was required to avoid a failure of the image registration. For this purpose, I reduced the field of view in this direction to a similar (smaller) coverage like for the training images (database 2). Figure 7.4 illustrates an example for this preprocessing step. I then employed a multi-atlas weighted majority voting, as described earlier in chapter 5.

7.2.1.3 Deep learning-based approaches

To choose suitable hyperparameters, the 27 patients from database 1 were randomly split into 80% for training and 20% for validation purposes. The trained model was then applied to all images from database 2. I chose the 2D method in this investigation. All deep learning-based approaches are described in detail in chapter 6.4.

7.2.2 Results

Without cropping the field of view for the atlas-based approach, the registration between training and testing images failed in all cases and hence, also the segmentation failed. I, therefore, manually cropped the images to a similar field of view, as shown in figure 7.4.

Figure 7.5 illustrates typical examples of atlas-based auto-segmented parotid glands. Most of the contours followed the boundaries of the parotid glands closely. There were some, usually isolated, voxels wrongly predicted as parotids, as indicated by the red arrows in figure 7.5. However, these could be easily removed in a fast postprocessing step.

For the deep learning-based approaches, the network would fail to segment the parotid glands without any adjustments of the image intensities. Despite both being T2w sequences, there were some substantial differences in both types of images, such as a 2D compared to a 3D acquisition and a magnetic field strength of 1.5 T compared to 3 T. While the image registration in the atlas-based segmentation was not sensitive to these differences, the deep learning-based method was. For this reason, any potential image property must be reflected in the training data, such as contrast, signal-to-noise

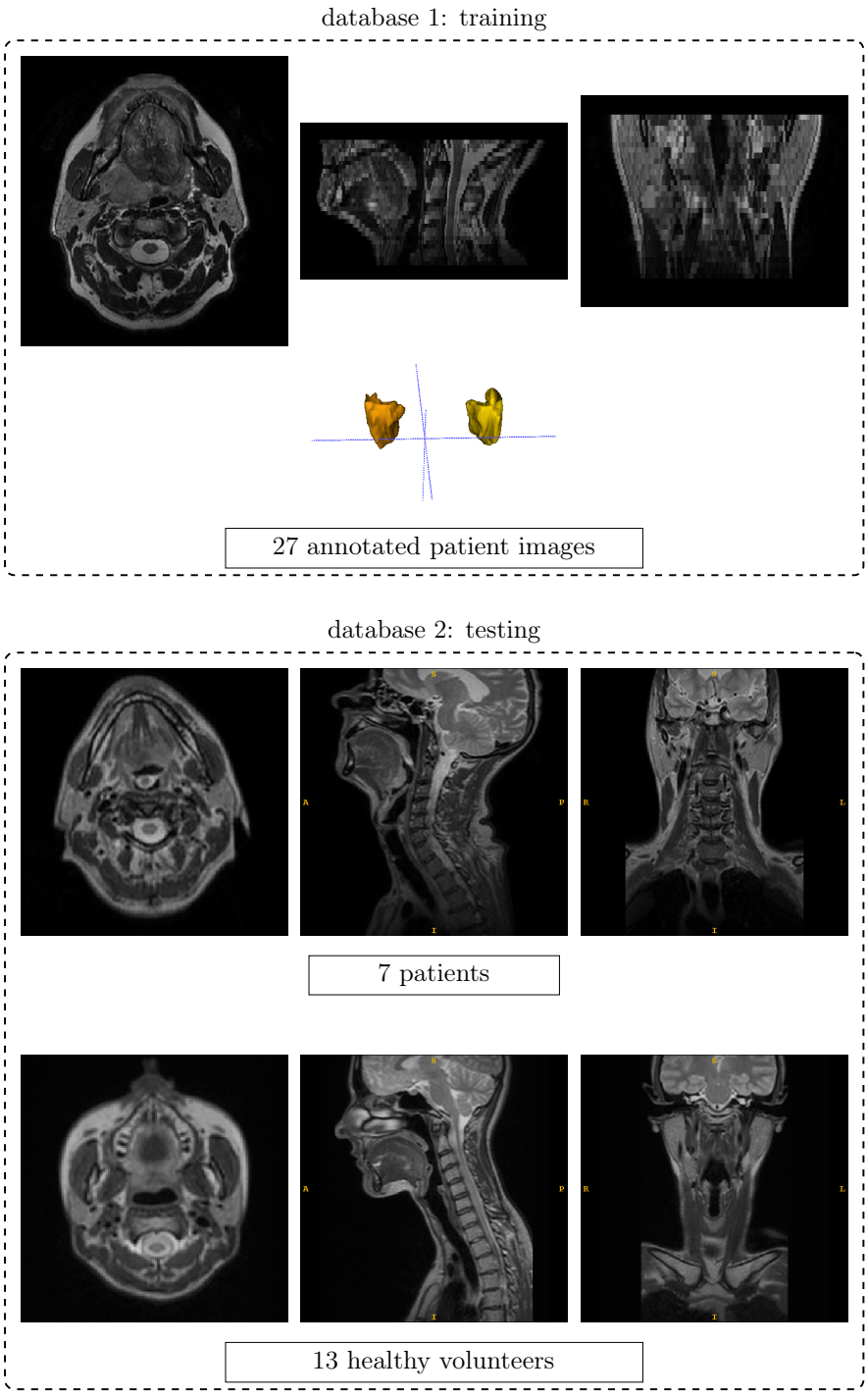


Figure 7.3: Test and training data: The first column exemplifies an axial, sagittal and coronal image of one patient from the training data used in this study. The second column illustrates axial, sagittal and coronal slices of a patient (first row) and a healthy volunteer (second row) from the testing data, respectively.

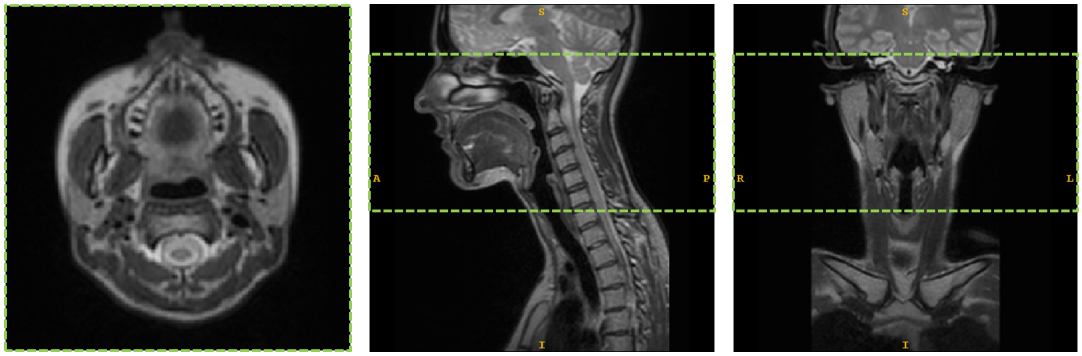


Figure 7.4: Preprocessing step for atlas-based segmentation: the field of view is cut (green dotted box) to match it to the training images.

ratio, sharpness or image quality. To mitigate this problem, I resampled the training images in such a way that they looked more like the testing images. For this purpose, I first downsampled them by a factor of 2x2 in the axial plane and then upsampled them again to the original resolution. I then retrained the network with the resampled training images. Figure 7.6 illustrates typical example cases comparing the atlas-based method to the deep learning-based method. In general, except for the brain region, the deep learning-based contours appear less fuzzy at the edges compared to the atlas-based contours. At the top borders of the image within the brain, the deep learning-based method tends to mispredict some voxels as parotids. However, these could easily be removed in a postprocessing step or by including slices within this region in the training data.

7.2.3 Discussion and conclusion

This thesis demonstrated that both, atlas- and deep learning-based method could accurately auto-segment parotid glands, which could then be used for RT treatment planning. In this chapter, I explored the potential limitations beyond a well-defined dataset on a new dataset qualitatively. A quantitative assessment of auto-segmentation accuracy requires observer delineations of ROIs. Unfortunately these were not available so this chapter provided a purely qualitative assessment. A quantitative assessment is left for future work.

The atlas-based method could cope with changes in resolution or image quality well. However, it was susceptible to the field of view and required matching this field of view in a preprocessing step.

The deep learning-based approach, on the other hand, was sensitive to the image sequence properties. However, it was straightforward to incorporate expected varieties within the training data. In this case, it was sufficient to simply blur the images with

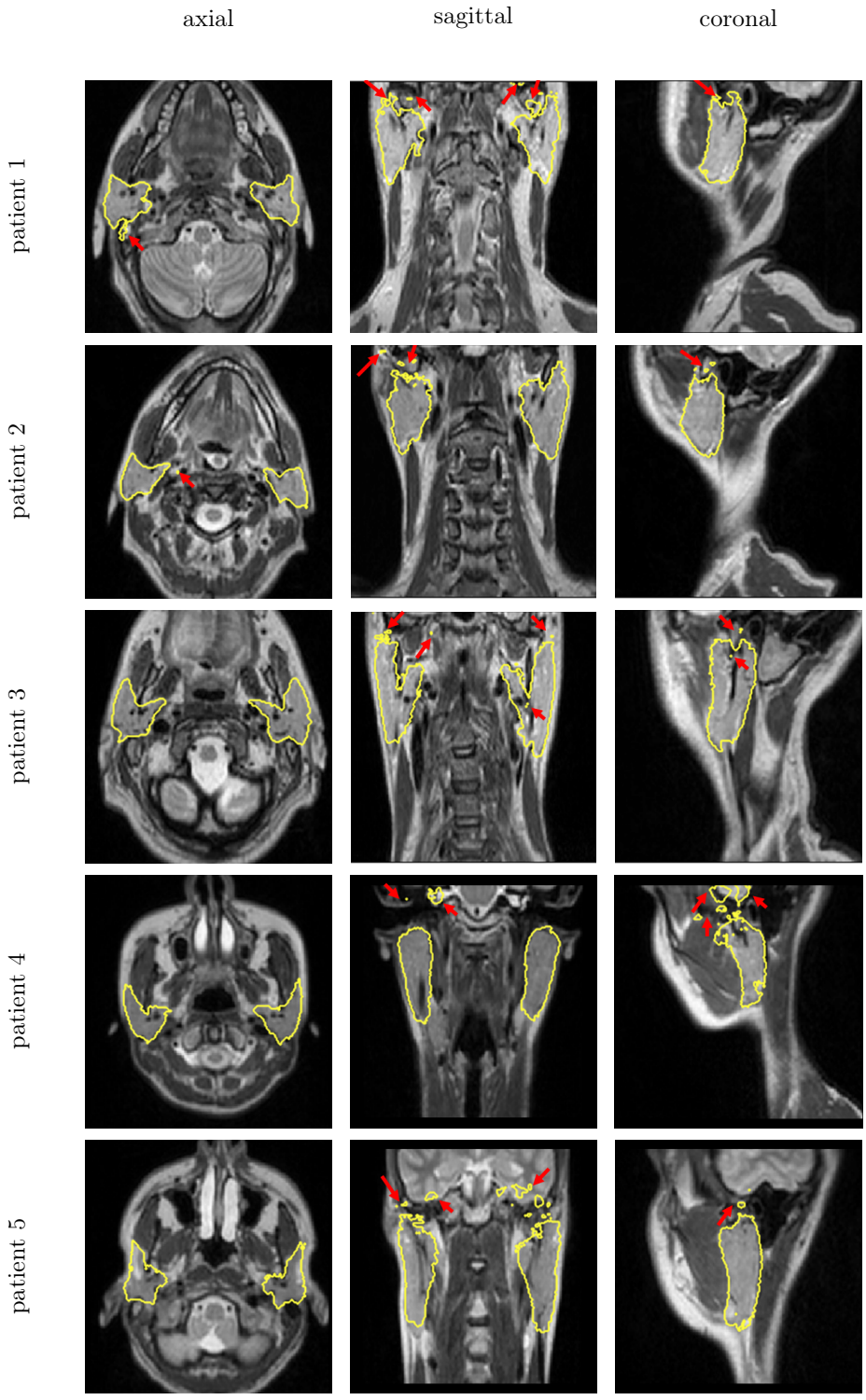


Figure 7.5: Segmentation examples: this figure provides 5 example cases of atlas-based auto-segmented parotid glands, denoted with the yellow line. The rows each show a typical example, whereas the columns illustrate the axial, sagittal and coronal cross-sections, respectively. There were some, usually isolated, voxels wrongly predicted as parotids, as indicated by the red arrows

