Maximilian Sackl, BSc

Automated Hippocampus Segmentation of High-Resolution MR Images with Deep Neural Networks

MASTER'S THESIS

to achieve the university degree of
Diplom-Ingenieur

Master's degree programme
Biomedical Engineering

submitted to

Graz University of Technology

Supervisor

Univ.-Prof. Dipl.-Ing. Dr. techn. Thomas Pock
Institute for Computer Graphics and Vision, Technical Univeristy Graz

Advisors

Dipl.-Ing. Dr. techn. Darko Štern
Department of Biophysics, Medical University of Graz

Assoz.-Prof. Univ.-Doz. Dipl.-Ing. Dr. techn. Stefan Ropele
Department of Neurology, Medical University Graz

Graz, Austria, June 2020

Abstract

Alzheimer's disease (AD) is a neurodegenerative disease and the most common form of dementia, with over 50 million affected worldwide. The Hippocampus (HC) is a tiny structure responsible for memory consolidation and among the first brain regions suffering damage. *HC* atrophy (tissue loss), assessed with magnetic resonance imaging (*MRI*), is the most important imaging marker of *AD* and outcome measure in clinical trials. Proper assessment of the *HC*-volume is therefore crucial. However, manual segmentation – the gold standard – requires an adept annotator, is a taxing and time consuming task and prone to inter-rater bias. These problems can be reduced by automated segmentation, where *FreeSurfer* is the most frequently used software in clinical research. For these segmentations, T1-weighted (T1w) *MR* images are standard, however, their capability to clearly separate the *HC* from surrounding tissue and *CSF* is limited. Thus, due to its enhanced contrast, T2-weighted (T2w) images have been proposed to serve as a better ground truth (GT) for segmenting the HC.

This thesis aims to assess if deep learning approaches can outperform *FreeSurfer*'s segmentation and if *T2w* high-resolution images can further improve the labelling result. A dataset with corresponding pairs of high-resolution 3T *T1w* and *T2w* scans was specifically acquired from healthy subjects, of which 28 hippocampi were manually labelled ($=GT$). The labelling was done on *T2w* images to benefit from the superior contrast and resolution. Different convolutional neural networks (CNNs) are proposed for segmentation and all results are compared to our *GT*. As a reference approach, *FreeSurfer* showed a Dice similarity coefficient (DSC) of 78.06%. Using these baseline labels also for training a *CNN* on *T1w* images reduced the computation time drastically and improved the *DSC*. Training with the *GT*-labels further increased results (*DSC* of 85.62%), while using the *T2w* images yielded overall the best result (*DSC* of 91.91%). As appropriate *T2w* scans are not feasible in clinical practice, T2-enhanced training of our *CNNs* was evaluated, including training on *T2w* and fine-tuning on *T1w* images, or by using generative adversarial networks to transform *T1w* scans to mimic T2-contrast. However, these results are similar to only using *T1w* images. Concluding all experiments, *T2w* images are suggested to obtain a reliable *HC* segmentation and a ground truth, since they provide better contrast of *HC* related features and show less flow- or susceptibility artefacts.

Kurzfassung

Die Alzheimer-Krankheit (AK) ist mit über 50 Millionen Erkrankten weltweit die häufigste Form der Demenz. Im Zuge dieser neurodegenerativen Erkrankung wird der Hippocampus (HC), eine kleine Struktur der Temporallappen die maßgeblich für die Bildung des Langzeitgedächtnisses verantwortlich ist, bereits zu Beginn stark geschädigt. Konkret führt die AK zu einem Gewebsverlust (Atrophie) im Bereich des medialen Temporallappens und speziell des HC. HC-Atrophie lässt sich mittels Magnetresonanztomographie (MRT) sehr gut darstellen und ist der wichtigste Bildgebungsmarker für die degenerative Komponente der AK, sowie eine häufig verwendete Messgröße für die Wirksamkeit von Therapiestudien. Die exakte Bestimmung des HC-Volumens ist daher essentiell. Der Goldstandard, die manuelle Segmentierung, ist jedoch sehr komplex und zeitintensiv sowie anwenderabhängig. Die mit manueller Segmentierung verbundenen Probleme lassen sich durch Verwendung einer standardisierten, automatischen Methode reduzieren. FreeSurfer (FS) ist in der klinischen Forschung die dafür meist genutzte Applikation. Dafür werden standardmäßig T1-gewichtete (T1w) MR-Bilder verwendet, die jedoch das Potential zur exakten Differenzierung zwischen dem HC und dem angrenzenden Gewebe sowie des Liquors limitieren. Aufgrund besserer Kontrasteigenschaften von T2-gewichteten (T2w) Bildern werden diese zur Erstellung einer akkuraten Ground Truth (GT) Segmentierung empfohlen.

Der Fokus dieser Diplomarbeit liegt im Wesentlichen auf zwei Fragestellungen. Erstens soll evaluiert werden, ob es mittels maschinellem Lernen möglich ist eine bessere Segmentierung des HC, verglichen mit *FreeSurfer*, zu erreichen. Zweitens soll untersucht werden, ob und in welchem Maße hochaufgelöste T2w Bilder das Ergebnis beim Lernen verbessern können. Dafür wurde ein neues Datenset, bestehend aus korrespondierenden, hochaufgelösten T1w und T2w MR-Bildern von gesunden Probanden aufgenommen. Von diesen akquirierten Daten wurden 28 Hippocampi händisch eingezeichnet. Um die bestmöglichen GT-Labels zu erhalten, wurden dazu die T2w Bilder verwendet. In dieser Diplomarbeit werden verschiedene Convolutional Neural Networks (CNNs) zur Segmentierung des HC vorgeschlagen und die Ergebnisse mit den manuellen GT-Labels evaluiert. Dazu wurden Dice Similarity Coefficients (DSCs) berechnet. Als Basis für jegliche Vergleiche wurden die Ergebnisse der *FreeSurfer* Segmentierung herangezogen (DSC=78.06%).

Werden die FS-Masken in Kombination mit den T1w Bildern zusätzlich dazu verwendet die Segmentierungsnetzwerke zu trainieren, konnte nicht nur die Berechnungszeit drastisch reduziert, sondern auch das Ergebnis verbessert werden. Werden hingegen die GT-Labels, gemeinsam mit den T1w Bildern, zum Trainieren des Modells verwendet, konnte das Ergebnis weiter verbessert werden (DSC=85.62%). Das beste Resultat konnte erzielt werden, indem stattdessen die T2w Bilder segmentiert wurden (DSC=91.91%). Da die Akquisition von entsprechenden T2w Scans bei klinischen Untersuchungen leider nicht möglich ist, wurde zusätzlich das Potential von T2-verstärktem Training der Netzwerke evaluiert. Dafür wurden die Segmentierungsnetzwerke mit bereits auf T2w Bildern trainierten Gewichten initialisiert und mittels T1w Bildern optimiert. Außerdem wurden in einem Zusatztask mittels Generative Adversarial Networks (GANs) synthetische T2w Bilder, basierend auf den jeweiligen T1w Bildern, generiert, um damit die T1w-basierte Segmentierung zu verbessern.

Diese Zusatzexperimente lieferten sehr ähnliche Ergebnisse zu der reinen Segmentierung von T1w Bildern. Unter Berücksichtigung aller erworbenen Erkenntnisse lässt sich folgern, dass mit hochaufgelösten T2w Scans eine verlässliche GT Segmentierung erstellt und mittels CNNs automatisch reproduziert werden kann. Dies lässt sich auf die verbesserte Bildqualität sowie die kontrastreichere Darstellung relevanter anatomischer Strukturen zurückführen. Weiters haben die durchgeführten Experimente gezeigt, dass die Ergebnisse von *FreeSurfer* mittels maschinellem Lernen sowohl beschleunigt als auch verbessert werden können.

Affidavit

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly indicated all material which has been quoted either literally or by content from the sources used.

The text document uploaded to TUGRAZonline is identical to the present master's thesis dissertation.

____________________ ____________________

Date Signature

Acknowledgments

This page is dedicated to all the people who supported and encouraged me not only since I have started working on this thesis but also during my whole studies.

First and foremost, I would like to thank Professor Stefan Ropele, my advisor at the Medical University of Graz, who not only made the whole thesis possible but also provides the possibility to continue with this work. Thank you for your continuous support throughout this time as well as your valuable guidance.

Secondly, I would like to thank Dr. Darko Štern, of the Medical Image Processing research group, for his ideas and feedback, in particular during the final phase of this work. Moreover, I want to thank all colleagues from this group, especially Stefan, Christian and Franz for their continuous help and enlightening insights to ensure the success of this work.

In addition to that, I would like to extend my gratitude to Professor Thomas Pock, my supervisor at the Technical University of Graz, for his support and feedback.

I also want to thank my colleagues at the Neuroimaging Research Unit, Medical University Graz, for all the insightful discussions and for the fun we had since we have been working together.

Furthermore, special thanks to all my friends for motivating and encouraging me as well as for providing their support.

I would also like to express my deepest gratitude to my parents and the rest of my family for their unconditional and continuous support, love and encouragement.

Finally, I want to express my greatest heartfelt appreciation to my girlfriend Mirjam, for all her endless support and encouragement. Thank you for always being there for me and for believing in me!

Contents

List of Figures

List of Tables

1 Introduction

Alzheimer's disease (AD) is the most common form of dementia accounting for $60 - 80\%$ of all dementia patients [2]. Hippocampal volume and especially its atrophy (shrinkage), assessed with Magnetic resonance imaging (MRI), has shown to be a valuable marker for dementia related diseases, most notably to aid early diagnosis and as an outcome measure for therapy studies.

The brain (or encephalon) is our most complex organ and builds together with the spinal cord the central nervous system (CNS). Receiving and processing all sensory input provided by our senses, makes it the most important body part regarding information processing. The encephalon controls our actions, enables us to express unique behaviour and have complex thoughts. It is composed of two major tissue groups: grey matter (GM) and white matter (WM). The cerebrum (or telencephalon) is the largest part of the brain and responsible for higher brain functions such as sophisticated thoughts, logic and creativity. It is divided by the *medial longitudinal fissure* into two sides; the left and right hemisphere. The outermost layer of the cerebrum is called *cerebral cortex*, consists of *GM* and can be separated into four lobes: *frontal lobe*, *parietal lobe*, *temporal lobe* and *occipital lobe*. The temporal lobe, especially the medial temporal lobe (MTL), which includes the hippocampus (HC), plays a key role in long-term memory formation. The hippocampi are essential for memory consolidation, involved in emotions and enables navigation.

Dementia is an umbrella-term for conditions characterised by several deficits of cognitive abilities, such as decline in memory, thinking, orientation and judgement, which are strongly related to the *HC*. This cognitive deficit is typically followed, and sometimes preceded, by deterioration in behaviour, motivation or emotional control, causing a disability to perform everyday tasks [2]. In addition to *AD*, other dementia causes include severe types such as vascular dementia (VaD), dementia with Lewy bodies (DLB) and forms which conduce to frontotemporal dementia [28, 106]. While all forms of dementia show similar syndromes and are caused by neuronal damage, the different types of dementia depend on the affected brain cell type and the brain region where the damage occurs.

In Alzheimer's disease excessive accumulation of amyloid-β (Aβ), a fragmented protein also known as (neuritic) plaques, and abnormally twisted tau protein (τ-protein), also called (neurofibrillary) tangles, are the hallmarks of the disease. Alzheimer-related stages of these pathological changes can be analysed with the *Braak staging* [13, 14]. Alterations occurs even years before symptoms are notable, with early *MTL* (*HC* and entorhinal cortex (ErC)) participation followed by progressive neocortical impairment [13, 27]. The brain damage in *VaD* is supposed to happen because of insufficient blood supply of the brain tissue. In other causes of dementia, like *DLB* and Parkinson's disease, an unusual clumping of the alpha-synuclein (α-synuclein) protein occurs. Abnormal brain proteins can also be caused by genetic conditions, such as Huntington's disease, which yield symptoms of dementia. Independent of the specific cause, these types are often generally described as neurodegenerative due to the progressive nerve cell damage and the resulting decline in the patients condition. [85]

In total, roughly 50 million people worldwide suffer from dementia with an increase of around 10 million each year [106], whereas dementia and *AD* is most frequent in Western Europe, followed by North America [77]. These tremendous numbers make dementia, along with *AD*, a condition causing one of the highest costs for health care and long-term care costs to society. The Alzheimer's Association estimates the national costs of the United States of America in 2020 to reach 305 billion dollar with additional costs for informal care of $ 244 billion [3]. Alzheimer's Disease International (ADI) estimated the global financial burden in 2015 due to dementia, including medical and social care costs as well as costs for informal care, at $ 818 billion [77]. Moreover, Wimo et al. analysed that the global costs increased by 35% in the time from 2010 to 2017 and forecast to reach costs of $ 1500 billion already by 2025 [104].

Despite all these numbers, dementia is currently under-diagnosed, under-disclosed and under-treated in primary care [78, 106]. Also, diagnosis is usually made only at a relatively late stage of the disease [106] and the explicit causes for dementia are still unclear. Due to overlapping alterations a clear identification of the specific type is also hardly possible. Unfortunately, there is no single test to diagnose *AD* or dementia. Physicians rather use a combination of cognitive tests, imaging to rule out other causes of the symptoms, and lumbar puncture to assess tau-levels to come up with a diagnosis like "possible Alzheimer's dementia" or "probable Alzheimer's dementia".

The most frequently used test is the Mini Mental State Examination (MMSE) which takes only 10 minutes and is comprised of tasks and questions that evaluate cognitive capabilities such as memory, language or attention. For *AD* patients the average MMSE-score decline is around 2 – 4 per year [22]. An increasingly used alternative is the Montreal Cognitive Assessment (MoCA) test, which is a 10 minute screening tool to assess mild cognitive impairment (MCI) [73]. Another common test, to quickly check if further evaluation is needed, is the Mini-Cog test, where only three common objects need to be recalled after drawing a complete clock showing a specified time. The Consortium to

Establish a Registry for Alzheimer's Disease (CERAD) neurophysiological battery is a collection of five subtests to reliable measure cognition in normal ageing and *AD* [19]. It is comprised of the Animal Naming, Boston Naming Test (BNT), *MMSE*, Constructional Praxis and Word List Memory and takes 40 minutes or more. The ADAS-cognitive subscale (ADAS-Cog) ([81]) was specifically designed to assess the severity of cognitive dysfunctions in *AD* patients and became the most commonly used test in research and clinical *AD* trials [88]. Skinner et al. [88] improved the *ADAS-Cog* with regard to the responsiveness in *MCI*, while Verma et al. [96] further improved the *ADAS-Cog* using item response theory (IRT) yielding to the ADAS-Cog using IRT (ADAS-CogIRT).

Despite neuropsychological tests, to assess the cognitive capabilities, an evaluation of the skills to perform everyday tasks is also crucial. This is done in terms of activities of daily living (ADL) for which several scales exist. While the Bristol ADL Scale (BADLS)([16]) is presumably the gold standard for evaluating non-cognitive deficits in dementia, the most popular one, however, is the Neuropsychiatric Inventory ([24]) [28].

If the diagnosis of a dementia syndrome has been made, the most likely cause for it has to be found. The reason is, that the syndromes can be caused by several forms of dementia. However, they can also be caused by other reasons such as vitamin deficiencies, medication side effects, infections or metabolic disorders. Therefore, in addition to the various questionnaires, laboratory tests and brain imaging are performed to rule out these causes, that might be reversible [28]. Laboratory tests include blood work or urinalysis and are primarily used to check functioning of various organs and screen for infections.

While an accurate and reliable clinical diagnosis with more than 90% [65] is possible with such questionnaires, a definite *AD* diagnosis can still only be achieved postmortem after a histopathological validation. Moreover, all of these questionnaires and clinical tests suffer from several disadvantages. They are biased by education, language and culture [15] and the condition of the subject at test time. Furthermore, professional examinations usually happen only at a late stage of the disease, at which the brain is already damaged.

Achieving a diagnosis as early as possible is of utmost interest yielding many benefits not only for the patients itself, but also for caregivers and society. This could be achieved by observing prodromal *AD* or subjects at risk of evolving *AD*. Amnestic *MCI* is the most commonly accepted recommendation among them [21]. Mild cognitive impairment refers to subjects with significant but isolated memory derogation in comparison to references of the same age. Careful tracking of *MCI* cases is supported by annual conversion rates from *MCI* to *AD* of approximately $10 - 15\%$, yielding a translation to *AD* of $50 - 75\%$ over the coarse of five years[65]. The most promising way to indicate already early stages of Alzheimer's or *MCI* is the use of biomarkers.

Biomarkers, shorthand for biological markers, are measurable parameters that can be reliably used for accurate detection of a disease or pathological conditions, assess the chance of developing the disease or even understand how patients respond to certain treatments. Therefore, a lot of research is going on to study and establish biomarkers for early *AD* detection. The most promising candidates involve brain imaging, cerebrospinal fluid (CSF) proteins and blood and urine tests [21, 34, 45, 56, 75, 91, 110]. Genetic risk profiling can be useful to rule out rare causes due to genetic deficiencies. These laboratory and neuroimaging biomarkers can be categorised into *pathophysiological* and *topographical* markers [34].

Olsson et al. [75] conducted the most comprehensive meta-analysis of *CSF* and blood biomarker literature in *AD* as well as *MCI*. A total of 231 articles, comprising 15699 *AD* cases and 13018 controls were included and examined for 15 biomarkers from blood and *CSF*. Three of these 15 markers are known as *core CSF biomarkers* for *AD* diagnosis:

- 42-aminoacid form of amyloid-β (Aβ42), found at low concentrations in *AD* patients as a result of cortical amyloid deposition.
- total tau (T-tau), found at high concentrations through cortical neuronal loss.
- phosphorylated tau (P-tau), found at high concentrations reflecting the formation of cortical tangles.

These core biomarkers are able to differentiate *AD* patients from controls with good success. Moreover, they can be effectively used to differentiate between *MCI* and stable *MCI*. Cerebrospinal fluid neurofilament light (NFL) protein, together with plasma *T-tau*, were also strongly associated with *AD*. Because of the consistency of the mentioned markers, they are suggested to be used in clinical practice and research.

Neuroimaging, is also among the most promising ways for early detection of Alzheimer's disease and for following up its progression. Several imaging techniques, such as structural *MRI* or computed tomography (CT), functional imaging with functional MRI (fMRI) or positron emission tomography (PET) and even molecular imaging including *fMRI* and *PET* are studied. Of all the underlying principles, *MRI* is the most alluring for several reasons. One of them is the high availability and utility, which is why it is renowned as an essential examination for dementia in most centres. Another major benefit is, that *MRI* uses harmless high power radio waves to create detailed views from inside the body. Different tissue contrasts, emphasising for example *WM* or *GM*, can be achieved with variations of the sequence parameters.

In comparison, *CT* uses X-rays to image cross-sections of various body parts and is mainly used to check for tumours, head trauma or stroke. While being faster and less noisy than *MRI*, *CT*, however, relies on ionising radiation. The latter is especially true for imaging methods including radioactive substances as used in *PET*.

Imaging of the brain alone is not used for a definite diagnosis but rather supports the diagnosis in various ways. It can not only be used to rule out other causes of symptoms such as tumours, haemorrhages and strokes, but also to differentiate between subtypes of dementia. For example shrinkage of specific brain areas can be an indicator for a certain subtype. In *AD*, dementia is caused by progressive loss of brain tissue, which can show up in a vast variety in brain imaging. Or, a detailed evaluation of brain vessel damage could allow for diagnosis of *VaD*. Both scenarios can be shown with magnetic resonance (MR) imaging. Moreover, a baseline and progression of degeneration can be accomplished.

In the last decade imaging markers have shown to reliably differentiate mild and moderate dementias from other neurodegenerative dementias and healthy controls [21, 34, 38, 91, 110]. In amyloid imaging via *PET*, a radiolabelled tracer is used to detect *Aβ* deposition into plaques. Teipel et al. [91] performed a pooled evaluation of published neuropathological validation studies, reporting a pooled sensitivity of 92% and a pooled specificity of 95% for amyloid aggregation detection with the use of F-labelled amyloid *PET* tracers. If the C-labelled Pittsburgh compound B (PIB) tracer is used, *PIB PET* scans showed a pooled sensitivity of 73% and a pooled specificity of 100% compared to histopathological presence/absence of amyloid. The main clinical usage of amyloid *PET* is limited to early stages of dementia. However, for the prediction of conversion speed from *MCI* to *AD* neuronal injury markers like *CSF* tau, FDG *PET* or structural MRI are more powerful.

Structural imaging with *MR* of patients with probable *AD*, in comparison to healthy aged controls, is consistent with established histopathological data and is able to show the most specific and sensitive characteristics of *AD*. These characteristics are atrophy (tissue loss) of the *HC* and the *ErC* and correspond even better if additional reduction in the temporal neocortex volume is present [21]. In comparison with *MCI* patients, these atrophic regions significantly expand to the temporal association neocortex, neighbouring the hippocampal regions. Moreover, *MCI* shows a significant atrophy of the *HC* compared to healthy controls. Frisoni et al. [38] also report typical early sites of atrophy, shown via *MRI*, to be along the perforant pathway, which is comprised of the *ErC*, the hippocampus formation and the posterior cingulate cortex. This coincides with early deposition of hyperphosphorylated tau and early deficits in memory. The authors also report atrophy of the medial temporal structures not only as a valid diagnostic marker already at the *MCI* stage, so before dementia occurs, but also as diagnostic criteria for most prevalent non-Alzheimer dementias.

Atrophy is an unavoidable and unstoppable consequence of neurodegeneration. The topography of this cerebral tissue loss correlates cross-sectionally and longitudinally with neuropsychological (cognitive) deficiencies. Moreover, these structural brain changes also precisely map upstream to the Braak stages of neurofibrillary tangle deposition [34, 55, 95, 103]. As such, *MRI* measures of atrophy can be used as a valid indicator of disease state and progression [38]. The latter is supported not solely by the change but also by the rate of change of structural measures, such as whole-brain-, *ErC*-, *HC*- and temporal lobe (TL) volumes. Additionally, enlargement of the ventricles is closely correlated with alterations in cognitive performance. These structural measures also change with progression of the disease over a broad range of *AD* severity. In stages from *MCI* to moderate dementia in *AD*, they have shown to be more sensitive to such changes as imaging or *CSF* markers of *Aβ* deposition. Therefore, hippocampal- and whole-brain atrophy rates are sensitive neurodegeneration markers and thus increasingly used as outcome measures in drug trials evaluating possible disease-modifying therapies [38]. Hippocampal volume in mild dementia stages of *AD* has been reported to be already 15 – 30% smaller compared to controls [38]. In amnestic *MCI* cases, the reduction in *HC* volume is lower, with a range of 10 – 15% [87]. Annual rates of hippocampal atrophy of 4.66% (95% CI [3.92 – 5.40]) in *AD* patients versus 1.41% (95% CI [0.52 – 2.3]) in healthy controls were reported by a meta-analysis from [7]. In comparison, the estimated mean whole-brain atrophy rate in subjects with *AD* is $1.9 \mp 0.9\%$ [89].

In 2015, Teipel et al. [91] showed that the differential pattern of brain-wide atrophy can differentiate pathologically confirmed *AD* cases from healthy subjects with a sensitivity of 97% and 94% specificity. Distinction from dementia patients with frontotemporal lobar degeneration or *DLB*, among further underlying pathological changes, resulted in a sensitivity and specificity of 91% and 84% respectively. While atrophy, plaques and tangles are all associated with dementia, Johnson et al. [55] pointed out that epidemiological autopsy studies of dementia subjects and controls showed that atrophy is the strongest correlate with dementia at all stages. Within the evaluation of Frisoni et al. [38], hippocampal atrophy measured on high-resolution T1-weighted *MR* images is the best established and validated *MRI* marker of *AD*. By visual inspection (not to be confused with segmentation) assessed atrophy of the *MTL*, rated by widely used rating scales [47], showed a sensitivity and specificity of around 80 – 85% in distinguishing *AD* patients from controls. Values for diagnosing amnestic *MCI* are slightly lower but also the predictive power of these scales to anticipate the decline in *MCI* are good [38]. Volumetric approaches such as visual scoring or manual segmentation of *HC* and *ErC* volume in single-centre studies of *MCI* have shown prediction accuracies for conversions to *AD* of around 80% [91]. Yuan et al. [112] report in their meta-analysis that medial temporal atrophy achieves a pooled sensitivity and specificity, for prediction of amnestic *MCI* cases translating to *AD*, of ≈ 73% and ≈ 81% respectively.

Regional and global atrophy, evaluated with *MRI* based segmentation, has been suggested as surrogate outcomes for disease-modifying trials due to a potential increase in study power [55]. Also [38] suggest, that for clinical trials, enrichment of *MCI* groups with *AD* cases, based on hippocampal atrophy as an inclusion criterion, may significantly increase study power for drugs aiming to delay dementia development.

Segmentation can be described as delineation of one structure from another. Crucial to evaluating brain changes, including atrophy, is a consistent assessment of the volume and shape of corresponding brain structures. This is especially important for multi-study comparisons or longitudinal studies.

While many scales for *visual rating* of atrophic changes exist and *manual segmentation* measures of hippocampus volume are still considered as the gold standard for classifying *MCI* and its conversion to *AD* [79], all of them have drawbacks. The main disadvantages are long assessment times, required know-how as well as inter- and intra-rater reliability.

Novel image processing algorithms enable the possibility for time efficient *automated segmentation* of the hippocampi or other neurodegeneration related structures with high anatomical accuracy, while ensuring consistency throughout the evaluation process. Frisoni et al. [38] also pointed out that a standardised acquisition and evaluation procedure as well as deployment of robust methods for automated assessment would increase the utility of structural imaging and other biomarkers. Within the last couple of years, machine learning and especially deep learning have become increasingly utilised in fields like image processing, pattern recognition and medical image analysis, yielding to applications for image reconstruction, segmentation and classification.

2 Related Work

Contents

This section will give an overview of related literature to the relevant topics of this thesis. At first, commonly used magnetic resonance imaging (MRI) sequences to image brain structures involved in Alzheimer's disease (AD), especially the medial temporal lobe (MTL), are reported. Methods for manual and automated segmentation of those important structures are shown in the next section. In the last part, a short overview of image-to-image translation methods is given.

2.1 Imaging of the Hippocampus

MRI is the imaging gold standard for most diseases of the central nervous system (CNS). In brain or dementia related structures, such as the *MTL*, two most commonly used imaging contrasts are applied to highlight different tissue features; T1- and T2-contrast.

T1-weighted (T1w) sequences, showing a T1-contrast of the imaged region, are the standard to visualise the whole head (brain). In research, T1 scans with a resolution of $1 \times 1 \times 1\,\text{mm}^3$, also referred to as 1 mm isotropic resolution, are basically part of every clinical *MRI* study ([33, 54]). While this is considered standard in research, such scans are actually regarded as high-resolution T1 images from a clinical point of view, as conventional sequences usually have a slice thickness of $3 - 5\,\text{mm}$. For high-resolution T1 scans, the three-dimensional (3D) magnetisation-prepared rapid gradient-echo (MPRAGE) ([71, 98]) is the most popular sequence and is used in many applications.

T2-weighted (T2w) sequences are used in clinical examinations to assess certain types of pathologies, for example lesions in multiple sclerosis (MS) or microbleeds. However, T2-contrast is also increasingly used in research ([33, 54, 60, 113, 114]), especially as higher resolutions can be achieved more easily. T2 contrast is best achieved with spin echo (SE) or fluid-attenuated inversion recovery (FLAIR) sequences. For the former, turbo spin echo (TSE) ([8, 35, 105, 114]) or their dual-echo variants ([50, 54, 60, 92]), creating proton density (PD)- and *T2w* images, are utilised. *TSE* sequences are also called fast spin echo (FSE) sequences as these are just manufacturer depending names. However, both of them refer to ***R**apid **A**cquisition with **R**elaxation **E**nhancement* (RARE) – the generic name.

The advent of emerging imaging methods have increased the reliability to delineate even subfields of the hippocampal formation in in-vivo measurements. Specialised T2-weighted sequences with anisotropic voxels and a limited field of view (FoV) are able to depict the layered structure of the hippocampus [115]. This exchange, of a lower resolution along the main axis of the hippocampus (HC) for a high coronal in-plane resolution, is motivated by the inner structure of the *HC*, which resembles a Swiss roll. Changes of the spiral formation occur less rapidly alongside its major axis, which is almost parallel to the anterior-posterior direction [52].

The first requirement for a standardised assessment of the volume and atrophy (tissue loss) of *AD* related structures is a common acquisition procedure for the underlying images. Harmonising the input data is crucial for the required manual segmentation protocols and especially for automated methods. To achieve this, the Alzheimer's Disease Neuroimaging Initiative (ADNI) was launched in 2003 by several institutes as well as private companies and non-profit organisations. It is a longitudinal multicentre study specifically designed to establish biomarkers (clinical, imaging, genetic and biochemical) for early *AD* detection and tracking of the disease. Further information can be found in [54, 100] or on their web-page[1]. By 2015 *ADNI* data includes more than 1500 subjects (age range of 55 – 90 years) consisting of cognitve normal (CN) elderly, people with mild cognitive impairment (MCI) and patients with early *AD*, which have been used in over 1500 publications. In late 2016 (*ADNI*-3) began with the aim to study the rate of change of cognition, function, brain structures and biomarkers.

The *ADNI* thoroughly evaluated *3D* T1-weighted sequences for morphometric analysis and also decided to use *3D MPRAGE* sequences, with an isotropic resolution of $1 \times 1 \times 1\,\mathrm{mm}^3$, for their structural whole brain scans. Moreover, they apply several sequences, for T2-contrast, to assess certain pathologies and for the high-resolution visualisation of the *HC* [44, 100]. For the latter, an in-plane resolution of $0.39 \times 0.39\,\mathrm{mm}^2$ together with a slice thickness of 2 mm is used and utilised for the subfield evaluation.

While the *ADNI* data sounds promising, there are some issues which make it impractical for us to use. Because of the large slice thickness of 2 mm the T2-weighted image data does currently not match our high requirements for the T2-weighted visualisation of the *HC*, despite the very high in-plane resolution. This thesis is a first line of work and aims to proof a concept of a machine learning based automated *HC* segmentation. To include data

[1] http://adni.loni.usc.edu

showing the *HC* best and to achieve consistency within the dataset by avoiding imaging data which already expresses age or disease related changes we restrict our data to healthy, young subjects. However, the *ADNI* data includes multi-centre data from a wide age range and medical conditions. Therefore, a unique dataset is acquired and utilised in this thesis.

2.2 Segmentation of Medical Images

Segmentation is the process of identifying and differentiating certain structures or regions from another. This partitioning is usually done based on homogeneous or similar areas defined by common attributes like intensity, colour, texture, shape, depth or suchlike. While the applications and needs for good segmentations are manifold, there is no single approach to all delineation problems. On the contrary, it strongly depends on the specific task and to date, the best option — the gold standard — still remains manual labelling.

The following described methods, for both manual and automated segmentation, are limited to the hippocampal formation and *AD*-related structures, such as the *MTL*.

2.2.1 Manual HC-Segmentation Protocols

Since atrophy of *MTL* structures has become an established biomarker for different types of cognitive disabilities, a lot of effort was made to construct common criteria for how to manually label them on *MRI* data. Such segmentation protocols exist for different *MRI* acquisitions, including field strengths and contrasts.

Segmentation protocols for hippocampal subfields exist for both, 3T ([25]) and 7T ([8, 90, 105]) scanners. To the authors knowledge, instructions based on T1-contrast are only available for 3T images ([70]). While for lower field strengths much more instructions are available, they often do not distinguish between the different subfields of the *HC* and provide less detailed delineation explanations. This is, because 7T imaging provides more consistent slice-by-slice visualisation, which is especially useful for tiny structures such as the hippocampal formation, due to ultra-high resolutions and increased signal-to-noise ratio (SNR).

Until 2017 many of the 7T protocols were limited to the *HC* body. Moreover, except of Wisse et al. [105] and Suthana et al. [90] none of the other available protocols incorporated novel findings from Ding and Van Hoesen [30, 31] about subfields of the perirhinal cortex (PrC) and the hippocampal formation. Berron et al. [8] incorporated these findings and the benefits of high-field *MRI* and proposed a protocol for manual segmentation of *MTL* subregions. To establish these instructions, 22 subjects (mean age of 26 years) underwent *MRI* examination on a 7T scanner (Siemens, Erlangen, Germany), in which T2-weighted partial *TSE* images with an in-plane resolution of $0.44 \times 0.44\,\mathrm{mm}^2$ and a slice thickness of 1 mm were acquired, oriented perpendicular to the long axis of the hippocampus. Of this dataset, a total of 24 hemispheres was then manually labelled by two expert raters and analysed for intra- and inter-rater reliability. The utility of the protocol has been tested with 35 participants, including 29 novices, in a hosted segmentation workshop [8].

To achieve *ADNI*'s goal of a standardised evaluation for imaging biomarkers, also attempts for a harmonised segmentation protocol have been conducted [9]. The authors have performed a review of 56 segmentation instructions, yielding 12 protocols (using a *3D* T1-weighted sequence) that were eventually included and evaluated for their accuracy and translation into practice. Finally, harmonised landmarks and differences were extracted. This resulted in the Harmonized Protocol (HarP)[2] for manual segmentation of the whole *HC* boundaries on magnetic resonance (MR) scans [10–12].

While the European Alzheimer's Disease Consortium (EADC)-*ADNI* effort focused on the delineation of the *HC* as a single structure, Yushkevich et al. [114] performed a similar study, however, with different *MR* modalities and focusing on (para-) *HC* subfields.

Fischbach-Boulanger et al. [35] investigated which of T1- or T2-weighted imaging is the superior sequence for visually assessing hippocampal atrophy. Therefore, visual ratings (medial temporal lobe atrophy (MTA) score) of 100 *MCI* and 50 *AD* hippocampi were independently acquired by two senior and two junior radiologists. Inter- and intra-rater reproducibility, accordance with a quantitative volumetric measure – obtained with *FreeSurfer*'s, discriminative power between the *MCI* and *AD* groups as well as correlation with several cognitive tests were used as the quality criteria. They have shown, that *MTA* scores for T1- and T2-weighted images show similar variability and consent with *FreeSurfer*'s volumetric measure. However, they suggest T2-weighted images for assessing *HC* atrophy in *AD*, as better discriminative power between the disease groups and higher correlation with several neuropsychological tests was achieved with T2-weighted images.

2.2.2 Automated HC-Segmentation

Automated segmentation, as a tool for fast, consistent and reliable assessment of various parameters (e.g. volume, cortical thickness...), especially for application to large data sets, has been pointed out by many authors among the medical imaging community. With the focus on the hippocampal formation, there are two main techniques used for segmentation — *atlas-based* and *learning-based* approaches.

