TU Dortmund University Department of Physics

Dissertation for the attainment of the academic degree
Doctor rerum naturalium

Enhancing precision radiotherapy: Image registration with deep learning and image fusion for treatment planning

Alexander Ratke

June 2023

This thesis is submitted by Alexander Ratke, born in Ljubinskoe (Russia) on 16 June 1995, to the Department of Physics of TU Dortmund University on 16 June 2023.

Assessors:

Prof. Dr. Kevin Kröninger
PD Dr. Christian Bäumer

Chairperson of the examination committee:

Prof. Dr. Frithjof Anders

Representative of the scientific staff:

Dr. Jörg Debus

Date of the oral exam: 29 August 2023

This work is dedicated to
Prof. Dr. Bernhard Spaan,
who sadly deceased too early
on 9 December 2021.

Abstract

Artificial intelligence is advancing in everyday life and supports its user by generating fast results in areas like communication or image recognition. This thesis aims at exploiting the abilities of deep-learning techniques for deformable image registration (DIR) to improve image alignment in medicine. An unsupervised registration and fusion workflow is developed and evaluated for 39 head scans, produced with computed tomography (CT) and magnetic resonance imaging (MRI). The three-part workflow starts by preprocessing the scans to unify the image formats and to perform affine transformation and rigid registration. Then, a deep-learning model trained for DIR is applied to these images. To obtain an appropriate configuration of the model, parameter tuning is required. The evaluation with the mutual-information metric indicates an improvement in image alignment of up to 14 % when using deep-learning-based DIR. Lastly, image fusion combines the registered CT and MRI scans with a wavelet-based method to merge the information of decomposed images. The workflow is designed for unimodal, e.g. T_1- and T_2-weighted MRI scans, and multimodal, e.g. CT and MRI scans, image pairs. Since medical imaging is an important basis of treatment-planning processes, the registered and fused images obtained from this workflow are expected to enhance precision radiotherapy.

Kurzfassung

Künstliche Intelligenz wird im Alltag immer präsenter und unterstützt den Anwender durch schnelle Ergebnisse in Bereichen wie Kommunikation oder Bilderkennung. Ziel dieser Arbeit ist es, die Fähigkeiten von Deep-Learning-Techniken für deformierbare Bildregistrierung zu nutzen, um die Übereinstimmung von medizinischen Bildern zu verbessern. Ein automatisierter Registrierungs- und Fusionsworkflow wird für 39 CT- und MRT-Aufnahmen des Kopfes entwickelt und evaluiert. Der dreiteilige Workflow beginnt mit einer Vorverarbeitung der Aufnahmen, um die Bildformate zu vereinheitlichen und eine affine Transformation sowie rigide Registrierung durchzuführen. Dann wird ein für DIR trainiertes Deep-Learning-Modell angewendet. Um eine geeignete Konfiguration des Modells zu erhalten, ist eine Parameteruntersuchung durch Variationen erforderlich. Die Auswertung mit der Transinformationsmetrik zeigt eine Verbesserung der Bildübereinstimmung um bis zu 14 % bei Verwendung von Deep-Learning-basierter DIR. Schließlich werden bei der Bildfusion die registrierten CT- und MRT-Aufnahmen mit einer Wavelet-basierten Methode kombiniert, um die Informationen zerlegter Bilder zusammenzuführen. Der Workflow ist für unimodale, z. B. T_1- und T_2-gewichtete MRT-Aufnahmen, und multimodale, z. B. CT- und MRT-Aufnahmen, Bildpaare konzipiert. Da die medizinische Bildgebung eine wichtige Grundlage der Behandlungsplanung darstellt, ist durch diesen Workflow eine Verbesserung der Präzision in der Strahlentherapie zu erwarten.

Contents

1 Introduction

In medical imaging, information on the anatomy of the human body is provided by various modalities, e.g. computed tomography (CT), magnetic resonance imaging (MRI) or positron emission tomography (PET). The significant differences between these modalities are caused by their physical concepts, which lead to different types of contrast for the same human tissue. For example, MRI provides images with high soft-tissue distinctness, while CT has better bone contrast and geometric fidelity. [1] The characteristics of CT are the basis for treatment planning in radiotherapy due to the availability of mass attenuation coefficients for dose calculation. In combination with superior soft-tissue contrast in MRI scans, especially between the tumour and the healthy tissue, the quality of treatment plans can benefit from a more precise delineation of target volumes [1–3].

The images of different modalities are typically superimposed with rigid registrations to outline regions in treatment-planning systems [4]. However, this process does not consider the displacement of organs due to the immobilisation of the patient or the distortion of MRI scans, which is supposed to originate from the magnetic fields [1, 4]. In the case of pediatric patients, further challenges, which impact the quality of treatment planning, arise through patient immobilisation, physical growth and a prolonged course of disease [5]. Therefore, deformable image registration (DIR), characterised by individual displacements of each image pixel, is expected to improve multimodal treatment-planning processes with more precise image alignment [3, 4].

In recent years, the interest in deep-learning-based medical image registration has rapidly increased [3]. In detailed reviews [6, 7] about the developments and applications of deep learning in medical image analysis, the outcome of most publications related to image registration is summarised. Methods for image registration depend on the modalities of the input images, their dimensions, and the region of interest. Moreover, the broad field of deep learning involves multiple approaches to the architecture of neural networks. [6] While generative adversarial networks (GANs) [8] are often used to generate artificial data, which aim at approximating the target data, convolutional neural networks (CNNs) are useful for image deformation or recognition by extracting image features with the help of convolution operations [3, 6]. The type of training process is an important choice. The reviews reveal that 31 % and 21 % of the publications include unsupervised and supervised methods, respectively. Furthermore, 70 % of the publications focus on unimodal image registration. Unsupervised multimodal methods mainly deal with the images of the abdominal region [9–12] or the generation of synthetic images with GANs [13–15], avoiding the challenge of defining accurate image-similarity measures for the multimodal case. [7] Consequently, there is a lack of direct application of unsupervised multimodal methods for deep-learning-based DIR. [6]

A popular deep-neural-network structure for CNNs is U-Net [16], which efficiently learns image features on a small-sized data set [6]. While the advantage of this method was exploited for fast atlas-based DIR of brain MRI scans [17], the investigations in this thesis employ a U-Net-shaped CNN for DIR of head CT and MRI scans. For the input images of a deep neural network, preprocessing is required, which normalises the images, for example, with affine transformation and rigid registration [4]. This ensures that the deep neural network does not have to deal with linear transformations, but can focus on diffeomorphic deformations [4]. This thesis contributes to that research field by developing a workflow, which includes the preprocessing, the deformable registration and the fusion. The outcome has the potential to facilitate treatment planning with fast and direct application.

This thesis includes a detailed description of the information needed to obtain a comprehensive overview of the research. The basics of relevant medical imaging modalities are given in Chapter 2. The CT and MRI scanners and their scanning sequences, which impact image reconstruction, are described besides the physical foundation. Chapter 3 includes a collection of image-processing techniques and image-matching metrics used in the upcoming investigations. The general concept of digital images is also explained. The information on the available data sets is contained in Chapter 4 besides the description of image segmentation and adjustments for the preprocessing. The deep neural network is introduced in Chapter 5. Moreover, various studies are performed to find optimal parameter settings of the network. The registered images are then used for image fusion, presented in Chapter 6. The methodology to combine the information from multiple images and results for multimodal and unimodal fusion are described. Lastly, an outlook on the clinical integration of the image-registration and image-fusion workflow is given in Chapter 7. The conclusion in Chapter 8 summarises the main results.

2 Medical imaging

The ability to use non-invasive techniques for the visualisation of the human anatomy originated over 100 years ago with X-rays. Since then, further imaging technologies have been developed and improved for medical healthcare, resulting from the increase in computing power of modern computers. Cross-sectional images are essential in clinical settings for medical diagnosis, treatment planning in radiotherapy or surgery. Computed tomography and magnetic resonance imaging are two common imaging techniques, which offer unique advantages. [18] The CT procedure is described in Section 2.1, while general information on the MRI technique is summarised in Section 2.2.

2.1 Computed tomography

CT scans, providing high spatial resolution, are advantageous for the identification of subtle abnormalities. The images are well suited for the visualisation of structures with different densities, making them indispensable for the evaluation, for example, of bone fractures or tumours. In addition, CT scans are useful for detecting foreign bodies, like projectiles, in the case of traumatic injuries. However, these objects and other factors can produce image artefacts, which decreases the quality of the CT scans. Artefacts can occur due to metallic objects, patient movement or hardware issues. Therefore, research is ongoing to minimise the impact of artefacts. Furthermore, exposure to ionising radiation requires the observance of the radiation dose for patient safety. The balance between radiation exposure and image quality is important to avoid unnecessary radiation and to obtain sufficient quality for diagnostics, treatment planning or monitoring. [18]

Based on conventional radiography, the CT procedure involves X-rays to create detailed cross-sectional images of the body, forming a three-dimensional scan. The X-rays traverse the patient and attenuate depending on the tissue type. The attenuated intensities are recorded with detectors, positioned opposite to the X-ray source. For the scan acquisition, projections of the X-rays from different angles are required to generate the two-dimensional images with special computer algorithms. [19]

The development and optimisation of the CT procedure produced several generations of CT scanners. Each generation was built with the aim of increasing the image quality. [20] The general construction of CT scanners is described in Section 2.1.1. Methods used for image reconstruction are presented in Section 2.1.2, while CT-specific properties of image representation are captured in Section 2.1.3.

2.1.1 CT scanner

The basis of the CT scanner is an annular gantry in which the X-ray tube and the detector module are located. For patient positioning, a patient table is connected to the gantry, moving the patient in height, y direction, and into the gantry, z direction. [18] Different generations of the gantry design were developed with the aim of reducing the acquisition time. The first versions included a fixed configuration of an X-ray tube and an opposite detector module, which rotated around the patient, placed in the centre of the gantry, to perform a two-dimensional scan in the axial plane, x–y plane. The movement of the patient through the gantry enabled the incremental acquisition of three-dimensional scans. The limitation of this step-and-shoot sequence was caused by the cable connection between the measuring components in the gantry and the energy sources, which allowed a maximal rotation of 360°. The introduction of the slip-ring technology led to contactless energy transfer, providing continuous rotation of the inner components. [20] The spiral sequence is performed with a constant rotation of the measuring components while constantly moving the patient through the gantry. This technique decreased the acquisition time and is commonly used in clinical facilities. [21]

An essential part of the CT scanner is the X-ray tube, which produces electromagnetic radiation. The vacuum tube consists of a cathode–anode pair for the emission and collision of accelerated electrons. The filament of the cathode is heated up to release thermal electrons, whose velocity increases due to the acceleration voltage, U_a, between cathode and anode. High temperatures require cooling of the tube, which is realised with coolant. The radiation energy depends on U_a, which ranges between 70 kV and 150 kV for medical-diagnostic purposes. The anode is disk-shaped with a tilt from the edge to the centre for the deflection of the radiation, producing a fan beam. Two constructions of the X-ray tube overcome the issue of electrons always colliding at the same position. Either the anode rotates inside the tube or the anode is connected to the rotating tube, including the cathode. The latter is beneficial for direct cooling of the anode. [21] Furthermore, the radiation beam is regulated with various collimators and filters. The former ensure that only photons with a certain direction pass the patient. The latter are employed to filter low-energy and scattered photons to prevent unnecessary radiation exposure to the patient. In modern CT scanners, the collimators are used to perform simultaneous scans of multiple axial slices. [18]

The attenuated X-ray photons arrive at the detector module, which consists of concatenated units. These are arranged as an array to cover the fan beam. [18] The units can be gas detectors, in which a gas, like xenon, is ionised through the X-ray photons. The electrons produced are captured with high voltage between a cathode and an anode. [20] In modern CT scanners, however, scintillator detectors with a scintillator medium and a photosensor are used. First, the X-ray photons are converted into light whose intensity is proportional to the energy of the photons per unit time. Then, the photosensor converts the visible light into current, which is amplified and digitised. The measured signal needs

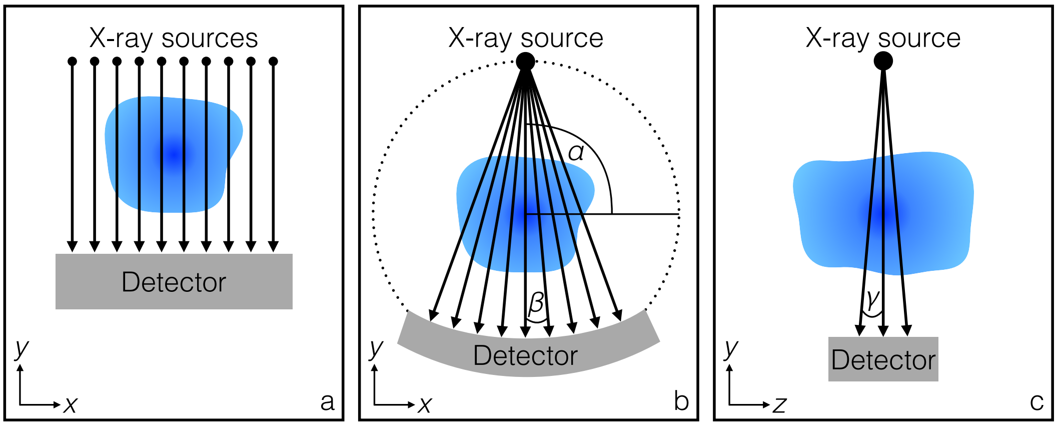


Figure 2.1: Illustration of the different trajectories of the radiation beam. The detector modules are represented as grey areas. Subfigure a: The parallel trajectory requires an individual X-ray source for each beam. Subfigure b: The fan-beam geometry is provided by a single X-ray source. The angle is different for each beam. Subfigure c: Several fan beams are emitted along the z axis for cone beams.

to be corrected as it does not correspond to the intensity of the X-ray photons. Disruptive effects, e.g. the offset and afterglow of the detector and fluctuations in the X-ray tube, are eliminated to obtain the attenuation values. [18] The digital signals are processed with advanced algorithms in the computer system to reconstruct the cross-sectional images. The operation console allows the clinician to adjust the scan parameters and to monitor the scanning sequence. [21]

2.1.2 Image reconstruction

The principle of reconstructing a cross-sectional image from the attenuated radiation intensity

$$I(S) = I_0 \exp\left(-\int_0^S \mu(s)\,\mathrm{d}s\right) \tag{2.1}$$

is based on Lambert–Beer's law. The X-ray photons traverse the patient and interact with matter, which causes a reduction of the initial intensity, I_0. The path, S, of the photons is subdivided into several sections, s, with different attenuation coefficients, $\mu(s)$, depending on the density of the matter. The measurement of $I(S)$ has to be repeated for several angles with a minimal scan range of 180° to increase the quality of the reconstructed image. [20]

The simplest reconstruction algorithm is the filtered back projection for radiation beams with parallel trajectories (see Figure 2.1a) [18]. The measured intensities divided by the initial intensity provide the line integral of the trajectory projection,

$$p(\theta,\eta) = -\ln\left(\frac{I}{I_0}\right) = \int_{-\infty}^{\infty} \mu(\eta, s)\,\mathrm{d}s\,, \tag{2.2}$$

for the angle θ and the distance η from the centre, which corresponds to the Radon transform [22]. The one-dimensional Fourier transform

$$P(\theta,\omega) = \int_{-\infty}^{\infty} p(\theta,\eta)\mathrm{e}^{-2\pi\mathrm{i}\omega\eta}\,\mathrm{d}\eta = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mu(\eta, s)\mathrm{e}^{-2\pi\mathrm{i}\omega\eta}\,\mathrm{d}\eta\,\mathrm{d}s \tag{2.3}$$

of Equation (2.2) is performed to connect the Radon transform to the frequency domain, ω. As a mapping of the attenuation coefficients for the (x, y) coordinates is desired to generate the image, the coordinate substitution of η and s yields

$$P(\theta,\omega) = \int_{-\infty}^{\infty} \mu(x,y)\exp\bigl(-2\pi\mathrm{i}\omega\,(x\cos\theta + \sin\theta)\bigr)\,\mathrm{d}x\,\mathrm{d}y = F\bigl(\omega\cos\theta,\omega\sin\theta\bigr)\,, \tag{2.4}$$

which corresponds to the two-dimensional Fourier transform, $F(\omega\cos\theta,\omega\sin\theta)$. [18] This is the Fourier slice theorem, which means that the frequency domain can be filled with projections from different angles. The inverse two-dimensional Fourier transform leads to the $\mu(x, y)$ distribution, but the scan with discrete angles requires interpolation. Therefore, a high-pass filter can be applied to $P(\theta,\omega)$ in the frequency domain. Several filters exist, which affect the definition and noise of the reconstructed image. Thus, a filter has to be set in the scan protocol before the acquisition. [21]

In practice, the radiation beam corresponds to a fan-beam geometry (see Figure 2.1b) with non-parallel trajectories of the projections. Therefore, the focus point

$$\vec{s}(\alpha) = \begin{pmatrix} R_\mathrm{f}\sin\alpha & -R_\mathrm{f}\cos\alpha & z \end{pmatrix}^\top\,, \tag{2.5}$$

which corresponds to the position of the X-ray source, is parameterised as a circular trajectory at the position z with the rotation angle α and the circumference R_f. The directions of the beams, defined by the angle β, differ within the fan. The coordinate transformations $\theta = \alpha + \beta$ and $\eta = -R_\mathrm{f}\sin\beta$ convert the fan-beam trajectories, (α,β), into the parallel trajectories, (θ,η). One option for the reconstruction is to perform a rebinning that includes interpolation and allows the filtered back projection for parallel trajectories to be applied. Another method directly substitutes $\mathrm{d}\theta$ and $\mathrm{d}\eta$ in the formula for the filtered back projection by $\mathrm{d}\alpha$ and $\mathrm{d}\beta$, which, however, leads to considerable computation. [18]

Furthermore, the extension to multi-slice scans introduced cone beams (see Figure 2.1c) with an additional z component of the focus point for the slices, which are tilted from the axial plane by the angle γ. The reconstruction is based on the filtered back projection for the fan-beam geometry, but includes a multiplication correction of the

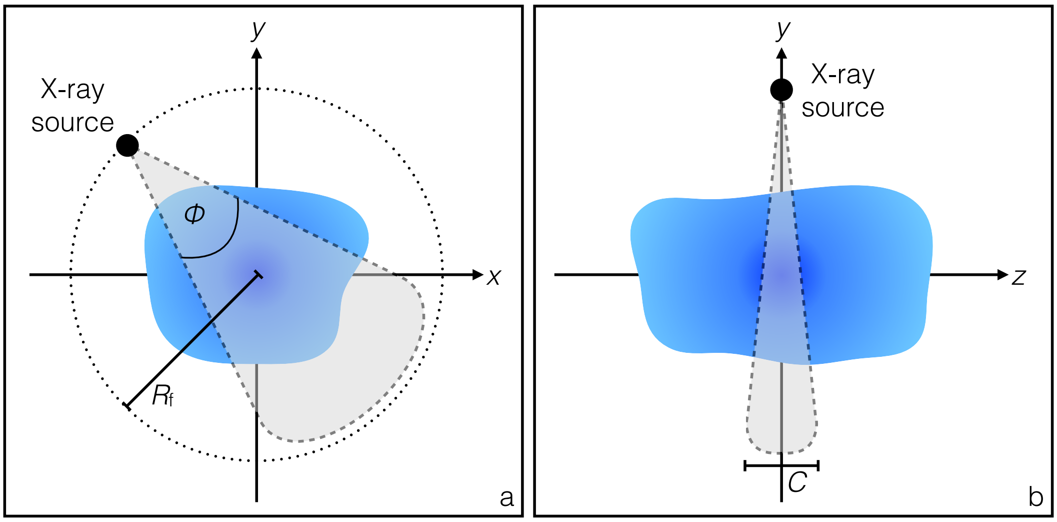


Figure 2.2: Illustration of the cone-beam geometry for CT scans. Subfigure a: The lateral view in the x–y plane shows the circular trajectory of the X-ray source with the circumference R_f and the fan angle Φ. Subfigure b: The longitudinal depth of the cone beam is limited by the collimator, C.

projections with $\cos\gamma$ and a three-dimensional back projection. A drawback of the three-dimensional method is the vulnerability to artefacts due to incomplete data in circular scans. Regarding the common scan technique with spiral trajectories, the focus point differs from Equation (2.5) by replacing z with the constant movement, $d' = d/(2\pi)$, of the patient table with the table distance d per revolution. The lateral angle, Φ, and the longitudinal collimator value, C, of the cone beam, which are visualised in Figure 2.2, define the pitch value, $p = d/C$, of the spiral trajectory. This value affects the image quality and has to be limited. Several algorithms exist for the reconstruction of multi-slice spiral CT scans. One class of algorithms reduces the data to two-dimensional circular scans, which can be used with the filtered back projection for fan-beam geometry. As this method is limited to the angle of the cone beam, algorithms with three-dimensional filtered back projection are more appropriate. This class of algorithms is computationally intense, but produces higher-quality images. [18]

2.1.3 Image display

The reconstructed CT scans contain the attenuation coefficients, $\mu(x, y)$, for each pixel position, which are visualised in greyscale. For this, the greyscale values

$$v_\mathrm{CT} = \frac{\mu - \mu_\mathrm{water}}{\mu_\mathrm{water}} \cdot 1000 \tag{2.6}$$

are computed by transforming the attenuation coefficients into the Hounsfield unit with regard to the attenuation in water. Thus, the greyscale value for water is 0 HU. While the value −1000 HU is obtained for air, the highest value of 3000 HU is reached for bones. The Hounsfield unit given in per mill is advantageous due to the dense distribution of the v_{CT} values of several organs near 0 HU. Another property of image display of CT scans is the limitation to relevant values as the human eye is not capable of differentiating 4000 greyscale values. Therefore, values below or above a specified range, e.g. the bone window from 300 HU to 1800 HU or the brain window from 35 HU to 85 HU, are displayed in black or white, respectively. [20]

2.2 Magnetic resonance imaging

Contrary to CT imaging, the patients are not exposed to ionising radiation in the case of MRI, ensuring a safer imaging technique. The strong magnetic fields used during acquisition can impact the physical well-being of some patients especially as the scan procedure is time-consuming. [18] Moreover, patients with metallic foreign bodies, e.g. implants, are not able to perform an MRI scan due to safety concerns related to the magnetic fields. However, MRI is particularly well suited for the visualisation of soft tissue, such as organs, muscles and the brain, as it provides high-resolution images that support medical diagnostics. [21]

In the context of medical imaging, nuclear magnetic resonance is employed to generate MRI scans with strong magnetic fields. Atomic nuclei in the body, like hydrogen, are excited in the presence of magnetic fields, resulting in nuclear spin resonance. The return of excited nuclei to equilibrium emits high-frequency signals, which are detected with receiver coils. This signal measurement involves the determination of tissue-dependent relaxation times, which provide valuable information on the structure and composition of the tissue. [21]

In Section 2.2.1, the hardware components that are required for the MRI procedure are described. Section 2.2.2 contains the physical concept of nuclear spin resonance, representing the basis of MRI. Lastly, the reconstruction method to produce MRI scans is presented in Section 2.2.3.

2.2.1 MRI scanner

In general, the MRI scanner, generating magnetic fields and emitting high-frequency signals, is constructed of three components: primary magnet, gradient system and high-frequency transmitter. These components are arranged in a housing, which can be cylindrical or open. The former shape is commonly used in clinics with electromagnetic coils, operating between 1.5 T and 3 T. To prevent the impact of external electromagnetic interference, the examination room is built as a Faraday cage. Body-part-specific receiver coils detect the high-frequency signals. Similar to CT scanners, a patient table is used to position the patient. [18]

The basis of the scan procedure is provided by the primary magnetic field, B_0, produced with superconducting electromagnetic coils. The field has to be static and homogeneous to evenly polarise the hydrogen nuclei in the patient. Therefore, additional shim coils are implemented to control and adjust the homogeneity in the isocentre of B_0. The coils of the superconducting electromagnet are vacuumed in a cryogenic vessel, which includes liquid helium for cooling up to 4 K. The superconductivity allows stable magnetic fields to be generated over a longer period of time, which cannot be switched off abruptly. An emergency switch ensures that the superconductivity can be interrupted by heating up a part of the coils. Gaseous helium can then escape through a safety valve. [18]

The gradient system generates linear gradient fields, G_x, G_y and G_z, in the x, y and z directions to enable spatial coding of the signals with position-dependent magnetic fields. The amplitude and the slew rate define the quality of the imaging, which is fast for high amplitudes and high slew rates of the gradient fields. The system consists of electromagnetic coils, which have cylindrical symmetry and are placed under the primary magnet. The fast-switching gradient coils require high current; therefore, water cooling is used to reduce the temperature of the gradient system. [18]

The high-frequency system is subdivided into two parts: transmitter and receiver coils. The former excite the hydrogen nuclei in the patient by applying high-frequency pulses with the Larmor frequency

$$\omega_0 = \gamma B_0 \,, \tag{2.7}$$

which includes the particle-specific gyromagnetic ratio, γ, and is proportional to B_0. A high-frequency amplifier is additionally used to support the transmitter system. The latter coils detect weak high-frequency signals, which are emitted by the excited nuclei in the body. Due to its sensitivity, the receiver system is inactive during the emission of the pulses from the transmitter coils. Moreover, the receiver system, consisting of several surface coils, is placed near the patient to reduce noise effects. The form and number of the coils depend on the body region. [18]

2.2.2 Physical concept

Atomic nuclei, which can contain protons and neutrons, have a nuclear spin, $\vec{I}$, if the number of nucleons is odd. For some cases of evenly distributed nucleons, e.g. odd numbers of both protons and neutrons, atoms also have a nuclear spin. The spin results in a magnetic moment, $\vec{\mu} = \gamma\vec{I}$, which provides the interaction with external magnetic fields for nuclear spin resonance. Regarding MRI, hydrogen is important due to its large presence in the human body. Its protons have the spin quantum number $I = \pm 1/2$, representing low- and high-energy states. The high gyromagnetic ratio of protons in hydrogen exhibits high sensitivity to the magnetic field. In the presence of B_0, the thermal equilibrium of the atomic nuclei is disturbed since the magnetic moment of the protons leads to Larmor precession with the frequency ω_0.

Due to the nuclear spin, two energy states, resulting in parallel and antiparallel distribution of the atomic nuclei, are possible. The Boltzmann distribution, describing the energy-state distribution, yields a magnitude of 10^6. This means that only a few nuclei out of millions contribute to the macroscopic magnetisation, $\vec{M}$, which is proportional to the measured signal in MRI. An increase in B_0 amplifies the high-frequency signal. [18]

The macroscopic magnetisation with a static magnetic field requires an alternating magnetic field, $B_1(t)$, to cause nuclear spin resonance. For this, the field $B_1(t)$, which is perpendicular to B_0 in z direction, is applied as high-frequency pulses to tilt $\vec{M}$ through a precession around the z axis. The amplitude and the duration of $B_1(t)$ affect the flip angle, α. While a flip angle of 90° tilts $\vec{M}$ into the transverse x–y plane, lower α values lead to a smaller tilt of the longitudinal component M_z into the transverse plane. The requirement for this resonance is that the frequency of the pulses corresponds to the Larmor frequency. After the pulse is switched off, the precession of the magnetisation decreases as the magnetisation $\vec{M}$ returns into the z direction. This emits the characteristic signals, which are detected with electromagnetic coils, sensitive to the rotating transverse component M_{xy}. [21]

Relaxation describes the process for a system to return to equilibrium. Two mechanisms, which affect the image contrast of the MRI scan, are the basis of the MRI procedure. The longitudinal relaxation is related to the return of the longitudinal magnetisation

$$M_z(t) = M_0\left(1 - \exp\left(-\frac{1}{T_1}t\right)\right) \tag{2.8}$$

to the initial state, M_0. The relaxation time T_1 indicates that the magnetisation is approximately 63 % in equilibrium, while $M_z(3T_1) \approx 0.95M_0$. The longitudinal relaxation, which results from the interaction of the spins with the molecular environment, depends on the tissue type and B_0. In contrast, the transverse relaxation represents the exponential decay of the transverse magnetisation

$$M_{xy}(t) = M_0 \exp\left(-\frac{1}{T_2}t\right) \tag{2.9}$$

with T_2 as the relaxation time for the decrease of M_{xy} to 37 %. This relaxation results from spin–spin interactions, leading to the divergence of M_{xy}. Due to inhomogeneous magnetic fields, the transverse relaxation is affected by additional dephasing effects, which produce the free induction decay in the receiver coils. The signal is a decreasing oscillation whose envelope decreases exponentially with T_2^*, similar to Equation (2.9). The two relaxation processes are visualised in Figure 2.3. The T_1 and T_2 relaxation times strongly depend on the tissue type and the magnetic field. Both times provide valuable information on the behaviour of the protons in the tissue, influencing the signal. [21]

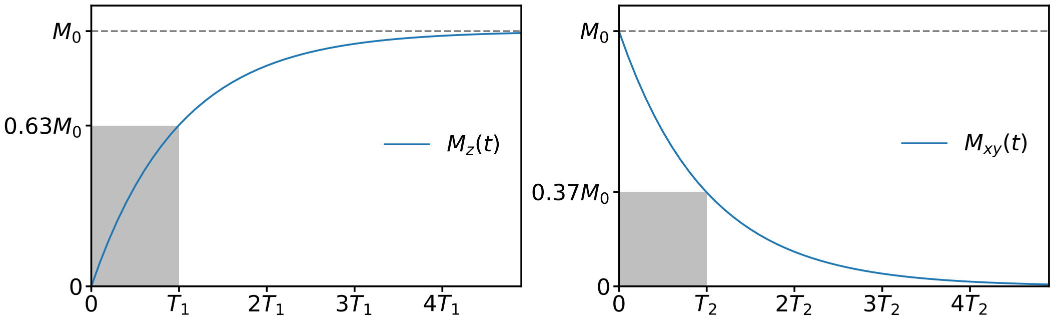


Figure 2.3: Relaxation processes during the MRI acquisition. The distributions of the longitudinal (left) and transverse (right) relaxations are shown. The grey areas indicate the states of the magnetisation components $M_z(t)$ and $M_{xy}(t)$ for the relaxation times T_1 and T_2, respectively.

2.2.3 Image acquisition

The gradient system of the MRI scanner generates three linear magnetic fields, which are superposed upon B_0. This allows the spatial domain to be subdivided into volume elements, defining the spatial resolution of the MRI scan. The signals, evoked by the high-frequency pulses, depend on the frequency

$$\omega(x, y, z) = \omega_0 + \gamma \left(xG_x(t) + yG_y(t) + zG_z(t) \right), \tag{2.10}$$

which is used for the spatial coding. A certain sequence of gradient fields ensures that the signal of the corresponding volume elements can be measured. First, specific slices within the volume can be selected for the signal measurement with high-frequency pulses, which affect only magnetic moments with the resonance frequency of the pulses. This can be controlled by applying the gradient field G_z during the pulse. The thickness of the selected slices can be adjusted through the bandwidth of the high-frequency pulse and the slew rate of G_z. Then, the spatial coding in the x–y plane is performed with a phase and frequency coding. After the selection of the slice, the gradient field G_y is applied for a certain time to obtain different phases along the y direction. As the signal of each phase has to be detected individually, the field G_y has to be applied several times with different strengths. The gradient field G_x is used during the measurement of the signal. It changes the precession frequency of the magnetisation along the x direction, which characterises the signal by the individual frequency $\omega(x) = \omega_0 + \gamma x G_x$. The spatial coding ensures that a discrete frequency domain is filled with the measured signals to generate the MRI scan with the Fourier transform. [21]

The advantage of the MRI procedure is the variability of gradient and pulse sequences, which highly affects the image contrast of the resulting scan. The echo time, T_e, and the repetition time, T_r, are two essential parameters to control the impact of both relaxation processes. The inversion-recovery sequence is used to determine the relaxation time T_1 with 180° pulses, which invert the direction of the longitudinal magnetisation. The magnetisation is forced to return to the initial state during an inversion time, T_i. To detect the signal, a 90° high-frequency pulse is applied to the recovered magnetisation, producing a free induction decay whose amplitude corresponds to the longitudinal magnetisation. A repetition of the 90° pulse is necessary to measure the exponential distribution of the longitudinal relaxation with different T_i values. The repetition time, considering the recovery of the magnetisation, defines the interval for the next 180° pulse. [21] In the spin-echo sequence, a 90° high-frequency pulse is applied to tilt the longitudinal magnetisation into the transverse plane. A dephasing is caused by the inhomogeneous magnetic fields. Therefore, an additional 180° pulse is used after the time $T_\mathrm{e}/2$ to flip the magnetisation, which means that the dephasing is reversed. After the time T_e, the magnetisation is refocused, producing a spin-echo signal. The application of the second pulse is repeated several times to determine T_2 from the decreasing amplitudes of the spin-echo signals. [21] Moreover, the gradient-recalled sequence contains angles of $\alpha \leq 90°$ to not flip the magnetisation completely into the x–y plane, reducing the time to return to the initial state. This sequence is often used in combination with the inversion-recovery sequence. [18]

The choice of the repetition and echo times correlates with the weighting of the MRI scan. The inversion-recovery sequence is appropriate for generating T_1-weighted images, while the spin-echo sequence is more variable in terms of image contrast. A large repetition time and an echo time that approximately corresponds to T_2 are required for T_2-weighted MRI scans via the spin-echo sequence. For T_1-weighted MRI scans, a short echo time has to be set, while the repetition time should match T_1. Due to large repetition times of some seconds, the acquisition of a two-dimensional image can take several minutes. Therefore, multi-slice acquisition can be performed for a three-dimensional scan, where other slices that are at a sufficient distance from each other are measured during the waiting time. [21]

3 Digital image processing

In medical imaging, the acquired information is presented as digital images to evaluate the anatomy of the patient. These can be used for the indication of diseases or the planning and monitoring of a treatment, like surgery, radiotherapy or chemotherapy. In Section 3.1, the construction of digital images is introduced, which includes typical image types and medical terms. Basic operations for image processing as well as deep-neural-network operations are summarised in Section 3.2. Afterwards, three metrics aiming at measuring similarity in the image-registration process are described in Section 3.3.

3.1 Digital images

In computer science, images are realised by arranging rectangular picture elements (pixels) in columns and rows. The numbers of columns and rows define the width and the height of the image, respectively. The resolution, indicated by the number of pixels per unit length, connects the represented object in the image with its real dimensions. Each pixel of the grid-shaped image contains a specific value or a set of values, which depends on the image type, e.g. greyscale images with one channel or colour-scale images with three channels. The information of the images is expressed by a finite set of numerical values that defines the range of pixel values as integers or floating-point numbers. Usually, the range is defined by 2^k values with k as the bit depth, increasing image quality with higher values. [23]

In general, an image, $I(x, y)$, is a two-dimensional matrix of a set of pixel values for the coordinates x and y. Since the medical scan of a patient contains a series of two-dimensional images, the scan is represented as a three-dimensional array, $I(x, y, z)$, with an additional axis for the coordinate z. The coordinate system of the array with its origin at $(0, 0, 0)$ in the upper left corner defines the location of each pixel. An illustration of a three-dimensional array is shown in Figure 3.1a. An axial image is displayed in the x–y plane, where the x and y axes run from left to right and from back to front, respectively. The other images or slices of a three-dimensional scan are concatenated along the z axis from top to bottom.

In this thesis, the PYTHON programming language [24] is used for image processing. The NUMPY library [25] includes several processing tools to access, manipulate and change the information of the images in the form of arrays. The IMAGEJ programme [26] and the MATPLOTLIB library [27] are used for the visualisation of the images.

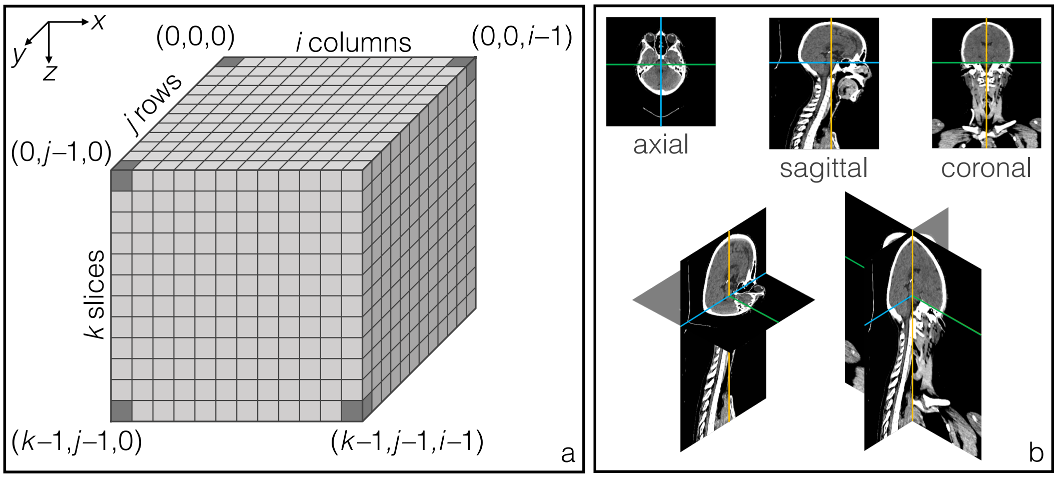


Figure 3.1: Illustration of a three-dimensional array and medical perspectives. Subfigure a: The pixels are arranged in a three-dimensional array with i columns, j rows and k slices. The coordinate system defines fixed discrete pixel positions and has its origin in the upper left corner. Subfigure b: The axial, sagittal and coronal perspectives with regard to a medical scan as well as their combination are visualised to show the three-dimensional mapping.

3.1.1 Image types

Several types of depiction exist for digital images, which depend on the number of channels and the bit depth. Images with one channel are called greyscale images and contain one value per pixel. In contrast, colour-scale images contain a set of three or four values per pixel in the form of the RGB or CMYK colour model. Furthermore, images can also include negative pixel values. These special images are helpful during the application of image-processing techniques that generate negative values. [23]

Binary images with $k = 1$ are a subset of greyscale images. This means that only two integers are candidates for the pixel value. The result is a black-and-white representation, which separates the foreground from the background with the pixel value 0. Other bit depths, for example, are 8, 12, 14 or 16, increasing the range of pixel values. An 8-bit image contains 256 integers to display different grey values. The number of integers can extend to 65 536 for $k = 16$. [23]

The bit depth of colour-scale images is similar to that of greyscale images, but is multiplied by the number of channels. The typical numbers 24, 36 or 42 are used to represent the RGB colour space with the same integer range for each channel. For example, a 24-bit image has 256 possible integers per pixel and channel. The CMYK colour space is realised by four channels with $k = 32$ and an integer range of $[0, 255]$ per channel. [23]

3.1.2 Medical images

The standard to manage the outcome of medical imaging is the DICOM [28] format, which stores information on the patient, the scan sequence and the image properties. As mentioned in Chapter 2, the studies in this thesis employ CT and MRI scans for image registration and fusion.

The pixel values of the CT scans are constructed of 12-bit integers with a range of $[0, 4095]$. The speciality of the CT scans is the connection of the pixel values, x, to the Hounsfield unit, which can be accessed with the linear calibration function

$$y = ax + b\,. \tag{3.1}$$

The Hounsfield values, y, are calculated via a scaling factor, a, and a shift parameter, b. For the CT scans in this thesis, the numbers are $a = 1$ and $b = -1024$, defining the Hounsfield range $[-1024, 3701]$ (see Section 2.1.3). This facilitates the identification of tissue, e.g. fluids and bones, for image processing. Contrary to CT, MRI scans have an arbitrary distribution of pixel values. Although 16 bits are allocated to these scans with signed integers, the pixel values vary between 0 and 9999.

The three-dimensional scan of a patient can be visualised through different perspectives, exemplified in Figure 3.1b. For images of the head, the side view is realised with the sagittal plane, which spans in the x–z plane. The coronal plane is given by the y and z axes and splits the head into a rear and front part. The common perspective of the x–y plane is called the axial plane. [23]

A set of parameters that is stored in the DICOM files provides information on the resolution of the scans. The pixel spacing, given in the range of millimetres, specifies the dimensions of the pixels in the x and y directions. In addition, the slice thickness and distance provide information on the extension along the z axis. The slice thickness can either be larger or smaller than the slice distance, which depends on the scan parameters, set during the acquisition. In the latter case, some parts of the patient's anatomy are not scanned, reducing the resolution of the scan.

3.2 Image processing

A modification of the images can be achieved by various techniques, depending on the objective. The simple geometric transformations described in Section 3.2.1 are useful for changing images regarding the position, size or form. To manipulate the pixel values of images, filter operations can be applied to influence the edges by softening or sharpening (see Section 3.2.2). Furthermore, image operations for special deep neural networks with the aim of extracting image features are explained in Section 3.2.3.