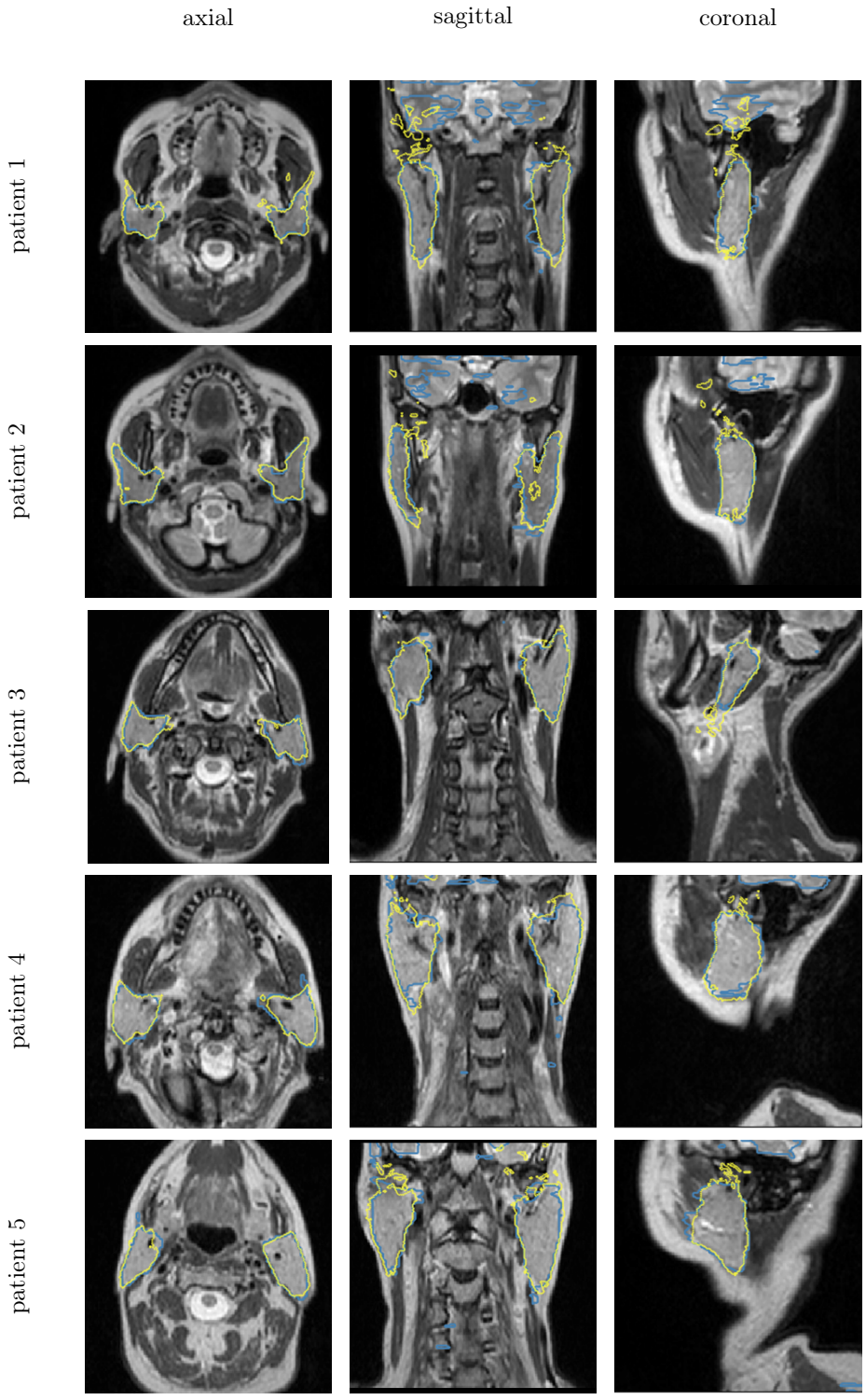


Figure 7.6: Segmentation examples: this figure provides 5 example cases for atlas- and deep learning-based auto-segmented parotid glands (yellow line: atlas-based, blue line: deep learning-based). The rows each show a typical example, whereas the columns illustrate the axial, sagittal and coronal cross-sections, respectively.

a downsampling and subsequent upsampling operation. For more complicated cases, where the contrast could have changed substantially, a promising solution would be the cross-modality learning approach, where one could transfer the expertise gained in the delineation of one type of sequence to any desired sequence through synthetic image generation (see chapter 6.6.)

While the accuracy was similar for both atlas-and deep learning-based methods, the deep learning-based methods outperformed the atlas-based ones to a large degree concerning the computation time (minutes or hours compared to seconds). Furthermore, with more input data, the computation time for atlas-based segmentation would increase due to a larger number of registrations, while in the case of deep learning, this would only affect the training time.

As an additional point, deep learning can cope with more unusual structures and can better handle varieties in shapes or locations, when shown many example cases. Therefore, deep learning-based approaches can be expected to outperform atlas-based approaches when applied to, and in the presence of, irregular structures such as tumours.

In conclusion, both atlas- and deep learning-based approaches seem promising applications to alleviate the enormous burden of manual segmentation and potential limitations can be fixed easily as we have shown in this chapter for two substantially varying MR sequences. With the collection of more data, I am confident that deep learning can outperform atlas-based methods in terms of accuracy and, more strongly, the computation time. Moreover, deep learning can better detect abnormalities, as well as segment normal tissues which have been infiltrated by abnormalities.

Chapter 8

Final discussion, conclusion and future directions

This chapter summarises the main findings of this thesis and their implications on future research, as well as potential clinical implementations. Moreover, the work is put in the broader perspective of automation in radiation therapy and I give a prognosis on its future.

8.1 Scope of this thesis

8.1.1 The role of auto-segmentation, MRI and adaptive RT

The segmentation of OARs is essential for the treatment planning of RT. To minimise side effects, such as swallowing dysfunction or dry mouth, it is crucial to accurately localise these organs and take them into account when planning the optimal configuration of beam intensities and arrangements. Conventional RT treatments rely on a snapshot of the patient's anatomy in time for the full treatment course of about six weeks, neglecting potential changes in the patient's anatomy due to weight loss, tumour or organ shrinkage and swelling. With the introduction of MRI-guided treatment systems, one can better visualise the soft-tissue contrast, which is crucial for many structures in the head and neck. Besides the advantage that MRI provides for image quality due to a better visualisation of soft tissues, as well as an extensive range of possible contrasts, it allows for functional imaging which can visualise surrogates, for instance for tumour metabolism, hypoxia, diffusion and perfusion. With these technological advances, real-time adaptive RT could be realised where, instead of one snapshot in time, frequent updates on the patient's anatomy and pathophysiological processes can be exploited for treatment adaptation.

To fully utilise the promises of these techniques, automation of the workflow is crucial. The current practice of manual segmentation by clinicians is time-consuming and tedious, especially for HNC patients due to many OARs to spare and target volumes to outline. Besides, there is generally no ground truth and, therefore, segmentation is subject to substantial inter- and intra-observer variabilities.

8.1.2 Aim of this thesis

The aim of this thesis was, therefore, to automate this tedious segmentation process. To the best of our knowledge, at the beginning of this thesis, there was no method available which could accurately and rapidly segment OARs on MR images of HNC patients. While previously atlas-based strategies were the state-of-the-art methods for auto-segmentation and are established now in many commercial treatment planning systems, artificial intelligence, in particular deep learning, has in recent years gained popularity for a variety of tasks. This popularity was mainly driven by the general availability of increased computational power, enabling to calculate and store the usually memory-heavy deep learning-derived models.

8.2 Main findings and applicability of this thesis

Following atlas-based (chapter 5) or deep learning-based (chapter 6) methods, I could demonstrate that it was feasible to automatically segment OARs with a geometric accuracy comparable to the measured inter-observer variability.

8.2.1 Atlas-based segmentation

I demonstrated that atlas-based segmentation was capable of generating clinically acceptable contours on both, T1w (chapter 4) and T2w (chapter 5) MR images. However, it was not suitable for daily adaptations due to relatively long computation times. Furthermore, this registration-based approach was sensitive to global features such as the head posture or the field of view, as I have shown in chapter 7. In addition to that, atlas-based segmentation performs well for organs which have similar shapes and locations for different patients. As this approach is based on an image registration between different patients, this property would become a challenge in tumour segmentation.

8.2.2 Deep learning-based segmentation

The computational burden of conventional auto-segmentation algorithms can be alleviated with deep learning-based methods, with prediction times in the order of seconds or even sub-seconds (see chapter 6). Furthermore, deep learning-based methods are more flexible and can be used to segment ROIs which vary in shape and location, such as involved lymph nodes or tumour volumes, as we have demonstrated recently [55].

8.2.3 Applicability of developed algorithms

In this thesis, I chose one clinical indication, HNC, and segmented OARs with a focus on the parotid glands. The parotid glands are crucial organs to spare, such that severe side effects can be reduced or avoided. In addition, they can vary significantly in shape and location in a patient cohort and are typically close to target regions, rendering auto-segmentation a challenging task. I am confident that the developed methods can be generalised for other clinical indications and ROIs. We have already demonstrated in related work that one could apply the same deep learning algorithm to the segmentation of involved lymph nodes on diffusion-weighted MR images [55].

8.3 Validation of auto-segmentation algorithms

8.3.1 Suitable evaluation measures

Currently, auto-segmentation algorithms are commonly assessed by measures of volume overlaps or geometrical distances to a reference segmentation. However, for an application of auto-segmentation within an RT planning scenario, these geometric metrics are not necessarily meaningful. For example, a disagreement between manual and automated delineations in a region with a large dose gradient is likely to have a much more significant impact on an over-dosage to OARs or under-dosage to target volumes than in a region that only receives low doses. This emphasises the need for the establishment of additional, more meaningful evaluation metrics. I have demonstrated in this thesis (see chapter 4) that geometric measures alone are not sufficient to predict the impact of inaccurate segmentation on RT planning. One, thus, needs to determine the impact of segmentation errors on the planned dose distributions for a thorough evaluation of auto-segmentation algorithms prior to their clinical implementation.

The computational burden of needing to calculate treatment plans for a dosimetric evaluation could, for instance, be solved by learning from a large database of combined information on the locations of the ROIs and the planned dose distributions. Moreover, one could employ an RT-specific loss function in the training of deep learning-based algorithms which could, for example, incorporate the distance to overlap with the target volume.

8.3.2 Lack of ground truth

A challenge for a quantitative evaluation of auto-segmentation algorithms remains with the definition of "expert" performance. The lack of objective reference annotations hinders an unbiased quantification. This is problematic not only for the evaluation but also for the development of auto-segmentation algorithms itself, as manual annotations are commonly used as input in a supervised manner. This way, an algorithm can only mimic the human performance and replicate potential errors. To mitigate this limitation, one could combine the prediction of multiple experts to obtain a more accurate estimation of the ground truth, as for example achieved with the STAPLE method [155].

