FACULTY OF
COMPUTER SCIENCE

A Deep Learning Framework for Predicted 4D MRI

Dissertation

zur Erlangung des akademischen Grades

Doktoringenieur (Dr.-Ing.)

angenommen durch die Fakultät für Informatik
der Otto-von-Guericke-Universität Magdeburg

von

Gino Gulamhussene, M.Sc.

geb. am 16.08.1989 in Staßfurt

1. Gutachter	Prof. Dr.-Ing. Christian Hansen Fakultät für Informatik Otto-von-Guericke University
2. Gutachter	Prof. Dr.-Ing. Frank Gerrit Zöllner Medizinische Fakultät Mannheim Universität Heidelberg
3. Gutachter	Prof. Dr.-Ing. Dietrich Paulus Fachbereich Informatik Universität Koblenz
Betreuer	Prof. Dr.-Ing. Christian Hansen and Dr.-Ing. Marko Rak

Magdeburg, 07. Oktober 2024

Gino Gulamhussene

A Deep Learning Framework for Predicted 4D MRI

Dissertation, 07. Oktober 2024

Gutachter: Prof. Dr.-Ing. Christian Hansen , Prof. Dr.-Ing. Frank Gerrit Zöllner und Prof. Dr.-Ing. Dietrich Paulus

Betreuer: Prof. Dr.-Ing. Christian Hansen und Dr.-Ing. Marko Rak

Otto-von-Guericke-Universität Magdeburg

Virtual and Augmented Reality Group

Department of Simulation and Graphics

Faculty of Computer Science

Universitätsplatz 2

39106 and Magdeburg

Kurzfassung

Organbewegungen stellen eine ungelöste Herausforderung bei bildgesteuerten Interventionen wie Strahlentherapie, Biopsien oder Tumorablationen dar. In dem Bestreben, dieses Problem zu lösen, hat sich das Forschungsgebiet der zeitaufgelösten volumetrischen Magnetresonanztomographie oder 4D MRT entwickelt. Die derzeitigen Techniken sind jedoch für die meisten Interventionen ungeeignet, da sie nicht über eine ausreichende zeitliche und/oder räumliche Auflösung verfügen und lange Aufnahme- und Rekonstruktionszeiten haben.

In dieser Arbeit wurde ein öffentlicher Datensatz erstellt, welcher aus dynamischen 2D und statischen 3D Leber magnetic resonance imaging (MRI) von 20 gesunden Probanden besteht und für die Entwicklung und das Testen von 4D MRI Methoden genutzt werden kann. Auf diesem Datensatz wurde eine klassische Sortier- und Stacking-Methode, welche Template-Updates nutzt entwickelt und getestet. Sie dient als Referenz für die anschließende Entwicklung eines auf Deep Learning basierenden Ansatzes. Es wird gezeigt, dass Template-Updates die Robustheit der Methode gegenüber ouf-of-plane Bewegungen der verfolgten Landmarken verbessern und gleichzeitig die Rekonstruktion beschleunigen.

Ein neuartiges, auf Deep Learning basierendes Framework für die Erstellung von 4D Leber MRI wurde auf demselben Datensatz entwickelt und getestet. Es verwendet ein Deep-Learning-Netzwerk, bei dem Input und Output nach dem hier vorgeschlagenen Konzept der transitiven Informationsbrücken organisiert werden. Dies macht die Methode zu einer durchgängig trainierbaren Lösung für das 4D MRI Rekonstruktionsproblem. Die netzwerkunabhängige Eigenschaft des Ansatzes wurde mit verschiedenen Architekturen getestet. Das Framework erreicht Rekonstruktionszeiten von unter einer Sekunde für hochauflösende 3D Leber MRI mit großem field of view (FOV) und ermöglicht so echtzeit 4D Leber MRI.

Darüber hinaus wird eine Transfer-Learning-Strategie vorgeschlagen, um die Menge der Trainingsdaten und damit die für die 4D MRI Rekonstruktion benötigte vorherige Aufnahmezeit zu reduzieren. Damit wird dem medizinischen Hintergrund Rechnung getragen, dass lange Aufnahmezeiten medizinisch und wirtschaftlich nicht praktikabel sind. Es wird eine Ensembling-Strategie vorgeschlagen, bei der mehrere Modelle ein 4D MRI vorhersagen. Es wird gezeigt, dass der Mittelwert der Vorhersagen eine verbesserte Qualität und einen verbesserten Vorhersagefehler aufweist. Zudem wird

die Standardabweichung der Vorhersagen zur Berechnung einer Unsicherheitskarte verwendet. Darüber hinaus wird die Wiederverwendung von Trainindaten als transitive Informationsbrücken vorgeschlagen, um die Zeit für die vorherige Erfassung von Trainingsdaten zu reduzieren. Es wird gezeigt, dass diese Wiederverwendung auch die Vorhersagequalität verbessert.

Abstract

Organ motion poses an unresolved challenge in image-guided interventions like radiation therapy, biopsies or tumor ablation. In the pursuit of solving this problem, the research field of time-resolved volumetric magnetic resonance imaging or 4D MRI has evolved. However, current techniques are unsuitable for most interventional settings because they lack sufficient temporal and/or spatial resolution and have long acquisition and reconstruction times.

In this work a public data set of dynamic 2D and static 3D liver MRI of 20 healthy subjects was established for the development and testing of 4D MRI methods. On this dataset a classical sorting and stacking method that utilizes template updates was developed and tested as a baseline for the subsequent development of a deep learning based approach. It is shown that template updates improve robustness against out of plain motion of tracked landmarks, while speeding up the reconstruction at the same time.

A novel deep learning based framework for the generation of 4D liver MRI was developed and tested on the same dataset. It uses a deep learning network where the input and output are organized according to the here proposed concept of transitive information bridges. This makes the method an end-to-end trainable solution to the 4D MRI reconstruction problem. The approaches network agnostic property was tested with different architectures. The framework achieves sub-second reconstruction times for high resolution, large FOV 3D liver MRI, thus facilitating real-time 4D liver MRI.

Furthermore a transfer learning strategy is proposed to reduce the amount of training data and thus the prior acquisition time needed for 4D MRI reconstruction. This addresses the medical background where long acquisition times are medically and economically not feasible. An ensembling strategy is proposed in which multiple models predict a 4D MRI. It is shown that the mean of the predictions has an improved quality an prediction error and the standard deviation is used to calculate an uncertainty map. Furthermore, also to reduce prior acquisition times needed for training data, the reuse of training samples as transitive information bridges is proposed. It is shown that this re-utilization also improvement prediction quality.

Contents

FOV field of view

SNR signal-to-noise ratio

MRI magnetic resonance imaging

MR magnetic resonance

RF radio frequency

CT computer tomography

US ultra sound

MWA microwave ablation

RFA radio frequency ablation

MSE mean squared error

RMSE root mean squared error

MDISP mean displacement

TRE target registration error

std standard deviation

RQ research question

SB sample bridge

TL transfer learning

ANOVA analysis of variance

ROI region of interest

sar specific absorption rate

CNN convolutional neural networks

DL deep learning

1 Introduction

1.1 Motivation

During the last decade, 4D MRI has gained considerable interest in research, because it promises clinical access to information on the respiratory motion of the thorax and abdomen free of radiation. Respiratory motion information is vital for many medical applications in diagnostic (Merchavy et al., 2016), treatment planning (Han et al., 2018) and execution (Colvill et al., 2016). In this wide field, this thesis is motivated by the potential of the use of 4D MRI in MRI guided percutaneous interventions on the liver like radio frequency-, microwave- and cryoablation, as well as biopsies and brachytherapy.

These interventions have in common the insertion of a needle instrument through the skin of a patient and the advancement of the needle tip to a target location. In case of the biopsy this is done to retrieve cell samples for diagnostics. In the other interventions the goal is to effectively kill all tumor cells, by either heating, freezing, radiating the cells with ionizing radiation or disintegrating their cell membrane. In the case of a liver intervention, the target is moving due to breathing. The liver is a good example of the basic problem of organ motion, because it exhibits high variability in deformation due to the breathing motion. It is a challenge to navigate the needle towards a tumor solely based on what the radiologist can see in planning data and live image data, both viewed on a display in the intervention room. Having to navigate the needle within a moving organ towards a moving tumor adds to the challenge. In other words, inter- and intra-organ motion in the abdomen and thorax poses a challenge in image guided interventions in this body area (Cleary et al., 2010; Ha et al., 2018; Xing et al., 2006; Gueulette et al., 2005; Lambert et al., 2005). That is a problem, because insufficient compensation of irregular organ motion during image-guided interventions can lead to inaccuracies in the instrument's navigation to the target and thus to deteriorated treatment results. For example, a needle that diverges from the planned intervention path during advancement of the needle, might injure risk structures, or an initially misplaced needle might need to be re-positioned, prolonging the intervention and thus increasing the risk

for complications or, if not re-positioned, it may cause the under-ablation of tumor tissue.

Furthermore, several works were published proposing computer systems to aid image guided interventions. They mostly seek to lower the mental load of the radiologist. However, none of these methods have integrated a correction for organ deformation, as this is still part of active research and no readily usable solutions to this problem exist (Mewes et al., 2019; Heinrich et al., 2019).

On the other hand, motion models could be used to account for organ motion in these methods and they are already used in radiation therapy (Ha et al., 2018; Xing et al., 2006). Motion models can be trained on 4D MRI or 4D CT data (Tanner et al., 2012). The notion of 4D means that the imaging modality acquires several 3D volumes over a period of time. Hence, resolving the target organ in three space dimensions as well as in the time dimension, i.e., 3D+t. While 4D MRI does not involve ionizing radiation, it is not easily available from a medical point of view and still an area of active research. Part of this is due to the significant amount of data needed to reconstruct different breathing states, which needs a lot of costly scanner time. However, the properties of high soft tissue contrast and being free of ionizing radiation make MRI the imaging modality that is most prominent in the research for a 4D imaging technique for the abdomen or thorax. Development of 4D MRI methods is becoming more advanced and seeks to make 4D MRI readily available for use in clinical scenarios, like image guided radiation therapy or needle guidance during percutaneous cancer intervention on the liver.

In this avenue several problems have to be addressed. First of all, MRI is inherently slow compared to computer tomography (CT) or ultra sound (US) and is even slower for large FOV. Yet, due to the size of the target organs like the liver or lung, 4D MRI methods need a large FOV. And to be usable they need a high temporal resolution as well, which is difficult, due to MRI being slow. Because of that, current 4D MRI techniques are unsuitable for most interventional settings because they are limited to specific breathing phases, i.e., they are only reconstructing a single averaged breathing cycle, neglecting variability in organ movement and deformation, lack temporal and/or spatial resolution, and have long prior acquisition times and/or reconstruction times.

1.2 Research Gap and Problem Definition

During an initial literature research, which is discussed in chapter 3, limitations of current 4D MRI methods were identified. Three of them are addressed in this thesis:

1. **No real-time 4D MRI method:** While there exist 4D MRI method that have some of the characteristics that are necessary to be applicable in the aforementioned scenarios, no method exhibits all of these requirements. These necessary characteristics are:

   a) **Large FOV:** The method needs a large FOV, to fit the entire target organ, e.g., liver or lung.

   b) **High resolution:** The method needs a spatial resolution good enough to discern inner structures of the organ, like vessels. ($2\,\mathrm{mm} \times 2\,\mathrm{mm} \times 4\,\mathrm{mm}$)

   c) **Real-time imaging:** It must be real-time capable, having a high temporal resolution and realize imaging and reconstruction of separate 3D volumes in sub-seconds.

   d) **Time-resolved:** It also must be time-resolved, representing the actual, current organ deformation, rather than an averaged breathing phase.

   e) **Short prior acquisition time:** Most time-resolved methods need training or reference data, which is acquired beforehand. However, the method must facilitate short prior acquisition times ($leq2\,\mathrm{min}$), for it to be both medically feasible and economical.

   f) **High image quality:** Most methods with short prior acquisition times trade that off for image quality. However, the image quality must be good enough to allow for the inference of fine organ deformations, which in turn are important to determine the position of a targeted tumor. This means, the images cannot be blurry or contain ghosting artefacts.

2. **No uncertainty estimation:** Regarding 4D MRI methods that utilize machine learning, there exists little research regarding uncertainty estimation of these methods. Yet, especially in case of image generation there is a need for an estimate of how confident or uncertain the method is about the anatomical and temporal correctness of its output, e.g., regarding the position of a tumor or risk structure.

3. **No public data set and common benchmark:** While there exist large public datasets and benchmarks for other medical image processing tasks like segmentation in the liver or prostate, there is no public data set for the development, testing, and direct comparison of 4D MRI methods that facilitate further research in the area as well as comparability of methods.

1.3 Research Questions

Based on the limitations of the current state of the art, the aim of the present work was to develop a real-time capable, time-resolved 4D MRI method with large FOV and high spatial and temporal resolution. Because deep learning has shown enormous successes in a variety of medical image processing tasks, this thesis focused on the realization of such a method based on deep learning. The first research question therefore was:

R1:

"Can deep learning methods be used to generate real-time 4D MRI with high spatiotemporal resolution, base on a real-time 2D MRI sequence?"

Seeing the need for such a method to be readily available and be both medical and economically feasible, a second focus of this thesis was to reduce the necessary reference or training data of such a method. Hence, the second research question was:

R2:

"Can the training data requirement for the training of such a deep learning based 4D MRI method be limited to below $3\,\mathrm{min}$, while achieving high prediction quality?"

1.4 Thesis Structure

The thesis is structured as follows.

Chapter 2 gives the medical and technical background, necessary to understand this thesis. This includes the basic principle of magnetic resonance (MR) imaging, as well as the basic ideas of convolutional neural networkss (CNNs) and UNets, which are the basis of the deep learning based 4D MRI approach presented.

Chapter 3 presents the state of the art of 4D MRI methods. A classification of current methods is made and advantages and limitations are discussed, which lead to the presented research gap.

Chapter 4 describes the acquisition and structure of data set, which was used for the development of the 4D MRI methods presented in this thesis, and was published to be of use for other researchers.

Chapter 5 describes a classic sorting approach to the 4D MRI problem both as a prototype and method to compare the deep learning approach against and to check the validity of the data base.

Chapter 6 presents a novel real-time deep learning based 4D MRI prediction framework.

Chapter 7 focuses on reducing the prio acquisition time of training data needed for the deep learning framework.

Chapter 8 focuses on increasing the image quality while reducing the prior acquisition time further as well as presenting a way of uncertainty assessment within the proposed framework.

Chapter 9 summarizes the findings, limitations and future work that arise from the thesis as a whole. A complete list of publications of the author is given as well.

1.5 Publication List

A list of all publications this thesis is based on is given in the following. Note that the last two conference papers did not directly contribute to this thesis. They are marked with **"not used for this thesis"**.

Journal Papers

Gino Gulamhussene, Fabian Joeres, Marko Rak, Maciej Pech, and Christian Hansen (2020). *4D MRI: Robust sorting of free breathing MRI slices for use in interventional settings*. PloS one, 15(6), e0235175.

Gino Gulamhussene, Anneke Meyer, Marko Rak, Oleksii Bashkanov, Jazan Omari, Maciej Pech, and Christian Hansen (2022). *Predicting 4D Liver MRI MRI for MR-guided Interventions*. Computerized Medical Imaging and Graphics, 101, 102122.

Gino Gulamhussene, Marko Rak, Oleksii Bashkanov, Fabian Joeres, Jazan Omari, Maciej Pech, and Chistian Hansen (2023). *Transfer-learning is a key ingredient to fast deep learning-based 4D liver MRI reconstruction*. Scientific Reports, 13(1), 11227.

Conference Papers

Gino Gulamhussene, Oleksii Bashkanov, Jazan Omari, Maciej Pech, Christian Hansen, and Marko Rak (2023, October). *Using Training Samples as Transitive Information Bridges in Predicted 4D MRI*. In Workshop on Medical Image Learning with Limited and Noisy Data (pp. 237-245). Cham: Springer Nature Switzerland.