3.2.1 Geometric transformations

The transformation of a source image, S, is performed with a transformation function, A, to generate the target image, T. The function is an $n \times n$ matrix, which contains parameters to compute the target pixel positions

$$\vec{x}' = A \cdot \vec{x} \tag{3.2}$$

based on the source positions, $\vec{x}$. The size, n, of A depends on the dimension of the source image, which leads to $n = 3$ and $n = 4$ for two- and three-dimensional images, respectively. The number of parameters in A varies with the number of geometric operations that are chosen for the transformation. [23] Transformations of an image can be classified into three types: rigid, affine and deformable. The rigid transformation is characterised by the fact that the pixel distances are maintained, which is accomplished by translation, rotation and reflection. By adding operations like scaling and shearing, the transformation type turns from rigid to affine. This means that the pixel distances change, while the parallelism of lines is preserved. The deformable transformation warps the image with unique displacement vectors for each pixel. Thus, the number of displacement parameters increases with the dimension and the number of pixels. [4]

Operations For the translation operation, one parameter, t_i, per dimension guarantees the displacement along the accompanying axis, i. The rotation of the image around the centre requires an angle, α. The two-dimensional transformation functions for rigid operations are defined as

$$A_{\text{translation}} = \begin{pmatrix} 1 & 0 & t_x \\ 0 & 1 & t_y \\ 0 & 0 & 1 \end{pmatrix} \quad \text{and} \quad A_{\text{rotation}} = \begin{pmatrix} \cos\alpha & -\sin\alpha & 0 \\ \sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{3.3}$$

Regarding the affine operations, the parameter s_i scales the image along the axis i, while the shearing parameter b_j of axis j is applied to distort the image along the axis i. In general, the transformation functions

$$A_{\text{scaling}} = \begin{pmatrix} s_x & 0 & 0 \\ 0 & s_y & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{and} \quad A_{\text{shearing}} = \begin{pmatrix} 1 & b_x & 0 \\ b_y & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{3.4}$$

are used for two-dimensional images. [23] Furthermore, the reflection of images always refers to an axis in two dimensions or to a plane in three dimensions. This operation is useful for data augmentation in machine-learning processes (see Section 5.3.1).

Mapping The source-to-target mapping enables the generation of T by computing the modified pixel positions as defined in Equation (3.2). This method can lead to insufficient results, e.g. if multiple pixel values of S are assigned to the same pixel in T. Hence, the target-to-source mapping is the superior option, avoiding these complications. For each pixel of T, the position in S is calculated with the inverse transformation function, A^{-1}. Since the reverse computation leads to positions between several pixels, this mapping requires an interpolation for the choice of the pixel values in T. Several interpolation methods, using neighbouring pixel information, are available for the transformation. The nearest-neighbour method takes the value of the next pixel by rounding up the calculated position in S. The target image can appear pixelated due to the single pixel information. The quality of T improves by increasing the number of pixel values in the interpolation process. The bi-linear interpolation uses the four pixel values that are adjacent to the calculated pixel position by applying linear interpolations along the x and y axes. A more detailed and computationally intense interpolation is achieved with the bi-cubic method, occupying the 16 nearest pixel values for the determination of cubic polynomial functions. [23]

3.2.2 Manipulation

In image processing, direct manipulation of the pixel values can be used to produce filtered images, which can support the search for edges or contours, or to generate binary images. Regarding image segmentation, algorithms like the flooding algorithm are useful to fill closed contours in binary images [23].

Global filters The application of filters aims at modifying the image to reduce noise or to increase its smoothness. Linear filters, affecting the whole image, have the disadvantage that desired features are also manipulated. In contrast, non-linear filters allow the specification of the properties to be filtered. Therefore, the non-linear operation computes the new pixel values according to a filter region by using a kernel with a specified size. As the centre of the kernel is placed on the pixel under consideration, the kernel dimensions should be odd. The maximum and minimum filters, for example, select the highest and lowest values of the kernel, respectively. The median filter determines the median value of the kernel. [23]

Morphological filters To change the form and size of the image content, morphological filters are useful. These filters are especially designed for binary images, but they are also applicable to greyscale and colour-scale images. The basic operations, called dilation and erosion, are related to the increase and decrease in size, respectively. These operations include a structural kernel with an origin, which is applied to each pixel position of the image. Regarding the image type, the pixel values of both image and kernel can either be zero or unity in case of binary images. For the dilation, the values of the kernel are transferred to the image if the value of the origin and the corresponding pixel value agree. In contrast, the erosion, where the kernel is mapped onto the image if all values of the kernel are in agreement with the pixel values in the corresponding area, reduces the size. Furthermore, the combination of dilation and erosion leads to two other operations: opening and closing. The former consists of the successive application of erosion and dilation, which removes image features smaller than the kernel and smooths the image afterwards. For the latter, dilation and erosion are applied to close gaps that are smaller than the kernel. [23]

Thresholding The threshold technique converts greyscale or colour-scale images into binary images. For this, a pixel value has to be determined as the threshold for the separation into foreground and background. The computation of the threshold is often performed with the image histogram. The isodata method, for example, calculates the mean of the pixel values based on the histogram. This divides the histogram into two parts, for which the means are again calculated to compute their average, representing the threshold. The iterative procedure is repeated until the threshold does not change. [23] Other methods, like Yen [29] and Li [30], are based on the entropy of the pixel-value distribution. While the Yen method is designed to determine a threshold for a maximal entropy through the probability distribution of the pixel values, the Li method minimises the cross-entropy according to the Kullback–Leibler divergence [31].

3.2.3 Deep-neural-network operations

Deep learning is one option to apply deformable transformations to images for direct and fast image registration. Deep neural networks are used to extract image features through simple operations, which are implemented in the networks and contribute to their architecture. In this thesis, two types of operations are needed; while convolution represents the basis, image contraction and expansion enable resizing.

Convolution The aim of extracting features from images is achieved with convolution operations. The foundation of a convolution is a k-dimensional kernel, having the same dimension as the image. The sequence to obtain two-dimensional feature maps, shown in Figure 3.2, is described in the following. The kernel of size $n \times m$ contains the weights w_{nm}, which are multiplied by the corresponding pixel values during the sliding procedure. For the pixel position p_{ij}, the $n \cdot m$ multiplied values are added to determine the value of the feature map f_{ij}. The properties of the convolution are defined by the stride and the padding. The stride specifies the step width of the kernel, while the padding controls the computation outside the image by adding a specified value. Both parameters impact the size of the resulting feature map. For example, a stride of two

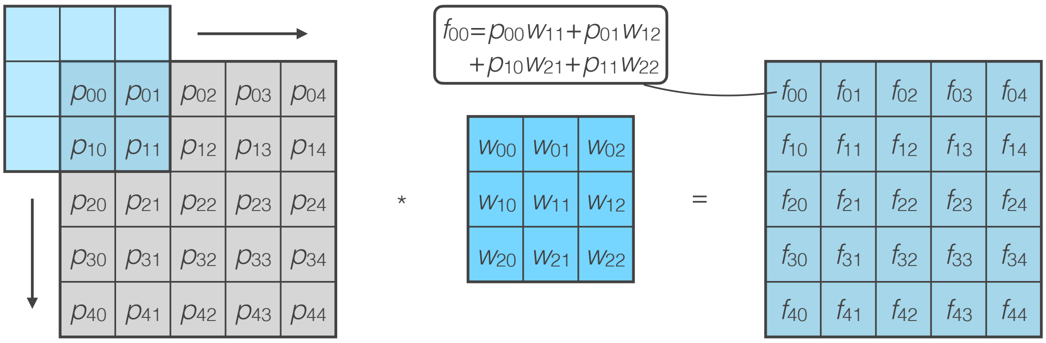


Figure 3.2: Procedure of the convolution operation. The kernel (blue) is applied to the image (grey) to produce the feature map (blue greyish). For this, the kernel slides over the image to determine the values f_{ij}. The size of the feature map depends on the stride and the padding. The process is visualised for a stride of one and zero padding. For each pixel position, the weight parameters w_{nm} and the accompanying pixel values p_{ij} are multiplied, and the sum is calculated afterwards. An example is shown in the rounded box for f_{00}.

skips every second pixel, and zero or same padding places the centre of the kernel at the origin of the image by extending the image with zeros or values of the image border. [32]

Resizing In deep neural networks, the extraction of image features is done with several levels of convolution operations and a reduction of the size of the resulting feature maps. For this, a pooling operation, which includes an empty window of size $\tilde{n} \times \tilde{m}$, is applied in the same way as a convolution, but following a specified rule. Max-pooling, for example, selects the highest value of the feature map according to the window dimensions. The size of the feature map in the next level is affected by the stride and the padding. Contrary to pooling, upsampling reverses the process to obtain feature maps of similar or the same size from the beginning. This operation, for example, doubles the size of the feature map and transfers each value in a 2 × 2 pattern to the feature map in the next level.

3.3 Image-matching metrics

A major challenge of multimodal approaches in DIR is the variation of the intensity distributions associated with different types of tissue. The metrics have to be chosen with the intention of measuring the alignment of image pairs. Two types of metrics are typically used for registration purposes: intensity-based metrics and feature-based metrics. The former employ the intensity distributions of the images, while the latter require additional information about image features, such as contours. [4]

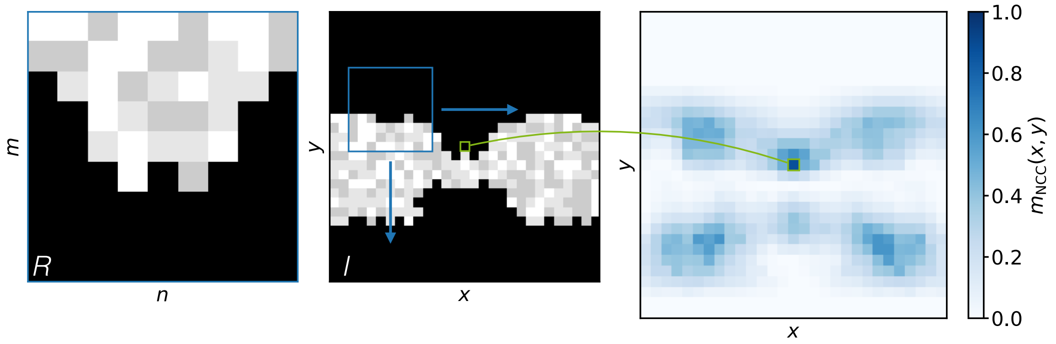


Figure 3.3: Illustration of the computation of the normalised cross-correlation with a reference template. The template, R, is overlaid with the image, I, for each coordinate (x, y) to calculate the metric. The $m_{\mathrm{NCC}}(x, y)$ values indicate the best alignment between template and image for the pixel coordinate with the highest value (green bordered).

3.3.1 Normalised cross-correlation

One option to assess the alignment of an image, I, with a reference template, R, is the intensity-based computation of the normalised cross-correlation [33]. The $n \times m$ template is passed across the image to calculate the metric

$$m_{\mathrm{NCC}}(x, y) = \frac{\left(\sum_{r=-i}^{i} \sum_{s=-j}^{j} I'(x, y, r, s) R'(x, y, r, s)\right)^2}{\sigma_I(x, y) \sigma_R(x, y)} \tag{3.5}$$

for each pixel coordinate (x, y) with the functions

$$F'(x, y, r, s) = F(x + r, y + s) - \bar{F}\,, \tag{3.6}$$

$$\sigma_F(x, y) = \sum_{r=-i}^{i} \sum_{s=-j}^{j} \left(F(x + r, y + s) - \bar{F}\right)^2. \tag{3.7}$$

The indices i and j in Equation (3.5) are the local coordinates in the $n \times m$ region for the current position at (x, y). The term F' calculates the difference between the pixel value at $(x + r, y + s)$ and the average value $\bar{F}$ of the template or the corresponding $n \times m$ region in the image. The term $\sigma_F(x, y)$ represents the variance. The m_{NCC} value ranges between 0 and 1. It indicates better and worse alignment for values towards unity and zero, respectively. Due to its normalisation, the metric is robust against differences in the pixel values. [23] An example of a reference-matching procedure is presented in Figure 3.3. Besides global templates, the metric in Equation (3.5) can also be used to measure the alignment of two images, I and J. Stationary $n \times m$ templates for each pixel position are defined in both images to calculate m_{NCC}. Then, the mean value quantifies the image alignment of I and J.

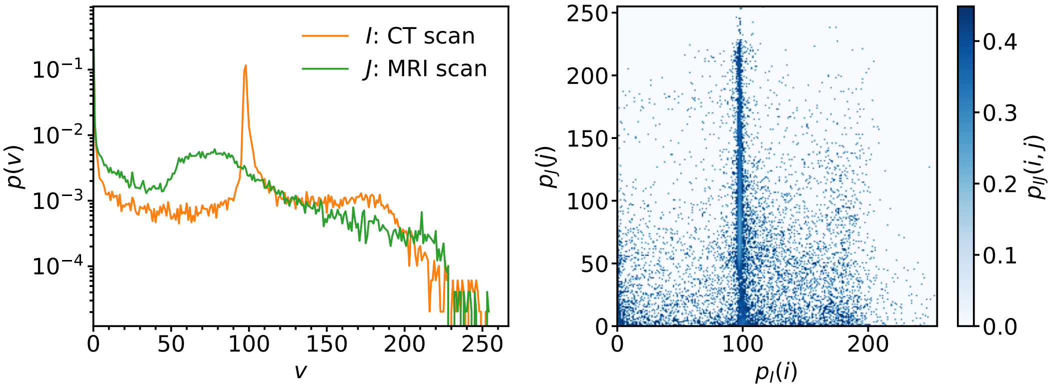


Figure 3.4: Probability distributions of the pixel values of two images for the computation of their mutual information. The distribution $p(v)$ is shown (left) individually for the images I and J, corresponding to CT and MRI scans, respectively. The joint distribution $p_{IJ}(i, j)$ of I and J is visualised (right) as a two-dimensional histogram.

3.3.2 Mutual information

Similar to m_{NCC}, the mutual-information metric is intensity based [4]. It is defined as an entropy measure of the pixel values of two images, I and J, according to the Kullback–Leibler measure [31, 34]. The metric,

$$m_{\mathrm{MI}}(I, J) = \sum_{i\in I}\sum_{j\in J} p_{IJ}(i, j) \ln\left(\frac{p_{IJ}(i, j)}{p_I(i)\,p_J(j)}\right), \tag{3.8}$$

assesses the image similarity by exploiting the probability distributions of the pixel values i and j. The individual probability distributions, $p_I(i)$ and $p_J(j)$, are computed as the number of pixels divided by the total number of pixels. The joint distribution, $p_{IJ}(i, j)$, results from the two-dimensional histogram of the pixel-value distributions of I and J. This technique provides a statistical measure, which is independent of the absolute values [4]. Image similarity increases for higher values of m_{MI} [34]. Example distributions are shown in Figure 3.4 for one image pair, consisting of CT and MRI scans.

3.3.3 Dice similarity coefficient

Another possibility to determine the alignment of two images, I and J, is the Dice similarity coefficient [35]

$$m_{\mathrm{DSC}}(S_I, S_J) = \frac{1}{N_s}\sum_i \frac{2\left|S_I(s_i) \cap S_J(s_i)\right|}{\left|S_I(s_i)\right| + \left|S_J(s_i)\right|}\,. \tag{3.9}$$

This feature-based metric requires the segmented versions, S_I and S_J, of the images to determine the overlap, $\left|S_I(s_i) \cap S_J(s_i)\right|$, of the segments, s_i, relative to their individual

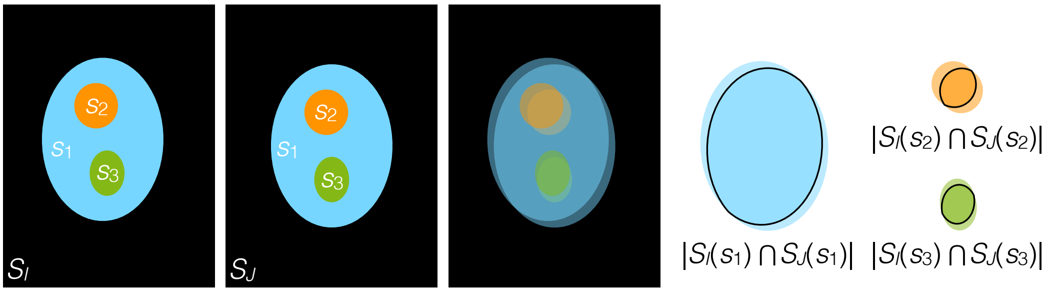

s_2
s_1
s_3
S_I
s_2
s_1
s_3
S_J
$|S_I(s_2) \cap S_J(s_2)|$
$|S_I(s_1) \cap S_J(s_1)|$
$|S_I(s_3) \cap S_J(s_3)|$

Figure 3.5: Illustration of the computation of the Dice similarity coefficient with segmented images. The segmented images, S_I and S_J, contain the segments s_1, s_2 and s_3, which are overlaid to determine the overlap fractions $\left|S_I(s_i) \cap S_J(s_i)\right|$, required for the calculation of $m_{\mathrm{DSC}}(S_I, S_J)$.

volumes, $S_I(s_i)$ and $S_J(s_i)$. The mean value of all segments assesses the image alignment. An example of the computation of m_{DSC} is presented in Figure 3.5. Consequently, the meaningfulness of this metric is limited by the number of segments and their volume. The larger the region of an image covered with segments, the larger the information content of the metric. A segment pair, representing the same tissue with the same pixel value, can be generated using contours or automated algorithms (see Section 4.2). An increase in the overlap of the segments is measured for m_{DSC} values towards unity, while the value $m_{\mathrm{DSC}} = 0$ means no overlap.

4 Image preprocessing

This chapter portrays the evolution of the data from the acquired to the preprocessed state by means of an unsupervised registration workflow. It is developed and constructed on the given data sets to standardise a set of multimodal images in terms of the image format and alignment with affine transformation and rigid registration. In Section 4.1, the origin, the composition and the handling of the data sets are described in detail. Afterwards, the main components, which are self-written algorithms for image segmentation and adjustment, are introduced in Sections 4.2 and 4.3, complementing the workflow. The outcome is presented with quantitative measures and qualitative comparisons of the preprocessed data in Section 4.4. The preliminary step is a necessity for DIR with a deep neural network, which is described in Chapter 5.

4.1 Data sets

The data are provided by the West German Proton Therapy Centre Essen (WPE). The clinic is specialised in proton therapy of static tumours in different regions, such as the head, the spine or the pelvis. In contrast to photon therapy, the unique irradiation properties of protons offer the advantage of high protection of organs at risk. Hence, one focus of the WPE lies on the treatment of children. [36]

The data belong to the *KiProReg* register study (ID: DRKS00005363) from 14 October 2013, created at the German Clinical Trials Register. The purpose of this study is to preserve the results of treatment planning and application for future evaluation of the efficacy of proton therapy. The collection includes information on patients of all sexes suffering from tumours. Further requirements for the admission are a maximum age of 17 and the indication for radiotherapy. For the latter, the patients are treated at the WPE with proton therapy instead of other approaches, e.g. photon therapy. Recorded consent was issued by the parents and patients. Furthermore, the Ethics Committee of the University Duisburg–Essen approved the realisation of the patient registration (13-5544-BO, 10 September 2013). [37]

Within the scope of this thesis, two data sets are used to study multimodal image registration and fusion. The number of patients differs between the data sets; therefore, a smaller and a larger data set are present. Both contain CT and MRI scans of patients with brain tumours. The CT scans are obtained with a Brilliance BigBore Scanner from Philips GmbH Health Systems. The scans are performed in the spiral mode with a peak kilovoltage of 120 kV. A laser system from LAP GmbH supports patient positioning for

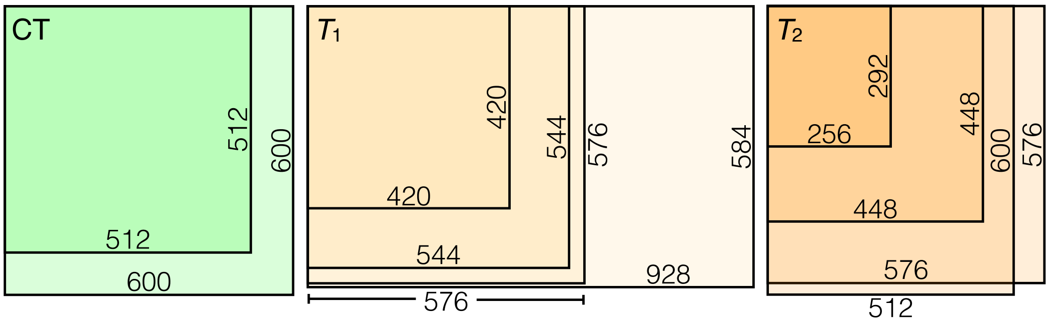


Figure 4.1: Illustration of the existing image frames and aspect ratios for the 14-patient data set. The frames of the CT scans as well as the T_1- and T_2-weighted MRI scans are shown with the accompanying numbers of pixels.

both medical imaging and treatment. For the MRI scans, the Vantage Titan$^{\mathrm{TM}}$ from Canon Medical Systems is used to produce differently weighted scans. The open bore machine is operated at 1.5 T. The gradient-recalled scan sequence, which is combined with the inversion-recovery sequence, and the spin-echo sequence are used for T_1 and T_2 weighting, respectively. All scans are stored in the DICOM format (see Section 3.1.2).

4.1.1 The 14-patient data set

The *KiAPT* study [38] of the WPE is aimed at investigating the use of synthetic CT scans for adaptive proton therapy on children with brain tumours. The *KiAPT* study is connected to the *KiProReg* study and was approved by the Ethics Committee of the University Duisburg–Essen (18-8320-BO, 1 October 2018). Data of the study were provided for this thesis to form the 14-patient data set. For each patient, scans of the different modalities were performed on the same day before treatment application.

The CT and MRI scans of each patient have individual properties related to their spatial extent. The head width of the pediatric patients varies approximately between 120 mm and 146 mm. Concerning the axial slices, the difference in pixel spacing offers a variety of possible image sizes. The aspect ratio differs between the scans, but a ratio of 1 : 1 is predominant. An illustration is shown in Figure 4.1. Furthermore, the number of CT slices varies between 258 and 364, whereas the ranges of 156 to 250 and 67 to 108 are observed for the T_1- and T_2-weighted MRI scans, respectively. Detailed information on the characteristics is listed in Table A.1 in the appendix. The CT slices have a constant thickness of 1 mm, which is not the case for the T_2-weighted MRI scans. There, the slice thickness can be 2 mm, like for all the T_1-weighted MRI scans, or 3 mm. A further quantity is the slice distance, which is relevant for the image resolution. The distance is equal to the slice thickness for all CT scans of the data set. Contrary to that, the majority of the T_1-weighted MRI scans include overlapping slices due to a spacing of 1 mm. For the T_2-weighted scans, the anatomical coverage is incomplete because of distances larger than the slice thickness. An overview of the slice properties is shown in Figure A.2.

Apart from the technical perspective, the comparison of the CT and MRI scans given in Figure 4.2 provides a visual representation of the spatial differences, indicating the benefits of each image type. The visibility of the anatomical structures is influenced by the modality-dependent image contrast. Bones, for example, are clearly present in the CT slices due to high pixel values (see Figure 4.2a). This is not the case for the MRI scans in Figures 4.2b and 4.2c, where bones are expressed by low pixel values, resulting dark in the images. The inverted effect appears for fluids in the CT and T_2-weighted MRI scans, which is the reason for bright eyes in the latter case. High contrast distinction between healthy tissue and tumours is especially important for treatment planning in radiotherapy. This effect is given by the T_2-weighted MRI scans in Figure 4.2c, where the quality of tumour distinction is higher compared to the CT and T_1-weighted MRI slices. Noise around the head is visible in the MRI slices, whereas the CT slices contain fragments of the patient table or the headrest. Moreover, the images, in particular the three-dimensional views, illustrate the position and the size of the patient related to the image frame. A slight tilt of the head is present in the MRI scans, which can occur during the acquisition without a fixing headrest. This complicates the search for corresponding slices, but images with similar anatomical structures still can be found for different slice numbers.

The 14-patient data set is used for the development of data preprocessing algorithms to eliminate the previously described disagreements between the modalities. Regarding the image registration, the data set is first chosen for preliminary studies on the deep neural network. Then, the images of this data set are defined as testing data for a comprehensive study on DIR with deep learning.

4.1.2 The 25-patient data set

Additional data were provided by the WPE for the 25-patient data set. These patients are part of the *KiProReg* study and fulfil the requirements of the register study. The CT and MRI scans are used as the planning scans for proton therapy. The scans were usually performed on the same day, but in exceptional cases, the MRI scan was acquired at most 30 days before the CT scan.

For this data set, the variation of the image frames is shown in Figure 4.3. An aspect ratio of 1 : 1 is mainly observed, and various image sizes are available for both MRI types. The number of CT slices as well as their thickness and distance are comparable to the 14-patient data set, but the number of slices of the MRI scans is reduced. The number ranges from 96 to 232 for T_1-weighted and from 33 to 94 for T_2-weighted MRI scans. Further image properties are presented in Table A.2 and Figure A.2 in the appendix. In general, the images of the data set are similar to the images in Figure 4.2. However, differences appear especially for T_2-weighted MRI scans because of the low number of

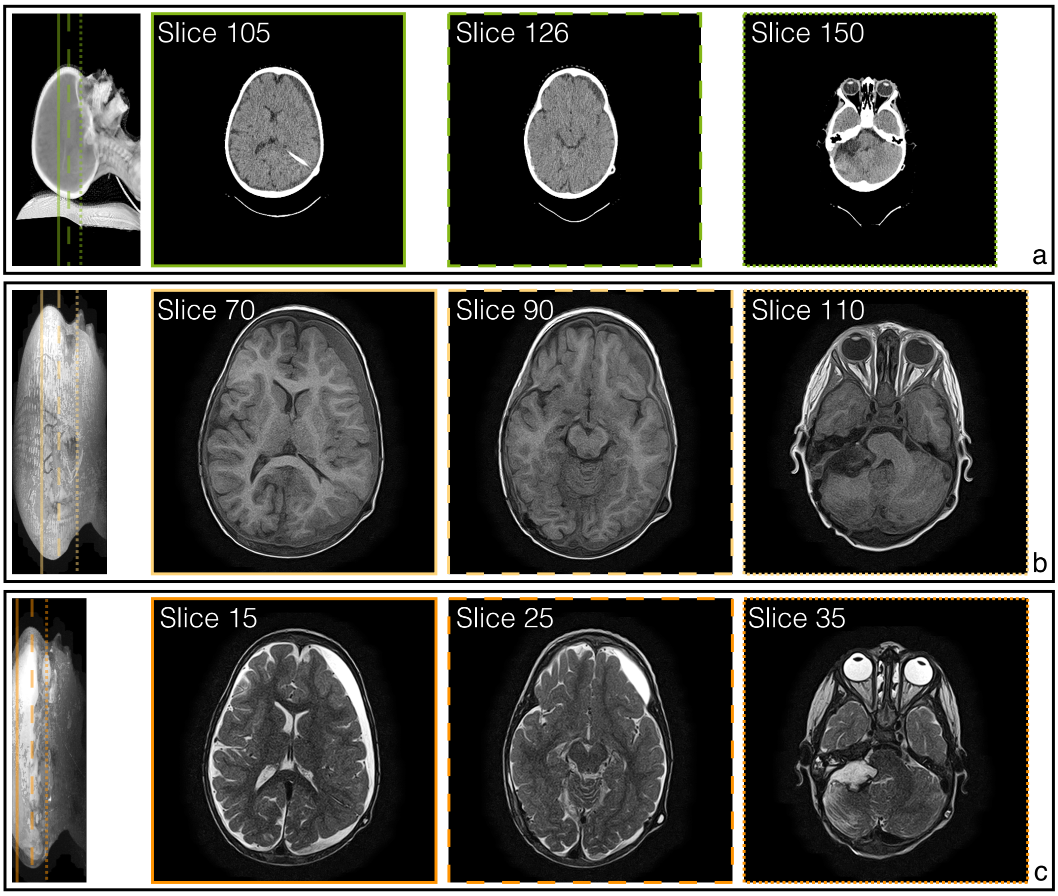


Figure 4.2: Visualisation of similar anatomical structures for each scan of Patient 5 of the 14-patient data set. The three-dimensional perspective (left) and three axial slices (colour bordered) are shown for the CT scan (green) as well as the T_1-weighted (light orange) and T_2-weighted (orange) MRI scans. The lines in the three-dimensional images represent the position of the corresponding axial slices. Subfigure a: The Slices 105, 126 and 150 show the image property of the CT scan in the axial plane. The depiction of the bones (white) is enhanced, while the visibility of fluids (dark grey) is reduced. Subfigure b: For the T_1-weighted MRI scan, the Slices 70, 90 and 110 are presented, showing higher contrast of soft tissue (grey). In addition, bones are displayed dark, like fluids. Subfigure c: The T_2-weighted MRI slices differ from those of the CT and T_1-weighted scans in terms of image contrast. Fluids (white) in the brain appear bright, such as the ventricles in Slice 15 and the eyes in Slice 35. The width of the three-dimensional image containing 75 slices is doubled for visibility.

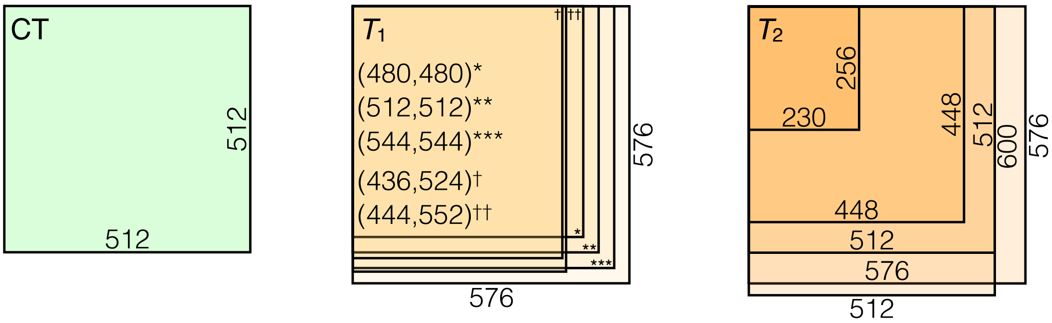


Figure 4.3: Illustration of the existing image frames and aspect ratios for the 25-patient data set. The frames of the CT scans as well as the T_1- and T_2-weighted MRI scans are shown with the accompanying numbers of pixels.

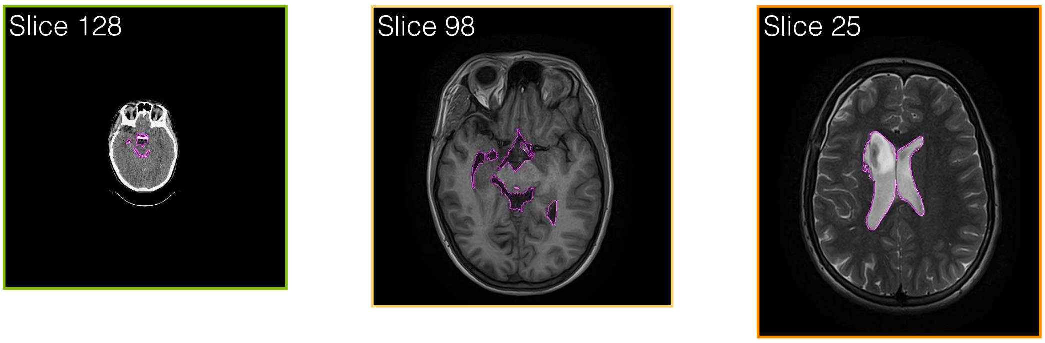


Figure 4.4: Contours of the ventricular system for Patient 22 of the 25-patient data set. Slices are shown for the CT (green), T_1-weighted (light orange) and T_2-weighted (orange) MRI scans with their contours (magenta).

slices and their large distance. For these scans, the head region of the patient is partially covered and the first slice starts abruptly in the brain. Usually, all scans were acquired in supine position, but six CT scans are available in prone position. The average head width of the patients is (123 ± 13) mm.

A special feature of the 25-patient data set is the availability of contours of the ventricular system (VS), which were outlined manually by a medical physicist and validated by a clinician at WPE. Therefore, the VS contours are ideally suited for validation in image processing. The contours are shown in Figure 4.4.

After the development of a preprocessing workflow with the smaller data set, the 25-patient data set is used to apply and test the workflow. Then, the preprocessed data are part of a comprehensive study in terms of image registration with deep learning. Afterwards, the registered image pairs of both data sets are merged with the fusion method, studied in Chapter 6.

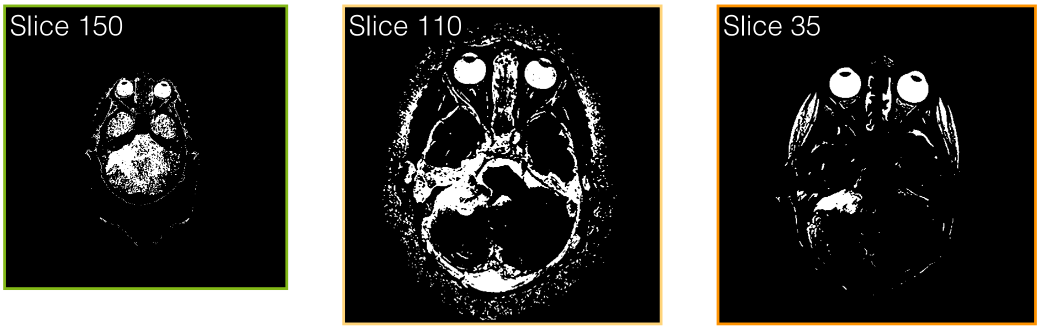


Figure 4.5: Preliminary segments for Patient 5 of the 14-patient data set. The segments (white) are generated with intensity-based threshold methods. Slices of the CT scan (green) as well as of the T_1-weighted (light orange) and T_2-weighted (orange) MRI scans are shown, including the eyes.

4.2 Segmentation

The process of image segmentation follows the simple principle of separating a feature in the image from the background, expressed by black pixels, $i_\mathrm{b} = 0$. This creates a binary image (see Section 3.1.1) in which the pixels of the segment represent the foreground with the same positive integer, i_f. In scans of medical imaging, several anatomical features are present with different intensities, e.g. eyes or fluids. This offers the option of generating a number of segments with individual labels for each feature of the image.

In the following, an algorithm, whose basis was created in a master's thesis [39] supervised by the author, is introduced for the generation of eye segments. These segments are produced automatically, in an unsupervised way, based on pixel values and their distributions. The eyes provide the advantage that intensity-based segmentation is easily possible due to their almost spherical shape and the sharp-edged border to neighbouring tissue (see Figure 4.2). For the 25-patient data set, the conversion from a VS contour to a segment is described afterwards. The VS is placed in the middle of the head, which increases the spatial coverage of the segment set.

4.2.1 Automated generation

The algorithm starts by performing plain segmentation on the three-dimensional image to separate fluid-like structures from other tissue in the head. For the separation, two thresholds for the pixel values are determined, which define the selected range. The methods, depending on the modality, are discussed in detail afterwards. The thresholds are applied to the image to produce a preliminary segment, which consists of several non-contiguous fragments with the pixel value $i_\mathrm{f} = 1$. In Figure 4.5, the preliminary segments of the CT and both MRI scans are shown for one patient.

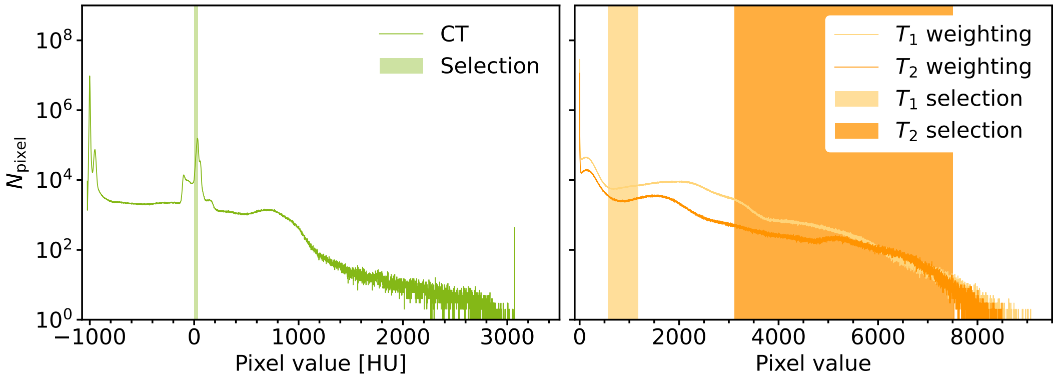


Figure 4.6: Distributions of the pixel values in logarithmic scale for Patient 5 of the 14-patient data set. The selected ranges are shown for the CT (left) and MRI (right) distributions. The position of the ranges differs between T_1-weighted (light orange) and T_2-weighted (orange) scans in the right plot.

CT scans The separation between fluids and other tissue is performed by means of the Hounsfield scale. Therefore, the distribution of the pixel values is investigated, as exemplified in Figure 4.6. The thresholds are determined in a specific search area, which is set from −30 HU to 40 HU. In this area, the steep slope coming from an increase in the number of pixels is the desirable range for the segmentation. The difference of adjacent pixel values, $i[n+1] - i[n]$, is calculated to determine sign changes, which indicate the start and the end of the slope in the search area. Then, the thresholds, t_1 and t_2, are detected by subtracting and adding 15 HU from the centre of the slope, respectively. If no thresholds are detected, the procedure is repeated with an extended search area until the centre of the slope is detected and the thresholds are determined. The outcome of such intensity-based segmentation is shown in Figure 4.5.

MRI scans The arbitrary scale of pixel values disturbs the separation of fluid-like structures with fixed pixel values, like for CT scans. However, the distributions of the pixel values are similar for every patient. Therefore, threshold computation with algorithms implemented in the `filters` module of the SCIKIT-IMAGE Python package [40] is used to determine the range of pixel values for the segmentation. The distributions for the T_1- and T_2-weighted scans as well as the selected regions are shown in Figure 4.6 for one patient. The lower threshold, t_1, for T_1-weighted scans is determined with the `threshold_li()` function, whereas the `threshold_isodata()` function is used to compute the upper threshold, t_2. Additionally, a median filter, which smooths the image, is applied to the preliminary segment. For T_2-weighted scans, the `threshold_yen()` function is applied to the images to get t_1 based on the distribution. For all T_2-weighted scans, t_2 is set at 7500. The results of the algorithm-based segmentation is shown in Figure 4.5.

After the plain segmentation, image processing tools from the OpenCV library [41] are used to separate the fragments of the preliminary segment, which is shown in Figure 4.7. This supports the identification of the eyes in the further process. Undesirable connections between the fragments can appear due to the loose intensity-based segmentation. For this, an opening operation (see Section 3.2.2) is applied to each slice of the segmented image with the `morphologyEx()` function. The kernel of this morphological function is generated with the `getStructuringElement()` function as a square of 5 × 5 pixels. The effect is presented in Figure 4.7a, where the noise around the eyes is reduced compared to the preliminary segments in Figure 4.5.

Subsequently, a selection is performed to reduce the amount of fragments and to obtain possible eye segments. The pixel spacing and the slice thickness are used to determine the volume of a single pixel. Then, the volumes of two spheres,

$$V_{\text{min}} = \frac{4\pi r_{\text{min}}^3}{3} f \quad \text{and} \quad V_{\text{max}} = \frac{4\pi r_{\text{max}}^3}{3} f\,, \tag{4.1}$$

with the radii $r_{\text{min}} = 7.5$ mm and $r_{\text{max}} = 12.5$ mm are calculated to take different sizes of the eyes into account. If the slice distance of the image is not equal to 1 mm, the shape of the eyes is rather elliptical than spherical due to the compression or dilation of the slices. The slice thickness is divided by the slice distance of the image to calculate the correction factor, f, which is used to scale both spheres along the z axis. The volumes are divided by the volume of a single pixel to define a range for the number of pixels. Then, the fragments are labelled with the multidimensional `label()` function from the SciPy package [42] to calculate the number of pixels of each fragment (see Figure 4.7b). The selection of possible eye segments is performed by sorting the fragments according to the number of pixels. A fragment is removed if the number is outside the defined range.