8.4 Change of clinical practice

8.4.1 Clinical implementation of auto-segmentation

I am convinced that adaptive RT will be one of the main driving forces for full automation of routine tasks in the clinical RT workflow in the near future due to the necessity for full exploitation of a large quantity of imaging data. This will pave the way for automation, in particular, of segmentation, in the general clinical RT workflow.

While deep learning-based algorithms are still "black boxes" to a large degree [168], it is easy for a human observer to verify the resulting automated delineation. For these reasons, auto-segmentation could be implemented in the clinic now, as a starting point for clinicians with the possibility to edit the proposed contours. I believe that auto-segmentation will gradually replace manual delineation within the next five to ten years, as clinicians gain confidence in the technique by observing its results. This would allow them to focus on different tasks.

Several vendors are already offering auto-segmentation algorithms in their commercial treatment planning systems, such as Eclipse (Varian Medical Systems, Palo Alto, California), Monaco (Elekta AB, Stockholm, Sweden) and RayStation (Raysearch, Stockholm, Sweden), and are currently investigating deep learning-based algorithms.

8.4.2 Margins and uncertainties

Currently, large margins are used for the CTVs to account for uncertainties, for instance, in the patient setup or the delineation of ROIs. With rapid daily image segmentation, these margins can be reduced since uncertainties in the location of ROIs will decrease with daily image guidance and adaptation.

Moreover, Bayesian deep learning-based methods can simulate the delineation uncertainty intrinsically [45, 133] and therefore remove the need to employ large population-based margins for uncertainties in the delineation (expansion from the CTV to the PTV). Instead, they could facilitate the introduction of non-isotropic margins, whose local extension depends on the uncertainties of the target outline at the considered spatial point. Furthermore, combining uncertainty predictions with better imaging techniques may allow for shrinkage of the large CTVs which are currently necessary to include potential regional involvement of lymph nodes and sub-clinical tumour spread. With this, toxicities to the patient could be significantly reduced and a better quality of life after treatment with RT may be achieved.

8.5 Deep learning algorithms are data-hungry

8.5.1 Need for large and consistent databases

A significant barrier to deep learning in medical imaging and, in particular, RT are small databases and a large variety in imaging protocols and guidelines. For this reason, an easy pitfall is to overfit the limited data at hand but missing out on generalisability. Extensive validation of such algorithms is essential to ensure safe clinical practice. Hence, a crucial step is to establish large, publicly available databases with consistent ground truth annotations. These large databases could prove as RT-community-wide testing frameworks. Hospitals need to work together to gather these large databases. With initiatives such as the MR-Linac consortium, such approaches may become a reality as imaging data will be shared and standardised according to specific guidelines.

8.5.2 Generative adversarial networks for image synthesis

I have shown that GANs hold a tremendous promise in overcoming limitations due to data sparsity (see chapter 6.6). GANs can help in exploring and discovering the underlying structure of training data and learn to generate new images. These approaches can be used to augment the small annotated datasets, typically available in medical imaging. Furthermore, GANs can, for instance, enable MR-only treatment workflows where a synthetic CT replaces the conventionally used planning CT with no additional imaging dose to the patient. This way, the need for image registration between MR and CT images would be limited, which is prone to uncertainty [159].

Typically, multiple MR sequences are routinely acquired due to the complementary information they provide. GANs could reduce the time spent on the acquisition of multiple sequences by reducing the number of acquisitions through learning from large databases of previous patients with multiple sequences [165]. However, with image generation methods, caution has to be taken as the training data distribution can differ substantially from the actual data distribution in a real-world application. Applications of GANs are still in their infancy but are likely to become the hot topic of the coming years.

8.6 Artificial intelligence in RT beyond auto-segmentation

8.6.1 Incorporating more knowledge

Recent trends move towards incorporating more information into deep learning-based algorithms [104, 112]. This may, for instance, include genetic data, patient demographics

and multi-modality imaging data. At this stage, fully exploiting multiple modalities is challenging since individual modalities typically possess different properties, such as noise levels and dimensionality. Additive neural networks extract features from each network and combine them at a later stage in a shared network. However, weighing the importance of individual modalities is still challenging and subject to future research [102].

8.6.2 Decisions and outcome prediction

In the more distant future, artificial intelligence could be applied to decision-making systems, helping clinicians to base their decisions on a collection of larger datasets which are difficult to process by humans. These applications could be recommendations on a specific therapy, or prediction and prognosis of a certain type of treatment. For example, one could predict the likelihood of patient survival and treatment-related toxicities with the delivered dose to target volumes and OARs. For this endeavour, a more systematic collection of data would be required compared to how it is currently typically done. Those tasks pose a larger challenge in terms of their validation compared to auto-segmentation, as quality-assurance is challenging. While deep learning-based approaches developed with large databases have the potential to guide clinicians in their decision making, it is difficult to justify a decision without knowing the reasons why the algorithm has drawn a specific conclusion. It might, hence, be beneficial to split larger tasks into smaller sub-tasks to be able to quality-assure each step on the way. Furthermore, active research is underway to open the black box of deep learning algorithms [32, 168].

8.6.3 Caveat

While deep learning-based approaches have demonstrated promising results and there is a large and increasing number of publications in this field, care has to be taken as to whether these approaches can generalise well. Furthermore, deep learning-based approaches are not always the most suitable approach, so one has to carefully select where applications other than deep learning-based approaches are sufficient or could even perform better.

8.7 Conclusion

In conclusion, I am confident that auto-segmentation will soon be routinely used in the clinical workflow and that, more generally, many repetitive tasks of the RT treatment process can be automated to assist humans. I believe that automated systems can reduce human errors, improve consistency and create more time for other tasks. Automation can augment, rather than replace, clinicians and can, therefore, have an enormous impact on the quality of treatment outcome.

Appendix A

Publications

A.1 Paper

The following paper has been peer-review and was published as

J. P. Kieselmann, C. P. Kamerling, N. Burgos, M. J. Menten, C. D. Fuller, S. Nill, M. J. Cardoso, and U. Oelfke. "Geometric and dosimetric evaluations of atlas-based segmentation methods of MR images in the head and neck region". *Physics in Medicine and Biology* **63** 145007 (2018).

A.2 Conference abstracts

A.2.1 MR in RT 2017

This conference abstract has been peer-reviewed and presented at MR in RT 2017.

PRESENTATION TITLE

Geometric and dosimetric evaluation of three atlas-based segmentation methods for head and neck cancer patients on MR images

AUTHOR(S)

Jennifer P Kieselmann[1]*; Cornelis P Kamerling[1]; Ninon Burgos[2]; Martin J Menten[1]; Simeon Nill[1]; M Jorge Cardoso[2]; Uwe Oelfke[1]
[1]Joint Department of Physics, The Institute Of Cancer Research and The Royal Marsden NHS Foundation Trust, London, UK
[2]University College London, Translational Imaging Group, London, UK

ABSTRACT

Purpose: To evaluate three atlas-based segmentation methods in terms of their geometrical accuracy and estimate the consequential dosimetric impact on head and neck (HN) radiotherapy treatment planning (TP).

Materials & Methods: A clinician manually segmented right (RP) and left parotids (LP), spinal cord (SC) and mandible (MD) on pre-treatment T1w MR images of 13 HN patients. We used these segmented images as database to perform three atlas-based auto-segmentation approaches of the organs-at-risk (OARs) following a leave-one-out cross-validation strategy: best-atlas (bestAtlas) and two multi-atlas approaches. The latter ones are based on two distinct atlas-fusion methods, a weighted majority voting (maWMV) [1] and an iterative approach called STEPS (maSTEPS) [2].
The manual segmentations were used as ground truth. We determined the geometrical differences between automated and manual segmentations by calculating Dice similarity coefficients (DSC), standard and 95% Hausdorff distances (HD and HD95, respectively). We investigated the significance of the differences between the three approaches with a paired t-test at a significance level of $p<0.05$.
We determined the dosimetric impact of the geometrical differences found in the first part of this work on treatment plans obtained using the TP system Monaco (Elekta AB, Sweden) with the MR-Linac machine model. For this purpose, we mapped the segmented OARs from the MRIs to the corresponding pre-treatment CTs utilising the deformable image registration tool ADMIRE in Monaco. A clinician segmented the primary (tumour including involved lymph nodes) and secondary (elective target including non-involved lymph nodes) CTVs, the brainstem, optical nerves and lenses on an arbitrary subset of 6 patients.
For each set of auto-segmented OARs, we generated a clinically acceptable plan for a 9-beam IMRT treatment (65 and 54 Gy in 30 fractions), defined according to the INSIGHT trial protocol [3]. Mean dose to the parotids must not exceed 30 Gy, dose to $1cm^3$ of the SC needs to be below 46 Gy and hot spots in the mandible should not exceed the prescribed target dose.
Using these plans we investigated the dosimetric effects on the manually segmented OARs by testing their fulfilment of the imposed planning constraints.

Results: Figure 1 presents the geometrical accuracy of the three sets of auto-segmented OARs. The treatment plans, optimised for the OARs auto-segmented with maWMV and maSTEPS, resulted in a fulfilment of all imposed constraints on the ground truth OARs for all 6 patients. For the bestAtlas, a violation of the constraints on one of the manually segmented parotids occurred for 2 patients with a mean parotid dose exceeding the constraint by more than 2 Gy. Figure 2 illustrates the DVHs and a slice of the dose distribution for two example cases.

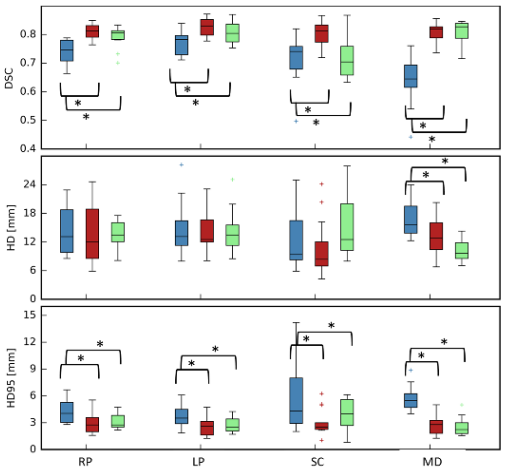


Figure 1: Boxplot illustrating the geometrical differences of the three automated segmentations to the manual segmentation: bestAtlas (blue), maWMV (red) and maSTEPS (green). The vertical axes refer to the DSC, HD and HD95 from top to bottom, whereas the horizontal axis refers to the four investigated organs at risk, the LP, RP, SC and MD, respectively. Smaller values for HD and HD95 and larger values for DSC mean higher accuracy. A star illustrates statistical significance.

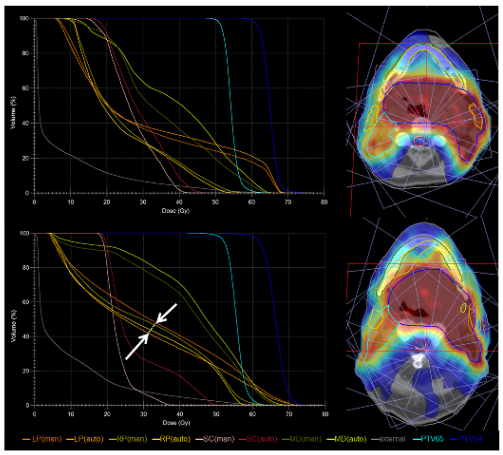


Figure 2: Two examples of DVHs and an illustration of the dose distribution of one image slice for manual and auto-segmented OARs, one satisfying (top row, maWMV) and one violating (bottom row, bestAtlas) the planning constraints. The white arrows indicate the region where constraints are violated. Even though differences for the MD in the top row are large, all constraints are satisfied here.