Gino Gulamhussene, Jonathan Spiegel, Arnab Das, Marko Rak, Christian Hansen (2023, June). *Deep Learning-based Marker-less Pose Estimation of Interventional Tools using Surrogate Keypoints*. BVM Workshop (pp. 292-298). Wiesbaden: Springer Fachmedien Wiesbaden. **not used for this thesis**

Gino Gulamhussene, Arnas Das, Jonathan Spiegel, Marko Rak, Christian Hansen (2023, June). *Needle Tip Tracking During CT-guided Interventions using Fuzzy Segmentation*. In BVM Workshop (pp. 285-291). Wiesbaden: Springer Fachmedien Wiesbaden. **not used for this thesis**

Gino Gulamhussene, Jonathan Spiegel, Christian Hansen (2021, October). *Using Deep Learning for Dose-Reduced Marker-less Instrument Tracking in CT Guided Interventions*. In 5th Conference on Image-Guided Interventions - Digitalisierung in der Medizin. **not used for this thesis**

Datasets

Gino Gulamhussene, Fabian Joeres, Marko Rak, Cindy Lübeck, Maciej Pech, and Christian Hansen (2019). *2D MRI liver slices with navigator frames. A test data set for image based 4D MRI reconstruction*. https://doi.org/10.24352/UB.OVGU-2019-093

Gino Gulamhussene, Anneke Meyer, Marko Rak, Oleksii Bashkanov, Jazan Omari, Maciej Pech, and Christian Hansen (2021). *2D MRI liver slices with navigator frames. A test data set for image based 4D MRI reconstruction (Part II)*. https://doi.org/10.24352/UB.OVGU-2021-071

2 Background

Synopsis

This chapter introduces the clinical as well as the technical background. In the medical part, first, the Liver, which is used as example target organ throughout this thesis is described and respiration, which is fundamental to the entire work as a restricting factor, is formalised. An outline of a state of the art clinical workflow of an image-guided intervention is given. In the technical part, the fundamentals of magnetic resonance imaging are explained as well as associated concepts, which are relevant for this work. A background of deep learning is given and finally the evaluation measures used in the thesis are described.

2.1 Target Organ Liver

The liver is used in this thesis as a target organ to exemplify the presented methods. Akinyemiju et al. (2017) state that liver cancer is among the leading causes of cancer deaths globally and that liver cancer cases increased by 75% between 1990 and 2015. This trend has continued in more recent times. Until 2020, the incidences of liver cancer and cancer in general have been further increasing (Galle et al., 2018; Bray et al., 2018; Sung et al., 2021). Besides primary liver tumors the liver is the second most common site for metastatic disease. More than 50% of all patients with malignant diseases develop liver metastases with significant morbidity and mortality (Pereira, 2007).

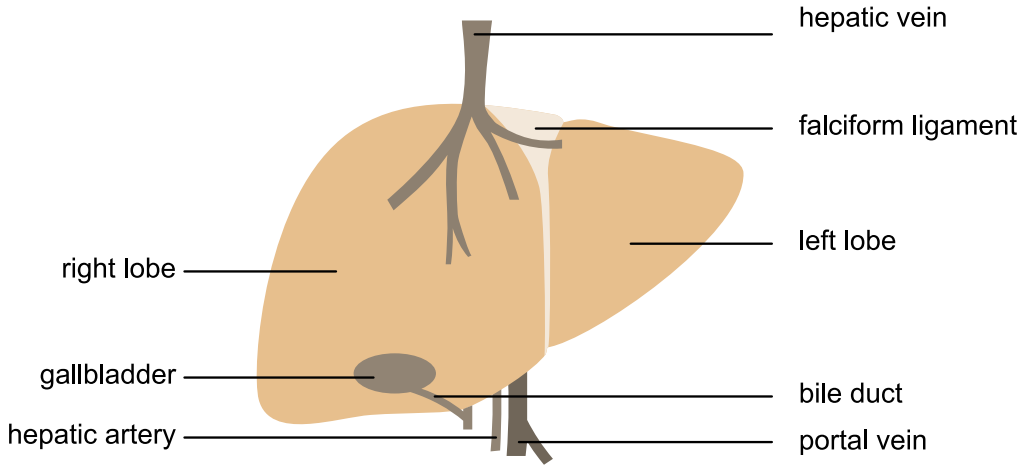

Fig. 2.1: The liver with its macro anatomy is divided in left and right liver lobe. Indicated are also the four in going and outgoing vessels systems, which all branch into the liver. Based on *Anatomy and physiology of the liver – Canadian Cancer Society* (2015)

The liver performs various functions, including the production of bile, metabolization, vitamin and mineral storage and the filtering of the blood (*Anatomy and physiology of the liver – Canadian Cancer Society* 2015). It is located in the abdominal cavity, below the rib cage and below the diaphragm that separates the chest cavity from the abdominal cavity and is surrounded by other organs. The lung and heart are located directly above the liver, on the other side of the diaphragm. The stomach and colon are below the liver. The liver contains four vessel systems that branch into the liver. The portal vein, which supplies 80% of the livers blood is draining from the spleen and intestines. The hepatic artery supplies the remaining 20% of the blood which is highly oxygenated and comes from the heart. The outflow of the liver is provided by the hepatic vein. (Sibulesky, 2013). The fourth is the bile duct

system. The liver is divided by the falciform ligament into a larger right lobe and a smaller left lobe (see Fig. 2.1)

2.2 Formalizing Breathing

Breathing is an important aspect in image guided interventions, as it interferes with the procedure. When a person inhales, the chest and with it the lungs expand while the diaphragm moves and pushes the abdominal organs like the liver down which also lets the stomach expand. While moving, the liver also deforms. This happens periodically with the breathing. Because there are interventions, for example on the liver, that are performed under free breathing, assistance system for these interventions need to account for breathing and hence must uses some kind of formalisation of the breathing act. There are two main approaches to do this. The first way is to derive a one dimensional relative or absolute breathing signal. The second way is to derive a multi-dimensional absolute breathing state.

While the terms breathing, respiration, and ventilation have different meanings in different scientific fields, in this thesis, all three terms are treated as synonym and mean the act of breathing in and out.

One dimensional Absolute Breathing Signal

The one dimensional absolute breathing signal tells how deep the inhale or exhale currently is. The absolute signal can be obtained by breathing through a hose and measuring the air volume. Another more easy and less accurate way is to use a belt that measures the expansion of either the chest or the stomach. The belt stretches when the chest or stomach expand and this stretch can be registered. A third way is to use interventional imaging and derive the signal from the image. For example measuring using the body cross section in an image slice. Because this formalization reduces all movement and deformation of the inner organs to one dimension, i.e., the current breathing depth, all breathing cycles can only be differentiated by their end-inhale and end-exhale breathing depths. All information about differences due to deformation is lost.

One dimensional Relative Breathing Signal

The one dimensional relative breathing signal tells the current so called breathing phase. When subdividing a breathing cycle into several phases of equal length, we get breathing phases or respiratory phases. In most of the relevant literature, the breathing cycle is divided into 6 or 10 phases (see Chapter 3 Tab.3.2). The simplest example is the division of the breathing cycle into four phases: 1) the inhalation phase, 2) the end-inhalation phase (or just end-inhale), 3) the exhalation phase, 4) the end-exhalation phase (or just end-exhale). These four phases define the breathing cycle from end-exhale to end-inhale and back again to end-exhale. However, a breathing cycle can be arbitrarily divided into more phases. The 6 phases example would have following phases: 25% inhale, 50% inhale, end-inhale, 25% exhale, 50% exhale, end-exhale.

The signal is derived from either image data or from raw data before image reconstruction. This is often done in MRI sequence programming. For that data is acquired during free breathing and retrospectively sorted or binned into a number of previously defined breathing phases.

Importantly, with this formalization approach each breathing cycle is described as identical and can't be distinguished at all after the formalization. In that way, it is the simplest and least descriptive formalization, describing each respiration as the repetition of identical breathing cycles broken down into identical breathing phases. While this simplicity is a great benefit of this approach, it can also be seen as a limitation, because a person breathes irregularly. Lets imaging two different breathing cycles of the same person. One was a deep breath, the other a shallow one. The liver in both breathing cycles at end-inhale will have a different place and deformation. This is a problem because if the breathing information is used to derive the position of a target, the derived target position will be wrong.

Another limitation is that all methods using this approach work only retrospectively. This is due to the relative nature of the signal. Because the breathing phase is determined relative to the complete breathing cycle, it must be captured in total first before deriving the individual phases.

Multi-dimensional Absolute Breathing State

In contrast to the breathing phase the breathing state is multi-dimensional and absolute. It is multi-dimensional to preserve the information about the position and

deformation of the inner organs that are effected by respiration. See Fig. 2.2 for a visualization on the non linear deformation of the liver, caused by breathing.

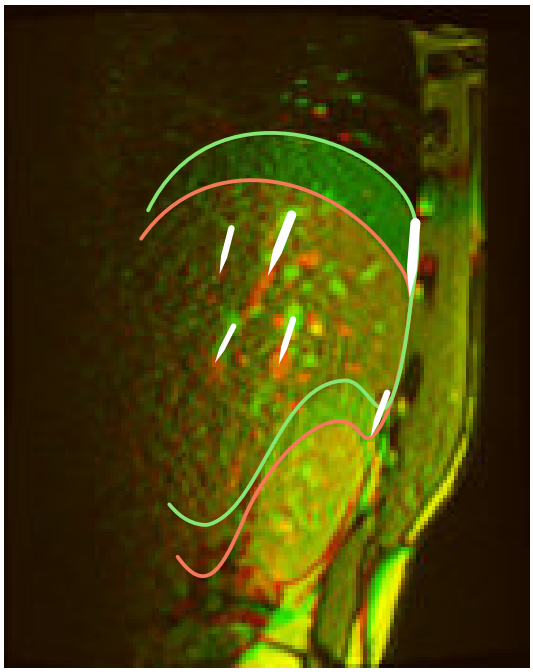

Fig. 2.2: Visualisation of the multi-dimensionality of breathing deformation. Used are two liver MRI slices of one subject at the same position and at exhale and inhale time points from the data set established in this thesis. The dataset is described in chapter 4. The inhale image is superimposed (in red) on the exhale image (in green). The liver contour of both inhale and exhale are drawn as well. White arrows show the movement of vessel cross sections. Note the apparent differences in arrow length as well as direction, which indicates the multi-dimensionality of liver deformation, caused by breathing.

The formulation is also absolute in the sense that the breathing state can be determined at a singular point in time without needing the reference of a completed breathing cycle. The current position and deformation of all inner organs is what can be call the breathing state. However, one can keep the representation of the breathing state as simple as is needed for the use case. For example if the target organ of an intervention is the liver than it might be sufficient to only account for the liver motion and deformation and hence define the breathing state solely by the position and deformation of the liver. Furthermore, the position and deformation can be described with more or fewer degrees of freedom. The extreme case of one degree of freedom is equal to the one dimensional absolute breathing signal describing only the position of the liver. In all other cases the breathing state is a high-dimensional concept and thus not easily depicted visually. Because of that, in this thesis, The temporal change in the breathing state is shown in simplified form as a one-dimensional breathing curve. Throughout the work presented in this thesis,

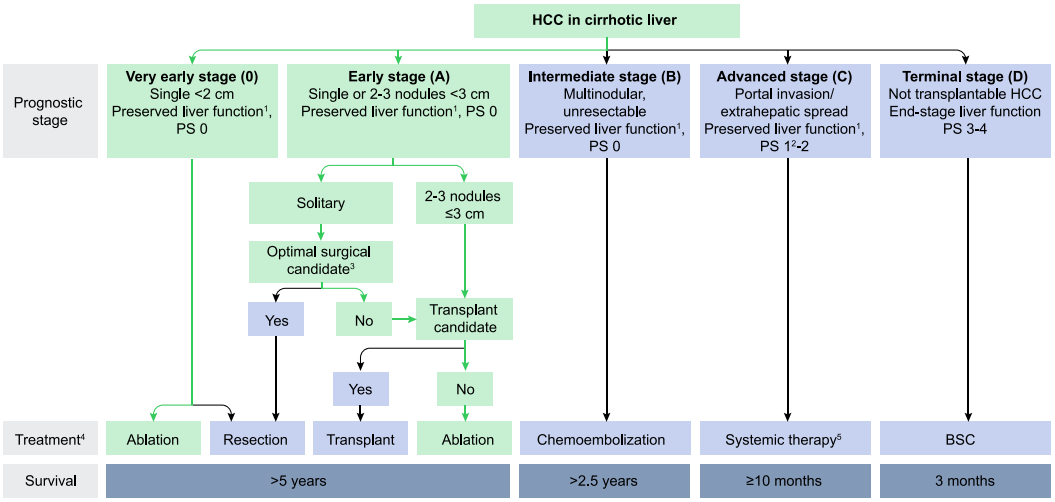

Fig. 2.3: The adapted Barcelona Clinic Liver Cancer (BCLC) staging and treatment strategy according to the EASL guidelines. The green highlights were added to the diagram retrospectively and mark the decision paths that lead to minimally invasive image guided local thermal ablation, in short, image guided interventional treatment of the tumor specifically. Taken from Galle et al. (2018)

breathing is formalized as a succession of breathing states not of simple breathing phases. The formalization relies on imaging. Location and deformation are derived from interventional real-time images.

2.3 Image-Guided Interventions

After a tumor was diagnosed, a multidisciplinary tumor board discusses the treatment options following the EASL guidelines (Galle et al., 2018) (see Fig. 2.3). Besides image-Guided interventions, the treatment options include the systemic therapy, liver resection (hepatectomy), transplantation (Malek et al., 2014). Systemic therapy is associated with side effects such as nausea, vomiting and worsening of liver function (Malek et al., 2014). Liver resection and transplantation are radical surgeries. While resection is contraindicated in cases of tumor related macrovascular invasion, transplantation is only performed when the Milan criteria are met and no extrahepatic metastases or vascular invasion was identified (Galle et al., 2018).

The background to the present work is thhe percutaneous local ablation. It is suited especially for older patients and patients with weakened hepatic function, multiple smaller tumors (Galle et al., 2018) or even larger tumors (Bale et al., 2010). Ablation, also called thermoablation, has several advantages over surgical resection: lower morbidity, increased preservation of surrounding tissues, reduced cost and

shorter hospitalization times (Pereira, 2007). Percutaneous interventions were made possible through the advent and advances of cross-sectional imaging (Chu et al., 2014) like CT, MRI, and US.

Whenever the term intervention is used in the present work, it refers to minimally invasive, local, percutaneous interventions. The seminal technique was percutaneous ethanol injection (PEI). It induces coagulative necrosis of the lesion as a result of cellular dehydration, protein denaturation, and chemical occlusion of small tumour vessels. (Galle et al., 2018) It was followed by local thermal ablations, which are classified as either hyper-thermic treatments or cryoablation. While the first is the heating of tissue at 60 to 100 °C to irreversibly damage it and includes radio frequency ablation (RFA), microwave ablation (MWA), and laser ablation (Galle et al., 2018), cryoablation is the freezing of tissue at −20 to −60 °C. Most of these procedures are performed percutaneous (Galle et al., 2018). The aim is always to irreversible damage the cell, which leads to cell death. For example, the cell death caused by RFA is based on the frictional heat generated using high-frequency alternating current, which produces coagulative necrosis of the tumour. It allows the extension of the necrosis to a safety margin. RFA has been evaluated as first-line therapy in early hepato-cellular carcinoma (HCC). Overall survival in very early HCC ($< 2\,\mathrm{cm}$) treated by RFA was demonstrated to be at least equal to surgical treatment in a Markov model and in a cost-effective analysis based on data from a systematic review (Galle et al., 2018). A limitation of RFA is its susceptibility to the heat-sink effect in proximity to large blood vessels, in which heat is transported away from the ablation zone by the flowing blood. This decreases the hyperthermia and thus the efficacy of the RFA. This makes RFA less suitable for tumor tissue that is adjacent to vasculature (Chu et al., 2014).

2.3.1 General Clinical Workflow

The general workflow was described by Mewes (2019) for MWA interventions in MRI (see Fig. 2.4), however the general workflow is the same for other ablation techniques like RFA and for the imaging modality CT.

1. It starts with the preparation of the patient. This includes the patient education, positioning of the patient on the table of the MRI as well as the intubation anaesthesia. Then a flexible coil is placed on the operating field. The patient is translated into the MRI bore.