Atlas-based methods can be subdivided into topological/deterministic or probabilistic/statistical atlases. While *deterministic atlases* are constructed only by a single subject, *probabilistic atlases* include a small amount of subjects, all co-registered to a standard space where the voxelwise frequency for each label is calculated. An *atlas* consists of two combined images, one intensity image (also template) and its corresponding delineation (also segmented/labelled image or atlas labels). For the segmentation procedure, unseen (target) images initially need to be registered with the atlases. Subsequently, the atlas labels can be transferred to the target image through several segmentation strategies. Registration is a fundamental biomedical image processing problem and refers to the process of establishing spatial alignment between images, by finding a transformation from one image space to the other.

[2] http://www.hippocampal-protocol.net

Label propagation, the easiest and fastest option is used for a single-atlas, where only labels of one sample exist. *Label fusion*, the process of combining propagated labels, is a major element in multi-atlas segmentation (MAS) methods, where several single-atlases are utilised. This approach has the benefit of comprising anatomical variability into the segmentation results and reducing atlas-specific registration errors. However, with the drawback of at least linear increased computational costs [51].

Probabilistic-atlas methods use a defined standard space to which the unseen images are registered during the first step. In this image space it is then possible to compile statistics about intensity, global position or neighbouring structures. This can be done as part of a Bayesian framework. In the subsequent segmentation step, each voxel is then labelled as the structure of interest or background. This is based on a-priori estimated probabilities obtained from the atlases. Fischl [36] investigated such methods, as one of the first, for neuroanatomical segmentation, yielding the well-known software package *FreeSurfer*.

FreeSurfer has become one of the most popular software for automated segmentation of neuroanatomical structures on *MR* images, taking the spot as the benchmark tool for whole cortical brain parcellation. This cortex segmentation consists of 21 classes, also includes the *HC* as a single structure, and takes up to 10 hours for one subject on a single central processing unit (CPU). Since the first version (v5.3 [36]), a lot of improvement and additional features have been implemented. The successor, version 6.0, was released in early 2017 and provides the possibility of a more sophisticated *HC* segmentation.

The latter is possible because of a new high-resolution post-mortem atlas, referred to as “ex-vivo atlas”, including hemispheres of 15 subjects with an average isotropic resolution of 0.13 mm^3, and a specifically designed protocol for labelling 13 hippocampal substructures [52]. Although ex-vivo (out of the living) data acquisition allows ultra-high resolutions and eliminates motion artefacts, it also results in a falsely brain tissue contrast. Additionally, missing perfusion (blood flow) and the fixation process cause different visualisation compared to in-vivo (within the living) measurements. Also registration of post-mortem scans becomes a harder task due to different *FoVs* and unique local features. Lastly, ex-vivo data is much harder to obtain, which yields to an overall trade-off between the advantages and disadvantages of ex-vivo atlases.

Since 29th April 2020 the latest version of *FreeSurfer* (v7.0) is available. While further improving the runtime, the hippocampal subfield segmentation remains unchanged.

Atlas-based approaches strongly depend on the underlying registration of the target images and the template. This is particularly a hard task if different contrasts/modalities are used. Detailed information of atlas-based approaches be can found in the following review-papers Cabezas et al. [17], Dill et al. [29], González-Villà et al. [39], Iglesias and Sabuncu [51].

Deep learning is a sub-class of *machine learning (ML)* algorithms used to model high-level abstractions of data by utilising artificial neural networks (ANNs), which are inspired by the neural system of humans. Learning-based approaches have been extensively utilised within recent years as powerful, specialised hardware has become more affordable and a lot of, still ongoing, effort is made to establish publicly available databases, providing imaging data and their corresponding annotations.

Similar to probabilistic atlas-based methods, *ANNs*, aim to leverage data from multiple subjects, known as training set, and learn features like intensity or shape for example. Especially convolutional neural networks (CNNs) ([62]), which are *ANNs* based on convolution operations, are widely used in medical image segmentation or classification. Recent reviews about the use of *CNNs* in medical image analysis can be found in Ker et al. [57] or Anwar et al. [4]. Another overview, focusing on convolutional architectures for the task of *3D* segmentation of the *HC*, was given back in 2015 by Lai [63] under supervision of Prof. Rueckert.

As for all data-driven methods, these techniques need an extensive amount of labelled data. Ronneberger et al. [80] have implemented a specialised training strategy for biomedical image segmentation, which allows to localise and can be trained with even little annotated data. Their proposed architecture, known as *U-net*, strongly depends on data augmentation using elastic deformations. The benefit of data augmentation concerning invariance has previously been shown by Dosovitskiy et al. [32]. As annotated data is a common bottleneck in the medical imaging domain, the U-net architecture is used in many applications ([18, 37, 42, 99]).

Other *CNNs* based methods, specifically used for segmentation of the hippocampus, have been reported in [20, 37, 108] where convolutional long-short-term memory (LSTM), multi-modal data and dilated deep supervision is utilised, respectively.

Among recent literature, also deep learning approaches comparable to *FreeSurfer* exist. A *3D* deep *CNN* for automated segmentation of 25 brain structures in T1-weighted *MR* images, called DeepNAT, was proposed in 2018 by [97]. Tools, only applicable to segment the hippocampus, have been proposed by [93] and [42], called HippoDeep and *HippMapp3r*, respectively. While the latter has been validated against other state-of-the-art methods (FIRST, *FreeSurfer*, HippoDeep) and was shown to be an efficient and robust tool for *HC* segmentation to assess neurodegenerative changes, it lacks flexibility. In detail, input images are required to have a certain orientation and acquisition sequence (contrast). Thus, it fails segmenting T2-weighted input images. Moreover, only whole *HC* segmentation, but no subfield labelling, is possible due to missing high-resolution data and different modalities.

2.3 Image Synthesis

Image-to-image translation refers to a class of computer vision problems, with the ultimate goal to map an input image to an output image. Examples problems are: image denoising, where image enhancement is achieved by noise reduction (improving *SNR*) of the input; image super-resolution, to create output images with higher resolution compared to the input; or image synthesis, in which the input is mapped to a different output modality, usually keeping the same resolution but expressing different styles/characteristics. In this thesis image-to-image translation refers to the class of image synthesis problems and are henceforth used interchangeably. Typical computer vision applications of image-to-image translation are changing season of landscape images or applying different painting-styles (van Gogh or Monet) or colouring black-white images and vice versa [53, 116].

In the medical imaging field, image synthesis is used in the context of modality transfer, including translation of computed tomography (CT) to *MR* images ([46, 48, 99]) or different translations specific to *MRI* (T1-/T2-contrast mappings ([86]) or generation of 7T images based on a 3T input ([74])).

Generative adversarial networks, first introduced by Goodfellow et al. [41], are frameworks which utilise an adversarial process to estimate generative models. In this adversarial process, a *generative model G* and a *discriminative model D* are simultaneously trained. In the initially proposed implementation of generative adversarial networks (GANs) ([41]), the *generator G* learns to generate new sample images from input noise. While G estimates the data distribution and is trained to maximise the probability of D being wrong, the *discriminator D* evaluates whether the current sample is more likely to be from the training data instead of being created by G. D is trained to maximise the probability for this correct labelling for both data sources.

The standard implementation of *GANs* ([41]), however, showed instability issues during training. Many of these problems, however, have been tackled with the utilisation of Wasserstein GANs (WGANs) [5]. The training of these *WGANs* was then even further improved by introducing a gradient penalty [43]. However, the output of image synthesis tasks statistically depends on the input images. This property can be exploited by means of conditional GANs (cGANs) ([68]), where the source image is provided as an input to condition the generator.

Motivated by [53], Yang et al. [109] introduced *cGANs* ([68]) into the field of brain *MRI*. In detail, Yang et al. [109] utilised *cGAN* for translation of T1-weighted *MR* images to mimic T2 contrast and applied the proposed cross-modality generation framework to registration and segmentation problems. During the course of this thesis, Dar et al. [26] also proposed a new method, which is based on *cGANs*, for multi-contrast *MRI* synthesis. The huge impact of *GAN* based methods in image-to-image translation problems is due to the concept of an adversarial loss.

3 Motivation

Various factors raise the need for a feasible and reliable segmentation of the hippocampus (HC). High-resolution T1-weighted scans, as utilised in clinical magnetic resonance (MR) studies, show only limited capability to clearly separate the *HC* from surrounding tissue and *CSF*. Moreover, they suffer from distinct visualisation of fine structures and subfields of the hippocampal formation, which are also beneficial for the manual ground truth labelling. T2-weighted *MR* scans are able to visualise these structures due to its complementary contrast information as well as higher resolutions [35, 107]. However, such scans are not feasible in clinical practice and are also not common in research studies.

As T1-weighted structural brain scans are standard during magnetic resonance imaging (MRI) examinations, a lot of T1 image data is available, however, necessary ground truth segmentations (GT-labels) are often missing. While the gold standard, manual annotation, is an exhaustive and time consuming task, many automated applications are impractical, like *FreeSurfer* with processing times of up to 10 hours per subject. Especially for evaluating disease-modifying drugs, which aim to slow down the progression of Alzheimer's disease (AD), consistent and reproducible segmentations are crucial to compare results across different studies. To ensure this consistency, a standardised approach for hippocampal volume assessment is required.

To address the poor visualisation of the hippocampal formation and to create accurate ground truth labels, a unique dataset is specifically acquired. This dataset is comprised of corresponding pairs of high-resolution T1- and T2-weighted 3T *MR* images from healthy volunteers. Based on novel T2-based instructions, 28 hippocampi were manually labelled on our T2-weighted images to establish the corresponding GT-labels. This dataset is subsequently preprocessed to cope with possible artefacts emerged from or during the generation process. Deep learning has become a powerful tool in medical imaging to learn complex models, based on annotated training data, which can then be applied to previously unseen input data. Thus, it is used for the automated segmentation task in order to limit the acquisition of the ultra high-resolution T2-weighted images as well as the exhaustive manual annotation by a trained annotator for the generation of the dataset, while exploiting the superior contrast and resolution of the T2 scans.

The aim of this thesis is to evaluate the hypothesis if deep learning with convolutional neural networks (CNNs) is able to outperform the hippocampal segmentation achieved with *FreeSurfer* — the clinically used benchmark in brain cortex parcellation, including the *HC* and its subfields. Additionally, the benefit of having ultra high-resolution T2-weighted scans and their impact on the training process of the *HC* segmentation task is evaluated.

To address the first question, *FreeSurfer* is used to generate the automated baseline segmentations of our manually labelled T1-weighted images. For our learning-based experiments, we utilise deep artificial neural networks (ANNs) to perform the segmentation task on our custom dataset. In more detail, deep convolutional neural networks (DCNNs) are trained with both image modalities and our manual ground truth labels. The results of the segmentations performed on T1 input images are used to evaluate the performance of the learning-based segmentations in comparison to the mask achievable with *FreeSurfer*.

To assess the impact of our T2-weighted images on the segmentation accuracy, the same architecture, which is utilised in the T1-based setup, is trained with the T2 input images. Moreover, the same setup is used with a combined input, consisting of the T1- and T2-weighted images, to evaluate the contribution of the T2 contrast.

As the advantage of the T2-weighted images can unfortunately only be exploited for training our *DCNNs*, methods incorporating the T2 scans into the training process are proposed. With the assumption of having a distinct benefit with ground truth labels that have been created based on T2 contrast and using the T2 images only during training, a specific deep learning approach – generative adversarial networks (GANs) – are utilised to create synthetic T2 images of the underlying T1-weighted images. This image-to-image translation is executed as an auxiliary task prior segmentation and meant to enhance the final segmentation result rather than focusing on a perfectly mimicked T2 contrast.

With the assumption of having a distinct benefit with ground truth labels that have been created based on T2 contrast and using the T2 images only during training, a *GANs* is proposed to create synthetic T2 images of the underlying T1-weighted images. This auxiliary task is used to create synthetic images, mimicking T2 contrast of the T1-weighted input images, on the fly prior the actual segmentation via *DCNN*.

3.1 Outline

The following chapter provides background information related to the main tasks of this thesis, including *MRI*, registration and the image synthesis with *GANs*.

Chapter 5 describes the acquisition process of our customised dataset and applied preprocessing steps. Moreover, the *FreeSurfer* pipeline as well as the learning-based models used for segmentation and the image-to-image translation task are introduced.

In Chapter 6 the experimental setup is given, including the training process, a list of all conducted experiments and their evaluation process. Subsequently, quantitative and qualitative results are shown and thoroughly discussed in chapter Chapter 7.

Eventually, Chapter 8 concludes the thesis and provides an outlook for future work.

4 Theory and Background

Contents

This chapter provides supplementary information about the principles of the utilised concepts in this thesis. At first, a basic understanding of the physics behind magnetic resonance imaging (MRI) and some concepts related to its contrast mechanisms are described. Following, a brief overview of medical image registration and its necessary elements like allowed transformations, the similarity metric as well as the optimiser is given. The last part presents common terms and concepts of neural networks to provide a basic understanding for the utilised architectures.

4.1 Magnetic Resonance Imaging - MRI

This section provides a short insight into basic theory of *MRI*, explaining the concepts required to understand contrast mechanisms and the differences between both sequences utilised in this thesis. Moreover, the benefits and drawbacks of each sequence are presented.

4.1.1 Spins and Signal

Clinical *MRI* utilises hydrogen atoms (H^1) respectively their nuclei, which include only a single proton. As every elementary particle, protons have the intrinsic property of a nuclear spin. Based on their atomic mass, protons have a spin quantum number $s = \frac{1}{2}$, yielding two states: m_s of $+\frac{1}{2}$ = “spin up” or $-\frac{1}{2}$ = “spin down”. This spin causes two effects; an *angular momentum* $\boldsymbol{L}$ and a *magnetic moment* $\boldsymbol{\mu}$ which can be expressed in terms of the *gyromagnetic ratio*:

$$\gamma = \frac{\boldsymbol{\mu}}{\boldsymbol{L}} \ . \tag{4.1}$$

The spinning proton or *spin* (both are used synonymously) can be considered as a bar like magnet. Based on the Boltzmann distribution, the parallel ("spin up") configuration in an external magnetic field is at a minimal better energy level, which yields slightly more nuclei with a "spin up". This results in a macroscopic equilibrium magnetisation $M_z = M_0$ per volume unit:

$$M_z = \rho \cdot \frac{\gamma^2 \hbar^2}{4kT} \cdot B_0 \ , \tag{4.2}$$

with ρ... proton density, γ... gyromagnetic ratio, $\hbar$... reduced Planck constant, k... Boltzmann constant, T... absolute temperature and B_0... external magnetic field.
This equilibrium magnetisation is aligned in parallel to the external B_0 field (along the z-axis). However, in magnetic resonance (MR) scanners the normal of the receiver coil is always perpendicular to the z-axis, which implies that only signals in the transversal (XY-) plane can be measured. The acquired magnetic flux Φ is then proportional to the transversal component (M_t) of M_z. Therefore, to actually measure a signal in the detector, the equilibrium magnetisation M_z needs to be "flipped" into the transversal plane, where it induces a voltage. This excitation of the spin-system is achieved with high frequency (HF), also called radio frequency (RF), -pulses of the right power. The induced *HF* signal can be described by:

$$U(t) = k \cdot M_t \cdot \omega_0 \cdot \sin(\omega_0 t) \ . \tag{4.3}$$

The angular frequency ω_0, known as the *Larmor frequency*, describes the precession of the magnetic moment about an external magnetic field. It depends on the gyromagnetic ratio and the magnitude of the applied magnetic field:

$$\omega_0 = \gamma \cdot B_0 \ . \tag{4.4}$$

Common field strengths for clinical *MRI* are 1.5 or 3 tesla (T).

4.1.2 Contrast Parameters

MRI has great diagnostic value due to the possibility of using specific sequences to either enhance or attenuate different tissues in a *MR* image. Three main tissue parameters are responsible for the corresponding intensity (brightness) in the resulting *MR* image and hence its contrast. Two characteristic times known as *spin-lattice relaxation time* (*T1*) and *spin-spin relaxation time* (*T2*), and the proton density (PD).

The ***PD*** defines the amount of excitable spins (protons) per volume unit and therefore defines the maximum signal of a tissue.

T1 is the time constant of the *longitudinal relaxation*, which is related to the dissipation of absorbed energy into the surrounding molecular lattice. It determines how fast the tissue returns to its equilibrium state after an excitation and depends on the tissues thermal conductivity and strength of the static magnetic field B_0.

T2 is the time constant of the *transversal-* or *T2 relaxation* and determines how fast the *MR* signal attenuates, due to spin-spin interactions, after an excitation. This attenuation happens due to dephasing of spins without any energy dissipation to the vicinity and is more or less independent of B_0. The spins rather exchange energy among each other because of fluctuating (fast changing) local magnetic field variations. These fluctuations happen because the spins, as tiny magnets themselves, change the field strength for one another. Thus, as the precession frequency depends on the magnetic field B_0, the spins precess at different speeds causing a loss in phase coherence.

However, such field strength variations can also be due to constant inhomogeneities of the external B_0-field. They are caused by the hardware itself and also by the body of the examined subject. Therefore, the overall signal decay happens with the apparent time constant *T2** and is also know as the free induction decay (FID):

$$\frac{1}{T_2^*} = \frac{1}{T_2} + \frac{1}{T_{2i}} = \frac{1}{T_2} + \gamma \Delta \left| B_0 \right| \ , \tag{4.5}$$

with T_2... spin-spin time const., T_{2i}... time const. due to inhomogeneities, γ... gyromagnetic ratio and $\gamma \Delta |B_0|$... contribution of intravoxel field inhomogeneities.

Both, T1 and T2 relaxation occur simultaneously and are independent of each other. However, T2 relaxation ($100 - 300$ ms) is much faster so that the *MR* signal is gone before M_z would be recovered to the equilibrium state because of T1 relaxation ($0.5 - 5$ s) [101]. Relaxation times are tissue dependent with a coarse division [69] of body tissues in:

	Tissue		
Type	Examples	T1	T2
fluids:	CSF or oedema	1500 – 2000 ms	700 – 1200 ms
water-based tissue:	brain, muscle or cartilage	400 – 1200 ms	40 – 200 ms
fat-based tissue:	fat or bone marrow	100 – 150 ms	10 – 100 ms

Table 4.1: Tissue depending T1 and T2 relaxation times.

As the mentioned tissue parameters (ρ, T1, T2, T2*) are properties of the imaged tissue, additional timings need to be utilised to make use of the mentioned contrast behaviours.

The ***repetition time* TR** is the time between successive *RF* excitation pulses. It is essential for the T1 contrast of the image as it determines how much time the excited spins have to regain the equilibrium magnetisation M_z. This in turn, defines the available longitudinal magnetisation for the next excitation. A “short” TR (in comparison to T1; usually <600 ms), will result in a pronounced T1-weighting. If a “long” TR (above ≈ 1500 ms) is chosen, almost all tissues will have time to fully recover, whereby the T1 influence on the contrast gets minimal.

The ***echo time* TE** is the time between an excitation and the actual measurement, so the read out of the signal. It defines the T2 influence onto the image contrast. "Short" echo times (<30 ms), yield only little T2-weighting as spin-spin relaxation processes just began. Therefore, to get a good T2-contrast, TE should be "long" (>75 ms) together with a "long" TR (>1500 ms) to minimise the T1 influence. A short overview of the possible combinations of TR and TE and the resulting contrasts is shown in Table 4.2.

	"Short" TR	**"Long" TR**
"Short" TE	T1-weighting	PD-weighting
"Long" TE	poor contrast	T2-weighting

Table 4.2: Summary of possible contrasts depending on the combination of sequencing times (TR and TE).

Depending on the chosen timings, three main types of contrasts can be achieved, namely *T1-weighted*, *T2-weighted* or *PD-weighted* images. In T1-weighted images, tissues with long T1 give the least signal whereas short T1 yields the most signal. This results in dark and bright pixels, respectively. T2 shows a complementary contrast where tissues with long T2 yield the highest signals, resulting in bright pixel intensities. For *PD*-weighted images also high PD-values result in high signal intensities.

4.1.3 Pulse Sequences

With varying the imaging technique, also known as *pulse sequences*, and controlling their timings, a large range of contrasts can be achieved with *MRI*. A pulse sequence is a carefully scheduled series of excitation (*RF*) pulses combined with gradient pulses to allow for spatial encoding of the measurements. The (temporal) correspondence of the applied pulses is commonly visualised with pulse sequence diagrams (PSDs). Moreover, such diagrams allow to infer the strategy of the spatial encoding of a pulse sequence.

Common terms for labelling a *PSD* and their description are shown in Table 4.3.

Terms	**Description**
RF	excitation or refocussing pulse
Slice / SS / $\mathrm{G_{SS}}$ / $\mathrm{G_z}$	slice selection (gradient)
Phase / PE / $\mathrm{G_{PE}}$ / $\mathrm{G_y}$	phase encoding (gradient)
Readout / FE / $\mathrm{G_{FE}}$ / $\mathrm{G_x}$	frequency encoding (gradient)
Signal / Echo	measured signal

Table 4.3: Explanation of common labels used to label PSDs.

In general, a pulse sequence in *MRI* consists of three components: (i) a preparatory module, (ii) the acquisition, and (iii) the recovery. The *preparation module* can include a variety of preparation pulses such as spectral saturation (for fat saturation), flow or spatial saturation pulses. In the *acquisition* phase any set of pulses and gradients to generate for example a spin echo (SE) (Fig. 4.2) or gradient echo (GE) (Fig. 4.1) are included. The *recovery*, is in principle a “dead time (TD)” without any signal generation to allow the system to return to equilibrium.

4.1.3.1 Gradient Echo vs. Spin Echo

In general, there are two principle sequence families: gradient recalled echo (GRE) or *GE* and *SE* sequences. With *GE* sequences either T1-, T2*- or *PD*-weighted images can be achieved. *SE* sequences can produce T1-, T2- or *PD*-weighted scans.

As mentioned, the spin-spin relaxations (T2 and T2*) happen due to dephasing of the excited spins. In *MRI*, the moment where the dephased spins are back in phase is called an *echo*. At this point in time, the spins add up creating the strongest signal and are therefore measured in this state. The principle how this state of phase coherence is achieved, is the main difference between a *GE* and a *SE* sequence.

A **gradient echo** is produced by a single *RF* excitation pulse together with a *de*-phasing and *re*-phasing gradient pair. In comparison, a **spin echo** is generated by means of a *RF* pair, where the first (90° *RF*-) pulse excites the spins. After exactly $TE/2$, a refocussing (180° *RF*-) pulse is applied which creates the *SE* and also reverts the dephasing caused by B_0 inhomogeneities (only T2-decay of signal in Fig. 4.2). Correction of B_0 inhomogeneities cannot be done with *GE* sequences, which results in a T2* decay of gradient echoes (see T2*-decay in Fig. 4.1). The random, intrinsic T2-dephasing cannot be reverted with either of both sequences, as described in Section 4.1.2.

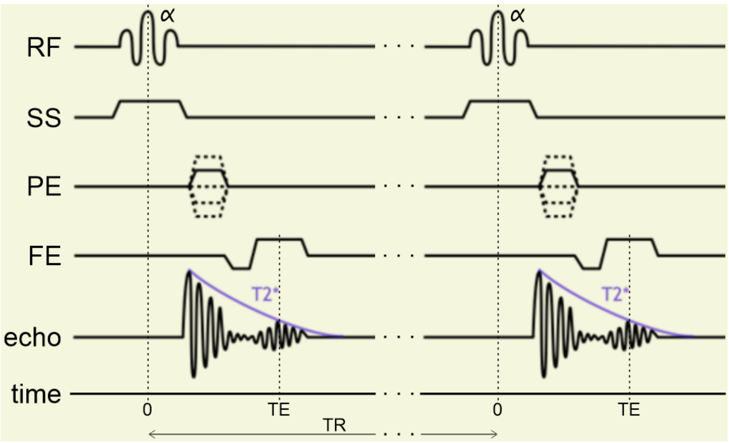
4.1.3.2 Gradient Echo and Spin Echo -Sequences

Following, the *PSD* of a basic *GE* (Fig. 4.1) and *SE* (Fig. 4.2) sequence is described and shown.

Common to both sequences (*GE*, *SE*), initially an excitation (*RF*-) pulse is used to flip the magnetisation. While in *SE* sequences a 90° excitation pulse is used, *GE* sequences use a variable flip angle α smaller than that. This excitation pulse is applied in combination with a selective gradient (SS) to excite either a single slice or a whole slab, depending whether a two-dimensional (2D) or three-dimensional (3D) sequence is used. The phase encoding (PE) gradient, applied in both sequence types, is used for spatial encoding of the measurement along the y-axis. If a *3D* acquisition is used, together with the phase encoding an additional slice encoding gradient (along the z-axis) is applied. As mentioned before, the principle to create the echo differs now for both sequences.

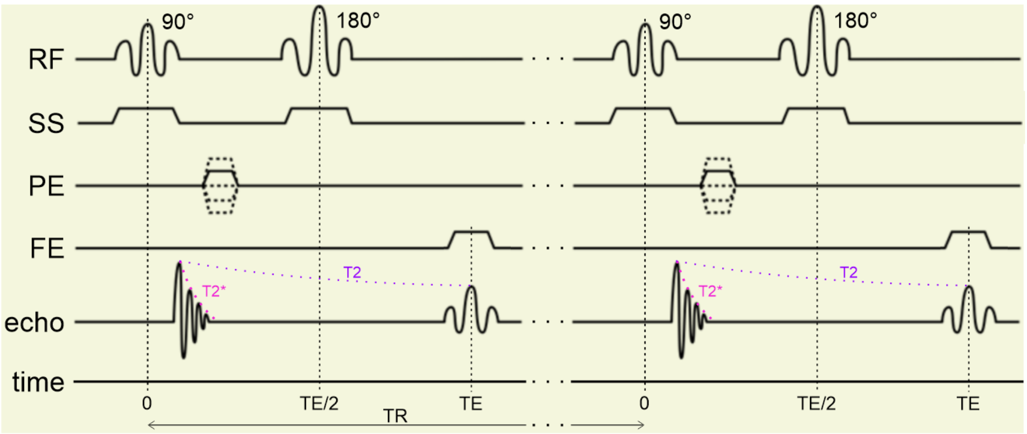
Gradient echo sequences (Fig. 4.1) use a negative *de*-phasing gradient before the frequency encoding (FE) gradient, which is applied during the readout (RO) of the signal. This *de*- and *re*-phasing gradient pair (see FE in Fig. 4.1) at first causes an additional dephasing, which is then, however, rephased with the first half of the positive FE gradient, such that the maximal signal is measured at TE (the middle of the positive FE gradient).

Spin echo sequences (Fig. 4.2) make use of a 180° *RF*- refocussing pulse, which is again applied with the selective gradient of the excitation, to create an echo after TE/2.



modified from http://xrayphysics.com/sequences (accessed 28.05.2020)

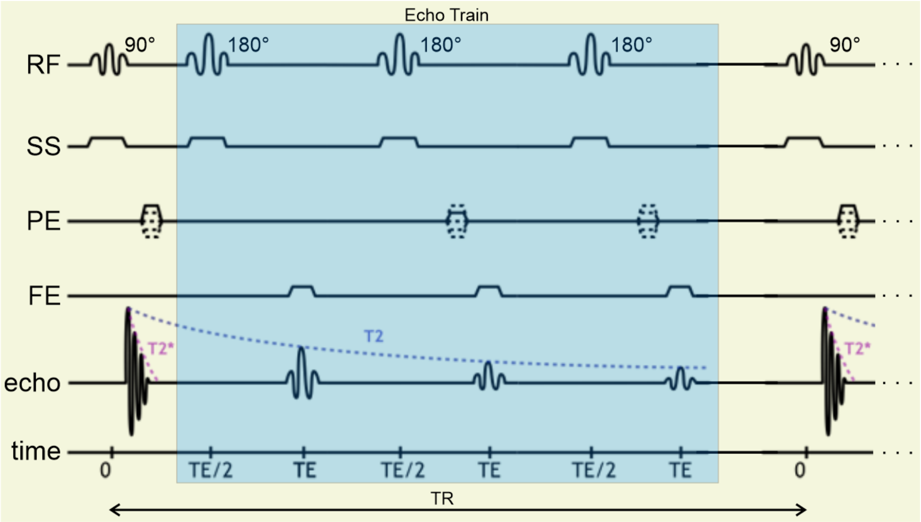
Figure 4.1: Pulse-Sequence-Diagram of a basic GE sequence.



modified from http://xrayphysics.com/sequences (accessed 28.05.2020)

Figure 4.2: Pulse-Sequence-Diagram of a basic SE sequence.

To utilise the long repetition times in *SE* sequences, a rapid acquisition method, called **turbo spin echo (TSE)** or fast spin echo (FSE), can be employed. Here, not only one 180° refocussing pulse is used after one 90° excitation pulse, but rather a series of refocussing pulses to create a whole *echo train*. With this method, several lines of the image are acquired with a single 90° *RF*- excitation. Thus, also the overall scan time is reduced. Figure 4.3 shows the *PSD* of an example *TSE* sequence with one repetition cycle and an echo train length (ETL) of three.



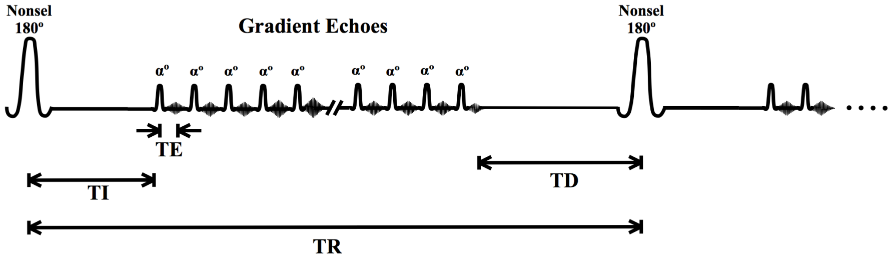
modified from http://xrayphysics.com/sequences (accessed 28.05.2020)

Figure 4.3: Pulse-Sequence-Diagram of a TSE sequence with an ETL = 3.

4.1.3.3 Inversion Recovery Sequences

Inversion recovery (IR) is an imaging technique which has such an additional preparation module prior the "standard" imaging sequence. As the name implies, an additional 180° *RF-inversion pulse* is used to flip the longitudinal magnetisation M_z in the opposite (negative) direction. This technique introduces another sequence timing, called *inversion time TI*. It is the time between the 180° *RF*- inversion pulse and the actual *RF* excitation pulse. During the TI periode, inverted spins of all tissues undergo longitudinal (T1) relaxation. Therefore, TI can be used to adjust the degree of separation based on the tissues intrinsic T1 times. This allows for example for fat or water suppression. The *IR* principle is used as a preparation in several *MR* sequences, including fluid-attenuated inversion recovery (FLAIR) or our utilised magnetisation-prepared rapid gradient-echo (MPRAGE).

Figure 4.4 shows a schematic representation of the pulses in an *MPRAGE* sequence. Note, that after one inversion pulse a series of *GEs* is rapidly acquired to exploit the long recovery time ($\approx 2000\,\text{ms}$). Usually, small flip angles α and short echo times around 5 – 12° and 2 – 4 ms are used, respectively.



from http://mriquestions.com/ir-prepped-sequences.html (accessed 30.05.2020); permitted use as a "Courtesy of Allen D. Elster, MRIquestions.com"

Figure 4.4: Simplified schematic of pulses used in a MPRAGE sequence.

4.1.4 Summary

The big advantage of *SE* sequences is the insensitivity to field inhomogeneities and therefore good image quality. This is tied to longer scanning times which, in combination with the often utilised very high resolutions, account for the sensitivity to displacement artefacts. In comparison, *GE* sequences are much faster due to the missing refocussing pulse and therefore possible shorter repetition times. The short echo times and usually lower resolution make them less sensitive to motions artefacts. However, shorter TEs might also reduce the signal-to-noise ratio (SNR). General artefacts inherent to *MRI* are shown specifically for our data and are therefore reported in Section 5.1.1.3.

4.2 Registration

In this section a brief overview of medical image registration and its elements, which need to be defined depending on the specific application, is given. These elements include the allowed transformations, used similarity metric as well as the optimiser to find the best mapping.

It is common and often necessary in medical imaging to acquired images with different modalities or at several time points. The latter is already needed in longitudinal research studies and is especially crucial to assess atrophy of the hippocampal formation. However, various setups yield somewhat different orientations of the imaged structures or a changed structure itself, due to atrophy for example. Furthermore, already within the same *MRI* examination or sequence, motion artefacts can cause displacements of corresponding images that need to be corrected.

Registration of images involves the process of finding an optimal geometric transformation that maximises the correspondence between two images. Registration methods of different complexity and use cases exist. A categorisation, based on increasing complexity, could be feature based rigid registration (points or surfaces), intensity based rigid registration and generally deformable (non-rigid) registration.

Geometric functions can for example be categorised by the properties they preserve. A short outline is given in the next paragraph.

Transformations. Starting with *rigid* (also Euclidean) transformations rotations, translations and reflections are allowed. Thus, in *3D* up to six degrees of freedom (DOF) can be manipulated. Here, Euclidean distances, so distances and angles, are preserved between every pair of points. *Affine* transformations, are the next group which include all rigid transformations. Moreover, scaling and shear mapping is possible, which yields up to 12 *DOF*. This group still preserves parallelism. Affine transformations can be expressed as a combination of linear transformations followed by a translation, which is why all linear transformations are also affine. The class of non-linear transformations introduce elastics, splines or similar. Which of these models is the most appropriate strongly depends on the specific application and the registration task.

Interpolation. Computation of the transformation is practically done via *inverse mapping*. Here, the voxel intensities from the moving image are interpolated and put into the corresponding coordinate space of the fixed image. The similarity metric is subsequently calculated, also in the space of the fixed image.

The simplest interpolation method is *nearest neighbour* re-sampling. Instead of calculating an intensity value, merely the value of the closest voxel is taken, which means that the original intensity range is preserved. This yields to a piecewise-constant interpolation and is commonly applied to binary masks. However, in images it results in a “blocky” appearance. A more sophisticated method is *linear interpolation*. While those results are less “blocky”, the procedure is slower and also some high frequency image information is lost. There is a wide variety of other interpolation methods. However, linear interpolation is used in many tasks because of its compromise between accuracy and computational efficiency.

Similarity Metric. In order to compare or estimate the correspondence of two images some kind of measure is needed. Such measures are generally called similarity metrics/measures or sometimes registration basis. While two main approaches exist, namely feature- and voxel-based similarity measures, only voxel/intensity -based will be elaborated due to its application in this thesis.