Conclusions: This preliminary study demonstrates the feasibility of auto-segmentation following atlas-based approaches in terms of geometric accuracy, as well as achieving clinically acceptable treatment plans. Multi-atlas-based approaches outperformed a simple best-atlas approach. HD was not reliable to predict dosimetric impact. Although there appears to be a slight correlation between geometric (DSC and HD95) and dosimetric measures, the geometric measures alone are not sufficient to predict the dosimetric impact of inaccuracies on TP for the data utilised in this study.

Acknowledgements: We would like to thank Y. Ding and M. Jomaa from the MD Anderson Cancer Center (Houston, Texas) for providing patient data. We acknowledge I. Petkar, G. McCormick, A. Hunt and F. Wang for the manual segmentations. This work has been kindly supported by the Oracle Cancer Trust. The ICR is supported by Cancer Research UK under programme C33589/A19727. ICR/RMH is part of the Elekta MR-Linac Atlantic Research consortium.

References:

[1] Cardoso *et al*, IEEE Transactions on Medical Imaging **34** (2015): 9, 1976-88
[2] Cardoso *et al*, Medical Image Analysis **17** (2013): 6, 671-684
[3] Welsh *et al*, Radiation Oncology **10** (2015): 112

A.2.2 MR in RT 2018

This conference abstract has been peer-reviewed and presented at MR in RT 2018.

6th MR in RT Symposium 2018

PRESENTATION TITLE

Application of a deep convolutional neural network to segment the parotid glands on MR images

AUTHOR(S)

Jennifer P Kieselmann*; Simeon Nill; Uwe Oelfke

Joint Department of Physics, The Institute Of Cancer Research and The Royal Marsden NHS Foundation Trust, London, UK

ABSTRACT

Purpose: Deep convolutional neural networks (CNNs) have shown great success in solving computer vision tasks, such as object detection and segmentation. To date, there has been only limited application of CNNs to the segmentation of organs-at-risk (OARs) on magnetic resonance (MR) images. We developed and evaluated a CNN-based approach to automatically segment the parotid glands on MR images of head and neck cancer (HNC) patients.

Materials & Methods: The proposed method builds upon recent developments in the computer vision community. We implemented the 2D U-net architecture [1], illustrated in Figure 1, using Keras [2] and TensorFlow [3]. Our database consisted of pre-treatment T1-weighted MR images of 13 HNC patients. Each MR image comprised 30 axial slices with 512 x 512 pixels, a slice distance of 4 mm and an in-plane resolution of 0.5 x 0.5 mm^2. Each slice was downsampled by a factor of 4 before being fed into the network. A clinician manually contoured the parotid glands, which served as the ground truth.
For each patient, a leave-one-out cross-validation approach was followed with the test data comprising slices of the patient's image and the training data slices from the remaining patients' images. The training data were further split into 90% training and 10% validation in order to optimise hyper-parameters, such as the number of training epochs and the learning rate. One epoch referred to one training cycle over the full training data. We chose 50 epochs and optimised a weighted cross-entropy cost function, accounting for the imbalance between background and foreground pixels. We used the Adam optimiser [4] with a learning rate of 0.00001. Geometric differences between the ground truth and the CNN-derived segmentations were evaluated by calculating the 3D Dice similarity coefficients (DSC) and 95-percentile Hausdorff distances (HD95). Due to their symmetry, we simultaneously segmented both parotid glands and divided them into the left and right part in a post-processing step.
Run time was determined for programme execution on a single NVIDIA Titan Xp GPU with 12 GB VRAM.

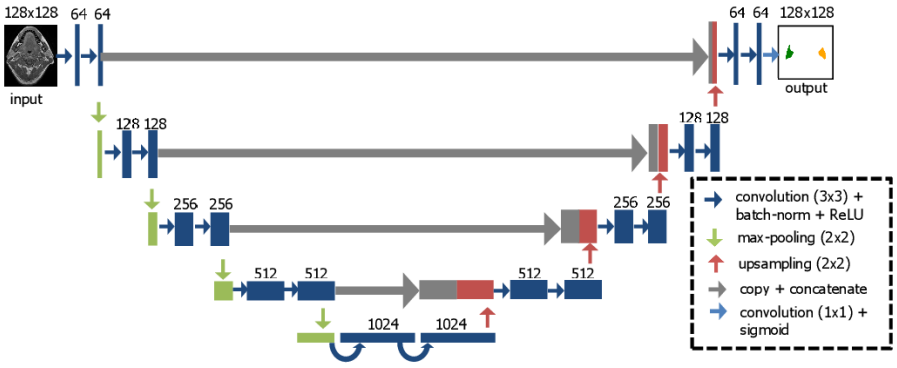


Figure 1: The architecture of U-net [1] with its cascade of operations. The legend in the dotted box lists the applied operations. Operations on the left hand side represent the encoding, on the right hand side the decoding path. Numbers indicate the number of features.

Results: Figure 2 shows three representative slices from three different patients. Mean DSC ± one standard deviation were 0.82 ± 0.03 for the left and 0.78 ± 0.04 for the right parotid. The mean HD95 were 2.78 ± 0.96 mm for the left and 3.18 ± 1.24 mm for the right parotid. Mean training time was 35.86 ± 0.22 min, whereas mean inference time for the full volume of one image was 0.84 ± 0.05 s.

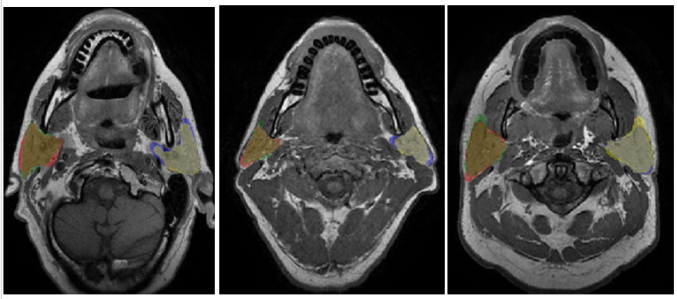


Figure 2: Three representative MR-slices from three different patients. Coloured areas represent manually (red) and auto-segmented (green) right parotid glands, as well as manually (blue) and auto-segmented (yellow) left parotid glands.

Conclusion: This preliminary study demonstrates great potential for the application of CNNs to segment OARs for RT treatment planning purposes with the accuracy comparable to state-of-the-art approaches such as atlas-based segmentation (ABS). In comparison to ABS, the segmentation time is much smaller (sub-second compared to minutes or hours). A limiting factor of this study was the amount of training data. With the availability of more training data, the addition of T2-weighted MRI, extended dimensionality to 3D and techniques such as data augmentation we expect to further improve accuracy.

References:

[1] O. Ronneberger et al., MICCAI, Vol. 9351, 2015, pp. 234–241.
[2] F. Chollet et al, Keras, GitHub, 2015, https://github.com/keras-team/keras
[3] M. Abadi et al., TensorFlow: Large-scale machine learning on heterogeneous systems, 2015.
[4] D. Kingma and J. Ba, preprint arXiv:1412.6980, 2014.

A.2.3 MR in RT 2019

This conference abstract has been peer-reviewed and presented at MR in RT 2019.

Cross-modality deep learning: Contouring of MRI data from annotated CTs only

Jennifer P Kieselmann, Oliver J Gurney-Champion, Brian Hin, Clifton D Fuller, Simeon Nill, Uwe Oelfke

Purpose:

The clinical introduction of MR-guided RT systems enables daily treatment adaptations. To fully utilise benefits from daily adaptations, daily delineation of volumes of interest (VOI) is required. This time-consuming process, known to be subject to large inter-expert variabilities (IEV), would benefit from automation. Therefore, we aim to automate delineation of VOIs using a deep convolutional neural network (CNN). Due to a lack of annotated training data, only few studies on using CNNs to segment VOIs on MR images have been reported. A general approach to tackle the lack of training data is to augment data with random rotations and translations. However, these simple methods do not capture large variabilities existing in the full population of patients' anatomies. Furthermore, this is only feasible if there is already sufficient annotated data available and both annotation and training need to be repeated for every novel MR contrast setting. In this study we solve these shortcomings by exploiting the wealth of publicly available annotated head and neck (H&N) CT images and generating synthetic MR images via an image style transfer network, instead. We then use these synthetic MR images to train a 2D CNN to automatically contour the parotid glands on real MR data.

Materials and Methods

Imaging data consisted of a CT and an MR database of H&N patients. The MR database contained 27 T2-weighted pre-treatment images (256x256x30 voxels of 1x1x4 mm^3; 3T scanner) acquired at the MD Anderson Cancer Center (Houston, USA). The CT database contained 202 CT images from the Cancer Imaging Archive and MICCAI H&N segmentation challenge. Each database included manual contours of the parotids. We performed constrast stretching for both CT and MR images and mapped intensities to a common range. The unpaired CT and MR images were then fed into a 2D CycleGAN network to generate synthetic MR images for each of the 202 CT images. Annotations of the synthetic MR images were generated by propagating the CT contours. We consequently selected synthetic MR axial slices to train a 2D CNN using the U-Net architecture. Prediction was performed on the 27 3D real MRIs. Performance was evaluated by calculating the Dice similarity coefficient (DSC) and mean surface distance (MSD) between manual and auto-generated contours and compared to the pairwise IEV between 3 observers.

Results

With an average DSC of 0.76 ± 0.07 and an average MSD of 2.57 ± 0.91 mm, the accuracy was close to the MR IEV of 0.84 ± 0.06 (DSC) and 1.50 ± 0.77 mm (MSD), respectively.

Conclusions

This technique of cross-modality learning can be of great value for segmentation problems where not a lot of annotated training data is available. We anticipate using this method with any small MR training dataset to generate synthetic MR images of the same type via image style transfer from CT images. Furthermore, as this technique allows for fast adaptation of annotated datasets from one imaging modality to another, it could prove to be useful for translating between large varieties of MRI contrasts due to differences in imaging protocols within and between institutions.

A.2.4 ICCR 2019

This conference abstract has been peer-reviewed and presented at ICCR 2019.

19th International Conference on the use of Computers in Radiation Therapy

Transfer learning from CT to MRI: auto-segmentation of the parotid glands using a deep convolutional neural network approach

Jennifer Kieselmann[†1], Oliver Gurney-Champion[1], Brian Hin[1], Simeon Nill[1], Clifton Fuller[2], Uwe Oelfke[1]

[1]*Joint Department of Physics, The Institute of Cancer Research and The Royal Marsden NHS Foundation Trust, 15 Cotswold Road, London, SM2 5NG, UK*

[2]*Department of Radiation Oncology, MD Anderson Cancer Center, Houston, TX, USA*

Introduction

Delineation of volumes of interest (VOIs) for radiotherapy (RT) is a time consuming process and, with the aim of daily treatment adaptations with, e.g., magnetic resonance (MR) guided RT systems, will have to be repeated for each treatment fraction. Additionally, it is known that delineation is subject to large inter-expert variabilities (IEV). We aim to automate delineation of the parotids for adaptive MR-guided RT with a deep convolutional neural network (CNNs). To date, only few studies on using CNNs to segment VOIs on MR images have been reported because of a lack of annotated training data. In this work we exploited the wealth of publicly available contoured head and neck (H&N) CT images to train a 2D CNN and subsequently fine-tune the learned model by training the network on MR images using transfer learning.