The algorithm continues with the identification of the eye segments. A spherical or elliptical template, depending on the slice properties, is generated for each of the remaining fragments. The template imitates the fragment for the case of a spherical or elliptical shape with the same volume. Then, the centres of mass of the fragment and the template are determined with the `ndimage.measurements.center_of_mass()` function from SciPy. The positions are used to overlay the template with the fragment. This enables the calculation of m_{DSC}, measuring the resulting overlap between the fragment and the template. In Figure 4.7c, the overlap of one eye segment and the corresponding template shows a high level of agreement. The overlap between a random fragment and its template is illustrated in Figure 4.7d, indicating less agreement. Besides the calculation of m_{DSC}, another identification criterion is provided by the location of the fragments. Both eyes have similar y and z coordinates, but they are located in different areas of the images regarding the x axis. After all fragments are overlaid with their template, the two fragments that are located in opposite areas and achieve the highest m_{DSC} values are identified as eye segments. The labels $l_{\text{LE}} = 10$ and $l_{\text{RE}} = 20$ are assigned to the left and right eyes, respectively.

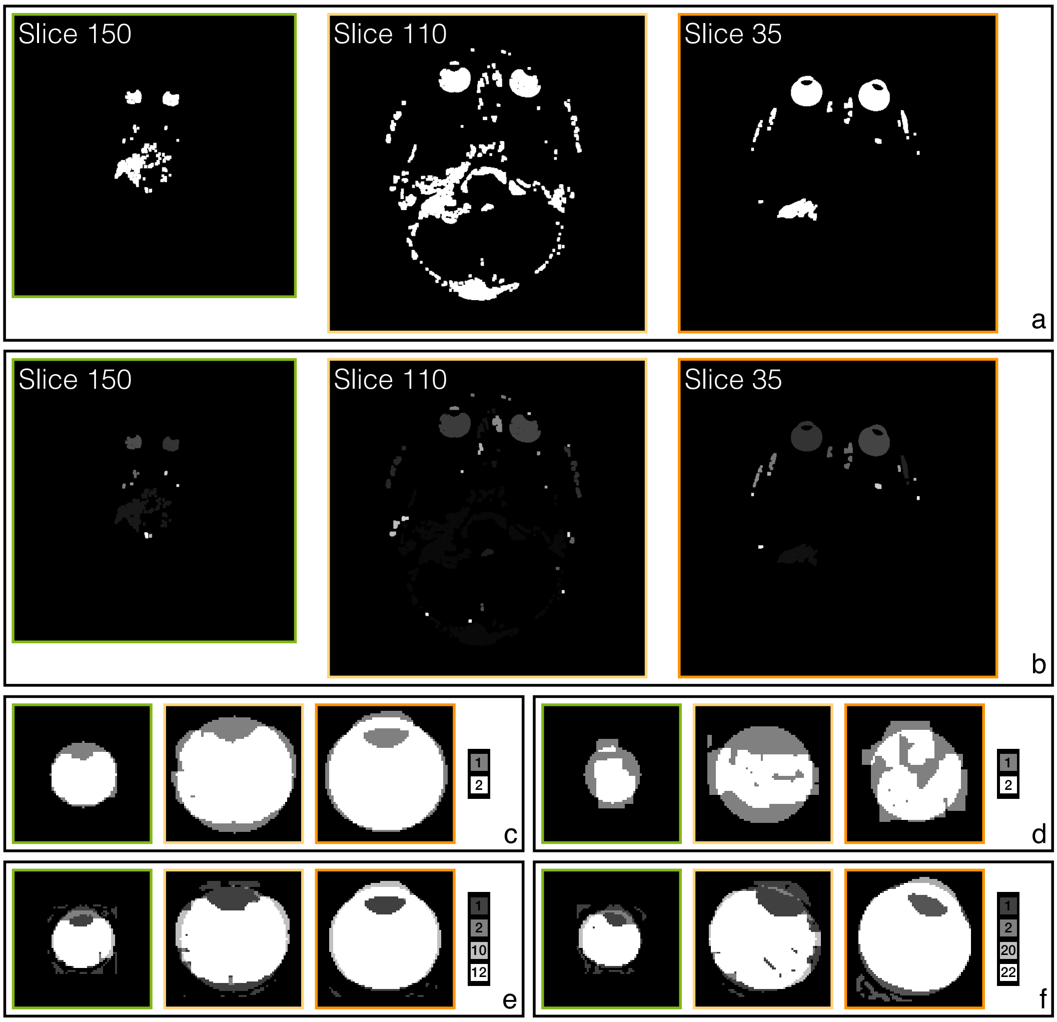


Figure 4.7: Visualisation of the algorithm steps for the eye segmentation of Patient 5 from the 14-patient data set. Slices are presented for the CT scan (green) as well as for the T_1-weighted (light orange) and T_2-weighted (orange) MRI scans. Subfigure a: The processing of the preliminary segment using the opening operation shows the separation of the eye segments from surrounding noise. Subfigure b: The amount of fragments is reduced by a size comparison, and the remnants are labelled (grey to white). Subfigure c: The comparison between an eye segment and its spherical or elliptical template shows many overlapping areas. Subfigure d: The overlap is decreased for a random fragment, which is not shaped like an eye. Subfigures e and f: The cleaning of the left (e) and right (f) eye segments is necessary to correct uncertainties on the surface. After adding the preliminary segment and the template, pixels with the values $i = \{2, 10, 12\}$ (e) and $i = \{2, 20, 22\}$ (f) are selected to be part of the eye segments.

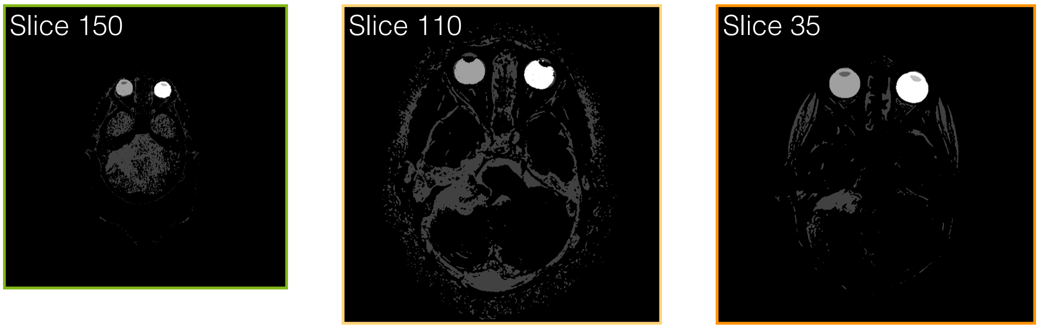


Figure 4.8: Eye segments for Patient 5 of the 14-patient data set. Both eye segments (white and grey) are generated automatically. The segmented slices of the CT scan (green) as well as the T_1-weighted (light orange) and T_2-weighted (orange) MRI scans are superimposed with the preliminary segments (dark grey) from Figure 4.5.

The last step is performed to clean the eye segments, especially the surface of the eyes. In some cases, the outermost pixels are removed during the process, e.g. after the opening operation. For this, the preliminary segment from the beginning is added to the identified eye segments, which again include the unintentionally removed pixels belonging to the eyes (see Figures 4.7e and 4.7f). Thus, the segmented image contains the pixel values $i = \{1, 10, 11, 20, 21\}$. Furthermore, two spheres with the diameters of the eye segments are generated with Equation (4.1) and added to the image. Besides the labels of both eye segments, the pixel values $i = \{2, 12, 22\}$ are required to be part of the eye segments with the labels l_{LE} and l_{RE}.

Ultimately, gaps in the eye segments are filled with the `fillPoly()` function from OpenCV, which requires contour points. These are computed with the `findContours()` function. The `CHAIN_APPROX_NONE` method is used in the `RETR_FLOODFILL` mode. This ensures that all points are included in the contour detection, while the detected contour is filled with the same value through the flooding algorithm. The segmented slices of the CT and both MRI scans are shown in Figure 4.8 for one patient.

4.2.2 Converted contours

The elaborate and time-consuming process of manual image contouring of anatomical features provides their spatial locations. The contours are superior to automatically generated segments. For the 25-patient data set, clinicians outlined the VS in the CT and MRI scans of each patient. The information on the contours is stored in separate DICOM files. Each file contains the location of the VS contours in Cartesian coordinates.

The contour information is retrieved from the file by searching under the specified keyword. The extracted data containing all indexed contours are stored in a list, which

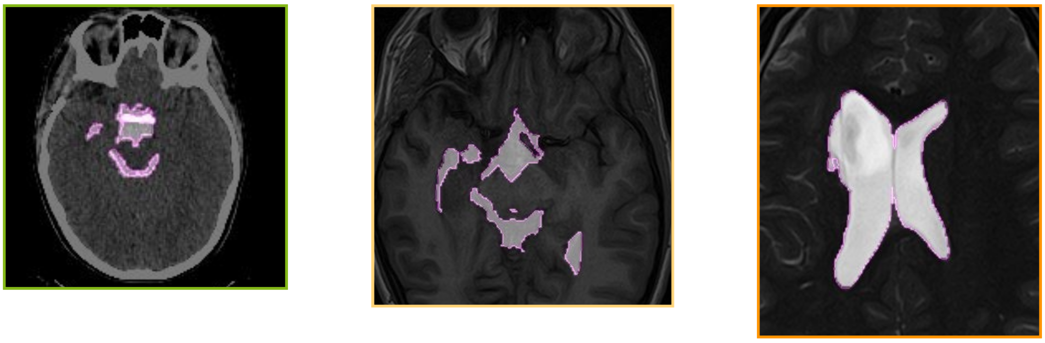


Figure 4.9: Visualisation of the contour conversion for the same slices as in Figure 4.4. The slices of the CT scan (green) as well as of the T_1-weighted (light orange) and T_2-weighted (orange) MRI scans are enlarged for visibility. The VS segments are indicated by the bright area (white), surrounded by the contours (magenta).

facilitates the access to the coordinates. Thus, the following algorithm, which is based on the DICOM-CONTOUR [43] library, is applied successively to all contours. First, the (x, y, z) coordinates of the point, n, are used to calculate the distance,

$$d = \sqrt{(x_{n+1} - x_n)^2 + (y_{n+1} - y_n)^2 + (z_{n+1} - z_n)^2}\,, \tag{4.2}$$

to the nearest point, $n + 1$. If the condition $d \geq 2\,\text{mm}$ is fulfilled, additional points are added to reduce gaps after the conversion from Cartesian to pixel coordinates. These points are computed by subdividing the space between adjacent points into pieces of at most 1 mm length. Next, the DICOM file of the respective slice is used to access slice properties, such as the pixel spacing, x_{spacing} and y_{spacing}, and the origin of the image, x_{origin} and y_{origin}. For the conversion to pixel coordinates, the information is included to calculate the pixel coordinates,

$$p = \begin{pmatrix}(x - x_{\text{origin}})/x_{\text{spacing}} & (y - y_{\text{origin}})/y_{\text{spacing}}\end{pmatrix}^{\top}\,. \tag{4.3}$$

Due to the rounding of the pixel coordinates, the possibility of producing a non-contiguous contour exists. Therefore, the closing operation (see Section 3.2.2) is applied to the contour. For this, the `morphologyEx()` function is used with a square element of 3 × 3 pixels, generated with the `getStructuringElement()` function. Lastly, the contour is filled with the label $l_{\text{VS}} = 30$, representing the ventricular system as a segment. The outcome is shown in Figure 4.9.

4.3 Adjustments

The main part of the preprocessing is the equalisation of the images of the data sets. The images must fulfil requirements to be provided for deep neural networks. These requirements refer to the image formats and the image alignment. The former are equalised to a specific aspect ratio with fixed pixel dimensions, while the latter is obtained with affine transformations and rigid registrations. Algorithms contributing to achieve the desired state of the data are introduced in the following.

4.3.1 Affine transformation

The acquired scans of a patient differ between the modalities, which is visualised in Figure 4.2. The number of slices and their distances as well as the pixel spacing produce different representations in terms of the size. Therefore, the dissimilarity is corrected by standardising the slice and pixel properties with affine transformations including scaling.

The aim of the scaling is to set the pixel and slice properties, such as the pixel spacing, the slice thickness and the slice distance, to 1 mm for all scans. For this, the linear transformation,

$$A_{\text{affine}} = \begin{pmatrix} s_x & 0 & 0 & t_x \\ 0 & s_y & 0 & t_y \\ 0 & 0 & s_z & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \tag{4.4}$$

contains the scaling, s, and translation, t, parameters for the x, y and z coordinates. The scaling factors, s_x and s_y, are provided by the pixel spacing of the image (see Figure A.1). To relocate the centre of the scaled image to the original position in the x–y plane, the translation parameters are defined as $t_x = N_x(1 - s_x)/2$ and $t_y = N_y(1 - s_y)/2$ with the numbers of pixels, N_x and N_y. The transformation also contains the scaling factor s_z, which is defined as the inverse of the slice distance, d_s, of the respective scan (see Figure A.2). There is no translation along the z axis.

The operation is performed with the `transform` module of SCIKIT-IMAGE. The `warp()` function is applied to the image with the inverted map, A^{-1}_{affine}, of the transformation in Equation (4.4). The output shape of the scaled image is defined as $(x, y, d_\text{s}z)$ to increase the number of slices especially for T_2-weighted MRI scans. Moreover, the constant mode, which fills the pixels outside the image with zero, is used. The pixel values of the scaled image are interpolated with the bi-cubic method. For the segmented image, the nearest-neighbour method is preferred to avoid polluting the image with values besides l_{LE}, l_{RE} and l_{VS}. In Figure 4.10, the effect of the affine transformation on the slice and pixel properties is shown for the T_2-weighted MRI scan. The number of slices increases (see Figure 4.10a), while the size in the axial plane (see Figure 4.10b) is reduced.

4.3.2 Format equalisation

The next algorithm relates to the image format, which varies a lot between the modalities as well as within the same modality. The dimensions are listed in Tables A.1 and A.2. To

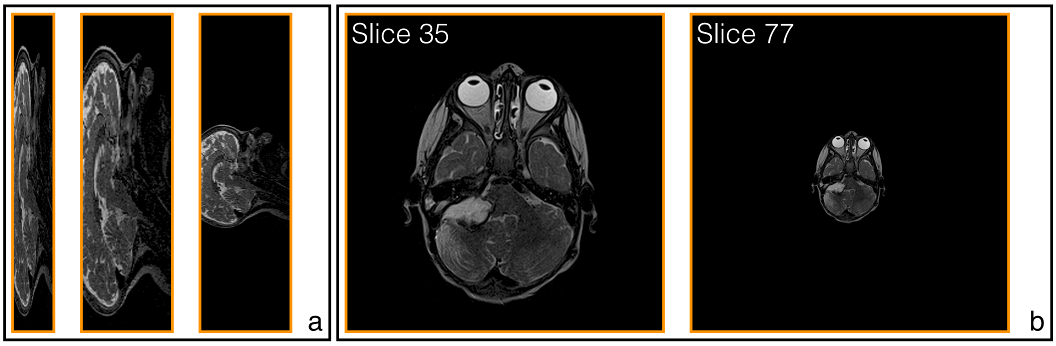


Figure 4.10: Visualisation of the affine transformation for the T_2-weighted MRI scan of Patient 5 from the 14-patient data set. Subfigure a: The perspective of the y–z plane is shown for $x = 288$. The scaling of the slice distance from 2.2 mm (left) to 1 mm (middle) increases the number of slices. In addition, the adjustment of the pixel spacing scales down the image (right). Subfigure b: Axial slices are shown before (left) and after (right) the application of the transformation. The slice position is changed from 35 to 77.

eliminate this inequality, the image is cut to a specific aspect ratio with fixed dimensions. The procedure includes the positioning of the images within the new frame.

First, the slices are limited to the head region. The CT and MRI scans contain many slices without anatomical information, which are placed above the head of the patient. Therefore, the number of non-zero pixels is counted for each slice. This allows the condition to be set that the minimum number of non-zero pixels is 0.2 % of the total number of pixels. For example, slices with 512 × 512 pixels need at least 525 pixels with values greater than zero. The first slice fulfilling the condition is used as the initial slice, while the slices above are removed. In addition, the lower part of the three-dimensional image often contains the shoulder of the patient. To preserve memory for further computations, this body part is also removed. For this, each slice of an image is scanned to detect the patient outlines from left to right and posterior to anterior. Such distributions are shown in Figure 4.11. If broad shoulders appear in the slices, the occurring intersection between both outlines is used as identification criterion for slice reduction.

Then, the images are cut to an aspect ratio of 3 : 4 with 192 × 256 pixels since this format perfectly suits the shape and the size of the head regarding pixel and slice properties of 1 mm. The process is illustrated in Figure 4.12. Temporary segmented images are created by applying threshold segmentation to the images. The bone window and the `threshold_mean()` function from SCIKIT-IMAGE are used for the CT and MRI scans, respectively. The median filter is additionally applied to reduce noise and to smooth the temporary image (see Figure 4.12a). In the next step, each temporary image is scanned to determine the rough proportion and location of the patient's head concerning the slice positions $N_x/2$ and $N_y/2$. The first and last non-zero pixels along the respective axis represent the outermost points to frame the patient's head. The position of the frame is

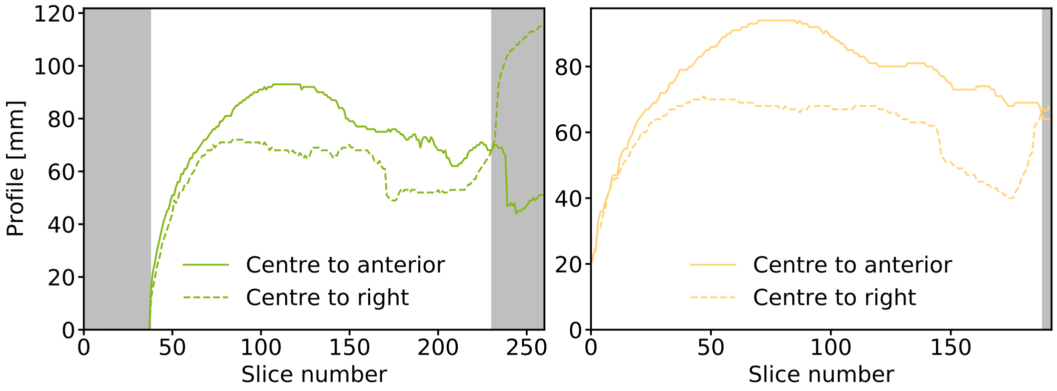


Figure 4.11: Distributions of the profile from two perspectives for Patient 5 of the 14-patient data set. The profiles are shown for the CT (left) and T_1-weighted MRI (right) scans. Features, like the ear or the nose, are visible, especially in the right plot. The grey areas represent the slices that are removed.

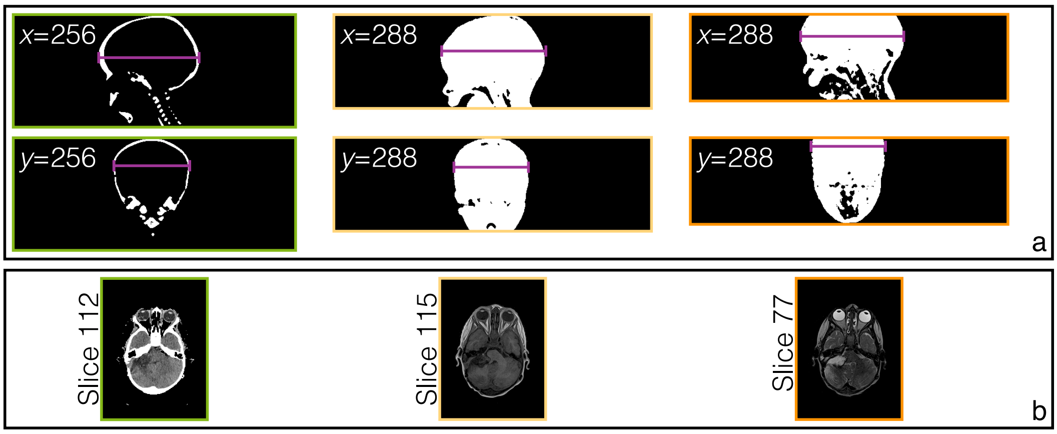


Figure 4.12: Effect of the adjustment of the image format on the scans of Patient 5 from the 14-patient data set. The results are presented for the CT scans (green) as well as the T_1-weighted (light orange) and T_2-weighted (orange) MRI scans. Subfigure a: The perspectives of the y–z (top) and x–z (bottom) planes are shown, including lines (magenta) to represent the largest distances of the outermost points. Subfigure b: The adjusted images of the scans have an aspect ratio of 3 : 4 with 192 × 256 pixels in the axial plane.

transferred to the image, which helps to adjust the format to the 3 : 4 shape. The outcome of this adjustment is presented in Figure 4.12b.

After the application of the rigid registration, described in the next section, the image format is further equalised by limiting the number of slices to 64. The slices are chosen to cover most of the head region. Therefore, the slice that contains the centre of mass of the eye segment with the label l_{LE} is determined to be the middle slice of the fully preprocessed image. The remaining slices are filled with 32 and 31 slices above and beneath the determined slice, respectively.

4.3.3 Rigid registration

The dissimilarity between the scans is mainly reduced due to the affine transformation and the format equalisation. However, the position and the slices still disagree between the scans. Rigid registration is additionally applied to increase the structural alignment between the source, S, and target, T, images. For this, translation and rotation operations are used as geometric transformations. These operations require coordinates of both S and T to calculate the displacements and angles. Thus, the centres of mass,

$$\vec{c}_{\mathrm{LE}} = \begin{pmatrix} x_{\mathrm{LE}} & y_{\mathrm{LE}} & z_{\mathrm{LE}} \end{pmatrix}^{\top} \quad \text{and} \quad \vec{c}_{\mathrm{RE}} = \begin{pmatrix} x_{\mathrm{RE}} & y_{\mathrm{RE}} & z_{\mathrm{RE}} \end{pmatrix}^{\top}, \tag{4.5}$$

of the eye segments are determined for both S and T. The segmented images with their labels l_{LE} and l_{RE} are used to obtain the respective pixel coordinates. These two points are sufficient to calculate the parameters of the translation operation. In contrast, at least three points are necessary to apply three-dimensional rotations to S. Therefore, the algorithm of the rigid registration is subdivided into two parts.

In the first part, the translation of the three-dimensional source image is performed with the linear transformation

$$A_{\text{translation}} = \begin{pmatrix} 1 & 0 & 0 & t_x \\ 0 & 1 & 0 & t_y \\ 0 & 0 & 1 & t_z \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{4.6}$$

containing the translation parameters, t_i, for each direction. They form the vector $\vec{t} = \vec{c}_{\mathrm{LE}}(T) - \vec{c}_{\mathrm{LE}}(S)$. The implementation uses the `warp()` function from the SCIKIT-IMAGE `transform` module. The interpolation methods are the same as for the affine transformation, described in Section 4.3.1. Since the translation along the z axis can move the source image out of the frame, the reshape mode is activated. This increases or decreases the number of slices depending on the direction of t_z in order to preserve the information. Afterwards, the differences in the number of slices between the target and source images are eliminated by adding zero-element slices to the image with the fewest slices. The outcome of such translation operation is shown in Figure 4.13 for the T_2-to-CT registration of one patient.

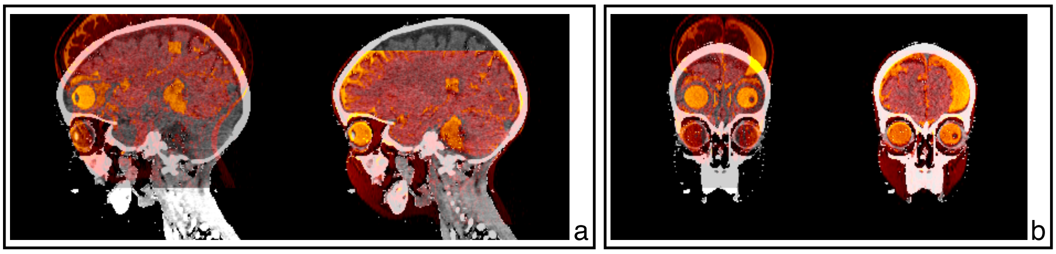


Figure 4.13: Effect of the translation on the T_2-weighted MRI scan of Patient 5 from the 14-patient data set. The overlay of the MRI (orange colours with transparency of 50 %) and CT (greyscale) scans is shown for two perspectives. Subfigure a: A slice of the y–z plane is presented to illustrate the translation with $t_z = 33$ mm and $t_y = 0$ mm. Subfigure b: The translation in the x direction is $t_x = 1.5$ mm, which is slightly visible in the x–z plane.

The second part includes the rotation of the images to increase image alignment. The centres of mass defined in Equation (4.5) are updated after the translation part. Then, the angles

$$\theta_{xy} = \tan^{-1}\left(\frac{y_{\mathrm{RE}} - y_{\mathrm{LE}}}{x_{\mathrm{RE}} - x_{\mathrm{LE}}}\right) \quad \text{and} \quad \theta_{xz} = \tan^{-1}\left(\frac{z_{\mathrm{RE}} - z_{\mathrm{LE}}}{x_{\mathrm{RE}} - x_{\mathrm{LE}}}\right) \tag{4.7}$$

are calculated to individually straighten out the source and target images based on the eyes. The operation is performed with the `ndimage.rotate()` function from SciPy. The common interpolation methods are used depending on the image type. Furthermore, the three-dimensional rotation is fixed to the centre $\vec{c}_{\mathrm{LE}}$. Consequently, the positions of the eyes are identical in the source and target images.

The acquisition of the MRI scans was often performed without a headrest, which decreases patient immobilisation. This leads to high discrepancies between the CT and MRI scans of up to 15° in the y–z plane. Due to the aforementioned lack of a third point for the calculation of θ_{yz}, an iterative process of several rotations in the y–z plane is developed to correct these discrepancies. For this, the Dice similarity coefficient, m_{DSC}, of the segments of S and T is used to determine the overlap after each rotation. For the scans of the 25-patient data set, the VS segments are used for the calculation of m_{DSC} besides the eye segments. Regarding the 14-patient data set, where no VS segments are available, a temporary segment of the entire head is generated with the `threshold_mean()` function from scikit-image. Rotations are applied to the source image in steps of 1°, ranging from −6° to 6°. If the highest m_{DSC} value is achieved for one of the boundary angles, the range of rotations is extended by six angles until the rotation angle with the highest m_{DSC} value is found. That angle is then applied to the source image, which finishes the rigid registration. In Figure 4.14, the steps that contribute to the rotational part of the rigid registration are presented.

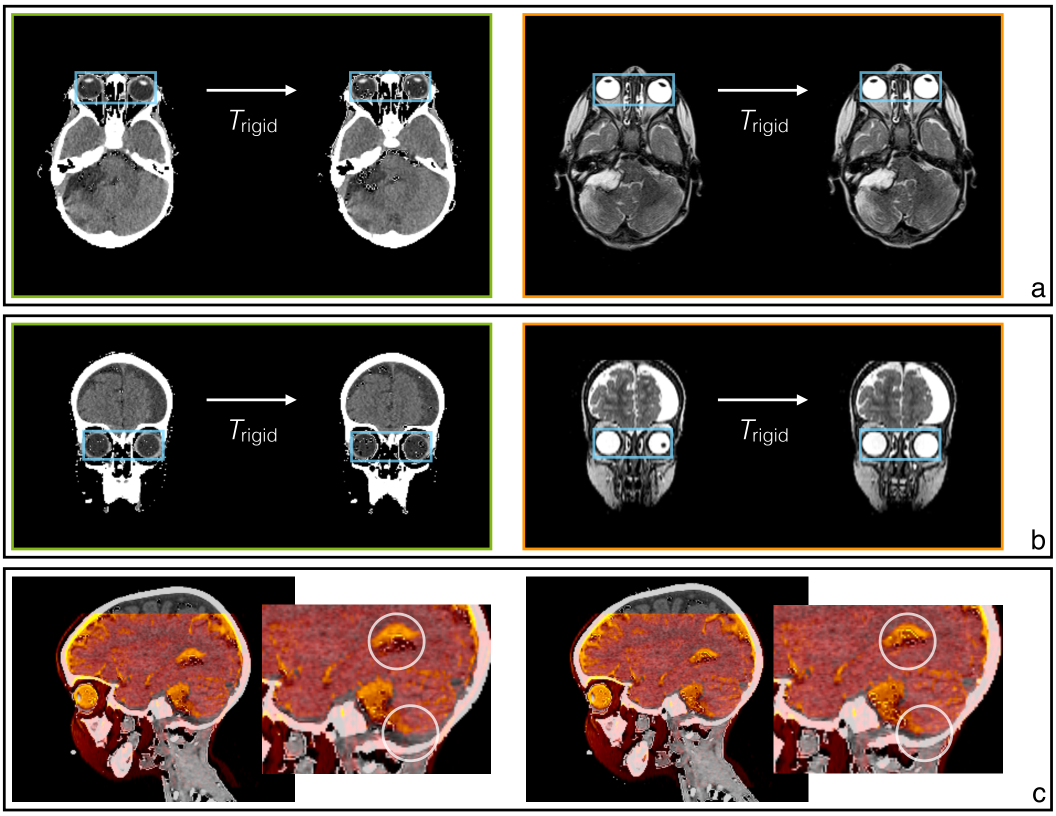


Figure 4.14: Effect of the rotation for the T_2-to-CT registration of Patient 5 from the 14-patient data set. The rectangles (blue) and the circles (white) are added to the images for comparison of non-rotated and rotated images. Subfigure a: The axial slices indicate the results of the rotation with $\theta_{xy}(\text{CT}) = 3.6°$ (left) and $\theta_{xy}(T_2) = 4.9°$ (right). Subfigure b: The perspectives of the x–z plane are presented to illustrate the rotation with $\theta_{xz}(\text{CT}) = -1.8°$ (left) and $\theta_{xz}(T_2) = -1.3°$ (right). Subfigure c: Images of the y–z plane are shown for the rotation $\theta_{yz} = 2°$. The overlap of the CT (greyscale) and MRI (orange colours with transparency of 50 %) slices shows that the image agreement between the non-rotated (left) and rotated (right) images increases, which is indicated by the white circles in the enlarged areas.

4.4 Results

The algorithms presented in Sections 4.2 and 4.3 are developed for the preprocessing of head CT and MRI scans. The process of standardising the data is aimed at fast and unsupervised application. Therefore, the algorithms are concatenated to form an efficient workflow. In the following, this workflow and its results of preprocessed data are evaluated in terms of image quality and image-similarity measures.

4.4.1 Preprocessing workflow

The workflow starts with the initial phase, i.e. collecting the data from the DICOM files of the scan. Since each slice of a scan is individually stored, all files provide access to the entire data. Information from the scan is used to generate a three-dimensional array representing the acquired image. The images do not necessarily have the same orientation, e.g. patients in supine or prone positions. Therefore, the slices are arranged to obtain uniformly orientated images. Besides the pixel values, slice properties—like the thickness, distance and location—and the pixel spacing are used for the image segmentation and adjustment. In the next step, the pixel values are manipulated to reduce the noise and to enhance the clarity in the images. The threshold of 0 HU is applied to the images of the CT scan, which means that pixels with lower values are added to the background. The threshold is chosen because of the position of the brain and bone window (see Section 2.1.3). An arbitrary pixel value of 600 is found to be best for the MRI scans of both data sets since they were acquired with the same scanner. These manipulated images are used for the deep-learning training process only.

Prior to the application of geometric transformations, the workflow continues with the segmentation phase. Segmented images containing eye segments with the labels l_{LE} and l_{RE} are generated for each scan. The algorithm described in Section 4.2.1 is applicable to any CT, T_1-weighted or T_2-weighted scan of the head as the automated generation is intensity based. It is performed on both data sets. The workflow also takes into account manually outlined anatomical features, which are then converted into segments with the algorithm introduced in Section 4.2.2. Besides the ventricular system, contours of other features can be used to generate segments and to extend this set. The reason to include image segmentation in the workflow is the possibility of measuring m_{DSC}, which quantifies the image alignment between two images. The segmented images, including either the eyes and the ventricular system or the eyes only, are essential for the image adjustments (see Section 4.3) in the further processes of the workflow. All transformations are applied to both the non-segmented and the segmented images.

The adjustment of the images includes the successive execution of the algorithms presented in Sections 4.3.1 and 4.3.2. Both algorithms are individually applied to each scan, while the rigid registration introduced in Section 4.3.3 requires an image pair with source and target images as input. Within the scope of the author's thesis, two mul-

Table 4.1: Run times of the first four workflow phases. The durations of the initial, segmentation, scaling and format phases are provided for the CT and both MRI scans of the 14-patient and 25-patient data sets.

Data set	Phase	CT	T_1 weighting	T_2 weighting
14-patient	Initial	(25.68 ± 2.52) s	(7.30 ± 0.89) s	(17.10 ± 2.47) s
	Segmentation	(49.05 ± 11.88) s	(40.84 ± 13.05) s	(13.91 ± 4.87) s
	Scaling	(37.18 ± 5.86) s	(23.90 ± 5.61) s	(26.87 ± 6.85) s
	Format	(10.40 ± 1.14) s	(9.53 ± 3.87) s	(7.63 ± 1.59) s
25-patient	Initial	(37.43 ± 7.20) s	(15.74 ± 5.75) s	(10.47 ± 5.30) s
	Segmentation	(56.99 ± 9.87) s	(40.51 ± 12.59) s	(15.03 ± 5.61) s
	Scaling	(37.73 ± 6.15) s	(21.09 ± 2.42) s	(22.51 ± 6.91) s
	Format	(11.85 ± 1.95) s	(8.21 ± 0.95) s	(6.91 ± 2.11) s

timodal registrations, T_2-to-CT and T_1-to-CT, and one unimodal registration, T_1-to-T_2, are investigated. At the end of the rigid-registration phase, the image type is changed to 8-bit integers with 256 greyscale values to improve similarity between the scans. This data type is found to be optimal for the deep neural network (see Section 5.2.1). The image adjustment is the main part of the workflow, which prepares an unimodal or multimodal image pair for the deep-learning-based DIR by scaling, rotating and positioning.

4.4.2 Quantitative evaluation

In the following, the average run time of each step of the preprocessing workflow is investigated for both data sets. Afterwards, the accuracy of the rigid registration, which is calculated with m_{DSC} (see Section 3.3.3), is presented to measure the alignment between source and target images.

The computation of the preprocessed data is performed with AMD EPYC 7742 CPUs. The initial, segmentation, scaling and format phases of the workflow are applied simultaneously to the CT, T_1- and T_2-weighted MRI scans. The duration of these phases is listed in Table 4.1. The run times of the 14- and 25-patient data sets are similar. The longest processing is required by the CT scans of the 25-patient data set with in total 144 s. In contrast, the T_2-weighted MRI scans of the same data set need the shortest time, 55 s on average.

The run times of the rigid-registration phase are provided in Table 4.2. For this phase, three different variants of source-to-target registrations are possible. The multimodal applications take longer than the unimodal rigid registration because the rotation in the y–z plane, described in Section 4.3.3, is skipped for the T_1-to-T_2 registration. The average run time for multimodal rigid registration is 81 s, while the unimodal variant

Table 4.2: Run times of the last workflow phase. This phase handles rigid registrations and requires a source-to-target image pair. Two multimodal combinations, T_2-to-CT and T_1-to-CT, and one unimodal combination, T_1-to-T_2, are investigated. The durations of each registration direction are given for both data sets.

Data set	Phase	T_2-to-CT	T_1-to-CT	T_1-to-T_2
14-patient	Rigid	(88.02 ± 13.03) s	(83.80 ± 11.21) s	(56.83 ± 12.12) s
25-patient	Rigid	(73.86 ± 7.24) s	(78.56 ± 7.43) s	(51.87 ± 6.47) s

takes 54 s. Consequently, preprocessed images of one patient are produced in less than four minutes, including image segmentation and adjustment of three different image types.

The accuracy of the preprocessing is measured with the help of segmented images, which are generated in the workflow. For each segment, the overlap between two images is calculated with m_{DSC}. As the same image format is required for the computation, the value of m_{DSC} is calculated before the application of the rigid registration (pre-rigid) and afterwards (post-rigid). The results for both data sets are presented in Figure 4.15. An increase in image alignment is directly apparent for the preprocessed images due to the mean values, which are closer to unity in the post-rigid case. The post-rigid values are either higher than or equal to the pre-rigid values. This results from the construction of the rigid-registration phase, which aims at improving image agreement. The application of the translation or rotation transformations depends on the m_{DSC} values. The respective operation is not applied to the image if lower values are achieved compared to the previous numbers. The third patient of the 14-patient data set, for example, has the same pre-rigid and post-rigid values. A difference between both data sets is present since the improvement of image agreement is higher for the 25-patient data set. For the T_2-to-CT registration, the mean values for the 14- and 25-patient data sets increase from 0.24 to 0.81 and from 0.12 to 0.82, respectively. Furthermore, the results of the unimodal registration indicate that the agreement between T_1- and T_2-weighted MRI scans is already high before the rigid registration, especially for the 14-patient data set with $m_{\text{DSC}}(\text{pre-rigid}) = 0.51$. Ultimately, the proposed workflow produces preprocessed CT and MRI scans of the patients with the same format and similar alignment. The overlap is improved by 52 % on average.

4.4.3 Qualitative evaluation

In contrast to quantitative evaluation, the quality of the preprocessed images is investigated by image overlay displays [4], which facilitates the comparison of the image features. For this, the registered image is overlaid with the target image using the IMAGEJ programme [26]. The image overlay displays of one patient are shown in Figure 4.16.

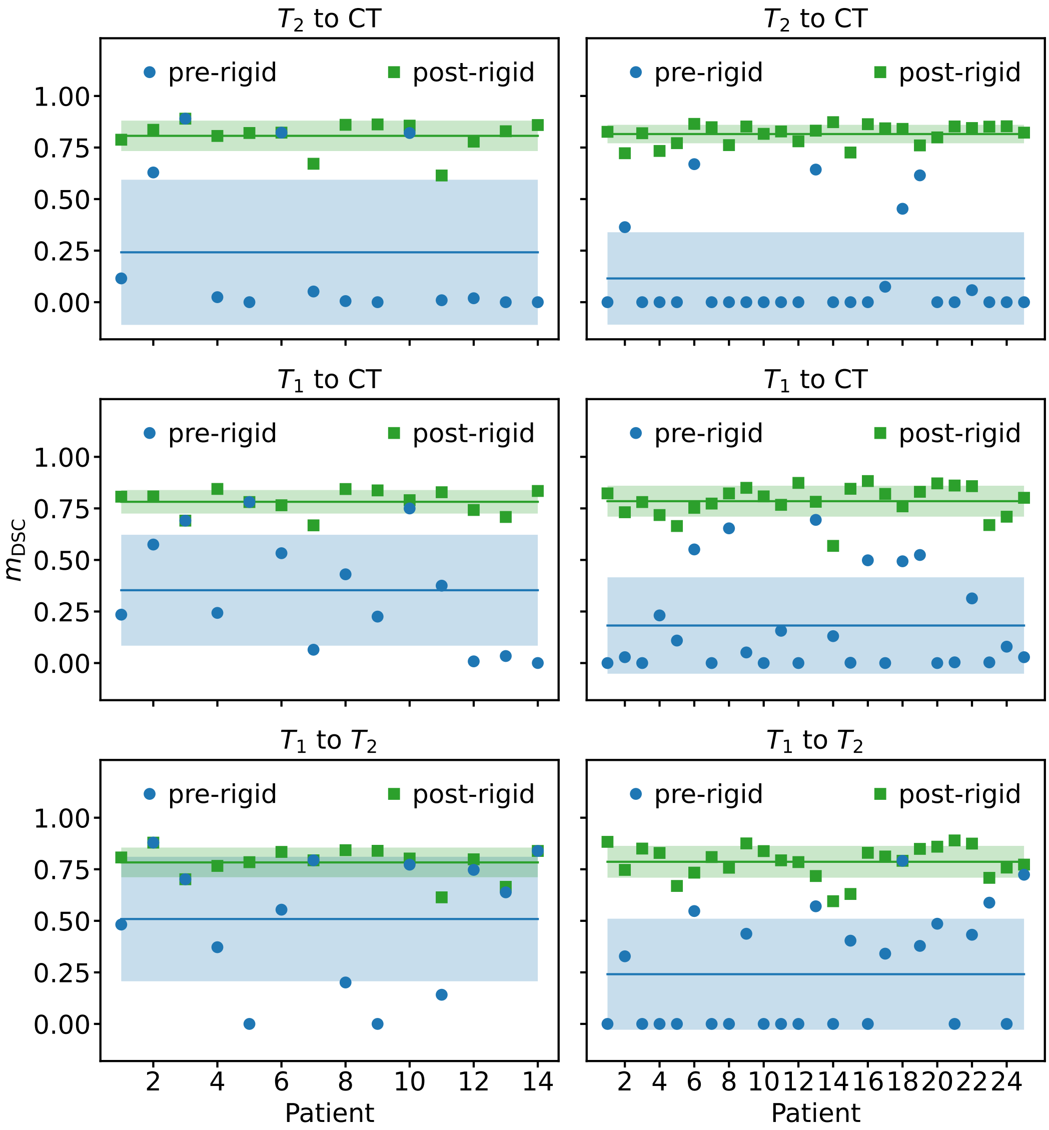


Figure 4.15: Accuracy measurement with m_{DSC} for the 14-patient (left) and 25-patient (right) data sets. The calculation is performed on the segmented images before (blue) and after (green) the application of the rigid registration. The results are depicted for the T_2-to-CT (top), T_1-to-CT (middle) and T_1-to-T_2 (bottom) registrations. In addition, the mean values (solid lines) and their uncertainties (bands) are shown.