Feature-based registration requires preceding feature extraction (landmarks or segmentation). Thus, errors during the feature extraction stage will propagate through the whole registration process. To avoid such errors, voxel intensities can be directly used to estimate the degree of common information. The benefit of this method is that no features need to be specifically extracted. The simplest intensity-based measure is based on the sum of squared differences (SSD).

$$S_{SSD} = -\frac{1}{N} \sum_{\mathbf{x}_A \, \epsilon \, \mathbf{\Omega}^{\mathbf{T}}_{I_A, I_B}} \left(I_A(\mathbf{x}_A) - I_B^{\mathbf{T}}(\mathbf{x}_A) \right) \, , \tag{4.6}$$

with N as the number of voxels in the overlapping domain $\mathbf{\Omega}^{\mathbf{T}}_{I_A,I_B}$ around the voxel $\mathbf{x}_A$. While it can be shown that *SSD* is the optimum measure for images of the same modality diverging only by Gaussian noise, its usage is restricted to mono-modal applications [83].

A less restrictive formulation, only based on the assumption of a linear relationship between both images, can be done by the (normalised) cross correlation (CC). One implementation of the *CC* is included in the Advanced Neuroimaging Tools (ANTs), which is an open source software library build on an Insight ToolKit (ITK) framework. The *ANTs CC* [6] is defined as:

$$CC(\mathbf{x}) = \frac{\sum_i \left(\left(I(\mathbf{x}_i) - \mu_{I(\mathbf{x})} \right) \left(J(\mathbf{x}_i) - \mu_{J(\mathbf{x})} \right) \right)^2}{\sum_i \left(I(\mathbf{x}_i) - \mu_{I(\mathbf{x})} \right)^2 \sum_i \left(J(\mathbf{x}_i) - \mu_{J(\mathbf{x})} \right)^2} \, , \tag{4.7}$$

where $\mathbf{x}$... the centre of an NxN squared window, μ... the mean value and $\mathbf{x}_i$... the running index of that window.

Moreover, information-theoretical measures exist which are based on information content or entropy of the registered images. One such information theory concept, that got applied to image registration [23, 67, 102], is the (normalised) mutual information (MI), which gets maximal when both images are aligned. *MI* is considered the most suitable option for multimodal registration, where images show very dissimilar contrast relationships.

Optimiser. Given a similarity metric, the registration algorithm tries to find a transformation that maximises the similarity between the source and target image. Optimisation of voxel-based similarity measures usually require iterative schemes such as gradient descent (GD), Gauss-Newton or quasi-Newton methods. Among the latter, conjugate direction methods or Broyden-Fletcher–Goldfarb-Shannon (BFGS) methods are commonly used.

Global optimisation schemes, such as exhaustive search, are not feasible for image registration. However, while local optimisation techniques are much more efficient, they can also get stuck in local optima. To increase the capture range of local optimisation schemes, multi-scale approaches can be used. Here, images are downsampled and registered. The solution of the coarse scale is then used as the initialisation of the next scale of the pyramid.

4.3 Machine Learning

This section presents a brief overview of the applied learning-based methods utilised in this thesis. Therefore, at first common terminology is given.

Artificial intelligence (AI) is an umbrella term used for all algorithms that allow machines to mimic the intelligence of humans, including machine learning (ML) and deep learning methods.

Machine learning is a subfield of artificial intelligence which applies statistical methods to find patterns in previously extracted features, learns from them and applies the acquired "skills" to new data while improving with experience. It can be described as the application of algorithms to extract features by analysing the given data, which can then be used to solve or make decisions about the given problem. In comparison to regular software, which is explicitly implemented to perform specific tasks, in *ML* only a set of tools (algorithms) is provided such that *machines can learn how to solve a given problem without being explicitly told how to do so.* The latter, italic part corresponds to a paraphrased definition of Arther Samuel from 1959 [61]. *ML* models can be divided into two main groups; *discriminative models* and *generative models.*

Discriminative Models. Discriminative approaches model the decision boundary between different classes. Therefore, mappings from the input space X to the output space Y are modelled via the conditional probability distribution $\mathbb{P}(y|x)$ of the labels y given data x.

Generative Models. Generative models on the other hand directly model the distribution of a class $\mathbb{P}(x|y)$. Thus, the joint probability distribution $\mathbb{P}(x, y)$. In the case of unlabelled data, the data distribution $\mathbb{P}(x)$ is modelled. In case of a classification task, the prediction is then determined by means of the Bayes theorem:

$$\mathbb{P}(y|x) = \frac{\mathbb{P}(x, y)}{\mathbb{P}(x)} = \frac{\mathbb{P}(x|y)\mathbb{P}(y)}{\mathbb{P}(x)} \ . \tag{4.8}$$

Because of the implicit or explicit modelling of the data distribution, generative models can also be used to generate new data based on this learned underlying distribution.

4.3.1 Deep Learning.

Deep learning is part of the field of *ML* and refers to a method used to model high-level abstractions in data by the use of artificial neural networks (ANNs). In more detail, it corresponds to a specific approach for setting up and training *ANNs*, where the input is passed through a series of non-linear transformations to acquire the model output. Thus, the network exhibits a certain depth and is therefore referred to as *deep.*

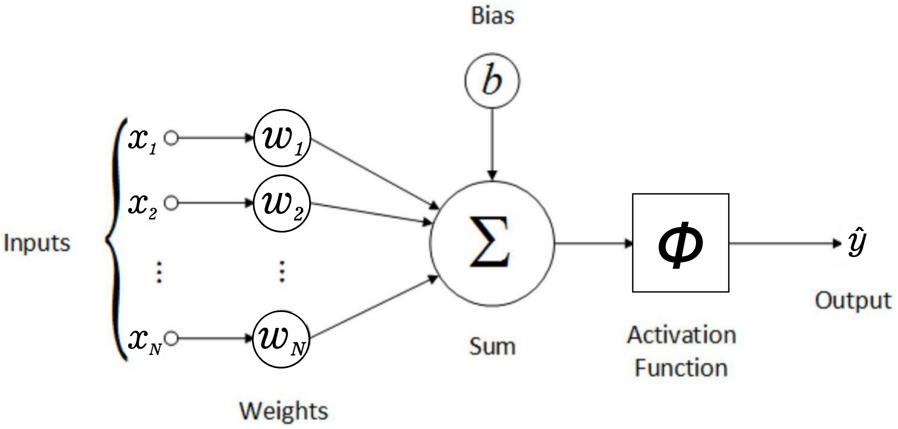
4.3.2 Supervised Learning.

One of the most frequently used *ML* type is *supervised learning*. As the name implies, supervised learning is the task of estimating a mapping function that maps a given input to a certain output. Therefore, an algorithm learns from training data which can be seen as a supervisor. Consequently, the training data consists not only of the input but also of the corresponding target output. In this thesis only supervised learning approaches are used.

4.3.3 Artificial Neural Networks

4.3.3.1 Artificial Neuron

The design of *ANNs* is inspired by the human neural system. Such interconnected networks are constructed in form of so called *layers*. The building blocks for all such layers are *artificial neurons*, which are composed of four parts: (i) the input, (ii) the weights and bias, (iii) a weighted summation, and (iv) an activation, as visualised in Fig. 4.5.



modified from Jayesh Bapu Ahire, 2018, https://medium.com/@jayeshbahire/the-artificial-neural-networks-handbook-part-4-d2087d1f583e (accessed 09.06.2020)

Figure 4.5: Schematic visualisation of a perceptron.

Mathematically an artificial neuron can be described as follows. Let the input be the set of N features $X = \{x_1, \ldots, x_N\}$ with the corresponding set of parameters $\theta = \{b, W\}$, which consists of the bias term b and the set of N weights $W = \{w_1, \ldots, w_N\}$. This yields the intermediate output, the weighted sum, of the neuron:

$$z = \sum_{n=0}^{N} w_n x_n + b \; . \tag{4.9}$$

To use a more convenient formulation, a constant input $x_0 = 1$ can be added to the input set such that the input and parameter set can be written as vectors with dimension $N + 1$,

$\boldsymbol{x} = (x_0, x_1, \ldots, x_N)$ and $\boldsymbol{\theta} = (b, \boldsymbol{w})$, respectively. The weighted sum of one neuron can then be calculated as the dot product between both vectors:

$$z = \boldsymbol{\theta} \cdot \boldsymbol{x} = \sum_{n=0}^{N} \theta_n \cdot x_n \ . \tag{4.10}$$

The weights can be interpreted in such a way, that positive weights activate the neuron while negative weights cause inhibition. The bias can be seen as an activation function dependent threshold of the sum that needs to be reached in order for the neuron to fire.

This weighted sum is then applied to an *activation function* ϕ to produce the neurons final output.

$$\hat{y} = \phi(z) = \phi(\boldsymbol{\theta} \cdot \boldsymbol{x}) = \phi(\boldsymbol{\theta}; \boldsymbol{x}) \tag{4.11}$$

In the simplest case the activation function is a *unit step* which maps the output to 0 or 1:

$$\hat{y} = \begin{cases} 1, & \text{if } \boldsymbol{\theta} \cdot \boldsymbol{x} \geq 0 \\ 0, & else. \end{cases} \tag{4.12}$$

A single artificial neuron with such an activation (Heaviside) is often referred to as *perceptron*, which is only capable of learning linear separable patterns. Therefore, more complex activation functions are needed for the model to perform also complex mappings. Non-linear activations allow to model arbitrarily complex mappings between the input and the output. Common functions, which have also been applied in this thesis, are shown next in Section 4.3.3.2.

4.3.3.2 Activation Functions

Activation functions are essential to introduce non-linearity into the model, which enables the network to learn complex, non-linear features of the data.

Linear activation functions lack in capability to model complex function, because independent of how many hidden layers are used, the network will always model only a linear function (from the input to the output), as a linear combination of linear functions is again linear.

A commonly used non-linear activation function, which does not suffer from this problem and is also computationally efficient is the rectified linear unit (ReLU) [72]. Based on the weighted sum z of an artificial neuron, it is given by:

$$\phi(z) = \max(0, z) \ , \tag{4.13}$$

and can be visualised as in Fig. 4.6a. However, based on this definition, *ReLUs* yield a zero for non-positive values $\phi(z)|_{z \leq 0} = 0$. This means that the neuron is deactivated and also the gradient will be zero. Therefore, the weights of deactivated neurons will not get updated during training. Ultimately, neurons which enter this zero-state will stop contributing to the model, which is called *dying ReLU* problem and again yields complications during training.

A solution to this problematic behaviour was introduced by Maas et al. [66] with a modified version called leaky rectified linear unit (leaky ReLU). Here, negative values are mapped as a linear function with an additional parameter α, controlling the tiny slope of the negative part, which ensures non-zero gradients over its entire domain. The *leaky ReLU* activation is given by:

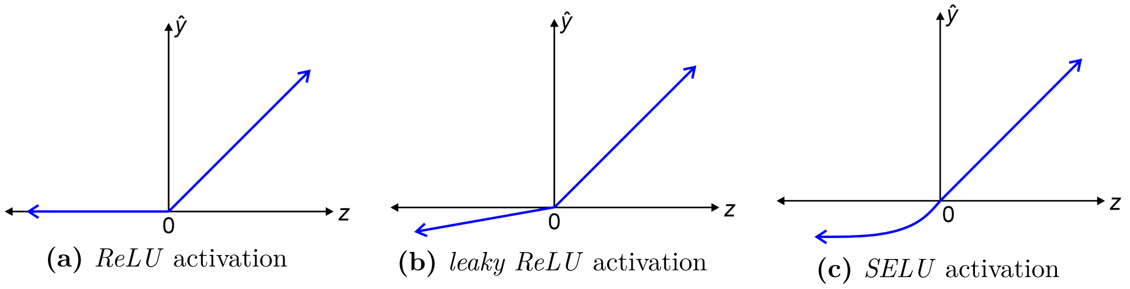
$$\phi(z) = \begin{cases} \alpha\, z, & z < 0 \\ z, & z \geq 0 \end{cases}, \tag{4.14}$$

and visualised in Fig. 4.6b. However, depending on α very small gradients can still prevent the network from learning. In general, small gradients result in little (no significant) change, which drastically slows down training or even hinders the network from learning. This effect is especially problematic with deep networks as the gradient information gets smaller while being back propagated through the network to the input, which is known as *vanishing gradients*.

Scaled exponential linear unit (SELU) is a relatively new activation function, which tackles not only vanishing gradients, but also their counterpart *exploding gradients* [59]. This activation was introduced in the context of self-normalising neural networks (SNNs), in which neuron activation automatically converges towards zero mean and unit variance. This improves training of neural networks and additionally converges faster. While the output behaviour for positive values is similar as in (leaky) *ReLU*, the negative part is handled via a scaled exponential. *SELU* activation is given by:

$$\phi(z) = \lambda \begin{cases} \alpha(e^z - 1), & z < 0 \\ z, & z \geq 0 \end{cases}, \tag{4.15}$$

where $\lambda \approx 1.051$ and $\alpha \approx 1.673$ are calculated values given in the original paper [59].



(a) *ReLU* activation (b) *leaky ReLU* activation (c) *SELU* activation

Figure 4.6: Visualisation of *ReLU*, *leaky ReLU* and *SELU* activation.

4.3.3.3 Feedforward Neural Networks

A single layer is now composed by many of the previously described artificial neurons. Together with an input and an output layer a simple *ANN* can be constructed. While the first and last layer is always called *input layer* and *output layer*, respectively, all layers in between are referred to as *hidden layers* (cmp. Fig. 4.7). Hidden layers include an arbitrary amount of artificial neurons, also termed (hidden) units or nodes.

Feedforward neural networks (FNNs) are networks that do not form any cyclic connection or loop between nodes. Information rather flows only in a forward direction from the input nodes via the hidden nodes to the output nodes, as shown in Fig. 4.7.

This is contrary to the group of *recurrent neural networks (RNNs)* where cycles can be formed. However, they are not used in this work.

4.3.3.4 Multi-layer FNN

To build more complex and powerful models, than achievable with a single hidden layer, multi-layer networks can be utilised. In such networks several hidden layers are stacked after one another, and are thus called *deep neural networks.* The network depth is herby given by the number of layers, whereas the network width is defined by the amount of artificial neurons per layer. If all nodes of a layer are connected with all nodes of the successive layer, it is referred to as a *fully connected layer.*

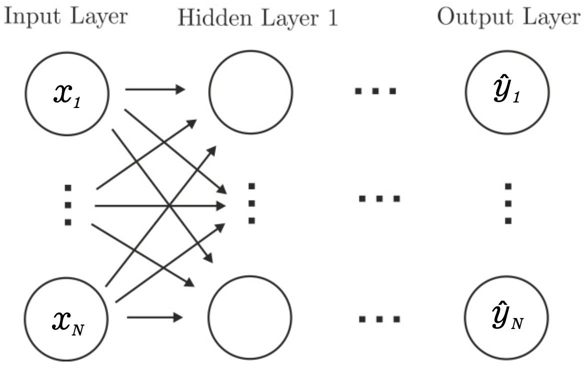


Figure 4.7: Schematic structure of a feedforward neural network.

Multi-layer *FNNs* achieve universal approximation capabilities [49] by composing the various mapping functions of each layer. A feedforward network defines a mapping of an input $\boldsymbol{x}$ to an output $\hat{\boldsymbol{y}}$ via a parametric function f:

$$\hat{\boldsymbol{y}} = f(\boldsymbol{x}; \boldsymbol{\theta}) \ , \tag{4.16}$$

and tries to find the best function approximation by learning the parameters $\boldsymbol{\theta}$.

For multi-layer networks, the composition of the functions is described by a directed acyclic graph, which is associated with the feedforward network [40]. Given a *FNN* with depth D and let f^d be the function of the current layer d, the composed function $f(\boldsymbol{x};\boldsymbol{\theta})$ can recursively be given as:

$$f(\boldsymbol{x};\boldsymbol{\theta}) = f^d(f^{d-1}(\boldsymbol{x};\boldsymbol{\theta}^{d-1});\boldsymbol{\theta}^d)\ ,\ \text{with } 1 < d \leq D. \tag{4.17}$$

The parameters $\boldsymbol{\theta}$ are iteratively optimised, based on the provided training data, with the aim to approximate the optimal parameters $\boldsymbol{\theta}^\star$ of the function, representing the ideal solution $y = f(x;\boldsymbol{\theta}^\star)$. This process is called training and is briefly described next.

4.3.4 Optimisation

The optimisation of neural networks is achieved by minimising a measure of difference (the "error") between the true data y and the predictions $\hat{y}$, which is given by the *objective-* or *loss function* $\mathcal{L}$. The minimisation is achieved by iteratively adjusting the model parameters depending on their individual contribution to the error. Rumelhart et al. [84] proposed an efficient, iterative gradient-based algorithm called *error backpropagation* or simply *backprop* to optimise the parameters of a neural network. Computing the gradients of the objective function with respect to the parameters allows to adjust $\boldsymbol{\theta}$ such that the loss $\mathcal{L}$ reaches a local minimum. At this point the training procedure stops as the network is said to have converged.

The iterative optimisation process of neural networks can be structured into three steps. During the *forward-propagation step* the network input is passed through each layer of the model. Based on the input, the weights and the utilised activation function the artificial neurons are activated or deactivated. The difference between the network output $\hat{y}$ and the ground truth y is computed via the defined objective function $\mathcal{L}$. Next, the gradients of $\mathcal{L}$ with respect to the model parameters $\boldsymbol{\theta}$ are calculated in each iteration τ for each sample x_n and its corresponding target y_n in the training sets X and Y, respectively:

$$g_{\boldsymbol{\theta}^\tau_n} = \nabla_{\boldsymbol{\theta}^\tau}\mathcal{L}(f(x_n;\boldsymbol{\theta}^\tau_n), y_n)\ , \tag{4.18}$$

where f was defined in Eq. (4.17).
The gradient over all samples in X is then given as:

$$g_{\boldsymbol{\theta}^\tau} = \frac{1}{N}\sum_{n=1}^{N} g_{\boldsymbol{\theta}^\tau_n}\ . \tag{4.19}$$

In the last step, the gradient descent step, the parameter set is updated as following:

$$\boldsymbol{\theta}^{\tau+1} := \boldsymbol{\theta}^\tau - \eta \cdot g_{\boldsymbol{\theta}^\tau}\ , \tag{4.20}$$

where η is a non-negative weighting parameter, called *learning rate*, which controls the magnitude of the update and $\tau + 1$ indicates the next iteration.

This formulation is called *GD* optimisation and is computationally very demanding as each update requires the gradient computation for all samples in the dataset. Stochastic gradient descent (SGD) is a *GD* variation in which only a small random subset $X_k \subset X$, called *mini-batch*, is used to calculate the gradients for one update step. While *SGD* has shown improved convergence performance [1, 64], both methods rely on a proper value for the learning rate η, which defines how much the parameters θ are modified per iteration. However, finding a proper learning rate is not trivial. Therefore, adaptive learning rate optimisers were implemented which do not rely globally fixed learning rate.

Adaptive moment estimation (ADAM) ([58]) is a popular adaptive optimiser, which incorporates a momentum that accelerates learning into the update process. To estimate the moments, *ADAM* uses exponentially moving averages on the computed gradients of the current mini-batch. The first and second order moment $\boldsymbol{\mu}$ and $\boldsymbol{v}$ respectively, is defined as:

$$\boldsymbol{\mu} := \beta_1 \cdot \boldsymbol{\mu} + (1 - \beta_1) \cdot g_{\boldsymbol{\theta}} \tag{4.21}$$

$$\boldsymbol{v} := \beta_2 \cdot \boldsymbol{v} + (1 - \beta_2) \cdot {g_{\boldsymbol{\theta}}}^2 \ , \tag{4.22}$$

where $\beta_1, \beta_2 \in [0, 1)$ are exponential decay rates for the moment estimates. The bias-corrected estimate of $\boldsymbol{\mu}$ and $\boldsymbol{v}$ is given by:

$$\hat{\boldsymbol{\mu}} = \frac{\boldsymbol{\mu}}{1 - \beta_1} \tag{4.23}$$

$$\hat{\boldsymbol{v}} = \frac{\boldsymbol{v}}{1 - \beta_2} \ . \tag{4.24}$$

The resulting update rule for the parameters is then given as:

$$\boldsymbol{\theta}^{\tau+1} = \boldsymbol{\theta}^{\tau} - \frac{\eta}{\sqrt{\hat{\boldsymbol{v}}} + \epsilon} \cdot \hat{\boldsymbol{\mu}} \ , \tag{4.25}$$

where ϵ is used to prevent divisions by zero. [58] suggested the following default values $\eta = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-8}$.

4.3.4.1 Loss Functions

Common measures to asses the difference between the prediction $\hat{y}$ and the corresponding ground truth y are the $\mathcal{L}_1$ and $\mathcal{L}_2$ loss.

The mean absolute error (MAE) is represented by $\mathcal{L}_1$ and is calculated by the mean absolute difference between $\hat{y}$ and y for all samples in a given set of size N:

$$\mathcal{L}_1 = MAE = \frac{1}{N} \sum_{n=1}^{N} |\hat{y}_n - y_n| \ . \tag{4.26}$$

The mean squared error (MSE) is represented by $\mathcal{L}_1$ and is calculated by the mean absolute difference between $\hat{y}$ and y for all samples of given set N:

$$\mathcal{L}_2 = MSE = \frac{1}{N}\sum_{n=1}^{N}(\hat{y}_n - y_n)^2 \ . \tag{4.27}$$

The cross-entropy is a measure from information theory and measures the difference between two probability distributions. It is commonly used for classification and is for a given sample y and its prediction $\hat{y}$ calculated as:

$$H = -\frac{1}{N}\sum_{n=1}^{N} y_n \log(\hat{y}_n) \ . \tag{4.28}$$

4.3.5 Convolutional Neural Networks

In fully connected *FNNs*, the input is a single vector containing all values and every node of a layer is connected to each neuron of its successive layer. This results in a vast amount of parameters that need to be calculated and stored, which in turn would also require an infeasibly amount of data to reduce overfitting. While the achievable performance is better the data requirements and computational effort limit the usability of these networks.

Convolutional neural networks (CNNs) [111] are specifically designed networks that assume a specific type of input, namely with a spatial or temporal structure such as images. This allows to encode image related properties into the architecture and make the forward function more efficient, which tremendously reduces the parameter count. As the name implies, *CNNs* are networks that have at least one layer utilising a *convolution* operation instead of the general matrix multiplication [40].

CNNs are built-up by blocks of convolution and pooling operations. This block typically consists of three stages. First, the *convolution layer* or also a set of such layers. The resulting set of linear functions is then applied to non-linear activation functions. Together, these two stages are sometimes referred to as the *detector stage.* This stage is usually followed by the third stage, a *pooling layer.*

4.3.5.1 Convolution Operation

A *convolution* is a special kind of linear operation. The discrete convolution for a *3D* input image I can be defined as:

$$S(i,j,k) = (I * W)(i,j,k) = \sum_m \sum_n \sum_o I(m,n,o)W(i-m, j-n, k-o) \ , \tag{4.29}$$

where W... represents the *filter* or *kernel* and the output is often referred to as *feature map.*

The motivation behind the utilisation of convolutional layers is based on three ideas.

Sparse Interactions. In *CNNs* the learned filters are smaller than the input itself, which introduces *sparse interactions/connectivity.* This yields to connections between spatially neighbouring pixels, where the receptive field is defined via the size of the convolution kernel. Moreover, fewer parameters need to be used, reducing the memory consumption as well as improving the statistical efficiency [40].

Parameter Sharing. In fully connected (or dense) layers each weight is used exactly once when calculating the output. In comparison, convolutional layers learn only one set of parameters, which is, depending on the border handling, applied to every pixel in the input image. Therefore, as the value of the weight applied to one input unit is tied to the value of a weight applied elsewhere it can be seen as tied weights [40]. While this does not affect the forward propagation runtime, it further reduces the memory consumption of the model.

Equivariant Representations. Based on the particular form of parameter sharing, convolution layers have a property called *equivariance* to translation. Equivariance of a function means that the output changes in the same way as the input changes. For (*MR*) images, convolutions create *3D* maps where certain features appear in the input. If the (anatomical) structure causing these features is translated within the input image, its representation in the output will be translate in the same way.

4.3.5.2 Pooling Operation

Pooling layers transform its input into a summary statistic while decreasing the image size. Pooling also enforce invariance to small translations, further reduce the amount of parameters and increase the receptive field for the successive layers [40]. Commonly used pooling functions are *max-pooling* or *average-pooling.* Here, only either the maximum or average value from a certain neighbourhood (defined by the kernel size) of the input is preserved in the corresponding, smaller output image.

4.3.6 Generative Adversarial Networks

Generative adversarial networks (GANs) represent a neural network based generative modelling approach, which were first introduced by Goodfellow et al. [41]. Such networks are composed of two parametric models, called the *generator* G and the *discriminator* or *critic* D respectively. Both models are usually implemented as deep neural networks and trained alternately in a competitive manner. This can be interpreted in a way such that G generates new images which look as real as possible to fool the discriminator. In more detail, G estimates the distribution of the training data and is trained to maximise the probability of D being wrong. D on the other hand only evaluates whether the current input sample is more likely to be from the training data instead of being created by G. The discriminator network is therefore trained to maximise the probability for a correct labelling independent of the input source. This can be formulated as a minimax two-player game (see Eq. (4.30)) between the generator and the discriminator.

More formally, let $\mathbb{P}_r$ be the *real data distribution* and let $\mathbb{P}_g$ be the *distribution of the synthetic data*, generated via G based on the input noise z with distribution $\mathbb{P}_z$. A generated, synthetic sample $\hat{y}$ is then given by $\hat{y} = G(z)$, with $\hat{y} \sim \mathbb{P}_g$ and $z \sim \mathbb{P}_z$. G is defined as a differentiable function represented by a multi-layer network with corresponding parameters θ_g. D is similarly defined as a multi-layer network with parameters θ_d, $D(y;\theta_d)$, and a single scalar output. $D(y)$ represents the probability that y originates from the real data.

The objective function $V(D, G)$ of the mentioned minimax game is now given by:

$$\min_G \max_D V(D,G) = \mathbb{E}_{y\sim\mathbb{P}_r}[\log D(y)] + \mathbb{E}_{z\sim\mathbb{P}_z}[\log(1 - D(G(z)))] \ . \tag{4.30}$$

The optimisation of this cost function is realised by alternately updating the parameters of the generator θ_g and the discriminator θ_d respectively. In practice, an unbalanced update scheme is suggested where the discriminator is updated n_{critic} times for each generator update, which yields D to be near the optimal solution D^* given that G changes slowly. Assuming an optimal generator ($\mathbb{P}_g = \mathbb{P}_r$), the discriminator is unable to distinguish between both distributions which yields $D(y) = 1/2$ [41]. Formally, the optimal discriminator D for a given generator G is given as:

$$D_G^*(y) = \frac{\mathbb{P}_r(y)}{\mathbb{P}_r(y) + \mathbb{P}_g(y)} \tag{4.31}$$

Goodfellow et al. [41] showed that given such an optimal discriminator, training of G is equivalent to minimising the *Jensen-Shannon divergence* between $\mathbb{P}_g$ and $\mathbb{P}_r$.

The standard implementation of *GANs* ([41]) turned out to be hard to train because of instability problems. The problem of *mode collapse* can be caused if the updates of D and G are not well synchronised and G overpowers D. In such a case, G collapses too many values from different z to the same data value and loses the ability to model $\mathbb{P}_r$. Another issue are vanishing gradients related to the Jensen-Shannon divergence, which produces more reliable but also smaller gradients as D improves [5].

4.3.6.1 Wasserstein GAN

To address stability related issues of the standard *GAN* implementation, Arjovsky et al. [5] evaluated a variety of measures to assess the distance or divergence between the model- and data distribution $\rho(\mathbb{P}_g, \mathbb{P}_r)$ as well as their convergence influence. Based on their findings, they have proposed to use the *Wasserstein distance* for training *GANs*. The Wasserstein-1 or Earth Mover (EM) distance measures how much energy is needed (“mass” needs to be transported) to transform one distribution (e.g. $\mathbb{P}_r$) into another distribution (e.g. $\mathbb{P}_g$). The *EM* distance correlates to the “cost” of the ideal transport plan. Formally, this measure is defined as:

$$W(\mathbb{P}_r, \mathbb{P}_g) = \inf_{\gamma \in \Pi(\mathbb{P}_r, \mathbb{P}_g)} \mathbb{E}_{(y,\hat{y}) \sim \gamma}[||y - \hat{y}||] \ , \tag{4.32}$$

where $\Pi(\mathbb{P}_r, \mathbb{P}_g)$ is the set of all joint distributions $\gamma(y, \hat{y})$ whose marginals are $\mathbb{P}_r$ and $\mathbb{P}_g$, respectively.

As it is highly intractable to determine all joint distributions to find the infimum, the Kantorovich-Rubinstein duality can be used to reformulate Eq. (4.32) as follows:

$$W(\mathbb{P}_r, \mathbb{P}_g) = \sup_{||f||_L \leq 1} \mathbb{E}_{y \sim \mathbb{P}_r}[f(y)] - \mathbb{E}_{\hat{y} \sim \mathbb{P}_g}[f(\hat{y})] \ , \tag{4.33}$$

where the supremum is taken over all 1-Lipschitz functions with a mapping of $f : X \longrightarrow \mathbb{R}$.

Moreover, the 1-Lipschitz constraint can be replaced by a K-Lipschitz assumption yielding an approximation of the Wasserstein distance up to a constant $(K \cdot W(\mathbb{P}_r, \mathbb{P}_g))$. Therefore, if a parametrised set of functions $\{f_w\}_{w \in \mathcal{W}}$ which are all K-Lipschitz for a constant K is used, Eq. (4.33) can be reformulated [5]. If the parametrised family of functions is replaced by our discriminator model D with its parameters θ_d the following problem can be considered:

$$\max_D \mathbb{E}_{y \sim \mathbb{P}_r}[D(y)] - \mathbb{E}_{\hat{y} \sim \mathbb{P}_g}[D(\hat{y})] \ . \tag{4.34}$$

To return to the framework of *GANs*, where $\hat{y}$ is generated by the generator G based on the input noise z sampled from the distribution $\mathbb{P}_z$, Eq. (4.34) can be formulated as:

$$\min_G \max_D \mathbb{E}_{y \sim \mathbb{P}_r}[D(y)] - \mathbb{E}_{z \sim \mathbb{P}_z}[D(G(z))] \ , \tag{4.35}$$

which yields the objective function of the Wasserstein GAN (WGAN).

As the parameters θ_d are assumed to be within the compact space $\mathcal{W}$, backpropagation estimation yields:

$$\mathbb{E}_{z \sim \mathbb{P}_z}[\Delta D(G(z))] \ , \tag{4.36}$$

which results in a training scheme based on the standard algorithm proposed for *GANs* ([41]). To ensure the required parameter constraint such that all functions D will be K-Lipschitz, [5] suggested *weight clipping* of the discriminator weights to a fixed box of $\mathcal{W} = [-0.01, 0.01]$ after each update step.

While for training the standard implementation of *GANs* ([41]) the balance between updating D and G was delicate, the problem of mode collapse was not observed in *WGAN* [5]. This is because in *WGANs* it is now possible to train the D till optimality, where it still provides a loss for the generator which can be trained. However, Arjovsky et al. [5] also report that training D in *WGANs* becomes unstable with momentum based optimisers such as Adam (58). This, however, can be circumvented if RMSProp ([94]) is used instead.

4.3.6.2 Wasserstein GAN with Gradient Penalty

While Arjovsky et al. [5] stated that using weight clipping in order to ensure the K-Lipschitz requirement is prone to failure as it can easily lead to *vanishing gradients*, they also encouraged others for further investigation.

Gulrajani et al. [43] demonstrated problems related to weight clipping, namely *vanishing-* or *exploding gradients*, because of interactions between the weight clipping and the cost function. Moreover, they observed that the capacity of D is reduced because of weight clipping, which limits its ability to learn complex functions. Therefore, they proposed an improved objective function, which penalises the gradients of the discriminator D to ensure the Lipschitz constraint and is therefore called Wasserstein GAN with gradient penalty (WGAN-GP).

Their modified cost function is defined as follows:

$$W(\mathbb{P}_r, \mathbb{P}_g) = \underbrace{\mathbb{E}_{y\sim\mathbb{P}_r}[D(y)] - \mathbb{E}_{z\sim\mathbb{P}_z}[D(G(z))]}_{\text{WGAN loss}} + \underbrace{\lambda \cdot \mathbb{E}_{\hat{y}\sim\mathbb{P}_{\hat{y}}}\left[(\|\nabla_{\hat{y}} D(\hat{y})\|_2 - 1)^2\right]}_{\text{gradient penalty}} \ , \tag{4.37}$$

where λ describes the *penalty coefficient* and $\mathbb{P}_{\hat{y}}$ is implicitly defined by sampling uniformly along straight lines between point pairs sampled from $\mathbb{P}_r$ and $\mathbb{P}_g$. Gulrajani et al. [43] found in their work that a penalty coefficient of $\lambda = 10$ works well in a variety of architectures and datasets.

Looking at the term of the gradient penalty (Eq. (4.37)) it can be seen that the penalty encourages the norm of the discriminator gradient to approach one, which was empirically found to produce slightly better results [43].

4.3.6.3 Conditioned GAN

The improved training procedure of *WGAN-GP* yields a powerful generative model. However, common to all three versions of the explained *GANs* samples are generated based on random noise z.

Mirza and Osindero [68] adapted the principle of unconditioned *GANs* by conditionig the model on additional information to guide the data generation process. Therefore, the random input noise $z \sim \mathbb{P}_z$ can be replaced with a prior x with distribution $\mathbb{P}_x$, which can be any meaningful information such as data from different modalities or class labels.

Isola et al. [53] evaluated the capability of conditional GAN (cGAN) in the application of image-to-image translation tasks. Depending on the implementation, either both, the generator G and the discriminator D, or only G can be subjected to the input prior x sampled from a distribution $\mathbb{P}_x$. Limiting the prior $x \sim \mathbb{P}_x$ only to the generator, the formulation of the objective in *WGAN-GPs* (Eq. (4.37)) can be modified to:

$$\mathcal{L}(G, D) = \mathbb{E}_{y\sim\mathbb{P}_r}[D(y)] - \mathbb{E}_{x\sim\mathbb{P}_x}[D(G(x))] + \lambda \cdot \mathbb{E}_{\hat{y}\sim\mathbb{P}_{\hat{y}}}\left[(\|\nabla_{\hat{y}} D(\hat{y})\|_2 - 1)^2\right] \ . \tag{4.38}$$

4.3.6.4 Incorporation of Additional Loss

Other work ([76]) reported benefits if the adversarial loss is combined with a pixel-wise loss such as $\mathcal{L}_2$. Isola et al. [53] also utilised this approach by incorporating $\mathcal{L}_1$distances into the objective function. As a consequence, now the task of the generator is not only to fool the discriminator but also to generate samples to be near the ground truth in a sense of $\mathcal{L}_1$. The job of the discriminator remains unchanged. As $\mathcal{L}_1$distances were preferred to encourage less blurring, combining the $\mathcal{L}_1$loss into the generator yields the final objective utilised in this thesis:

$$G^* = \min_G \max_D \mathcal{L}(G, D) + \gamma \cdot \mathcal{L}_1 \ , \tag{4.39}$$

with $\mathcal{L}_{L1} = \mathbb{E}_{x,y}\left[\|y - G(x, y)\|_1\right]$ and a weighting parameter γ for the pixel-wise loss.