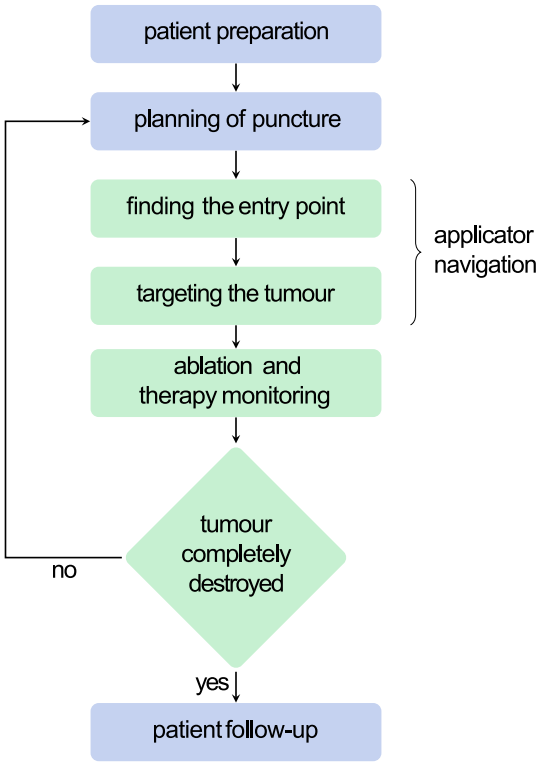

Fig. 2.4: Workflow of an image-guided percutaneous thermal ablation in green and pre- and postinterventional steps in blue. Based on Mewes (2019)

2. The planning step comes directly before the intervention. Morphological T1 and T2-weighted datasets of the target region are acquired. The applicator path is planned from outside the intervention room in the control room. Using the anatomical data acquired, the optimal entry point and path to the target position is identified. The radiologist identifies a path that is easily accessible, contains no structure of risk, and is short. The MRI planes are adjusted accordingly in order to visualize the complete instrument path. The planning data and interventional image data both are presented on a display in the intervention room near the MRI bore.

3. The actual intervention starts with finding the planned entry point on the patient. This is described in detail in section 2.3.2. After finding the entry point non-sterile, the intervention area is sterilized and the rest of the patient is covered in surgical drape. The access point is locally anaesthetised.

4. The instrument is advanced to the target under constant interventional live imaging, also called fluoroscopy. For that a fast T1-weighted MRI sequence is used. This is also described in detail in 2.3.2.

5. The tumor is ablated for typically 8 to 10 minutes in case of MWA, depending on tumor size and applicator specifications. The process is monitored using thermometry imaging (Roujol et al., 2010; Kägebein et al., 2018b). An additional T1-weighted dataset is acquired to check for the completeness of the ablation. If the tumor is not fully covered by the ablation zone, the applicator is repositioned and the ablation is continued. Otherwise, the necrosis zone is verified in a final control scan using contrast agent. After the ablation, the patient is moved out of the bore and extubated in the wake-up room. The whole intervention time is approximately 120 to 180 minutes. A follow-up is set for three and twelve months later.

2.3.2 Instrument Navigation

The instrument navigation from the entry point on the patient skin to the targeted tumor can be seen as containing two stages. First, finding the entry point and second, advancing the needle-like instrument to the target.

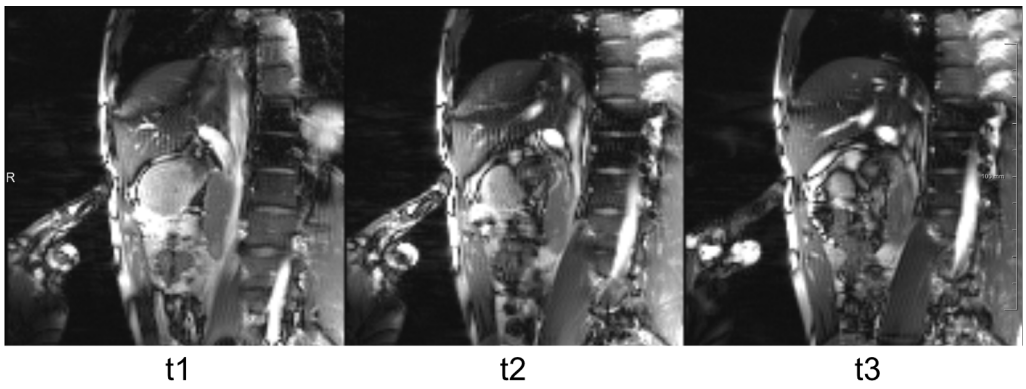

Fig. 2.5: The radiologist moves a finger on the patient skin near the presumed entry point and verifies the correct entry position in a real-time MRI image stream, while orienting on landmarks seen in both the real-time image as well as the planning data. The figure shows three different time points (t1, t2, and t3) of the real-time image stream. Based on Mewes (2019)

Finding the Entry Point

The most common method to find the entry point is the finger tipping method (Mewes, 2019) using the index finger (Fischbach et al., 2011; König et al., 2003; Fritz et al., 2011) or a saline filled syringe (Lewin et al., 2000). During continuous imaging in the planned imaging plane the radiologist moves the index finger along the patient until it can be seen in the live interventional image. In practice that means, during continuous acquisition, a new image is presented at least every 2 s. The finger is then moved along the imaging plane until it reaches the planned entry point. For orientation, the radiologist is using anatomical land marks that are visible in the planning data in which the entry point was planned. Once the entry point is found the finger remains in place while the patient is moved out of the bore. Then the position sterile draped, local anaesthetic administered, the skin incision is made, and the MR-compatible needle is inserted and advanced subcutaneously, then the patient is moved into the magnet. (Fischbach et al., 2011; Koenig et al., 2001).

Targeting the Tumor

Once the needle is inserted, it is guided along the planned path towards the targeted location, e.g., the tumor. The most common and simple method is the freehand technique (Rothgang et al., 2013), which requires continuous MRI fluoroscopy. Three parallel or orthogonal planes are acquired to provide orientation along the planned path (Rothgang et al., 2013). Within the MR image the needle is only visible as a needle artifact. The fluoroscopy images together with the planning images

are presented to the radiologist within the intervention room on an in-room display (Rothgang et al., 2013). The three main advantages of the freehand method are that it allows for fast reactions to positional changes of organs, due to breathing, that it allows for complex non-linear paths to the target, by means of bending the needle and by that passing around a risk structure and that it does not require additional equipment, like tracking cameras or a physical needle guide.

The workflow is comparable with the US-guided percutaneous intervention (Rothgang et al., 2013). However, the manual adaptation of the planned applicator path requires re orientations of the MR imaging plane, which is a major workflow disturbance (Mewes, 2019). This is also the case when the applicator unintentionally leaves the imaging plane. This happens more often in the beginning of the needle advancement, because it is difficult to infer the needle orientation in the first centimetres of the path, because the needle artefact is only visible inside the patient's body. Also the contrast of the interventional image might be a different one than the contrast of the planning image, which makes it harder to mentally match landmarks between interventional imaging and planning data. This also includes the fact that the tumour could be visible in the planning data but not in the interventional data. This can be either due to the selected MRI sequence during planning, because the visibility of the tumor is dependent of the MR contrast, or due to the use of contrast agents during planning.

There exist also mechanical guidance systems that are directly connected to the MRI, consisting of a remote manipulator and allowing for dynamic imaging plane adjustments according to the current needle orientation (Tsekos et al., 2005). It's missing haptic feedback is still a major drawback that has to be actively researched on (Schreiter et al., 2023).

Another solution was proposed by Kägebein et al. (2018a). They developed an automatic MR image alignment, using a so called Moiré phase markers rigidly attached to a needle instrument, which are tracked by a bore ceiling mounted camera. They developed a special MRI sequence in which the imaging plane is adapted based on the needle tracking information. The limitation of this approach is it's small field of view and the marker rig attached to the instrument, which could interfere with the workflow of the radiologist.

In conclusion, even so image guidance for percutaneous interventions exist, there is still need for improvement regarding the orientation those systems provide for the radiologist. For example and in particular providing the radiologist with the current position of the tumor, even if it is not visible in the used interventional MRI sequence.

2.4 Magnetic Resonance Imaging

2.4.1 Physical basics

Nuclear Magnetic Resonance

Atomic nuclei that have an odd number of protons and/or neutrons have a nuclear angular momentum also known as nuclear spin. In the human body, hydrogen nuclei occur bound in the form of water. This is predominantly the light isotope with only one proton and no neutrons. So it has a nuclear spin. If the hydrogen nucleus is exposed to an external static magnetic field B_0, the nucleus has additional potential energy. In an MRI machine the external static magnetic field is generated by a superconducting magnet (see Fig. 2.6). The hydrogen nucleus behaves like an atomic gyroscope due to the external static magnetic field B_0. Due to the conservation of angular momentum the magnetic moment μ is precessing with the Larmor frequency f_0 around the main axis of the static magnetic field (see Fig. 2.7). In total, a number of all hydrogen spins are aligned in the direction (parallel) and a number against (anti-parallel) the main axis z of the magnetic field. When considering the entire spin ensemble, a macroscopic net magnetisation along the z-direction can be measured.

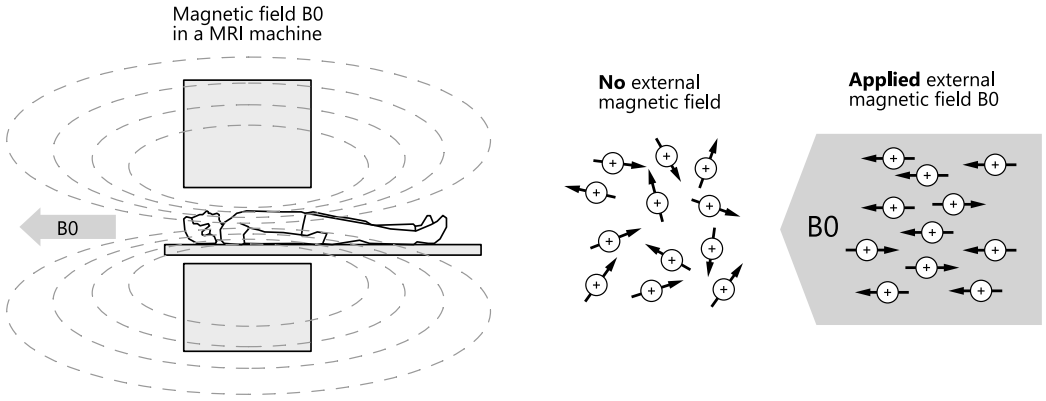

Fig. 2.6: Left: The super conduction magnet in the MRI machine produce a strong magnetic field B_0. Right: This external magnetic field causes the proton spins to align. Parts based on Pooley (2005)

By using an electromagnetic radio frequency (RF) pulse with the excitation frequency f_T, which is generated by an RF antenna (transmitting coil), the longitudinal magnetisation tilts out of its equilibrium state. This is only true if the exitation freqtuency f_T is the resonance frequency of the spins, which means $f_T = f_0$. The extent to

which the longitudinal magnetisation is tilted from its rest position in the direction of the transverse x-y plane depends on the so-called flip angle α (Kägebein, 2018). This in turn is determined by the duration of the RF pulse. The tilting leads to a magnetisation in the x-y plane and is referred to as transverse magnetisation M_xy. The transverse magnetisation rotates with the frequency f0 and induces an alternating voltage into a correspondingly positioned HF antenna (receiving coil).

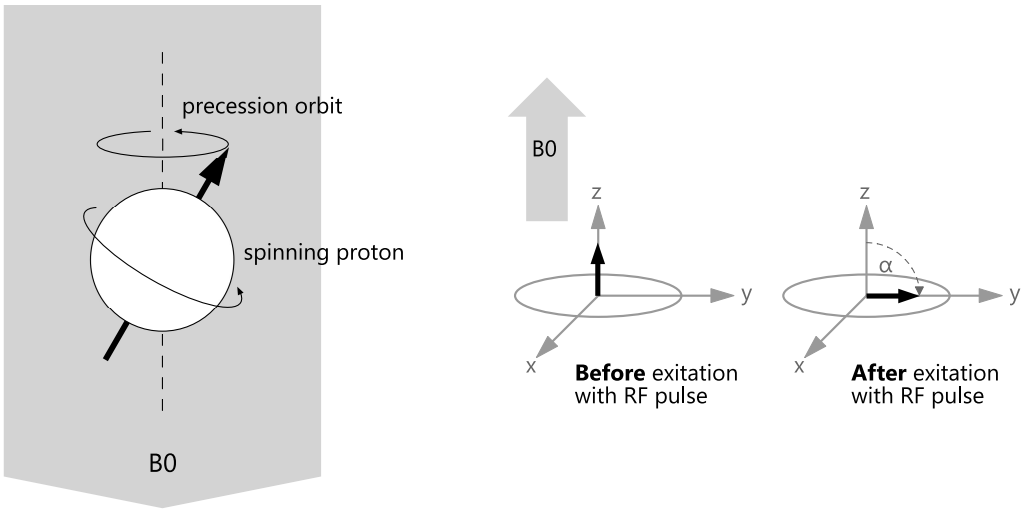

Fig. 2.7: Left: Precession of the proton spin with the Larmor frequency around the main axis of B_0. Right: Prior to an RF pulse, the net magnetization (black arrow) is longitudinal aligned with B_0. An RF pulse at the Larmor frequency will allow energy to be absorbed by the protons, exciting it, thus causing the net magnetization to tilt away from the z axis. Parts based on Pooley (2005)

Relaxation

There are two central interactions between the hydrogen nuclei and their environment. First, the spin-lattice relaxation and second, the spin-spin relaxation (Brown et al., 2014). The spin-lattice relaxation describes the return of the spin ensemble to the thermodynamic equilibrium state, which corresponds to the minimum energy state of the system. The necessary thermal energy exchange happens with the crystal lattice, whereby the spin ensemble can align itself parallel to the external static magnetic field again. The time constant that determines this process is the time T_1, which depends on the main magnetic field B_0 and the composition of the human tissue (Brown et al., 2014).

The second effect, the spin-spin relaxation, also leads to a reduction in transverse magnetisation. The magnetic field present locally at the individual spin is a combination of the external static magnetic field and the magnetic field of the neighbouring

spins. As a result, the local precession frequency varies, which in turn leads to a difference in dephasing for the individual spins. The corresponding transverse magnetisation of the spin ensemble thus decreases steadily over time. The descriptive time constant is the time T_2 , which is primarily dependent on the local composition of the human tissue.

In reality, there are additional local fluctuations of the external magnetic field, which leads to a faster decay of the transverse magnetisation. Together with T_2, this leads to the total relaxation time T_2^* (Brown et al., 2014).

Spatial Encoding

The basic principle of MR imaging is the measurement of the transverse magnetisation M_xy at a specific point in time (echo time (TE)) after the RF pulse has been emitted. The transversal magnetisation varies depending on the tissue under consideration and thus the times T_1 and T_2^*. Assuming that this signal intensity was measured with a receiving coil, each excited spin ensemble in the human body would contribute to the signal with the locally present transverse magnetisation M_xy. The gradient fields G_x, G_y and G_z generated by three separate gradient coils form a practical solution approach. They produce gradient fields. Their strength changes gradually in the respective spatial direction, hence the name gradient field. Irrespective of their orientation, the gradient coils generate a magnetic field aligned with the external magnetic field in the z-direction and thus leads to local variation or moderation of the total field (Bernstein et al., 2004).

This property is used to encode the local magnetisation M_xy. The coding process can be very diverse and that is what MR sequence programming is about. However, for a 2D sectional image, i.e., for an image slice, it follows the following basic principle, which has three streps.

1. Slice Selection: To realize a layer selective excitation, a slice selective gradient is applied (see Fig. 2.8). For example, consider the gradient field G_z is applied, the Larmor frequency f_0 of the spins varies as a function of the z position within the bore that the hydrogen nuclei is located. Accordingly, the RF pulse can only excite those spins whose precession frequency f_0 is equal to the exciting frequency f_T of the pulse. Note that the selected slice can have any orientation by using a some linear combination all three gradients G_x, G_y and G_z.

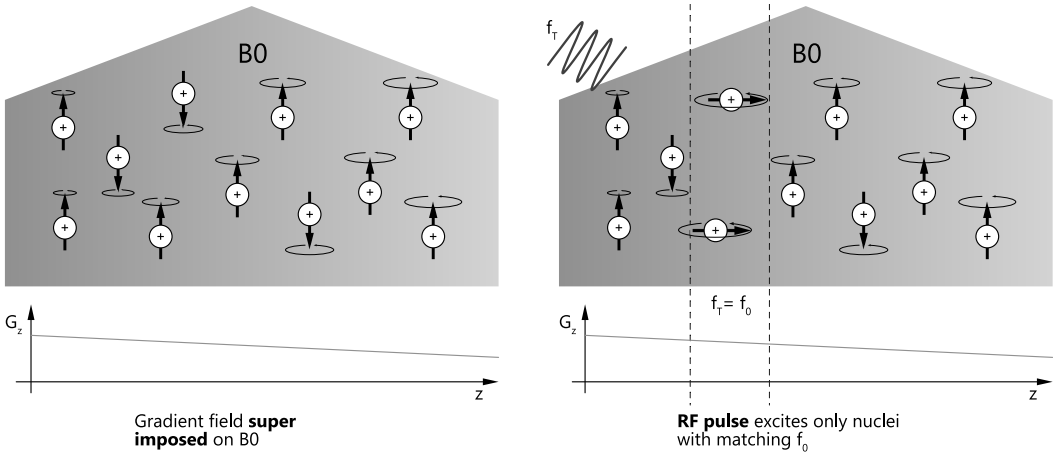

Fig. 2.8: Left: A gradient field is superimposed onto B_0 changing the local field strength. The precession frequency of the proton spin depends on the local field strength. Right: An RF pulse is applied, exciting only a narrow slice, i.e., contraining protons with a matching Larmor frequency f_0

2. Phase Encoding: Between the excitation and the measurement of the transverse magnetisation, a second gradient is applied. The positional dependence of its field strength is orthogonal to the earlier applied slice selective gradient. Following our example, where only the G_z gradient was used as slice selective gradient, now the G_y gradient would be applied to vary the frequency of the spins for only a short time as a function of the y position of the nuclei within the magnetic field. After switching off the gradient, the spins continue to precess with the same frequency f_0, but have different phases now. The used gradient is therefore referred to as the phase encoding gradient.