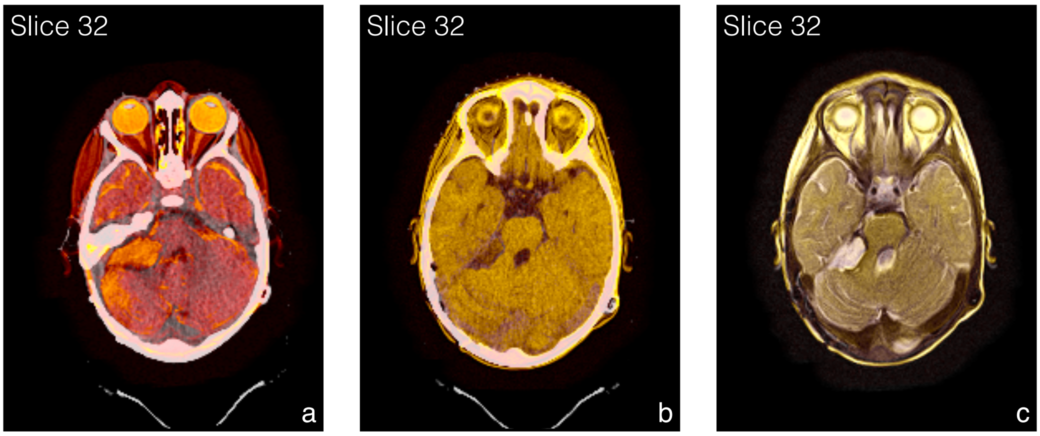


Figure 4.16: Image overlay displays of preprocessed images of Patient 5 of the 14-patient data set. The target (greyscale) and registered source (colours with transparency of 50 %) images are shown. Subfigure a: The result of the T_2-to-CT registration is presented with the T_2-weighted MRI slice in an orange-red colour scale. Subfigure b: For the T_1-weighted MRI slice, a yellow-orange colour scale is used to indicate the impact of the T_1-to-CT registration. Subfigure c: The unimodal registration of T_1-weighted MRI scans (yellow-orange colours) is shown with the T_2-weighted MRI scan as the target image.

In Figure 4.16a, the image overlay represents the result of the T_2-to-CT registration. The MRI features are illustrated in an orange-red colour scale, providing visual distinctness to the CT features in greyscale. The visibility of bones is restricted for MRI, but the soft tissue indicates the image agreement with the high contrast of bones, represented in the CT scan. The improved alignment between both images is especially apparent for the eyes. The image overlay display for the T_1-to-CT registration is shown in Figure 4.16b. A yellow-orange colour scale is used for the preprocessed T_1-weighted MRI scan. Similar to the T_2 weighting, soft tissue indicates the increase in image agreement with the CT scan. The unimodal registration of the T_1-weighted MRI scan is presented in Figure 4.16c with the same colour scale. The acquisition of the two MRI weightings was performed successively on the same scanner, which implies high image agreement before the application of rigid registrations. However, an increase is still achievable, as quantified in Figure 4.15.

In general, the image overlays visualise the extent of the preprocessing, which adjusts the images to match in size and position. High improvement is obtained for all registration combinations, but small non-rigid displacements are not taken into account. Corrections of such effects are studied with DIR using deep learning in Chapter 5. Since simple image overlays provide low variability for visual evaluation, image fusion is investigated in detail in Chapter 6.

5 Deformable image registration

Image preprocessing is specifically designed for image-format equalisation and image positioning by using rigid registration. This type of registration disregards the mobility of organs or the distortion of images. The former is more distinctive for images of the lung than for the brain, but shrinking tumours from radiotherapy still influence the brain morphology. Distortion effects, which are supposed to originate from the magnetic fields [1, 4], appear in particular for the MRI scans. These challenges require individual displacements of each pixel in addition to the rigid registration. In this chapter, deformable image registration is investigated with a deep neural network. In Section 5.1, the algorithm and the implementation of the network are introduced. Several studies, described in Section 5.2, are performed to optimise the setting of the network for multimodal DIR, which includes input configurations and parameter tuning. Finally, a multimodal model generated with data augmentation for the optimal network setting is presented in Section 5.3.

5.1 Deep neural network

In general, deep-learning models are perceptrons with multiple connected layers, which consist of nodes. These imitate neurons in the brain by carrying information to subsequent layers [6]. This type of neural network, illustrated in Figure 5.1, is an extension of machine learning, and methods for image deformation or recognition regarding medical image analysis have evolved in recent years [7].

The models are constructed from input, inner and output layers. The input, expressed by images, vectors or distributions, depends on the application, e.g. speech recognition or image processing [44]. Supervised methods require additional ground truth data to generate the output. For the case that ground truth data, e.g. simulated samples, are not available, unsupervised training is performed with the input only. [32] The main computation takes place in the inner layers, l, where the input is processed to obtain features, $\vec{f}(l)$, with mathematical operations. The depth of the network is given by the number of inner layers, n_l, and their individual features, f_i. The calculation of the features is performed with a linear operation, containing a linear transformation matrix, $W(l)$, and a displacement vector, $\vec{b}(l)$. The matrix includes weight parameters, W_{ij}, and the vector comprises bias components, b_{ij}. The tail end of the network is represented by the output layer. The output vector, $\vec{o}$, which is obtained by a final linear operation with the transformation W_o and the displacement $\vec{b}_o$, contains the predictions to solve

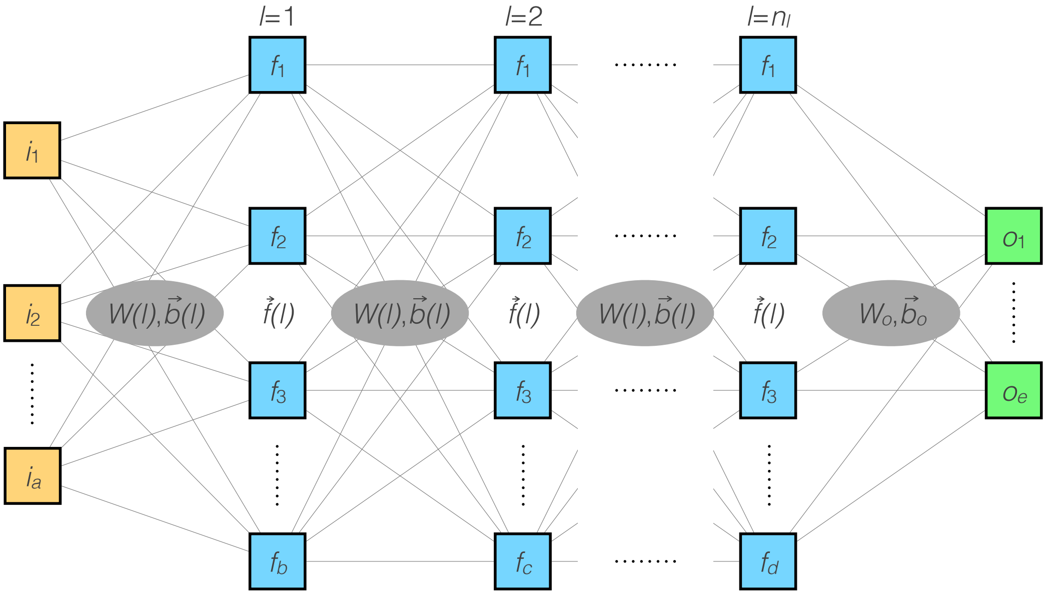


Figure 5.1: Illustration of the general structure of a deep neural network. The inputs (orange) are processed through the inner layers (blue), which predict the output (green). The processing of the input vector, $\vec{\imath}$, is performed with linear transformations, $W(l)$, and bias displacements, $\vec{b}(l)$, in each layer, l, to extract the features, $\vec{f}(l)$. The depth is defined by the number of layers, n_l, and the number of features (b, c and d). A final operation with W_o and $\vec{b}_o$ concludes the network for the prediction of the output, $\vec{o}$.

the particular task. The first layer, $l = 1$, serves as an example for the calculation of the features, $\vec{f}(1)$. Let the layer consist of b features, which result from the input vector, $\vec{\imath}$, with a samples. Then, the operation is

$$\vec{f}(1) = W(1) \cdot \vec{\imath} + \vec{b}(1) \Leftrightarrow \begin{pmatrix} f_1 \\ f_2 \\ f_3 \\ \vdots \\ f_b \end{pmatrix} = \begin{pmatrix} W_{11} & \cdots & W_{1a} \\ W_{21} & \cdots & W_{2a} \\ W_{31} & \cdots & W_{3a} \\ \vdots & \ddots & \vdots \\ W_{b1} & \cdots & W_{ba} \end{pmatrix} \cdot \begin{pmatrix} i_1 \\ \vdots \\ i_a \end{pmatrix} + \begin{pmatrix} b_{11} \\ b_{12} \\ b_{13} \\ \vdots \\ b_{1b} \end{pmatrix}. \tag{5.1}$$

The features of subsequent layers are generated likewise to connect all nodes. Consequently, all features contribute to the prediction of the network output vector,

$$\vec{o} = W(o) \cdot \Big[W(l) \cdot \cdots \cdot \Big(W(2) \cdot \Big[W(1) \cdot \vec{\imath} + \vec{b}(1) \Big] + \vec{b}(2) \Big) + \cdots + \vec{b}(n_l) \Big] + \vec{b}(o)\,, \tag{5.2}$$

which is determined by concatenating all linear mappings.

Since a neural network based on linear operations is not able to solve complex tasks, activation functions, e.g. hyperbolic functions or rectified linear functions [45], are additionally applied at each node to break the linearity of Equation (5.1). This suppresses irrelevant information, while important features responding to the activation contribute more to the output. The networks can be constructed with many layers and features, which increases the number of parameters in $W(l)$ and $\vec{b}(l)$. The general task of the network is to adapt these parameters in the training process with the stochastic-gradient-descent method, which aims at minimising an objective function, $\mathscr{L}$. Backpropagation is used to determine the partial derivatives $\partial\mathscr{L}/\partial W$ and $\partial\mathscr{L}/\partial b$ consecutively for each layer in the output-to-input direction. Deep neural networks with such a structure are capable of optimising their parameters to solve a particular non-linear problem after many iterations. [32]

In this thesis, a CNN is employed to solve the task of image registration of two images. This class of deep neural networks uses a series of convolution operations (see Section 3.2.3) in each layer to extract image features, responding to the respective convolution kernel [6]. The method allows image deformations to be computed on the basis of the features and their positions in the images. Several types of CNNs were developed for image detection, segmentation or registration purposes [46]. One popular way to construct a CNN is U-Net [16] because of its unique architecture of contraction and expansion paths [6]. These characteristics are efficient for small data sets [16]. Consequently, a U-shaped CNN is investigated and optimised to be included in the application-related registration workflow.

5.1.1 Algorithm

The CNN formed according to U-Net breaks with the conventional arrangement of the inner layers to reach the output layer. The U-shape splits the network into two symmetric paths for encoding and decoding. Both paths are subdivided into several layers, containing various operations to contribute to the goal of the respective path. Its general structure is presented in Figure 5.2a.

Input At least one pair of normalised images is required for the operation of the CNN. As the task of the network is the alignment of both images, these have to be classified before entering the inner layers. Therefore, one image is determined as the source image, S, while the other is set as the target image, T. The network is not restricted to a specific dimension and can perform two-dimensional or three-dimensional image registration. Furthermore, the input can be divided into batches if it contains several image pairs. The images within a batch are then concatenated to form a group with a specific batch size, b, which runs jointly through the network. This increases the possibility to generalise the model [32].

Inner layers The encoder path is especially designed to extract image features of both S and T by applying convolutions (see Section 3.2.3) in each layer. The basis of

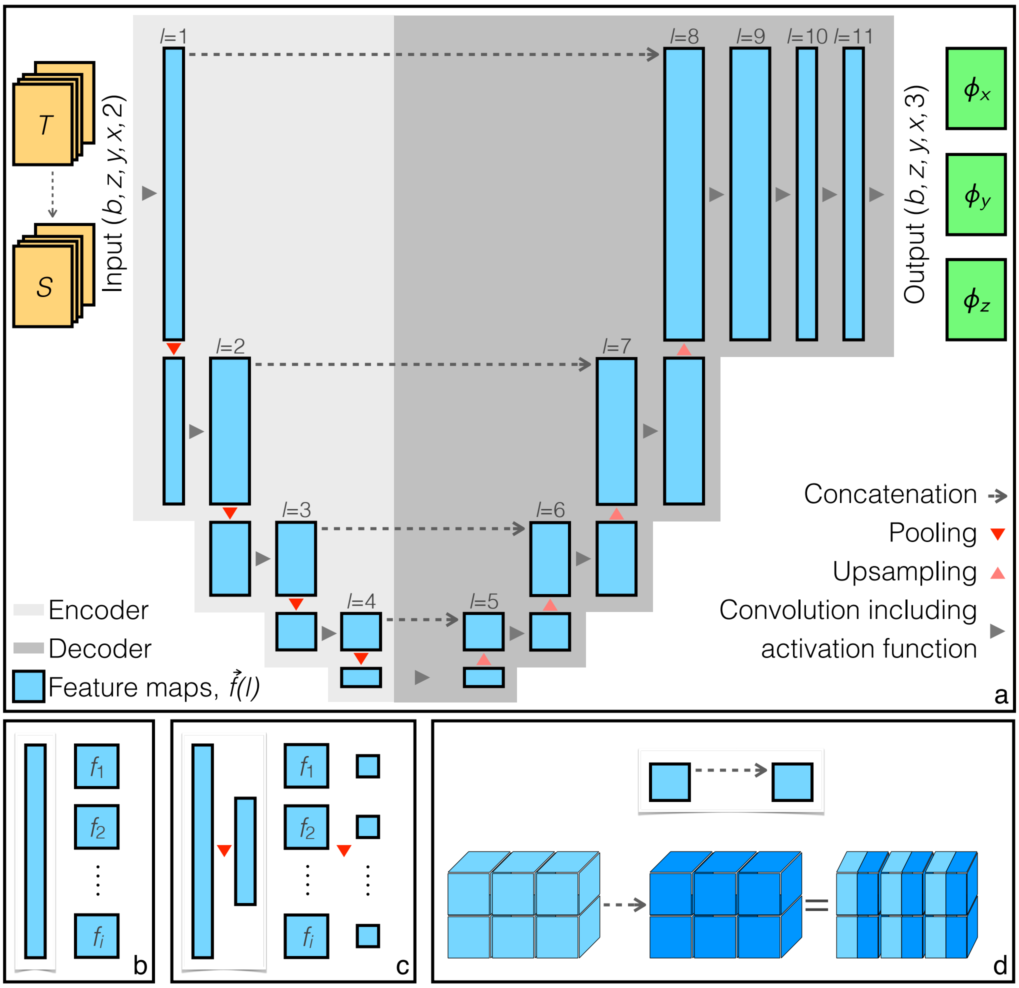


Figure 5.2: Illustration of the CNN algorithm with the U-Net structure. Subfigure a: The input with the shape $(b, z, y, x, 2)$ consists of source, S, and target, T, images, which are concatenated in the last axis. For the batch configuration, b image pairs are stored in the first axis of the input. The first four inner layers, $l = 1$ to $l = 4$, are part of the encoder path, where convolution and pooling operations are applied successively. The convolutions including an activation function produce the feature maps $\vec{f}(1)$ to $\vec{f}(4)$, while the pooling manipulates their shape afterwards. The decoder path, built of seven layers, $l = 5$ to $l = 11$, restores the shape of the feature maps $\vec{f}(5)$ to $\vec{f}(11)$ after the application of the convolution. Finally, the last convolution operation generates the network output in the form of displacement vector fields, ϕ_i. Subfigure b: The number of resulting feature maps, f_1 to f_i, has to be set before the training for each layer. Subfigure c: The pooling operation halves the size of the feature maps. Subfigure d: The concatenation method applied to the feature maps of the encoder and decoder paths is visualised.

the convolutions is the kernel weights, which are initially passed across the input images in the first level to produce feature maps. Then, these feature maps are further convoluted in the subsequent levels. The transition between two layers of the encoder path is occupied by the max-pooling operation (see Section 3.2.3) to reduce the size of the feature maps. Traversing the encoder path yields the features from the input images, but the information on their location is lost. The decoder path is constructed to regain the information on the location of the image features. For this, the original size of the feature maps from the first level has to be restored, which is performed in each level with the upsampling method (see Section 3.2.3). Furthermore, a connection between the enlarged feature maps and the feature maps of the respective encoder level is made through concatenation. Convolutions are then applied to combine the concatenated information into feature maps. After each convolution, the leaky rectified linear function [45] is applied to activate units of the feature maps below zero with a small gradient.

Output Eventually, the output is generated with final convolutions of the last feature map $\vec{f}(11)$ with three kernels. This network is designed to generate deformation vector fields, ϕ_i, predicting the spatial transformation of S. Each pixel of the two-dimensional or three-dimensional source image gets a unique displacement for each direction. The application of the deformation vector fields (see Section 3.2.1) is performed with target-to-source mapping to generate the deformed image, D. The linear interpolation is used to compute the values of D for each pixel. In addition, the respective segmented image is deformed with the same displacements, but the nearest-neighbour method is applied to maintain the labels of the segments .

Network training In general, a neural network with the aim of solving a particular task must be trained to learn and improve its parameters. This means that several iterations are required for the optimisation process, which uses an objective function to control the quality of the output after each iteration [32]. Here, the loss function

$$\mathscr{L}(T, D, \phi) = \mathscr{L}_{\text{sim}}(T, D) + \lambda\, \mathscr{L}_{\text{reg}}(\phi) \tag{5.3}$$

is set as the objective function to be minimised. The loss function is subdivided into a similarity term, $\mathscr{L}_{\text{sim}}(T, D)$, and a regularisation term, $\mathscr{L}_{\text{reg}}(\phi)$. The former consists of a metric that quantifies the degree of similarity between T and D. The latter, which is regulated with the parameter λ, quantifies the smoothness of ϕ by calculating the differences of the displacements of adjacent pixels. During the network training, the minimisation of $\mathscr{L}$ is achieved with the stochastic-gradient-descent method, which is implemented with a specific learning strategy, including a customisable learning rate, α.

5.1.2 Implementation

The algorithm described in Section 5.1.1 is based on VoxelMorph [17, 47], which was developed for fast DIR of brain MRI scans. In this thesis, the deep neural network is employed to investigate the advantages, like fast and direct application, for multimodal use of three-dimensional head CT and MRI scans. Thus, the input of the CNN is a pair of three-dimensional images, equalised with the preprocessing workflow (see Section 4.4.1). The operations are performed with the TensorFlow [48] machine-learning platform (`tf`), which is supported by the Keras [49] deep-learning interface.

Regarding the computations in the inner layers of the CNN, operations are implemented with the `tf.keras.layers` module. The convolutions are applied with the `Conv3D()` function using a kernel size of 3 × 3 × 3 pixels with a regular stride of one and the `same` option for padding. The number of resulting feature maps has to be specified (see Figure 5.2b). In addition, the `he_normal` method [50] is set for the weight initialisation of the kernels. As activation, the `LeakyReLU()` function [45] is used with the gradient 0.2. The `MaxPool3D()` function contributes to the contraction in the encoder path by halving the size of the feature maps with a window size of 2 × 2 × 2 pixels and no padding (see Figure 5.2c). A window of the same size is chosen for the `UpSampling3D()` function in the decoder path to expand the shape of the feature maps from (b, z, y, x, n) to $(b, 2z, 2y, 2x, n)$. The doubling is followed by the application of the `Concatenate()` function, which, for example, connects $\vec{f}(5)$ with $\vec{f}(4)$ in their last axis to obtain a feature map with the shape $(b, z, y, x, n_{\vec{f}(5)} + n_{\vec{f}(4)})$ (see Figure 5.2d). The final convolution operation for $l = 11$ uses the `RandomNormal` method [51] for the kernel-initialiser option, generating normal-distributed kernels with the mean 0 and the standard deviation 10^{-5}. That step produces the output with the shape $(b, z, y, x, 3)$. The number of layers and their resulting feature maps are customisable. The default configuration of the encoder path includes four levels, which produces 16, 32, 32 and 32 nodes. The number of layers in the decoder path is seven by default. The resulting feature maps $\vec{f}(5)$ to $\vec{f}(11)$ contain 32, 32, 32, 32, 32, 16 and 16 nodes. Hence, this setting contains 327 331 trainable weight and bias parameters. Exemplary feature maps generated in the encoder and decoder paths are presented in Figure 5.3.

The output of the CNN is further processed with the source image to compute D with the `map_fn()` function, which performs pixel-by-pixel displacements of S with the mapping from ϕ. Afterwards, the loss function in Equation (5.3) is calculated. Here, the normalised cross-correlation described in Section 3.3 is chosen to be part of the similarity term

$$\mathscr{L}_{\text{sim}}(T, D) = (1 - m_{\text{NCC}}(T, D)) \, . \tag{5.4}$$

As the metric quantifies high image similarity for values near unity, the subtraction is necessary to meet the target of minimising the loss function. The implementation of $m_{\text{NCC}}(T, D)$ is done with the `tf.nn` module, which contains primitive neural-network operations, like the `conv3d()` function. This function is used to calculate the metric

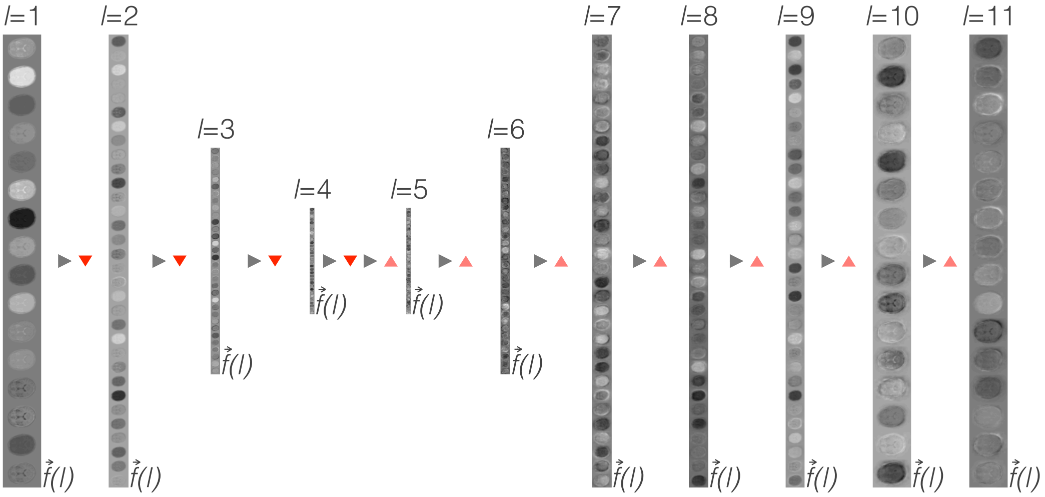


Figure 5.3: Exemplary slices of the feature maps for the layers $l = 1$ to $l = 11$ of the CNN, which is shown in Figure 5.2. The operations for convolution (grey triangle), pooling (red triangle) and upsampling (pink triangle) affect the appearance and the size of the feature maps $\vec{f}(l)$.

for T and D with a $9 \times 9 \times 9$ filter in which all elements are unity. The stride number is one, while the `same` option is used for the padding. The L2 norm [32] is implemented as regularisation by computing the gradients of ϕ, which are approximated through the differences of the displacements [47]. The regularisation term is defined as

$$\mathscr{L}_{\text{reg}}(\phi) = \frac{1}{3} \sum_{i \in \{x,y,z\}} \frac{1}{n_p} \sum_{p} \left(\phi_i(p_i + 1) - \phi_i(p_i)\right)^2 . \tag{5.5}$$

The differences are determined for $p-1$ pixels, starting with the second components. The means over all pixels and spatial directions yield the penalty value, which is multiplied with λ. The result is added to $\mathscr{L}(T, D, \phi)$.

For the optimisation process during the network training, a learning strategy from the `tf.keras.optimizers` module can be implemented in the CNN. The `Adam()` function based on the Adam algorithm [52] is set by default. This algorithm takes previous steps of the optimisation into account and determines the direction for the trainable parameters in the next iteration [32]. In addition, the algorithm includes adaptive learning rates during the optimisation. The default configuration of the CNN is $\alpha = 10^{-4}$.

When the training of a registration model is finished after the specified number of iterations, the model with the lowest value of the loss function is set as the best run. This model can be applied to any preprocessed image that is similar to the source image. For the generation of the deformed image, the same implementation is used as for the network training.

5.2 Parameter tuning

The main task of dealing with deep neural networks is to find the optimal configuration of the parameters. As described in Section 5.1, the CNN contains a customisable architecture as well as variable functions, e.g. the loss and optimiser functions. These are viable options for the parameter tuning, which is split into two studies. In Section 5.2.1, effects on the CNN performance are investigated regarding the image type of the inputs. In addition, different optimiser functions are tested, and the impact of a dropout rate is examined. An extensive study with variants of the CNN architecture is presented in Section 5.2.2. Within the scope of this thesis, the parameter tuning is performed on multimodal data sets with CT and T_2-weighted MRI scans. The T_2 weighting is chosen as the source image because of its higher soft-tissue distinctness compared to the T_1 weighting. Results of unimodal DIR with T_1- and T_2-weighted MRI scans are presented later in this chapter (see Section 5.3.2).

5.2.1 Preliminary studies

The minor investigations are carried out on the 14-patient data set with a division into training and validation data of approximately 80 % to 20 %. This means that eleven image pairs are used for the network training. The trained models are additionally applied to the remaining three image pairs to validate the results. The assignment of the patients to the data subsets is done randomly and results in Patient 4, 12 and 13 as the validation data. The number of iterations is set to 200 for the network training of each registration model in the following investigations.

The training performance and the registration accuracy are two aspects that are checked in the evaluation process. The former is determined by comparing the loss-function distributions. Since the aim of the network training is the minimisation of Equation (5.3), the distribution with the lowest values indicates an improved training performance. The latter aspect quantifies the image similarity by calculating the mutual-information metric (see Section 3.3.2) between the target and deformed images, $m_{\mathrm{MI}}(T, D)$, as well as between the target and source images, $m_{\mathrm{MI}}(T, S)$. The relative deviation

$$\Delta m_{\mathrm{MI}}(T, D, S) = \frac{m_{\mathrm{MI}}(T, D) - m_{\mathrm{MI}}(T, S)}{m_{\mathrm{MI}}(T, S)} \tag{5.6}$$

represents the increase or decrease in image similarity for positive or negative values, respectively.

The mutual-information metric is not chosen as $\mathscr{L}_{\mathrm{sim}}(T, D)$ due to fluctuating distributions of the loss function during the training. The distributions of $\mathscr{L}(T, D, \phi)$ for the normalised cross-correlation and the mutual information are shown in Figure A.3 in the appendix.

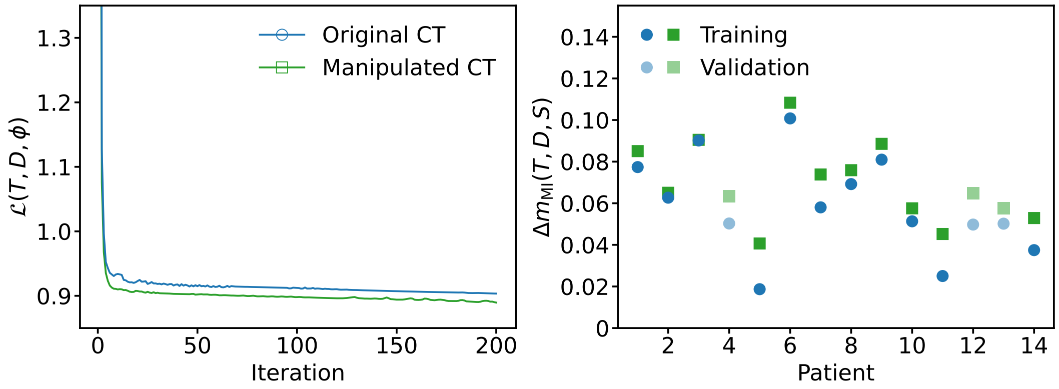


Figure 5.4: Registration results regarding CT-scan manipulation with the 14-patient data set. The distributions of the loss function (left plot) are shown for the network training of two input variations. The input contains either original (blue) or manipulated (green) CT scans. The approach for the manipulation is discussed in detail in the main text. Lower values of $\mathscr{L}(T, D, \phi)$ indicate an improved registration performance. The difference in the mutual-information metric (right plot) after and before DIR is calculated for both variants of CT scans (blue circles and green squares). Higher values of $\Delta m_{\mathrm{MI}}(T, D, S)$ express an increase in image similarity.

CT-scan manipulation Multimodal image registration is challenged by defining accurate image-similarity measures [6] for $\mathscr{L}_{\mathrm{sim}}(T, D)$ in Equation (5.3). This issue is caused by the different distributions of the pixel values (see Figure 4.6). The largest discrepancy between CT and MRI scans is the opposite portrayal of bones (see Figure 4.2), which complicates the choice of an appropriate metric. However, the normalised cross-correlation (see Section 3.3.1) is used in this thesis due to its robustness with respect to fluctuations of pixel values in the target and source images [33]. To investigate the effect of the bone discrepancy, two registration models are trained with the default configuration of the CNN. These settings are listed in Table A.3. The input of one model, consisting of preprocessed CT and MRI scans, is unchanged, while the other model contains manipulated preprocessed CT scans. The manipulation is done by subtracting 300 from the pixel values higher than 300 HU. This maintains the morphology, but softens the high difference in the pixel values between CT and MRI coming from bone tissue. As shown in Figure 5.4, the distribution of the loss function for the model with manipulated CT scans is slightly shifted towards lower values. In addition, the CT manipulation causes an improvement in image similarity for all patients after DIR. This effect is measured for the training and validation data, where an average increase of 1 % and 1.2 %, respectively, is determined. Therefore, further investigations in Section 5.2 are performed with manipulated CT scans as part of the input during network training.

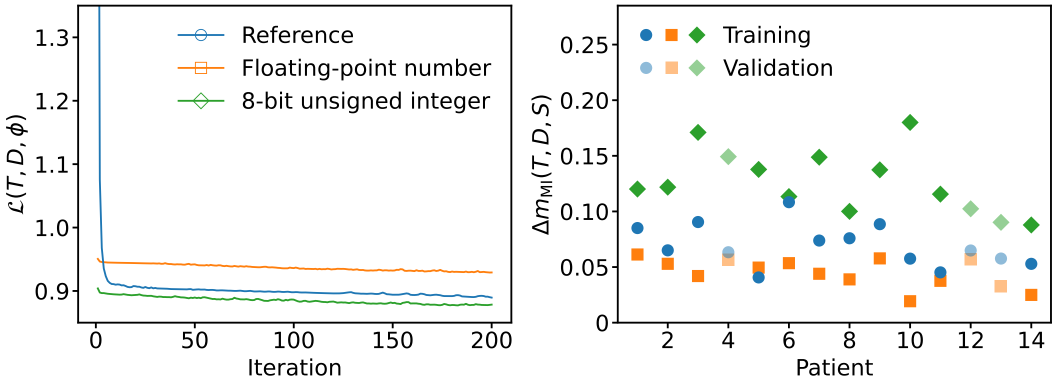


Figure 5.5: Registration results for different data types of the input images with the 14-patient data set. The loss-function distributions (left plot) are shown for each variant of the data type. The distribution of the reference model (blue) is the same as in Figure 5.4 for manipulated CT scans. The loss values of the models trained with the floating-point (orange) and 8-bit (green) data types are depicted as well. An improved registration performance is obtained for lower values of $\mathscr{L}(T, D, \phi)$. The evaluation of the deformed MRI scans (right plot) with the mutual-information metric, $\Delta m_{\mathrm{MI}}(T, D, S)$, is calculated for each patient. Higher values of the metric indicate an increase in image similarity.

Input type The data type of images defines the range of pixel values, which depends on the modality. For the data sets in this thesis, the data type is different for the CT and MRI scans (see Section 3.1.2). In the preprocessing, the normalisation of the images should include the choice of a unified data type. Therefore, the effect on DIR of changing the data type of the input is investigated. The registration model trained with manipulated CT scans in the previous study is the reference for the evaluation. This model contained the CT and MRI scans with their initial data types as input. For the variation, floating-point numbers and 8-bit unsigned integers are taken into consideration. The former defines the range of pixel values between zero and unity. The latter expresses the pixel values in 256 greyscale values. The network training is conducted with the default configuration of the CNN (see Table A.3). In Figure 5.5, the outcome regarding loss function and image similarity is presented for the three registration models. The comparison of the loss-function distributions indicates that the registration model with 8-bit unsigned integers surpasses the reference model, whereas the performance of the training decreases with floating-point numbers. Moreover, the evaluation of the deformed MRI scans shows an improvement in image similarity for the 8-bit data type. For the training data, an average increase of 5.9 % is measured with $\Delta m_{\mathrm{MI}}(T, D, S)$ of the reference and 8-bit settings. The results of the validation data are similar to those of the training data and achieve an improvement of 5.2 % on average. Consequently, 8-bit-integer images are used as input for the network training in further studies. Therefore, the conversion of the CT and MRI scans to that data type is subsequently implemented at the end of the

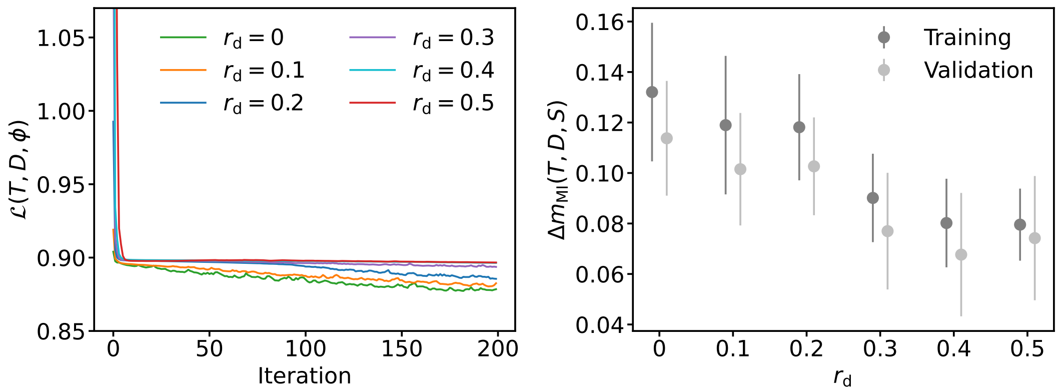


Figure 5.6: Impact of the dropout method on the performance of the CNN using the 14-patient data set. The loss-function distributions (left plot) are shown for each setting of the dropout rate, r_d. Lower values of $\mathscr{L}(T, D, \phi)$ indicate an improved registration performance. The mutual-information metric (right plot) indicates the change in image similarity for the images of each patient, depending on r_d. The distributions are presented as the mean values and the uncertainty of training and validation data. Higher values of $\Delta m_\mathrm{MI}(T, D, S)$ express an increase in image similarity.

image-preprocessing workflow, as mentioned in Section 4.4.1.

Dropout layer The construction of the CNN connects the feature maps of previous and subsequent layers. An approach aiming at increasing the network performance is the regularisation of the feature maps with dropout layers [32]. This method sets a specific fraction of feature maps to zero, which is defined by the dropout rate, r_d. The rate, typically varied between 0.2 and 0.5, is only active during network training [32]. In addition, the remaining feature maps are scaled up with a factor of $1/(1 - r_\mathrm{d})$ to compensate missing connections. The implementation is done with the `Dropout()` function from the `tf.keras.layers` module. These layers are applied after each convolution operation in the encoder and decoder paths of the CNN. The dropout rates $r_\mathrm{d} \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$ are selected to check for improvements in the performance. For this, the registration models with different dropout rates are trained and compared to the reference model ($r_\mathrm{d} = 0$) with manipulated CT scans and 8-bit data type. The distributions of the loss function and the evaluation of the deformed MRI scans with the mutual-information metric are shown in Figure 5.6. Interestingly, the effect of a dropout layer on the CNN performance differs from the expectations. The increase of r_d leads to a deterioration in the CNN performance since the task of minimising $\mathscr{L}(T, D, \phi)$ is best accomplished by the reference model. Thus, dropout layers are disadvantageous for the image-registration techniques used in this thesis. Further investigations with the CNN disregard dropout layers during network training.

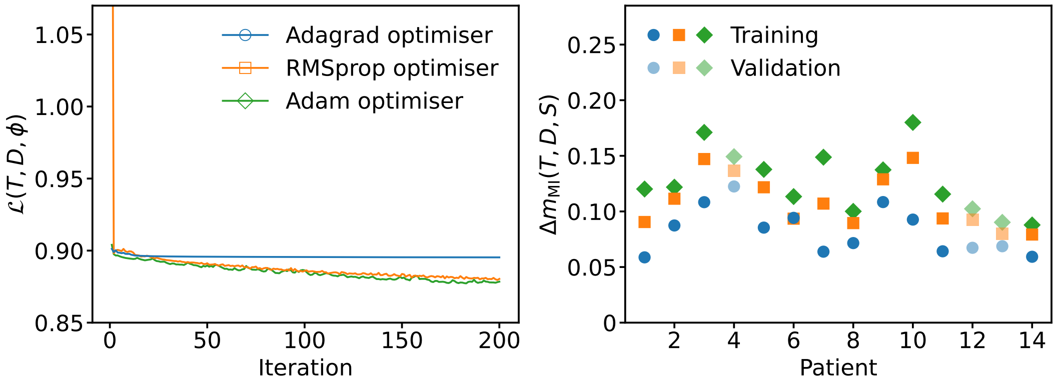


Figure 5.7: Registration results for different optimiser functions of the CNN with the 14-patient data set. The loss-function distributions (left plot) are shown for the Adagrad (blue), RMSprop (orange) and Adam (green) optimisers. An improvement in registration performance is obtained for lower values of $\mathscr{L}(T, D, \phi)$. The mutual-information metric, $\Delta m_{\mathrm{MI}}(T, D, S)$, is calculated for each patient to assess the image similarity (right plot) between the CT and deformed MRI scans. Higher values of the metric indicate an increase in image similarity.

Optimiser function The optimisation of the weights in the convolution kernels depends on the implemented function. As described in Section 5.1.2, the Adam algorithm [52] is the standard function in this thesis. Since the goal of the parameter tuning is to find an appropriate configuration of the network, two other optimiser algorithms, Adagrad [53] and RMSprop [54], are tested. Similar to the Adam optimiser, the Adagrad and RMSprop algorithms include continuously reduced adaptive learning rates during network training [32]. To investigate the impact of these algorithms, three registration models are trained for 200 iterations with the initial learning rate $\alpha = 10^{-4}$. In Figure 5.7, the trained models are evaluated in terms of the loss function and the similarity between the CT and deformed MRI scans. The distributions of the loss function indicate that the RMSprop algorithm achieves similar results to the Adam algorithm. Moreover, the performance of the registration model including the Adagrad optimiser decreases, which is quantified with the mutual-information metric. Compared to the model with the Adam optimiser, average deteriorations of 4.9 % and 2.8 % are determined for the training and validation data, respectively. In addition, the loss-function distribution of the RMSprop model is slightly shifted towards higher values of $\mathscr{L}(T, D, \phi)$ compared to that of the Adam model. By calculating the difference of $\Delta m_{\mathrm{MI}}(T, D, S)$ between the RMSprop and the Adam algorithms, a decrease of 2.0 % for the training data and 1.1 % for the validation data is found. This confirms that the standard optimiser is superior to the Adagrad and RMSprop optimisers and represents the most appropriate setting for the CNN.

Conclusion The preliminary studies aim at finding appropriate settings related to the input images and to two CNN parameters. The challenge for multimodal image registration is the difference in the distributions of the pixel values. Therefore, adjustments of the input images are investigated, which improve the image alignment after the deformable registration. Consequently, the input images should include manipulated CT scans and the 8-bit data type. Furthermore, two approaches for optimising the CNN performance are examined by varying optimiser functions and using various dropout rates. The results indicate that other variations besides the default setting decrease the performance. Since no improvements are determined regarding these CNN parameters, other parameters of the network are taken into consideration in the extensive study in Section 5.2.2.