5 Methods

Contents

The content of this chapter aims to explain the whole workflow of this thesis. It starts with the creation of the dataset, which is comprised of three steps. The image acquisition, followed by the manual delineation of the hippocampus as well as preprocessing to cope with possible motion artefacts. This generated ground truth set is subsequently processed with the *FreeSurfer* pipeline to acquire the reference segmentation currently used in clinical research. Next, the data augmentation / preparation and utilised network architectures are described. Finally, the evaluation process along with the used metrics is explained.

5.1 Dataset

This section gives an insight into the generation of the dataset. At first the magnetic resonance imaging (MRI) hardware and sequences utilised for the acquisition of magnetic resonance (MR) images are described. Following, the actual acquisition process is explained and both of our chosen image sequences are motivated. A description of the acquired data is given and a qualitative example of a recorded T1- and T2-weighted image pair and their relation to each other is shown. Moreover, differences between both image modalities, unique challenges and limitations of the imaged data are reported.

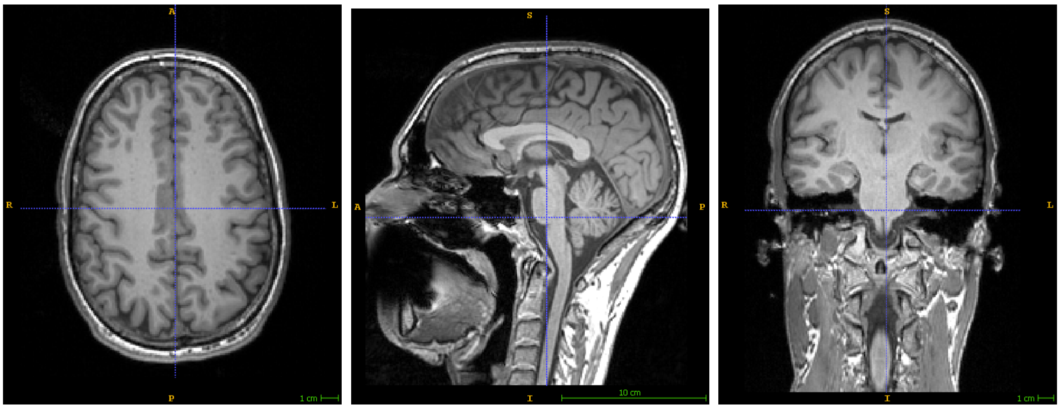
The next part provides insights into the manual annotation process. The used protocol to label the hippocampus (HC) as well as its specific application to our acquired data is explained in detail. An outline of the used software and faced problems is given at the end of this part.

5.1.1 MRI Acquisition

5.1.1.1 Scanning Setup

Hardware. The acquisition of the *MR* data has been done at the Medical University of Graz (MUG). A Siemens Prisma 3T *MR* scanner was used throughout the whole data acquisition process. For our measurements, a 20 channel head coil was used for signal reception.

Scanning Procedure. Before the actual *MRI* examination, participants had to complete a short questionnaire about medical conditions that would contradict the exposure to high magnetic fields. Exclusion criteria are any form of metal implants, claustrophobia or electrical devices such as pacemakers. Admitted subjects were placed into the head coil in a "head-first-supine" manner. Moreover, headphones and foam cushioning were used to lower noise and provide additional head support, respectively reduce the space for possible but unwanted movement of the head. At first, so called "localiser" measurements were performed, which are usually 3 – 5 low resolution slices in three planes to cover the head. This is done to get an overview of the imaged anatomy and to plan the subsequent acquisition of the study sequences, such that the desired region of interest is covered. To achieve approximately the same orientation of all brain scans, which is essential for intra- and inter-subject comparison in an *MRI* study, anatomical landmarks are used to align the three standard planes with the current subject. The sagittal plane is hereby crucial and oriented along to the longitudinal fissure (vertical cursor in Fig. 5.1a), which separates the brain into both hemispheres (Fig. 5.1).



(a) axial view **(b)** sagittal view **(c)** sagittal view

Figure 5.1: Visualisation of the used landmarks to plan the MR examination.

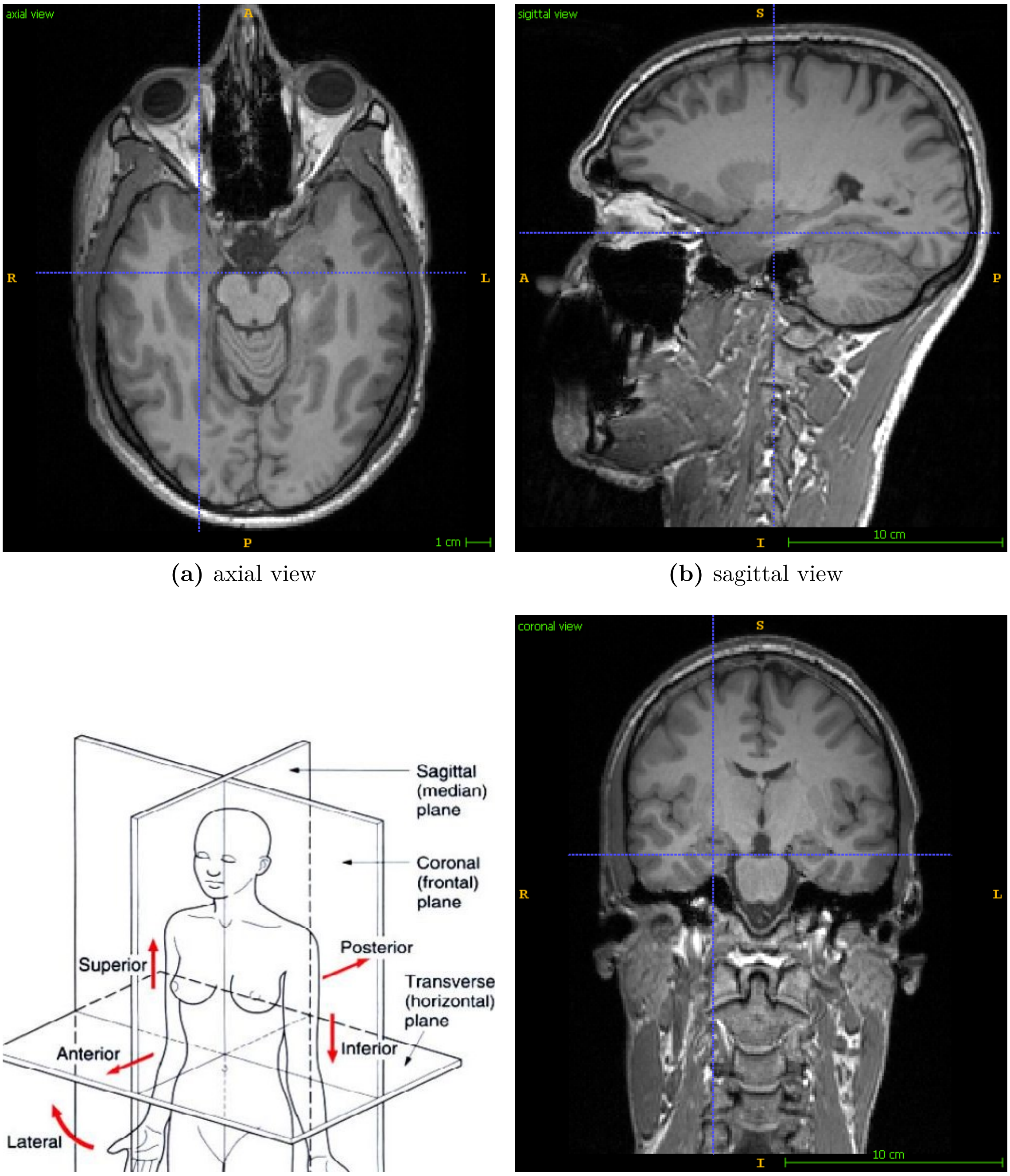
After the initial adjustment, two types of images were acquired for our study: (i) a structural T1-weighted scan of the whole brain, and (ii) a detailed T2-weighted scan of the

HC.

5.1.1.2 Sequences

T1-weighted Scan. The first scan produces a high-resolution T1-weighted image of the whole head, which is a standard scan in clinical research. It was acquired with a three-dimensional (3D) magnetisation-prepared rapid gradient-echo (MPRAGE) sequence, which is a spoiled gradient echo (GE) sequence for ***R**apid **A**cquired **G**radient **E**cho* (RAGE) images. *Spoiling* refers to a method used for fast *GE* sequences and is the disruption of transverse coherence (transverse magnetisation) at the end of each repetition cycle. *3D* acquisition means, that the whole volume is excited at once instead of slice by slice. *MPRAGE* sequences can produce very high-resolution T1-weighted images, providing very good anatomical detail of the brain. In our case these scans have an isotropic resolution of $1 \times 1 \times 1\,\mathrm{mm}^3$ with a field of view (FoV) of $224 \times 256 \times 176$ and take 6.5 minutes to complete. The purpose of this T1 scan is to get a fast, structural image of the whole brain that can be further used for a variety of analyses. For example, the *FreeSurfer* pipeline to perform regional analysis, volumetry or similar. In our case the T1 scans were also processed with *FreeSurfer* v6.0 to get the baseline segmentations of the measured hippocampi. We also use this scan to localise the *HC* for the subsequent T2-weighted scan. After the whole T1 scan is acquired, both hippocampi were located on the axial (Fig. 5.2a) and sagittal (Fig. 5.2b) views and the graphical windows of the scanner software, indicating the coverage / region of interest (ROI) of the T2 sequence, were adjusted to (ideally) start with the first slice of each *HC*. Moreover, the planes of the principle directions were rotated to match the orientation of the *HC*.

T2-weighted Scan. The second scan provides an ultra-high-resolution T2-weighted image slab, which covers only a small region of the brain, including the *HC*. While only covering a small portion of 4 cm along the sagittal plane (slice direction) (Fig. 5.3c), the whole oblique coronal plane (cross section) (Fig. 5.3b) is acquired. The reason for this cropped *FoV* ($352 \times 512 \times 40$) is that a full brain T2-weighted scan, at such a high resolution is not feasible. However, for the purpose of hippocampal segmentation it is sufficient to visualise only the *HC* and its surrounding with such detail and contrast. The T2-weighted scan has a very high in-plane resolution of $0.47 \times 0.47\,\mathrm{mm}^2$ and also a slice thickness of 1 mm. This anisotropic resolution results from the trade-off to achieve the very high in-plane resolution, while covering as much of the hippocampal formation as possible. Moreover, this is motivated by the inner structure of the *HC*, which resembles a Swiss roll. Major structural variations occur within its cross-section, whereas changes along its main axis happen less rapidly. This special structure is also the reason for an *oblique coronal* orientation of the T2 scan, with the slice direction approximately aligned with the long axes of both hippocampi. The latter implies, that oblique coronal plane across the *HC* is perpendicular to these major axes. Here, oblique stands for a slanting or inclined orientation compared to the principal direction (see Fig. 5.8 or Fig. 5.9).



(a) axial view

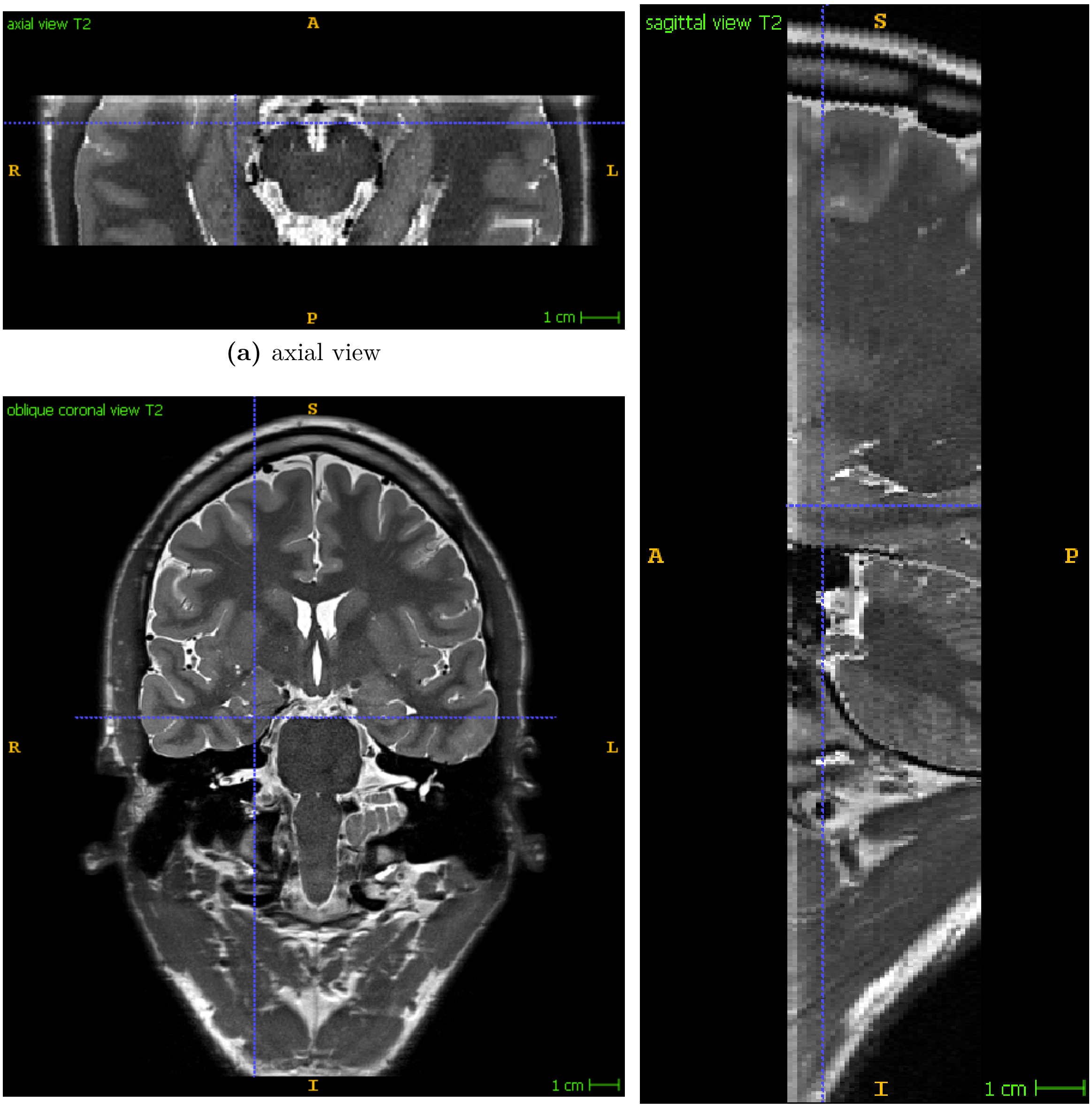
(b) sagittal view

(c) standard anatomical planes [2]

(d) coronal view

Figure 5.2: Example visualisation of a T1-weighted *MR* scan for all views. [2] adapted from http://completesoccertraining.blogspot.com/2012/09/terminology-understandance.html (accessed 25.05.2020)

For the acquisition of this scan a two-dimensional (2D) turbo spin echo (TSE) sequence was utilised which takes around 8.5 minutes. An “interleaved” slice order was used, which means that at first the odd slices and then the missing even slices are acquired and correspondingly merged during the image reconstruction. The slice thickness of 1 mm together with a distance factor of 0 % yields a resulting slice distance of 1 mm. The spin echo (SE) is needed to achieve a T2-contrast of the image. Moreover, the *2D*-based acquisition allows, for the limited *FoV*, only capturing the region around the hippocampal formation.



(a) axial view

(b) oblique coronal view

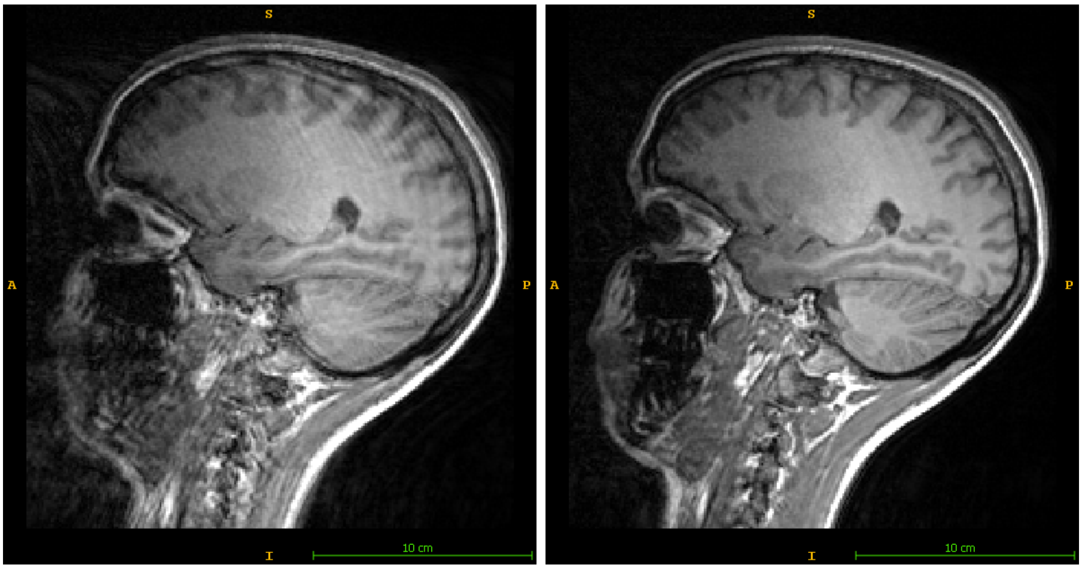
(c) sagittal view

Figure 5.3: Example visualisation of a T2-weighted *MR* scan for all views.

5.1.1.3 Challenges and Artefacts

During the described acquisition process several challenges, partly unique to this dataset, can occur. The first step, the placement of the patient and especially the additional cushioning, is already crucial. If the head is forced into an unnatural or overstretched position, it is almost impossible to keep still during the roughly 20 minutes of total scan time. A steady position of the head throughout the whole scanning procedure is crucial, especially during each of both sequences. Motion of the head between the acquisition of the T1 and T2 scan results in a misalignment of these scans with respect to each other. Any form of movement during a measurement itself yields ghosting effects or blurring and a worse image quality of the resulting *MR* scan. Without motion, the background should be black and the brain visualised in a clear manner with high signal-to-noise ratio (SNR) (cmp. Fig. 5.2).

In contrast, Fig. 5.4 exemplarily shows the mentioned motion artefacts based on an initial (Fig. 5.4a) and repeated T1-weighted scan of the same subject. Both scans are contaminated with motion induced artefacts, whereas the second scan is slightly better implying less movement during the scan.



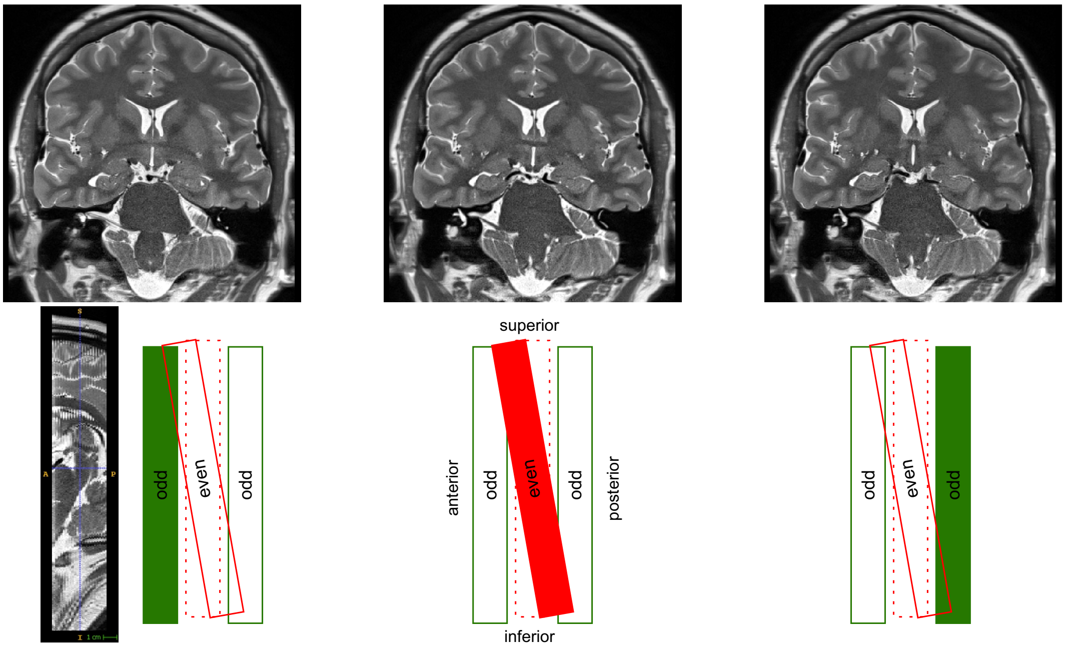
(a) severe motion **(b)** considerable motion

Figure 5.4: Visualisation of motion induced artefacts like ghosting, blurring and bad *SNR*. (a) shows an initial T1 scan with distinctive artefacts due to severe motion. (b) shows the same slice, with the same contrast settings of a repeated scan, with less but still considerable motion artefacts.

Motion during the *3D MPRAGE* acquisition results in a worse image quality (*SNR*) of the whole scan, compared to the achievable quality without any motion artefacts. For the *2D TSE* sequence, used for the detailed visualisation of the *HC*, motion additionally results in alternating jumping between odd- and even slices. This jiggling is due to the interleaved acquisition of the *TSE* sequence, which is used to avoid cross-talk of neighbouring slices.

Thus, motion during such a "sequential" acquisition of odd and even blocks, causes an intra-volume misalignment of the T2-slices as well as lower image quality. An essential factor for the susceptibility to motion artefacts is the scan time because the scanned subjects have to stay still for an extended time, which gets harder the longer it takes. Moreover, our ultra-high in-plane resolution is prone to show even the slightest kind of motion. Already displacements by only 0.5 mm, which occur even due to physiological motion (like breathing, blood and cerebrospinal fluid (CSF) pulsation, ...), result in a shift by one pixel!

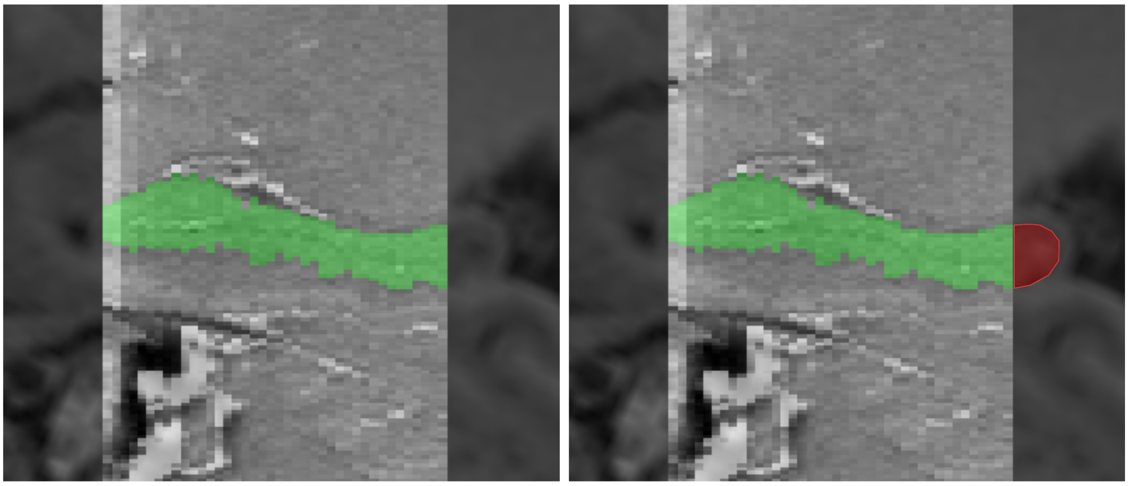
Figure 5.5 visualises motion induced misalignment between the odd and even block of an acquired T2 slab. The shown case exhibits severe motion, most likely caused by nodding of the head. As indicated with the schematic visualisation of the slices (bottom row), is the even block (shown in red) not parallel to the odd block anymore. The misorientation results in an "overlap" at the superior part of the centre slice (Fig. 5.5b) and its previous (Fig. 5.5a) slice. This can also be seen when the upper half of the coronal slice in Fig. 5.5a and Fig. 5.5b is compared. If the even slice (Fig. 5.5b) is compared to its subsequent odd slice (Fig. 5.5c) it becomes visible that here the inferior part of the slices is similar. This can now be seen when comparing the lower half of the coronal slice in Fig. 5.5b and Fig. 5.5c.



(a) previous odd slice **(b)** even center slice **(c)** subsequent odd slice

Figure 5.5: Misalignment between odd and even block of acquired T2 slab. The upper row shows three successive oblique coronal slices of an example T2 MR scan. The bottom row, schematically displays the relative orientation of these slices to each other in a sagittal view.
Dashed even rectangles (- - -) indicate the parallel, desired, configuration of both blocks. Filled rectangles highlight the current slice, which is shown above the schematic. Green refers to slices of the odd image block, whereas the even (centre) slice in between two odd slices is shown in red.

Another error source connected to the acquisition process itself, is the adjustment of the T2-weighted image slab to match the beginning (first slice) of both hippocampi. As this is done based on the T1 scan it is very hard and often not possible to exactly start the T2 scan with the first slice of each *HC*. However, this only results in a not complete coverage of the *HC* head but does not affect the image quality itself. Moreover, a coverage of the whole *HC*, including the complete hippocampal tail (HT), was only possible in a few cases due to the limited amount of 40 slices. This is shown with one example in Fig. 5.6 below. In this figure, the green mask represents the manual ground truth labelling of the left *HC* while the red annotation approximately shows the missing part of the *HT*, which is only covered with the whole head *3D* T1 scan.



(a) unlabelled missing HT **(b)** annotated missing HT

Figure 5.6: Visualisation of incomplete coverage of the *HT*. Bright slices show the T2-weighted scan, which is overlaid onto the aligned T1 scan. The green mask is the manual ground truth of the left HC. The left image (a) shows that, due to the missing slices of the *HT*, the complete tail cannot be manually labelled. The missing part is additionally roughly outlined in red (b).

Figure 5.6 also visualises another effect of the limited *FoV* of our *TSE* sequence. The first odd and even slice (slice 1 & 2) are much brighter, so tissues show higher intensities, compared to the other slices. This is due to effects of magnetisation transfer (MT) and inference from excitations of neighbouring slices. The width of the excited slice depends among other things on the bandwidth and the slice profile of the applied pulses. Ideally, pulses would have a rectangular (top hat) slice profile such that only spins of the current slice are affected by applied radio frequency (RF) excitations. To achieve a rectangular shape in the frequency domain, the amplitude of the applied pulse in the time domain needs to resemble a sinc function ($sinc = \frac{sin(x)}{x}$). However, such a sinc function would have an infinite duration in the time domain. Therefore, truncation of the pulses is needed which in turn causes non-rectangular slice profiles. The resulting partial excitation of neighbouring slices is called *cross talk*, and reduces the *SNR* due to saturation effects (intensities get darker). Cross talk is especially problematic in *SE* sequences because slice profiles for 180° *RF*- pulses are worse. Therefore, the very first slices of the odd and even acquisition block of the T2-weighted images show much brighter intensities, with a negative gradient of the effect along the first slice.

Additionally to the artefacts caused by the acquisition process, inherent errors of the *MRI* methodology respectively sequence (*GE* or *SE*) related artefacts are present. This includes *flow artefacts* caused by the blood and *CSF*, *partial volume effects* or *susceptibility artefacts*. The reason for partial volume artefacts is a mixture of different tissues within the same voxel. Therefore, if the voxel size and especially its slice thickness is too big for the measured organ, the resulting image slice will show an average of different signal intensities. This causes an inaccurate visualisation of the underlying tissue properties.

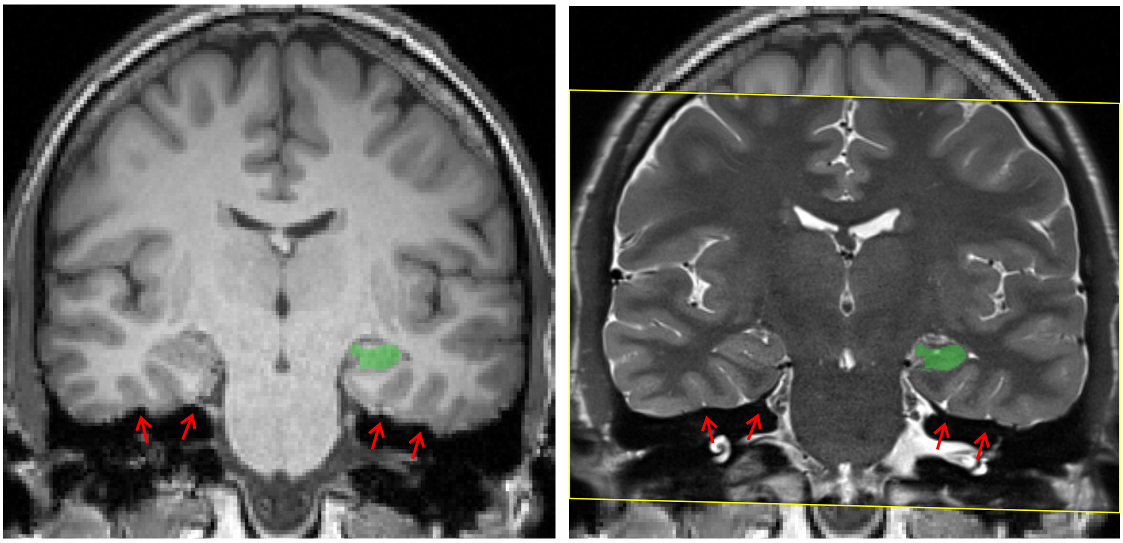
Susceptibility artefacts also occur in *GE* and *TSE* sequences, however, they are much more pronounced in *GE* acquisitions. *SE* sequences are almost insensitive to static field inhomogeneities due to the additional *RF* pulse. *Magnetic susceptibility* defines to which extend any material becomes temporarily magnetised when subjected to a strong magnetic field. Of all tissues, bone shows the lowest susceptibility values whereas iron-containing molecules, like blood, have the highest. Other tissues mainly express susceptibilities in a mid-range. Air is to a large part composed of oxygen, which is paramagnetic, and provides no *MR* signal due to the missing protons. However, air locally influences the magnetic properties of its vicinity, yielding to susceptibility artefacts. Different susceptibility values at tissue boundaries (e.g. bone-tissue or bone-air) cause slight changes in the local magnetic field. These changes, create micro-gradients across affected voxels which speed up the spin dephasing, similar to the T2* effect. The phase change caused by susceptibility, at a specific location, is given by:

$$\Delta\varphi(\vec{x}) = \gamma \cdot G_i(\vec{x}) \cdot \Delta r \cdot TE \ , \tag{5.1}$$

where G_i is the internal magnetic field gradient and Δr denotes the voxel size. Equation (5.1) shows that a higher resolution reduces susceptibility artefacts, which also applies to our T2-weighted scans. Figure 5.7 shows an example of susceptibility artefacts at the tissue boundaries to the cranial base.

Another prominent artefact in *MRI* is caused by fat and is based on the *chemical shift*. The term “chemical shift” refers to the dependence of the resonance frequency of a proton (spin) on its molecular surrounding. Compared to water molecules (H_2O), fat and their triglyceride chains have much more hydrogen atoms (H) which are each surrounded by many other atoms. The neighbouring electron clouds of these other atoms locally reduce the external magnetic field (B_0), which yields a reduced Larmor frequency of spins in fat molecules compared to those of water. This difference is know as the chemical shift and is 3.5 parts per million (ppm) between fat and water [69]. As this shift depends on the strength of B_0, the resulting frequency difference also depends on the field strength of the used scanner. At 3T scanners it is about 440 Hz.

This chemical shift induces now two problems. For one, *chemical shift artefacts* which are mapping errors (pixel shifts) of fat containing pixels due to the incorrectly lower frequency. This artefact is present in *GE* and *SE* sequences and occurs only along the frequency-encoding direction. The overall pixel shift, additionally depends inversely on the receiver bandwidth of the scanning sequence – a lower bandwidth causes a greater pixel shift. However, increasing the bandwidth would decrease the *SNR*. The second fat - water shift related artefact is called *phase cancellation artefact*, among other names. It is limited to only *GE* sequences, but can easily avoided as it occurs only at specific echo times.



(a) original T1 coronal slice

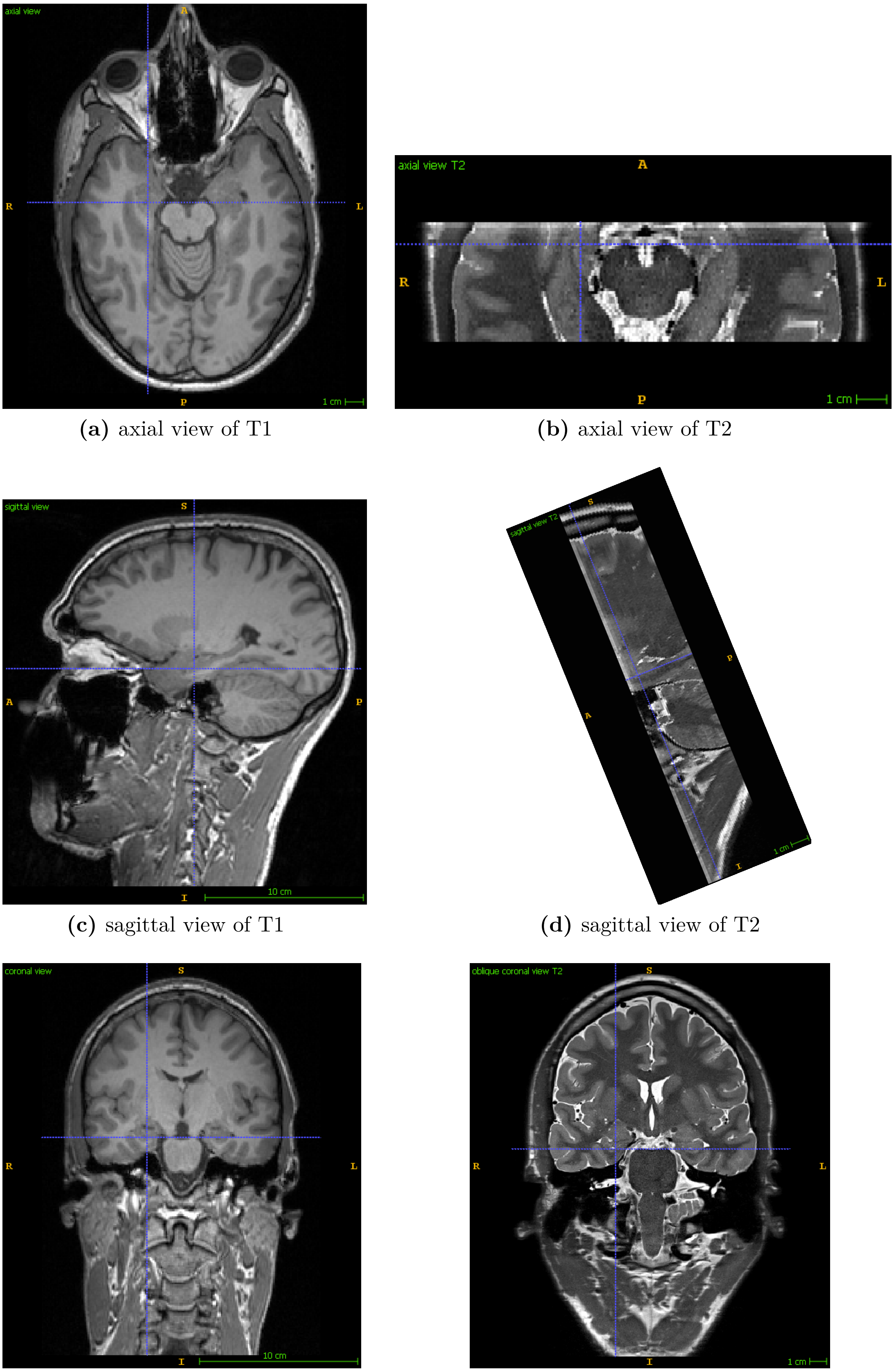
(b) T1 with T2 overlay

Figure 5.7: Visualisation of susceptibility artefacts; (a) shows a coronal slice of the original T1, while (b) additionally has the resliced T2 overlaid (border indicated as yellow frame). Red arrows indicate susceptibility artefacts at the cortex boundary.