Materials & Methods

Data acquisition and preparation: Imaging data consisted of a CT and an MR database of H&N cancer patients. The MR database contained 27 T2-weighted pre-treatment images (256x256x30 voxels of 1x1x4 mm^3; 3T scanner) acquired at the MD Anderson Cancer Center (Houston, Texas, USA). A clinician manually contoured the parotid glands on all images, using the treatment planning system Raystation (Raysearch, Stockholm, Sweden). The CT database contained 202 downloaded CT images from the Cancer Imaging Archive [1] and the MICCAI H&N segmentation challenge [2], together with manual segmentations of the parotid glands. We downsampled CT and MR images by a factor of 2x2 in-plane and rescaled intensities of CT images (level 40, window size 350 Hounsfield Units). We mapped image intensities for both CT and MR images to a common range.

Automated segmentation: We implemented a 2D U-Net [3] in Python using Keras [4] and Tensorflow [5]. Commonly in transfer learning, the pre-trained weights referring to shallower levels, describing edges and primary shapes, are kept unchanged. Only deeper levels, describing complex image composition, are adapted as these are more specific to the underlying problem. The information we would like to transfer from the CT to the MR segmentation problem are the variety of shapes and locations of the parotids. Due to the architecture of the U-Net, the desired information regarding the location of objects (i.e. parotids) in the image is likely learned in the deeper levels, whereas information on shape is contained in the shallower levels. We therefore explored three different approaches: (1) keeping the weights fixed in the encoding part (shallower levels), (2) keeping the weights fixed in the decoding part (deeper levels) and (3) retraining the full network. The CT network was trained for 70 epochs, optimising a Dice loss objective function. We used the Adam optimiser [6] with a learning rate of 1e-05. The fine-tuning was performed with training for 60 epochs and a learning rate of 1e-04. For the fine-tuning, we performed a 9-fold cross validation (for each fold 24 patients for training, 3 for testing the network).

Evaluation: We evaluated the performance by calculating geometric differences between the manual and the CNN-derived segmentations in the original image resolution, using the Dice similarity coefficient (DSC) and the mean surface distance (MSD) as metrics. The run time was determined for programme execution on a single Tesla V100 with 16 GB VRAM. We furthermore compared its performance to the IEV, determined by comparing manual segmentations from three different experts [7].

Results

Keeping the weights fixed for the (1) encoder or (2) decoder part of the network led to a poor accuracy in the prediction of the segmentation with average DSC of 0.12 and 0.53, respectively. We therefore chose to fine-tune all pre-trained weights of the U-Net. Figure 1 shows how our network closely follows the expert delineation in three representative slices from three different patients. Table 1 shows that the mean DSC and MSD of our network (approach (3)) are close to the IEV. Mean training time in the fine-tuning was 16.99 ± 0.16 min. The mean inference time for the full parotid glands on all slices of one image was 0.70 ± 0.15 s.

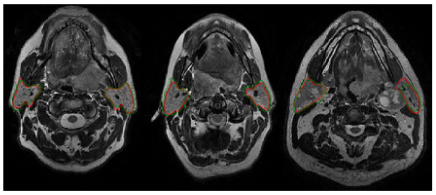


Figure 1: *Representative MR slices from three different patients. The coloured lines represent manually (red) and auto-segmented (green) parotid glands.*

VOI	approach	DSC	MSD [mm]
Right parotid	*TL*	0.80 ± 0.06	2.31 ± 1.41
	IEV	0.84 ± 0.04	1.40 ± 0.45
Left parotid	*TL*	0.82 ± 0.05	1.83 ± 0.94
	IEV	0.83 ± 0.04	1.59 ± 0.63

Table 1: *Quantitative results comparing the different transfer learning (TL) methods to the inter-expert variability (IEV).*

Discussion & Conclusions

This study demonstrates the suitability of the application of CNNs to segment OARs for RT treatment planning purposes, where the accuracy is comparable to state-of-the-art approaches such as atlas-based segmentation but the computation time was much smaller (sub-second compared to minutes/hours). We believe this technique will play a great role in re-contouring for daily treatment adaptation in MR-guided RT. In future work we will investigate how to further improve the transfer of feeding information on the learned anatomy into the training process and apply this approach to other VOIs.

References

[1] A. Grossberg et al., doi: 10.7937/K9/TCIA.2017.umz8dv6s
[2] P. Raudaschl et al., Med Phys (2015), 44(5), 2020-2036.
[3] O. Ronneberger et al., MICCAI, Vol. 9351, 2015, pp. 234–241.
[4] F. Chollet et al. Keras, GitHub, 2015, https://github.com/keras-team/keras.
[5] M. Abadi et al., TensorFlow: Large-scale machine learning on heterogeneous systems, 2015.
[6] D. Kingma and J. Ba, preprint arXiv:1412.6980, 2014.
[7] J. Kieselmann et al., Phys. Med. Biol. (2018), 63 145007.

†Corresponding Author: jennifer.kieselmann@icr.ac.uk

References

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mane, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viegas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. "TensorFlow: Large-scale machine learning on heterogeneous distributed systems" (2016). arXiv: 1603.04467.

[2] P. Aljabar, R. A. Heckemann, A. Hammers, J. V. Hajnal, and D. Rueckert. "Multi-atlas based segmentation of brain images: atlas selection and its effect on accuracy". *NeuroImage* 46.3 (2009), pp. 726–738.

[3] J. L. Barker, A. S. Garden, K. K. Ang, J. C. O'Daniel, H. Wang, L. E. Court, W. H. Morrison, D. I. Rosenthal, K. S. C. Chao, S. L. Tucker, R. Mohan, and L. Dong. "Quantification of volumetric and geometric changes occurring during fractionated radiotherapy for head-and-neck cancer using an integrated CT/linear accelerator system". *International Journal of Radiation Oncology Biology Physics* 59.4 (2004), pp. 960–970.

[4] W. J. Beasley, A. McWilliam, A. Aitkenhead, R. I. Mackay, and C. G. Rowbottom. "The suitability of common metrics for assessing parotid and larynx autosegmentation accuracy". *Journal of Applied Clinical Medical Physics* 17.2 (2016), p. 5889.

[5] J. Bergstra and Y. Bengio. "Random search for hyperparameter optimization". *Journal of Machine Learning Research* (2012).

[6] M. A. Bernstein, K. F. King, and X. J. Zhou. *Handbook of MRI pulse sequences*. 2004. ISBN: 9780080533124.

[7] F. Bloch. "Nuclear induction". *Phys. Rev.* 70 (7-8 1951), pp. 460–474.

[8] T. Bortfeld. "IMRT: a review and preview." *Physics in Medicine and Biology* 51.13 (2006), R363–79.

[9] K. Brock, M. Sharpe, L. Dawson, S. Kim, and D. Jaffray. "Accuracy of finite element model-based multi-organ deformable image registration". *Medical Physics* 32.6 (2005), pp. 1647–1659.

[10] R. W. Brown, Y. C. N. Cheng, E. M. Haacke, M. R. Thompson, and R. Venkatesan. 2014. ISBN: 9781118633953.

[11] *Cancer Research UK*. URL: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/head-and-neck-cancers (visited on 04/19/2019).

[12] C. E. Cardenas, J. Yang, B. M. Anderson, L. E. Court, and K. B. Brock. "Advances in auto-segmentation". *Seminars in Radiation Oncology* 29.3 (2019), pp. 185–197.

[13] M. J. Cardoso, M. Modat, R. Wolz, A. Melbourne, D. Cash, D. Rueckert, and S. Ourselin. "Geodesic information flows: spatially-variant graphs and their application to segmentation and fusion". *IEEE Transactions on Medical Imaging* 34.9 (2015), pp. 1976–1988.

[14] M. J. Cardoso, M. J. Clarkson, G. R. Ridgway, M. Modat, N. Fox, and S. Ourselin. "LoAd: A locally adaptive cortical segmentation algorithm". *NeuroImage* 56.3 (2011), pp. 1386–1397.

[15] M. J. Cardoso, K. Leung, M. Modat, S. Keihaninejad, D. Cash, J. Barnes, N. C. Fox, and S. Ourselin. "STEPS: Similarity and Truth Estimation for Propagated Segmentations and its application to hippocampal segmentation and brain parcelation". *Medical Image Analysis* 17.6 (2013), pp. 671–684.

[16] J. W. Chan, V. Kearney, S. H. Ms, S. Wu, M. Bogdanov, M. Reddick, N. D. Ms, A. Sudhyadhom, J. Chen, and S. S. Yom. "A convolutional neural network algorithm for automatic segmentation of head and neck organs-at-risk using deep lifelong learning". *Medical Physics* (2019).

[17] H. Chen, X. Qi, L. Yu, and P.-A. Heng. "DCAN: deep contour-aware networks for accurate gland segmentation". *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 2016, pp. 2487–2496.

[18] V. Cheplygina. "Cats or CAT scans: transfer learning from natural or medical image source datasets?" *Current Opinion in Biomedical Engineering* 9 (2018), pp. 21–27. arXiv: 1810.05444.

[19] F. Chollet et al. *Keras*. 2015. URL: https://keras.io.

[20] N. N. Chung, L. L. Ting, W. C. Hsu, L. T. Lui, and P. M. Wang. "Impact of magnetic resonance imaging versus CT on nasopharyngeal carcinoma: Primary tumor target delineation for radiotherapy". *Head and Neck* 26.3 (2004), pp. 241–246.

[21] D. Cireșan, A. Giusti, L. M. Gambardella, and J. Schmidhuber. "Deep neural networks segment neuronal membranes in electron microscopy images". *NIPS*. 2012. ISBN: 9781627480031.

[22] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, L. Tarbox, and F. Prior. "The cancer imaging archive (TCIA): Maintaining and operating a public information repository". *Journal of Digital Imaging* (2013).

[23] M. Clements, N. Schupp, M. Tattersall, A. Brown, and R. Larson. "Monaco treatment planning system tools and optimization processes". *Medical Dosimetry* (2018).

[24] O. Commowick and G. Malandain. "Efficient selection of the most similar image in a database for critical structures segmentation." *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Vol. 10. Pt 2. Springer. 2007, pp. 203–210. ISBN: 978-3-540-75758-0.

[25] M. Conson, L. Cella, R. Pacelli, M. Comerci, R. Liuzzi, M. Salvatore, and M. Quarantelli. "Automated delineation of brain structures in patients undergoing radiotherapy for primary brain tumors: From atlas to dose-volume histograms". *Radiotherapy and Oncology* 112.3 (2014), pp. 326–331.

[26] J.-F. Daisne and A. Blumhofer. "Atlas-based automatic segmentation of head and neck organs at risk and nodal target volumes: a clinical validation." *Radiation and Oncology* 8 (2013), p. 154.

[27] G. Delouya, L. Igidbashian, A. Houle, M. Bélair, L. Boucher, C. Cohade, S. Beaulieu, É. J. Filion, G. Coulombe, M. Hinse, C. Martel, P. Després, and P. F. Nguyen-Tan. "18F-FDG-PET imaging in radiotherapy tumor volume delineation in treatment of head and neck cancer". *Radiotherapy and Oncology* (2011).

[28] J. Deng, W. Dong, R. Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. "ImageNet: A large-scale hierarchical image database". *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2009, pp. 248–255. ISBN: 978-1-4244-3992-8.

[29] L. R. Dice. "Measures of the amount of ecologic association between species". *Ecology* 26.3 (1945), pp. 297–302. arXiv: 1932409 [10.2307].

[30] G. X. Ding, D. M. Duggan, C. W. Coffey, M. Deeley, D. E. Hallahan, A. Cmelak, and A. Malcolm. "A study on adaptive IMRT treatment planning using kV cone-beam CT". *Radiotherapy and Oncology* (2007).

[31] P. Dirix, K. Haustermans, and V. Vandecaveye. "The value of magnetic resonance imaging for radiotherapy planning". *Seminars in Radiation Oncology* 24.3 (2014), pp. 151–159.