3. Frequency Encoding: During the measurement or readout process of the transverse magnetisation, the third gradient is applied constantly. This gradient is also aligned orthogonal to the two previous gradients. Note again, the orientation of the gradients is the spatial direction in wich the field strength varies, it is not necessarily the same as the direction of their magnetic fields. In the example the G_x gradient would be used. The local precession frequency changes depending on the x position. The gradient used in this step is known as the readout gradient.

The measured raw data is entered into a matrix known as k-space. The one-time execution of the described process of slice selection, phase encoding, and frequency encoding plus readout, fills one k-space line. For the reconstruction of an image, it is necessary to fill the k-space appropriately. Therefore the process is repeated several times. The interval between two consecutive RF pulses is referred to as

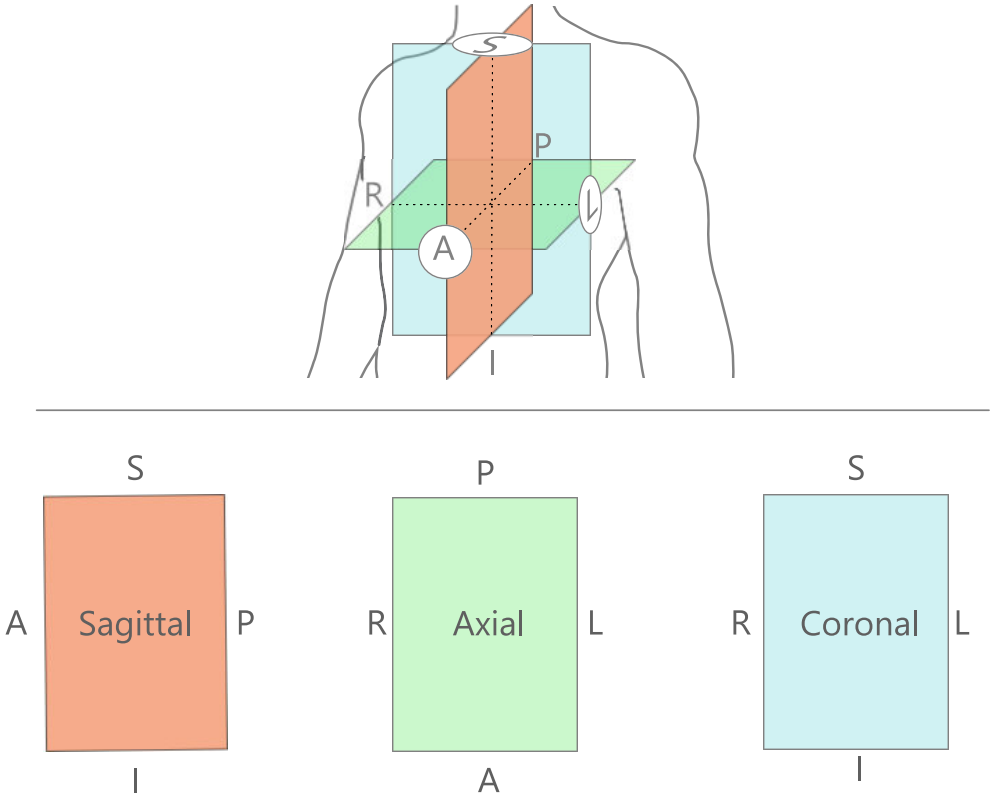

Fig. 2.9: Three main image or slice orientations are used in medical imaging. Sagittal, axial and coronal. The axes are defined by anatomical terms of location and direction, anterior (A) to posterior (P), right (R) to left (L) and inferior (I) to superior (s)

the repetition time (TR). As soon as the k-space is sufficiently filled, the location-dependent signal intensity can be decoded using an inverse Fourier transformation. The complex signal intensity $I(x, y)$ obtained is proportional to the local transverse magnetisation M_xy. The magnitude of the complex signal is used for the display of morphological MRI images. The display of a phase image, is primarily used for the evaluation of physical parameters, e.g., temperature or flow velocity.

2.4.2 Medical Image Orientation

In medical tomographic imaging the imaging planes have a distinct naming scheme, depending on their orientation (see Fig. 2.9). The sagittal orientation is from anterior (A) to posterior (P) and inferior (I) to superior (S). The axial orientation is from right (R) to left (L) and anterior (A) to posterior (P). The last orientation is the coronal one, it is from right (R) to left (L) and inferior (I) to superior (S).

2.5 Deep Learning

2.5.1 CNN

In medical image processing, CNNs play a important role and are extensively utilized. The concept of CNNs was initially introduced by LeCun et al. (1989) and gained prominence with the introduction of the AlexNet by Krizhevsky et al. (2017) 23 years later. CNNs draw conceptual inspiration from the visual cortex and leverage the grid-like organization of information in digital images. This is achieved by connecting neurons of one layer solely to neurons of the previous and subsequent layers within a defined neighborhood.

CNNs possess a significant computational advantage over fully connected networks due to their design. At the core of CNNs lie two components: convolution and pooling operations.

The convolution operation, utilizing a uniform filter kernel across the entire image, inherently introduces redundancy in the weights of convolutional layers. This redundancy is further exploited through the implementation of shared weights within a single convolutional layer. By leveraging shared weights, CNNs achieve greater space efficiency in storing network architecture information.

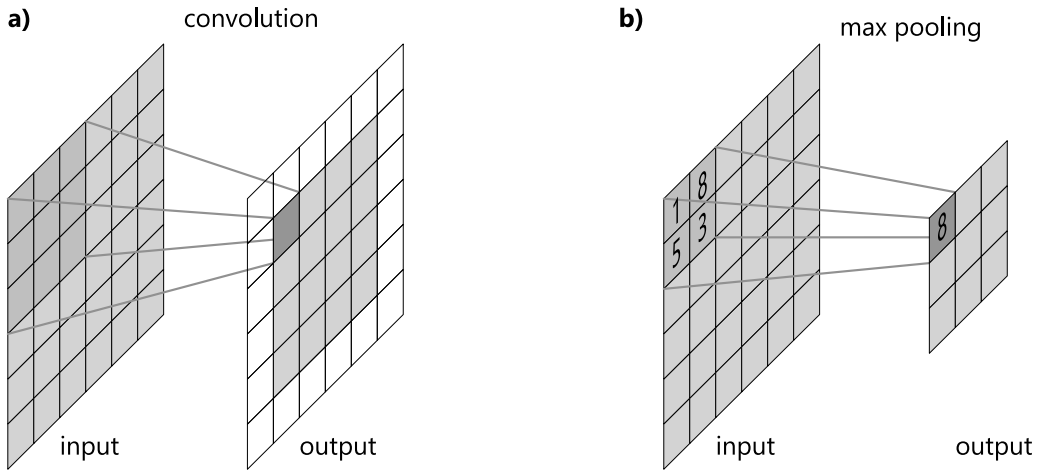

Fig. 2.10: a) Example convolution with filter kernel size 3x3, stride 1 and no padding. If now padding is applied, the output is smaller than the input. b) Example of max pooling with a size of 2x2 and stride 2

The CNNs architecture is structured in layers, with each layer comprising a collection of convolutional kernels or filters denoted as K with a given size s×s (see Fig. 2.10 a)), along with associated weights (W) and biases (B) corresponding to the size of

the kernel. This organized layer structure facilitates hierarchical feature extraction from input data, making CNNs highly effective in tasks such as image classification, segmentation, and image generation.

In CNNs, each filter operates on its input to produce a feature map, which is then non-linearly transformed element-wise. Specifically, the filters in layer l utilize the feature maps from the previous layer (X_{l-1}) as input to generate their own feature maps (X_l). Different boarder handling strategies are used. The most common one is the same padding, which pads the input, such that the output has the same size as the input.

As mentioned, pooling operations play a crucial role in CNNs by enhancing spatial context across the network's layers. These operations condense local regions of input feature maps into a single value, typically by extracting either the maximum (max pooling) or the average value (average pooling) (see Fig. 2.10 b)).

The parameters (W and B) of a neural network are iteratively learned through optimization of the loss function using gradient descent. This process involves calculating the gradient of the loss function with respect to the parameters, a task accomplished using the backpropagation algorithm, introduced by Rumelhart et al. (1986). This whole process is also refered to as training.

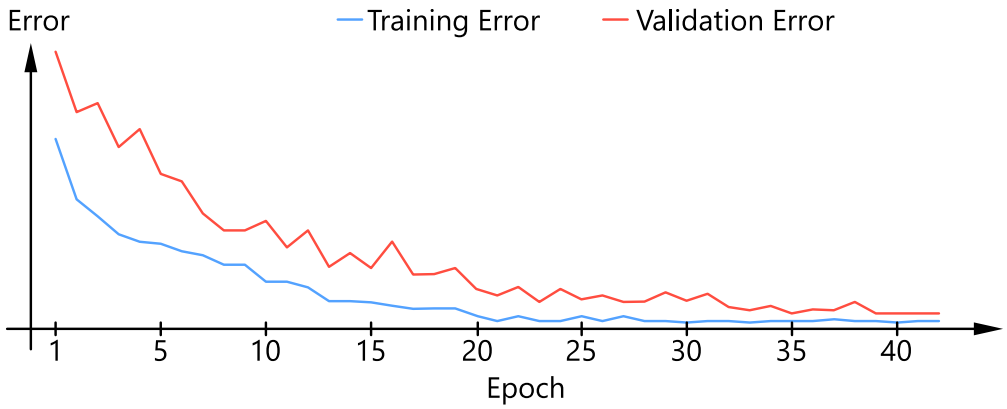

Fig. 2.11: Training a model sees the training error as well as the validation error decrease.

2.5.2 Training

Prior to training a model, the available data is split into a training set, a validation set and a test set. The training set contains the actual training data point which are used to compute the prediction error that is backpropagated. The validation data is

used to ensure the model does not overfit and instead achieves some meaningful generalization on the whole data set distribution. The test set is hold out until the end of training and hyper parameter tuning. The actual training is normally done in epochs. In each epoch the prediction error of the model is computed for all training data, also called training error or training loss. Often the computation is done in batches. After each epoch a validation error is computed on the validation data. The training is successful if the training as well as the validation error decrease over the course of epochs (see Fig. 2.11).

2.5.3 Overfitting

On the other hand, overfitting is the notion that a network learns a model with very high variance in order to perfectly model the seen training data (Shorten et al., 2019) while at the same time loosing its generalization to the whole distribution of the data. Overfitting is a common problem in deep neural networks and can be spotted when comparing the models training error curve with its validation error curve (see Fig. 2.12). There are several strategies to counter overfitting. Some of them are given in the following.

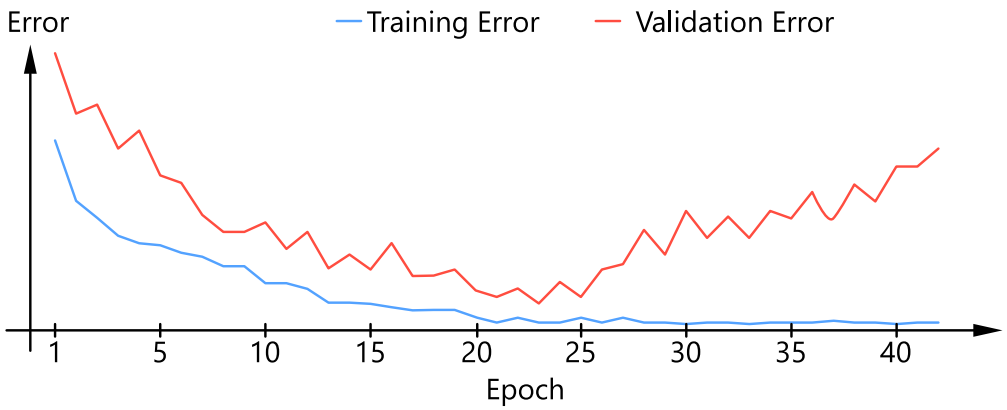

Fig. 2.12: A models overfitting to seen training data can be identified by comparing the training error curve with the validation error curve. While the validation error initially improves with the training error, it starts to worsen after some training epoch. At that point the model starts to overfit.

2.5.4 Dropout

Dropout is a way for addressing overfitting and was proposed by Srivastava et al., 2014. The key idea is to randomly drop units, i.e., neurons of the network, during training. This results in a "thinned" network. In that way dropout samples from an exponential number of different “thinned” networks during training. At test time, the effect of averaging the predictions of all thinned networks is approximated by using a single "un-thinned" network that has smaller weights. This significantly reduces overfitting.

2.5.5 Batch Normalization

Batch normalization (Ioffe et al., 2015) is another regularization technique that normalizes the set of activations in a layer (Shorten et al., 2019). Training Deep Neural Networks encounters the problem that the distribution of each layer’s inputs changes during training. This slows down the training by requiring lower learning rates and careful parameter initialization. Ioffe et al. (2015) refer to this phenomenon as internal covariate shift. To address it they proposed to make normalization a part of the model architecture and performing the normalization for each training mini-batch. To normalize a input batch, the batch mean is subtracting from each activation and dividing by the batch standard deviation. This allows for the use of much higher learning rates, relaxes the need for careful initialization, and in some cases eliminates the need for Dropout.

2.5.6 Augmentation Strategies

Of course, one way to address the problem of overfitting is to use large amounts of training data, however, that is not always practical like in the case of medical image analysis (Shorten et al., 2019). Data augmentation is another common means of countering overfitting. It artificially increases the number of training samples by altering the image data in predefined ranges. A survey on data augmentation strategies is given by Shorten et al. (2019). There are several image data augmentation techniques like geometric transformations, color space transformations, kernel filters, mixing images, random erasing, feature space augmentation, adversarial training, GAN-based augmentation, neural style transfer, and meta-learning schemes (Shorten et al., 2019). In this work only geometric transformations are utilized as data augmentation strategies, namely rotation, translation and scaling.

With any data augmentation strategy it is important to consider it's safety with regards to the label. For example rotation can be used for augmentation, but in a digit recognition task such as presented by MNIST (LeCun et al., 2010), rotating a "6" by $180\,\mathrm{deg}$ is not safe as the digit looks like a "9" after the augmentation. The safety is also dependent on the task. For example in a demonising task, the input and label can be safely rotated by the same amount, thus ensuring the correspondence of input and label is preserved. In the task dealt with in this paper, for example, it is relevant that the augmentation only changes the images to the extent that they are still plausible. This means that the shape and size of the organs as well as the orientation must remain valid. For example, if subjects are always lying on the back during MRI imaging and the intervention, than the augmentation should not result in an image that would only be produced if the patient or subject would stand upright within the MRI bore. Besides rotation, translation, i.e., shifting images left, right, up, or down can be a very useful transformation to avoid positional bias in the data. Again this translation must be in ranges such that the result could plausibly be produced the MRI imaging protocol. Finally, scaling can be applied to the input and label to change the size of the structures that are visible in the image.

2.5.7 U-Net

The name of the U-Net stems from its topology that looks like a U shape (see Fig. 2.13). It is due to the two symmetric paths, the encoder or contraction path, which extracts features from the input and compresses it to a denser representation and the decoder, or expansive path, which up-samples the dens or latent features back to the original size of the input image. The encoder and decoder are connected by the two central convolutional layers, in between which the feature maps represent the latent feature space. In the U shape analogy this is at the bottom of the U. Both encoder and decoder comprise distinct stages operating at varying resolution levels. In the original formulation, each stage consists of a convolutional block containing two convolutional layers with a 3×3 kernel and a stride of 1. After each convolutional layer, an activation function, in this case a rectified linear unit (ReLU) function is applied, thereby introducing non-linearity to the network.