5.1.2 MRI Data

Based on the previously described acquisition procedure (Section 5.1.1) a total of 33 healthy volunteers have been scanned, resulting in 33 corresponding pairs of T1- and T2-weighted *MR* images. An example of these images is shown in Fig. 5.2 and Fig. 5.3 for the T1 and T2 scans, respectively. Below, a side by side comparison of an original T1- (Fig. 5.8 left column) and T2-weighted (Fig. 5.8 right column) image pair is shown.

The spatial relation of the T1- and T2 scans, as explained in Section 5.1.1.2, is shown in Fig. 5.9. The T2-weighted image slab, overlaid onto the T1-weighted image, is shown with bright intensity to visualise the special *oblique coronal* matching the long axis of the hippocampus (shown as green mask).



(a) axial view of T1

(b) axial view of T2

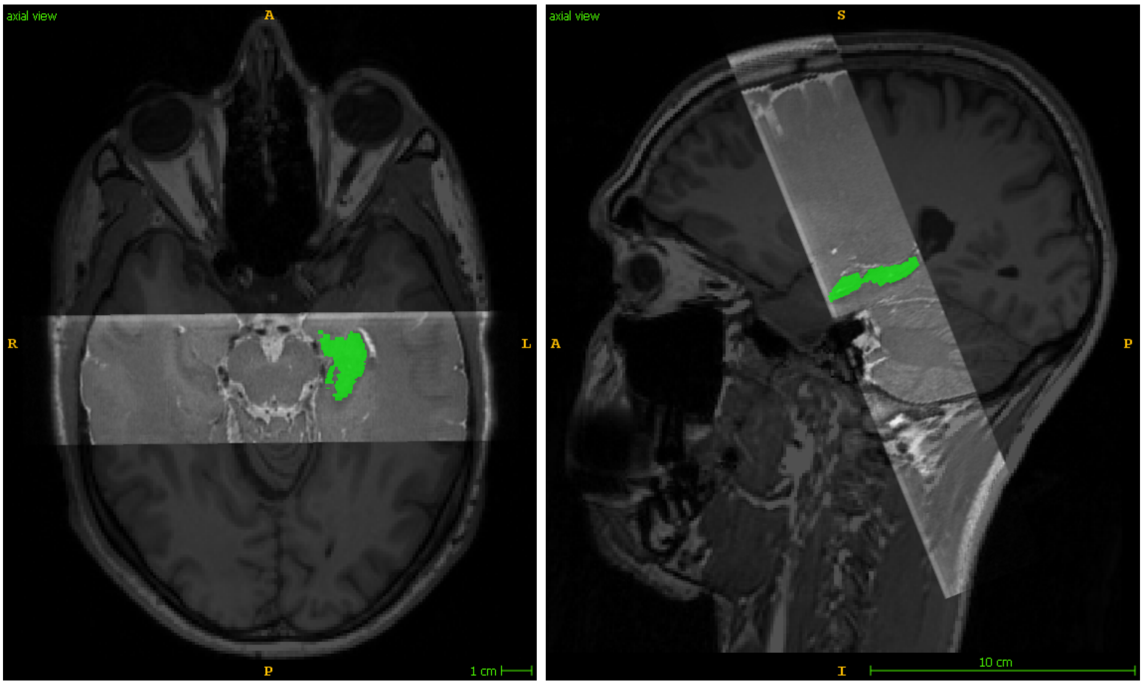
(c) sagittal view of T1

(d) sagittal view of T2

(e) coronal view of T1

(f) coronal view of T2

Figure 5.8: Comparison of an example T1- (left) and T2-weighted (right) MR scan. Images show roughly the same slice, as an comparison of the exact same slice would required interpolation.

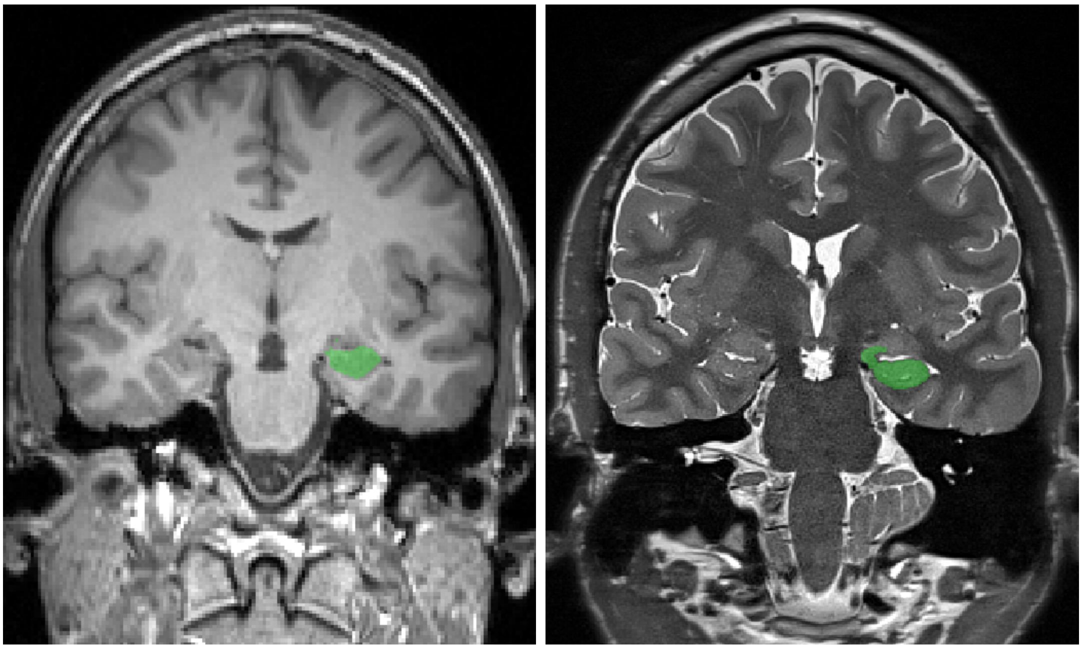


(a) Axial view

(b) Sagittal view

Figure 5.9: Spatial relation of the acquired T1- and T2-weighted MR scans. The manual ground truth annotation of the left HC is shown in green.

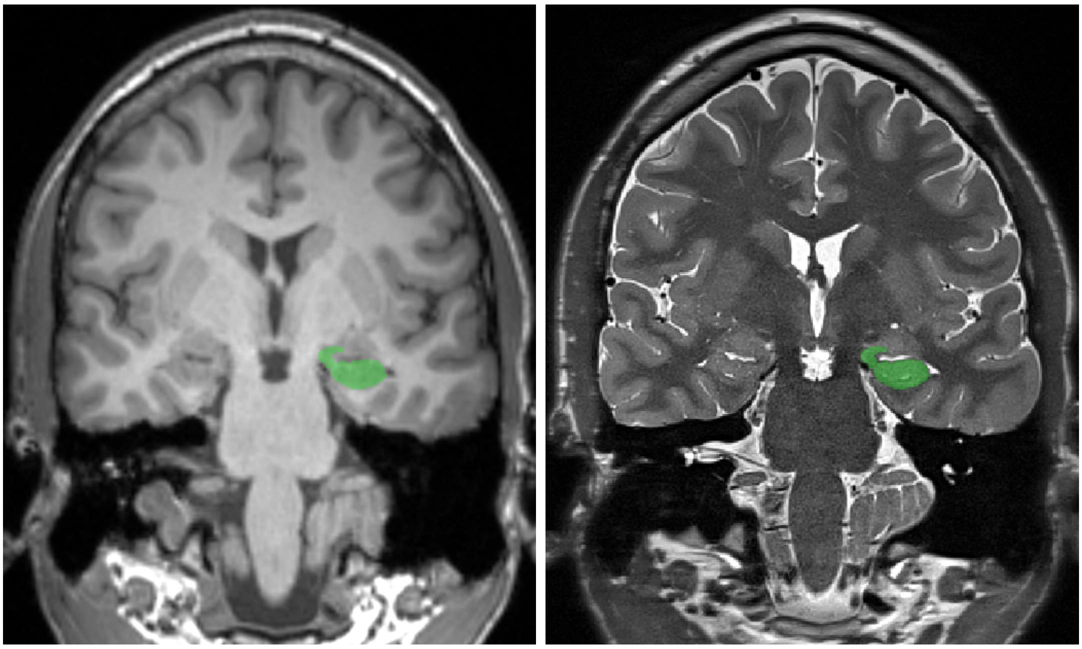
To show and describe the differences between the T1- and T2-weighted scans a cross-sectional slice at roughly the same position is shown below in Fig. 5.10 and Fig. 5.11. The former shows the cross-sections of the original scans in their corresponding directions of the acquisition. This means that the T1-weighted scan is shown in the standard coronal plane, whereas for the T2-weighted scan its oblique coronal slice is depicted. Due to this divergent orientation of both cross-sections also slightly different anatomical structures can be seen. Figure 5.11 shows approximately the same cross-section through the brain but for this the T1 scan had to be resliced and thus interpolated. The manual ground truth mask of the left *HC* is just shown as a reference where the *HC* is located and how it looks like.



(a) coronal T1 slice

(b) oblique coronal T2 slice

Figure 5.10: Cross-section through both *original* scans at roughly the same position. Anatomical variations are due to the different orientation of the coronal (T1) and oblique coronal (T2) planes.



(a) interpolated coronal T1 slice

(b) oblique coronal T2 slice

Figure 5.11: Cross-section through both scans at roughly the same position. T1 scan is rotated and interpolated to align with the T2 oblique coronal direction.

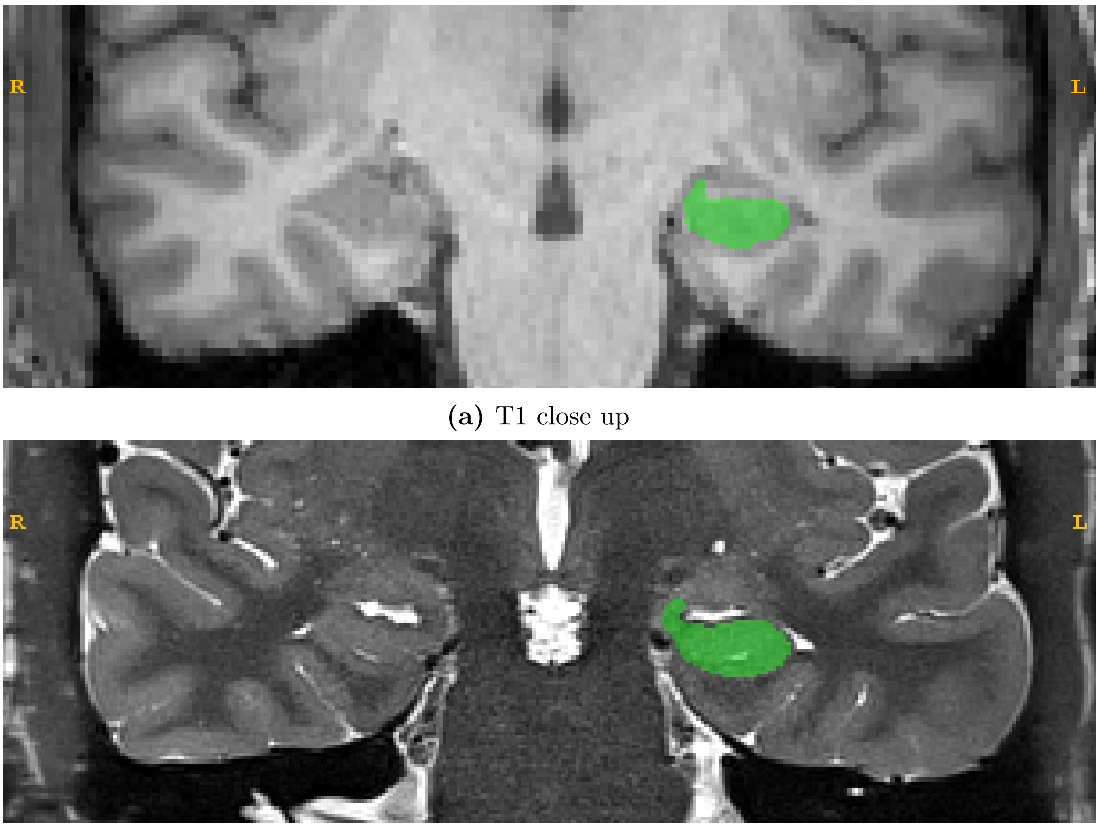
5.1.3 Manual Labelling

Crucial for an accurate, repetitive and comparable segmentation of the *HC* and the assessment of its volume are precise but comprehensible instructions on how to manually label the corresponding structures. Berron et al. [8] have developed such a protocol for manual segmentation of medial temporal lobe (MTL) subregions in 7T *MRI*, including novel anatomical findings about the hippocampal subfields [30, 31]. The authors provide detailed instructions together with slice-by-slice segmentation examples for understandable and alleviated learning of the annotation procedure, which have also been tested for usability in a segmentation workshop. Compared to our T2 scans, their protocol is based on similar T2-weighted *TSE* images with slightly higher in-plane resolution ($0.44 \times 0.44\,\text{mm}^2$) but also bigger slice distances of 1.1 mm. Furthermore, they acquired 55 slices, which ensures a complete coverage of the *HC*. Their images also show somewhat improved *SNR*, because of the higher field strength of 7T.

Because of the good correspondence with our acquired T2-weighted images, the manual ground truth annotation of our data is based on this protocol [8]. Generation of the manual ground truth is limited to our *2D* T2-weighted scans, as in [8], because of the described advantages if utilising this image modality.

5.1.3.1 Elaboration of T1 vs. T2 Contrast Differences

Figure 5.12 shows a close up view of the area around the *HC* on both, the native T1- ($1\,\text{mm}^3$) and T2-weighted ($0.47 \times 0.47 \times 1\,\text{mm}^3$) scans. Despite the mentioned artefacts (Section 5.1.1.3) both modalities show distinct differences in the image contrast. This is due to the different acquisition sequences (*GE* vs *SE*), which rely on dissimilar methods to create the measured echoes (cmp. Section 4.1). To be more precise, T1-weighted (Fig. 5.12a) images show primarily a complementary contrast compared to its T2 (Fig. 5.12b) counterpart. In general both images show dark pixels for air, bone and fast-flowing blood. Bright intensities are expressed, among others, by fat and fatty bone marrow, whereas such tissues are brighter in T2-weighted scans. A major drawback for the task of *HC* segmentation in T1 images is, that *CSF* cannot really be distinguished from the background; it shows very low signal and is basically noise. Therefore, it is very hard to delineate the border of the *HC* at such *CSF*-tissue boundaries. In comparison, *CSF* yields the highest signals in T2-weighted images and therefore yields bright pixels. This allows for an accurate labelling of the hippocampal border in such regions and also *CSF* inclusions within the *HC* itself. Moreover, the transitions from white matter (WM) and grey matter (GM) are more pronounced in T2, which is due to slightly better contrast and the higher resolution.



(a) T1 close up

(b) T2 close up

Figure 5.12: Close up comparison (T1 vs T2) of the HC and surrounding structures. Slices are not identical as both modalities are shown in their native orientation; for T1: coronal slice at $1\,\mathrm{mm}^3$ isotropic resolution while for T2: an oblique coronal slice at $0.47 \times 0.47 \times 1\,\mathrm{mm}^3$. Green shows the manual ground truth label of the left HC.

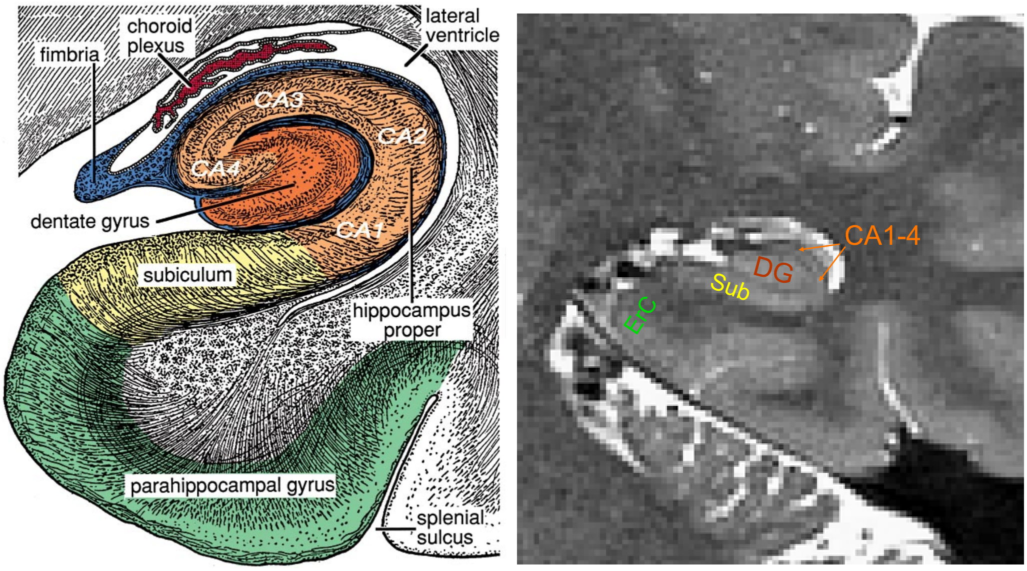
5.1.3.2 Annotation Protocol

The protocol by Berron et al. [8] was especially designed to annotate various subfields of the *MTL*, including cortical subregions as well as the *HC* and its subfields. The *hippocampal formation* consists of three structures (Fig. 5.13) (i) the dentate gyrus (DG), (ii) the cornu ammonis (CA) areas, and (iii) the subiculum (Sub).

The *CA* areas can be subdivided into four subfields, namely $CA1 - CA4$, which actually are the "true" *HC* also termed *hippocampus proper*. A more common division of the hippocampal formation with respect to segmentation is the grouping into hippocampal head (HH), hippocampal body (HB) and *HT*. Based on [8], the *HH* and *HB* are composed of the *DG*, the *CA* fields and the *Sub*. It should be noted that the groupings into *HH*, *HB*, *HT* or between different subfields are usually not separated by a distinct border but rather are continuous transitions from one structure into the other. This is why specific rules to consistently asses such ambiguous regions are needed and why delineating, especially the hippocampal subfields, is an exhaustive and tedious task.

In this thesis the terms *HC* and *hippocampal formation* are used interchangeably. Moreover, we have limited our manual annotations to three labels; the *HH*, the combination of both the *HB* and the *HT* as well as *CSF*. *CSF* is separately labelled as it is not part of the hippocampal tissue and should therefore be excluded. Hippocampal subfields, or bordering cortical regions are not explicitly labelled as this would need an additional extensive amount of time and are also not important for the aim of this thesis.

The application of the segmentation protocol to our data, including problematic areas, is briefly explained in the next section.



(a) annotated schematic[4] **(b)** cropped section of a T2-weighted slice

Figure 5.13: Detailed visualisation of the hippocampal formation setup. (a) shows a schematic coronal view of the MTL including the hippocampal formation and its surrounding structures. (b) shows a similar view of a real MR image with annotations of the most important parts.

5.1.3.3 Application of the Segmentation Protocol

Common practice for labelling the hippocampal formation is to exclude certain structures that can occur around or within the *HC*. Namely, alveus, fimbria, blood vessels and *CSF*. Alveus and fimbria are structures that cover the *HC*, which is generally enclosed by *WM*. All of them, as well as blood vessels, appear hypointense (dark) in T2-weighted images and are excluded from anatomical masks. In comparison, *CSF* and cysts shows hyperintense (bright) intensities.

To follow the instructions ([8]) and the subsequent elaboration, it should be noted that the segmentation is performed on the oblique coronal slices in an anterior - posterior (front to back / *HH* to *HT*) direction.

[2] vanat.cvm.umn.edu/ (accessed 25.05.2020)

Detailed instructions how to label the *HH*, *HB* and *HT* can be found elsewhere (Berron et al. [8], section 2.5.2.2). However, problematic regions that needed special attention are pointed out below.

Superior *HH* Outline. At the beginning of the *HH*, before the *Sub* reaches the medial surface, the medial and superior (upper) boundary of the *Sub* is determined by the entorhinal cortex (ErC). However, this border is not always clearly visible.

The superior boundary in general remains a problematic region throughout the whole segmentation of the *HH*. Additionally to the superior outline of the *Sub*, starting around the mid-slices of the *HH*, the upper contour of the hippocampal subfields (CA1 and CA3) also do not exhibit distinct borders. Moreover, this supero-medial area of the *Sub* and *CA* regions often exhibits signal cancellation due to neighbouring blood vessels.

Inferior Subiculum Transition. Moreover, around the first third of the *HH*, the *Sub* splits into an superior and inferior (lower) part. Also the inferior boundary of the *Sub* is often not clearly visible and suffers from signal loss. While the *Sub* inferiorly transitions into the *ErC* within the *HH*, the perirhinal cortex (PrC) becomes the inferior structure around the beginning of the *HB*. Independent of the inferior structure, these merging areas are hard to label due to signal loss that occurs also in the 7T reference images, but is more pronounced in our 3T data. However, the *Sub* should maintain roughly the same thickness, which can be exploited to estimate a better annotation.

Hippocampal Tail. Within the *HT* it is also sometimes hard to determine the exact infero-medial border. Moreover, the supero-lateral borders can be a bit unclear and especially the end (last slice) of the tail is problematic. However, the latter region is not often present due to the short coverage of the *HT* in our data.

General Remarks. For the whole annotation process it is recommended to attempt *smooth curvature* even if hypointense edges are discontinuous. This should not only be tried to achieved in-plane but also for smooth transitions between slices. Therefore, switching back and forth through the oblique coronal slices is beneficial and can also help to determine missing borders of the current slice based on the outlines of the previous and successive slices.

5.1.3.4 Annotation Process

The ground truth annotations were performed by the author of this thesis and a second master students, namely Alina Dima. Both of us have been studying the instructions by [8] in order to get familiar with the required anatomical background and landmarks necessary to conduct the segmentation. Moreover, prior the annotation of the data in our dataset, a view cases have been labelled independently and together to get to know with the segmentation software and to identify possible problematic areas. This initial learning phase was done under supervision of Prof. Stefan Ropele, who has many years of experience in the field of neuroimaging via *MRI*. Final labelling, including corrections, took between 5 – 10 hours per *HC* mask.

Annotations were performed with *DispImage*, an Interactive Data Language (IDL) based in-house segmentation tool made by Prof. Ropele. DispImage allows for slice wise labelling of structures, by drawing a closed contour around the *ROIs*. The "value" of the current label can be individually set, which was used to create different labels for the *HH*, the *HB* and *HT* as well as *CSF*. Moreover, annotations can be done at even a sub-pixel resolution, which uncouples the drawing of the contour from the pixel size. For sub-pixel segmentations, images are internally upsampled prior the actual labelling. The contour is parametrised as a closed polygon, defined by the points of the polygon chain. This set of *2D* points, together with the information of the current slice, is saved as a *.roi file to enable subsequent editing. Moreover, a proper mask, based on the original resolution of the underlying image can be exported.

The way how our masks were generated can be briefly described as stated below and was executed with a Python script. An original image pixel is considered to be part of the mask if more than 50 % of the 16 times upsampled image, so at least eight sub-pixels of the current pixel, are within the contour. This approach ensures a certain granularity of the final mask, while also requiring a minimal occupancy of the pixel.

Based on this procedure, 29 hippocampi have been segmented from 24 different subjects, whereas from 5 of them both hippocampi were labelled. One mask (subject), however, was eventually removed from the dataset due to severe motion and really bad *SNR*. Therefore, our manual ground truth annotations comprise a total of 28 annotated hippocampi; 14 for each hemisphere.

5.2 Registration

Registration of images is the process of accomplishing spatial correspondence between two images. The reasons for registration of our dataset are twofold. Patient movement in between both *MR* scans, namely the T1-weighted *MPRAGE* and the T2-weighted *TSE*, causes sligthly diverse orientations of both scans. However, as already described, yields motion during our *TSE* acquisition a misalignment of the odd and even slice slabs. Both of these misalignments were reduced with voxel-based registration steps, which are described below.

5.2.1 T2 Alignment

Patient motion during the acquisition of our interleaved based T2-weighted scans can cause misalignment of the odd and even slices within the T2 volume. A possible result of such a misalignment, are anti-parallel odd and even slice stacks, which was already shown in Fig. 5.5. In such a case (Fig. 5.5), there exist small parts of the actual T2 volume that are not pictured at all, whereas for some other regions the sub-volume is imaged twice. For the latter, tissue information is contained twice; in the current and in a neighbouring slice. However, tissue information of the sub-volumes that are never covered by any imaged odd or even slice of the T2 slab is simply lost and cannot be recovered without additional *MR* scans until even each sub-volume was covered. This effect is schematically visualised in Fig. 5.14.

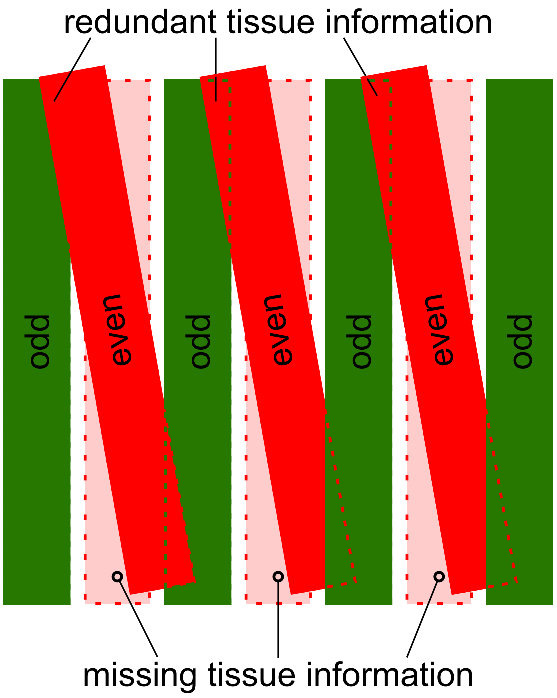


Figure 5.14: Motion induced intra-T2-volume information loss. *Dashed* even slices (- - -) should indicate the correct configuration the even block. Transparent light red filling of the dashed slices, indicates the parts of the corresponding even slice that is not images at all (missing tissue information). Filled green refers to slices of the odd image block, where the dashed (- - -) line should indicate the outline of the part that is images twice.

The aforementioned information loss needs to be considered in the intra-volume alignment of the T2 scan. To cope with anti-parallel slices, out-of-plane rotation (around the x-axis) is necessary, which requires interpolation between slices. However, due to our anisotropic resolution, with a big slice thickness of 1 mm compared to our in-plane resolution of $0.47 \times 0.47\,\text{mm}^2$, and the possible missing information this interpolation would annihilate the details of our ultra-high in-plane resolution.

If the patient's motion, however, preserves the parallelism of the odd and even slice slabs only in-plane transformations need to be considered. Thus, interpolation only occurs within the high resolution of the oblique coronal T2 slices. Therefore, we have limited our T2 slab registration to cope only with in-plane misalignment. This approach is also supported by the inner structure of the hippocampal formation, which exhibits changes mainly within the cross-section and not between slices.

In this first step of our registration, the T2 volumes were aligned by 2D (in-plane) rigid transformations. The registration scheme is performed in such a way that, starting with the second slice, each slice is individually aligned to its previous slice. This yields the intra-T2-volume aligned images, so the T2 images with aligned odd and even slices.

5.2.2 T2 to T1 Registration

In the next step, the previously aligned T2 scans (denoted as T2') are then registered to their T1-weighted counterpart. This registration is now an inter-modality (T2 to T1) registration problem, which is harder to optimise because of the different image intensity for the same anatomical structure. As there are now two intra-volume aligned *3D* images, there is no need for slice-wise (2D) registration. Therefore, *3D* rigid transformations are used, which means that only one transformation for the whole T2' volume needs to be calculated. However, transformations are still limited to in-plane translations and rotations.

The resulting "pre-"registered T2 images (denoted as T2") together with the T1 scans are now used in an auxiliary image synthesis task. A stacked convolutional neural network (CNN) was trained to mimic the T2 contrast of the corresponding T1 input images. For the optimisation a $\mathcal{L}_1$loss between the network output and pre-registered T2 images (T2") was used. Details about the architecture and its training procedure will be given at a later stage in this thesis. Note, that the intention of this auxiliary task was not to achieve a perfect representation of the T2 contrast including all anatomical structures, but rather increase the similarity of the two images that are registered to each other in our second registration step. Based on the assumption that this image synthesis task does not introduce any spatial transformations, the synthesised T2 image (T2s) should be well aligned with the original T1 image. Therefore, the registration is repeated between T2" and T2s, which is now an easier to optimise intra-modality registration.

Eventually, these three single transformations are combined and applied as a single transformation to the slices of the original T2 image, which yields the final registered T2-weighted image.

General Remarks All three image registration steps used in-plane rigid transformations. ANTs cross-correlation (Eq. (4.7)) was used as the similarity measure together with a conjugate gradient descent approach to optimise the transformations. This local optimisation scheme is applied in combination with a multi-scale approach to increase the capture range and diminish the influence of local optima.

5.2.3 Ground Truth Label Registration

The combined transformations, from the image registration steps, are also used to transform the manual ground truth labels, which have been created in the original T2 image space (Section 5.1.3), to the corresponding T1 space.

Together, all these registrations yield the final dataset including the resampled T1 images, the registered T2 images as well as the transformed ground truth labels.

For all experiments, except of the procession with the *FreeSurfer* pipeline, the original T1 images were resampled and cropped to match the resolution ($0.47 \times 0.47 \times 1\,\mathrm{mm}^3$) and size of the T2-weighted images.

5.3 FreeSurfer Pipeline

To establish a baseline for the automated segmentation of our T1-weighted images, the same 23 subjects that are included in our dataset were processed with the *FreeSurfer* pipeline. Specifically *freesurfer-Linux-centos6_x86_64-stable-pub-v6.0.0* on a Linux Ubuntu 18.04.3 LTS machine. *FreeSurfer*'s "hippocampal-subfields-T1" protocol was used, which calculates the labels for the hippocampal subfields based on *FreeSurfer*'s high-resolution "ex-vivo atlas". To use the *hippocampal-subfields-T1* protocol, each subject has to be pre-processed with the standard *FreeSurfer* pipeline, called *recon-all*.

The *recon-all* script was executed with the "-all" flag, which performs cortical and subcortical parcellation of the brain. The "hippocampal-subfields-T1" flag can be additionally set already with the initial procession or be separately executed afterwards. In total, the segmentation of one subject took around 9 hours, which can be reduced to 4.5 h by using the "parallel" flag.

A thorough description of all performed steps and settings of *FreeSurfer*'s *recon-all* pipeline can be found on their web-pages [5] [6]. Detailed information about the hippocampal-subfields protocol can be found in Iglesias et al. [52] or online [7].

5.3.1 FreeSurfer Label Post-Processing

As common for atlas-based segmentation methods, the input images are registered into the space of the reference template (the atlas). Segmentation results from the *recon-all* pipeline are provided in the standard FreeSurfer (FS) space (FSvoxelSpace), which corresponds to *FreeSurfer*'s T1 image space. *FreeSurfer* T1 images have dimensions of $256 \times 256 \times 256$ pixels at an isotropic resolution of $1 \times 1 \times 1\,\mathrm{mm}^3$. The segmentation result is given as an image in which each pixel intensity represents a unique label. A MATLAB script was implemented to first extract the necessary labels of the hippocampal formation, based on *FreeSurfer*'s lookup table (LUT)[8], and to subsequently create the binary masks.

In order to compare the *FS* segmentations to our manual ground truth labels, they also need to be registered to the corresponding space. Therefore, the inverse transformation of the initial transformation, which was internally applied by *FreeSurfer* to bring the original T1 images into the template space, was utilised. In detail, a rigid transformation with

[5] https://surfer.nmr.mgh.harvard.edu/fswiki/recon-all

[6] https://surfer.nmr.mgh.harvard.edu/fswiki/ReconAllTableStableV6.0

[7] https://surfer.nmr.mgh.harvard.edu/fswiki/HippocampalSubfields

[8] https://surfer.nmr.mgh.harvard.edu/fswiki/FsTutorial/AnatomicalROI/FreeSurferColorLUT

mutual information (MI) as similarity metric and trilinear interpolation was applied. The same transformations were used for the corresponding segmentation masks.

There is one additional post-processing step required before the *FS* masks can be compared with our manual ground truth masks. *FreeSurfer* uses the *3D* whole head T1-weighted images for segmentation. Therefore, also the whole hippocampus is labelled. However, our T2-weighted images only cover a certain area of the brain which sometimes does not include the complete *HC*. Thus, equivalently to our T1-weighted images, the *FS* segmentations also needed to be cropped and resampled to match the *FoV* of the T2-weighted scans.

5.4 Network-Based Setup

In this section insights into the training setup of our neural networks are given. The preparation of the input data and the data augmentation is described. Moreover, all utilised network architectures for the segmentation and the image-to-image translation task are briefly reported.