5.2.2 Extensive study

Parameter tuning is performed to find optimal settings of the CNN parameters. The results of the preliminary studies have provided useful information on the construction of the input images for the network training, whereas the variation of optimiser functions and the addition of dropout layers have remained ineffective. However, unconsidered CNN parameters could still lead to an improvement in the registration performance, which is sought by a more detailed parameter tuning. Therefore, extensive tests are carried out on both data sets to train and evaluate several registration models. Each model is trained with 80 % of the CT and MRI scans from the 25-patient data set, representing the training data. The remaining five image pairs are used for validation by applying the trained models to these images. The fivefold cross-validation technique assures that the images of each patient are at least once assigned to the training and validation data. The concept is visualised in Figure 5.8a. To assess the impact of the parameter tuning, the network training is performed with 200 iterations and the same convolution weights for the initial run. The configuration with the parameter settings that are found to be suitable for more precise DIR is set for the training of a registration model with the entire 25-patient data set. This model is then applied to the testing data, consisting of the CT and MRI scans of the 14-patient data set. The evaluation focuses on quantitative and qualitative comparison of the registered images. First, the registration accuracy and performance are determined to quantify the quality of the outcome, where the results of all five folds are combined within one configuration. Then, image overlay displays of the registered images are generated for visual assessment.

Variations The parameter tuning aims at investigating the effect of different CNN configurations, which are controlled by four parameters: batch size, regularisation parameter, learning rate and architecture. Besides the default setting of the batch size (see Table A.3), another value of b is chosen to analyse the generalisation of the registration models. As the training data consist of 20 image pairs, one possibility is to train models

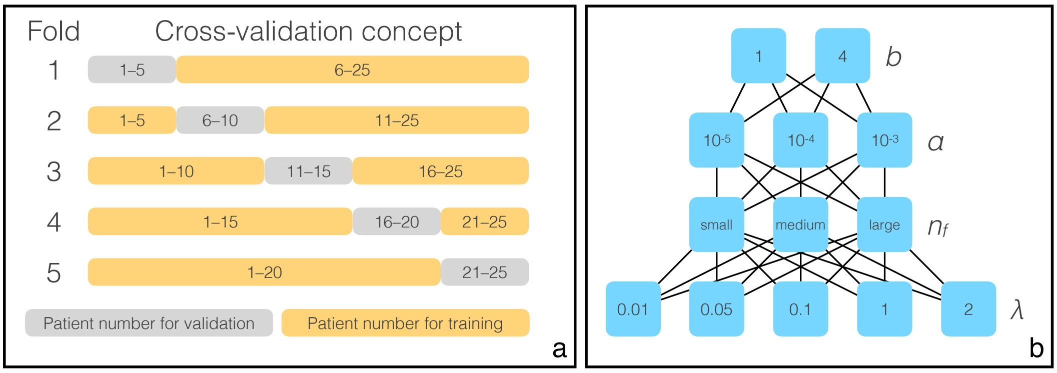


Figure 5.8: Illustration of the cross-validation technique and the parameter tuning of the extensive study. Subfigure a: The fivefold cross-validation requires five registration models to be trained with the same configuration, but with different data. The 25-patient data set is subdivided into the training and validation data with 20 and 5 image pairs, respectively. Subfigure b: The parameter tuning includes the variation of the batch size (b), the learning rate (α), the architecture (n_f) and the regularisation parameter (λ). Regarding n_f, the number of feature maps in the respective layers is described in the main text.

Table 5.1: Variations of the CNN architecture regarding the number of feature maps. The numbers are given for the feature maps of each layer.

Name	Encoder	Decoder
Small	[8 – 16 – 16 – 16]	[16 – 16 – 16 – 16 – 16 – 8 – 8]
Large	[24 – 48 – 48 – 48]	[48 – 48 – 48 – 48 – 48 – 24 – 24]

with $b = 4$, splitting the input into five groups of four image pairs each. The scaling factor λ, included in the loss function in Equation (5.3), regulates the smoothness of the displacements. The settings $\lambda \in \{0.01, 0.05, 0.1, 1, 2\}$ are chosen to investigate the impact of λ values around unity. The investigation of three optimiser algorithms in the previous section included the fixed learning rate $\alpha = 10^{-4}$ for the optimisers. In this study, two other values, $\alpha = 10^{-5}$ and $\alpha = 10^{-3}$, are taken into account for the Adam optimiser. Moreover, the size of the network architecture is varied by means of the number of resulting feature maps in the levels of the encoder and decoder paths. The medium-architecture model corresponds to the default setting (see Table A.3), whereas a small-architecture model and a large-architecture model, listed in Table 5.1, are tested with fewer and more feature maps, respectively. The parameter tuning leads to 90 CNN configurations, which is visualised in Figure 5.8b. Due to the application of fivefold cross-validation, 450 models are trained in total.

Quantitative evaluation: Accuracy The training and validation data are used to calculate the value of $m_{\mathrm{DSC}}(T, D)$ as defined in Equation (3.9) between the segmented images in their target and deformed states. The change in accuracy is then determined by calculating the difference

$$\Delta m_{\mathrm{DSC}}(T, D, S) = m_{\mathrm{DSC}}(T, D) - m_{\mathrm{DSC}}(T, S) \tag{5.7}$$

with the value of $m_{\mathrm{DSC}}(T, S)$, representing the overlap after the preprocessing workflow. The segments that contribute to the computation of the metric are the left and right eyes as well as the ventricular system with the labels l_{LE}, l_{RE} and l_{VS}. While positive values of $\Delta m_{\mathrm{DSC}}(T, D, S)$ imply an increase in the overlap of the segments, a deterioration is observed for negative values. The accuracy of registration models with the batch size $b = 1$ is shown in Figure 5.9, separated into training and validation. The distributions of the configurations with $b = 4$, which are presented in Figure A.4, indicate similar tendencies. In general, a decreasing trend of the accuracy is visible for lower values of the regularisation parameter. These models reach large negative values, which means that the image alignment is degraded by up to 20 % for configurations like the large-architecture model with the settings $b = 1$, $\lambda = 0.01$ and $\alpha = 10^{-4}$. Contrary to that, models yield better accuracy if the factor of the regularisation in the loss function is $\lambda = 1$ or $\lambda = 2$. The small-architecture model with the settings $b = 1$, $\lambda = 2$ and $\alpha = 10^{-3}$, for example, can improve the image alignment by up to 3.5 %. The variation of the learning rate shows that models with $\alpha = 10^{-3}$ mostly achieve lower accuracy than models with $\alpha = 10^{-4}$ or $\alpha = 10^{-5}$. For some models with $\alpha = 10^{-3}$, the distribution of the loss function is unstable because of an increase in the loss value after many epochs. Furthermore, the comparison between the results of the training and validation data illustrates the same tendencies, e.g. low values of λ lead to a decrease in accuracy. When all results are combined, an average value of $\Delta m_{\mathrm{DSC}}(T, D, S)$ greater than zero is determined for four CNN configurations:

1. Small-architecture model with $b = 1$, $\lambda = 2$ and $\alpha = 10^{-4}$
2. Small-architecture model with $b = 4$, $\lambda = 2$ and $\alpha = 10^{-3}$
3. Large-architecture model with $b = 1$, $\lambda = 2$ and $\alpha = 10^{-5}$
4. Large-architecture model with $b = 4$, $\lambda = 2$ and $\alpha = 10^{-4}$

The qualitative evaluation with image overlay displays (see Figure 5.13) will show that the largest deformations take place in the back of the head. The segmented images do not cover this region of the head; therefore, these deformations are not considered in the computation of $\Delta m_{\mathrm{DSC}}(T, D, S)$. To obtain a stronger statement on the registration accuracy, the mutual-information metric from Equation (5.6) is calculated for each deformation. The results, listed in Tables A.4 and A.5, indicate the same tendencies

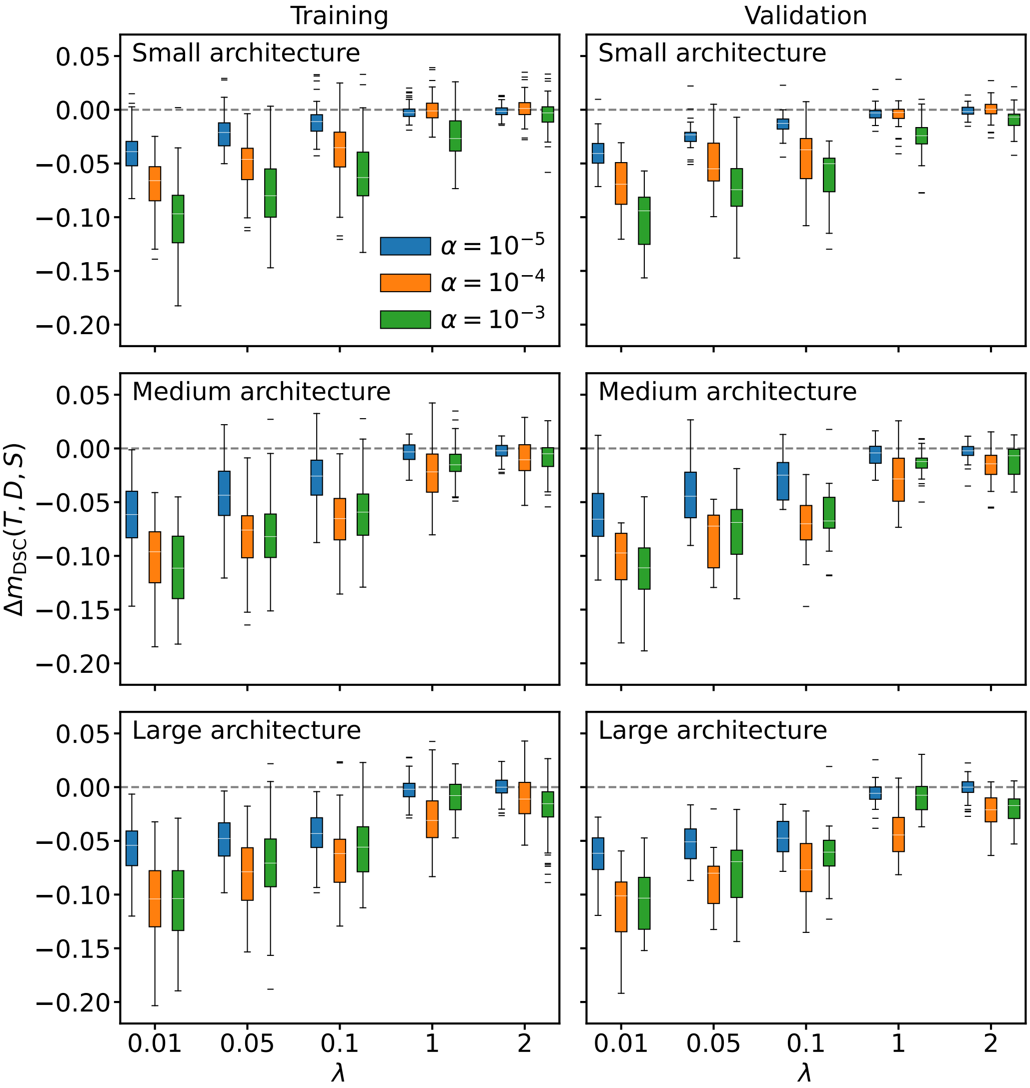


Figure 5.9: Registration accuracy based on the Dice similarity coefficient for models with the batch size $b = 1$ using the 25-patient data set. The value of $\Delta m_{\mathrm{DSC}}(T, D, S)$ is measured as the change in the overlap of segments before, S, and after, D, deformation with regard to the target image, T. The results are shown for the training (left) and validation (right) data as well as for the small-architecture (top), medium-architecture (middle) and large-architecture (bottom) models. Each plot contains the variations of λ and α, regulating the smoothness of the deformations and the step size of the optimiser, respectively. The white lines inside the boxes represent the median values. The dashed lines indicate the border for the increase (positive values) or decrease (negative values) in image alignment.

as $\Delta m_{\mathrm{DSC}}(T, D, S)$, but the measurement is more accurate. An improvement in image similarity is present for positive values of $\Delta m_{\mathrm{MI}}(T, D, S)$, which is the case for higher values of λ, achieving up to 12 % on average.

Quantitative evaluation: Performance Two checks recommended by the AAPM Task Group No. 32 [4] are performed on the deformation vector fields to assess the registration performance. The inverse-consistency method measures the independence of the registration direction. This means that the CT-to-MRI registration should achieve results similar to the training direction, the MRI-to-CT registration. For this, a trained MRI-to-CT model is applied to the image pair consisting of a CT and MRI scan as source and target images, respectively. The addition of the deformation vector fields of both the MRI-to-CT and the CT-to-MRI directions determines the uncertainty for each pixel, which is expected to vary around zero. The outcome of the parameter tuning for the registration models with $b = 1$ is depicted in Figure 5.10. The application of the trained models has the same effects on the validation data, which is evident from the similar distributions. Furthermore, there is a trend towards larger uncertainties for lower values of λ in all architectures, and the distributions are, in addition, similar between the three learning rates. Regarding the small-architecture models, the mean values decrease towards zero from low to high λ values, which indicates an improvement in registration performance. All parameter settings of this architecture are consistent within the uncertainties. In contrast, the mean values of the medium architecture and the large architecture neither increase nor decrease in a significant way when varying the λ parameter. Also, the low uncertainties with respect to the parameter settings $\lambda = 1$ and $\lambda = 2$ do not fulfil the inverse consistency, which means that the registration performance of these architecture models is lower than that of the small-architecture models. The distributions of the configurations with $b = 4$ are shown in Figure A.5, presenting the same tendencies. Another check quantifies the change in the pixel volume by calculating the Jacobian determinant, $m_{\mathrm{JD}}(\phi)$. For each pixel of the deformation vector field, the gradients, which are necessary to compute the determinant, are approximated as the difference in the displacements in x, y and z directions. A determinant below unity indicates volume reduction, while a determinant above unity implies the opposite. For the assessment of the registration performance, the mean value is computed, which is expected to vary by unity. Stronger uncertainties of unity indicate inauthentic deformations, resulting from a decreased registration performance. The distributions of the Jacobian determinant for the configurations with $b = 1$ are shown in Figure 5.11. The uncertainties are larger for low-λ models, which hints at unstable registration performance. The determinants of the small-architecture model with $\alpha = 10^{-3}$, for example, range from 0.5 to 1.5, which is inauthentic for deformations in the brain. For $\lambda = 1$ and $\lambda = 2$, the mean values are close to unity with small uncertainties, which is expected for scans of the same cohort. Comparable conclusions are made for the models trained

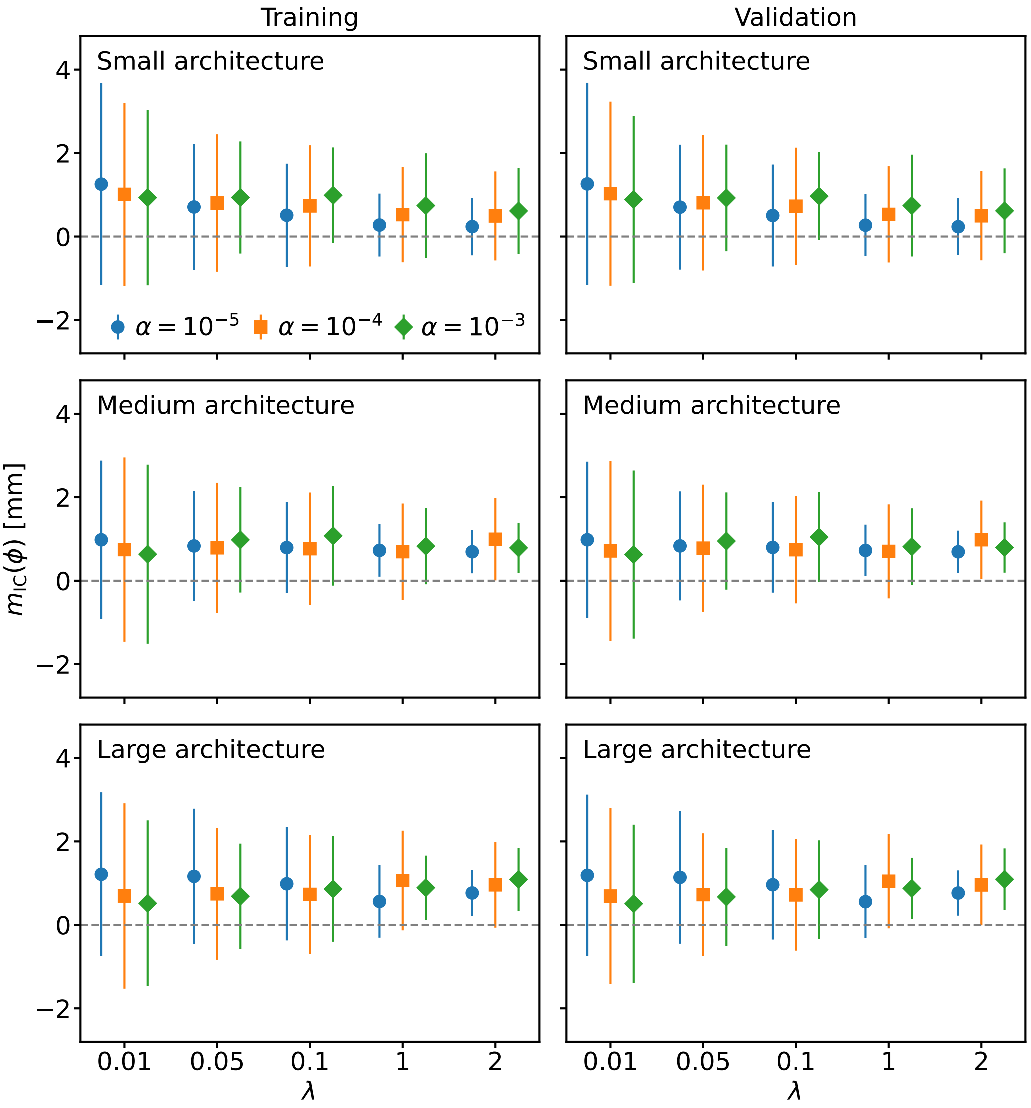


Figure 5.10: Registration performance based on the inverse-consistency method for models with the batch size $b = 1$ using the 25-patient data set. The sum of the MRI-to-CT and the CT-to-MRI deformation vector fields leads to individual values for each pixel, which explains the large error bars. The results are shown for the small-architecture (top), medium-architecture (middle) and large-architecture (bottom) models. The mean values and uncertainties of the respective five-fold data are presented for the variation of the regularisation parameter λ and the learning rate α, separated into training (left) and validation (right). The dashed lines indicate the expected value.

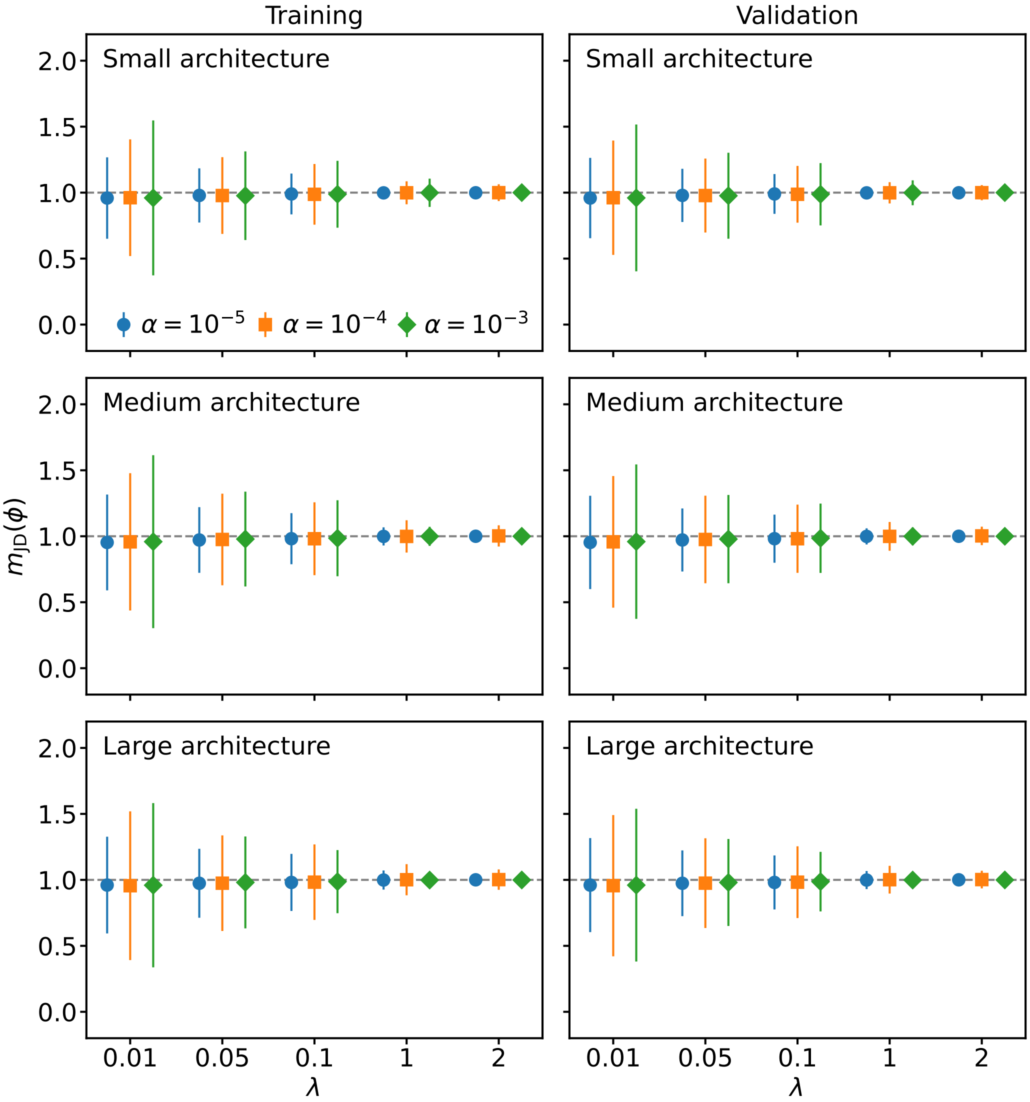


Figure 5.11: Registration performance based on the Jacobian determinant for models with the batch size $b = 1$ using the 25-patient data set. The determinant is calculated individually for each pixel with the corresponding displacements from the deformation vector field. The results are shown for the small-architecture (top), medium-architecture (middle) and large-architecture (bottom) models. The mean values and uncertainties of the respective five-fold data are presented for the variation of the regularisation parameter λ and the learning rate α, separated into training (left) and validation (right). The dashed lines indicate the expected value.

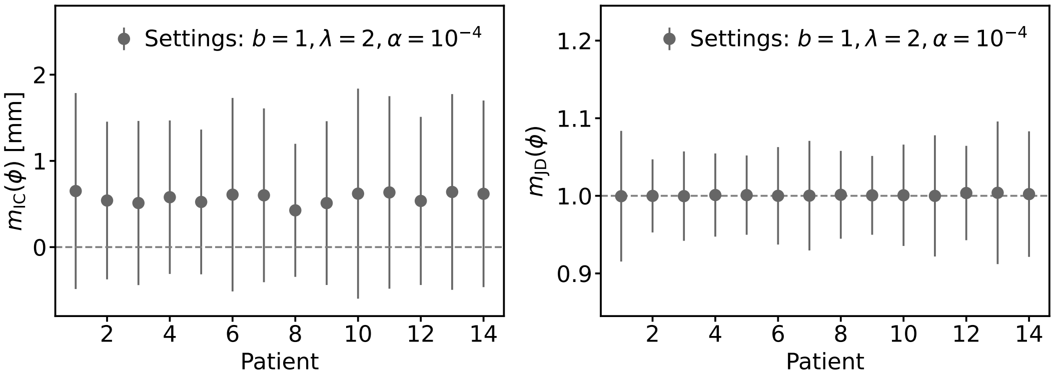


Figure 5.12: Registration performance of a model with the CNN configuration for more precise DIR using the 14-patient data set. The inverse-consistency method (left) and the Jacobian determinant (right) are used to evaluate the performance of the small-architecture model with the batch size $b = 1$, the regularisation parameter $\lambda = 2$ and the learning rate $\alpha = 10^{-4}$ for each patient. The dashed lines indicate the expected values.

with the batch size $b = 4$ (see Figure A.6). In summary, the evaluation indicates that an appropriate CNN configuration should include the requirement $\lambda \geq 1$ for improved registration performance. With respect to the results on the registration accuracy, the small-architecture model with the parameter settings $b = 1$, $\lambda = 2$ and $\alpha = 10^{-4}$ is the most suitable CNN configuration for more precise DIR.

Quantitative evaluation: Testing The CNN configuration that is found to be optimal on the 25-patient data set is used to train a registration model with the scans of the 14-patient data set. The network is constructed with the small architecture and the settings $b = 1$, $\lambda = 2$ and $\alpha = 10^{-4}$. The registration performance is evaluated with the inverse-consistency method and the calculation of the Jacobian determinant, shown in Figure 5.12. Regarding the former check, the results meet the expectations because of the equal behaviour compared to the training data. The values of the inverse-consistency method are consistent with zero, evidenced by the average value of $\bar{m}_{\text{IC}} = (0.57 \pm 1.00)$ mm. For the second check, determinants close to unity are measured for all patients, with small uncertainties, similar to the results of the training data. The average of the mean values is determined to $\bar{m}_{\text{JD}} = 1.00 \pm 0.07$.

Qualitative evaluation Similar to the qualitative evaluation of the preprocessed images in Section 4.4.3, image overlay displays are used to asses the visual impact of DIR on the training and testing data. The overlay is generated with the ImageJ programme [26] by merging the bone tissue of the preprocessed CT scans with the MRI scans. In Fig-

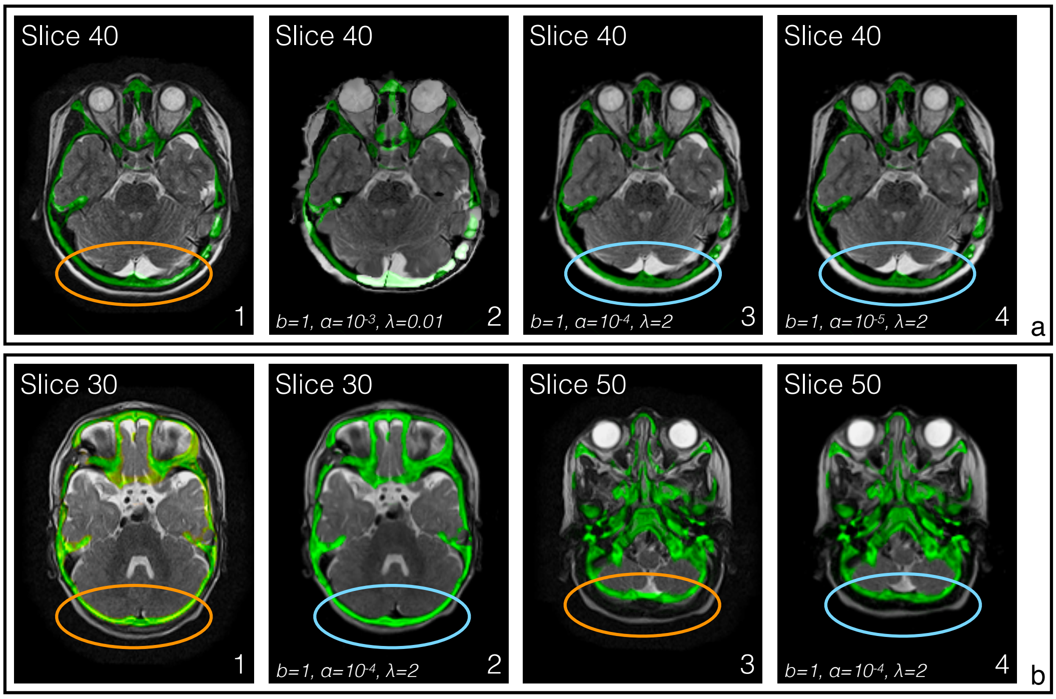

Slice 40
1
Slice 40
b=1, a=10^{-3}, λ=0.01
2
Slice 40
b=1, a=10^{-4}, λ=2
3
Slice 40
b=1, a=10^{-5}, λ=2
4
a
Slice 30
1
Slice 30
b=1, a=10^{-4}, λ=2
2
Slice 50
3
Slice 50
b=1, a=10^{-4}, λ=2
4
b

Figure 5.13: Image overlay displays of MRI (greyscale) slices and the bone structures of CT (green colours with transparency of 50 %) slices. The ellipses indicate the regions with large displacements after DIR (blue) compared to the preprocessing (orange). Subfigure a: The overlay of preprocessed images (1) is shown for Patient 24 of the 25-patient data set. A medium-architecture model (2), a small-architecture model (3) and a large-architecture model (4) illustrate the effect of DIR with different CNN configurations. Subfigure b: The effect of DIR on the testing data is shown with the small-architecture model. The preprocessed overlaid slices (1, 3) of Patient 3 of the 14-patient data set are presented for comparison with the deformed MRI slices (2, 4).

ure 5.13a, the outcome of three CNN configurations, which is compared to the overlay of the corresponding preprocessed CT and MRI scans, is presented for one Patient of the training data. The largest discrepancy between the preprocessed CT and MRI scans appears in the back of the head, where the positions of the tissue in the MRI slice differ from the skull in the CT slice. Furthermore, the structures in the front region of the head agree more for the preprocessed images. These effects are related to the patient positioning and the preprocessing. The former aspect can be attributed to the fact that the organs are placed differently for the acquisition in prone position or supine position. The latter uses the centres of the eye segments as coordinates for the calculation of the transformation parameters of rigid registrations. However, the difference in size occurs despite the fact that the pixel spacing of both the MRI and CT scan is set to 1 mm. This indicates distortion effects of MRI with radial degradation towards the outer regions of the body [55–58], which should be corrected with DIR. One overlay in Figure 5.13a shows the result of the medium-architecture model with the parameter settings $b = 1$, $\lambda = 0.01$ and $\alpha = 10^{-3}$. The deformations of the low-λ model are spiky and distorted, deteriorating the MRI scans after DIR. In contrast, the increase in the value of λ smooths the deformations, which is visible for two overlays with a regularisation parameter of 2 in Figure 5.13a. The small-architecture model with $b = 1$ and $\alpha = 10^{-4}$ and the large-architecture model with $b = 1$ and $\alpha = 10^{-5}$ quantitatively achieve the largest improvements in image alignment. While the small-architecture model perfectly aligns the scalp of the MRI scan with the bone of the CT scan in the back of the head, the large-architecture model with $b = 1$ and $\alpha = 10^{-5}$ does not achieve an appropriate agreement with the CT scan. Consequently, the quantitative evaluation stating that the quality of the deformations decreases with lower values of the regularisation parameter and that the small-architecture model with $b = 1$, $\lambda = 2$ and $\alpha = 10^{-4}$ performs the most precise DIR is confirmed. In addition, the overlays of this model applied to the images of one Patient of the testing data are presented in Figure 5.13b. The same effects, which means that most deformations happen in the back of the head, are apparent for Slices 30 and 50. The application of the registration model for DIR leads to a significant improvement in image alignment.

Conclusion The extensive study includes the variations of four parameters of the CNN with different settings, which leads to 90 configurations. An increase in the batch size is expected to generalise the registration model. The results show no significant improvements compared to the setting $b = 1$, which might be caused by the low amount of training data. Similar to that, the variation of the learning rate yields no clear differences. While a value of 10^{-3} achieves lower accuracy in most cases, the results of the settings $\alpha = 10^{-5}$ and $\alpha = 10^{-4}$ often dominate. In contrast, the regularisation parameter has the largest impact on the quality of the deformed images. Both the quantitative and the qualitative evaluation indicate an improvement in image alignment for the setting $\lambda \geq 1$.

These registration models are able to precisely deform the MRI scans, whereas low-λ models generate strongly distorted MRI scans. Regarding the CNN architecture, three variants consisting of different numbers of feature maps in the encoder and decoder paths are tested. The increase in the number of feature maps slightly deteriorates the registration accuracy of the low-λ models, but that effect is less prominent for models with $\lambda \geq 1$. In summary, the small-architecture model with the parameter settings $b = 1$, $\lambda = 2$ and $\alpha = 10^{-4}$ is found to be the configuration with the most precise DIR, achieving similar results on the testing data.

5.3 Results

Deformable image registration is performed with the CNN described in Section 5.1. The investigations in this thesis focus on the application to multimodal images, including parameter tuning of the CNN by varying the input composition as well as various model parameters. The result (see Section 5.2) is the determination of one CNN configuration that is used to train a registration model with data augmentation in Section 5.3.1. The research is completed by unimodal registrations, presented in Section 5.3.2.

5.3.1 Multimodal registration

The parameter tuning includes both data sets, providing 39 CT and MRI scans of the head. As the run time of the 200-iteration training with 20 image pairs amounts to approximately two hours on an NVIDIA A40 GPU, the parameter tuning would be time-consuming for a larger data set. Therefore, DIR with an extended training data set using data augmentation is investigated in the following with the most suitable CNN configuration.

Data augmentation is a common method to strengthen the training of a neural network by extending the available data set. The individual images are used multiple times during the training process, but each image differs slightly from the original image. [32] Here, two transformation operations are performed to obtain three differently extended data sets with 156, 468 and 780 image pairs, each resulting from the 39 original scans. First, the images are mirrored horizontally, vertically and diagonally for an augmentation factor, f_A, of four, which includes the original images. Then, each image is additionally rotated two times by the angles ±20°, which forms the extended data set with 468 image pairs ($f_\mathrm{A} = 12$). For the largest extension with $f_\mathrm{A} = 20$, rotations by the angles ±10° and ±20° are applied.

Besides the extended data sets, a registration model with the 39 original scans ($f_\mathrm{A} = 1$) is trained for comparison. The data sets are randomly shuffled to achieve an even distribution of original and artificially generated images. The number of iterations is set to 200 for the network training, which runs for approximately five hours on 39 image

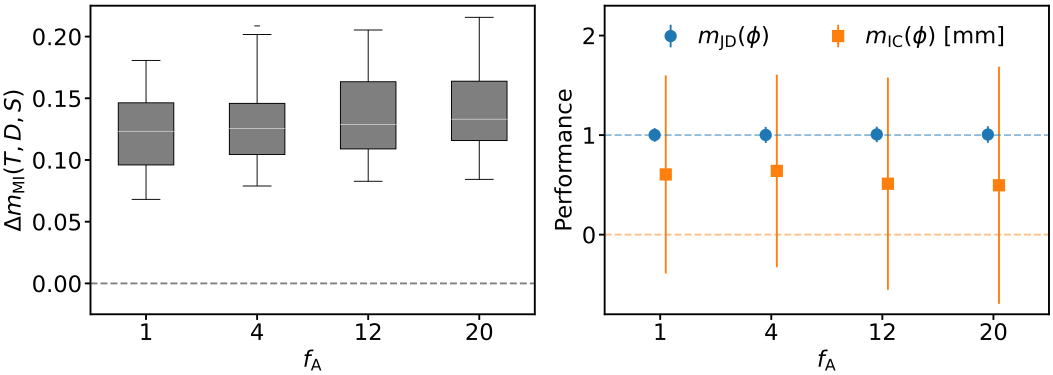


Figure 5.14: Quantitative evaluation of the registration with different data-augmentation factors. Left plot: The mutual-information metric is measured as the change in the metric before (S) and after (D) deformation with regard to the target image (T). The white lines inside the boxes represent the median values. Right plot: The Jacobian determinant (blue) and the inverse-consistency measure (green), which are determined individually for each pixel with the deformation vector field (ϕ), are presented as mean values and their uncertainties. The dashed lines indicate the expected values.

pairs. The run time of the training increases roughly linearly with f_{A}.

Quantitative evaluation Similar techniques as in Section 5.2 are used to evaluate the results of the registration models. The Dice similarity coefficient is avoided since the evaluation in Section 5.2.2 points out that the largest deformations happen in the back of the head, where segments are unavailable. Therefore, the image similarity is measured with the mutual-information metric defined in Equation (3.8). The results are shown in Figure 5.14 for the registration models with different augmentation factors. A slight increase in Δm_{MI} is observed for higher f_{A} values, which means that data augmentation with $f_{\mathrm{A}} = 20$ achieves the largest improvement in image alignment. The mean value over all image pairs rises from $\Delta \bar{m}_{\mathrm{MI}} = 0.12$ for $f_{\mathrm{A}} = 1$ to $\Delta \bar{m}_{\mathrm{MI}} = 0.14$ for $f_{\mathrm{A}} = 20$. Moreover, the performance of the registration models is quantified with the Jacobian determinant and the inverse-consistency measure, visualised in Figure 5.14. The determinants, yielding unity as the average value, indicate an equal distribution of volume changes through the deformations. Furthermore, an improved inverse consistency is obtained for the registration model with $f_{\mathrm{A}} = 20$ compared to the other factors. The obtained mean value $\bar{m}_{\mathrm{IC}} = 0.50 \pm 1.19$ is comparable to the result of the testing data in Figure 5.12. In summary, the measurements agree with the results of the quantitative evaluation in Section 5.2.2, implying adequate registration performance without excessive deformations. Data augmentation leads to a slight improvement in deep-learning-based DIR, but the run time also increases with the augmentation factor.

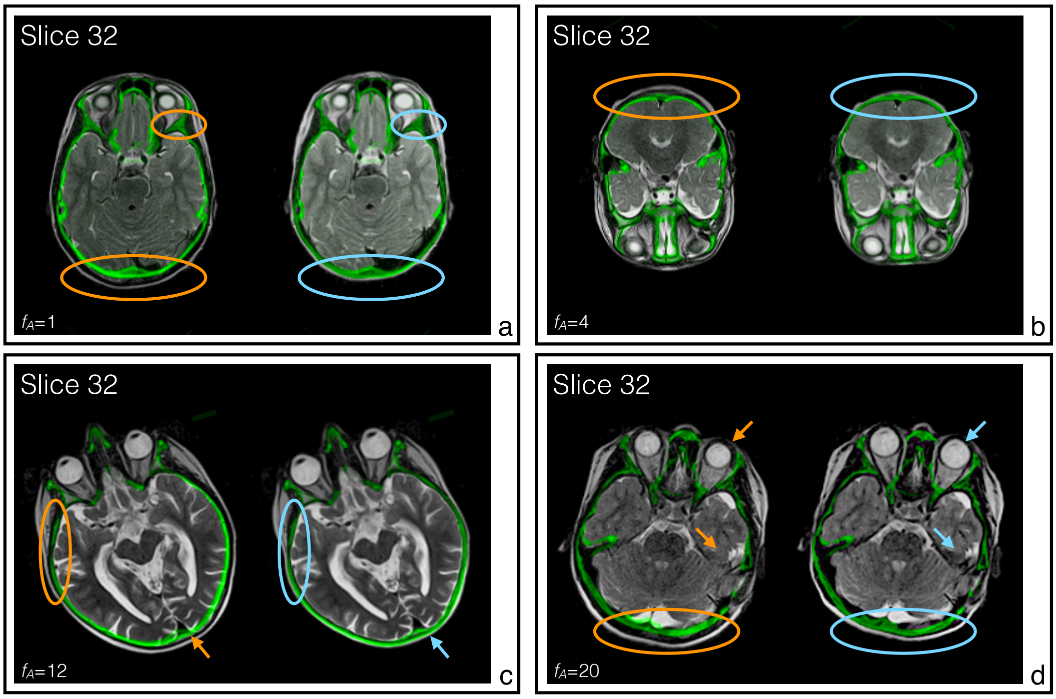


Figure 5.15: Image overlay displays of MRI (greyscale) and the bone structures of CT (green colours with transparency of 50 %) slices for data augmentation. The overlays of preprocessed (left) and registered (right) images are shown in the subfigures with ellipses and arrows, indicating the regions with large displacements after DIR (blue) compared to the preprocessing (orange). The images of Patient 8 (Subfigure a) and the diagonally mirrored images of Patient 3 (Subfigure b) are presented for the 14-patient data set. For the 25-patient data set, the horizontally mirrored images with 20° rotation of Patient 19 (Subfigure c) and the images rotated by −10° of Patient 24 (Subfigure d) are illustrated.