Following each stage in the encoder, a down-sampling operation is performed to expand the network's receptive field. In the original U-Net design, down-sampling is achieved through 2×2 max pooling with a stride of 2. However, alternative implementations of the U-Net in the literature utilize strided convolutions, with strides larger than one, for down-sampling instead of max pooling. Moreover, the number

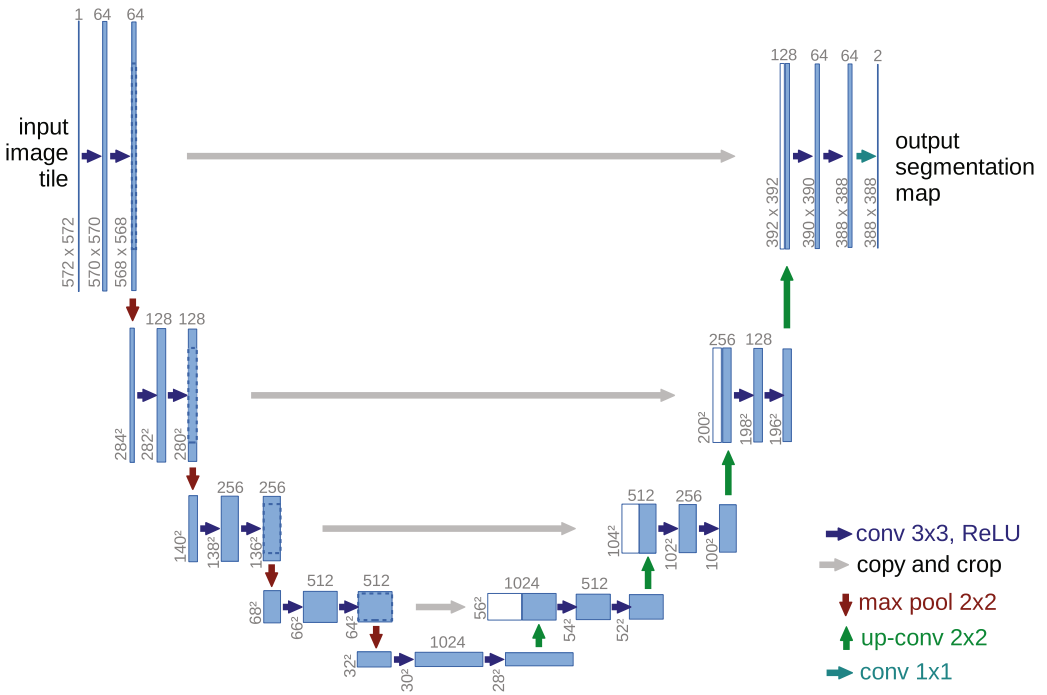

Fig. 2.13: Original U-net architecture. Blue boxes correspond to a multi-channel feature map. The number of channels is denoted on top of the box. The x-y-size is provided at the lower left edge of the box. White boxes represent copied feature maps. The arrows denote the different operations. Figure from Ronneberger et al. (2015), reprinted with permission from Springer.

of feature channels is doubled at each stage of the network's encoder, including the bottom-most layer, to enhance its capacity.

Similarly, each stage of the decoder initiates with an up-sampling operation, and the number of feature channels is halved in each convolutional block. In the original U-Net, up-sampling is accomplished using deconvolutions, enabling the learning of nonlinear up-sampling. The up-sampled feature channels are concatenated with feature maps that come from the encoder path at the corresponding resolution level just before feeding them into the convolutional block. This facilitates the transmission of finer details from earlier stages of the network, thus enabling the generation of a more detailed final result, e.g., a segmentation map as or a generated image as output. The network concludes with a final 1×1 convolution layer that produces the output channels.

2.6 Evaluation Measures

To quantitatively assess the performance of a method that generates images and for the statistical analysis of differences between methods, two image based error measures will be used, namely root mean squared error (RMSE) and mean displacement (MDISP), and one key point based error measure target registration error (TRE). They will be used to express the similarity or dissimilarity between a generated image and the ground truth. In this thesis, methods will generate MRI slices in one of the following two ways. The first way is to find MRI slices with similar breathing state in a larger set of MR images, which are then averaged. The second one will be to predict MRI slices using reference MRI slices in a deep learning approach. The computation time that a method needs for the generation will also be used as a quantitative measure for comparison. The four measures are described in the following.

RMSE

The RMSE between two images, e.g., predicted slice and ground truth is calculated as expressed in Eq. 2.1 by computing the voxel-wise intensity difference d_i and then taking the root of the mean of the squared differences.

$$\text{RMSE} = \sqrt{\frac{1}{W \cdot H} \sum_{i=0}^{W \cdot H} d_i^2}, \tag{2.1}$$

where **W** and **H** are the width and height of the images. It is common practice to report the RMSE in the evaluation of 4D MRI methods. However, the comparability of the measure across works is limited because different image normalization might be used. Moreover, image differences that are due to the appearance or presence of structures on the one hand and the displacements of structures on the other hand both contribute to a change in the measure. However, especially in the case of predicted or synthetic images it might be interesting to differentiate between those to sources of error or dissimilarity. That is, it is one type of error if a method predicts an anatomically incorrect image of the abdomen and it is another error if the method predicts an anatomically correct image of the abdomen but in the wrong breathing state, i.e., particular structures within the anatomy are not correctly located. To overcome this limitation two additional measures are used. Namely the MDISP and the DN_RMSE, described in the following.

MDISP

The mean displacement error (MDISP) quantifies the difference of two MRI slices of the same subject and organ at the same slice position by attributing the difference to deformation and displacement. The MDISP is computed by first performing a B-spline deformable registration using simpleITK (Lowekamp et al., 2013) to obtain a dense deformation field between the two images, e.g., a predicted image and a ground truth image.

The parameterization of the deformable registration algorithm was empirically determined as follows. The ANTSNeighborghoodCorrelation (radius = 2) option was used as the similarity measure. It visually yielded better registrations than the other options MeanSquarse, MattesMutualInformation, and correlation. To make the registration more robust and speed efficient, a pyramid scheme with two levels was utilized. In the first level, the images were smoothed with a sigma of 0.25 before halving their resolution using linear interpolation. In the second level, the original image was used with no smoothing. The grid size of the deformation mesh was 4×4 in the first level. It was doubled to 8×8 in the second level. A gradient descent optimizer (learning rate = 0.25, number of iterations = 20, convergence minimum value = $1e^{-7}$, convergence window size = 10, estimate learning rate = True, maximum step size in physical units = 0.25) was used.

The resulting dense deformation field was then sampled in a 16×16 grid (8×8 voxel spacing) to obtain a sparse displacement field of displacement vectors. The sparse displacement field is masked to omit vectores that originate outside of the liver. For that the target organ (liver) was manually segmented for a breath hold MRI volume for each subject. The segmentation was used as the mask. The average Euclidean norm of the remaining displacement vectors was calculated in mm.

The MDISP is a better measure for comparison across works than the RMSE because the displacement of structures is independent of image normalization. However, the measure has it's limitations. First, the displacement field between a generated image and the ground truth is not always well defined. For example, when a prediction contains structures not present in the ground truth or vice versa when structures are missing. An extreme example is an empty prediction, which would lead to an MDISP of zero, which of course, would not reflect the actual similarity. Another practical limitation is the need to parameterize the deformable registration. In this work the parameterization was choosen to yield the best results for the used data. In other work the optimal parameterization could be different. This dependence on the parameter set makes the comparability difficult again.

DN_RMSE

To alleviate some of the shortcomings of RMSE and MDISP, a new measure is proposed: the deformation-normalized root mean squared error (DN_RMSE). The idea is to compute the RMSE after the image in question is deformably registered to the ground truth using the same pipeline as used for MDISP. In other words, DN_RMSE tries to express the similarity purely based on appearance and the presence of anatomical structures and not on deformation or displacement. It can be used to interpret small MDISP values better. Not unlike MDISP, taken by itself, DN_RMSE is not conclusive. However, combined with MDISP, it aids in a better comparison of generated images within a single work.

TRE

The TRE is a medically crucial metric for accuracy. To compute it, the position of vessel cross-sections is marked in both the generated MRI image and the ground truth. In the presented work this was done manually. For that a self written tool was used that shows both images and allows to place markers in a series of images, making it easier to track the cross sections. Tracking was performed within slice positions. In practice, from one to six vessel cross-sections were tracked, first in the ground truth and after that in other image, e.g., a predicted image. The number of tracked vessels depended on the availability of visible landmarks, i.e., vessel cross-sections in a given liver and slice position. Based on that, the TRE is defined as the mean euclidean distance of corresponding marked positions in the generated image and the ground truth.

3 Related Work

Synopsis

This chapter gives an over view of the related work in the field of 4D MRI methods. A classification is made that divides the relevant work in two types: phase-resolved and time-resolved methods. The unique advantages and disadvantages of both types are discussed.

3.1 Structured Literature Research

A structured literature research was performed using the search engine pubMed. The search was limited to the time period between January 1st 2000 and August 31th 2022 and resulted in 307 initial retrieved papers. The search term was constructed using five categories. To account for different spellings, several synonyms per category were used. (see table 3.1). The categories and synonyms were established after performing an unstructured literature research. The categories represent the following:

A) The method must be a 4D method, i.e., both temporally and spatially resolved.

B) The method must be applicable on large organs or areas.

C) The imaging modality must be MRI.

D) It must be a technical method and not a paper just regarding the application of the method.

E) Some related concepts and work must be excluded because they are out of the scope of this work.

The search term was constructed to fulfill all categories A,B,C, and D and exclude category E. Equation 3.1 shows the outer structure of the search term.

$$A \cap B \cap C \cap D \cup E \tag{3.1}$$

All synonyms within a category were included by concatenating them using the OR operation. The final pubMed readable search term, including its inner structure looks as follows:

((4D[Title/Abstract] OR 4-dimensional[Title/Abstract] OR four-dimensional[Title/Abstract] OR time-resolved[Title/Abstract] OR respiration resolved[Title/Abstract] OR temporal phase-resolved[Title/Abstract] OR respiratory motion-resolved[Title/Abstract] OR respiratory phase-resolved[Title/Abstract])

AND

(abdominal[Title/Abstract] OR liver[Title/Abstract] OR lung[Title/Abstract] OR thorax[Title/Abstract] OR abdomen[Title/Abstract] OR pulmonary[Title/Abstract] OR large FOVs[Title/Abstract])

AND

(MRI[Title/Abstract] OR magnetic resonance imaging[Title/Abstract] OR MR[Title/Abstract] OR magnetic Resonance[Title/Abstract])

AND

(reconstruction[Title/Abstract] OR imaging[Title/Abstract] OR acquisition[Title/Abstract] OR prediction[Title/Abstract] OR approach[Title/Abstract] OR technique[Title/Abstract] OR method[Title/Abstract] OR procedure[Title/Abstract] OR strategy[Title/Abstract])

NOT

(angiography[Title/Abstract] OR flow[Title/Abstract] OR subtracted[Title/Abstract] OR cardiac[Title/Abstract])

AND

(2000/1/1:2022/8/31[pdat]))

The 307 results were further filtered. Any paper that did not present a method for the reconstruction of 4D MRI was omitted from the initial list of papers, leaving a total of 23 papers. Also included in the related work are 5 more papers that did not match the search term but were found during the non-structured literature research and were relevant related work.

Categories				
A	B	C	D	E
4D	abdominal	MRI	reconstruction	angiography
4-dimensional	liver	magnetic resonance imaging	imaging	flow
four-dimensional	lung	MR	acquisition	subtracted
time-resolved	thorax	magnetic Resonance	prediction	cardiac
respiration resolved	abdomen		approach	
temporal phase-resolved	pulmonary		technique	
respiratory motion-resolved	large FOVs		method	
respiratory phase-resolved			procedure	
			strategy	

Tab. 3.1: The categories and synonyms or terms within each category that were used to build the search term for the structured literature research. The term was chosen to include synonyms from the category A to D and to preclude synonyms from category E.

The unstructured and subsequent structured literature research revealed that there exist two approaches to the acquisition of 4D MRI, each with its unique advantages and disadvantages. The first approach is to acquire fast 3D MRI sequences in real-time as done by Kim et al. (2014), Dinkel et al. (2009) and Bled et al. (2011). The main advantage of this approach is that it does not rely on gating and thus supports the imaging of events that do not occur repeatedly, i.e., events that are not periodic. The main disadvantage of this approach is that it typically has either a small FOV, low spatial resolution or inferior image quality. Kim et al. (2014) use a FOV of $20\,\text{cm} \times 16\,\text{cm} \times 8\,\text{cm}$, Dinkel et al. (2009) use a voxel size of $3.1 \times 3.1 \times 4$ and have a bad image quality. This renders this approach incapable to capture the respiratory motion of large organs like the liver in a sufficient temporal and spatial resolution. They are thus not further discussed in this work.

The second approach is to reconstruct volumes for different breathing states or breathing phases of the organ or area of interest in retrospection. This is done by binning previously acquired data. This approach can be further divided into two main types: respiratory phase-resolved or time-resolved methods. They are descried in the following sections. Both types can share similarities. For example, the used surrogate or breathing signal, which is used for the data binning, can be either intrinsic or extrinsic. Intrinsic signals, rely on image or k-space information. Extrinsic signals, are externally recorded, e.g., using a breathing belt or tracking markers that are placed on the abdomen of the subject. Also using a flight of time or depth camera can be counted to extrinsic signals. In the following the two types and relevant works are discussed in detail. A summary of the works can be found in table 3.2.

3.2 Respiratory Phase-Resolved Methods

The first type can be described as respiratory phase-resolved. Methods of that type are mainly based on sequence programming and unique k-space sampling designs. The sparsely sampled k-space data is binned into respiratory phases and each bin then gives rise to the reconstruction of a single volume for that given respiratory phase. The total data acquisition of these methods usually takes around $5\,\text{min}$. The number of phases they can reconstruct is generally fixed (usually 10 or fewer phases) as well is the number of breathing cycles that can be reconstructed is generally restricted to a single averaged breathing cycle.

Tokuda et al. (2008) proposed an adaptive 4D MR imaging method based on navigator echo and multiple gating windows leading to a more efficient acquisition. Cai et al. (2011) acquire axial MRI slices to derive an image based intrinsic breathing signal, which they call body area. They use the signal retrospectively to sort the axial slices into four respiratory phases. Hu et al. (2012) proposed a triggering scheme that consists of a preparation stage and an acquisition stage. Immediately prior to MRI acquisition, the preparation stage monitors the respiration via an external respiratory belt. Based on the respiratory amplitude the respiration cycle is equally divided into N respiratory phases (they showed it for 4 bins). Each phase was associated with a trigger which started image acquisition for a single slice. A complete 4D MRI was acquired by acquirring slices at each position and in each respiratory phase. Yanle Hu et al. (2013) used single-shot acquisition with parallel imaging and partial k-space imaging to improve acquisition speed. They reconstruct four respiratory states of one breathing cycle, the mid point and end point of both inhalation and exhalation in a slice wise manner. Each slice is acquired once in all four breathing states. They use a triggering mechanism for MRI image acquisition based on the respiratory amplitude instead of respiratory phase as in other 4D techniques. The promise a high contrast by using the T2 weighted sequences. Tryggestad et al. (2013) have developed a longer-duration MRI and post processing technique to derive the average or most-probable state of mobile anatomy and meanwhile capture and convey the observed motion variability. They acquire sagittal and coronal slices and derive in in a two-pass approach respiratory interval-correlated volumes, by retrospectively sorting them into ten respiratory phase volumes. However, by using a prolonged acquisition, they lose the normally inherent advantage of this type of methods, its short acquisition time. Y. Liu et al. (2014) proposed 4D MRI construction based sorting of 2D data slices into respiratory phases, useing a sagittal body area surrogate to determine the respiratory phase.

Paganelli et al. (2015) acquire serial interleaved 2D multislice MRI data and use mutual information to automatically determine a stable reference phase. They then sort the image slices retrospectively without the need for navigator frames by directly comparing neighboring slices using mutual information to reconstruct eight breathing phases of one breathing cycle. Deng et al. (2016) implemented a continuous spoiled gradient echo sequence with 3D radial trajectory and 1D self-gating for respiratory motion detection. They sort data retrospectively into different respiratory phases based on their temporal location within a respiratory cycle. Ten phases are reconstructed via a self-calibrating CG-SENSE program. Based on the balanced steady-state free-precession (bSSFP) technique and 3D k-space encoding, Han et al. (2017) designed a novel rotating cartesian k-space

(ROCK) reordering method. It incorporates repeatedly sampled k-space center lines as the self gated motion surrogate, which is used to retrospectively bin the k-space data into different respiratory positions based on the amplitude of the surrogate. Each of the eight k-space data bins is then subsequently reconstructed using a joint parallel imaging and compressed sensing method with spatial and temporal regularization. Rank et al. (2017) proposed the 4D joint motion-compensated high dimensional total variation algorithm, which alternates between motion-compensated image reconstruction and artifact-robust motion estimation at multiple resolution levels. Lindt et al. (2018) acquire axial slices using a dynamic multi-slice 2D Turbo Spin Echo (TSE) sequence. They extract an image-based self-sorting signal by computing correlation coefficients between all acquired slices. Images are then sorted into 10 phases while missing data is interpolated.