[32] F. Doshi-Velez and B. Kim. "Towards a rigorous science of interpretable machine learning". Ml (2017), pp. 1–13. arXiv: 1702.08608.

[33] J. Duchi, E. Hazan, and Y. Singer. "Adaptive subgradient methods for online learning and stochastic optimization". *JMLR* (2011).

[34] V. Dumoulin and F. Visin. "A guide to convolution arithmetic for deep learning" (2016), pp. 1–31. arXiv: 1603.07285.

[35] E R Dougherty. "An introduction to morphological image processing". *SPIE Optical Engineering Press* (1992).

[36] J. M. Edmund and T. Nyholm. "A review of substitute CT generation for MRI-only radiation therapy". *Radiation Oncology* 12.1 (2017), p. 28.

[37] A. R. Eldesoky, G. Francolini, M. S. Thomsen, E. S. Yates, T. B. Nyeng, C. Kirkove, C. Kamby, E. S. Blix, M. H. Nielsen, Z. Taheri-Kadkhoda, M. Berg, and B. V. Offersen. "Dosimetric assessment of an atlas based automated segmentation for loco-regional radiation therapy of early breast cancer in the Skagen Trial 1: A multi-institutional study". *Clinical and Translational Radiation Oncology* 2 (2017), pp. 1–4.

[38] B. Emami, A. Sethi, and G. J. Petruzzelli. "Influence of MRI on target volume delineation and IMRT planning in nasopharyngeal carcinoma". *International Journal of Radiation Oncology Biology Physics* 57.2 (2003), pp. 481–488.

[39] E Faggiano, C Fiorino, E Scalco, S Broggi, M Cattaneo, E Maggiulli, I Dell'Oca, N Di Muzio, R Calandrino, and G Rizzo. "An automatic contour propagation method to follow parotid gland deformation during head-and-neck cancer tomotherapy". *Physics in Medicine and Biology* 56.3 (2011), pp. 775–791.

[40] B. G. Fallone, B. Murray, S. Rathee, T. Stanescu, S. Steciw, S. Vidakovic, E. Blosser, and D. Tymofichuk. "First MR images obtained during megavoltage photon irradiation from a prototype integrated linac-MR system". *Medical Physics* 36.6 (2009), pp. 2084–2088.

[41] E. C. Ford, P. E. Kinahan, L. Hanlon, A. Alessio, J. Rajendran, D. L. Schwartz, and M. Phillips. "Tumor delineation using PET in head and neck cancers: Threshold contouring and lesion volumes". *Medical Physics* (2006).

[42] A. Frederiksson. "Robust optimization of radiation therapy accounting for geometric uncertainty". PhD thesis. KTH Royal Institute of Technology, 2013. ISBN: 9789175017716.

[43] K. D. Fritscher, M. Peroni, P. Zaffino, M. F. Spadea, R. Schubert, and G. Sharp. "Automatic segmentation of head and neck CT images for radiotherapy treatment planning using multiple atlases, statistical appearance models, and geodesic active contours." *Medical Physics* 41.5 (2014), p. 051910.

[44] K. Fukushima. "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position". *Biological Cybernetics* (1980).

[45] Y. Gal and Z. Ghahramani. "Dropout as a bayesian approximation: Representing model uncertainty in deep learning". *33rd International Conference on Machine Learning*. Vol. 48. 2016. arXiv: 1506.02142.

[46] X. Geets, J.-F. Daisne, S. Arcangeli, E. Coche, M. De Poel, T. Duprez, G. Nardella, and V. Grégoire. "Inter-observer variability in the delineation of pharyngo-laryngeal tumor, parotid glands and cervical spinal cord: comparison between CT-scan and MRI." *Radiotherapy and Oncology* 77.1 (2005), pp. 25–31.

[47] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel. "ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness" (2018), pp. 1–22. arXiv: 1811.12231.

[48] X. Glorot and Y. Bengio. "Understanding the difficulty of training deep feedforward neural networks". *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS-10)* (2010).

[49] I. Goodfellow. "NIPS 2016 Tutorial: Generative adversarial networks" (2016). arXiv: 1701.00160.

[50] I. Goodfellow, Y. Bengio, and A. Courville. "Deep Learning - whole book". *Nature* (2016). arXiv: arXiv:1312.6184v5.

[51] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. "Generative adversarial networks" (2014), pp. 1–9. arXiv: 1406.2661.

[52] V. Grégoire, K. Ang, W. Budach, C. Grau, M. Hamoir, J. A. Langendijk, A. Lee, Q. T. Le, P. Maingon, C. Nutting, B. O'Sullivan, S. V. Porceddu, and B. Lengele. "Delineation of the neck node levels for head and neck tumors: A 2013 update. DAHANCA, EORTC, HKNPCSG, NCIC CTG, NCRI, RTOG, TROG consensus guidelines". *Radiotherapy and Oncology* 110.1 (2014), pp. 172–181.

[53] A. J. Grossberg, A. S. R. Mohamed, H. Elhalawani, W. C. Bennett, K. E. Smith, T. S. Nolan, S. Chamchod, M. E. Kantor, T. Browne, K. A. Hutcheson, G. Gunn, Gary BrandonA, A. S. Frank, D. I. Rosenthal, J. B. Freymann, and C. D. Fuller. "Data from head and neck cancer CT atlas. The Cancer Imaging Archive." (2017).

[54] A. L. Grosu, M. Piert, W. A. Weber, B. Jeremic, M. Picchio, U. Schratzenstaller, F. B. Zimmermann, M. Schwaiger, and M. Molls. "Positron emission tomography for radiation treatment planning". *Strahlentherapie und Onkologie* (2005).

[55] O. Gurney-Champion, J. Kieselmann, K. Wong, K. Harrington, and U. Oelfke. "Rapid and accurate automatic contouring of quantitative diffusion-weighted MRI using a deep convolutional neural network". *7th MR in RT symposium*. 2019.

[56] X. Han, M. S. Hoogeman, P. C. Levendag, L. S. Hibbard, D. N. Teguh, P. Voet, A. C. Cowen, and T. K. Wolf. "Atlas-based auto-segmentation of head and neck CT images". *MICCAI 2008* LNCS 5242.Part II (2008), pp. 434–441.

[57] K. He, X. Zhang, S. Ren, and J. Sun. "Deep residual learning for image recognition". *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 2016. ISBN: 9781467388504.

[58] T. Heimann and H. P. Meinzer. "Statistical shape models for 3D medical image segmentation: A review". *Medical Image Analysis* (2009).

[59] D. L. Hill, P. G. Batchelor, M. Holden, and D. J. Hawkes. "Medical image registration". *Physics in Medicine and Biology* (2001).

[60] G. E. Hinton, N. Srivastava, and K. Swersky. *RMSProp*. URL: https://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf.

[61] S. Hissoiny, B. Ozell, H. Bouchard, and P. Després. "GPUMCD: A new GPU-oriented Monte Carlo dose calculation platform." *Medical Physics* 38.2 (2011), pp. 754–764. arXiv: 1101.1245.

[62] J. K. Hoang, C. M. Glastonbury, L. F. Chen, J. K. Salvatore, and J. D. Eastwood. "CT mucosal window settings: A novel approach to evaluating early T-stage head and neck carcinoma". *American Journal of Roentgenology* (2010).

[63] A. K. Hoang Duc, G. Eminowicz, R. Mendes, S.-L. Wong, J. McClelland, M. Modat, M. J. Cardoso, A. F. Mendelson, C. Veiga, T. Kadir, D. D'Souza, and S. Ourselin. "Validation of clinical acceptability of an atlas-based segmentation algorithm for the delineation of organs at risk in head and neck cancer". *Medical Physics* 42.9 (2015), pp. 5027–5034.

[64] G. N. Hounsfield. "Computerized transverse axial scanning (tomography): I. Description of system". *British Journal of Radiology* (1973).

[65] M. A. Hunt, A. Jackson, A. Narayana, and N. Lee. "Geometric factors influencing dosimetric sparing of the parotid glands using IMRT". *International Journal of Radiation Oncology Biology Physics* 66.1 (2006), pp. 296–304.

[66] B. Ibragimov and L. Xing. "Segmentation of organs-at-risks in head and neck CT images using convolutional neural networks:" *Medical Physics* (2017).

[67] ICRU. "Prescribing, recording and reporting photon-beam intensity modulated radiation therapy (IMRT) (ICRU Report 83)". *Journal of the ICRU* (2010).

[68] J. E. Iglesias and M. R. Sabuncu. "Multi-atlas segmentation of biomedical images: A survey". *Medical Image Analysis* 24.1 (2015), pp. 205–219.

[69] P. Isola, J. Y. Zhu, T. Zhou, and A. A. Efros. "Image-to-image translation with conditional adversarial networks". *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*. 2017. ISBN: 9781538604571.

[70] D. A. Jaffray, D. G. Drake, M. Moreau, A. A. Martinez, and J. W. Wong. "A radiographic and tomographic imaging system integrated into a medical linear accelerator for localization of bone and soft-tissue targets". *International Journal of Radiation Oncology Biology Physics* (1999).

[71] A. P. Jellema, B. J. Slotman, P. Doornaert, C. R. Leemans, and J. A. Langendijk. "Impact of radiation-induced xerostomia on quality of life after primary radiotherapy among patients with head and neck cancer". *International Journal of Radiation Oncology Biology Physics* (2007).

[72] J. P. Kieselmann, C. P. Kamerling, N. Burgos, M. J. Menten, C. D. Fuller, S. Nill, M. J. Cardoso, and U. Oelfke. "Geometric and dosimetric evaluations of atlas-based segmentation methods of MR images in the head and neck region". *Physics in Medicine and Biology* (2018).

[73] D. P. Kingma and J. L. Ba. "Adam: A method for stochastic gradient descent". *ICLR: International Conference on Learning Representations* (2015). arXiv: arXiv:1412.6980v9.

[74] S. Klein, M. Staring, K. Murphy, M. A. Viergever, and J. P. Pluim. "Elastix: A toolbox for intensity-based medical image registration". *IEEE Transactions on Medical Imaging* (2010).

[75] S. Klein, M. Staring, and J. P. Pluim. "Evaluation of optimization methods for nonrigid medical image registration using mutual information and B-splines". *IEEE Transactions on Image Processing* (2007).

[76] M Köhler, T Vaara, M. V. Grootel, R Hoogeveen, R Kemppainen, and S Renisch. "MR-only simulation for radiotherapy planning treatment planning". *White paper: Philips MRCAT for prostate dose calculations using only MRI data* (2015), pp. 1–16.

[77] E. Kouwenhoven, M. Giezen, and H. Struikmans. "Measuring the similarity of target volume delineations independent of the number of observers". *Physics in Medicine and Biology* (2009).

[78] A. Krizhevsky, I. Sutskever, and G. E. Hinton. *ImageNet classification with deep convolutional neural networks*. 2012.

[79] M. La Macchia, F. Fellin, M. Amichetti, M. Cianchetti, S. Gianolini, V. Paola, A. J. Lomax, and L. Widesott. "Systematic evaluation of three different commercial software solutions for automatic segmentation for adaptive therapy in head-and-neck, prostate and pleural cancer". *Radiation Oncology* 7.1 (2012), p. 160.

[80] J. J. Lagendijk, B. W. Raaymakers, C. A. Van Den Berg, M. A. Moerland, M. E. Philippens, and M. Van Vulpen. "MR guidance in radiotherapy". *Physics in Medicine and Biology* 59.21 (2014), R349–R369.