5.4.1 Data Preparation

The generation of our dataset was already described in Section 5.1. However, for training a neural network, several preparation steps are required or rather are performed to improve the training and thus the results. All preparations were calculated and applied on the fly to prevent excessive data storage. Moreover, transformations were computed separately, combined and then applied at once such that only one interpolation step was needed. Applied transformations are listed below.

Normalisation. *MR* images often have some outliers or show different intensities for similar tissues. One such example was already shown in Fig. 5.6. To address this problem, a slice-wise intensity normalisation was performed. Therefore, intensity values were shifted by the mean and divided by the standard deviation of the current slice.

Flipping. In order to ensure similarity of all our input patches, we decided to perform all learning-based operations on equally oriented hippocampi. Specifically, hippocampi from the right hemisphere were flipped to be oriented in the same way as the hippocampi from the left hemisphere. This ensures that the network is trained on hippocampi from both hemispheres, while increasing the training data (shared information) compared to a separate segmentation of either the left- or right side. This is also supported by the inherent, intra-subject anatomical variations between the left and right *HC*. Therefore, even if both hippocampi from the same subject are used and flipped to the same side, it results in two different data samples with slightly different shape, orientation and coverage. This in turn helps the network to generalise.

Data Augmentation. Extensive data augmentation was used to enhance the variability and quantity of the training data, such that the model can again better generalise and is less prone to overfit. Augmentation is essential to induce invariance and robustness, when using small datasets [80]. In detail, the dataset images (all in original T2 dimensions) underwent ***translations*** in a range of $[-10, 10]$ millimetre, ***rotations*** by $[-20°, 20°]$ and ***scaling*** by a factor of $[0.8, 1.2]$. Moreover, ***elastic deformations*** were applied to better simulate anatomical variations by distorting points of a coarse 6×6 grid by values between $[-15, 15]$ millimetre and interpolating with third order B-splines.

All transformation values were randomly drawn from a uniform distribution within the given intervals, again limited to the coronal plane and applied to the whole *3D* image slab. Images were resampled with linear interpolation, whereas masks were resampled via nearest neighbour interpolation.

Specific to our segmentation experiments an additional transformation have been applied. Namely, ***intensity variations*** were introduced via affine augmentation of the form $f(x) = ax + b$ where a represents a uniform random scaling by $\pm 15\,\%$ and b is a random offset in a range of ± 0.15.

Cropping. After data augmentation, the images could already be used for training. However, as we only segment the hippocampal formation, just one side at a time, it is not necessary to use the whole images of the dataset. On the contrary, it would only dramatically increase computational costs, especially as we use volumetric data with such a high-resolution. Therefore, the dataset images and labels were cropped to a certain area around the *HC*. Thus, our segmentation task can be seen as a two step approach, consisting of a localisation task and the actual labelling. For our segmentation, this localisation task is also not that critical as only the rough location of the hippocampal formation is required to crop the input to a patch which includes the *HC*. As this thesis focusses on the segmentation task no special approach was implemented to perform the localisation task. However, possible methods could be another dedicated network, or template-based cropping approaches which kind of define the centre of mass of the hippocampus.

Since we are only able to train and validate our methods with annotated data, there was currently no need for a separate localisation method as the centre of mass could be simply calculated based on the manual ground truth annotations. Therefore, the true centre of mass was used to crop the input images and masks to an in-plane patch size of 96×64, while all 40 slices were kept.

Label Simplification and Smoothing. The last data preparation step was only applied to the manual ground truth masks. In detail, our three labels (*HH*, *HB* and *HT*, *CSF*) were unified into only one label representing the whole hippocampal formation. To alleviate annotation inconsistencies and provide smoother training labels, the binary label image was smoothed with a Gaussian filter ($\sigma = 0.8$) before resampling. As this results in non-integer values, which is not usable within our optimisation framework, a thresholding of 0.5 is applied to the resampled label image to generate again a binary label image.

5.4.2 Segmentation Architecture

A network architecture describes the applied operations and their hierarchical setup of a neural network. Our segmentation network is based the popular U-net [80] architecture. The U-net architecture consists of a contracting / downsampling path and an expanding / upsampling path. In our case, the U-net has an overall downsampling factor of 16, which was achieved by four contracting blocks. In each contracting block, a full *3D* $(3 \times 3 \times 3)$, reflect padded convolution with 64 filter channels, followed by a scaled exponential linear unit (SELU) [59] activation and an average pooling operation with stride 2 was used. The expanding blocks are essentially the reversal of the contracting blocks. Therefore, a deconvolution (up-convolution) layer with stride 2 to upsample the images as well as fully *3D* padded convolutions were applied. All up and downsampling operations were performed only in-plane such that the lower resolution, along the slice thickness, is not affected from the downsampling. In comparison to the original U-net [80], features from the contracting path were combined with the upsampling path via addition. Moreover, instead of learning the upsampling via deconvolution layers linear upsampling was used. The final amount of filter channels for this task was two, as we only have two labels, hippocampus and background.

This single architecture (Fig. 5.15) was used to pre-train the segmentation for the combined task including the image synthesis followed by the *HC* segmentation. For experiments, in which only the segmentation task was executed, this architecture was stacked to have two consecutive U-nets (Fig. 5.16) with intermediate supervision.

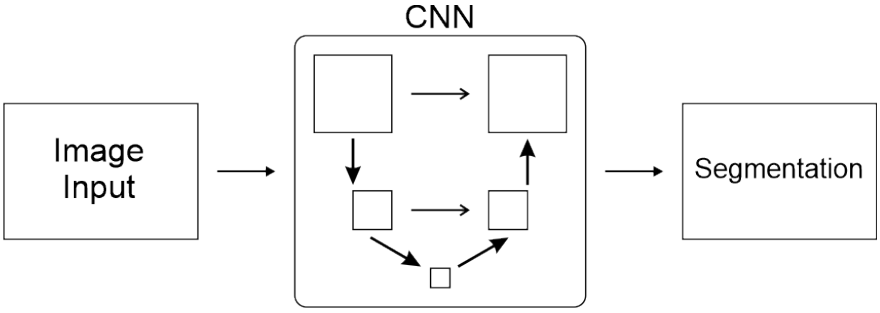


Figure 5.15: Schematic single CNN architecture.

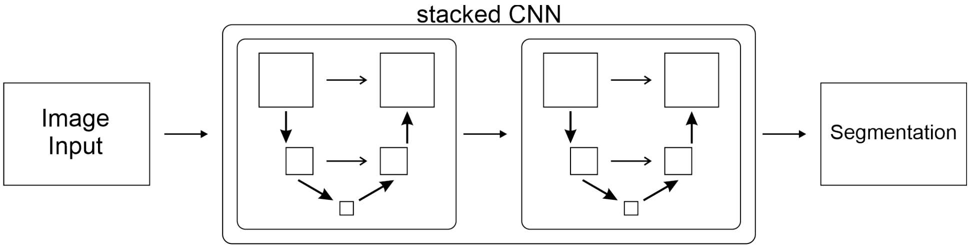


Figure 5.16: Schematic stacked CNN architecture.

5.4.3 Image Translation Architecture

This section refers to the architectures that have been used in our image-to-image translation task to synthesise T2-weighted images based on their T1-weighted counterparts. As described in the theory chapter about generative adversarial networks (GANs), this process represents a minimax two-player game between the *generator G* and the *discriminator D*.

5.4.3.1 Generator Network

The *generator* network was used to create synthetic T2 images and utilised the same architecture as used in the segmentation task. Moreover, for the generator only the single architecture setup was used for two reasons. However, the final amount of channels is now one, as we have grey-scale images. First, memory limitations of the utilised hardware and second, for better comparability between all experiments. The image synthesis task was only used as a first step, prior the successive segmentation task. Therefore, to limit the utilised architectures throughout all experiments to two stacked networks only the single setup was used for each subtask, respectively.

5.4.3.2 Discriminator Network

The *discriminator* architecture differs from the other U-net based networks used in this thesis. Fundamentally, it resembles an encoder network that maps the image input to feature representations in order to estimate whether the input is real or synthetic. Our discriminator architecture consists of five levels with an overall downsampling factor of 16, which was achieved by four contracting blocks. In each block a $3 \times 3 \times 3$ convolution with zero-padding, 64 filter channels and leaky rectified linear unit (ReLU) activation was applied. For downsampling, an average pooling layer was used.

6 Experimental Setup

In this section all related information to the experiments is given. At first the applied cross-validation (CV) setup is reported. Following, a description of the training process is given and at last, all conducted experiments are elaborated.

6.1 Cross-Validation Setup

In order to evaluate the results of deep learning-based methods and analyse their generalisation qualities, different data samples need to be used for training and evaluating of a neural network. Such splitting of the dataset ensures that the model is validated/tested only with previously unseen data and thus is not biased by this particular data sample. To prevent such a biased evaluation, we have divided our 28 labelled data samples into a *training set* including 18 samples and a *validation set* containing the remaining samples. As a countermeasure to the small dataset, a 3-fold *CV* was performed. This ensured that each data sample was used for training and inference at least once, while still not being used for both tasks at the same time. These *CV* sets were also balanced as good as possible with respect to the annotator and the labelled hemisphere. Based on initial experiments, we decided to use the sample of one particular subject only during inference. The reason for this decision was, that this scan is missing the first slices of the *HH* and therefore covers not only the whole *HT* but also two successive slices. Thus, approximately the last five slices look quite different from the remaining samples.

6.2 Training Procedure

This section presents insights into the actual training process of the neural networks. Utilised loss functions, optimisers and their parameters are reported. For all experiments Python v3.6 with the deep learning framework TensorFlow v1.13.1 were used. Training of the models was performed with a Nvidia Titan V with 12GB of VRAM and took 50 h for the image synthesis task and about 10 h for the segmentation per *CV* set.

6.2.1 Segmentation Task

The utilised architectures for the segmentation task have been described in Section 5.4.2. To train both architectures, either the single or the stacked setup, prepared data (Section 5.4.1) was fed with a batch size of two. A cross-entropy loss in combination with a default parametrised adaptive moment estimation (ADAM) ([58]) optimiser was utilised to train the segmentation task (Section 4.3.4). The initial learning rate was set to 10^{-5} with an exponential decay every 2500 iterations over the course of 60 000 iterations, which was the total training amount.

6.2.2 Image Synthesis Task

Based on the architecture setup of a *GAN*, two different networks, namely the generator G and the discriminator D, need to be trained for the image synthesis task. We have implemented our *GAN* as a Wasserstein GAN with gradient penalty (WGAN-GP) (Section 4.3.6.2). Both sub-networks (G and D) were trained simultaneously, however, the discriminator was updated $n_{critic} = 5$ times per generator iteration. The generator was trained for 80 000 iterations. Moreover, for both networks the same learning rate was applied, starting at 10^{-5} with an exponential decay every 2500 iterations.

Generator. For the generator, also the single setup of the described segmentation architecture was used (Section 5.4.3). Moreover, we utilised a conditioned generator, which means that the synthetic images are not generated based on noise, but rather on the corresponding T1-weighted images. Therefore, training of G was performed with T1 input images with a batch size of two and an *ADAM* optimiser. In comparison to the segmentation tasks, a combination of a $\mathcal{L}_1$and an adversarial loss (see Section 4.3.6.3) was applied for this task.

Discriminator. The discriminator architecture was introduced in Section 5.4.3. To optimise this network, again an *ADAM* optimiser was used, however, with different parameters. The settings were chosen similar to [43] with $\beta_1 = 0.5$, $\beta_1 = 0.9$, $\epsilon = 10^{-8}$.

6.3 Experiments

This chapter provides an outline of all conducted experiments including a brief summary of their inputs, utilised architecture and training method. Moreover, a motivation behind each experiment is given.

6.3.1 FreeSurfer Experiments

FreeSurfer (*FS*) was used to get an initial automated segmentation of our subjects, which can be utilised as a baseline result of the automated segmentation methods. The detailed setup was described in Section 5.3. The post-processed *FS* masks (Section 5.3.1), henceforth only referred to as *FS-labels* or *FS-masks*, were used in the following experiments.

In order to evaluate how good learning-based methods are able to reproduce the *FreeSurfer* segmentations, we have trained the stacked segmentation network (Section 5.4.2) together with the same T1-weighted images that were utilised with *FreeSurfer*. The processed *FS* masks were then used as training labels for this experiment. This setup is visualised in Fig. 6.1.

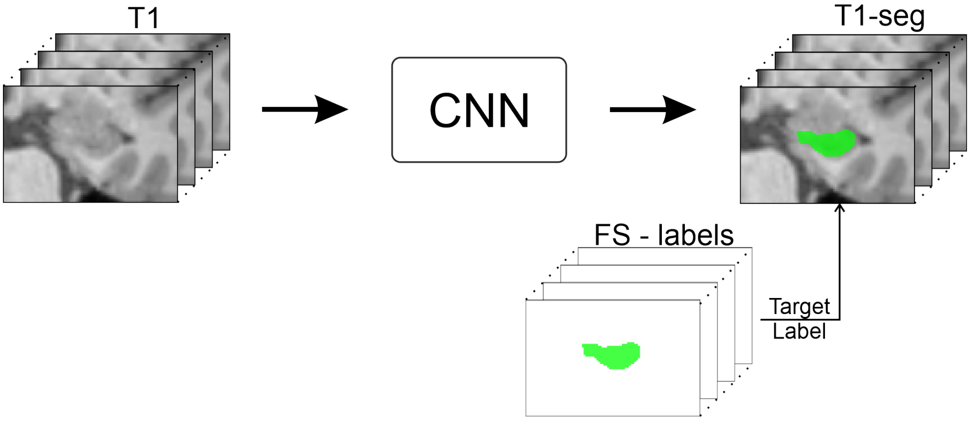


Figure 6.1: Training setup of network-based T1 segmentation with FS target labels.

6.3.2 Learning-based Segmentation Experiments

After evaluating the reproducibility of the *FreeSurfer* segmentations and their impact when used as target labels during training, the outcome of our manual annotations was tested. To analyse the performance of the networks in combination with our processed manual ground truth annotations, henceforth called ground truth (GT) labels, simple network based segmentation experiments were performed. Therefore, also the stacked segmentation architecture (Section 5.4.2) was used. However, instead of the FS-labels, our GT-labels were now applied. Moreover, to evaluate the impact of the different image modalities (T1- vs. T2-weighted images), the same network setup was trained three times with different inputs. To ensure comparability, always the same settings and only different input images (i) T1-weighted images, (ii) T2-weighted images, and (iii) T1- and T2-weighted images were used.

The setup for the T1-weighted and the T2-weighted input images is shown in Fig. 6.2. Note that either the upper (T1) or the lower (T2) configuration is trained at a time. For the experiment utilising T1- and T2-weighted input data, images of both modalities were simply concatenated along the channel dimension and fed as one input.

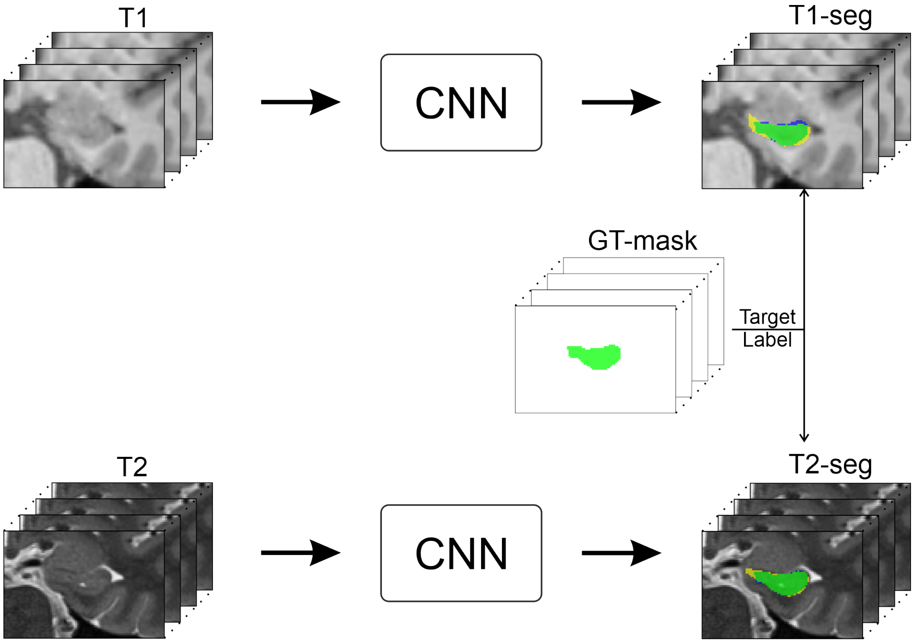


Figure 6.2: Training setup of network-based T1- or T2 segmentation with manual GT target labels.

6.3.3 T2-enhanced Training of Deep NNs

After performing the simple segmentation experiments, showing an outline of the learning-based segmentation results, setups incorporating the T2 image information into the training process were implemented. Therefore, two different approaches were tested (i) segmentation of T1 input with a model pre-trained on T2 input, and (ii) segmentation of synthetic T2 images with a model pre-trained on T2 input.

Pre-Trained T2 Segmentation Models

To get the required pre-trained segmentation models, the single (Fig. 5.15) and the stacked architecture (Fig. 5.16) setup were trained as described in Section 6.2.1 with T2-weighted input images and our GT-labels.

The model of the stacked architecture was already trained for the simple T2 segmentation experiment described in Section 6.3.2. For the current experiment, this model has been loaded and tuned with T1 input images. The tuning was performed for another 40 000 iterations with an initial learning rate of 10^{-5} and a step-wise exponential decay.

Synthetic T2 Image Segmentation

In this experiment the *GAN* architecture is utilised, as explained in Section 5.4.3. The training process was executed as described in Section 6.2.2.

After pre-training the image synthesis task and the T2-based segmentation, each one with the single architecture setup, both of these models were loaded for the actual synthetic T2 image segmentation experiment. From the pre-trained *GAN* setup, only the generator was loaded as there is no tuning involving the discriminator or an adversarial loss anymore. The trained generator G is merely loaded to create the synthetic T2 images on the fly for the successive segmentation network. This is necessary, as for our data augmentation the full sized images are required and not only the cropped patches ($96 \times 64 \times 40$) including the *HC*, which are, however, the output of the augmentation process and thus also reflect the output dimensions of the networks.

The generator model of the image synthesis is loaded first, while the pre-trained T2 segmentation is stacked after G of the synthesis task (see Fig. 6.3). As mentioned in the training process (Section 6.2.2), G is fed with the T1-weighted image patches of the dataset. Thus, the first network (pre-trained G) generates synthetic T2 images, based on their corresponding T1-weighted input images, which are then used as inputs for the subsequent segmentation network. The initial T1 input was concatenated with all feature channels of the image generation task and fed into the second network, the pre-trained T2-based segmentation architecture.

This setup was then also trained for another 40 000 iterations with an initial learning rate of 10^{-5} with a step-wise exponential decay. Similar to the normal segmentation experiments, a cross-entropy loss on the segmentation output together with an *ADAM* optimiser was used for training. The parameters of both networks, the generator and the segmentation architecture, were updated during this optimisation process.

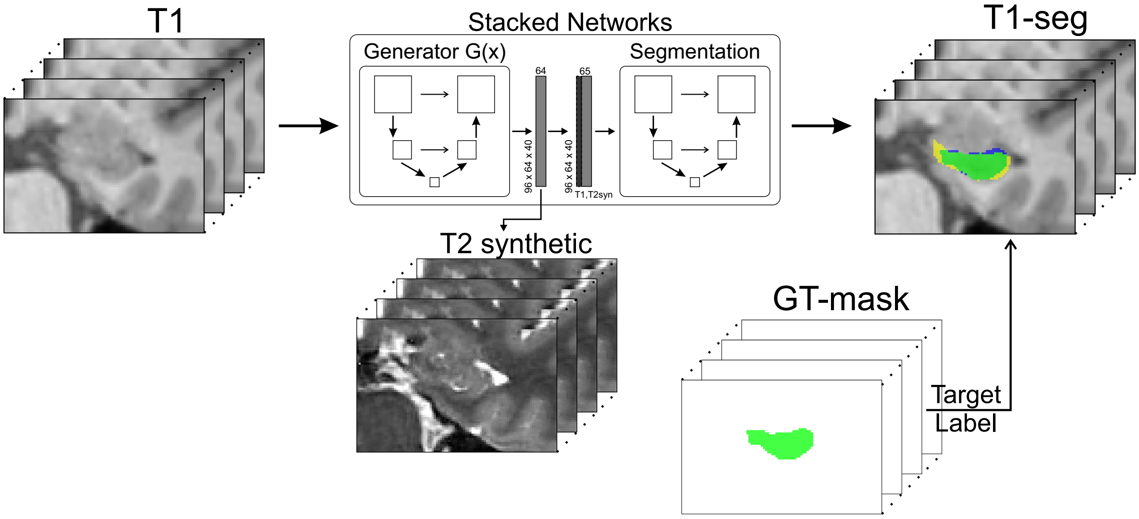


Figure 6.3: Training setup for the synthetic T2 image segmentation.

6.4 Evaluation

In this last section, metrics applied to evaluate the segmentation results of the conducted experiments are described. To evaluate the segmentations, two main measures were used; *Dice coefficients* and *surface distances.* Independently of these measures, segmentation masks were also visually spot-checked.

To assess the image quality of the synthetically created images mimicking T2 contrast, we use peak signal-to-noise ratio (PSNR). However, as this does not completely represent the perceived image quality, we mainly use visual inspection.

Quantitative and qualitative results will be presented in the next chapter 7.

6.4.1 Dice Similarity Coefficient

The Dice similarity coefficient (DSC), also known as F_1 score, is an accuracy measure calculated by means of the precision and the sensitivity. For binary classification it can be calculated as:

$$DSC = 2 \cdot \frac{precision \cdot sensitivity}{precision + sensitivity} = \frac{2TP}{2TP + FP + FN} \,, \tag{6.1}$$

with $precision = {}^{TP}/_{(TP+FP)}$, $sensitivity = {}^{TP}/_{(TP+FN)}$, TP ... True Positive, FP ... False Positive and FN ... False Negative.

DSCs were calculated between the automated segmentation results and our manual ground truth annotations.

6.4.2 Surface Distances

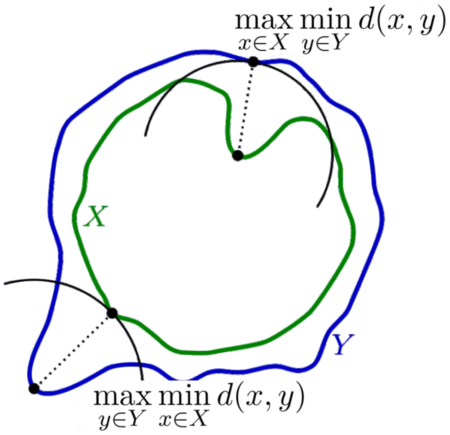
Additionally to Dice coefficients, surface distances were calculated of which mainly the *Hausdorff distance* was used for comparison of our results. Given two subsets of a metric space, the Hausdorff distance (d_H) describes the maximum distance of a point set to the nearest point in the other set and thus, is also referred to as d_{max}. More formally, the *bidirectional Hausdorff distance* between the point sets $X = \{x_1, x_2, \ldots, x_n\}$ and $Y = \{y_1, y_2, \ldots, y_n\}$ can generally be defined as:

$$d_H(X,Y) = d_{max}(X,Y) = \max \left\{ \max_{x \in X} \min_{y \in Y} d(x,y), \; \max_{y \in Y} \min_{x \in X} d(x,y) \right\} \,, \tag{6.2}$$

where $d(x, y)$... can be any appropriate distance metric [82].

In our case the two subsets correspond to the voxels of the segmentation boundary of either the automated segmentation masks or our manual ground truth annotations. As d_{max} only represents the maximal error between both masks, this measure alone is not representative for the overall quality of the predictions. Therefore, additionally the *mean-* and *median distances* between both sets were calculated.

Based on our in-plane resolution of $0.47 \times 0.47\,\text{mm}^2$, a *surface distance* below 0.47 mm, corresponds to an error by up to one pixel. A *median distance* of $d_{median} = 0\,mm$ indicates


$\max_{x \in X} \min_{y \in Y} d(x, y)$
X
Y
$\max_{y \in Y} \min_{x \in X} d(x, y)$

by Rocchini, CC BY 3.0, accessed 08.06.2020
https://en.wikipedia.org/wiki/Hausdorff_distance#/media/File:Hausdorff_distance_sample.svg

Figure 6.4: Exemplaric visualisation of the one-sided Hausdorff.

that more than 50 % of all border pixels are correct. The combination of all three surface measures, together with the *DSC* allows for a holistic evaluation of the segmentation predictions.

6.4.3 Peak Signal-to-Noise Ratio

To assess the image quality of the generated synthetic T2 images, the *PSNR* was used. The *PSNR* is defined as the ratio of the maximal possible signal power to the power of its corrupting noise. The noise is commonly described by means of the mean squared error (MSE), which is given for a *2D* image as:

$$MSE = \frac{1}{MN} \sum_{i=0}^{M-1} \sum_{i=0}^{N-1} [y(i, j) - G(x)(i, j)]^2 \ , \tag{6.3}$$

where M and N... are the amount of pixels per dimension, y... is the real T2-weighted image and $G(x)$... is the synthetic T2 image created by the generator of the GAN.

6.4.4 Presentation of Results

Following, the way how the acquired results of this thesis are presented is briefly described.

Quantitative Results are presented by means of *DSC* in percent and surface distances in millimetres, as described in Section 6.4. In general all scores are presented in form of *mean* ± *standard deviation* over the results of all three cross-validation sets. Note, a standard deviation of ± 0.00 indicates that the value of the corresponding metric is exactly the same for all evaluated samples of all *CV* sets.

Qualitative results are shown to visualise the resulting segmentation masks of the utilised methods. For these visualisations, the following colour encoding is used:

- green ... true positive (TP) (union),
- yellow ... false positive (FP),
- blue ... false negative (FN),

where TP + FP = automated segmentation mask and TP + FN = manual ground truth mask.

7 Results and Discussion

Contents

This section presents and discusses quantitative and qualitative results of all our experiments. This includes the segmentations acquired with *FreeSurfer* and our deep convolutional neural networks (DCNNs) as well as an example of the auxiliary T2 synthesis task.

An overview of the conducted experiments and their setup as well as the evaluation procedure was given in Section 6.3, Section 6.2 and Section 6.4, respectively.

7.1 FreeSurfer Related Segmentation

In this section results related to FreeSurfer (FS) are presented. First of all this includes the baseline segmentations performed with the software package *FreeSurfer* as has been described in Section 5.3. Moreover, all *DCNN*-based experiments in which the post-processed *FS* masks haven been utilised (Section 6.3.1).

7.1.1 FreeSurfer Segmentation Pipeline

In order to get initial segmentations of the hippocampi in our dataset, accordingly to clinical research, our data has been processed with *FreeSurfer*'s brain segmentation pipeline. These masks also served as the baseline results of our segmentations.

Table 7.1 shows by the low Dice similarity coefficient (DSC) and high surface distances that the hippocampal segmentation with *FreeSurfer* yields only moderate agreement

Task	**Input**	**Labels**		**Evaluation**			
				Dice Score	Surface Distances in mm		
	Image	training	eval.	in %	d_H	d_{mean}	d_{median}
FreeSurfer	T1	-	GT	78.06 ± 3.99	4.66 ± 1.28	0.54 ± 0.10	0.47 ± 0.00

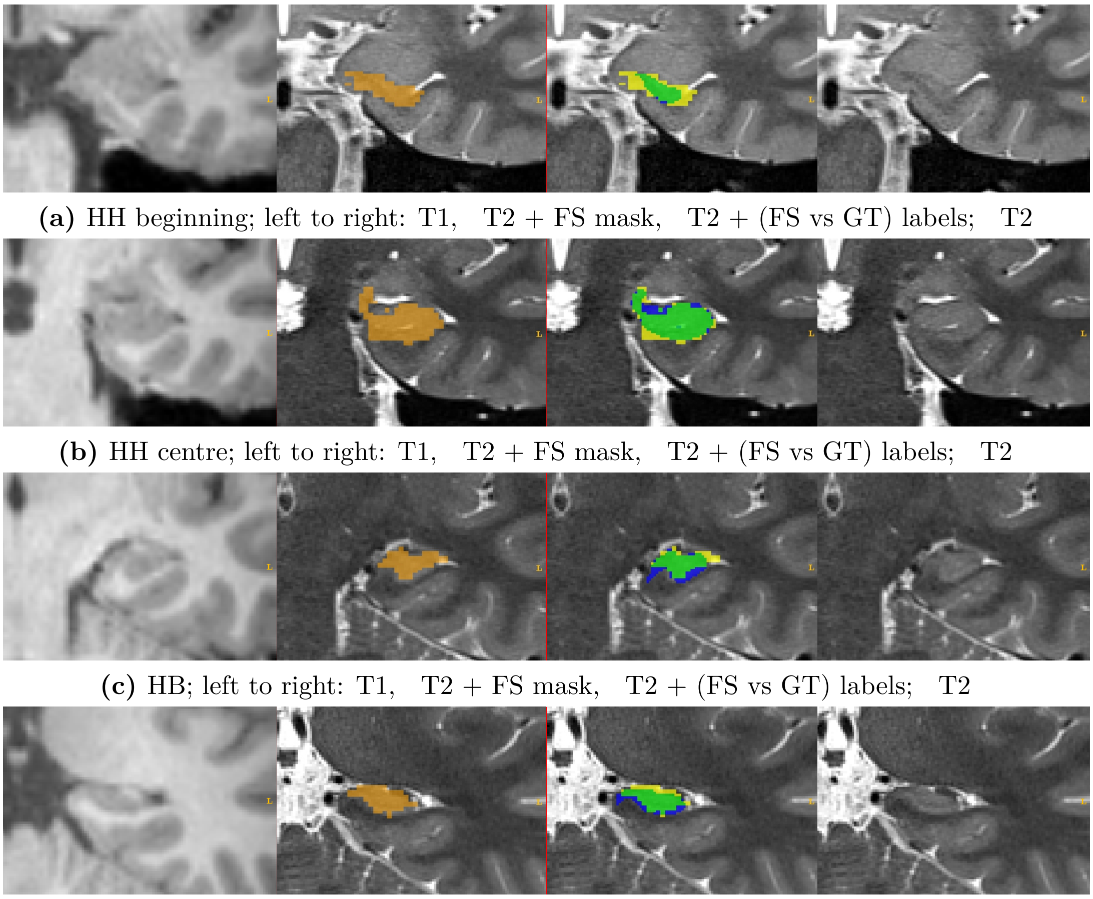
Table 7.1: Results of the *FreeSurfer* (atlas-based) segmentation.

with the manual ground truth annotations. The average of the *median distance* with $d_{median} = 0.47 \pm 0.00$ shows that for all evaluated samples, in all cross-validation (CV)-sets, more than 50 % of the border pixels of the *FS* labels are off by one pixel ($0.47 \times 0.47\,\mathrm{mm}^2$). The mean distance between the *FS* and GT set is with $d_{mean} = 0.54\,\mathrm{mm}$ a bit more than one pixel.

The mean pixel shift becomes more obvious when looking at the masks itself as shown in Fig. 7.1 and Fig. 7.2. These figures show the T1- and T2 images with qualitative results of an example hippocampus, segmented with the *FreeSurfer* pipeline (2. left column). Moreover, the comparison against the manual ground truth (GT) mask is given in form of an overlay visualised as true positive (TP), false positive (FP) and false negative (FN) pixels (2. right column).

Figure 7.1 shows coronal slices of the hippocampal head (HH) (beginning and centre), hippocampal body (HB) and hippocampal tail (HT). As one can see, the *FS* mask (orange) is in some areas too small, resulting in *FN* pixels (yellow), while in some regions too big, which yields *FP* pixels (blue). In general the hippocampus (HC) border and the transition of the subiculum (Sub) into the entorhinal cortex (ErC) are problematic, while the central part and general shape of the hippocampal formation is also followed quite well. Moreover, results show that it is not trivial to distinguish the *HC* from surrounding cerebrospinal fluid (CSF) on T1-weighted images, which were also used for the *FreeSurfer* pipeline. In contrary, *CSF* is easily identifiable on the T2-weighted images.

Figure 7.2 depicts the same subject and the same masks, however, in the sagittal view. This visualises the continuity of the masks along the slice direction. Moreover, it shows that the *FS* labels more slices of the *HH* and especially the *HT* compared to the manual ground truth.



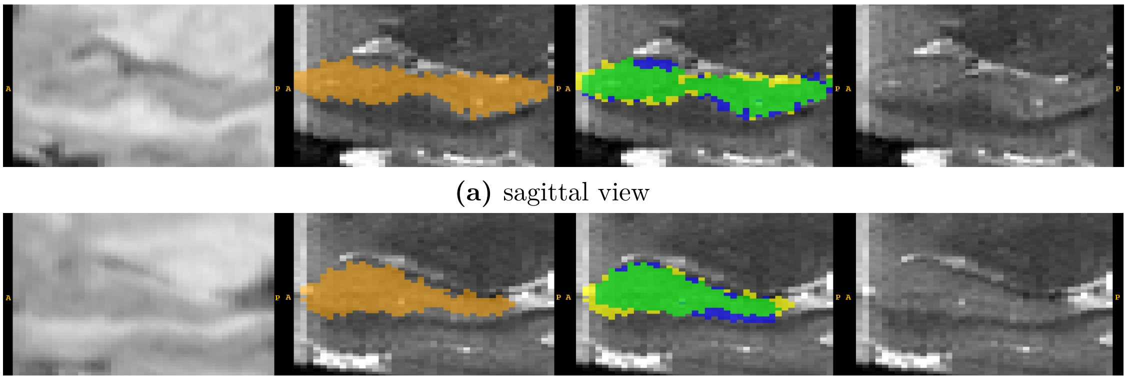
(a) HH beginning; left to right: T1, T2 + FS mask, T2 + (FS vs GT) labels; T2

(b) HH centre; left to right: T1, T2 + FS mask, T2 + (FS vs GT) labels; T2

(c) HB; left to right: T1, T2 + FS mask, T2 + (FS vs GT) labels; T2

(d) HT; left to right: T1, T2 + FS mask, T2 + (FS vs GT) labels; T2

Figure 7.1: Example FreeSurfer segmentation in coronal view. Left / 1. column: T1 image, all other images show T2 modality; orange/FS-mask; green/TP-pixels; yellow/FP-pixels; blue/FN-pixels



(a) sagittal view

(b) sagittal view

Figure 7.2: Example segmentations from FreeSurfer in sagittal view. Left / 1. column: T1 image, all other images show T2 modality; orange/FS-mask; green/TP-pixels; yellow/FP-pixels; blue/FN-pixels.

7.1.2 Learned FreeSurfer Segmentation

In this section results are presented, which were acquired by learning the *FreeSurfer* segmentation. Therefore, the stacked segmentation architecture was trained with the T1 input images and the post-processed *FS* masks. Table 7.2 summarises the results of this experiment and the standard *FreeSurfer* pipeline compared to the T2-based manual ground truth mask.