Harris et al. (2018) proposed to use a combination of prior 4D MRI volumes, which are acquired using a retrospective approach as prior information and kV imaging (x-ray) of a linear accelerator (LINAC) system, which is acquired in real time during the intervention, to generate on-board 4D MRI prospectively and in real time. To that end, they determined a end exhale volume from the prior 4D MRI and computed a synthetic CT volume from that. The CT volume was then registered, using deformable transformation and projection of the volume, to the real time x-ray. Finally the on-board 4D MRI is generated by deforming end prior exhale MRI volume according to the optained deformation field. Meschini et al. (2019) exploit a k-medoids clustering approach to sort 2D MRI into 4D MRI. Data slices were sorted using multiple automatically tracked internal landmarks using the scale-invariant feature transform (SIFT) without using a separate navigator image. Richter et al. (2020) implemented a wave-CAIPI k-space trajectory in a respiratory self gated 3D spoiled gradient echo pulse sequence. Trajectory correction applying the gradient system transfer function was used, and images were reconstructed using an iterative conjugate gradient SENSE algorithm. Navest et al. (2020) proposed to use a so called noise navigator as respiratory surrogate signal for 4D-MRI generation. It is based on the respiratory-induced modulation of the thermal noise variance measured by the receiver coils during acquisition and thus is inherently present and synchronized with MRI data acquisition. This eliminates the need for acquisition of an actual navigator frame. The k-space data was binned into ten equally sized respiratory phases using phase binning. Kavaluus et al. (2020) acquired a proprietary T2-weighted single-shot fast spin echo research sequence. The respiratory surrogate signal was observed within a linear navigator interleaved with the anatomical liver images. The navigator was set on head-feet direction on the superior surface of the liver to detect the edge of diaphragm. The navigator signal and 2D liver image data were

retrospectively sorted into 4D MRI using the position of the diaphragm seen in the navigator.

Yang et al. (2020) developed a method that uses the diaphragm in sagittal slices as an anatomic feature to guide the sorting of axial slices into ten breathing phases. Initially, both abdominal 2D sagittal cine MRI images and axial MRI images were acquired. The sagittal cine MRI images were divided into 10 phases as ground truth. Following that, the phase of each axial MRI image is determined by matching its diaphragm position in the intersection plane to the ground truth cine MRI. Then, those axial images with matched phases were sorted into 10-phase bins, which were identical to the ground truth cine images. 10-phase 4D-MRI were reconstructed from these sorted axial images. Eldeniz et al. (2021) developed a deep learning method to remove streaking artifacts and noise from free-breathing magnetic resonance imaging using a radial acquisition and k-space undersapmling. The method does not rely on high-quality ground truth. Self-navigation was used to bin k-space data into 10 respiratory phases. Short acquisitions were simulated by using subsets of radial spokes to reconstruct images with multicoil nonuniform fast Fourier transform (MCNUFFT), compressed sensing (CS). They developed a deep leaning network Phase2Phase that is trained to remove artifacts from the simulated low quality 4D MRI images. Keijnemans et al. (2022) developed a hybrid 2D/4D-MRI methodology that uses a simultaneous multislice accelerated MRI sequence, which acquires two coronal slices simultaneously and repeatedly cycles through slice positions over the image volume. Slices are sorted retrospectively into respiratory-correlated 4D-MRIs using an intrinsic end-exhale reference. Li et al. (2022) proposed a novel motion-aligned reconstruction method based on higher degree total variation and locally low-rank regularization to recover 4D MR images from the highly undersampled Fourier coefficients. They proposed a two-stage reconstruction framework alternating between a motion alignment and regularized optimization reconstruction, presenting a unified framework to exploit the spatial and temporal correlation of the 4D-MRI data. A fast alternating minimization algorithm based on variable splitting was utilized to solve the optimization problem efficiently.

In summary, one can note that the strength of respiratory phase-resolved methods, lies in the ability to capture periodic breathing state changes with a large FOV within a few minutes depending on the length of the motion cycle. Its weaknesses are its assumption of strictly periodic organ motion. In reconstructing only one average or dominant motion cycle of the target organ, which is not ensured to be physiologically meaningful, it cannot account for arbitrary or irregular breathing and thus irregular breathing cycles. Furthermore, this type introduces image artifacts Mickevicius et al. (2017) and Pang et al. (2016) that could hinder motion estimation

from the reconstructed 4D MRI. This however was addressed by Eldeniz et al. (2021). Another disadvantage is that it requires the MRI machine to allow the use of the non-standard sequences on with those methods often rely.

3.3 Time-Resolved Methods

4D MRI methods of this type can reconstruct variable and irregular breathing motion and are mainly based on clinically available MRI sequences. They often utilize a fast dynamic 2D sequence to acquire images at different slice positions, which in total cover the organ of interest. After that most methods apply retrospective gating to the acquired 2D image slices, to bin them by different breathing states of the organ and sort them based on their slice position in their respective volumes.

In 2007, M. v. Siebenthal et al. (2007) were the first to propose a time-resolved 4D MRI reconstruction framework suitable to account for arbitrary breathing motion of the liver. They acquired two sets of fast 2D MRI image sequences. First, a data set of images, which subsequently was sorted to form a 4D volume and, second, a dynamic reference sequence, which was used to sort the other images by and that showed the course of the respiratory movement of the liver within a navigator slices position. The first set was a series of 2D images alternating between spatially fixed navigators and spatially moving data slices. Using a similarity search strategy, these were then reconstructed into a 4D MRI that corresponds to a given sequence of navigators with respect to the breathing state. Thus, they reconstructed time-resolved 4D MRI from dynamic 2D navigators, accounting for arbitrary breathing. The shortcoming of this method was the long acquisition time needed to establish the data set in which the similarity search is performed and the time-expensive search during reconstruction itself.

Several works adapted the idea and tried to address the shortcomings within the framework, i.e., the long acquisition and reconstruction times. They applied machine-learning methods to interpolate navigators or data slices, effectively reducing acquisition time or improving image quality.

Wachinger et al. (2012) proposed a purely image-based retrospective respiratory gating method. They created 4D MR within the Siebenthal-framework but using a navigator based on Laplacian eigenmaps, a manifold learning technique, to determine the low-dimensional manifold embedded in the high-dimensional image space. This made the sorting more robust against ouf-of-plane motion of liver structures, which before posed a problem to the template based approach of M. v.

Siebenthal et al. (2007). Tanner et al. (2014) addressed the long acquisition times and propose to actively generate more suitable data images instead of only selecting from the available images that were actually acquired. Thus they can either reduce the number of data slices that actually actually have to be acquired or increase the image quality by boosting the number of reference images. The method is based on learning the relationship between the motion of navigator and data-slice by linear regression after dimensionality reduction. They use this to predict new data slices for a given navigator by warping existing data slices by their predicted displacement field. Among the approaches within the Siebenthal-framework, the method of Tanner et al. has the most similarities to the deep learning based 4D MRI method presented later in this thesis. This is because it is based on learning the relation between navigator and data slices. The essential difference is that their method is not an end-to-end learnable formulation for the 4D MRI reconstruction, i.e., the machine learning technique does not solve the complete problem of 4D reconstruction but rather a part of it. Thus the method still requires a time-expensive search for similar data slices within the prior acquisitions, making the method suited for retrospective reconstruction but not for prospective reconstruction.

Celicanin et al. (2015) also addressed the acquisition time, by simultaneously acquiring navigator and data slices, cutting the total acquisition time in half. They used a standard balanced steady state free precession sequence and modified it to simultaneously acquire two superimposed slices with different phase cycles, i.e., an image and a navigator slice. Instead of multiband RF pulses, two separate RF pulses were used for the excitation. Images were reconstructed using offline CAIPIRINHA reconstruction.

Karani et al. (2018) also addressed the long acquisition times and proposed the temporal interpolation of navigator slices using a CNN. They used that to effectively half the number of navigator acquisitions by replacing every second navigator with an interpolation. They showed that an extension of the temporal context from T=1 to T=2, improved the interpolation result, however, a further extension did not result in further improvements. Zhang et al. (2018) addressed both the long acquisition time as well as the reconstruction time by expanding on the idea of Karani et al.. They proposed the re-formulation of the temporal interpolation idea using the prediction of a motion field as an intermediate step, reducing the problem of blurry predictions and missing structures. Another advantage of their formulation was that it provides an unsupervised estimation of bi-directional motion fields that can be used to halve the number of registrations required during 4D reconstruction. That way they also reduced the reconstruction time. Qiu et al. (2019) proposed a method that incorporates Sliding Motion Compensation into the standard Low rank

+ Sparse reconstruction. The global superior-inferior displacement of the internal abdominal organs is inferred directly from the undersampled raw data and then used to correct the breathing induced sliding motion. The reconstructed temporal frames are roughly registered before applying the standard Low rank + Sparse decomposition.

Romaguera et al. (2019) proposed a fully automatic self-sorting 4D MRI volume construction method outside of the Siebenthal-framework. They used a 2D T2-weighted true FISP sequence, first extracting a pseudo-navigator signal for each 2D dynamic slice acquisition series. Then, a weighted graph was created using both spatial and motion information provided by the image based pseudo-navigator to ensure the temporal coherence of the reconstruction. The volume at a given time point was reconstructed following the shortest paths in the graph starting at the time point of a reference slice chosen based on its pseudo-navigator signal. Yuan et al. (2019) proposed a time-resolved large FOV 4D MRI reconstruction technique, which also does not operate within the Siebenthal-framework. It is based on sequence programming to shorten MRI acquisition times drastically. It attains high temporal resolutions (615ms) at moderate spatial resolution ($128 \times 128 \times 56$ voxels, $2.7 \times 2.7 \times 4.0\text{mm}^3$). The method is still a retrospective solution, because the volume reconstruction takes around $20\,\text{s}$. Also, the authors state that the huge amount of captured data (91 MR images/s) risks filling up the scanner's memory during longer imaging sessions.

In summary, time-resolved methods have three advantages. First, most of them relaxs the strict constraint of periodicity of the breathing motion, to a degree that quasi-periodic and even non-periodic changes in the organ can be captured. Although, an event or breathing state still has to occur multiple times to be reliably captured. The only exception here is the method of Yuan et al. (2019). The second advantage is the high temporal and spatial resolution. Hence these methods are well-suited to capture motion variation, e.g., deep or shallow, abdominal or thoracic breaths within one session with a sufficient spatio-temporal resolution. This method type can work with a time-resolved navigator or respiratory signal to ensure physiological correctness of reconstructed motion. The third advantage is its availability. Most methods work with all MRI machines and all standard clinically available 2D MRI sequences that are fast enough. The disadvantages of this type of method are that it is generally more time-intensive. The data acquisition takes up to $60\,\text{min}$. Also, none of the methods presented so far are real-time capable due to long reconstruction times of tens of seconds. Furthermore, a significant portion of the acquired data is often redundant. This, however, can advantageously be used to increase

the SNR of the reconstructed 4D images in classical approaches or boost the data basis for machine learning approaches.

	TR	P/R	Matrix size	Resolution in mm^3	Breath. cycle smpl.		vps		Recon. time in s/vol.	befAcq in min	RMSE median (95%)
					P	R	P	R	-		
Tokuda et al. (2008)	no	R	256x128x24	-	-	5	-	-	-	18	-
Cai et al. (2011)	no	R	256x166	1.5x1.5x5	-	4	-	-	-	-	-
Hu et al. (2012)	no	P	-	-	10	-	-	-	-	-	-
Yanle Hu et al. (2013)	no	R	250x176x32	1.5x1.5x5	-	4	-	-	-	3	-
Tryggestad et al. (2013)	no	R	175x190x9	2x2x5	-	10	-	-	-	13	-
Y. Liu et al. (2014)	no	R	256x166	2.5x2.5x5	-	10	-	-	-	-	-
Paganelli et al. (2015)	no	R	256x224x20	1.28x1.28x5	-	8	-	-	-	1.2	-
Deng et al. (2016)	no	R	-	-	-	10	-	-	-	8	-
Han et al. (2017)	no	R	416x250x125	**1.2x1.2x1.6**	-	8	-	-	75	5	-
Rank et al. (2017)	no	-	256x256x60	1.5x1.5x5	-	20	-	-	22.5	0.7	-
Lindt et al. (2018)	no	R	138x208x30	2x2x5	-	10	-	-	30	5	-
Harris et al. (2018)	no	P	-	1.67x1.67x1.67	10	-	-	-	-	-	-
Meschini et al. (2019)	no	R	256x224x20	1.28x1.28x5	-	8	-	-	262	1.2	-
Kavaluus et al. (2020)	no	R	-	1.33x1.33x3	-	8	-	-	-	15	-
Richter et al. (2020)	no	R	224x224x144	2.24x2.23x2.23	-	8	-	-	11	10	6.51
Navest et al. (2020)	no	R	-	-	-	10	-	-	-	-	-
Yang et al. (2020)	no	R	-	1.67^1.67x5	-	10	-	-	-	-	-
Eldeniz et al. (2021)	no	R	318x318x96	1.13x1.13x3	-	10	-	-	**2.7**	5	-
Keijnemans et al. (2022)	no	R	52x240x136	6.7x1.9x1.9	-	8	-	-	-	3	-
Li et al. (2022)	-	-	256x256x25	1.37x1.37x4	-	-	-	-	-	-	-
Wachinger et al. (2012)	no	R	-	-	-	-	-	-	-	-	-
M. v. Siebenthal et al. (2007)	**yes**	R	192x192x25	1.8x1.8x4	-	36	-	5	73	60	-
Tanner et al. (2014)	**yes**	R	224x224x53	1.3x1.3x5	-	36	-	4.4	-	10	-
Celicanin et al. (2015)	**yes**	R	120x128	1.87x1.87x6	-	20	-	3.33	-	-	-
Zhang et al. (2018)	**yes**	R	-	1.33x1.33x5	-	36	-	2.4	36.5	30	10.23 (13.74)
Karani et al. (2018)	**yes**	R	-	1.33x1.33x5	-	36	-	2.4	-	20	**4.09 (6.81)**
Romaguera et al. (2019)	**yes**	R	-	1.7x1.7x3	-	**46**	-	**7.5**	-	28	-
Qiu et al. (2019)	**yes**	R	256x256x53	1.34x1.56x4	-	-	-	1.6	-	-	-
Yuan et al. (2019)	**yes**	R	128x128x56	2.7x2.7x4	-	9.78	-	1.63	20	**0.33**	-

Tab. 3.2: Comparison of the related work regarding whether its time-resolved (TR), whether reconstruction is done pro-/retrospectively (P/R), matrix size, voxel resolution, how many phases of a breathing cycle can be resolved (breath. cycle smpl.) based on a 6 s breathing cycle, volumes per second (vps) in pro- and retrospective reconstruction (P/R), prior acquisition time (befAcq), reconstruction time, and RMSE. Values taken from respective publications. Best values bold.

3.4 Summary

To summarize, let us revisit the advantages and limitations of phase-resolved and time-resolved methods that also were anticipated in the introduction.

phase-resolved methods

pro short acquisition times of around $5\,\mathrm{min}$

pro high spatial resolution

con single breathing cycle

con no accounting for irregular breathing

con fixed number of breathing phases (<= 10)

con not readily usable in all MRI sites (sequence programming)

con no real-time capability, i.e., no live imaging

con low image quality

con no uncertainty or error estimation

time-resolved methods

pro multiple breathing cycles

pro accounting for irregular breathing

pro high temporal resolution

pro high spatial resolution

pro high image quality

pro use clinically available MR sequences

con long prior acquisition time (»10min)

con not real-time capable, i.e., no live imaging

con no uncertainty or error estimation

In a direct comparison the time-resolved methods have more advantages than limitations compared to the phase-resolved methods. For this reason a time-resolved method was developed during the work presented in this thesis, in which the following three limitations have to be addressed:

1. long prior acquisition times
2. long reconstruction times, i.e., no real time capability
3. no uncertainty or error estimation

4 MRI Data Base for 4D MRI Reconstruction

Synopsis

This chapter reports the establishment of a public MRI database for developing and testing 4D MRI methods. The MRI data, study information, and MR sequence protocols used in this study are available in the Open Science Repository for Research Data and Publications of OVGU (Creative Common License 4.0) in part one: https://doi.org/10.24352/UB.OVGU-2019-093 *and part two:* https://doi.org/10.24352/UB.OVGU-2021-071.