[81] K. Langen, S. Meeks, D. Poole, T. Wagner, T. Willoughby, P. Kupelian, K. Ruchala, J. Haimerl, and G. Olivera. "SU-FF-T-324: The use of megavoltage CT (MVCT) Images for dose recomputation". *Medical Physics* (2005).

[82] J. A. Langendijk, P. Doornaert, D. H. Rietveld, I. M. Verdonck-de Leeuw, C. René Leemans, and B. J. Slotman. "A predictive model for swallowing dysfunction after curative radiotherapy in head and neck cancer". *Radiotherapy and Oncology* (2009).

[83] T. Langerak, U. van der Heide, A. Kotte, F. Berendsen, and J. Pluim. "Improving label fusion in multi-atlas based segmentation by locally combining atlas selection and performance estimation". *Computer Vision and Image Understanding* 130 (2015), pp. 71–79.

[84] R. J. Lapeer, A. C. Tan, and R. Aldridge. "Active watersheds: Combining 3D watershed segmentation and active contours to extract abdominal organs from MR images". *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2002, pp. 596–603.

[85] Y. LeCun, Y. Bengio, and G. Hinton. "Deep learning". *Nature Methods* 13.1 (2015), p. 35.

[86] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. "Gradient-based learning applied to document recognition". *Proceedings of the IEEE* (1998).

[87] C. Lee, K. Langen, P. Kupelian, S. Meeks, R. Mañon, W. Lu, J. Haimerl, and G. Olivera. "Positional and volumetric changes in parotid glands during head and neck radiation therapy assessed using deformable image registration". *International Journal of Radiation Oncology Biology Physics* 69.3 (2007), S203–S204.

[88] H. Lester and S. R. Arridge. "A survey of hierarchical non-linear medical image registration". *Pattern Recognition* (1999).

[89] H Levene. "Robust tests for equality of variances". *Contributions to probability and statistics: Essays in ...* (1960).

[90] S.-J. Lim, Y.-Y. Jeong, C.-W. Lee, and Y.-S. Ho. "Automatic segmentation of the liver in CT images using the watershed algorithm based on morphological filtering". *Medical Imaging 2004: Image Processing*. 2004.

[91] G. P. Liney, B. Dong, J. Begg, P. Vial, K. Zhang, F. Lee, A. Walker, R. Rai, T. Causer, S. J. Alnaghy, B. M. Oborn, L. Holloway, P. Metcalfe, M. Barton, S. Crozier, and P. Keall. "Technical Note: Experimental results from a prototype high-field inline MRI-linac". *Medical Physics* 43.9 (2016), pp. 5188–5194.

[92] 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". *Medical Image Analysis* (2017). arXiv: 1702.05747.

[93] J. B. Maintz and M. A. Viergever. "A survey of medical image registration." *Medical Image Analysis* 2.1 (1998), pp. 1–36.

[94] M. Mazonakis, J. Damilakis, H. Varveris, P. Prassopoulos, and N. Gourtsoyiannis. "Image segmentation in treatment planning for prostate cancer using the region growing technique". *British Journal of Radiology* (2001).

[95] P. Metcalfe, G. P. Liney, L. Holloway, A. Walker, M. Barton, G. P. Delaney, S. Vinod, and W. Tomé. "The potential for an enhanced role for MRI in radiation-therapy treatment planning". *Technology in Cancer Research & Treatment* 12.5 (2013), pp. 429–446.

[96] F. Milletari, N. Navab, and S. A. Ahmadi. "V-Net: Fully convolutional neural networks for volumetric medical image segmentation". *Proceedings - 2016 4th International Conference on 3D Vision, 3DV 2016*. 2016. ISBN: 9781509054077.

[97] D. Močnik, B. Ibragimov, L. Xing, P. Strojan, B. Likar, F. Pernuš, and T. Vrtovec. "Segmentation of parotid glands from registered CT and MR images". *Physica Medica* 52 (2018), pp. 33–41.

[98] M. Modat, G. R. Ridgway, Z. A. Taylor, M. Lehmann, J. Barnes, D. J. Hawkes, N. C. Fox, and S. Ourselin. "Fast free-form deformation using graphics processing units". *Computer Methods and Programs in Biomedicine* 98.3 (2010), pp. 278–284.

[99] M. Modat, D. M. Cash, P. Daga, G. P. Winston, J. S. Duncan, and S. Ourselin. "Global image registration using a symmetric block-matching approach." *Journal of Medical Imaging* 1.2 (2014), p. 024003.

[100] S. Mutic and J. F. Dempsey. "The ViewRay system: magnetic resonance–guided and controlled radiotherapy". *Seminars in Radiation Oncology* 24.3 (2014), pp. 196–199.

[101] B. E. Nelms, W. a. Tomé, G. Robinson, and J. Wheeler. "Variations in the contouring of organs at risk: Test case from a patient with oropharyngeal cancer". *International Journal of Radiation Oncology Biology Physics* 82.1 (2012), pp. 368–378.

[102] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. "Multimodal deep learning". *28th international conference on machine learning (ICML-11)*. 2011, pp. 689–696.

[103] N. P. Nguyen, C. Frank, C. C. Moltz, P. Vos, H. J. Smith, U. Karlsson, S. Dutta, A. Midyett, J. Barloon, and S. Sallah. "Impact of dysphagia on quality of life after treatment of head-and-neck cancer". *International Journal of Radiation Oncology Biology Physics* (2005).

[104] D. Nie, L. Wang, Y. Gao, and D. Sken. "Fully convolutional networks for multi-modality isointense infant brain image segmentation". *Proceedings - International Symposium on Biomedical Imaging*. 2016. ISBN: 9781479923502.

[105] C. Nutting, D. P. Dearnaley, and S. Webb. "Intensity modulated radiation therapy: A clinical review". *British Journal of Radiology* (2000).

[106] C. M. Nutting, J. P. Morden, K. J. Harrington, T. G. Urbano, S. A. Bhide, C. Clark, E. A. Miles, A. B. Miah, K. Newbold, M. A. Tanay, F. Adab, S. J. Jefferies, C. Scrase, B. K. Yap, R. P. A'Hern, M. A. Sydenham, M. Emson, and E. Hall. "Parotid-sparing intensity modulated versus conventional radiotherapy in head and neck cancer (PARSPORT): A phase 3 multicentre randomised controlled trial". *The Lancet Oncology* (2011).

[107] T. Nyholm and J. Jonsson. "Counterpoint: Opportunities and challenges of a magnetic resonance imaging-only radiotherapy workflow". *Seminars in Radiation Oncology* 24.3 (2014), pp. 175–180.

[108] N Otsu. "A threshold selection method from gray level histograms". *IEEE Trans. Systems, Man and Cybernetics* 9 (1979), pp. 62–66.

[109] K. Otto. "Volumetric modulated arc therapy: IMRT in a single gantry arc." *Medical Physics* 35.1 (2008), pp. 310–317.

[110] S. Ourselin, A. Roche, S. Prima, and N. Ayache. "Block matching: A general framework to improve robustness of rigid registration of medical images". 2011.

[111] S. Ourselin, A. Roche, G. Subsol, X. Pennec, and N. Ayache. "Reconstructing a 3D structure from serial histological sections". *Image and Vision Computing* 19.1-2 (2001), pp. 25–31.

[112] E. Park, X. Han, T. L. Berg, and A. C. Berg. "Combining multiple sources of knowledge in deep CNNs for action recognition". *2016 IEEE Winter Conference on Applications of Computer Vision, WACV 2016*. 2016. ISBN: 9781509006410.

[113] A. Paszke, G. Chanan, Z. Lin, S. Gross, E. Yang, A. D. L. S. Antiga, S. Chintala, Z. DeVito, and A. Lerer. "Automatic differentiation in PyTorch". *NIPS*. 2017.

[114] A. C. Paulino, M. Koshy, R. Howell, D. Schuster, and L. W. Davis. "Comparison of CT- and FDG-PET-defined gross tumor volume in intensity-modulated radiotherapy for head-and-neck cancer". *International Journal of Radiation Oncology Biology Physics* (2005).

[115] V. Pekar, S Allaire, and A. Qazi. "Head and neck auto-segmentation challenge: segmentation of the parotid glands". *MICCAI 2010: A Grand Challenge for the Clinic* August (2010), pp. 273–280.

[116] L. A. Perez-Romasanta, M. Bellon-Guardia, J. Torres-Donaire, E. Lozano-Martin, M. Sanz-Martin, and J. Velasco-Jimenez. "Tumor volume delineation in head and neck cancer with 18-fluor- fluorodeoxiglucose positron emission tomography: Adaptive thresholding method applied to primary tumors and metastatic lymph nodes". *Clinical and Translational Oncology* (2013).

[117] D. L. Pham, C. Xu, and J. L. Prince. "Current methods in medical image segmentation". *Annu. Rev. Biomed. Eng.* 02 (2000), pp. 315–337.

[118] E. Prieto, P. Lecumberri, M. Pagola, M. Gómez, I. Bilbao, M. Ecay, I. Peñuelas, and J. M. Martí-Climent. "Twelve automated thresholding methods for segmentation of PET images: A phantom study". *Physics in Medicine and Biology* (2012).

[119] A. a. Qazi, V. Pekar, J. Kim, J. Xie, S. L. Breen, and D. a. Jaffray. "Auto-segmentation of normal and target structures in head and neck CT images: A feature-driven model-based approach". *Medical Physics* 38.11 (2011), p. 6160.

[120] B. W. Raaymakers, J. J. W. Lagendijk, J Overweg, J. G. M. Kok, A. J. E. Raaijmakers, E. M. Kerkhof, R. W. van der Put, I Meijsing, S. P. M. Crijns, F Benedosso, M van Vulpen, C. H. W. de Graaff, J Allen, and K. J. Brown. "Integrating a 1.5 T MRI scanner with a 6 MV accelerator: proof of concept". *Physics in Medicine and Biology* 54.12 (2009), N229–N237.

[121] C. R. Rasch, R. Steenbakkers, I. Fitton, J. C. Duppen, P. J. Nowak, F. A. Pameijer, A. Eisbruch, J. H. Kaanders, F. Paulsen, and M. van Herk. "Decreased 3D observer variation with matched CT-MRI, for target delineation in Nasopharynx cancer". *Radiation Oncology* 5.1 (2010), p. 21.

[122] C. R. Rasch, R. Steenbakkers, and M. van Herk. "Target definition in prostate, head, and neck." *Seminars in radiation oncology* (2005).

[123] P. F. Raudaschl, P. Zaffino, G. C. Sharp, M. F. Spadea, A. Chen, B. M. Dawant, T. Albrecht, T. Gass, C. Langguth, M. Luthi, F. Jung, O. Knapp, S. Wesarg, R. Mannion-Haworth, M. Bowes, A. Ashman, G. Guillard, A. Brett, G. Vincent, M. Orbes-Arteaga, D. Cardenas-Pena, G. Castellanos-Dominguez, N. Aghdasi, Y. Li, A. Berens, K. Moe, B. Hannaford, R. Schubert, and K. D. Fritscher. "Evaluation of segmentation methods on head and neck CT: Auto-segmentation challenge 2015". *Medical Physics* 44.5 (2017), pp. 2020–2036.

[124] J. L. Robar, A. Day, J. Clancey, R. Kelly, M. Yewondwossen, H. Hollenhorst, M. Rajaraman, and D. Wilke. "Spatial and dosimetric variability of organs at risk in head-and-neck intensity-modulated radiotherapy". *International Journal of Radiation Oncology Biology Physics* 68.4 (2007), pp. 1121–1130.

[125] T. Rohlfing, R. Brandt, R. Menzel, and C. R. Maurer. "Evaluation of atlas selection strategies for atlas-based image segmentation with application to confocal microscopy images of bee brains". *NeuroImage* 21.4 (2004), pp. 1428–1442.