Task	**Input**	**Labels**		**Evaluation**			
				Dice Score	Surface Distances in mm		
	Image	training	eval.	in %	d_H	d_{mean}	d_{median}
FreeSurfer	T1	-	GT	78.06 ± 3.99	4.66 ± 1.28	0.54 ± 0.10	0.47 ± 0.00
seg.	**T1**	**FS**	**GT**	$\mathbf{80.85 \pm 3.61}$	$\mathbf{5.09 \pm 2.06}$	$\mathbf{0.53 \pm 0.14}$	$\mathbf{0.45 \pm 0.08}$

Table 7.2: Comparison of the standard *FreeSurfer* results to the learned FS segmentation with *DCNNs*.

As shown by Table 7.2, the trained *DCNN* was not only able to reproduce a similar segmentation accuracy as achieved with the standard *FreeSurfer* pipeline, it also outperformed the *DSC* of the *FreeSurfer* results by almost 3 %. The mean and median surface distance, d_{mean} and d_{median} respectively, are slightly lower. Contrary, the Hausdorff distance is bigger indicating that the maximal distance (error) between the learned *FS* segmentation and the GT-labels got bigger compared to the standard *FreeSurfer* pipeline.

7.2 Learning-based Segmentation Experiments

Simple segmentation experiments comprise all deep learning based experiments that have been trained with our manual ground truth labels. All image modalities haven been used, namely T1-weighted and T2-weighted images separately as well as a combined experiment where both images were concatenated and used simultaneously.

Table 7.3 summarises the evaluated metrics of all learning based experiments and is therefore structured into two parts. At the very top, for easier comparison, again the scores achieved with the standard *FreeSurfer* pipeline are reported. Moreover, the achieved results by learning the *FS* labels are listed. In the lower part the new results of this section are presented.

Looking at the results of the learned **T1 input** segmentation (Table 7.3, lower part, 1. experiment), it can be seen that utilising T1-weighted input images together with the ground truth labels increased the segmentation accuracy by almost 5%, compared to training with FS-labels. Moreover, all surface distance measures dropped whereat especially the median distance decreased. This implies that only a few segmentation masks, out of all samples of all *CV* sets, exhibit a

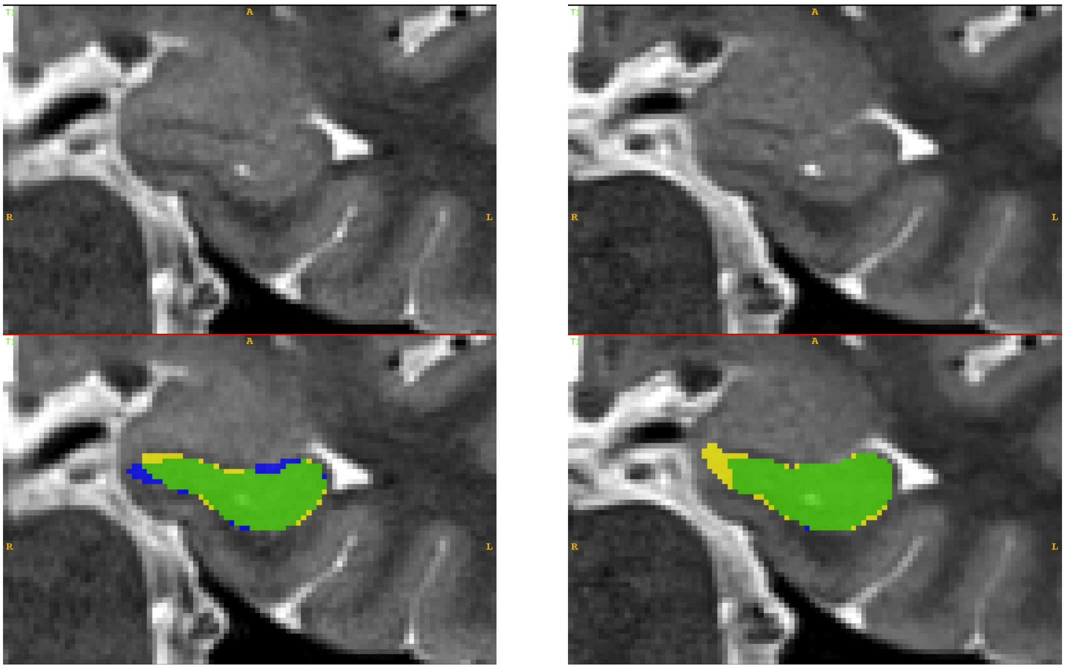
				Evaluation			
Task	**Input**	**Labels**		Dice Score	Surface Distances in mm		
	Image	training	eval.	in %	d_H	d_{mean}	d_{median}
FreeSurfer	T1	-	GT	78.06 ± 3.99	4.66 ± 1.28	0.54 ± 0.10	0.47 ± 0.00
seg.	T1	FS	GT	80.85 ± 3.61	5.09 ± 2.06	0.53 ± 0.14	0.45 ± 0.08
seg.	T1	GT	GT	85.62 ± 3.86	4.12 ± 1.93	0.36 ± 0.07	0.13 ± 0.21
seg.	**T2**	**GT**	**GT**	$\mathbf{91.91 \pm 0.87}$	$\mathbf{3.90 \pm 1.41}$	$\mathbf{0.19 \pm 0.03}$	$\mathbf{0.00 \pm 0.00}$
seg.	T1 + T2	GT	GT	91.64 ± 0.94	4.14 ± 1.85	0.20 ± 0.03	$\mathbf{0.00 \pm 0.00}$

Table 7.3: Learning-based segmentation results of different input image modalities trained with FS/man. GT labels. All results evaluated against manual ground truth (GT).

wrong segmentation boundary by at least one pixel for more than half of the boundary pixels.

The results achieved by segmenting T1-weighted images can be improved if **T2-weighted input** images are used instead for training the *DCNN*. In this setting, the overall best result is achieved with a *DSC* of 91.91 ± 0.87 %. Additionally, all surface distance measures showed the smallest values of all experiments. While $d_{max} = 3.90\,\text{mm}$ is the lowest overall Hausdorff distance, it still corresponds to a wrong *HC* outline by around five pixels. However, a possible explanation for this could be that the network performed a more consistent segmentation, for some slices or certain problematic regions, than given by the actual GT-labels.

An example for such a case, where the network provided a better segmentation, is shown in Fig. 7.3. In slice 6 (Fig. 7.3a) the GT labels properly cover the *HC* and the network segmentation also matches the GT. However, looking at slice 7 Fig. 7.3b retrospectively, there is an inconsistency in the GT label. This shift of opinion to the more accurate *HC* outline is motivated by two reasons. First, it could already be guessed based on the visible borders in the non-segmented image that the *HC* already continues at least to the medial tissue boundary. Second, in the region of the *HH* changes along the main axis of the hippocampal formation can only occur in a continuous manner. Therefore, it is not reasonable that the GT-mask gets as much smaller as shown in Fig. 7.3 within the next slice. On the contrary, it should continue along the medial boundary as the *DCNN* labelled correctly (see FP pixels at the medial border of the mask in Fig. 7.3b). Such errors in the manual GT are caused by general problems related to the delineation of the supero-medial region, as described in Section 5.1.3.3. Mainly, they are caused by neighbouring blood vessels, which cause signal distortions in their local vicinity, which can also be seen in Fig. 7.3b.

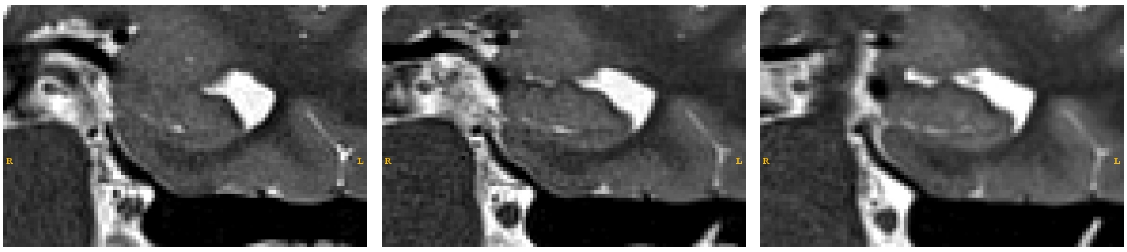


(a) coronal slice 6

(b) coronal slice 7

Figure 7.3: Visualisation of possible cause for high d_{max} values despite good overall segmentation accuracy. green ... TP-pixels; yellow ... FP-pixels; blue ... FN-pixels

Figure 7.4 shows an extreme case of this problem, where the completely supero-medial outline of the hippocampal formation is annihilated.



(a) slice 9

(b) slice 10

(c) slice 11

Figure 7.4: Severe signal distortion of the supero-medial HC border due to neighbouring vessel.

At last, the *DCNN* was also trained with both image modalities at the same time in order to roughly assess the impact of each input modality on the segmentation result. If compared to the simple T2 segmentation, this experiment showed only a marginally worse Dice score and Hausdorff distance while the mean and median distances did not change. This suggests, if both input images are provided the segmentation is still mainly based on the T2 input images. The slight decrease for *DSC* and d_{max} might be caused by inconsistencies within both image modalities.

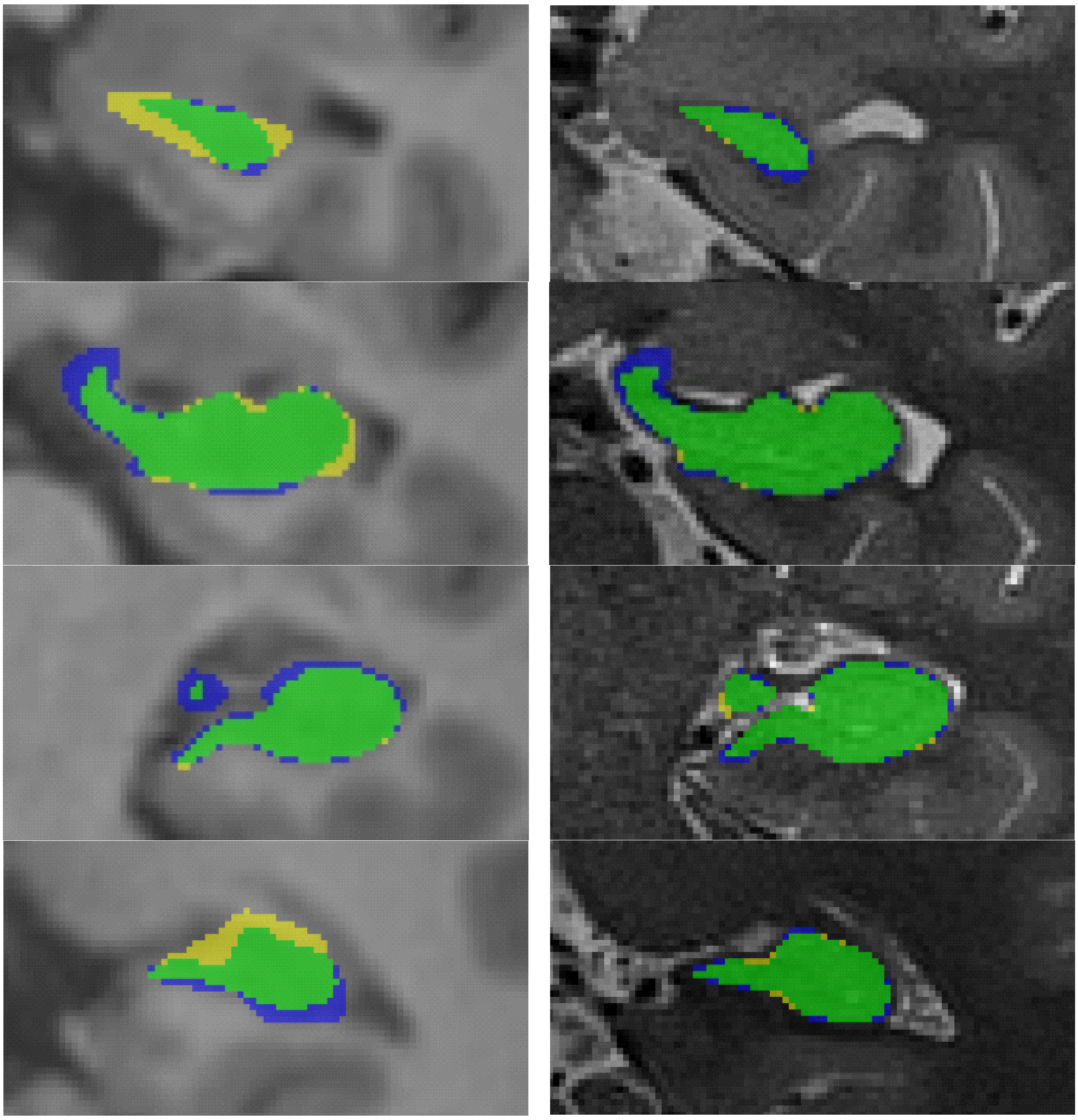
7.2.1 Qualitative T1 and T2 Segmentation Results

Figure 7.5 qualitatively shows the results of the simple *DCNN* based segmentation of either T1- or T2-weighted input images. For both modalities the same slices of the same subject are depicted. To give examples over the whole range of the hippocampal formation one slice of each block of 10 slices is shown. This means the upper most row visualises a slice from the slice block 1-10, the second row a slice from block containing slices 11-20 and so on.

Correspondingly to the quantitative results explained above, the automated segmentation masks created by our *DCNNs* show very high agreement with the manual GT (*TP*/green pixels) if T2 images are used as the input modality (Fig. 7.5b). Almost no pixels are labelled as *FP* (yellow) and only very view *FN* pixels (blue) can be seen.

Comparison of the generated mask created by segmenting the T1-weighted input images exhibit more discrepancies between the *DCNN* mask (TP + FP) and the manual GT (TP + FN), shown in Fig. 7.5a.

If the results of the T1- and T2-based segmentation are compared, one can see that in the centre of the hippocampal formation both segmentations are in good agreement not only with the GT but also among each other. However, especially the first slices of the *HH* and the last slices of the *HT* are wrongly labelled, visualised by the *FP* and *FN* pixels. An explanation for this could be that the very first and last slices of the *HC* are, due to the acquisition method, not present in every sample of the dataset. Moreover, as the hippocampal formation is in general better distinguishable on T2-weighted images, the available data covering these regions might be sufficient to train the T2-based segmentation, while for the T1-based *DCNN* it seems to be not the case.



(a) segmentation of T1 images

(b) segmentation of T2 images

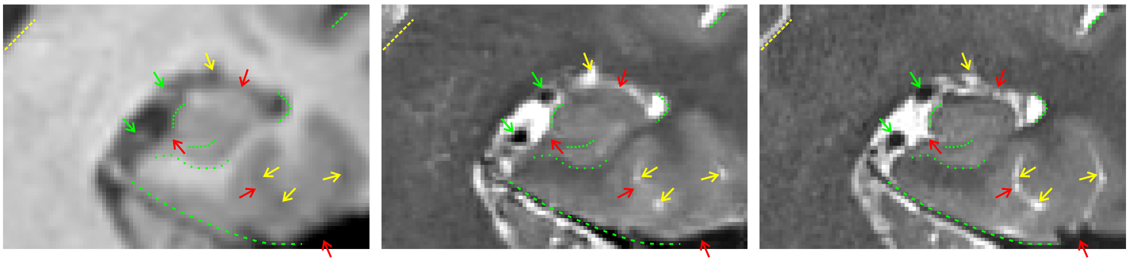
Figure 7.5: Exemplaric visualisation of a simple DCNN segmentation on T1 and T2 images, respectively. Labels: green... TP-pixels; yellow... FP-pixels; blue... FN-pixels.

7.3 Auxiliary Image Synthesis

This section presents the results for our auxiliary image synthesis task, which was described in Section 6.3.3.

In this work, the image synthesis task was not performed to generate perfectly mimicking T2-weighted images which can be further used for diagnosis or similar where an completely accurate T2-representation of the underlying T1-weighted images is crucial. This experiment was rather performed as an auxiliary task to incorporate the T2 images into the training process of a T1 based segmentation experiment. For this reason, and because optimising Wasserstein GANs (WGANs) is a non-trivial task extensive parameter tuning of this model was reduced to acquiring proper rather than perfect representations of the T2 contrast.

The image synthesis model achieved an average peak signal-to-noise ratio (PSNR) of 21.69 ± 0.66 dB over the whole *CV* set. However, in our case *PSNR* does not quite represent the perceived image quality. Therefore, we mainly performed visual inspection, as shown in Fig. 7.6.



(a) T1 **(b)** T2 synthetic **(c)** T2

Figure 7.6: Example T2 image synthesis result. Legend for annotations: green ... good feature replication; yellow ... average/not complete feature replication; red ... poor/missing feature replication.

By utilising a *WGAN* it was possible to create synthetic T2 representations of the underlying T1 data. When comparing the synthetic T2 image (Fig. 7.6b) with its ground truth image (Fig. 7.6c) we can see that especially the borders of the brain tissue got recovered correctly (dashed green lines). Moreover, the blood vessels (black “circles” in the T2 images) medial to *HC* were reconstructed correctly as indicated by the green arrows. This is quite impressive considering that blood flow related signal distortions are different for T1- and T2-weighted image acquisition. Looking at the synthetic representations of *CSF* (bright at T2 contrast) the model managed to recover such features to a moderate degree. Yellow arrows indicate *CSF* regions which have been moderately but not completely recovered. Red arrows point out regions which could not be reproduced with the current model. Looking particularly at the *HC*, the inferior (lower) outlines (dotted green lines) were mapped accurately, while the superior (upper) region exhibits errors (red arrow).

7.4 T2-enhanced Training of Deep NNs

As the simple segmentation experiments with *DCNN* showed good results, especially if the T2-weighted images were utilised, two experiments were conducted that incorporated the T2 images into the training process. The detailed setup of these experiments was described in Section 6.3.3. In the lower part of Table 7.4 results of both experiments are shown.

The first experiment, named *seg. with pre-trained T2seg*, describes a segmentation experiment which was performed on T1-weighted images by utilising a pre-trained segmentation model. In more detail, all parameters of the already optimised T2 segmentation model (from Section 7.2) were loaded to initialise the same architecture, which was subsequently fine-tuned with the T1-weighted magnetic resonance (MR) images.

The second experiment, named *seg. with T2 synthesis & pre-trained T2seg*, utilised the auxiliary image synthesis task in combination with a *DCNN* for the *HC* segmentation. Therefore, the parameters of the generator of the previously trained *WGAN* (Section 7.3) and the parameters of a pre-trained T2 image segmentation were loaded.

Additionally, as both experiments were eventually performed with T1-weighted input images, the simple segmentation results for the T1 modality are shown again in the top part of the table.

Task	**Input Image**	**Labels**	**Evaluation**			
			Dice Score in %	Surface Distances in mm		
				d_H	d_{mean}	d_{median}
segmentation	T1	GT	85.62 ± 3.86	4.12 ± 1.93	0.362 ± 0.07	$\mathbf{0.13 \pm 0.21}$
seg. with pre-trained T2seg	T1	GT	$\mathbf{85.90 \pm 3.85}$	$\mathbf{3.94 \pm 1.28}$	$\mathbf{0.357 \pm 0.11}$	0.15 ± 0.22
seg. with T2 synthesis & pre-trained T2seg	T1	GT	85.73 ± 4.23	4.14 ± 1.93	0.358 ± 0.13	0.17 ± 0.23

Table 7.4: *DCNN* segmentation results of T2-incorporation experiments.

As shown in Table 7.4, both experiments using T2-enhanced training yield similar results as the simple *DCNN* T1 segmentation. Regarding the *DSCs*, both experiments got inconsiderable better where the first experiment achieved the best accuracy among them with 85.90 ± 3.85 %. The Hausdorff distances also marginally improved for the first experiment, while d_{max} more or less stayed the same for the segmentation with image synthesis compared to the simple T1 experiment. For both experiments the median distances minimally increased when compared to the standard T1 segmentation, while the mean distances negligibly decreased. Generally for these two measures, the mean values indicated almost no change while the variances marginally increased.

8 Conclusion

In this thesis a deep learning (DL) -based approach for segmentation of the hippocampal formation was proposed to answer two main questions of interest:

1. are *DL*-based methods able to outperform the hippocampus (HC) segmentation from FreeSurfer (FS) ?
2. what is the impact of a high-resolution T2-weighted dataset on this segmentation task?

Therefore, a unique dataset containing corresponding pairs of high-resolution T1-weighted and ultra high-resolution T2-weighted 3T magnetic resonance (MR) images has been acquired, manually annotated and pre-processed. Next, baseline segmentations were calculated by applying *FreeSurfer*'s brain segmentation pipeline together with its *hippocampal-subfields-protocol* to our T1-weighted images. Additionally, several *DL*-based experiments were conducted, including segmentations of the T1- or T2-weighted dataset images with deep convolutional neural networks (DCNNs). In one experiment both image modalities were used in combination to train the segmentation task.

The *FreeSurfer* segmentations, which are currently used in clinical research, showed a segmentation accuracy, reported as Dice similarity coefficient (DSC), of $78.06 \pm 3.99\,\%$ when compared to our manual ground truth (GT). To evaluate the extend to which this segmentation can be reproduced with *DCNN*, an initial learning-based experiment was conducted. This experiment showed, when the *FS*-labels are used as target labels while training our *DCNN*, the *FreeSurfer* results can not only be replicated but also improved to a $DSC = 80.85 \pm 3.61\,\%$.

The application of the same model architecture, however, trained with our GT-labels yields a Dice score of $85.62 \pm 3.86\,\%$, which is again an increase of almost $5\,\%$. Utilising the T2-weighted images instead of our T1-weighted scans, again trained with our manual GT-labels, yields the overall best accuracy with a $DSC = \mathbf{91.91 \pm 0.87}\,\%$. If both image modalities, simultaneously, are provided as the network input, the results are almost identical to the findings achieved with just T2 input images, namely $DSC =$

$\mathbf{91.64 \pm 0.94}\,\%$. This suggests that all information required by our models, to perform the segmentation task in the best possible way, is already contained in the T2-weighted images and that no complementary information can be gained when using T1 and T2 images together.

T2-enhanced Training Experiments

Unfortunately, such ultra high-resolution T2-weighted images are not feasible even in clinical research. Therefore, we have conducted two additional experiments, where the T2-weighted images are not explicitly used as an input, but rather are utilised only during training of our *DCNNs* to enhance the T1-based segmentation.

For the first approach the same network architecture was utilised, however, instead of training only once with either of the image modalities, a pre-trained model was used. In particular, the parameters of the pre-trained *DCNN* from the "simple" T2-based segmentation were loaded to initialise the model and subsequently optimise with the T1-weighted *MR* images. This setup yields similar results as the "simple" T1-based segmentation (85.62±3.86 %), with an insignificant improvement to a $DSC = 85.90{\pm}3.85\,\%$. A possible explanation for the similar outcome could be that the loaded, initially T2-based, weights got outweighed by the T1-based fine-tuning of the model.

For the second experiment, we have therefore implicitly enforced the usage of the T2-weighted images not only by making use of T2-based parameters but also by combining the segmentation task with an auxiliary T2 image synthesis task. Therefore, a Wasserstein GAN (WGAN) was trained prior the actual segmentation to create synthetic T2 images based on their corresponding T1-weighted counterparts. The parameters of the pre-trained generator model of this *WGAN* were then loaded together with the parameters of the segmentation architecture, again pre-trained with the T2 inputs. The actual combined segmentation task was then trained with this initialised model and the T1-weighted input images. The segmentation results achieved with this experiment are in between the "simple" T1-based *DCNN* and the T1-based segmentation with the additional T2-based weight initialisation. In detail, a Dice score of $DSC = 85.73 \pm 4.23\,\%$ was achieved. While the generated synthetic T2 images express a proper T2-like contrast and is quite capable of recovering anatomical features like blood vessels or some cortical borders, it still has some difficulties (see Fig. 7.6). Especially recovering cerebrospinal fluid (CSF) and sometimes its neighbouring tissue is error-prone, which could be explained thereby that in T1-weighted images *CSF* is not shown distinctively but rather expressed only in form of background noise. However, as *CSF* is neighbouring a considerable portion of the *HC* this could be a reason for the T1-like outcome. Moreover, remaining misalignment between the corresponding pairs of T1- and T2-weighted images in the dataset could explain missing or shifted representations of some anatomical features.

To conclude all conducted experiments, we have shown that independently of the applied deep learning based segmentation, the results achieved with *FreeSurfer* can not only be improved by means of the segmentation accuracy but also with respect to the processing time. While even the application of the parallelised *FreeSurfer* pipeline takes up to 4.5 hours for a single subject, the segmentation with *DCNN* can be acquired within a couple of seconds.

The evaluation of the impact of the ultra high-resolution T2-weighed images can be answered based on our "simple" segmentation experiments, which perform just the segmentation task on either T1, T2 or T1+T2 inputs, as follows. While the *DL*-based segmentation of just the T1 images together with the T2-based manual *GT* annotations outperforms the *FreeSurfer* segmentation by around +7.5 %, the overall highest score can be achieved if instead also the T2-weighted images are used as the model input. This yields another increase by 6.3 % and an average $DSC = \mathbf{91.91 \pm 0.87}\,\%$.

Enhancing the *DCNN* based segmentation of T1-weighted images with our unique, corresponding T2-weighted image data did not show a significant improvement in the resulting segmentation accuracy. However, there are indications that this might be caused by the not perfectly aligned data pairs as well as the general complexity of training *WGAN*. Thus, future work is needed to address these problems.

8.1 Future Outlook

In order to aid Alzheimer's disease (AD) related research and in particular to establish a fast, consistent and accurate assessment of the hippocampal volume a lot of studies are currently ongoing. The presented work, was a proof of concept to evaluate the power of *DCNN* with respect to the task of automated *HC* segmentation and to assess the feasibility of acquiring an ultra high-resolution T2-weighted *MR* dataset.

As this was a first line of work in which lot of sub-tasks have been addressed, many useful insights were revealed. Not only our two main questions have been answered but also additional insights for future research could be identified, which can be categorised into three related parts: (i) Dataset improvement, (ii) Registration, and (iii) Deep learning-related improvements.

The first step to a better dataset could be improvements to the *MR* acquisition procedure to ensure complete coverage of the *HC* and to reduce motion artefacts. Essential for deep learning-based methods is not only the amount but especially the quality of the data. Thus, to improve the training data and hence the overall segmentation performance, a key will be not only to reduce ground truth noise but also to increase the amount of annotated hippocampi. Additionally, semi-supervised learning procedures could be utilised to make use of even unlabelled data.

As shown in this work, registration and especially the inter-modality registration of the T1- and T2-weighted image pairs is a crucial step and will therefore be an interesting topic of future research. Alternatively, different approaches for the image synthesis task could be utilised, which do not rely on previously registered images. This relates already to network architecture related improvements. Moreover, cycle generative adversarial networks (GANs) or types of recurrent neural networks (RNNs) could be an interesting point of research for the synthetic T2 generation task.

A List of Acronyms

α-synuclein	alpha-synuclein
τ-protein	tau protein
3D	three-dimensional
Aβ	amyloid-β
Aβ42	42-aminoacid form of amyloid-β
AD	Alzheimer's disease
ADAS	Alzheimer's Disease Assessment Scale
ADAS-Cog	ADAS-cognitive subscale
ADAS-CogIRT	ADAS-Cog using IRT
ADAS-Noncog	ADAS-non-cognitve subscale
ADI	Alzheimer's Disease International
ADL	activities of daily living
ADNI	Alzheimer's Disease Neuroimaging Initiative
ADRDA	Alzheimer's Disease and Related Disorders Association
ANN	artificial neural network
BADLS	Bristol ADL Scale
BNT	Boston Naming Test
CERAD	Consortium to Establish a Registry for Alzheimer's Disease
cGAN	conditional GAN
CN	cognitve normal
CNN	convolutional neural network
CNS	central nervous system
CPU	central processing unit
CSF	cerebrospinal fluid

CT	computed tomography
DCNN	deep convolutional neural network
DLB	dementia with Lewy bodies
DSC	Dice similarity coefficient
DSM-IV-TR	Diagnostic and Statistical Manual of Mental Disorders, fourth edition
EADC	European Alzheimer's Disease Consortium
ErC	entorhinal cortex
FID	free induction decay
FLAIR	fluid-attenuated inversion recovery
fMRI	functional MRI
FoV	field of view
FSE	fast spin echo
GAN	generative adversarial network
GE	gradient echo
GM	grey matter
GRE	gradient recalled echo
HarP	Harmonized Protocol
HC	hippocampus
HF	high frequency
IRT	item response theory
LSTM	long-short-term memory
MAS	multi-atlas segmentation
MCI	mild cognitive impairment
MMSE	Mini Mental State Examination
MoCA	Montreal Cognitive Assessment
MPRAGE	magnetisation-prepared rapid gradient-echo
MR	magnetic resonance
MRI	magnetic resonance imaging
MS	multiple sclerosis
MTA	medial temporal lobe atrophy
MTL	medial temporal lobe
NFL	neurofilament light
NINCDS	National Institute of Neurological Disorders and Stroke
P-tau	phosphorylated tau
PD	proton density
PET	positron emission tomography

PIB	Pittsburgh compound B
PrC	perirhinal cortex
RF	radio frequency
SE	spin echo
SNR	signal-to-noise ratio
SPECT	single photon emission computed tomography
T-tau	total tau
T1w	T1-weighted
T2w	T2-weighted
TL	temporal lobe
TSE	turbo spin echo
USA	United States of America
VaD	vascular dementia
WM	white matter

Bibliography

[1] N. Abramson, D. Braverman, and G. Sebestyen. Pattern recognition and machine learning. IEEE Transactions on Information Theory, 9(4):257–261, 1963. ISSN 15579654. doi: 10.1109/TIT.1963.1057854. (page 35)

[2] Alzheimer's Association. 2020 Alzheimer's disease facts and figures. Alzheimer's & Dementia, 16(3):391–460, mar 2020. doi: 10.1002/alz.12068. (page 1)

[3] Alzheimer's Association. 2020 Alzheimer's Disease Facts and Figures: On the Front Lines: Primary Care Physicians and Alzheimer's Care in America. Technical report, Alzheimer's Association, mar 2020. (page 2)

[4] Syed Muhammad Anwar, Muhammad Majid, Adnan Qayyum, Muhammad Awais, Majdi Alnowami, and Muhammad Khurram Khan. Medical Image Analysis using Convolutional Neural Networks: A Review, nov 2018. ISSN 1573689X. (page 14)

[5] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein GAN, jan 2017. (page 15, 38, 39, 40)

[6] Brian B. Avants, Nicholas J. Tustison, Gang Song, Philip A. Cook, Arno Klein, and James C. Gee. A reproducible evaluation of ANTs similarity metric performance in brain image registration. NeuroImage, 54(3):2033–2044, feb 2011. ISSN 10538119. doi: 10.1016/j.neuroimage.2010.09.025. (page 28)

[7] Josephine Barnes, Jonathan W. Bartlett, Laura A. van de Pol, Clement T. Loy, Rachael I. Scahill, Chris Frost, Paul Thompson, and Nick C. Fox. A meta-analysis of hippocampal atrophy rates in Alzheimer's disease, nov 2009. ISSN 01974580. (page 6)

[8] D. Berron, P. Vieweg, A. Hochkeppler, J. B. Pluta, S. L. Ding, A. Maass, A. Luther, L. Xie, S. R. Das, D. A. Wolk, T. Wolbers, P. A. Yushkevich, E. Düzel, and L. E.M. Wisse. A protocol for manual segmentation of medial temporal lobe subregions in 7 Tesla MRI. NeuroImage: Clinical, 15:466–482, may 2017. ISSN 22131582. doi: 10.1016/j.nicl.2017.05.022. (page 10, 11, 56, 57, 58, 59)

[9] Marina Boccardi, Rossana Ganzola, Martina Bocchetta, Michela Pievani, Alberto Redolfi, George Bartzokis, Richard Camicioli, John G. Csernansky, Mony J. De Leon, Leyla Detoledo-Morrell, Ronald J. Killiany, Stéphane Lehéricy, Johannes Pantel, Jens C. Pruessner, H. Soininen, Craig Watson, Simon Duchesne, Clifford R. Jack, and Giovanni B. Frisoni. Survey of protocols for the manual segmentation of the hippocampus: Preparatory steps towards a joint EADC-ADNI harmonized protocol. Journal of Alzheimer's Disease, 26(SUPPL. 3):61–75, jan 2011. ISSN 18758908. doi: 10.3233/JAD-2011-0004. (page 12)

[10] Marina Boccardi, Martina Bocchetta, Liana G. Apostolova, Josephine Barnes, George Bartzokis, Gabriele Corbetta, Charles Decarli, Leyla Detoledo-Morrell, Michael Firbank, Rossana Ganzola, Lotte Gerritsen, Wouter Henneman, Ronald J. Killiany, Nikolai Malykhin, Patrizio Pasqualetti, Jens C. Pruessner, Alberto Redolfi, Nicolas Robitaille, Hilkka Soininen, Daniele Tolomeo, Lei Wang, Craig Watson, Henrike Wolf, Henri Duvernoy, Simon Duchesne, Clifford R. Jack, and Giovanni B. Frisoni. Delphi definition of the EADC-ADNI harmonized protocol for hippocampal segmentation on magnetic resonance. Alzheimer's and Dementia, 11(2):126–138, feb 2015. ISSN 15525279. doi: 10.1016/j.jalz.2014.02.009. (page 12)

[11] Marina Boccardi, Martina Bocchetta, Rossana Ganzola, Nicolas Robitaille, Alberto Redolfi, Simon Duchesne, Clifford R. Jack, Giovanni B. Frisoni, George Bartzokis, John G. Csernansky, Mony J. De Leon, Leyla Detoledo-Morrell, Ronald J. Killiany, Stephane Lehericy, Nikolai Malykhin, Johannes Pantel, Jens C. Pruessner, Hilkka Soininen, and Craig Watson. Operationalizing protocol differences for EADC-ADNI manual hippocampal segmentation. Alzheimer's and Dementia, 11(2):184–194, feb 2015. ISSN 15525279. doi: 10.1016/j.jalz.2013.03.001. (page)

[12] Martina Bocchetta, Marina Boccardi, Rossana Ganzola, Liana G. Apostolova, Gregory Preboske, Dominik Wolf, Clarissa Ferrari, Patrizio Pasqualetti, Nicolas Robitaille, Simon Duchesne, Clifford R. Jack, Giovanni B. Frisoni, George Bartzokis, Charles Decarli, Leyla Detoledo-Morrell, Andreas Fellgiebel, Michael Firbank, Lotte Gerritsen, Wouter Henneman, Ronald J. Killiany, Nikolai Malykhin, Jens C. Pruessner, Hilkka Soininen, and Lei Wang. Harmonized benchmark labels of the hippocampus on magnetic resonance: The EADC-ADNI project. Alzheimer's and Dementia, 11(2): 151–160.e5, feb 2015. ISSN 15525279. doi: 10.1016/j.jalz.2013.12.019. (page 12)

[13] H. Braak and E. Braak. Neuropathological stageing of Alzheimer-related changes, sep 1991. ISSN 00016322. (page 2)

[14] H. Braak and E. Braak. Frequency of stages of Alzheimer-related lesions in different age categories. Neurobiology of Aging, 18(4):351–357, jul 1997. ISSN 01974580. doi: 10.1016/S0197-4580(97)00056-0. (page 2)

[15] Henry Brodaty, Lee Fay Low, Louisa Gibson, and Kim Burns. What is the best dementia screening instrument for general practitioners to use?, may 2006. ISSN 10647481. (page 3)

[16] R. S. Bucks, D L Ashworth, G K Wilcock, and K Siegfried. Assessment of activities of daily living in dementia: Development of the Bristol Activities of Daily Living Scale. Age and Ageing, 25(2):113–120, 1996. ISSN 00020729. doi: 10.1093/ageing/25.2.113. (page 3)

[17] Mariano Cabezas, Arnau Oliver, Xavier Lladó, Jordi Freixenet, and Meritxell Bach Cuadra. A review of atlas-based segmentation for magnetic resonance brain images.