Qualitative evaluation Image overlay displays are generated with the ImageJ programme to overlay green-coloured bone tissue of the CT scans and the MRI scans in greyscale. The preprocessed images are compared with deformably registered images for different data-augmentation factors. In Figure 5.15a, the overlays of Patient 8 of the 14-patient data set are shown for Slice 32. The results of the training, which includes 39 images pairs ($f_A = 1$), indicate larger deformations at the back of the head, while smaller displacements are applied in the front part. The overlays in Figure 5.15b illustrate the impact of data augmentation by a factor of four with a mirrored image of Patient 3 of the 14-patient data set. The comparison of the diagonally mirrored images points out that large deformations are performed independent of prone or supine positions. Furthermore, the effect of adding rotated images ($f_A = 12$) to the training data is visualised in Figure 5.15c for Patient 19 of the 25-patient data set. The horizontally mirrored image, including a rotation of 20°, is part of the training data with 468 image pairs. The overlap of bone tissue from the CT scan and soft tissue from the MRI scan, indicated by arrows, is reduced after DIR. Exemplary overlays for the largest augmentation factor $f_{\mathrm{A}} = 20$ are presented in Figure 5.15d for Patient 24 of the 25-patient data set. There, the images, rotated by an angle of $-10°$ without any reflection operation, are part of the training data with 780 image pairs. Besides the corrections at the back of the head, the registration also includes larger deformations of other regions, like the eyes and the temporal lobe. Ultimately, data augmentation, improving the registration of multimodal images, covers different patient positions, which can support small-sized data sets with regard to deep-learning techniques.

5.3.2 Unimodal registration

As explained in Chapter 1, a variety of publications deal with unimodal image registration for medical use, which is the reason why the focus of this thesis lies on the multimodal application. The CNN described in Section 5.1 is based on the VoxelMorph framework [17], which was applied to T_1-weighted MRI scans only. Therefore, unimodal registration of T_1- and T_2-weighted MRI scans is investigated in the following.

The training of the registration model is performed with the 14-patient data set, which is subdivided into training and testing data with eleven and three image pairs, respectively. Patients 2, 6 and 11 are randomly assigned to form the testing data. The T_2-weighted MRI scans are set as the target images, while the T_1-weighted MRI scans are the source images, which are deformed. The network training runs 200 iterations with a small CNN architecture (see Table 5.1) and the parameter settings $b = 1$ and $\alpha = 10^{-4}$. Since the regularisation of the loss function has the largest impact in the T_2-to-CT registration, three models with different values of λ are trained and compared. The variations are $\lambda \in \{0.1, 1, 2\}$. In addition, the normalised cross-correlation is included as the similarity term, defined in Equation (5.4), in the loss function.

Quantitative evaluation The models are compared in terms of the loss-function distribution and the image alignment, quantified by the mutual-information metric, given in Equation (5.6). In Figure 5.16, the distributions of the loss function indicate an improved performance for low-λ models since the minimisation of the loss function performs more successful for $\lambda = 0.1$. In addition, the same conclusions can be drawn with the assessment of $\Delta m_{\mathrm{MI}}(T, D, S)$ for each image pair, visualised in Figure 5.16. Regarding the training data, the results of the model with the setting $\lambda = 0.1$ surpass those of higher-λ models. The application to the testing data shows slight differences in the λ variation, which could be triggered by the small amount of training data. The image alignment improves by 17.5 % on average. Lastly, these results are contrary to the results of the multimodal registration regarding the λ values, but also here, improvements in

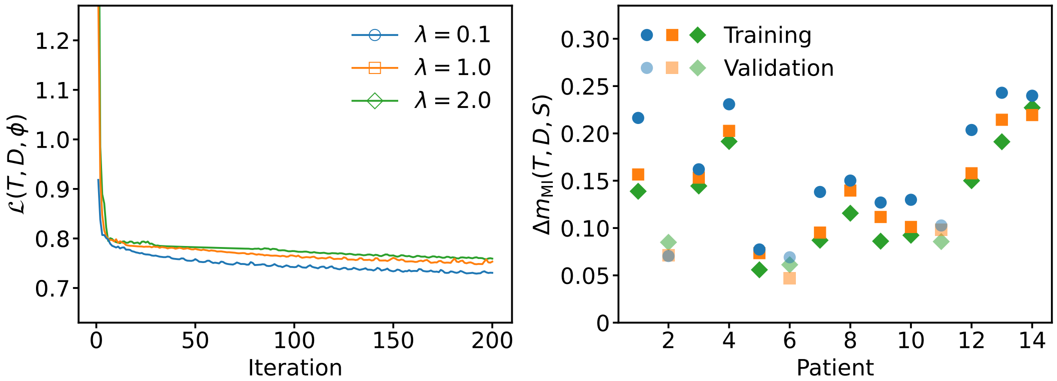


Figure 5.16: Results of the unimodal registration for the 14-patient data set. The distributions of the loss function (left plot) are shown for the network training with three different regularisation parameters, λ. Lower values of $\mathscr{L}(T, D, \phi)$ indicate an improved registration performance. The difference in the mutual-information metric (right plot) after and before DIR is determined for each patient and variation. Higher values of Δm_{MI} express an increase in image similarity.

image alignment are achieved with deep-learning-based DIR. Data augmentation can further increase the data set with duplicates of the original images, but the scaling factor needs to be chosen appropriately.

Qualitative evaluation Checkerboard displays are used to assess the impact of unimodal registrations of T_1- and T_2-weighted MRI scans in Figure 5.17. The display is generated in a grid pattern with alternating parts of both images. For visual comparison, the T_1- and T_2-weighted MRI scans are depicted in magenta and blue colours, respectively. Figure 5.17a includes the displays for Slice 32 of Patient 1 of the training data. The display with the preprocessed images shows the largest differences in image alignment. The scalp differs most between source and target images. This discrepancy is reduced with the lowest-λ model, while higher-λ models are unable to perform appropriate deformations. In Figure 5.17b, Slice 32 of Patient 11 represents the result of the testing data. There, similar issues occur in the display with preprocessed images, but none of the registration models is capable of registering the scalp of the T_1-weighted MRI scan completely. These findings reflect the quantitative evaluation with the mutual-information metric regarding the testing data. Consequently, unimodal DIR of MRI scans is feasible and clearly improves image similarity compared to the preprocessed images.

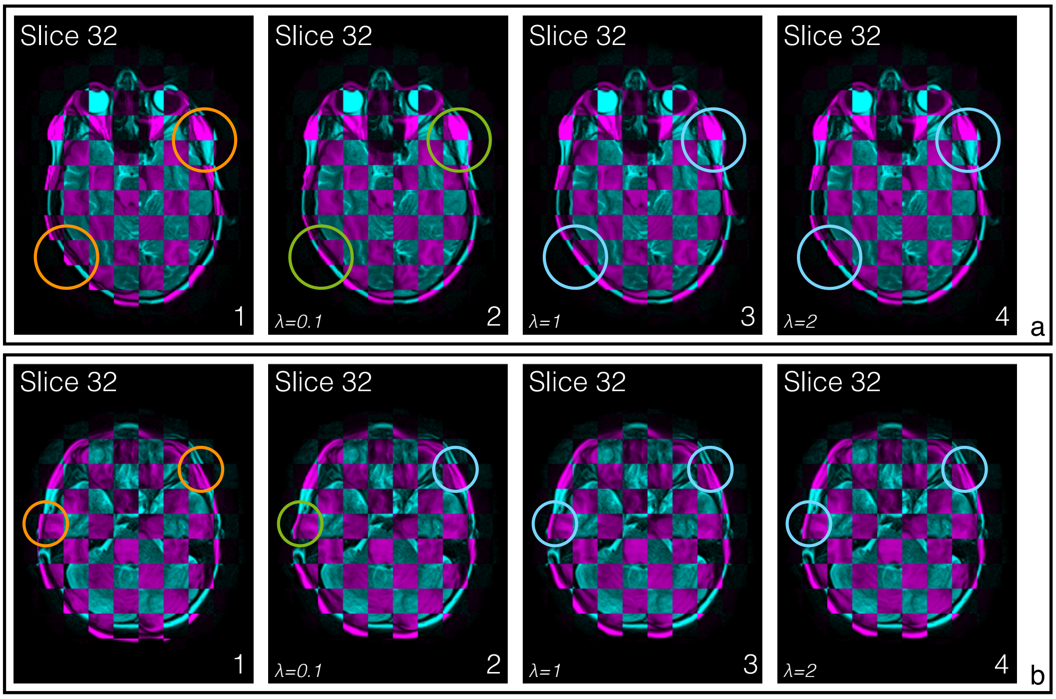


Figure 5.17: Checkerboard displays of T_2-weighted (blue colours) and T_1-weighted (magenta colours) MRI slices. The circles serve the comparison of the displacements after DIR (blue) compared to the preprocessing (orange). Subfigure a: The display is shown for Patient 1 of the 14-patient data set. The preprocessed images (1) as well as the corresponding images with DIR (2–4) illustrate the effect of DIR on the training data. Subfigure b: The display is shown for Patient 11 of the 14-patient data set. The preprocessed images (1) as well as the corresponding images with DIR (2–4) illustrate the effect of DIR on the testing data.

6 Image fusion

The main task of rigid and deformable image registration is the improvement in image alignment of a set of complementary individual images. Image fusion is able to merge these images, increasing the amount of information in one image. Fused images can visually facilitate clinical processes, e.g. treatment planning in radiotherapy. In this chapter, medical image fusion is investigated for the generation of unimodal and multimodal fused images. First, the methodology of merging a set of images is described in Section 6.1. Then, the application of the image-fusion method is presented in Section 6.2, including the results with regard to quantitative and qualitative evaluations.

6.1 Methodology

In general, the fusion of images requires a specific rule for the pixel values of the images. This rule determines the value of each pixel in the fused image. Furthermore, the application of image fusion necessitates the choice of an appropriate method, which depends on the domain of the images. One class of image-fusion methods uses the images in their spatial domain. There, the pixel values are directly included in the generation of the fused images. This procedure entails disadvantages, like distortion effects and low variability. Other methods are part of the transform-domain class, where a specific transformation is applied to the images for image decomposition, which accesses another domain of the images. The fusion rule is applied in that domain. Afterwards, the inverse transformation generates the fused image. The transform-domain technique offers the advantage of merging the images in the respective domain, leading to a higher variability. [59]

A variety of implementations of image-fusion techniques were developed for various applications [59–63]. The transform-domain technique, converting the images into the frequency domain, became a widely used method, which provides a more distinct representation of information [64]. The Fourier-transform method, for example, uses a discrete transformation function to compute the coefficients for the frequency-domain image [62]. During the merging process, the signals of different features can be differentiated to emphasise desired features [62]. In medical image fusion, the wavelet-transform method is often used due to the complexity of the images [59, 61]. The application of the discrete wavelet transform provides localisation in the frequency domain with higher resolution [64]. Therefore, the wavelet-transform method is employed in this thesis for image fusion of the registered images.

6.1.1 Wavelet transform

The discrete wavelet transform applied to images with discrete pixel positions originates from the continuous transformation

$$W_f(s,x) = \int_{-\infty}^{\infty} \psi_{s,x}(t) f(t)\,\mathrm{d}t \tag{6.1}$$

of a function $f(t)$ with the continuous wavelet transform

$$\psi_{s,x}(t) = \frac{1}{\sqrt{s}}\psi\left(\frac{t-x}{s}\right), \tag{6.2}$$

consisting of the scaling, $s \in \mathbb{R}^+$, and the translation, $x \in \mathbb{R}$, of a basis function, $\psi(t)$. The impractical implementation of Equation (6.1) is caused by the continuity of the parameters s and x. Therefore, both parameters are converted into the discrete parameters $s = 2^m$ and $x = n2^m$ with $m, n \in \mathbb{Z}$. Consequently, the discrete wavelet transform

$$\psi_{m,n}(t) = \frac{1}{\sqrt{2^m}}\psi\left(2^{-m}t - n\right) \tag{6.3}$$

includes a set of scaled and shifted basis functions due to the parameter substitution. [65] The basis functions used in this thesis are described in Section 6.1.2.

The implementation of the discrete wavelet transform corresponds to the sub-band coding with low-pass (L) and high-pass (H) filters, separating the signal into approximation and detail coefficients [64]. The number of coefficients is usually reduced afterwards, and the two-band coding additionally offers the possibility for multi-resolution decomposition by the iteration of the filters [64, 65]. The inverse transformation reverses the process by upsampling the coefficients. Both filters are then applied to the coefficients, and the addition of the results yields the approximated coefficients of the higher level. [64] Regarding the discrete wavelet transform, the wavelet function, $\psi(t)$, and a scaling function, $\phi(t)$, imitate the effect of the high-pass and low-pass filters, respectively. The scaling function extracts the essential information, while the wavelet function generates a representation with the details of the signal, $f(t)$. Due to the recursive property, the composition of the signal in level m is the superposition

$$f_m(t) = \sum_n c_{m+1,n}\phi_{m+1,n}(t) + \sum_n d_{m+1,n}\psi_{m+1,n}(t) \tag{6.4}$$

of the decomposed signals from the level $m+1$. The approximation, $c_{m+1,n}$, and the detail, $d_{m+1,n}$, coefficients are multiplied with the respective filter function for each discrete position n. [64, 65]

The application to an image, I, is performed in an alternate procedure along the x and y axes, visualised in Figure 6.1. First, the pixel values are decomposed line by line with scaled and shifted versions of $\phi(t)$ and $\psi(t)$, which produces the approximated,

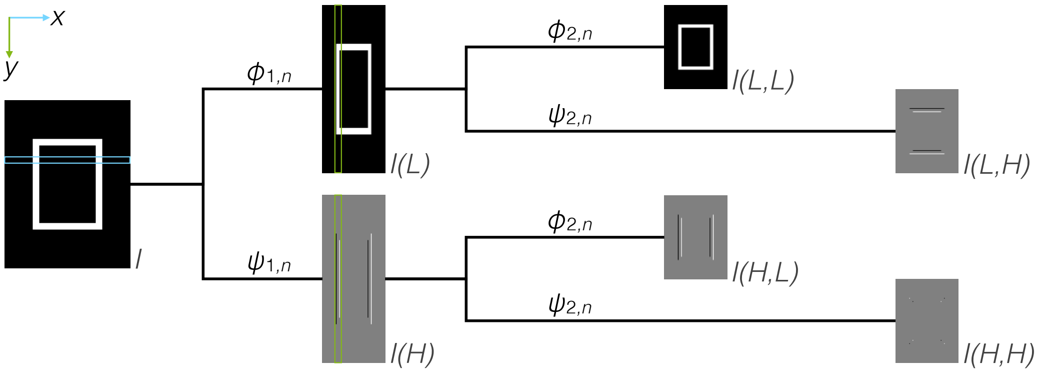


Figure 6.1: Effect of the discrete wavelet transform on an exemplary image with a white frame. The procedure includes two levels of decomposition of the image I. The scaling, $\phi_{m,n}$, and wavelet, $\psi_{m,n}$, functions corresponding to low-pass, L, and high-pass, H, filters are applied to each image of the levels $m = 1$ and $m = 2$. The decomposition in the first level is performed on each line (blue frame) of the x axis by iterating through the pixel positions n. In the second level, the same technique is applied for each column (green frame) of the y axis.

$I(L)$, and detailed, $I(H)$, images. Then, a column-by-column decomposition takes place for both images $I(L)$ and $I(H)$, creating one approximated image, $I(L,L)$, and three detailed images, $I(L,H)$, $I(H,L)$ and $I(H,H)$, in the second level. The main information is collected in the image $I(L,L)$, whereas the vertical, horizontal and diagonal details are located in the images $I(L,H)$, $I(H,L)$ and $I(H,H)$, respectively.

In this thesis, the `dwt2()` and `idwt2()` functions from the PyWavelets library [66] are used for image decomposition and reconstruction. The former function applies the discrete wavelet transform according to Figure 6.1. A two-dimensional image is decomposed by computing the coefficients of the approximated and detailed images. To reverse the process, the inverse discrete wavelet transform is performed with the latter function, including the fused coefficients of the decomposed images. Both functions require a specific wavelet function for operation.

6.1.2 Wavelet groups

The decomposition of the images depends on the scaling and wavelet functions, which form a wavelet group [65]. An important condition is that those functions including the scaled and shifted versions are orthonormal and defined on a finite interval. In general, the scaling function has to meet the functional equation

$$\phi(t) = \sqrt{2} \sum_{n} h_0(n) \phi(2t - n) \tag{6.5}$$

with the sequence $h_0(n)$, representing the coefficients of a discrete low-pass filter. The corresponding wavelet function

$$\psi(t) = \sqrt{2}\sum_n h_1(n)\phi(2t-n) \tag{6.6}$$

results from the coefficients $h_1(n)$ of a discrete high-pass filter and $\phi(t)$. [65] The impact of the functions of six wavelet groups, which are described in the following, is investigated within the scope of the image-fusion study in Section 6.1.3.

Haar The simplest wavelet is the Haar function [67], which is the basis of most other wavelet groups. The scaling function

$$\phi_{\text{Haar}}(t) = \begin{cases} 1, & 0 \leq t < 1 \\ 0, & \text{otherwise} \end{cases} \tag{6.7}$$

fulfils Equation (6.5) with the coefficients $h_0(0) = 1/\sqrt{2}$ and $h_0(1) = 1/\sqrt{2}$ [65]. Therefore, the wavelet function

$$\psi_{\text{Haar}}(t) = \begin{cases} 1, & 0 \leq t < 0.5 \\ -1, & 0.5 \leq t < 1 \\ 0, & \text{otherwise} \end{cases} \tag{6.8}$$

is determined through Equation (6.6) with the high-pass filter coefficients $h_1(0) = 1/\sqrt{2}$ and $h_1(1) = -1/\sqrt{2}$ [65].

Daubechies The group of Daubechies (db) wavelets [68] consists of twenty db$N$ functions, which are orthogonal and asymmetric. The degree of differentiability of these functions increases with the order $N = H/2$, resulting in smoother functions with $H = \{2, 4, \dots, 40\}$ coefficients. The db1 function is equivalent to the non-smooth Haar wavelet, while the db2 to db20 functions have oscillating distributions. [65, 69]

Coiflet Orthogonality and near symmetry are the properties of the Coiflet (coif) wavelets [70]. The number of coefficients, $H = \{6, 12, 18, 24, 30\}$, for the order $N = H/6$ leads to five coif$N$ functions. The distribution of the functions is smoother for higher values of N. [65, 69]

Symlet The Symlet (sym) group [69] contains 19 orthogonal functions with near-symmetric distributions. The sym$N$ functions with the order $N = \{2, 3, \dots, 20\}$ are versions of the db$N$ functions with increased symmetry. [69]

Table 6.1: Configurations of the biorthogonal wavelet functions.

bior1.1	bior1.3	bior1.5	bior2.2	bior2.4
bior2.6	bior2.8	bior3.1	bior3.3	bior3.5
bior3.7	bior3.9	bior4.4	bior5.5	bior6.8

Biorthogonal The computation of biorthogonal (bior) wavelet functions is performed with two scaling and two wavelet functions with the coefficients $(h_0, \tilde{h}_0)$ and $(h_1, \tilde{h}_1)$, respectively. This allows symmetric functions to be generated, depending on the number of coefficients of both scaling functions. Thus, 15 variations, listed in Table 6.1, are possible, in which the bior1.1 function is the Haar function. [71]

Reverse biorthogonal The reverse biorthogonal (rbior) wavelet group is computed similarly to the bior functions. The same configurations of the number of coefficients (see Table 6.1) are possible and the rbior functions are also symmetric. The rbior1.1 function corresponds to the Haar function. [71]

6.1.3 Fusion

The actual fusion of two images, I and J, is performed with the coefficients of their decomposed images, e.g. $I(L, L)$ is merged with $J(L, L)$. The determination of the new coefficients is restricted to the given fusion rule. Three operations [59] are used in the study:

- The minimum (min) rule selects the smallest coefficient of two pixels.
- The average (avg) rule computes the mean of two coefficients.
- The maximum (max) rule selects the largest coefficient of two pixels.

The image-fusion process is shown in Figure 6.2. Since the detailed images $X(L, H)$, $X(H, L)$ and $X(H, H)$ contain less information than the approximated image $X(L, L)$, the same fusion rule is applied to these images. Hence, the notation ruleA–ruleD specifies in the following which fusion rule is used for the approximated (ruleA) and detailed (ruleD) images.

6.1.4 Evaluation

Statistical approaches for quantitative metrics, like the signal-to-noise ratio, measure the fraction of noise in one image with regard to a reference image. However, a reference as ground truth is unavailable for the fusion of two images, I and J, generating the

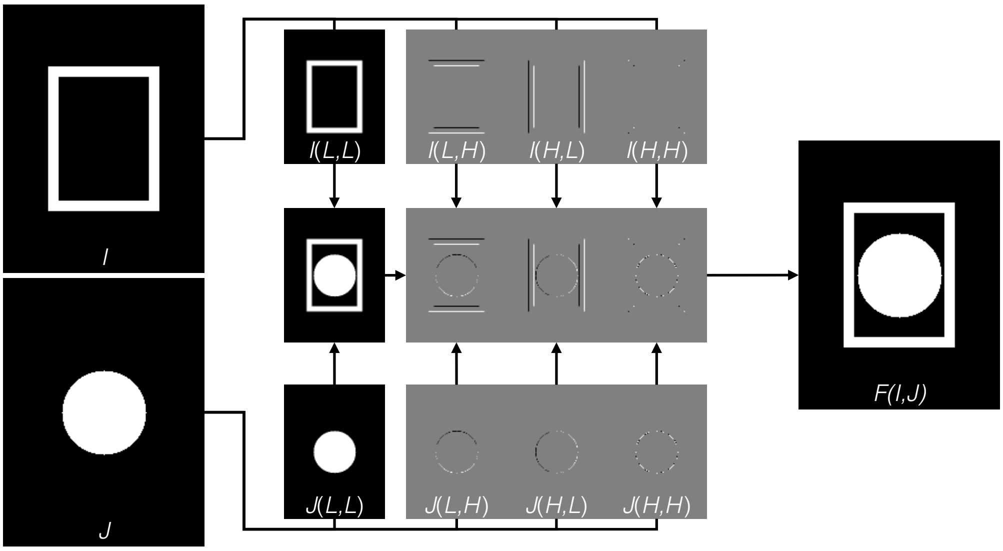


Figure 6.2: Illustration of the image-fusion method with the discrete wavelet transform using the Haar wavelet and the avg–avg rule. The images, I and J, are individually decomposed into approximated ($X(L,L)$) and detailed ($X(L,H)$, $X(H,L)$, $X(H,H)$) images, which are then merged. The inverse discrete wavelet transform generates the fused image, $F(I,J)$.

fused image $F(I,J)$. Therefore, the quality of the fused images is assessed with the mutual-information metric, m_{MI}, from Equation (3.8), which determines the amount of information in $F(I,J)$ from I or J. [72]

Two measures containing m_{MI} are used for the quantitative evaluation. The fusion-factor metric

$$m_{\mathrm{FF}}(I,J,F) = m_{\mathrm{MI}}(I,F) + m_{\mathrm{MI}}(J,F) \tag{6.9}$$

is computed by adding the individual measures. The definition of the mutual-information metric implies that the information content of the fused image improves for higher values of $m_{\mathrm{FF}}(I,J,F)$. In addition, the fusion-symmetry metric

$$m_{\mathrm{FS}}(I,J,F) = \left| \frac{m_{\mathrm{MI}}(I,F)}{m_{\mathrm{MI}}(I,F) + m_{\mathrm{MI}}(J,F)} - 0.5 \right| \tag{6.10}$$

is designed to determine the fraction of information from I and J in $F(I,J)$, indicating the symmetry. Contrary to Equation (6.9), the quality of the fusion increases for values towards zero. [73]

6.2 Results

Image fusion aiming at merging information from multiple images is performed with the discrete wavelet transform on the registered images of the 14- and 25-patient data sets from Section 5.3. The variety of wavelet functions and the resulting decomposed images as well as the fusion rules offer a high degree of variability. Therefore, the wavelet groups are investigated in Section 6.2.1 by comparing the decomposed images of various wavelet functions. After the selection of three functions producing the most distinctive decomposition, the results of multimodal and unimodal image fusion are presented in Sections 6.2.2 and 6.2.3, respectively. First studies on the application of image-fusion methods were carried out in a bachelor's thesis [74], which was supervised by the author.

6.2.1 Selection of wavelet functions

The investigation of the wavelet functions is done on the CT scan as well as the T_1- and T_2-weighted MRI scans of Patient 9 from the 14-patient data set. This patient is randomly selected. The db1, bior1.1 and rbior1.3 functions are omitted as these functions are equivalent to the Haar function. The remaining functions from the wavelet groups described in Section 6.1.2 are compared to select the most appropriate wavelets.

In Figure 6.3, the decomposed images of the CT scan are shown for the Haar, db2, coif1, sym2, bior1.3 and rbior1.3 functions. The approximated images, which contain most information, are very similar to each other, and differences are hardly noticeable. In contrast, the detailed images significantly differ between the wavelet functions, where the images of the Haar and bior1.3 functions clearly stand out. The vertical and horizontal details contain steeper gradients, which results in a stronger representation of edges. This effect is even more visible for the diagonal details. In addition, similar detailed images are generated with the db2 and sym2 functions, differing in their symmetry. The coif2 and rbior1.3 functions qualitatively provide the weakest decomposition of the CT scan. The images decomposed with higher-order wavelet functions of the db, coif, sym, bior and rbior groups are presented in Figure A.7. The distributions from lower to higher orders indicate a decrease in strength of image decomposition, especially for the detailed images. Consequently, the Haar, bior1.3 and db2 functions are selected for the image fusion including the CT scan, which is described in the following section.

The decomposed images of the T_1- and T_2-weighted MRI scans are shown in Figures A.8 and A.9, respectively. The former provide the same results as the CT scan. For the T_2-weighted MRI scan, the decomposition with the Haar and bior1.3 functions performs best according to the detailed images, while the rbior1.3 function produces images in which edge details are not well pronounced. The db2, coif1 and sym2 functions are similar, and the function with the clearest decomposition cannot be visually detected. Therefore, the same wavelet functions as those selected for the CT scan are used for image fusion with MRI scans.

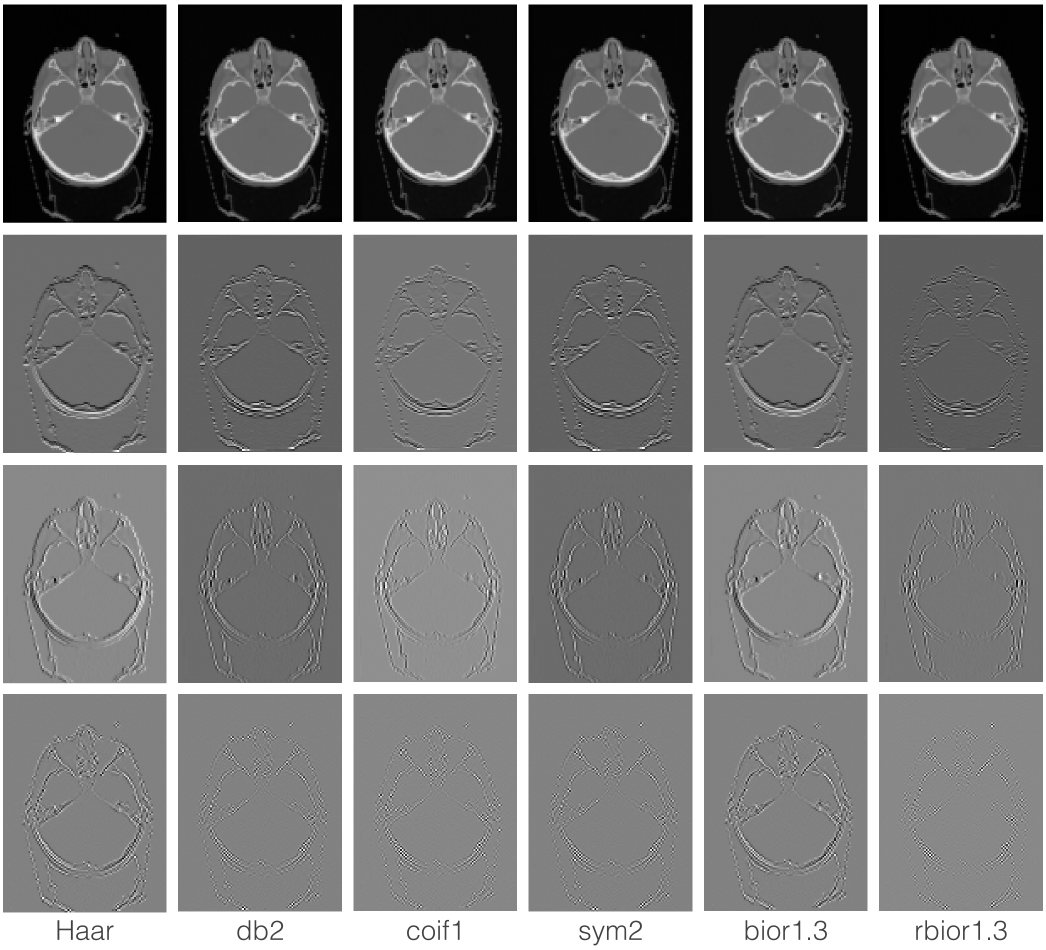


Figure 6.3: Image decomposition with the discrete wavelet transform for the preprocessed CT scan of Patient 9 of the 14-patient data set. The approximated images (top row) as well as the detailed images containing vertical (second row), horizontal (third row) and diagonal (bottom row) information are presented for six wavelet functions.

6.2.2 Multimodal image fusion

The deformably registered CT scans and T_2-weighted MRI scans are used to investigate the fusion of multimodal images. The T_1-weighted MRI scans are not considered for the fusion with CT scans because of the similar representation of fluids (see Figure 4.2). The T_2 weighting, however, facilitates the visual delineation of risk structures and tumours. The fused images are generated for the entire data, which include the 14 and 25 patients of both data sets. Furthermore, the fusion is performed with the Haar, bior1.3 and db2 functions for the nine combinations of the fusion rules. The evaluation is done with the two measures from Equations (6.9) and (6.10), based on the mutual-information metric. In addition, a visual comparison of the fused images is presented to assess the quality and utility.

Quantitative evaluation The distributions of the fusion-factor and fusion-symmetry metrics are shown in Figure 6.4, depending on the wavelet function and fusion rule. The metrics are calculated for each of the 39 fused images, and the mean values and their uncertainties are then determined. In general, both $m_{\mathrm{FF}}(I,J,F)$ and $m_{\mathrm{FS}}(I,J,F)$ indicate the same results in terms of the fusion rule. As the fusion-factor metric measures the amount of information from the CT and T_2-weighted MRI scans in the fused images, the fusion improves for the avg–avg rule. This rule ensures a high information content for all three wavelet functions. The Haar function in combination with the avg–avg rule, for example, yields $m_{\mathrm{FF}}(I,J,F) = 2.41 \pm 0.19$, which is almost identical to the values of the bior1.3 and db2 functions. In contrast, the combination of the approximated images with the minimum rule (min–min, min–avg, min–max) leads to the lowest $m_{\mathrm{FF}}(I,J,F)$ values. These rules by definition select the smallest coefficient of two pixels. Therefore, the fused images are expected to contain mostly the information from the CT scans without bone structures. A similar case is obtained for the max–min, max–mean and max–max rules, transferring most information from the T_2-weighted MRI scans to the fused images. The $m_{\mathrm{FF}}(I,J,F)$ values are slightly higher than those of the minimum rules, increasing the information content by including the bone structures from the CT scans. The results of the fusion-symmetry metric, which assesses how much information from the CT and T_2-weighted MRI scans is included in the fused image, indicate that the average rule (avg–min, avg–avg, avg–max) with regard to the approximated images provides the most symmetric distribution of information. For the Haar function, a value of $m_{\mathrm{FS}}(I,J,F) = 0.028 \pm 0.025$ is measured with the avg–avg rule. The values of the other fusion rules are higher due to the increased asymmetry, which is related to the uneven distribution of information content using the minimum or maximum rule for the approximated images. In summary, fused images with the highest information content and symmetry are generated with the avg–avg rule, while no significant impact of the wavelet function is observed.

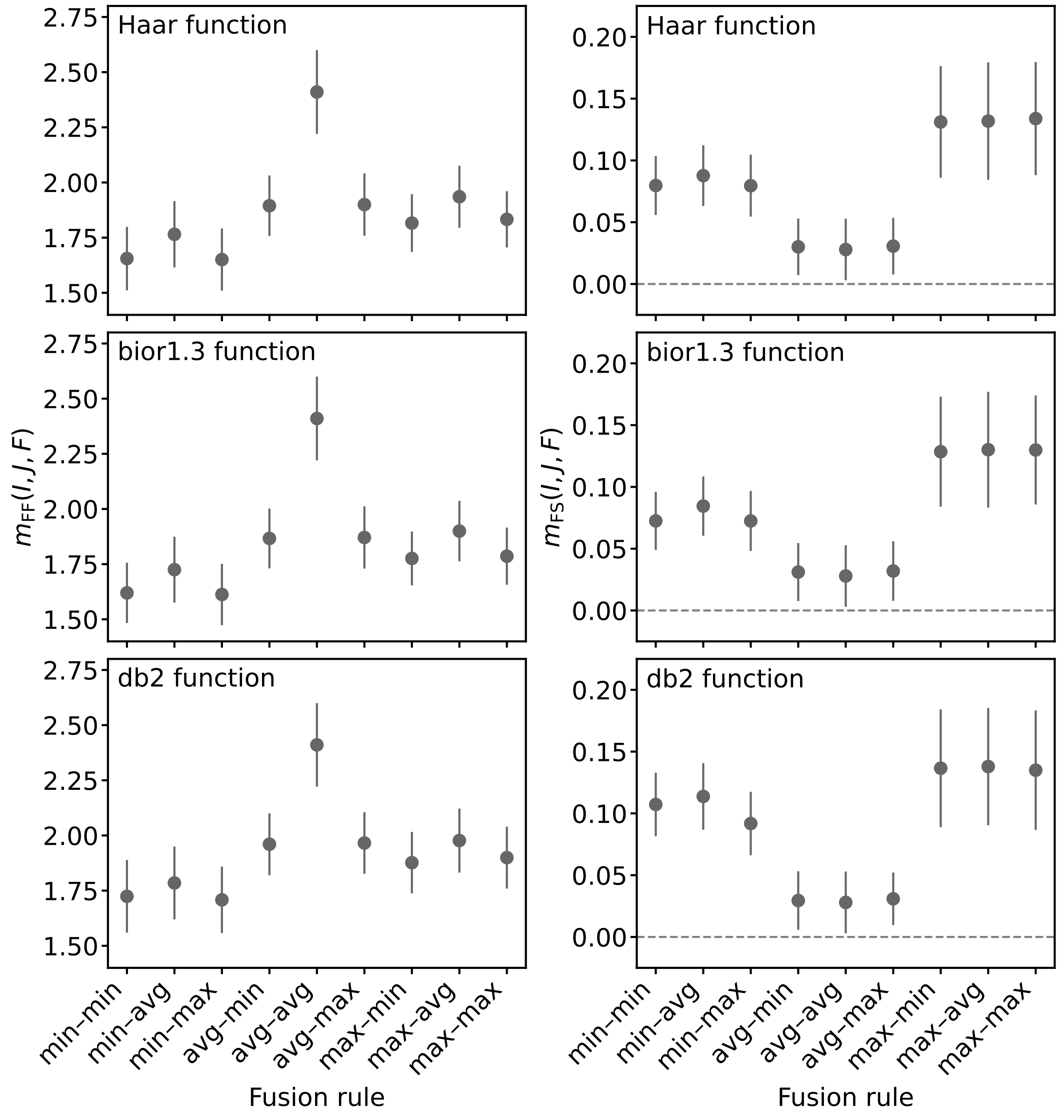


Figure 6.4: Fusion quality of the fused CT and T_2-weighted MRI scans for the patients of the 14- and 25-patient data sets. The fusion-factor (left column) and the fusion-symmetry (right column) metrics are shown for the Haar (top row), bior1.3 (middle row) and db2 (bottom row) functions. The metrics are determined as the mean values and their uncertainties of all patients.

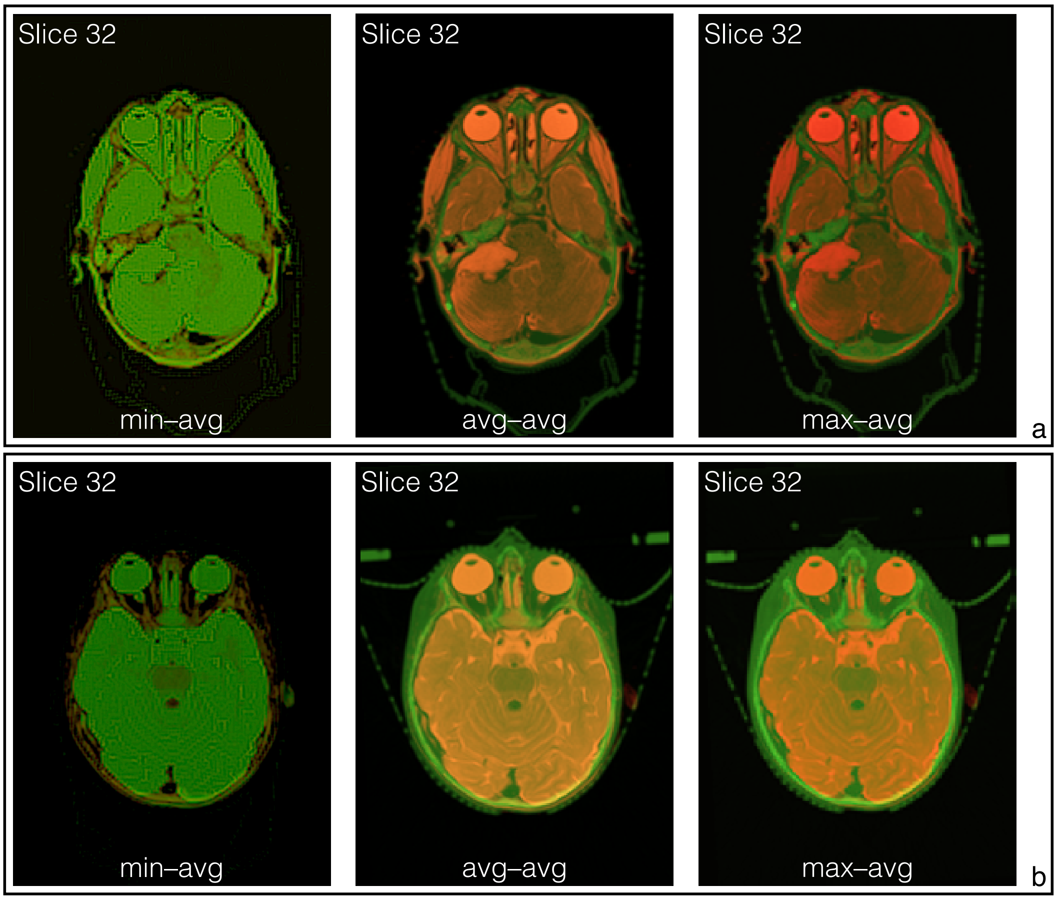


Figure 6.5: Fusion of multimodal images in the RGB colour model with the Haar wavelet function for three fusion rules. The information from the CT scans is depicted in a green colour scale, while red-coloured T_2-weighted MRI scans are used to indicate the impact. Subfigure a: The fused images are shown for Patient 5 of the 14-patient data set. Subfigure b: From the 25-patient data set, the images of Patient 11 are merged.

Qualitative evaluation Fused images generated with the min–avg, avg–avg and max–avg rules are shown in Figure 6.5 for Patients 5 and 11 of the 14- and 25-patient data sets, respectively. To obtain a visual distinctness, the CT scans are converted from greyscale to a green colour scale, while red-coloured T_2-weighted MRI scans are used for the fusion. The quality of the fused images that are based on the minimum rule for the approximated images is decreased compared to the average and maximum rules. The fused images with the minimum rule exclude the bone structures and contain green noise. High contrast is obtained with the max–avg rule. In these fused images, the bone structures of the CT scans stick out, but suppress information from the T_2-weighted MRI scans for these coordinates. The selection of the largest values triggers low transparency, which can be increased with the avg–avg rule. These fused images have the highest symmetry according to the fraction of information of both input images, as indicated by the fusion-symmetry metric. Thus, the avg–avg rule improves the possibility to compare structures at the same coordinates due to a higher transparency, enabling its use in treatment planning. Ultimately, multimodal image fusion with the discrete wavelet transform enables the images to be split into different parts, which are individually combined with specific fusion rules. The avg–avg rule produces fused images with highly symmetric information content, as quantitatively and qualitatively evaluated.

6.2.3 Unimodal image fusion

Image fusion is applied to the registered T_1- and T_2-weighted MRI scans from Section 5.3.2, which includes the images of the 14-patient data set only. The 14 image pairs, which result from the CNN configuration with the small architecture and the parameter settings $b = 1$, $\lambda = 0.1$ and $\alpha = 10^{-4}$, are merged with the Haar, bior1.3 or db2 functions. All nine combinations of the fusion rules are investigated with the same evaluation techniques as for the multimodal image fusion in Section 6.2.2.