About this chapter

Parts of this chapter have been published in: Gino Gulamhussene, Fabian Joeres, Marko Rak, Maciej Pech, and Christian Hansen (2020). "4D MRI: Robust sorting of free breathing MRI slices for use in interventional settings". PloS one, 15(6), e0235175. (Gulamhussene et al., 2020).

4.1 Introduction

Previously there has been no publicly available data set for the development and testing of 4D MRI methods. At the beginning of this thesis such a data set was generated and made publicly available for fellow researchers to develop and test their methods. This chapter describes the generation of a public data set. First, a general motivation of the data requirements is given and the resulting study design is described. Then section 4.2 describes what exact data was acquired and what the structure of the data is like. Section 4.3 describes both how the data was acquired and the MRI sequences that were used.

4.1.1 Data Requirements

The data set had to fulfill some requirements. First, the contained MR images must mimic the interventional image data, which is available during an interventional setup, described in 2.3 as closely as possible. This specifically means high acquisition speed and a contrast that is just good enough to detect the respiratory motion. Second, to make the data set suitable for the development and testing of a variety of 4D MRI methods that are compatible with a wide range of also external surrogates, no body array coil (surface array coil comprised of multiple elements) but only the bore fixed receiver coil was used, to ensure a free line of sight to the abdomen of the subject. This made the acquisition compatible with, for example, surrogates based on a scan of the abdomen's surface or marker tracking on the abdomen. This was important to make the gathered motion information available for a wide range of interventional scenarios where different surrogates may be used to track breathing.

4.1.2 Study Design

The data acquisition was carried out in two acquisition periods. The first ran from February to June 2018. The second was started and completed within November 2020. Healthy subjects were invited to participate in a MRI study to acquire images of the liver. Subjects had to fulfill all requirements to participate. For example, subjects with tattoos, non-removable piercings, braces, or metal implants were excluded from the study. 13 healthy subjects participated in the first round. Further seven healthy subjects participated in the second round. For each of the 20 subjects

two image sets were acquired. Over the period of several days three acquisition sessions per day were scheduled. The volunteers were invited and imaged on two different acquisition sessions on different days, to include variations that occur in between imaging sessions.

The ethics board of the Otto-von-Guericke-University Magdeburg/Germany approved our study "Studies with healthy subjects in 3 Tesla for methodological development of MRI experiments" (approval number 172/12), concluding that there were no ethical concerns and that this approving assessment was made based on unchanged conditions. All research was performed in accordance with relevant guidelines and regulations. Verbal and written informed consent was obtained from all subjects.

The data was made publicly available in the Open Science Repository for Research Data and Publications of the OVGU (Creative Common License 4.0) in two parts. Part one: https://doi.org/10.24352/UB.OVGU-2019-093 and part two: https://doi.org/10.24352/UB.OVGU-2021-071. The images were anonymized and uploaded in the DICOM image format. A detailed MRI acquisition protocol, and description of the high level data structure was attached to the data.

4.1.3 Study Protocol

Each acquisition session for a specific subject followed the same study protocol:

1. Determination of contraindications, like tatoo, piercing, etc.
2. Subject education
3. Obtaining verbal and written Consent
4. Subject instruction (Duration of the session, no special breathing, i.e., free breathing and no breath commands)
5. Positioning of the subject on the MRI table
6. Get subjects out of the MR room
7. Ask about well-being

4.2 Data Structure

An overview of the general structure of the data set is given in Fig. 4.1 and an overview of the data set on a subject level is given in Fig. 4.2. The data set comprises 20 subjects. For each subject two image sets were acquired on different days. Each of the two image sets per subject contains three parts (see gray boxes in Fig. 4.2):

1. two static 3D liver MRI
2. several data sequences (between 38 and 61)
3. two reference sequences

3D volumes and 2D slices of the same subject share common scanner coordinates. Throughout this thesis the reference sequence and navigator frames will be depicted as orange and the data frames as purple. All three parts to the data base are described in the following.

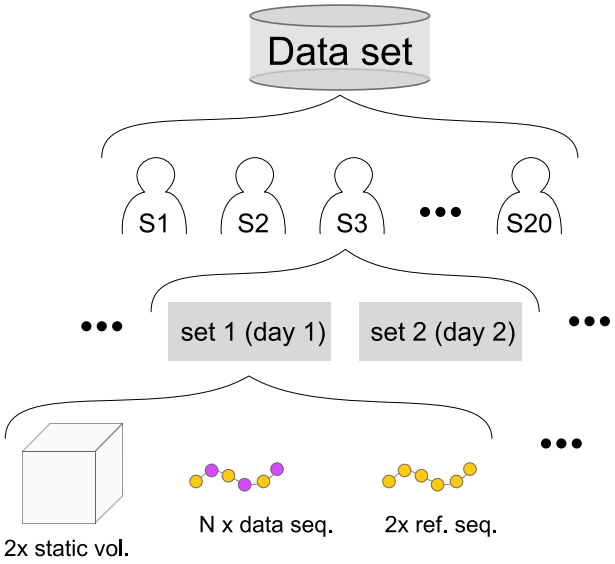

Fig. 4.1: Data set structure: The data set contains data of 20 subjects. Two image sets belong to each subject. Each image set contains two static volumes, several (N) data sequences and two reference sequences. Data sequences contain alternating navigator (orange dots) and data slices (purple dots). Reference sequences contain only navigator slices.

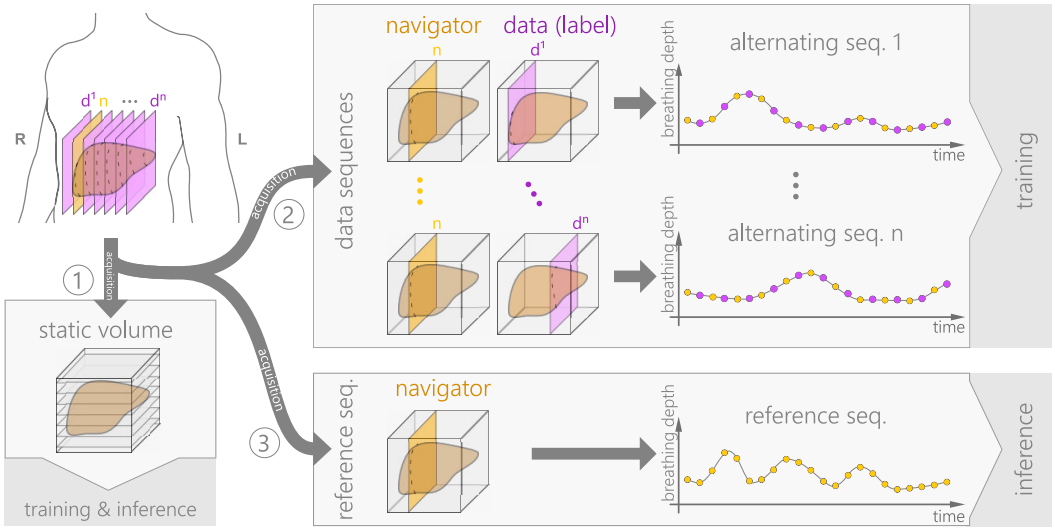

Fig. 4.2: Image set structure and acquisition: Each image set (two sets per subject) consists of three parts: 1) two static volumes, 2) several data sequences, alternating between navigators (orange dots) and data slices (purple dots), and 3) two dynamic reference sequences.

4.2.1 Static Volume

For each subject two static 3D volumes were acquired using a STAR VIBE MRI sequence with axial slice orientation. The sequence itself is described in more detail in section 4.3. Exaples of four subjects are given in Fig. 4.3. Because it is a static volume, it will be also refer to as breath-hold volume. Note, that technically this sequence is not acquired under breath-hold. The subjects are breathing freely during the STAR VIBE volume is acquired. However, this single volume mimics a breath-hold acquisition by binning k-space data into breathing states and reconstructing only one dominant breathing state. The two STARE VIBE volumes were acquired before and after the data sequence and reference sequence.

4.2.2 Reference Sequence

A reference sequence is a dynamic 2D MRI sequence of so-called navigator frames, which is acquired during free breathing. Note, MRI sequences that are dynamic are also refereed to as cine. In this case a TRUFI MRI sequence, described in section 4.3, was used, because it allows for a fast slice acquisition of 166 ms per slice. A schematic depiction can be found in Fig. 4.2 in the gray box 3). Example images for three subjects can be seen in Fig. 4.4. The reference sequence is

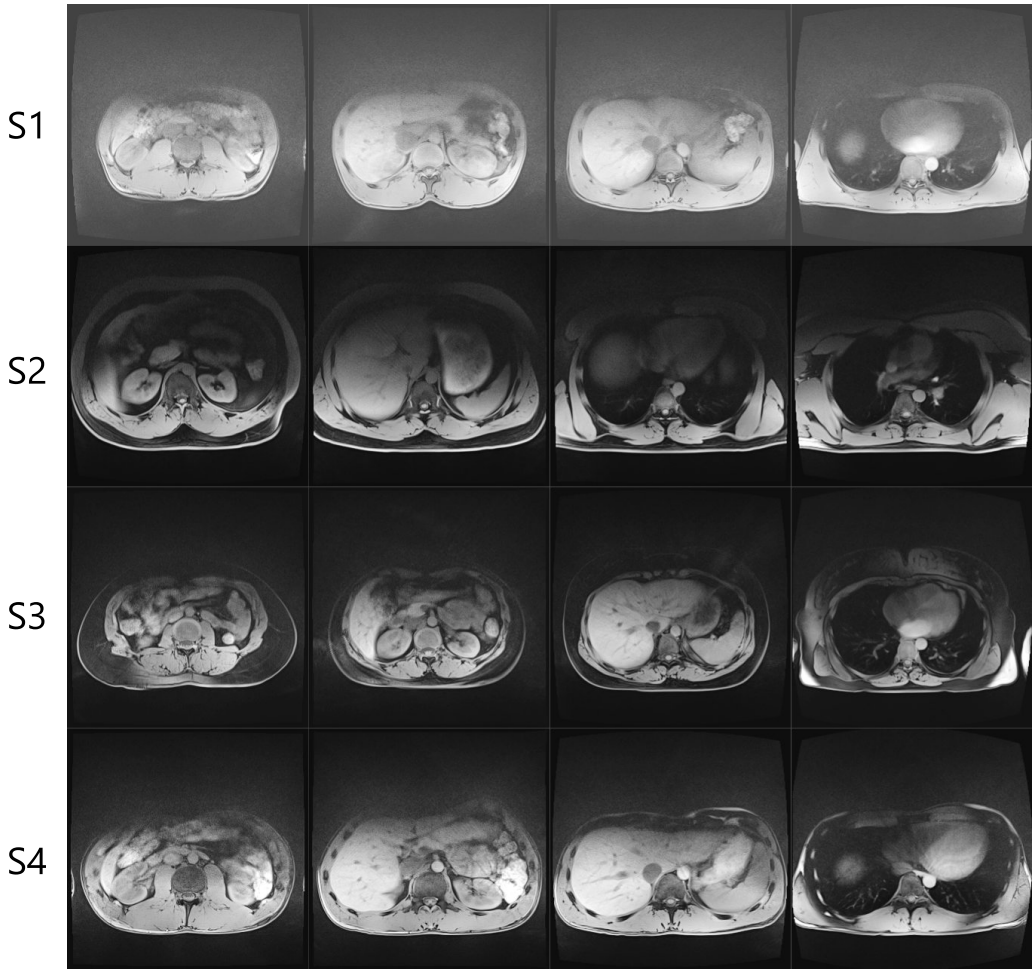

Fig. 4.3: Examples of STAR-VIBE volumes shown in four axial slices of four subjects (S1 - S4).

dynamic in time but static in position, i.e., the slice position does not change. It is the same fixed position as the one of the navigator in the data sequences. Navigator frames represent an image plane, in which the respiratory motion is visible via organ deformation and the positional change of vessel cross-sections. In this case, the navigator is a sagittal slice that intersects the right liver lobe.

This sequence is used for the 4D reconstruction, as a respiratory reference, i.e., a breathing signal or surrogate signal, because it contains a natural succession of different breathing states and pattern that depend for example on shallow or deep, thoracic or abdominal breathing, and is thus physiologically meaningful. The reference sequence represents the real-time imaging of interventional MRI and is used as breathing surrogate for the 4D MRI method. Like the static volume, the first reference sequence was acquired at the beginning and the second at the end

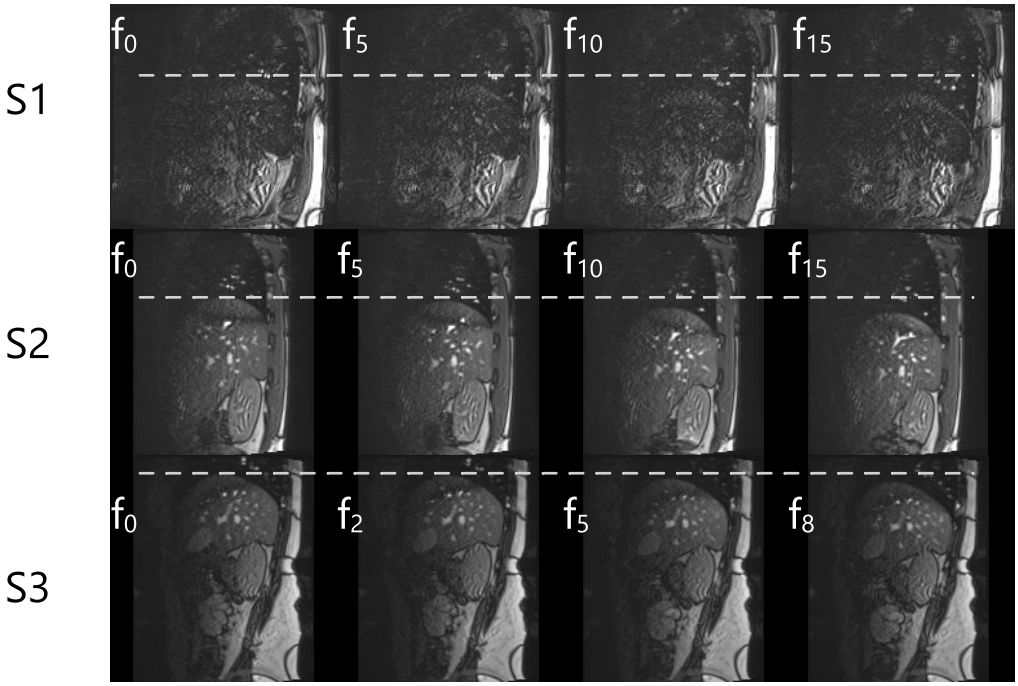

Fig. 4.4: Examples of TRUFI reference sequences of three subjects (S1 - S3). Shown are four frames representative of one inhale. The relative frame number is shown in the left upper corner of each image, e.g., as f_5. The dashed line serves as reference for the breathing motion.

of each subject session. A reference sequence comprises 513 images, i.e., time points, covering a time span of 85 seconds. This is about 20 breathing cycles. The sagittal position of the navigator slice is the same for all sequences per subject.

4.2.3 Data Sequence

A data sequence consists of navigator frames and data frames that were acquired alternatingly, i.e., in an interleaved fashion. It was also acquired during free breathing, using the TRUFI MRI sequence. An example of a data sequence of one subject is given in Fig. 4.5. The data sequence shows a different breathing curve than the navigator sequence, because it was imaged at another time. However, it contains similar or the same breathing patterns, albeit in a different order of succession.