Computer Methods and Programs in Biomedicine, 104(3):e158–e177, dec 2011. ISSN 01692607. doi: 10.1016/j.cmpb.2011.07.015. (page 13)

[18] Diedre Carmo, Bruna Silva, Clarissa Yasuda, Letícia Rittner, and Roberto Lotufo. Extended 2D Consensus Hippocampus Segmentation. Medical Imaging with Deep Learning, 2019. (page 14)

[19] M. J. Chandler, L. H. Lacritz, L. S. Hynan, H. D. Barnard, G. Allen, M. Deschner, M. F. Weiner, and C. M. Cullum. A total score for the CERAD neuropsychological battery. Neurology, 65(1):102–106, jul 2005. ISSN 00283878. doi: 10.1212/01.wnl.0000167607.63000.38. (page 3)

[20] Yani Chen, Bibo Shi, Zhewei Wang, Tao Sun, Charles D. Smith, and Jundong Liu. Accurate and consistent hippocampus segmentation through convolutional LSTM and view ensemble. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), volume 10541 LNCS, pages 88–96. Springer Verlag, 2017. ISBN 9783319673882. doi: 10.1007/978-3-319-67389-9_11. (page 14)

[21] Gaël Chetelat and Jean Claude Baron. Early diagnosis of Alzheimer's disease: Contribution of structural neuroimaging, feb 2003. ISSN 10538119. (page 3, 4, 5)

[22] Joseph R Cockrell and Marshal F Folstein. Mini-mental state examination. Principles and practice of geriatric psychiatry, pages 140–141, 2002. (page 2)

[23] A Collignon, F Maes, D Delaere, D Vandermeulen, P Suetens, and G Marchal. Automated multi-modality image registration based on information theory. Information processing in medical imaging, 3(6):263–274, 1995. doi: 10.1007/11784012_16. (page 28)

[24] J. L. Cummings, M. Mega, K. Gray, S. Rosenberg-Thompson, D. A. Carusi, and J. Gornbein. The neuropsychiatric inventory: Comprehensive assessment of psychopathology in dementia. Neurology, 44(12):2308–2314, dec 1994. ISSN 1526632X. doi: 10.1212/wnl.44.12.2308. (page 3)

[25] Marshall A Dalton, Peter Zeidman, Daniel N Barry, Elaine Williams, and Eleanor A Maguire. Segmenting subregions of the human hippocampus on structural magnetic resonance image scans: An illustrated tutorial. Brain and neuroscience advances, 1:2398212817701448, jan 2017. ISSN 2398-2128. doi: 10.1177/2398212817701448. (page 11)

[26] Salman Uh Dar, Mahmut Yurt, Levent Karacan, Aykut Erdem, Erkut Erdem, and Tolga Cukur. Image Synthesis in Multi-Contrast MRI With Conditional Generative Adversarial Networks. IEEE transactions on medical imaging, 38(10):2375–2388, oct 2019. ISSN 1558254X. doi: 10.1109/TMI.2019.2901750. (page 15)

[27] A. Delacourte, J. P. David, N. Sergeant, L. Buée, A. Wattez, P. Vermersch, F. Ghozali, C. Fallet-Bianco, F. Pasquier, F. Lebert, H. Petit, and C. Di Menza. The biochemical pathway of neurofibrillary degeneration in aging and Alzheimer's disease. Neurology, 52(6):1158–1165, apr 1999. ISSN 00283878. doi: 10.1212/wnl.52.6.1158. (page 2)

[28] K H Dening, A Burns, M B Sandilyan, T Dening, H Hayo, C Knifton, C Baker, J Leeks, E Moniz-Cook, and A Beatty. Evidence-Based Practice in Dementia for Nurses and Nursing Students. Jessica Kingsley Publishers, 1 edition, 2019. ISBN 9781784507978. (page 1, 3)

[29] Vanderson Dill, Alexandre Rosa Franco, and Márcio Sarroglia Pinho. Automated Methods for Hippocampus Segmentation: the Evolution and a Review of the State of the Art, oct 2015. ISSN 15392791. (page 13)

[30] Song Lin Ding and Gary W. Van Hoesen. Borders, extent, and topography of human perirhinal cortex as revealed using multiple modern neuroanatomical and pathological markers. Human Brain Mapping, 31(9):1359–1379, sep 2010. ISSN 10659471. doi: 10.1002/hbm.20940. (page 11, 56)

[31] Song Lin Ding and Gary W. Van Hoesen. Organization and detailed parcellation of human hippocampal head and body regions based on a combined analysis of Cyto- and chemoarchitecture. Journal of Comparative Neurology, 523(15):2233–2253, oct 2015. ISSN 10969861. doi: 10.1002/cne.23786. (page 11, 56)

[32] Alexey Dosovitskiy, Philipp Fischer, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with exemplar convolutional neural networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 38(9):1734–1747, 2016. ISSN 01628828. doi: 10.1109/TPAMI.2015.2496141. (page 14)

[33] A. T. Du, N. Schuff, D. Amend, M. P. Laakso, Y. Y. Hsu, W. J. Jagust, K. Yaffe, J. H. Kramer, B. Reed, D. Norman, H. C. Chui, and M. W. Weiner. Magnetic resonance imaging of the entorhinal cortex and hippocampus in mild cognitive impairment and Alzheimer's disease. Journal of Neurology Neurosurgery and Psychiatry, 71(4): 441–447, oct 2001. ISSN 00223050. doi: 10.1136/jnnp.71.4.441. (page 9, 10)

[34] Bruno Dubois, Howard H. Feldman, Claudia Jacova, Jeffrey L. Cummings, Steven T. DeKosky, Pascale Barberger-Gateau, André Delacourte, Giovanni Frisoni, Nick C. Fox, Douglas Galasko, Serge Gauthier, Harald Hampel, Gregory A. Jicha, Kenichi Meguro, John O'Brien, Florence Pasquier, Philippe Robert, Martin Rossor, Steven Salloway, Marie Sarazin, Leonardo C. de Souza, Yaakov Stern, Pieter J. Visser, and Philip Scheltens. Revising the definition of Alzheimer's disease: A new lexicon, nov 2010. ISSN 14744422. (page 4, 5)

[35] C. Fischbach-Boulanger, A. Fitsiori, V. Noblet, S. Baloglu, H. Oesterle, S. Draghici, N. Philippi, E. Duron, O. Hanon, J. L. Dietemann, F. Blanc, and S. Kremer. T1- or T2-weighted magnetic resonance imaging: what is the best choice to evaluate atrophy of the hippocampus? European Journal of Neurology, 25(5):775–781, may 2018. ISSN 14681331. doi: 10.1111/ene.13601. (page 10, 12, 17)

[36] Bruce Fischl. FreeSurfer, aug 2012. ISSN 10538119. (page 13)

[37] Lukas Folle, Sulaiman Vesal, Nishant Ravikumar, and Andreas Maier. Dilated deeply supervised networks for hippocampus segmentation in MRI, mar 2019. (page 14)

[38] Giovanni B. Frisoni, Nick C. Fox, Clifford R. Jack, Philip Scheltens, and Paul M. Thompson. The clinical use of structural MRI in Alzheimer disease, feb 2010. ISSN 17594758. (page 5, 6, 7)

[39] Sandra González-Villà, Arnau Oliver, Sergi Valverde, Liping Wang, Reyer Zwiggelaar, and Xavier Lladó. A review on brain structures segmentation in magnetic resonance imaging, oct 2016. ISSN 18732860. (page 13)

[40] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Deep learning. MIT press, London, 2016. ISBN 0262337371. (page 33, 36, 37)

[41] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. Advances in Neural Information Processing Systems, 3(January):2672–2680, 2014. ISSN 10495258. (page 15, 37, 38, 39)

[42] Maged Goubran, Edward Ntiri, Hassan Akhavein, Melissa Holmes, Sean Nestor, Joel Ramirez, Sabrina Adamo, Miracle Ozzoude, Christopher Scott, Fuqiang Gao, Anne Martel, Walter Swardfager, Mario Masellis, Richard Swartz, Bradley MacIntosh, and Sandra E. Black. Hippocampal segmentation for brains with extensive atrophy using three-dimensional convolutional neural networks. In ISMRM 27th Annual Meeting, pages 1–18, oct 2019. doi: 10.1002/hbm.24811. (page 14)

[43] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein GANs. In Advances in Neural Information Processing Systems, volume 2017-Decem, pages 5768–5778, 2017. (page 15, 40, 70)

[44] Jeffrey L. Gunter, Bret J. Borowski, Kaely Thostenson, Arvin Arani, Robert I. Reid, David M. Cash, David L. Thomas, Hui Zhang, Charles S. DeCarli, Nick C. Fox, Paul M. Thompson, Duygu Tosun, Michael Weiner, and Clifford R. Jack. [P4-242]: ADNI-3 MRI ACQUISITIONS. Alzheimer's & Dementia, 13(7S_Part_28): P1368–P1369, jul 2017. ISSN 1552-5260. doi: 10.1016/j.jalz.2017.06.2110. (page 10)

[45] Harald Hampel, Katharina Bürger, Stefan J. Teipel, Arun L.W. Bokde, Henrik Zetterberg, and Kaj Blennow. Core candidate neurochemical and imaging biomarkers

of Alzheimer's disease. Alzheimer's and Dementia, 4(1):38–48, jan 2008. ISSN 15525260. doi: 10.1016/j.jalz.2007.08.006. (page 4)

[46] Xiao Han. MR-based synthetic CT generation using a deep convolutional neural network method:. Medical Physics, 44(4):1408–1419, apr 2017. ISSN 00942405. doi: 10.1002/mp.12155. (page 15)

[47] Lorna Harper, Frederik Barkhof, Nick C. Fox, and Jonathan M. Schott. Using visual rating to diagnose dementia: A critical evaluation of MRI atrophy scales, apr 2015. ISSN 1468330X. (page 6)

[48] Yuta Hiasa, Yoshito Otake, Masaki Takao, Takumi Matsuoka, Kazuma Takashima, Aaron Carass, Jerry L. Prince, Nobuhiko Sugano, and Yoshinobu Sato. Cross-modality image synthesis from unpaired data using cyclegan: Effects of gradient consistency loss and training data size. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), volume 11037 LNCS, pages 31–41. Springer Verlag, sep 2018. ISBN 9783030005351. doi: 10.1007/978-3-030-00536-8_4. (page 15)

[49] Kurt Hornik. Approximation capabilities of multilayer feedforward networks. Neural Networks, 4(2):251–257, jan 1991. ISSN 08936080. doi: 10.1016/0893-6080(91) 90009-T. (page 33)

[50] Yuan Yu Hsu, Norbert Schuff, An Tao Du, Kevin Mark, Xiaoping Zhu, Dawn Hardin, and Michael W. Weiner. Comparison of automated and manual MRI volumetry of hippocampus in normal aging and dementia. Journal of Magnetic Resonance Imaging, 16(3):305–310, sep 2002. ISSN 10531807. doi: 10.1002/jmri.10163. (page 10)

[51] Juan Eugenio Iglesias and Mert R. Sabuncu. Multi-atlas segmentation of biomedical images: A survey. Medical Image Analysis, 24(1):205–219, aug 2015. ISSN 13618423. doi: 10.1016/j.media.2015.06.012. (page 13)

[52] Juan Eugenio Iglesias, Jean C. Augustinack, Khoa Nguyen, Christopher M. Player, Allison Player, Michelle Wright, Nicole Roy, Matthew P. Frosch, Ann C. McKee, Lawrence L. Wald, Bruce Fischl, Koen Van Leemput, and for the Alzheimer's Disease Neuroimaging Alzheimer's Disease Neuroimaging Initiative. A computational atlas of the hippocampal formation using ex vivo, ultra-high resolution MRI: Application to adaptive segmentation of in vivo MRI. NeuroImage, 115:117–137, jul 2015. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2015.04.042. (page 10, 13, 63)

[53] Phillip Isola, Jun Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, volume 2017-Janua, pages 5967–5976, 2017. ISBN 9781538604571. doi: 10.1109/CVPR.2017.632. (page 15, 40, 41)

[54] Clifford R. Jack, Matt A. Bernstein, Nick C. Fox, Paul Thompson, Gene Alexander, Danielle Harvey, Bret Borowski, Paula J. Britson, Jennifer L. Whitwell, Chadwick Ward, Anders M. Dale, Joel P. Felmlee, Jeffrey L. Gunter, Derek L.G. Hill, Ron Killiany, Norbert Schuff, Sabrina Fox-Bosetti, Chen Lin, Colin Studholme, Charles S. DeCarli, Gunnar Krueger, Heidi A. Ward, Gregory J. Metzger, Katherine T. Scott, Richard Mallozzi, Daniel Blezek, Joshua Levy, Josef P. Debbins, Adam S. Fleisher, Marilyn Albert, Robert Green, George Bartzokis, Gary Glover, John Mugler, and Michael W. Weiner. The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods, apr 2008. ISSN 10531807. (page 9, 10)

[55] Keith A. Johnson, Nick C. Fox, Reisa A. Sperling, and William E. Klunk. Brain imaging in Alzheimer disease. Cold Spring Harbor Perspectives in Medicine, 2 (a006213):24, apr 2012. ISSN 21571422. doi: 10.1101/cshperspect.a006213. (page 5, 6)

[56] Jason Karlawish, Clifford R. Jack, Walter A. Rocca, Heather M. Snyder, and Maria C. Carrillo. Alzheimer's disease: The next frontier - Special Report 2017. Alzheimer's and Dementia, 13(4):374–380, apr 2017. ISSN 15525279. doi: 10.1016/j.jalz.2017.02.006. (page 4)

[57] Justin Ker, Lipo Wang, Jai Rao, and Tchoyoson Lim. Deep Learning Applications in Medical Image Analysis. IEEE Access, 6:9375–9379, dec 2017. ISSN 21693536. doi: 10.1109/ACCESS.2017.2788044. (page 14)

[58] Diederik P. Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. In 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings. International Conference on Learning Representations, ICLR, dec 2015. (page 35, 39, 70)

[59] Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. Self-Normalizing Neural Networks. In Advances in Neural Information Processing Systems 30 (NIPS 2017), pages 971–980, 2017. (page 32, 66)

[60] Juha Koikkalainen, Hanneke Rhodius-Meester, Antti Tolonen, Frederik Barkhof, Betty Tijms, Afina W. Lemstra, Tong Tong, Ricardo Guerrero, Andreas Schuh, Christian Ledig, Daniel Rueckert, Hilkka Soininen, Anne M. Remes, Gunhild Waldemar, Steen Hasselbalch, Patrizia Mecocci, Wiesje Van Der Flier, and Jyrki Lötjönen. Differential diagnosis of neurodegenerative diseases using structural MRI data. NeuroImage: Clinical, 11:435–449, jan 2016. ISSN 22131582. doi: 10.1016/j.nicl.2016.02.019. (page 10)

[61] John R. Koza, Forrest H. Bennett, David Andre, and Martin A. Keane. Automated Design of Both the Topology and Sizing of Analog Electrical Circuits Using Genetic Programming. In Artificial Intelligence in Design '96, pages 151–170. Springer Netherlands, 1996. doi: 10.1007/978-94-009-0279-4_9. (page 29)

[62] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet classification with deep convolutional neural networks. Communications of the ACM, 60(6):84–90, 2017. ISSN 15577317. doi: 10.1145/3065386. (page 14)

[63] Matthew Lai. Deep Learning for Medical Image Segmentation, may 2015. (page 14)

[64] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11):2278–2323, 1998. ISSN 00189219. doi: 10.1109/5.726791. (page 35)

[65] Bennett P. Leifer. Early Diagnosis of Alzheimer's Disease: Clinical and Economic Benefits. Journal of the American Geriatrics Society, 51(5s2):S281–S288, may 2003. ISSN 0002-8614. doi: 10.1046/j.1532-5415.5153.x. (page 3)

[66] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. Rectifier nonlinearities improve neural network acoustic models. In in ICML Workshop on Deep Learning for Audio, Speech and Language Processing, 2013. (page 32)

[67] Frederik Maes, André Collignon, Dirk Vandermeulen, Guy Marchal, and Paul Suetens. Multimodality image registration by maximization of mutual information. IEEE Transactions on Medical Imaging, 16(2):187–198, 1997. ISSN 02780062. doi: 10.1109/42.563664. (page 28)

[68] Mehdi Mirza and Simon Osindero. Conditional Generative Adversarial Nets, nov 2014. (page 15, 40)

[69] Elizabeth A Moore, Martin J Graves, Martin R Prince, and Donald W McRobbie. MRI from Picture to Proton. Cambridge University Press, 2 edition, 2006. ISBN 0511348495. (page 21, 51)

[70] Matthew Moore, Yifan Hu, Sarah Woo, Dylan O'Hearn, Alexandru D Iordan, Sanda Dolcos, and Florin Dolcos. A comprehensive protocol for manual segmentation of the medial temporal lobe structures., jul 2014. ISSN 1940-087X. (page 11)

[71] John P. Mugler and James R. Brookeman. Three-dimensional magnetization-prepared rapid gradient-echo imaging (3D MP RAGE). Magnetic Resonance in Medicine, 15 (1):152–157, jul 1990. ISSN 15222594. doi: 10.1002/mrm.1910150117. (page 9)

[72] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve Restricted Boltzmann machines. In ICML 2010 - Proceedings, 27th International Conference on Machine Learning, pages 807–814, 2010. ISBN 9781605589077. (page 31)

[73] Ziad S. Nasreddine, Natalie A. Phillips, Valérie Bédirian, Simon Charbonneau, Victor Whitehead, Isabelle Collin, Jeffrey L. Cummings, and Howard Chertkow. The Montreal Cognitive Assessment, MoCA: A brief screening tool for mild cognitive impairment. Journal of the American Geriatrics Society, 53(4):695–699, apr 2005. ISSN 15325415. doi: 10.1111/j.1532-5415.2005.53221.x. (page 2)

[74] Dong Nie, Roger Trullo, Jun Lian, Li Wang, Caroline Petitjean, Su Ruan, Qian Wang, and Dinggang Shen. Medical Image Synthesis with Deep Convolutional Adversarial Networks. IEEE Transactions on Biomedical Engineering, 65(12):2720–2730, dec 2018. ISSN 15582531. doi: 10.1109/TBME.2018.2814538. (page 15)

[75] Bob Olsson, Ronald Lautner, Ulf Andreasson, Annika Öhrfelt, Erik Portelius, Maria Bjerke, Mikko Hölttä, Christoffer Rosén, Caroline Olsson, Gabrielle Strobel, Elizabeth Wu, Kelly Dakin, Max Petzold, Kaj Blennow, and Henrik Zetterberg. CSF and blood biomarkers for the diagnosis of Alzheimer's disease: a systematic review and meta-analysis. The Lancet Neurology, 15(7):673–684, jun 2016. ISSN 14744465. doi: 10.1016/S1474-4422(16)00070-3. (page 4)

[76] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context Encoders: Feature Learning by Inpainting. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, volume 2016-Decem, pages 2536–2544, 2016. ISBN 9781467388504. doi: 10.1109/CVPR.2016.278. (page 41)

[77] Martin Prince, Anders Wimo, Maëlenn Guerchet, Ali Gemma-Claire, Yu-Tzu Wu, and Matthew Prina. World Alzheimer Report 2015: The Global Impact of Dementia - An Analysis of Prevalence, Incidence, Cost and Trends. Technical report, Alzheimer's Disease International, 2015. (page 2)

[78] Martin Prince, Adelina Comas-Herrera, Martin Knapp, Guerchet Maëlenn, and Maria Karagiannidou. World Alzheimer Report 2016. Alzheimer's Disease International, pages 1–124, sep 2016. (page 2)

[79] V. P.Subramanyam Rallabandi, Ketki Tulpule, and Mahanandeeshwar Gattu. Automatic classification of cognitively normal, mild cognitive impairment and Alzheimer's disease using structural MRI analysis. Informatics in Medicine Unlocked, 18:7, jan 2020. ISSN 23529148. doi: 10.1016/j.imu.2020.100305. (page 6)

[80] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), volume 9351, pages 234–241. Springer Verlag, 2015. ISBN 9783319245737. doi: 10.1007/978-3-319-24574-4_28. (page 14, 65, 66)

[81] W. G. Rosen, R. C. Mohs, and K. L. Davis. A new rating scale for Alzheimer's disease. American Journal of Psychiatry, 141(11 (1356-1364),), 1984. ISSN 0002953X. doi: 10.1176/ajp.141.11.1356. (page 3)

[82] Günter Rote. Computing the minimum Hausdorff distance between two point sets on a line under translation. Information Processing Letters, 38(3):123–127, may 1991. ISSN 00200190. doi: 10.1016/0020-0190(91)90233-8. (page 74)

[83] Daniel Rueckert and Julia A Schnabel. Medical image registration. In Medical Image Registration, pages 131–154. Springer, Berlin, Heidelberg, 2001. ISBN 9781420042474. doi: 10.1051/epn:2000401. (page 28)

[84] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. Nature, 323(6088):533–536, 1986. ISSN 00280836. doi: 10.1038/323533a0. (page 34)

[85] Malarvizhi Babu Sandilyan and Tom Dening. Brain function, disease and dementia. Nursing standard (Royal College of Nursing (Great Britain) : 1987), 29(39):36–42, may 2015. ISSN 20479018. doi: 10.7748/ns.29.39.36.e9425. (page 2)

[86] Vasileios Sevetlidis, Mario Valerio Giuffrida, and Sotirios A. Tsaftaris. Whole image synthesis using a deep encoder-decoder network. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), volume 9968 LNCS, pages 127–137. Springer Verlag, oct 2016. ISBN 9783319466293. doi: 10.1007/978-3-319-46630-9_13. (page 15)

[87] Feng Shi, Bing Liu, Yuan Zhou, Chunshui Yu, and Tianzi Jiang. Hippocampal volume and asymmetry in mild cognitive impairment and Alzheimer's disease: Meta-analyses of MRI studies. Hippocampus, 19(11):1055–1064, nov 2009. ISSN 10509631. doi: 10.1002/hipo.20573. (page 6)

[88] Jeannine Skinner, Janessa O. Carvalho, Guy G. Potter, April Thames, Elizabeth Zelinski, Paul K. Crane, and Laura E. Gibbons. The Alzheimer's Disease Assessment Scale-Cognitive-Plus (ADAS-Cog-Plus): An expansion of the ADAS-Cog to improve responsiveness in MCI. Brain Imaging and Behavior, 6(4):489–501, dec 2012. ISSN 19317557. doi: 10.1007/s11682-012-9166-3. (page 3)

[89] J. D. Sluimer, H. Vrenken, M. A. Blankenstein, N. C. Fox, P. Scheltens, F. Barkhof, and W. M. Van Der Flier. Whole-brain atrophy rate in Alzheimer disease: Identifying fast progressors. Neurology, 70(19 PART 2):1836–1841, may 2008. ISSN 00283878. doi: 10.1212/01.wnl.0000311446.61861.e3. (page 6)

[90] Nanthia A. Suthana, Markus Donix, David R. Wozny, Adam Bazih, Michael Jones, Robin M. Heidemann, Robert Trampel, Arne D. Ekstrom, Maria Scharf, Barbara Knowlton, Robert Turner, and Susan Y. Bookheimer. High-resolution 7t fMRI of human hippocampal subfields during associative learning. Journal of Cognitive Neuroscience, 27(6):1194–1206, jun 2015. ISSN 15308898. doi: 10.1162/jocn_a_00772. (page 11)

[91] Stefan Teipel, Alexander Drzezga, Michel J. Grothe, Henryk Barthel, Gaël Chételat, Norbert Schuff, Pawel Skudlarski, Enrica Cavedo, Giovanni B. Frisoni, Wolfgang Hoffmann, Jochen René Thyrian, Chris Fox, Satoshi Minoshima, Osama Sabri,

and Andreas Fellgiebel. Multimodal imaging in Alzheimer's disease: Validity and usefulness for early detection, oct 2015. ISSN 14744465. (page 4, 5, 6)

[92] Jens M. Theysohn, O. Kraff, S. Maderwald, M. U. Schlamann, A. De Greiff, M. Forsting, S. C. Ladd, M. E. Ladd, and E. R. Gizewski. The human hippocampus at 7 T - In vivo MRI. Hippocampus, 19(1):1–7, jan 2009. ISSN 10509631. doi: 10.1002/hipo.20487. (page 10)

[93] Benjamin Thyreau, Kazunori Sato, Hiroshi Fukuda, and Yasuyuki Taki. Segmentation of the hippocampus by transferring algorithmic knowledge for large cohort processing. Medical Image Analysis, 43:214–228, jan 2018. ISSN 13618423. doi: 10.1016/j.media.2017.11.004. (page 14)

[94] Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural networks for machine learning, 4(2):26–31, 2012. (page 39)

[95] P. Vemuri, H. J. Wiste, S. D. Weigand, L. M. Shaw, J. Q. Trojanowski, M. W. Weiner, D. S. Knopman, R. C. Petersen, and C. R. Jack. MRI and CSF biomarkers in normal, MCI, and AD subjects: Predicting future clinical change. Neurology, 73(4):294–301, jul 2009. ISSN 1526632X. doi: 10.1212/WNL.0b013e3181af79fb. (page 5)

[96] Nishant Verma, S. Natasha Beretvas, Belen Pascual, Joseph C. Masdeu, and Mia K. Markey. New scoring methodology improves the sensitivity of the Alzheimer's Disease Assessment Scale-Cognitive subscale (ADAS-Cog) in clinical trials. Alzheimer's Research and Therapy, 7(1):64, nov 2015. ISSN 17589193. doi: 10.1186/s13195-015-0151-0. (page 3)

[97] Christian Wachinger, Martin Reuter, and Tassilo Klein. DeepNAT: Deep convolutional neural network for segmenting neuroanatomy. NeuroImage, 170:434–445, apr 2018. ISSN 10959572. doi: 10.1016/j.neuroimage.2017.02.035. (page 14)

[98] Jinghua Wang, Lili He, Hairong Zheng, and Zhong Lin Lu. Optimizing the Magnetization-Prepared Rapid Gradient-Echo (MP-RAGE) sequence. PLoS ONE, 9 (5):12, may 2014. ISSN 19326203. doi: 10.1371/journal.pone.0096899. (page 9)

[99] Yuenan Wang, Chenbin Liu, Xiao Zhang, and Weiwei Deng. Synthetic CT Generation Based on T2 Weighted MRI of Nasopharyngeal Carcinoma (NPC) Using a Deep Convolutional Neural Network (DCNN). Frontiers in Oncology, 9:1333, nov 2019. ISSN 2234943X. doi: 10.3389/fonc.2019.01333. (page 14, 15)

[100] Michael W. Weiner, Dallas P. Veitch, Paul S. Aisen, Laurel A. Beckett, Nigel J. Cairns, Robert C. Green, Danielle Harvey, Clifford R. Jack, William Jagust, Enchi Liu, John C. Morris, Ronald C. Petersen, Andrew J. Saykin, Mark E. Schmidt, Leslie Shaw, Li Shen, Judith A. Siuciak, Holly Soares, Arthur W. Toga, and John Q.

Trojanowski. The Alzheimer's Disease Neuroimaging Initiative: A review of papers published since its inception, sep 2013. ISSN 15525260. (page 10)

[101] Dominik Weishaupt, Victor D Köchli, and Borut Marincek. Wie funktioniert MRI? Eine Einführung in Physik und Funktionsweise der Magnetresonanzbildgebung. Springer, 6 edition, 2006. ISBN 978-3-540-89572-5. (page 21)

[102] William M. Wells, Paul Viola, Hideki Atsumi, Shin Nakajima, and Ron Kikinis. Multimodal volume registration by maximization of mutual information. Medical Image Analysis, 1(1):35–51, 1996. ISSN 13618415. doi: 10.1016/S1361-8415(01)80004-9. (page 28)

[103] J. L. Whitwell, K. A. Josephs, M. E. Murray, K. Kantarci, S. A. Przybelski, S. D. Weigand, P. Vemuri, M. L. Senjem, J. E. Parisi, D. S. Knopman, B. F. Boeve, R. C. Petersen, D. W. Dickson, and C. R. Jack. MRI correlates of neurofibrillary tangle pathology at autopsy: A voxel-based morphometry study. Neurology, 71(10):743–749, sep 2008. ISSN 1526632X. doi: 10.1212/01.wnl.0000324924.91351.7d. (page 5)

[104] Anders Wimo, Maëlenn Guerchet, Gemma Claire Ali, Yu Tzu Wu, A. Matthew Prina, Bengt Winblad, Linus Jönsson, Zhaorui Liu, and Martin Prince. The worldwide costs of dementia 2015 and comparisons with 2010. Alzheimer's and Dementia, 13 (1):1–7, jan 2017. ISSN 15525279. doi: 10.1016/j.jalz.2016.07.150. (page 2)

[105] L. E.M. Wisse, L. Gerritsen, J. J.M. Zwanenburg, H. J. Kuijf, P. R. Luijten, G. J. Biessels, and M. I. Geerlings. Subfields of the hippocampal formation at 7T MRI: In vivo volumetric assessment. NeuroImage, 61(4):1043–1049, jul 2012. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2012.03.023. (page 10, 11)

[106] World Health Organisation. Global action plan on the public health response to dementia 2017 - 2025. Technical report, World Health Organisation, 2017. (page 1, 2)

[107] Zhengwang Wu, Yaozong Gao, Feng Shi, Valerie Jewells, and Dinggang Shen. Automatic hippocampal subfield segmentation from 3T multi-modality images. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), volume 10019 LNCS, pages 229–236. Springer Verlag, oct 2016. ISBN 9783319471563. doi: 10.1007/978-3-319-47157-0_28. (page 17)

[108] Zhengwang Wu, Yaozong Gao, Feng Shi, Guangkai Ma, Valerie Jewells, and Dinggang Shen. Segmenting hippocampal subfields from 3T MRI with multi-modality images. Medical Image Analysis, 43(d):10–22, jan 2018. ISSN 13618423. doi: 10.1016/j.media.2017.09.006. (page 14)

[109] Qianye Yang, Nannan Li, Zixu Zhao, Xingyu Fan, Eric I-Chao Chang, and Yan Xu. MRI Cross-Modality NeuroImage-to-NeuroImage Translation, jan 2018. (page 15)

[110] Xianfeng Yang, Ming Zhen Tan, and Anqi Qiu. CSF and Brain Structural Imaging Markers of the Alzheimer's Pathological Cascade. PLoS ONE, 7(12), dec 2012. ISSN 19326203. doi: 10.1371/journal.pone.0047406. (page 4, 5)

[111] Yoshua Bengio Yann LeCun. Convolutional networks for images, speech, and time series.MIT Press, Cambridge. The handbook of brain theory and neural networks, 3361(10):1995, 1995. (page 36)

[112] Y. Yuan, Z. X. Gu, and W. S. Wei. Fluorodeoxyglucose-positron-emission tomography, single-photon emission tomography, and structural MR imaging for prediction of rapid conversion to alzheimer disease in patients with mild cognitive impairment: A meta-analysis. American Journal of Neuroradiology, 30(2):404–410, feb 2009. ISSN 01956108. doi: 10.3174/ajnr.A1357. (page 6)

[113] Paul A. Yushkevich, Brian B. Avants, John Pluta, Sandhitsu Das, David Minkoff, Dawn Mechanic-Hamilton, Simon Glynn, Stephen Pickup, Weixia Liu, James C. Gee, Murray Grossman, and John A. Detre. A high-resolution computational atlas of the human hippocampus from postmortem magnetic resonance imaging at 9.4 T. NeuroImage, 44(2):385–398, jan 2009. ISSN 10538119. doi: 10.1016/j.neuroimage.2008.08.042. (page 10)

[114] Paul A. Yushkevich, Robert S.C. Amaral, Jean C. Augustinack, Andrew R. Bender, Jeffrey D. Bernstein, Marina Boccardi, Martina Bocchetta, Alison C. Burggren, Valerie A. Carr, M. Mallar Chakravarty, Gaël Chételat, Ana M. Daugherty, Lila Davachi, Song Lin Ding, Arne Ekstrom, Mirjam I. Geerlings, Abdul Hassan, Yushan Huang, J. Eugenio Iglesias, Renaud La Joie, Geoffrey A. Kerchner, Karen F. LaRocque, Laura A. Libby, Nikolai Malykhin, Susanne G. Mueller, Rosanna K. Olsen, Daniela J. Palombo, Mansi B. Parekh, John B. Pluta, Alison R. Preston, Jens C. Pruessner, Charan Ranganath, Naftali Raz, Margaret L. Schlichting, Dorothee Schoemaker, Sachi Singh, Craig E.L. Stark, Nanthia Suthana, Alexa Tompary, Marta M. Turowski, Koen Van Leemput, Anthony D. Wagner, Lei Wang, Julie L. Winterburn, Laura E.M. Wisse, Michael A. Yassa, and Michael M. Zeineh. Quantitative comparison of 21 protocols for labeling hippocampal subfields and parahippocampal subregions in in vivo MRI: Towards a harmonized segmentation protocol. NeuroImage, 111:526–541, may 2015. ISSN 10959572. doi: 10.1016/j.neuroimage.2015.01.004. (page 10, 12)

[115] Michael M. Zeineh, Stephen A. Engel, Paul M. Thompson, and Susan Y. Bookheimer. Dynamics of the hippocampus during encoding and retrieval of face-name pairs. Science, 299(5606):577–580, jan 2003. ISSN 00368075. doi: 10.1126/science.1077775. (page 10)

[116] Jun Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In Proceedings of

the IEEE International Conference on Computer Vision, volume 2017-Octob, pages 2242–2251, 2017. ISBN 9781538610329. doi: 10.1109/ICCV.2017.244. (page 15)