Quantitative evaluation The fusion-factor and fusion-symmetry metrics are determined for the fused images of the 14 patients. The distribution of the mean values and their uncertainties, depending on the wavelet function and the fusion rule, are presented in Figure 6.6. Interestingly, the highest information content in the fused images is obtained for maximum rules according to the fusion-factor metric. The largest value is determined for the db2 function and the max–avg rule with $m_{\mathrm{FF}}(I, J, I_F) = 2.37 \pm 0.15$. This could be caused by the similar pixel-value distributions (see Figure 4.6) between both MRI scans. The choice of the largest coefficient has a smaller effect if the pixel values are very similar. Therefore, the information content is slightly decreased for the average rules. However, the fusion-symmetry metric indicates the most symmetric distribution of information for the average rules, like in the multimodal case. The values of $m_{\mathrm{FS}}(I, J, I_F)$ are similar for the wavelet functions, e.g. $m_{\mathrm{FS}}(I, J, I_F) = 0.021 \pm 0.020$ is yielded for the bior1.3 function with the avg–max rule.

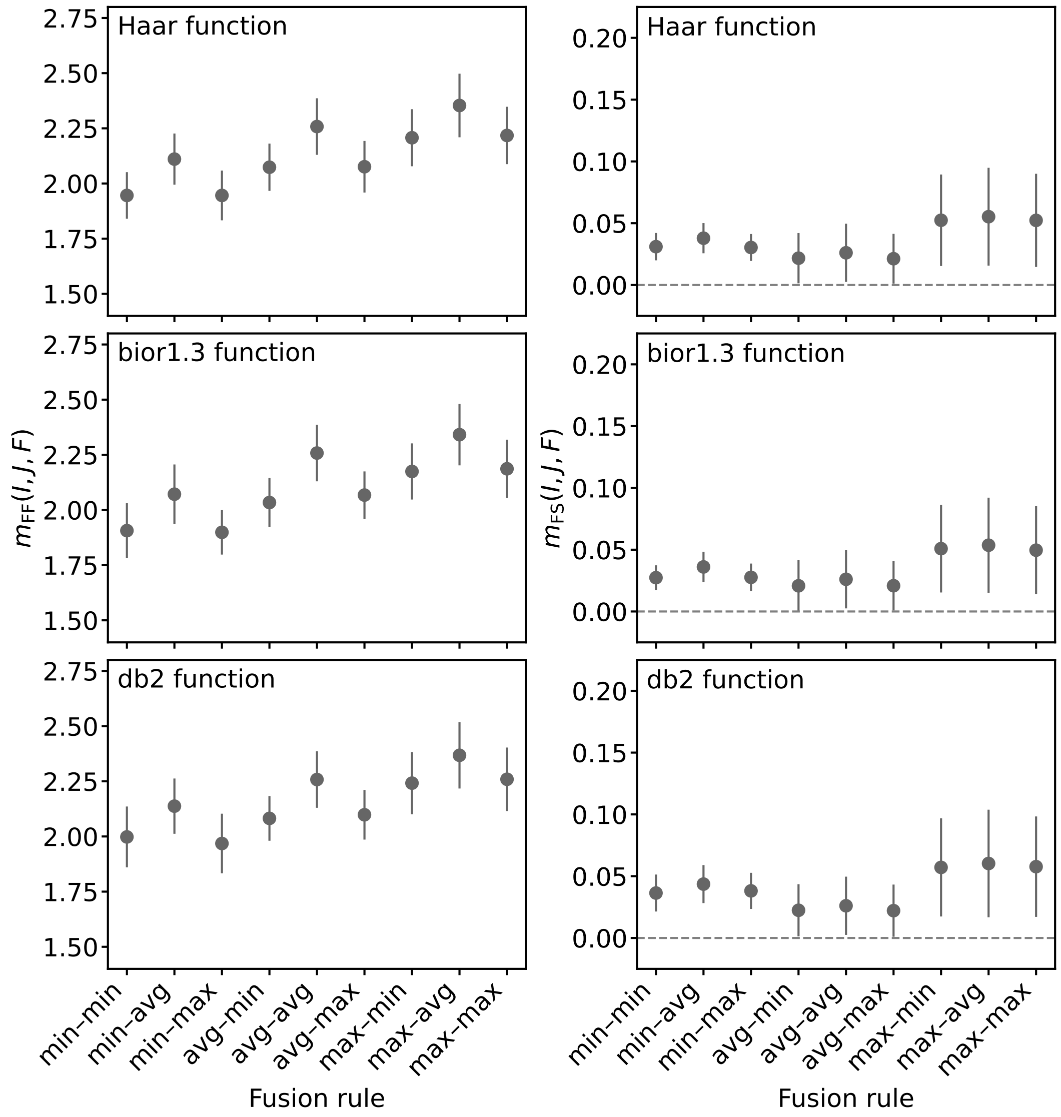


Figure 6.6: Fusion quality of the fused T_1- and T_2-weighted MRI scans for the patients of the 14-patient data set. The fusion-factor (left column) and the fusion-symmetry (right column) metrics are shown for the Haar (top row), bior1.3 (middle row) and db2 (bottom row) functions. The metrics are determined as the mean values and their uncertainties of all patients.

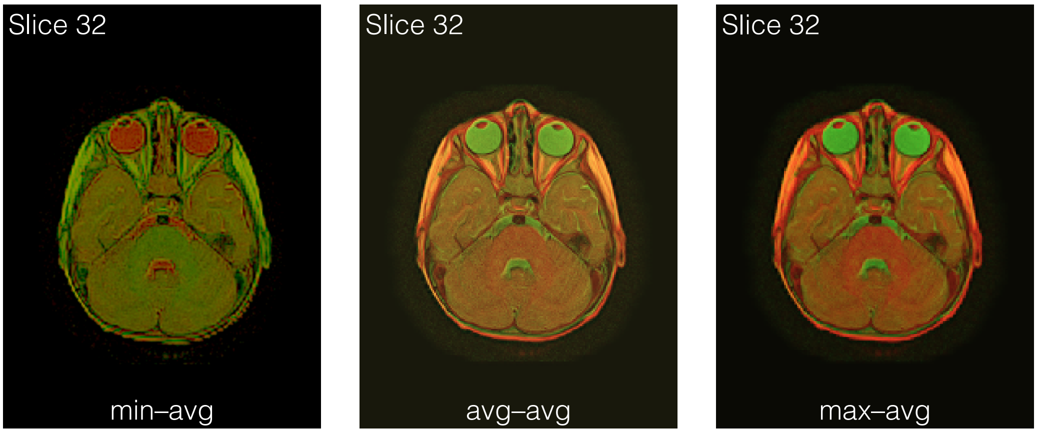


Figure 6.7: Fusion of unimodal images in the RGB colour model with the Haar wavelet function for three fusion rules. The green-coloured T_2-weighted MRI scan is merged with the T_1-weighted MRI scan in a red colour scale. The fused images are shown for Patient 2 of the 14-patient data set.

Qualitative evaluation The fused images of Patient 2 of the 14-patient data set are depicted in Figure 6.7 for the min–avg, avg–avg and max–avg rules. The Haar function is used for image decomposition with the discrete wavelet transform, and the T_1- and T_2-weighted MRI scans are converted to red and green colour scales, respectively. Visual effects similar to the multimodal fused images are noticeable (see Figure 6.5). The noisy image of the min–avg rule mostly contains the information from the T_2-weighted MRI scan, while high contrast is apparent for the max–avg rule. Furthermore, the avg–avg rule produces a fused image with a higher transparency, which increases the possibility to distinguish between both MRI scans. Ultimately, the choice of the fusion rule for image fusion of T_1- and T_2-weighted MRI scans depends on the user and the application of the images.

7 Potential for clinical integration

The research presented in this thesis is subdivided into three parts, which can be combined to form a single workflow. The outcome of this workflow, operating fast and unsupervised, has the potential to enhance precision radiotherapy. The application of this workflow allows radiotherapy planning to be facilitated with more precise registered images or medical diagnostics with fused images. The multimodal investigations are particularly interesting since research on this topic is barely present. Therefore, this chapter deals with the potential integration of the workflow in clinics. In Section 7.1, the individual subworkflows are recapitulated with regard to the operation and optimisation possibilities. Then, radiotherapy and treatment-planning systems that are commonly used in practice are described in Section 7.2. Lastly, suggestions for the integration of the workflow in the clinical setting are presented in Section 7.3.

7.1 Operation of the workflow

Within the scope of this thesis, several aspects of image registration and fusion are investigated for head CT and MRI scans. The focus lies on multimodal image pairs, but unimodal registration and fusion of T_1- and T_2-weighted MRI scans are also evaluated. The result is a two-part image-registration workflow, which seeks to improve image alignment with rigid and deep-learning-based deformable registrations. In addition, the workflow can be extended by image fusion, providing a single image with combined information from two separate images. The advantage of this all-in-one procedure is the direct and unsupervised operation without external markers or atlas images. Moreover, the workflow realised with the SNAKEMAKE [75] tool takes between three and four minutes to produce the desired images, starting from the initial DICOM files.

Preprocessing The main task of the preprocessing is the rigid registration of source and target images. The preprocessing workflow (see Section 4.4.1) is designed to allow the operator the choice of the source and target images, e.g. T_2-to-CT or T_1-to-T_2 registration. Prior to that, the images passed as input are individually modified to equalise pixel and slice properties within an image pair. A slice thickness and a pixel spacing of 1 mm are set as default, but these values can be varied by the operator. Moreover, automatic image segmentation is part of the subworkflow, supporting the rigid registration. One algorithm is implemented to generate segments of the eyes; another is developed to convert manually outlined structures into segments. In this thesis, contours of the ventricular system (see Section 4.2.2) are available, but any other contour is applicable. As the process of manually creating contours is elaborate, segments generated automatically with deep learning should be considered in future work, especially for multimodal applications. There exists usable software, like FreeSurfer [76], whose advantages are restricted to the preprocessing and segmentation of MRI scans only. This requires more detailed research on the proposed preprocessing subworkflow in terms of image-alignment precision and run time. For the improvement in image alignment, the rigid registration can be optimised by adding more segments for the determination of the image overlap between source and target images. Furthermore, the execution on a GPU instead of on a CPU is expected to drastically decrease the run time.

Deformable registration The focus of the all-in-one workflow is the deep-learning-based deformable registration of source and target images. The pure application of a trained model for deformable registration takes approximately 40 s, including the computation of the quantitative measures. The use of deep-learning-based techniques is challenged by the number of images for the training of an appropriate registration model. This model should be able to register images that are not included in the training process, but are of the same type. In this thesis, a small-sized data set of CT and MRI scans of 39 pediatric patients with brain tumours is used. After the parameter tuning of the CNN is performed for DIR of the CT and MRI scans, data augmentation, which improves the registration accuracy (see Section 5.3.1), is investigated with the optimal parameter settings. Therefore, the potential of fast and direct multimodal DIR should be exploited in the future by increasing the amount of real data. This is crucial because the physical growth or changes in the brain morphology require image-registration methods that are adaptable. Deep neural networks, for example, are beneficial for facilitating treatment planning in image-guided radiotherapy or adaptive radiotherapy. Furthermore, the CNN used in this thesis reflects only one type of deep neural networks. A variety of algorithms, developed and evaluated on unimodal image pairs, provide the potential for further studies of multimodal DIR.

Fusion The last step of the workflow generates fused images from the registered images, which are obtained in the deep-learning workflow. For the fusion, the wavelet transform is applied to separate vertical, horizontal and diagonal details from the images. This increases the variability of fusion combinations, which means that the operator is able to set specific wavelet function and fusion rules. The quality of the fused images and the balance of information from the input images depend on the setting. Since the run time of this subworkflow is approximately 16 s on average, several variants can be generated for comparison. The application of the fused images in clinical practice is more likely the visual support by comparing information from CT and MRI at once. This can simplify the assessment of risk structures, organs and healthy tissue.

7.2 Radiotherapy

The treatment technique related to high-energy irradiation of tumour tissue is one branch of cancer treatment besides surgery and chemotherapy. Radiotherapy used individually or in combination with the other techniques manages to treat various types of cancer. Improvements in computing power have produced a variety of radiotherapy methods that deliver treatments with high precision. One requirement for radiotherapy is a mapping of the patient's body, which is performed with medical imaging techniques, to plan the treatment with appropriate dose calculation. [1]

Medical imaging For the treatment with radiation, medical imaging is an essential instrument regarding the planning, the delivery and the monitoring of radiotherapy. Imaging modalities like CT, MRI and PET are employed for the acquisition of high-resolution images. These are used to precisely delineate tumour and surrounding healthy tissues, allowing radiation oncologists to design patient-specific treatment plans. The information obtained from the images helps to determine the target volume, which has to be irradiated, and to define critical structures to be spared. Target localisation is performed to ensure the precise delivery of radiation. Cone-beam CT and MRI are typically used to verify and adjust patient positioning before the treatment session. These images enable the comparison of the current anatomy with the planning images, ensuring accurate alignment with the radiation beams. Real-time imaging modalities during treatment are useful to guide radiation delivery. Fluoroscopy, for example, can assist in image-guided radiotherapy, where continuous X-ray imaging is employed to track tumour motion and to adjust the treatment beams accordingly. Adaptive radiotherapy involves modified treatment plans, needed since the anatomy of the patient changes in between treatment sessions. This technique relies on CT or MRI scans that are regularly acquired at specific points in time during the entire treatment to monitor and assess anatomical changes of both tumour and healthy tissues. Furthermore, images of the anatomical structures support the evaluation of the treatment response after radiotherapy. Several imaging modalities can provide valuable information on tumour regression and potential treatment-related side effects, which helps to evaluate the treatment efficacy and to assess possible adjustments. [18]

Treatment planning The requirement for radiotherapy is to deliver sufficient dose of radiation to the tumour, while sparing healthy tissue. This goal is realised by customised treatment plans, which take the specific anatomy of the patient into account. Treatment planning consists of multiple steps, including medical imaging, segmentation and target-volume definition, choice of irradiation technique, dose calculation, evaluation and optimisation. The necessary information on the attenuation of radiation in the human body can be accessed from the CT to simulate the dose distribution. For this, several algorithms exist to model the setup in detail and to reproduce physical interactions of radiation with the patient. To select an algorithm, the balance between computation time and accuracy must be considered. Dose calculation requires target-volume definition by delineating relevant structures based on the planning images. This includes not only the actual target volume, but also healthy tissue and structures placed near the target volume, which can be harmed by irradiation. As the therapy with photons, electrons or particles like protons is based on distinct physical processes, the choice of the irradiation technique impacts the planning procedure. In general, treatment planning has become increasingly computerised due to advancements in hardware and software. [18] Treatment-planning systems like RayStation [77] from RaySearch Laboratories [78] allows images to be registered, target volumes to be defined and dose calculations to be performed. In clinical practice, the images of different modalities are typically superimposed with rigid registrations to outline the volumes [4]. However, the displacement of organs due to the immobilisation of the patient or tumour change during treatment is not considered by this registration type. In the case of pediatric patients, physical growth and a prolonged course of disease are further challenges for treatment planning. Therefore, DIR is expected to improve multimodal treatment-planning processes. RayStation, for example, is able to perform DIR with its anatomically constrained deformation algorithm [79], but multimodal DIR is barely supported. The method implemented in RayStation solves the registration problem with a non-linear optimisation procedure using contoured information [79]. The creation of contours is elaborate and requires human intervention. The unsupervised workflow presented in this thesis operates without external information, which can facilitate clinical processes with the methods described in the following section.

7.3 Integration

Quality assurance and quality control of image-registration software is of great importance for treatment planning and delivery in radiotherapy since registered and fused images are used as input in delineation processes. Therefore, the uncertainties caused by this software should be easily accessible for medical applications. However, documentation from commercial systems is not always available, which makes it difficult to validate the performance of the image-registration software. The clinical integration of such software should be conveniently achievable through clear instructions for the clinicians. [4]

Quality assessment The uncertainties arising from the image-registration software are associated with the input images and the present algorithms. The former aspect relates to possible distortions in the images, e.g. distortion effects of MRI [55–58] with radial degradation towards the outer regions of the body. The latter aspect refers to limitations of the algorithms and incorrect selection of their parameters. Thus, accurate assessment of registration uncertainties depends on the software-integrated tools and the operator interaction. The image-registration accuracy can be determined with several quantitative measures, e.g. marker or contour comparisons, Jacobian determinant or inverse-consistency check. [4] The Jacobian determinant and the inverse-consistency check (see Section 5.2.2) are already implemented in the developed workflow presented in this thesis to assess the performance of the registration. Furthermore, contour comparisons are realised through the overlap of segments, which is determined with the Dice similarity coefficient, defined in Equation (3.9). The procedure of calculating m_{DSC} can be extended to any segment that the operator would like to outline during treatment planning. These quantitative metrics are easily accessible and can contribute to the assessment of the workflow performance. In addition, qualitative evaluation should provide visual assessment of the registration results by combining the registered images [4]. Several methods, like split screen, checkerboard or image overlay displays, support the verification of the registration accuracy [4]. While image overlay displays are employed in this thesis, a more complex image-fusion method is investigated (see Chapter 6), providing high variability. This adaptability is advantageous as the fused image can be generated to satisfy the requirements of the operator.

Commissioning Before the clinical use of the software, the image quality resulting from the registration processes needs to be validated. The related checks should cover various aspects, such as accuracy of image deformation and the system functionality. Quality assurance aims at accurate registration of the images that are used in the course of radiotherapy by taking uncertainties into account and avoiding spatial distortions. It should also ensure consistency and accuracy in propagating patient treatment-planning data across all image data sets used for plan creation. Although quality assurance, whose architecture depends on the individual goal of the software, is not always feasible on all systems, specific methods are recommended for commissioning. One aspect is that the registration techniques are reproducible before the clinical integration. The recommendations for commissioning include three methods, dealing with physical-phantom, digital-phantom and clinical-data tests to assess the registration results. A comprehensive commissioning process should include all three methods. Physical-phantom tests ensure accurate data representation and integrity verification across imaging and radiotherapy systems. Digital phantoms allow controlled testing of registration accuracy, and clinical data tests provide final validation using examples of images expected in clinical use. [4] To substantiate the registration results achieved in this thesis, images of both physical and digital phantoms must be used for a justified commissioning of the workflow.

Application Image registration is characterised by the fact that a mapping of an image pair is generated, allowing various applications, such as target-volume delineation and image fusion. By registering a source image to match a target image, the geometric transformations in the form of deformation vector fields create a link between these images. Structure mapping is one approach for the clinical application in treatment planning since the information is accessible and clearly defined. This method is performed by delineating tumours and healthy tissue, for example on MRI scans, which can be easily transferred to the CT scan. Thus, treatment planning can directly profit from the high soft-tissue distinctness. Another suggestion for the use of registered images is related to the dose-calculation procedure during the planning. The clinician can compute the dose distributions on the CT scan, which are then mapped and displayed onto the registered MRI scan. The visualisation of dose information on different modalities can indicate errors, providing evidence for possible adjustments on the dose calculation. Furthermore, the spatial integrity of registered images allows the information from both images to be combined. Several procedures for image fusion exist, like image overlay or checkerboard displays. [4] The wavelet-based fusion evaluated in this thesis is an additional method for clinical application, which benefits from various adjustment options regarding the fusion rules. In summary, the output of the proposed workflow is able to facilitate the treatment planning in clinical processes. Although the proposed workflow is an unsupervised method, its application still requires supervision by trained clinicians, especially for quality assessment.

8 Conclusion

Medical imaging modalities provide different types of anatomical information due to their physical concepts and contrast characteristics. CT is commonly used for treatment planning in radiotherapy due to its bone contrast and dose-calculation capabilities, whereas MRI offers high soft-tissue distinctness. Combining the benefits of both modalities can improve the accuracy of treatment plans by precisely delineating target volumes, like tumours and healthy tissue. [1] Rigid registration is a common practice to superimpose images from different modalities, but it does not take deformations between treatment sessions or distortion effects into account [1, 4]. DIR with deep-learning techniques can enhance multimodal treatment planning, but there is a lack of direct unsupervised multimodal methods for deep-learning-based DIR. This thesis faces the challenges of multimodal registration problems and provides a complete workflow for image registration and fusion of head CT and MRI scans. The data that are used to develop the workflow consist of the scans of 39 pediatric patients with tumours in the head.

The preprocessing is the first step of the workflow, generating rigidly aligned images. At the beginning, the DICOM files are used to collect the required information of each scan. This includes the image intensities for the creation of a three-dimensional array, representing stacked axial slices. Moreover, slice properties like the thickness and the distance as well as the pixel spacing of the axial slices are necessary in the image-adjustment process. Before the images are adjusted, the subworkflow performs image segmentation with an automated algorithm to produce segments of both eyes. These two eye segments are required for the translation and rotation operations during the rigid registration. Another algorithm for segmentation is applied if manually outlined structures are available. This algorithm converts the contours to segments, which can be added to the segmented image with eye segments. Each scan of the data is individually acquired to obtain high-quality images, but this causes a difference in image properties. The pixel spacing and the slice thickness are used to scale the images, providing similar representations of the object regarding the size. Afterwards, the rigid registration is performed to align the images with translation and rotation by considering large overlap of the segments, determined with the Dice similarity coefficient as the image-similarity metric. Data normalisation is required for deep-learning techniques. Therefore, the rigidly aligned images, produced by the preprocessing workflow, are prepared for DIR.

The workflow continues with the deformable registration of target and source images. The preprocessed images can be directly passed to the deep neural network to produce deformably aligned images within seconds. Although it sounds simple, there are many challenges to obtain an appropriate configuration of the deep neural network. In this thesis, a CNN [17, 47] with U-Net [16] architecture is employed to solve the multimodal registration problem with diffeomorphic transformations. The CNN is a complex construct containing several parameters and operations, which can be varied and optimised. Therefore, the main part of investigations regarding deep-learning-based DIR is the parameter tuning. Modified input images and typical deep-learning techniques, like the inclusion of dropout layers or the variation of the optimiser function, are checked. The architecture of the CNN as well as parameters regarding the loss function and the weight optimisation are varied in an extensive study, where many registration models are trained. The parameter tuning provides the configuration of the CNN that produces the most precise deformations according to quantitative and qualitative measures. Furthermore, the aspect of the amount of data is important for deep neural networks. By increasing the number of input images, the quality of the model improves due to more precise prediction of the deformations. This thesis investigates the impact of larger data sets in the network training by using data augmentation. Factors of up to 20 generate 780 image pairs, originating from the 39 original scans. The results indicate a slight improvement in the registration accuracy with higher augmentation factors. Besides the multimodal application, unimodal registration is investigated with the CNN. Lower values of the regularisation parameter achieve more precise deformations.

The last part of the workflow deals with the image fusion of the registered images. Typical fusion methods, like image overlay or checkerboard displays, are used in this thesis for qualitative evaluation. In addition, the wavelet-based method is investigated to check for differences compared to the simple applications. The wavelet transform has the ability to decompose the images into several components, which can represent vertical, horizontal or diagonal details of the images. The fusion takes place in that domain by combining the pixel values with a specific fusion rule, e.g. minimum, average or maximum. The average rule yields quantitatively and qualitatively the most homogeneous fused images, while other combinations of the fusion rules provide an asymmetric distribution of information. The results obtained from these investigations are similar for CT–T_2 and T_2–T_1 fused images.

In this thesis, the data sets contain scans of patients under the age of 18, which leads to differences in the head shape. The variation is useful for DIR with deep learning. As the shape and morphology of the head vary from age to age, especially for children, scans of each age group provide more information, such as anatomical structures and intensity distributions, for the deep neural network. Future work should deal with the training of a registration model based on a larger training sample, consisting of more individual data. This should lead to a further generalisation of the applicability of the model, as shown with data augmentation. Multimodal DIR has the potential to support treatment planning, but more investigations have to be performed in the future. One option is to vary the similarity part of the loss function by using other metrics, e.g. the Dice similarity coefficient. The latter metric requires segmented images, which should contain numerous segments to cover the majority of anatomical features. Ultimately, the significance of this thesis is highlighted by the lack of application-related research in the field of multimodal DIR. Many previous studies dealt with image processing using one modality, while the feasibility of direct multimodal DIR with deep learning is pointed out in this thesis. The proposed workflow combines rigid and deformable registrations as well as image fusion for fast and direct application. The results presented here enhance image alignment and pave the way for further studies on the training of a robust registration model with an extension to other locations and adult patients. The registration improvements made by deep-learning-based DIR are a benefit for more precise delineation of tumours and healthy tissue in radiotherapy treatment planning.

A Additional information

A.1 Details of the data sets

Information on the CT scans and the T_1- and T_2-weighted MRI scans from the 14- and 25-patient data sets is presented. The formats of the three-dimensional images are listed in Tables A.1 and A.2. The pixel spacing of each scan is shown in Figure A.1. In addition, the slice thickness and the slice distance of each scan are visualised in Figure A.2.

Table A.1: Image format of the CT scans and the T_1- and T_2-weighted MRI scans of the 14-patient data set. The number of slices, N_z, and the number of pixels, N_x and N_y, in the axial image, x–y plane, are listed for each scan.

Patient number	Dimensions ($N_z \times N_y \times N_x$) of the scans		
	CT	T_1 weighting	T_2 weighting
1	364 × 600 × 600	156 × 584 × 928	95 × 292 × 256
2	292 × 512 × 512	250 × 576 × 576	108 × 576 × 576
3	266 × 512 × 512	186 × 544 × 544	82 × 576 × 576
4	330 × 512 × 512	236 × 544 × 544	95 × 600 × 512
5	260 × 512 × 512	192 × 576 × 576	75 × 576 × 576
6	262 × 512 × 512	202 × 576 × 576	101 × 576 × 576
7	338 × 512 × 512	170 × 584 × 928	95 × 292 × 256
8	268 × 512 × 512	191 × 576 × 576	71 × 576 × 576
9	294 × 512 × 512	218 × 576 × 576	80 × 576 × 576
10	258 × 512 × 512	192 × 576 × 576	90 × 576 × 576
11	307 × 512 × 512	198 × 420 × 420	75 × 448 × 448
12	274 × 512 × 512	194 × 544 × 544	75 × 600 × 512
13	304 × 512 × 512	192 × 576 × 576	68 × 576 × 576
14	327 × 512 × 512	156 × 544 × 544	67 × 600 × 512

Table A.2: Image format of the CT scans and the T_1- and T_2-weighted MRI scans of the 25-patient data set. The number of slices, N_z, and the number of pixels, N_x and N_y, in the axial image, x–y plane, are listed for each scan.

Patient number	Dimensions ($N_z \times N_y \times N_x$) of the scans		
	CT	T_1 weighting	T_2 weighting
1	333 × 512 × 512	192 × 512 × 512	74 × 512 × 512
2	363 × 512 × 512	232 × 512 × 512	44 × 512 × 512
3	329 × 512 × 512	212 × 552 × 444	73 × 512 × 512
4	279 × 512 × 512	176 × 512 × 512	52 × 576 × 576
5	285 × 512 × 512	98 × 524 × 436	50 × 256 × 230
6	245 × 512 × 512	96 × 480 × 480	64 × 448 × 448
7	323 × 512 × 512	192 × 480 × 480	33 × 512 × 512
8	301 × 512 × 512	168 × 480 × 480	33 × 512 × 512
9	301 × 512 × 512	204 × 576 × 576	82 × 576 × 576
10	303 × 512 × 512	192 × 480 × 480	47 × 512 × 512
11	369 × 512 × 512	198 × 480 × 480	56 × 512 × 512
12	303 × 512 × 512	160 × 512 × 512	39 × 512 × 512
13	301 × 512 × 512	194 × 544 × 544	73 × 600 × 512
14	279 × 512 × 512	152 × 512 × 512	33 × 512 × 512
15	303 × 512 × 512	192 × 480 × 480	60 × 576 × 576
16	303 × 512 × 512	194 × 544 × 544	50 × 600 × 512
17	359 × 512 × 512	192 × 576 × 576	66 × 576 × 576
18	279 × 512 × 512	182 × 576 × 576	66 × 576 × 576
19	363 × 512 × 512	198 × 544 × 544	74 × 600 × 512
20	303 × 512 × 512	194 × 544 × 544	78 × 600 × 512
21	265 × 512 × 512	196 × 576 × 576	60 × 576 × 576
22	333 × 512 × 512	194 × 544 × 544	73 × 600 × 512
23	305 × 512 × 512	218 × 544 × 544	94 × 600 × 512
24	279 × 512 × 512	178 × 576 × 576	60 × 576 × 576
25	325 × 512 × 512	178 × 544 × 544	73 × 600 × 512

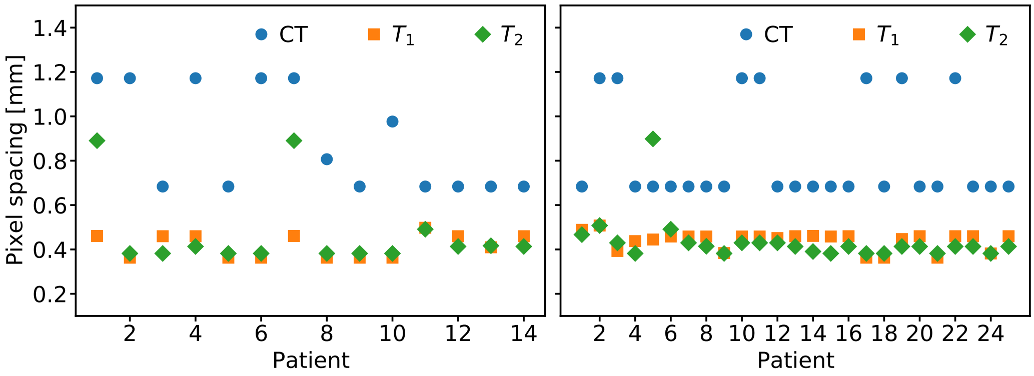


Figure A.1: Pixel spacing of the CT and MRI scans of each patient. The values of x_{spacing} and y_{spacing}, shown for the 14-patient (left) and 25-patient (right) data sets, are equivalent.

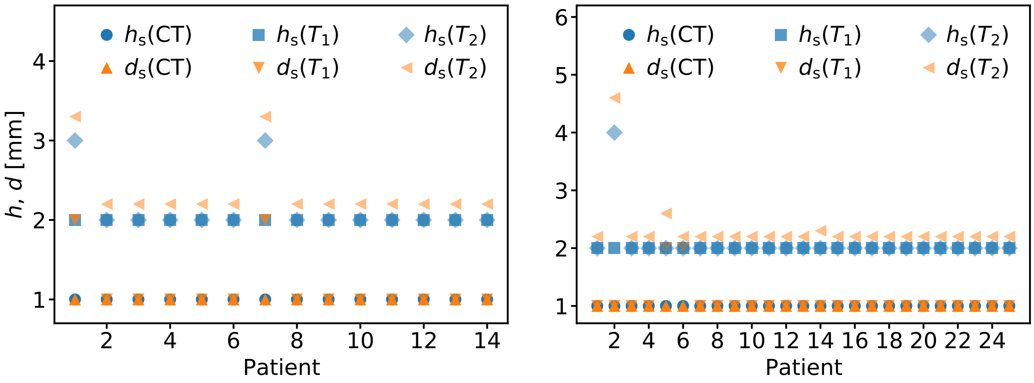


Figure A.2: Slice properties of the CT and MRI scans of each patient. The slice thickness, h_s, and the slice distance, d_s, are shown for the 14-patient (left) and 25-patient (right) data sets.

A.2 Default configuration of the neural network

The default configuration of the CNN from Section 5.1 is used as reference for the parameter tuning in Section 5.2. The settings are given in Table A.3.

Table A.3: Default configuration of the CNN regarding architecture and network parameters.

Parameter	Variable	Setting
	CNN architecture	
Number of feature maps	n_f(enc)	$[16 - 32 - 32 - 32]$
Number of feature maps	n_f(dec)	$[32 - 32 - 32 - 32 - 32 - 16 - 16]$
	Network training	
Optimiser function	—	Adam algorithm
Learning rate	α	10^{-4}
Batch size	b	1
Regularisation parameter	λ	1
	Network operations	
Convolution kernel	—	$3 \times 3 \times 3$
Pooling grid	—	$2 \times 2 \times 2$
Upsampling grid	—	$2 \times 2 \times 2$

A.3 Comparison of loss functions

The normalised cross-correlation and the mutual information, which are tested as $\mathscr{L}_{\text{sim}}(T, D)$, are compared in Figure A.3.

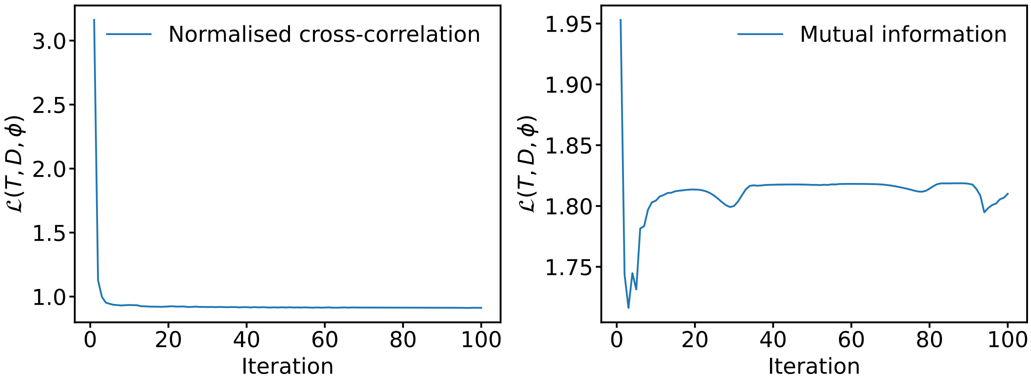


Figure A.3: Distributions of the loss function for two different metrics. The results of the normalised cross-correlation (left) and the mutual information (right) are shown.

A.4 Further results of the extensive parameter tuning

The evaluation results regarding registration accuracy and performance are shown in Figures A.4, A.5 and A.6. In addition, the values of the mutual-information metric are listed in Tables A.4 and A.5.

Table A.4: Mutual-information metric for parameter settings of the extensive study with the batch size $b = 1$.

α	$\lambda = 0.01$	$\lambda = 0.05$	$\lambda = 0.1$	$\lambda = 1$	$\lambda = 2$
Training with small-architecture model					
10^{-5}	-0.09 ± 0.02	-0.05 ± 0.01	-0.01 ± 0.03	0.06 ± 0.04	0.08 ± 0.05
10^{-4}	-0.12 ± 0.02	-0.05 ± 0.02	0.01 ± 0.03	0.09 ± 0.04	0.12 ± 0.05
10^{-3}	-0.08 ± 0.02	-0.05 ± 0.02	-0.03 ± 0.02	0.03 ± 0.02	0.04 ± 0.02
Training with medium-architecture model					
10^{-5}	-0.08 ± 0.02	-0.05 ± 0.01	-0.02 ± 0.02	0.04 ± 0.03	0.05 ± 0.03
10^{-4}	-0.10 ± 0.02	-0.04 ± 0.02	0.00 ± 0.03	0.06 ± 0.04	0.07 ± 0.05
10^{-3}	-0.10 ± 0.02	-0.05 ± 0.02	-0.04 ± 0.01	0.02 ± 0.01	0.02 ± 0.01
Training with large-architecture model					
10^{-5}	-0.09 ± 0.02	-0.05 ± 0.02	-0.02 ± 0.03	0.05 ± 0.04	0.07 ± 0.04
10^{-4}	-0.11 ± 0.02	-0.05 ± 0.02	-0.04 ± 0.03	0.09 ± 0.04	0.11 ± 0.05
10^{-3}	-0.09 ± 0.02	-0.06 ± 0.02	-0.04 ± 0.02	0.03 ± 0.02	0.04 ± 0.02
Validation with small-architecture model					
10^{-5}	-0.09 ± 0.02	-0.05 ± 0.01	-0.01 ± 0.02	0.06 ± 0.03	0.07 ± 0.03
10^{-4}	-0.10 ± 0.02	-0.05 ± 0.01	-0.01 ± 0.02	0.07 ± 0.03	0.10 ± 0.04
10^{-3}	-0.08 ± 0.02	-0.06 ± 0.01	-0.04 ± 0.02	0.03 ± 0.02	0.03 ± 0.02
Validation with medium-architecture model					
10^{-5}	-0.08 ± 0.02	-0.05 ± 0.20	-0.02 ± 0.02	0.04 ± 0.02	0.04 ± 0.02
10^{-4}	-0.09 ± 0.02	-0.04 ± 0.01	-0.01 ± 0.02	0.05 ± 0.02	0.06 ± 0.03
10^{-3}	-0.09 ± 0.02	-0.05 ± 0.02	-0.04 ± 0.01	0.01 ± 0.01	0.02 ± 0.01
Validation with large-architecture model					
10^{-5}	-0.09 ± 0.02	-0.05 ± 0.01	-0.02 ± 0.02	0.05 ± 0.03	0.06 ± 0.03
10^{-4}	-0.10 ± 0.02	-0.05 ± 0.02	0.00 ± 0.02	0.08 ± 0.03	0.09 ± 0.03
10^{-3}	-0.08 ± 0.02	-0.06 ± 0.02	-0.04 ± 0.01	0.03 ± 0.02	0.03 ± 0.02

Table A.5: Mutual-information metric for parameter settings of the extensive study with the batch size $b = 4$.