[126] O. Ronneberger, P. Fischer, and T. Brox. "U-net: Convolutional networks for biomedical image segmentation". *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 9351 (2015), pp. 234–241. arXiv: 1505.04597.

[127] W. van Rooij, M. Dahele, H. R. Brandao, A. R. Delaney, B. J. Slotman, and W. F. Verbakel. "Deep learning-based delineation of head and neck organs-at-risk: geometric and dosimetric evaluation". *International Journal of Radiation Oncology Biology Physics* (2019).

[128] S. Ruder. "An overview of gradient descent optimization algorithms". *ArXiv preprint* (2017). arXiv: arXiv:1609.04747.

[129] D. Rueckert. "Nonrigid registration using free-form deformations: Application to breast mr images". *IEEE Transactions on Medical Imaging* 18.8 (1999), pp. 712–721.

[130] M. A. Schmidt and G. S. Payne. "Radiotherapy planning using MRI". *Physics in Medicine and Biology* 60.22 (2015), R323–R361.

[131] N. Sharma, A. Ray, K. Shukla, S. Sharma, S. Pradhan, A. Srivastva, and L. Aggarwal. "Automated medical image segmentation techniques". *Journal of Medical Physics* 35.1 (2010), p. 3.

[132] G. Sharp, K. D. Fritscher, V. Pekar, M. Peroni, N. Shusharina, H. Veeraraghavan, and J. Yang. "Vision 20/20: perspectives on automated image segmentation for radiotherapy." *Medical Physics* 41.5 (2014), p. 050902.

[133] K. Shridhar, F. Laumann, and M. Liwicki. "A comprehensive guide to bayesian convolutional neural networks with variational inference" (2019), pp. 1–38. arXiv: 1901.02731.

[134] R. Sims, A. Isambert, V. Grégoire, F. Bidault, L. Fresco, J. Sage, J. Mills, J. Bourhis, D. Lefkopoulos, O. Commowick, M. Benkebil, and G. Malandain. "A pre-clinical assessment of an atlas-based automatic segmentation tool for the head and neck". *Radiotherapy and Oncology* 93.3 (2009), pp. 474–478.

[135] G. Song, J. Han, Y. Zhao, Z. Wang, and H. Du. "A review on medical image registration as an optimization problem". *Current Medical Imaging Reviews* (2017).

[136] C. Spearman. "Spearman's rank correlation coefficient". *Amer. J. Psychol.* 15 (1904), pp. 72–101.

[137] R. Steenbakkers, J. C. Duppen, I. Fitton, K. E. Deurloo, L. J. Zijp, E. F. Comans, A. L. Uitterhoeve, P. T. Rodrigus, G. W. Kramer, J. Bussink, K. De Jaeger, J. S. Belderbos, P. J. Nowak, M. Van Herk, and C. R. Rasch. "Reduction of observer variation using matched CT-PET for lung cancer delineation: A three-dimensional analysis". *International Journal of Radiation Oncology Biology Physics* (2006).

[138] B. W. Stewart and C. P. Wild. "World cancer report 2014". *International Agency for Research on Cancer, World Health Organization* (2014), pp. 422–425. arXiv: 1011.1669.

[139] Student. "The probable error of a mean". *Biometrika* 6.1 (1908), pp. 1–25.

[140] Y. Sun, X.-L. Yu, W. Luo, A. W. Lee, J. T. S. Weec, N. Lee, G.-Q. Zhou, L.-L. Tang, C.-J. Tao, R. Guo, Y.-P. Mao, R. Zhang, Y. Guo, and J. Maa. "Recommendation for a contouring method and atlas of organs at risk in nasopharyngeal carcinoma patients receiving intensity-modulated radiotherapy". *Radiotherapy and Oncology* 110.3 (2014), pp. 390–397.

[141] A. A. Taha and A. Hanbury. "Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool". *BMC Medical Imaging* (2015).

[142] Y. Tan, L. H. Schwartz, and B. Zhao. "Segmentation of lung lesions on CT scans using watershed, active contours, and Markov random field". *Medical Physics* (2013).

[143] D. N. Teguh, P. C. Levendag, P. W. J. Voet, A. Al-Mamgani, X. Han, T. K. Wolf, L. S. Hibbard, P. Nowak, H. Akhiat, M. L. P. Dirkx, B. J. M. Heijmen, and M. S. Hoogeman. "Clinical validation of atlas-based auto-segmentation of multiple target volumes and normal tissue (swallowing/mastication) structures in the head and neck." *International Journal of Radiation Oncology Biology Physics* 81.4 (2011), pp. 950–7.

[144] J. P. Thirion. "Image matching as a diffusion process: An analogy with Maxwell's demons". *Medical Image Analysis* (1998).

[145] N. Tong, S. Gou, S. Yang, D. Ruan, and K. Sheng. "Fully automatic multi-organ segmentation for head and neck cancer radiotherapy using shape representation model constrained fully convolutional neural networks". *Medical Physics* 45.10 (2018), pp. 4558–4567.

[146] S. Y. Tsuji, A. Hwang, V. Weinberg, S. S. Yom, J. M. Quivey, and P. Xia. "Dosimetric evaluation of automatic segmentation for adaptive IMRT for head-and-neck cancer". *International Journal of Radiation Oncology Biology Physics* 77.3 (2010), pp. 707–714.

[147] K. Van Leemput, F. Maes, D. Vandermeulen, and P. Suetens. "Automated model-based bias field correction of MR images of the brain." *IEEE Transactions on Medical Imaging* 18.10 (1999), pp. 885–896.

[148] A. Van Opbroek, M. A. Ikram, M. W. Vernooij, and M. De Bruijne. "Transfer learning improves supervised image segmentation across imaging protocols". *IEEE Transactions on Medical Imaging* (2015).

[149] H Veeraraghavan, N Tyagi, M Hunt, N Lee, and J Deasy. "SU-F-303-16: Multi-atlas and learning based segmentation of head and neck normal structures from multi-parametric MRI". *Medical Physics* 42.6 (2015), p. 3541.

[150] C. Veiga, J. McClelland, S. Moinuddin, A. Lourenço, K. Ricketts, J. Annkah, M. Modat, S. Ourselin, D. D'Souza, and G. Royle. "Toward adaptive radiotherapy for head and neck patients: Feasibility study on using CT-to-CBCT deformable registration for "dose of the day" calculations". *Medical Physics* (2014).

[151] S. K. Vinod, M. G. Jameson, M. Min, and L. C. Holloway. "Uncertainties in volume delineation in radiation oncology: A systematic review and recommendations for future studies". *Radiotherapy and Oncology* 121.2 (2016), pp. 169–179.

[152] P. W. J. Voet, M. L. P. Dirkx, D. N. Teguh, M. S. Hoogeman, P. C. Levendag, and B. J. M. Heijmen. "Does atlas-based autosegmentation of neck levels require subsequent manual contour editing to avoid risk of severe target underdosage? A dosimetric analysis". *Radiotherapy and Oncology* 98.3 (2011), pp. 373–377.

[153] H. A. Vrooman, C. A. Cocosco, F. van der Lijn, R. Stokking, M. A. Ikram, M. W. Vernooij, M. M. Breteler, and W. J. Niessen. "Multi-spectral brain tissue segmentation using automatically trained k-Nearest-Neighbor classification". *NeuroImage* (2007).

[154] K. Wardman, R. J. D. Prestwich, M. J. Gooding, and R. J. Speight. "The feasibility of atlas-based automatic segmentation of MRI for H \& N radiotherapy planning". *Journal of Applied Clinical Medical Physics* 17.4 (2016), pp. 146–154.

[155] S. K. Warfield, K. H. Zou, and W. M. Wells. "Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation". *IEEE Transactions on Medical Imaging* 23.7 (2004), pp. 903–921.

[156] T. A. van de Water, H. P. Bijl, H. E. Westerlaan, and J. A. Langendijk. "Delineation guidelines for organs at risk involved in radiation-induced salivary dysfunction and xerostomia". *Radiotherapy and Oncology* (2009).

[157] O. Weistrand and S. Svensson. "The ANACONDA algorithm for deformable image registration in radiotherapy." *Medical Physics* 42.1 (2015), pp. 40–53.

[158] L. Welsh, R. Panek, D. McQuaid, A. Dunlop, M. Schmidt, A. Riddell, D.-M. Koh, S. Doran, I. Murray, Y. Du, S. Chua, V. Hansen, K. H. Wong, J. Dean, S. Gulliford, S. Bhide, M. O. Leach, C. Nutting, K. Harrington, and K. Newbold. "Prospective, longitudinal, multi-modal functional imaging for radical chemo-IMRT treatment of locally advanced head and neck cancer: the INSIGHT study". *Radiation Oncology* 10.1 (2015), p. 112.

[159] J. M. Wolterink, A. M. Dinkla, M. H. Savenije, P. R. Seevinck, C. A. van den Berg, and I. Išgum. "Deep MR to CT synthesis using unpaired data". *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 10557 LNCS. 2017, pp. 14–23. ISBN: 9783319681269. arXiv: 1708.01155.

[160] K. H. Wong, R. Panek, S. A. Bhide, C. M. Nutting, K. J. Harrington, and K. L. Newbold. "The emerging potential of magnetic resonance imaging in personalizing radiotherapy for head and neck cancer: An oncologist's perspective". *British Journal of Radiology* (2017).

[161] M. Wu, C. Rosano, P. Lopez-Garcia, C. S. Carter, and H. J. Aizenstein. "Optimum template selection for atlas-based segmentation". *NeuroImage* 34.4 (2007), pp. 1612–1618.

[162] X. Yang, N. Wu, G. Cheng, Z. Zhou, D. S. Yu, J. J. Beitler, W. J. Curran, and T. Liu. "Automated segmentation of the parotid gland based on atlas registration and machine learning: A longitudinal mri study in head-and-neck radiation therapy". *International Journal of Radiation Oncology Biology Physics* 90.5 (2014), pp. 1225–1233. arXiv: 15334406.

[163] Y. Yang, E. C. Ford, B. Wu, M. Pinkawa, B. Van Triest, P. Campbell, D. Y. Song, and T. R. McNutt. "An overlap-volume-histogram based method for rectal dose prediction and automated treatment planning in the external beam prostate radiotherapy following hydrogel injection". *Medical Physics* (2013).

[164] Y. Yang, E. Schreibmann, T. Li, C. Wang, and L. Xing. "Evaluation of on-board kV cone beam CT (CBCT)-based dose calculation". *Physics in Medicine and Biology* (2007).

[165] X. Yi, E. Walia, and P. Babyn. "Generative adversarial network in medica imaging: a review" (2018). arXiv: 1809.07294.

[166] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. "How transferable are features in deep neural networks?" (2014), pp. 1–9. arXiv: 1411.1792.

[167] M. D. Zeiler. "ADADELTA: An adaptive learning rate method" (2012). arXiv: 1212.5701.

[168] M. D. Zeiler and R. Fergus. "Visualizing and understanding convolutional networks". *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 8689 LNCS.PART 1 (2014), pp. 818–833. arXiv: arXiv:1311.2901v3.

[169] J. Y. Zhu, T. Park, P. Isola, and A. A. Efros. "Unpaired image-to-image translation using cycle-consistent adversarial networks". *Proceedings of the IEEE International Conference on Computer Vision*. Vol. 2017-Octob. 2017, pp. 2242–2251. ISBN: 9781538610329. arXiv: 1703.10593.

[170] M. Zhuang, R. A. Dierckx, and H. Zaidi. "Generic and robust method for automatic segmentation of PET images using an active contour model". *Medical Physics* (2016).