α	$\lambda = 0.01$	$\lambda = 0.05$	$\lambda = 0.1$	$\lambda = 1$	$\lambda = 2$
Training with small-architecture model					
10^{-5}	-0.09 ± 0.02	-0.05 ± 0.01	-0.01 ± 0.03	0.04 ± 0.04	0.06 ± 0.04
10^{-4}	-0.11 ± 0.02	-0.05 ± 0.02	0.02 ± 0.03	0.08 ± 0.04	0.11 ± 0.04
10^{-3}	-0.08 ± 0.02	-0.02 ± 0.02	0.01 ± 0.03	0.03 ± 0.02	0.04 ± 0.02
Training with medium-architecture model					
10^{-5}	-0.08 ± 0.02	-0.05 ± 0.02	-0.03 ± 0.02	0.03 ± 0.02	0.04 ± 0.02
10^{-4}	-0.09 ± 0.02	-0.04 ± 0.02	0.00 ± 0.03	0.05 ± 0.04	0.05 ± 0.04
10^{-3}	-0.60 ± 0.02	-0.05 ± 0.02	-0.04 ± 0.01	0.01 ± 0.01	0.02 ± 0.01
Training with large-architecture model					
10^{-5}	-0.09 ± 0.02	-0.06 ± 0.01	-0.03 ± 0.03	0.04 ± 0.03	0.06 ± 0.03
10^{-4}	-0.11 ± 0.02	-0.05 ± 0.01	0.00 ± 0.04	0.06 ± 0.04	0.07 ± 0.05
10^{-3}	-0.11 ± 0.03	-0.06 ± 0.01	-0.04 ± 0.01	0.02 ± 0.02	0.02 ± 0.02
Validation with small-architecture model					
10^{-5}	-0.09 ± 0.02	-0.05 ± 0.01	-0.03 ± 0.02	0.04 ± 0.03	0.06 ± 0.03
10^{-4}	-0.09 ± 0.02	-0.06 ± 0.01	0.00 ± 0.02	0.07 ± 0.03	0.08 ± 0.04
10^{-3}	-0.08 ± 0.02	-0.05 ± 0.02	-0.02 ± 0.02	0.02 ± 0.02	0.03 ± 0.02
Validation with medium-architecture model					
10^{-5}	-0.08 ± 0.02	-0.05 ± 0.02	-0.03 ± 0.02	0.02 ± 0.02	0.03 ± 0.02
10^{-4}	-0.09 ± 0.02	-0.05 ± 0.01	-0.02 ± 0.02	0.04 ± 0.03	0.04 ± 0.03
10^{-3}	-0.09 ± 0.02	-0.05 ± 0.02	-0.04 ± 0.01	0.01 ± 0.01	0.01 ± 0.01
Validation with large-architecture model					
10^{-5}	-0.09 ± 0.02	-0.06 ± 0.01	-0.03 ± 0.02	0.04 ± 0.02	0.05 ± 0.02
10^{-4}	-0.09 ± 0.02	-0.05 ± 0.01	-0.01 ± 0.03	0.06 ± 0.03	0.06 ± 0.04
10^{-3}	-0.10 ± 0.02	-0.06 ± 0.02	-0.05 ± 0.01	0.02 ± 0.01	0.02 ± 0.02

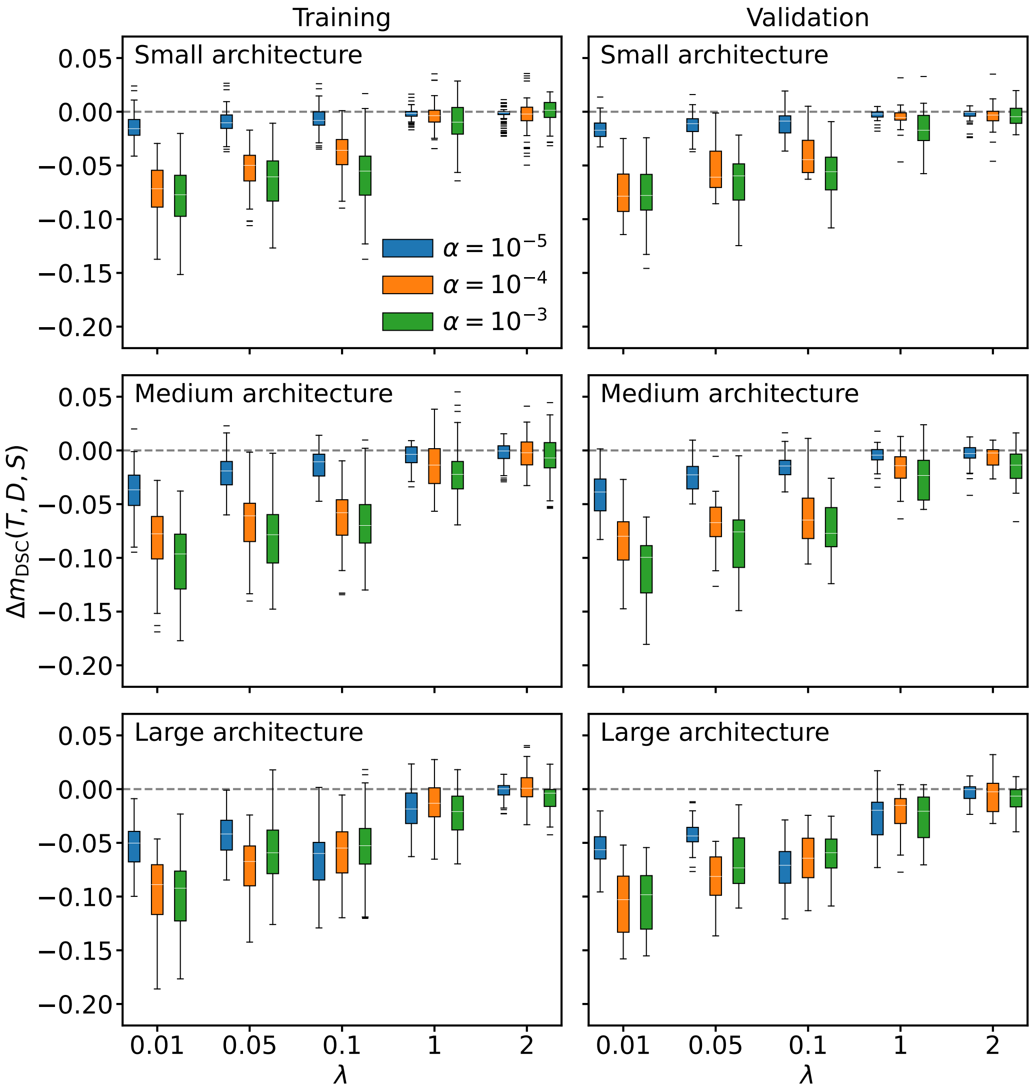


Figure A.4: Registration accuracy based on the Dice similarity coefficient for models with the batch size $b = 4$ using the 25-patient data set. The value of $\Delta m_{\mathrm{DSC}}(T, D, S)$ is measured as the change in the overlap of segments before, S, and after, D, deformation with regard to the target image, T. The results are shown for the training (left) and validation (right) data as well as for the small-architecture (top), medium-architecture (middle) and large-architecture (bottom) models. Each plot contains the variations of λ and α, regulating the smoothness of the deformations and the step size of the optimiser, respectively. The white lines inside the boxes represent the median values. The dashed lines indicate the border for the increase (positive values) or decrease (negative values) in image alignment.

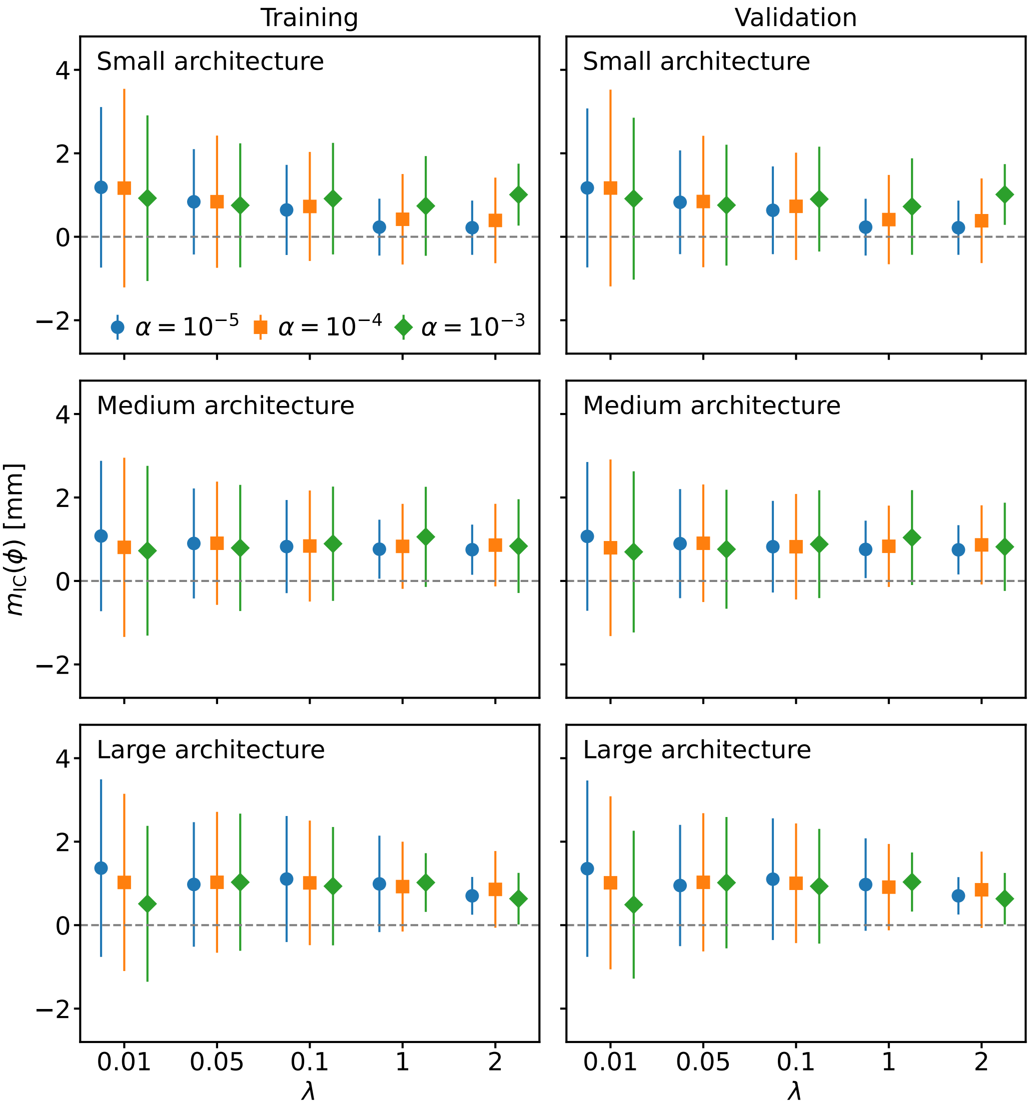


Figure A.5: Registration performance based on the inverse-consistency method for models with the batch size $b = 4$ using the 25-patient data set. The sum of the MRI-to-CT and the CT-to-MRI deformation vector fields leads to individual values for each pixel, which explains the large error bars. The results are shown for the small-architecture (top), medium-architecture (middle) and large-architecture (bottom) models. The mean values and uncertainties of the respective five-fold data are presented for the variation of the regularisation parameter λ and the learning rate α, separated into training (left) and validation (right). The dashed lines indicate the expected value.

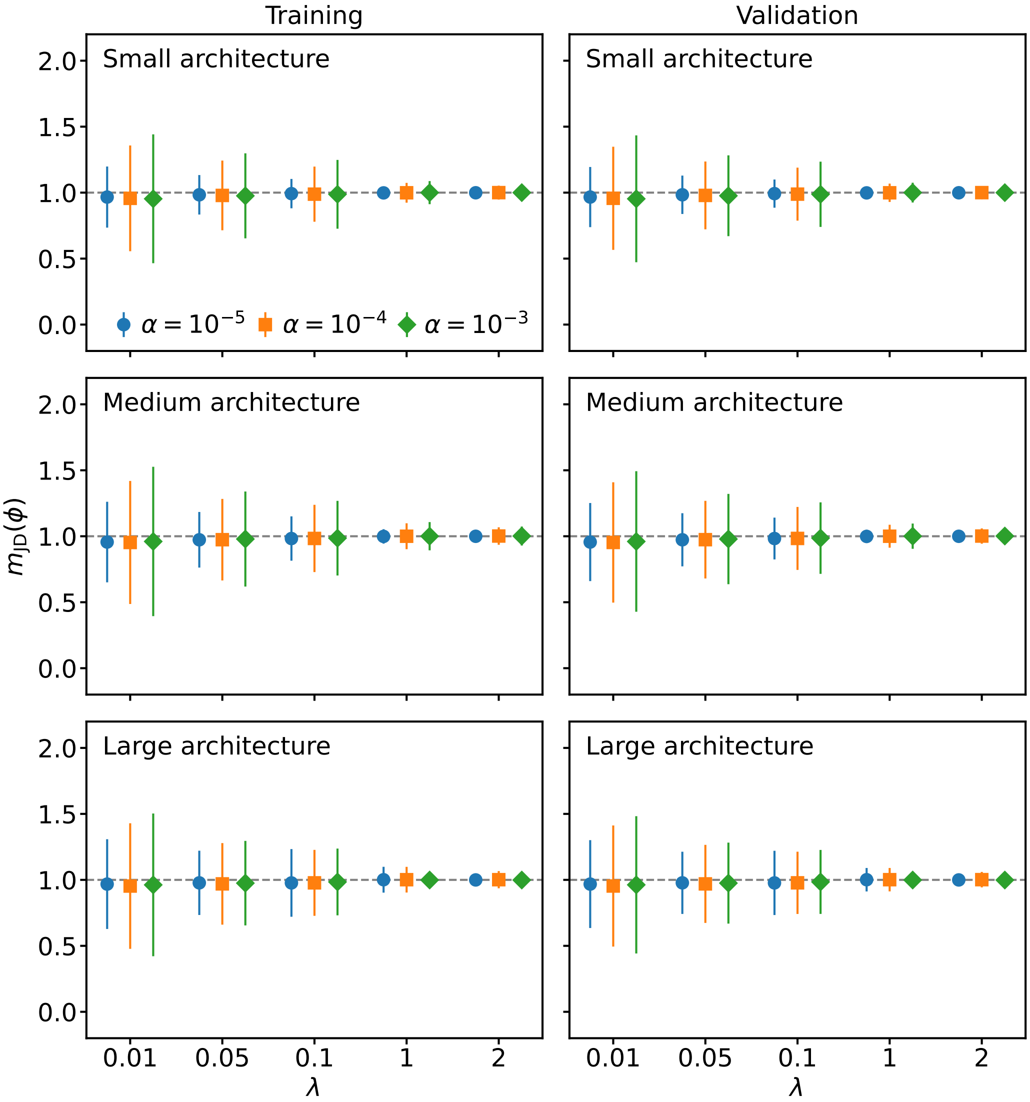
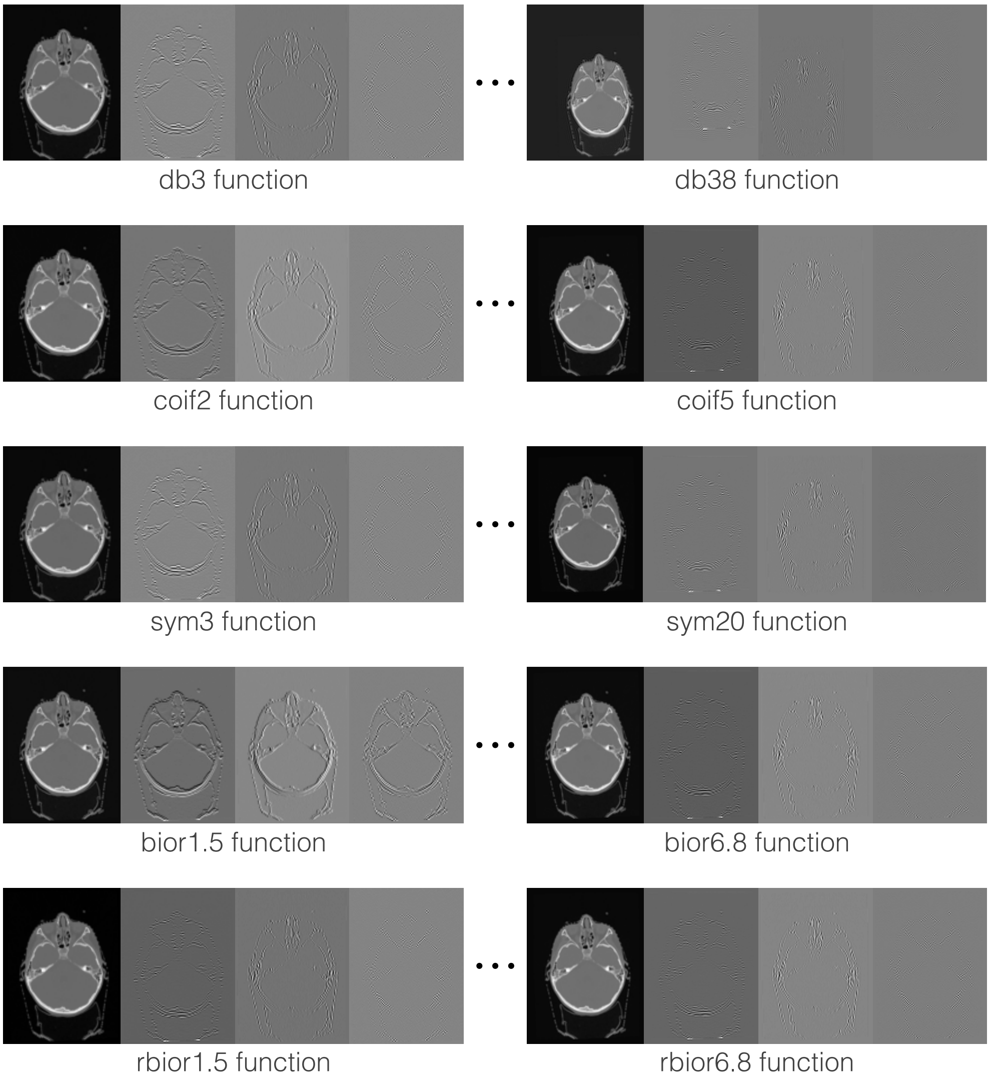


Figure A.6: Registration performance based on the Jacobian determinant for models with the batch size $b = 4$ using the 25-patient data set. The determinant is calculated individually for each pixel with the corresponding displacements from the deformation vector field. The results are shown for the small-architecture (top), medium-architecture (middle) and large-architecture (bottom) models. The mean values and uncertainties of the respective five-fold data are presented for the variation of the regularisation parameter λ and the learning rate α, separated into training (left) and validation (right). The dashed lines indicate the expected value.

A.5 Impact of wavelet functions

The decomposed images of several wavelet functions are shown in Figures A.7, A.8 and A.9.



db3 function

db38 function

coif2 function

coif5 function

sym3 function

sym20 function

bior1.5 function

bior6.8 function

rbior1.5 function

rbior6.8 function

Figure A.7: Image decomposition with the discrete wavelet transform for the preprocessed CT scan of Patient 9 of the 14-patient data set for higher orders of wavelet functions. The approximated images (first and fifth column) as well as the detailed images containing vertical (second and sixth column), horizontal (third and seventh column) and diagonal (fourth and eighth column) information are presented for five wavelet groups.

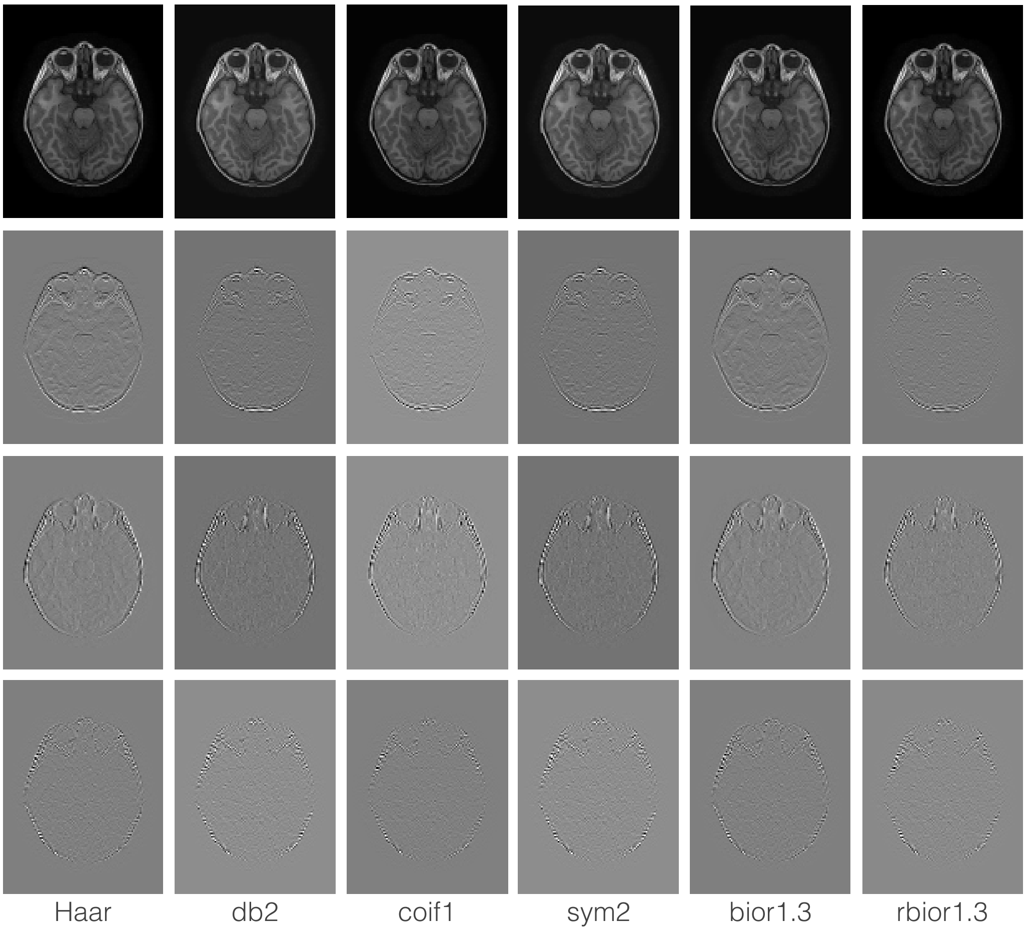


Figure A.8: Image decomposition with the discrete wavelet transform for the preprocessed T_1-weighted MRI scan of Patient 9 of the 14-patient data set. The approximated images (top row) as well as the detailed images containing vertical (second row), horizontal (third row) and diagonal (bottom row) information are presented for six wavelet functions.

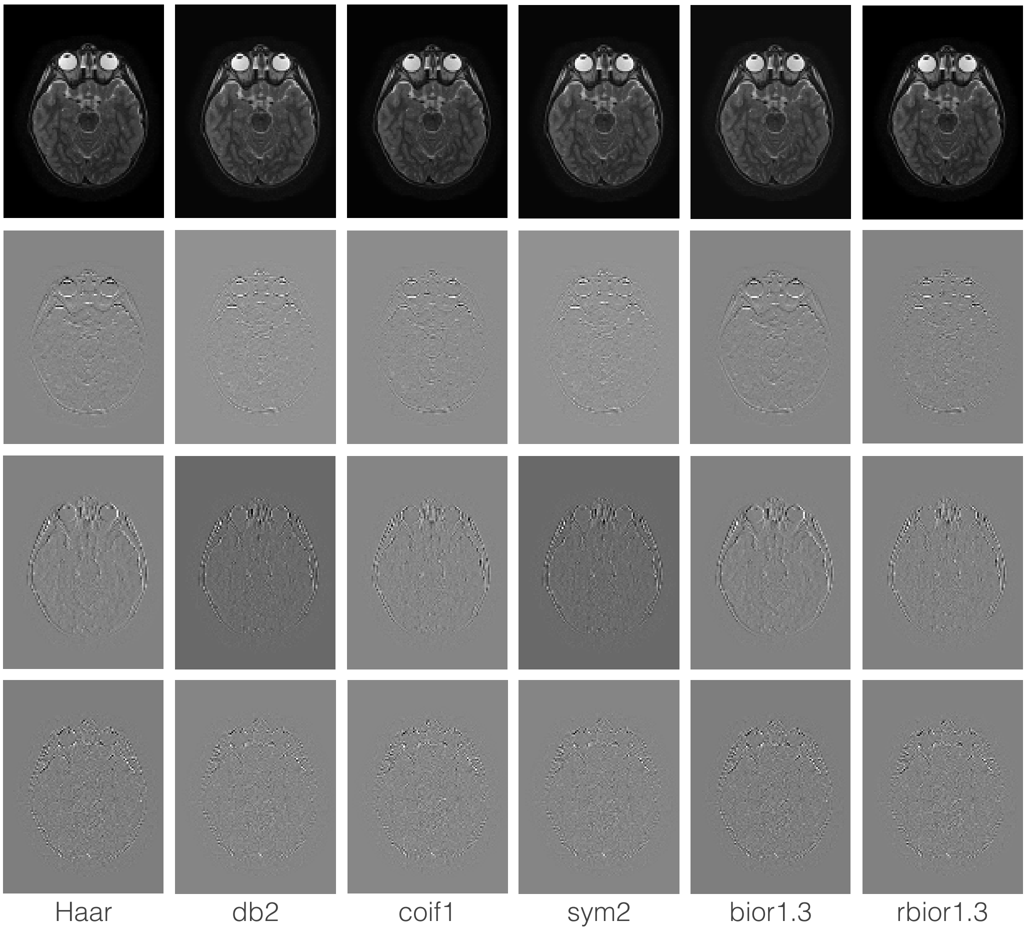


Figure A.9: Image decomposition with the discrete wavelet transform for the preprocessed T_2-weighted MRI scan of Patient 9 of the 14-patient data set. The approximated images (top row) as well as the detailed images containing vertical (second row), horizontal (third row) and diagonal (bottom row) information are presented for six wavelet functions.

Bibliography

[1] G.C. Pereira, M. Traughber and R.F. Muzic Jr. 'The role of imaging in radiation therapy planning: Past, present, and future'. In: *Biomed Res. Int.* 2014, 9 (2014). DOI: 10.1155/2014/231090.

[2] F.M. Khan. *The Physics of Radiation Therapy.* 4th ed. Lippincott Williams & Wilkins, 2009.

[3] H.-P. Chan et al. *Deep Learning in Medical Image Analysis. Challenges and Applications.* Advances in Experimental Medicine and Biology. Springer, 2020. DOI: 10.1007/978-3-030-33128-3_1.

[4] K.K. Brock et al. 'Use of image registration and fusion algorithms and techniques in radiotherapy: Report of the AAPM Radiation Therapy Committee Task Group No. 132'. In: *Med. Phys.* 44, e43 (2017). DOI: 10.1002/mp.12256.

[5] N.J. DeNunzio and T.I. Yock. 'Modern radiotherapy for pediatric brain tumors'. In: *Cancers* 12, 1533 (2020). DOI: 10.3390/cancers12061533.

[6] Y. Fu et al. 'Deep learning in medical image registration: A review'. In: *Phys. Med. Biol.* 65, 20TR01 (2020). DOI: 10.1088/1361-6560/ab843e.

[7] J. Zou et al. 'A review of deep learning-based deformable medical image registration'. In: *Front. Oncol.* 12, 1047215 (2022). DOI: 10.3389/fonc.2022.1047215.

[8] I. Goodfellow et al. 'Generative adversarial nets'. In: *NIPS'14: Proceedings of the 27th International Conference on Neural Information Processing Systems.* 2014, 2672.

[9] X. Cao et al. 'Deep learning based inter-modality image registration supervised by intra-modality similarity'. In: *9th International Workshop on Machine Learning in Medical Imaging.* 2018, 55. DOI: 10.1007/978-3-030-00919-9_7.

[10] Z. Xu et al. 'Adversarial uni- and multi-modal stream networks for multimodal image registration'. In: *23rd International Conference on Medical Image Computing and Computer-Assisted Intervention.* 2020, 222. DOI: 10.1007/978-3-030-59716-0_22.

[11] Z. Xu, J. Luo and J. Yan. 'F3RNet: Full-resolution residual registration network for deformable image registration'. In: *Int. J. Comput. Assist. Radiol. Surg.* 16, 923 (2021). DOI: 10.1007/s11548-021-02359-4.

[12] Z. Xu et al. 'Double-uncertainty guided spatial and temporal consistency regularization weighting for learning-based abdominal registration'. In: *25th International Conference on Medical Image Computing and Computer-Assisted Intervention.* 2022, 14. DOI: 10.1007/978-3-031-16446-0_2.

[13] J.M. Wolterink et al. 'Deep MR to CT synthesis using unpaired data'. In: *Second International Workshop on Simulation and Synthesis in Medical Imaging.* 2017, 14. DOI: 10.1007/978-3-319-68127-6_2.

[14] C. Tanner et al. *Generative adversarial networks for MR-CT deformable image registration.* 2018. arXiv: 1807.07349.

[15] R. Han et al. 'Deformable MR-CT image registration using an unsupervised, dual-channel network for neurosurgical guidance'. In: *Med. Image Anal.* 75, 102292 (2022). DOI: 10.1016/j.media.2021.102292.

[16] O. Ronneberger, P. Fischer and T. Brox. 'U-Net: Convolutional networks for biomedical image segmentation'. In: *18th International Conference on Medical Image Computing and Computer-Assisted Intervention.* 2015, 234. DOI: 10.1007/978-3-319-24574-4_28.

[17] G. Balakrishnan et al. 'VoxelMorph: A learning framework for deformable medical image registration'. In: *IEEE Trans. Med. Imaging* 38, 1788 (2019). DOI: 10.1109/TMI.2019.2897538.

[18] W. Schlegel, C.P. Karger and O. Jäkel. *Medizinische Physik. Grundlagen – Bildgebung – Therapie – Technik.* 1st ed. Springer Spektrum Berlin, Heidelberg, 2018. DOI: 10.1007/978-3-662-54801-1.

[19] R. Acharya et al. 'Biomedical imaging modalities: A tutorial'. In: *Comput. Med. Imaging Graph.* 19, 3 (1995). DOI: 10.1016/0895-6111(94)00043-3.

[20] T.M. Buzug. *Computed Tomography. From Photon Statistics to Modern Cone-Beam CT.* 1st ed. Springer Berlin, Heidelberg, 2008. DOI: 10.1007/978-3-540-39408-2.

[21] O. Dössel and T.M. Buzug. *Medizinische Bildgebung.* Biomedizinische Technik. De Gruyter, 2014. DOI: 10.1515/9783110252149.

[22] J. Radon. 'On the determination of functions from their integral values along certain manifolds'. In: *IEEE Trans. Med. Imaging* 5, 170 (1986). DOI: 10.1109/TMI.1986.4307775.

[23] W. Burger and M.J. Burge. *Digitale Bildverarbeitung. Eine algorithmische Einführung mit Java.* 3rd ed. Springer Berlin, Heidelberg, 2015. DOI: 10.1007/978-3-642-04604-9.

[24] *Python.* URL: https://www.python.org. Version: 3.8.10. Date of access: 14 June 2023.

[25] C.R. Harris et al. 'Array programming with NumPy'. In: *Nature* 585, 357 (2020). DOI: 10.1038/s41586-020-2649-2. Version: 1.19.5.

[26] C. Schneider, W. Rasband and K. Eliceiri. 'NIH Image to ImageJ: 25 years of image analysis'. In: *Nat. Methods* 9, 671 (2012). DOI: 10.1038/nmeth.2089. Version: 1.53.

[27] J.D. Hunter. 'Matplotlib: A 2D graphics environment'. In: *Comput. Sci. Eng.* 9, 90 (2007). DOI: 10.1109/MCSE.2007.55. Version: 3.4.3.

[28] *DICOM*. URL: https://dicom.nema.org/standard.htm. Date of access: 14 June 2023.

[29] J.-C. Yen, F.-J. Chang and S. Chang. 'A new criterion for automatic multilevel thresholding'. In: *IEEE Trans. Image Process.* 4, 370 (1995). DOI: 10.1109/83.366472.

[30] C.H. Li and C.K. Lee. 'Minimum cross entropy thresholding'. In: *Pattern Recognit.* 26, 617 (1993). DOI: 10.1016/0031-3203(93)90115-D.

[31] S. Kullback and R.A. Leibler. 'On information and sufficiency'. In: *Ann. Math. Stat.* 22, 79 (1951).

[32] M. Erdmann et al. *Deep Learning For Physics Research*. 1st ed. World Scientific Publishing Co. Pte. Ltd., 2021. DOI: 10.1142/12294.

[33] D.-M. Tsai and C.-T. Lin. 'Fast normalized cross correlation for defect detection'. In: *Pattern Recognit. Lett.* 24, 2625 (2003). DOI: 10.1016/S0167-8655(03)00106-5.

[34] F. Maes et al. *Image Registration Using Mutual Information*. Handbook of Biomedical Imaging: Methodologies and Clinical Research. Springer US, 2015, 295. DOI: 10.1007/978-0-387-09749-7_16.

[35] W.R. Crum, O. Camara and D.L.G. Hill. 'Generalized overlap measures for evaluation and validation in medical image analysis'. In: *IEEE Trans. Med. Imaging* 25, 1451 (2006). DOI: 10.1109/TMI.2006.880587.

[36] *Proton therapy*. URL: https://www.wpe-uk.de/en/proton-therapy/. Date of access: 14 June 2023.

[37] *Register Study Standard Proton Beam Therapy WPE - Children*. URL: https://drks.de/search/en/trial/DRKS00005363. Date of access: 14 June 2023.

[38] C. Bäumer et al. 'Adaptive proton therapy of pediatric head and neck cases using MRI-based synthetic CTs: Initial experience of the prospective KiAPT study'. In: *Cancers* 14, 2616 (2022). DOI: 10.3390/cancers14112616.

[39] E. Darsht. 'Optimierung und Evaluation eines neuronalen Netzes für anwendungsbezogene Bildregistrierung von CT- und MRT-Schädelaufnahmen'. Master's thesis. TU Dortmund University, 2022.

[40] S. van der Walt et al. 'Scikit-image: Image processing in Python'. In: *PeerJ* 2, e453 (2014). DOI: `10.7717/peerj.453`. Version: 0.19.3.

[41] *OpenCV*. URL: `https://opencv.org`. Version: 4.5.5. Date of access: 14 June 2023.

[42] P. Virtanen et al. 'SciPy 1.0: Fundamental algorithms for scientific computing in Python'. In: *Nat. Methods* 17, 261 (2020). DOI: `10.1038/s41592-019-0686-2`. Version 1.7.1.

[43] *Dicom-contour*. URL: `https://pypi.org/project/dicom-contour/`. Version: 2. Date of access: 14 June 2023.

[44] Y. LeCun, Y. Bengio and G. Hinton. 'Deep learning'. In: *Nature* 521, 436 (2015). DOI: `10.1038/nature14539`.

[45] A.L. Maas, A.Y. Hannun and A.Y. Ng. 'Rectifier nonlinearities improve neural network acoustic models'. In: *ICML Workshop on Deep Learning for Audio, Speech and Language Processing*. 2013.

[46] L. Geert et al. 'A survey on deep learning in medical image analysis'. In: *Med. Image Anal.* 42, 60 (2017). DOI: `10.1016/j.media.2017.07.005`.

[47] G. Balakrishnan et al. 'An unsupervised learning model for deformable medical image registration'. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018, 9252. DOI: `10.1109/CVPR.2018.00964`.

[48] TensorFlow Developers. *TensorFlow*. 2021. DOI: `10.5281/zenodo.4758419`.

[49] *Keras*. URL: `https://keras.io`. Version: 2.4.3. Date of access: 14 June 2023.

[50] K. He et al. 'Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification'. In: *2015 IEEE International Conference on Computer Vision*. 2015, 1026. DOI: `10.1109/ICCV.2015.123`.

[51] *RadnomNormal algorithm*. URL: `https://www.tensorflow.org/versions/r2.5/api_docs/python/tf/keras/initializers/RandomNormal`. Date of access: 14 June 2023.

[52] D.P. Kingma and J. Ba. 'Adam: A method for stochastic optimization'. In: *3rd International Conference for Learning Representations*. 2015. arXiv: `1412.6980`.

[53] J. Duchi, E. Hazan and Y. Singer. 'Adaptive subgradient methods for online learning and stochastic optimization'. In: *J. Mach. Learn. Res.* 12, 2121 (2011).

[54] *RMSprop algorithm*. URL: `https://www.tensorflow.org/versions/r2.5/api_docs/python/tf/keras/optimizers/RMSprop`. Date of access: 14 June 2023.

[55] J. Slagowski et al. 'Quantification of geometric distortion in magnetic resonance imaging for radiation therapy treatment planning'. In: *Int. J. Radiat. Oncol. Biol. Phys.* 102, e547 (2018). DOI: `10.1016/j.ijrobp.2018.07.1527`.

[56] F. Putz et al. ‘Magnetic resonance imaging for brain stereotactic radiotherapy’. In: *Strahlenther. Onkol.* 196, 444 (2020). DOI: 10.1007/s00066-020-01604-0.

[57] E.P. Pappas et al. ‘MRI-related geometric distortions in stereotactic radiotherapy treatment planning: Evaluation and dosimetric impact’. In: *Technol. Cancer Res. Treat.* 16, 1120 (2017). DOI: 10.1177/1533034617735454.

[58] K. Ulin, M.M. Urie and J.M. Cherlow. ‘Results of a multi-institutional benchmark test for cranial CT/MR image registration’. In: *Int. J. Radiat. Oncol. Biol. Phys.* 77, 1584 (2010). DOI: 10.1016/j.ijrobp.2009.10.017.

[59] N. Nahvi and D. Mittal. ‘Medical image fusion using discrete wavelet transform’. In: *Int. J. Eng. Res. Appl.* 4, 165 (2014).

[60] A.A. Suraj et al. ‘Discrete wavelet transform based image fusion and de-noising in FPGA’. In: *J. Electr. Syst. Inf. Technol.* 1, 72 (2014). DOI: 10.1016/j.jesit.2014.03.006.

[61] X. Xu, Y. Wang and S. Chen. ‘Medical image fusion using discrete fractional wavelet transform’. In: *Biomed. Signal Process. Control* 27, 103 (2016). DOI: 10.1016/j.bspc.2016.02.008.

[62] X. Zeng, Z. Luo and X. Xiong. ‘A fast fusion method for visible and infrared images using Fourier transform and difference minimization’. In: *IEEE Access* 8, 213682 (2020). DOI: 10.1109/ACCESS.2020.3041759.

[63] L. Qu et al. ‘Rethinking multi-exposure image fusion with extreme and diverse exposure levels: A robust framework based on Fourier transform and contrastive learning’. In: *Inf. Fusion* 92, 389 (2023). DOI: 10.1016/j.inffus.2022.12.002.

[64] N. Kehtarnavaz. ‘Frequency Domain Processing’. In: *Digital Signal Processing System Design.* 2nd ed. Academic Press, 2008. Chap. 7, 175. DOI: 10.1016/b978-0-12-374490-6.00007-6.

[65] A.N. Akansu and R.A. Haddad. *Multiresolution Signal Decomposition. Transforms, Subbands, and Wavelets.* 2nd ed. Elsevier, 2001. DOI: 10.1016/B978-0-12-047141-6.X5000-9.

[66] G. Lee et al. *PyWavelets.* 2022. DOI: 10.5281/zenodo.6347505.

[67] A. Haar. ‘Zur Theorie der orthogonalen Funktionensysteme’. In: *Math. Ann.* 69, 331 (1910). DOI: 10.1007/BF01456326.

[68] I. Daubechies. ‘Orthonormal bases of compactly supported wavelets’. In: *Comm. Pure Appl. Math.* 41, 909 (1988). DOI: 10.1002/cpa.3160410705.

[69] R.X. Gao and R. Yan. *Wavelets. Theory and Applications for Manufacturing.* 1st ed. Springer New York, NY, 2011. DOI: 10.1007/978-1-4419-1545-0.

[70] G. Beylkin, R. Coifman and V. Rokhlin. 'Fast wavelet transforms and numerical algorithms I'. In: *Comm. Pure Appl. Math.* 44, 141 (1991). DOI: `10.1002/cpa.3160440202`.

[71] S. Mallat. 'Wavelet Bases'. In: *A Wavelet Tour of Signal Processing*. 3rd ed. Academic Press, 2009. Chap. 7, 263. DOI: `10.1016/B978-0-12-374370-1.00011-2`.

[72] Q. Wang, Y. Shen and J. Jin. 'Performance Evaluation of Image Fusion Techniques'. In: *Image Fusion*. Elsevier, 2008. Chap. 19, 469. DOI: `10.1016/b978-0-12-372529-5.00017-2`.

[73] C. Ramesh and T. Ranjith. 'Fusion performance measures and a lifting wavelet transform based algorithm for image fusion'. In: *Proceedings of the Fifth International Conference on Information Fusion*. 2002, 317. DOI: `10.1109/ICIF.2002.1021168`.

[74] L. Kader. 'Untersuchung und Optimierung der fusionierten Bildgebung von CT- und MRT-Aufnahmen des Schädelbereichs'. Bachelor's thesis. TU Dortmund University, 2022.

[75] J. Köster and S. Rahmann. 'Snakemake—A scalable bioinformatics workflow engine'. In: *Bioinformatics* 28, 2520 (2012). DOI: `10.1093/bioinformatics/bts480`. Version 6.3.0.

[76] B. Fischl. 'FreeSurfer'. In: *NeuroImage* 62, 774 (2012). DOI: `10.1016/j.neuroimage.2012.01.021`.

[77] D. Bodensteiner. 'RayStation: External beam treatment planning system'. In: *Med. Dosim.* 43, 168 (2018). DOI: `10.1016/j.meddos.2018.02.013`.

[78] *RaySearch Laboratories*. URL: `https://www.raysearchlabs.com`. Date of access: 14 June 2023.

[79] O. Weistrand and S. Svensson. 'The ANACONDA algorithm for deformable image registration in radiotherapy'. In: *Med. Phys.* 42, 40 (2015). DOI: `10.1118/1.4894702`.

Acknowledgements

Starting a PhD on registration of medical images wasn't that obvious as my research field had been particle physics before. However, Prof. Spaan convinced me to face the challenge because: 'It's all about handling data, whether from particle physics or medical physics' – Bernhard Spaan. This conversation encouraged me to start this thesis in January 2020. Sadly, Prof. Spaan suddenly passed away in December 2021, leaving a void. I want to thank him for giving me this opportunity; so, I'm dedicating the thesis to him.

Prof. Kevin Kröninger took over as the first assessor of this thesis – many thanks to him and his advice, supporting me in the last one and a half years of my PhD. The second assessor, PD Christian Bäumer, was always a help by giving me tips from the perspective of a medical physicist.

The central part of my PhD was overshadowed by COVID-19, which forced me and the whole working group to do home office. Thankfully, our E5a group had regular coffee meetings via Zoom during that time. This was a small consolation that allowed keeping in touch for one and a half years. Returning to the office in October 2021 was limited at first, but it was nice to have little talks in the office again, especially with my office colleagues Henning Manke and later Michelle Stroth. I enjoyed our three-person medical physicist's office.

My PhD wouldn't have happened without the MERCUR grant St-2019-0007, which funded our graduate school, involving the Essen University Hospital, the West German Proton Therapy Centre Essen and the University of Duisburg–Essen. The graduate school included seminars and lectures each semester, which were very interesting, especially for someone without medical physics knowledge. I liked the Zoom meetings after the lessons with the people from the graduate school, where we had nice conversations. A highlight was the AAPM conference in Washington, D.C., where we were able to present our work.

I am indebted to those who took the time to read the draft of this thesis and gave useful comments.

Furthermore, I need to thank another colleague of the E5a group and friend of mine: Julian Boelhauve. Thank you for the cheerful time during my bachelor, master and PhD. Especially during the last months of writing up, you always motivated me and contributed much by reviewing this thesis.

Last but not least, I need to thank my family and friends for supporting me during nine years of studying at TU Dortmund University. Without them, I wouldn't have come this far.