Each data sequence consists in equal parts of data frames and navigator frames (between 150 and 200 each), see gray box 2 in Fig. 4.2. The acquisition of one slice took $166\,\mathrm{ms}$ per slice. Each data sequence was acquired for $1\,\mathrm{min}$ before moving the imaging plane of the data slice $4\,\mathrm{mm}$ to the left, while keeping the

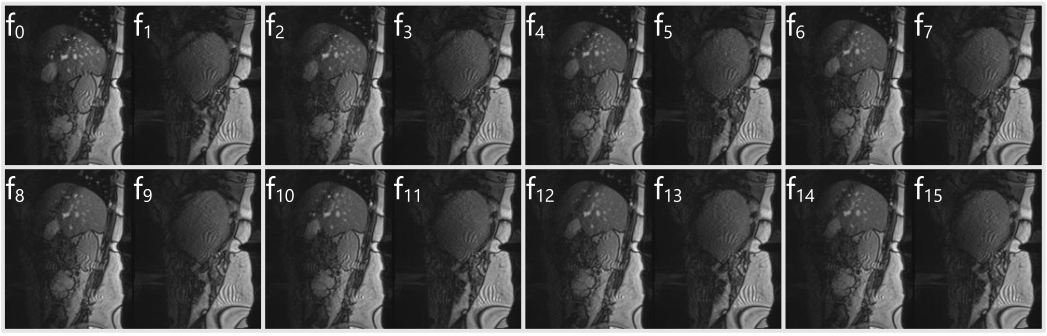

Fig. 4.5: Examples of TRUFI data sequences of one subject. Shown are 16 consecutive frames. To signify interleaved character of the sequence pairs of navigators and data frames are framed with gray boxes. The relative frame number is shown in the left upper corner.

navigator position fixed. This is repeated until the entire liver is covered. Keeping the navigator position exactly the same as in the reference sequence, renders temporal reconstruction possible. Sweeping the data slices over the target organ in 4 mm gaps during acquisition, renders spatial reconstruction possible. In other words, while the navigator position is the same for all data sequences, the slice positions for the data frames are distinct for each data sequence. Data sequences were acquired in the order from right to left. For each slice position of the reconstructed volume, i.e., at 4 mm distances, one data sequences is acquired. The total number of data sequences per subject ranges between 38 and 57 (mean = 46.68), depending on the size of the subjects' liver to capture its whole extent. Thus, the overall acquisition time for a subject ranged between 40 and $80\,\mathrm{min}$, plus $\sim 15\,\mathrm{min}$ per subject for imaging localizers, determining the navigator position and setting up the interleaved sequences.

In the sorting approach to 4D MRI, which will be discussed in chapter 5, the 4D MRI method sorts the data frames into a 4D MRI volume based on information extracted from the navigators that come befor and after the data frame.

In the deep learning approach to 4D MRI, which will be discussed in chapter 6, the entire data sequence can be thought of as made up from a series of pairs or training samples, i.e., navigator and data slices. The navigator slices will be part of the networks training input, and the data slices will be the training label images. Because the ground truth is an important part of the data in the deep learning approach, it and its limitations are discussed in a bit more detail here. The ground truth data is contained in the data sequence. Specifically, the data slices will be used in all relevant tests as the ground truth, regarding the the task: given navigator frame N what is the appropriate data frame D for slice position P? That means

the data set does not contain a ground truth for a whole volume as one 3D image. Rather for the full volume the ground truth is divided into multiple slices that are not coherent in time. Because of that the ground truth must be handled separately for each slice position. That is also the reason why this ground truth does not support the testing or training of a method directly on the whole reconstructed volumes but only for each slice position individually.

Note, M. v. Siebenthal et al. (2007) noted that a better acquisition scheme for the data sequence exist. Instead of first acquiring all data slices in one position before moving on to the next data slice position, it is beneficial to move the slice position after each acquisition while keeping the navigator position fixed. This has advantages, which are discussed later. However, the Siemens MRI machine used for this study did not allow for this acquisition scheme with on board software. A separate software, the SIEMENS Healthineers Access-I Framework, like done by Alpers (2023) in another context, or alternatively sequence programming is needed to do that.

4.3 MRI Sequences

The whole data set was acquired on a MAGNETOM Skyra MRI scanner (Siemens Medical Solutions, Erlangen, Germany). In this work the Philips based MRI sequences proposed by M. v. Siebenthal et al. (2007) were translated to their equivalent on a Siemens machine.

STAR-VIBE

Parameter	value
matrix size	$320 \times 320 \times 72 - 88$
slice thickness	3 mm
in plane resolution	$1.19 \times 1.19\,\text{mm}^2$
phase oversampling	0%
slice oversampling	44.4%
FOV read	380 mm
FOV phase	100%
TR	2.83 mm
TE	1.48 ms
flip angle	$9°$
slice partial Fourier	$\frac{7}{8}$

Tab. 4.1: Parameters used for the STAR-VIBE sequence in this work

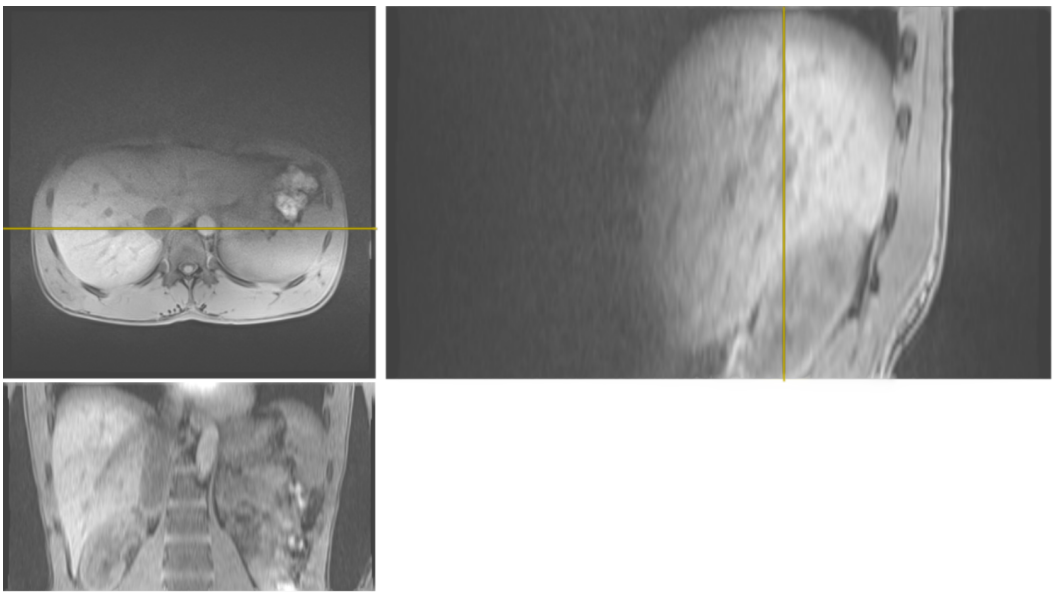

Fig. 4.6: An example liver volume from one subject, acquired using a STAR-VIBE MRI sequence. The liver tissue appears bright in contrast to the vessel-cross sections, which appear dark.

As can be seen in Fig. 4.6, the liver tissue in a STAR-VIBE volume appears bright, while the vessel-cross sections appear dark. In general major vessels are visible, smaller vessels are harder to identify. When the subject did breath more deeply during the STAR-VIBE acquisition, the volumes are more blurry. In these cases finer vessels are not visible.

The parameters used in this work for the STAR VIBE sequence can be found in Tab.4.1. No body array coil was used, limiting acquisition to the bore's fixed receiver coil. The acquisition of one volume took between one and two minutes.

TRUFI

As can be seen in Fig. 4.7, the liver tissue appears to be dark in contrast to the bright vessel cross-sections. The sequence parameters are given in Tab. 4.2. For faster measurement, a partial Fourier was used sampling 5/8 of the k-space asymmetrically in phase-encoding direction, i.e., roughly 60% of the k_y lines, resulting in 88 actually acquired k_y lines. This resulted in an acquisition time of $166\,\mathrm{ms/slice}$. No body array coil was used, limiting acquisition to the bore's fixed receiver coil.

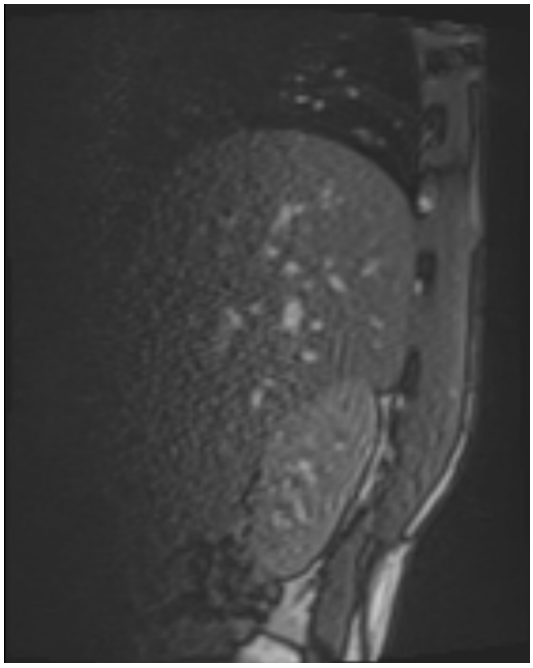

Fig. 4.7: An example liver slice from one subject, acquired using a TRUFI MRI sequence. The liver tissue appears dark in contrast to the bright vessel-cross sections.

Parameter	value
matrix size	140 x 176
slice thickness	4mm
in plane resolution	1.82mm x 1.82mm
FOV	255 mm x 320 mm
TR	39.96 ms
TE	1.49 ms
echo spacing	3.33 ms
flip angle	30°
slice partial Fourier	5/8
readout bandwidth	676 Hz/px
base resolution	176 k_x
phase resolution	80

Tab. 4.2: Parameters used for the TRUFI sequence in this work

5

4D MRI: Robust Sorting of free Breathing MRI Slices for use in Interventional Settings

Synopsis

This chapter reports on the development and evaluation of a 4D MRI method, which serves as a baseline for further development in this thesis. The main strengths of the method are its large FOV, high image quality and its ability to capture irregular breathing motion. It is based on the sorting and stacking approach and uses template updates and search regions for faster and more robust vessel cross-section tracking in the navigator slices in the presence of out-of-plane motion.

About this chapter

Parts of this chapter have been published in: Gino Gulamhussene, Fabian Joeres, Marko Rak, Maciej Pech, and Christian Hansen (2020). "4D MRI: Robust sorting of free breathing MRI slices for use in interventional settings". PloS one, 15(6), e0235175. (Gulamhussene et al., 2020)

5.1 Introduction

The previous chapter described the acquisition and establishment of a public data set for the development and testing of 4D MRI methods. This chapter describes the development of a 4D MRI reconstruction method based on the classical sorting and stacking approach using the data set. The method will serve as the image quality baseline for the development of the deep learning based method described in chapter 6.

The recent years have seen the introduction of 4D MRI methods (as discussed in chapter 3), however, none of these methods have met all the necessary requirements for interventional use as defined in chapter 1 section 1.2, nor can they serve as a strong baseline method for this thesis. Specifically, breathing phase resolved methods lack physiological correctness, while retrospective sorting and stacking approaches that are time-resolved lack robustness against out-of-plane motion. This work follows the retrospective sorting approach because, as highlighted in the related work chapter (3), it is the only method capable of capturing physiologically meaningful, non-periodic organ motion with high temporal and spatial resolution and large field of views. While the approach has a disadvantage of long acquisition time and susceptibility to out-of-plane motion, these can be overcome as demonstrated in this and following chapters. Our work builds specifically upon the method proposed by von Siebenthal et al. (M. v. Siebenthal et al., 2005; M. v. Siebenthal et al., 2007) and utilizes retrospective sorting of dynamic 2D TRUFI MRI slices. It is capable of imaging the whole liver during free breathing and capturing organ motion and deformations caused by respiration. It reconstructs a physiologically meaningful sequence of respiratory states by utilizing a dedicated navigator frame.

Our methodological contribution is, to introduce template updates and search regions to the sorting and stacking approach, to improve robustness against out-of-plane motion and to enhance the reconstruction speed.

5.2 Materials and Methods

This section, first describes the general concept of 4D MRI reconstruction used by von Siebenthal and the proposed adapted method (section 5.2.1). Then in section 5.2.2 the determination of the breathing state is described. After that, the sorting of data slices based on the breathing state is explained in sections 5.2.3. Finally section 5.2.4 describes how the method was improved by the utilization of

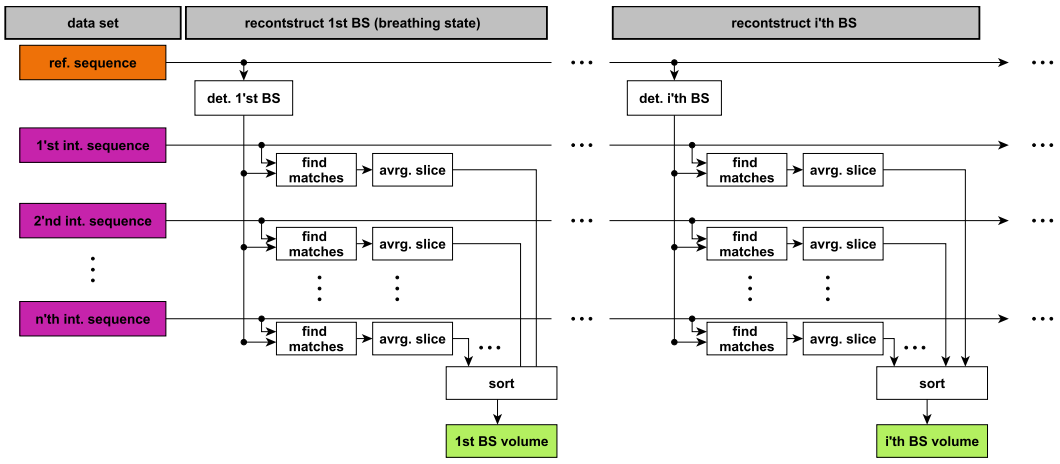

Fig. 5.1: Sorting and stacking scheme of 4D MRI reconstruction.

template updates to increase robustness against out-of-plane motion and search regions to speed up the sorting process.

5.2.1 4D MRI Reconstruction

The last chapter, already described the reference sequence containing navigator frames. The general idea of the method, described in this chapter is, to use the navigator frames to generate a 3D volume for every navigator frame. Each navigator frame defines a breathing state and from that, the method generates a volume with the corresponding breathing state. Repeated for every navigator frame in the reference sequence, this results in a 4D MRI sequence showing a physiological progression of respiratory states.

Figure 5.1 illustrates the scheme of volume reconstruction per navigator frame in detail. The scheme consists of four steps as follows. First, the breathing state of a navigator frame is determined (see section 5.2.2). Second, a search is performed in every data sequence to find all data frames that match this breathing state. The matching criterion used in the first two steps is described in following sections. Third, all found data frames from the same data sequence, i.e., same slice position, are averaged to produce one image slice to improve the signal-to-noise ratio (SNR). As a result, an averaged slice is obtained with the correct breathing state for each data sequence. Fourth, the averaged slices are inserted into the volume at their position, since each data frame has a unique and known position in the liver volume. This step is referred to as the "stacking" step, and it results in a volume with the

same breathing state as the navigator. Doing this for all navigator frames results in a continuous 4D MRI sequence.

The size of the reconstructed volumes, both in terms of FOV and spatial and temporal resolution, is determined by the MRI sequence used to acquire the data sequences. The FOV of the reconstructed volumes is based on the FOV of the acquired slices and the number of slice positions used to cover the entire liver. In this study, the FOV of the volumes is $255\,\mathrm{mm} \times 320\,\mathrm{mm} \times 152$ to $288\,\mathrm{mm}$, which corresponds to a matrix size of $140 \times 176 \times 38$ to 57 voxels.

5.2.2 Determine the Breathing State

Section 2.2, introduced the concept of breathing states. To determine the breathing state of the liver, the positions of vessel cross-sections in the navigator are tracked. These vessel positions convey both the liver's position as well as its deformation. To track vessels, first a subset of well-visible vessels with high contrast in the navigators of the subject were select. Typically 3 to 5 vessels were chosen. They represent and determine the breathing state. Manually defined region of interests (ROIs) around vessel cross sections were used to define template images and use them to find the vessel cross section positions via template matching. The matched positions of the templates represent the breathing state of the liver at the time the navigator was acquired. However, we are interested in the breathing states of the data frames, because we want to sort data frames according to their breathing state. To achieve this, the breathing state of a data frame was defined based on the breathing states of the two navigators acquired immediately before and after the data frame, i.e., the leading and the following navigator. See Fig. 5.2. This can be called a derived breathing state, which can be differentiated from the definition of a breathing state that is only based on a single navigator. In fact, the derived breathing state of a navigator is also used in the reference sequence, to correspond with the derived breathing state of a data frame. Of course, this is only an approximation, because only vessel cross-sections within the navigators are considered and from that conclusions about the breathing state of the liver as a whole are drawn.

5.2.3 Sorting Data Frames based on the Breathing State

As mentioned in section 5.2.1 a matching criterion is used to search for data frames within the data sequences that show the same breathing state as the navigator.

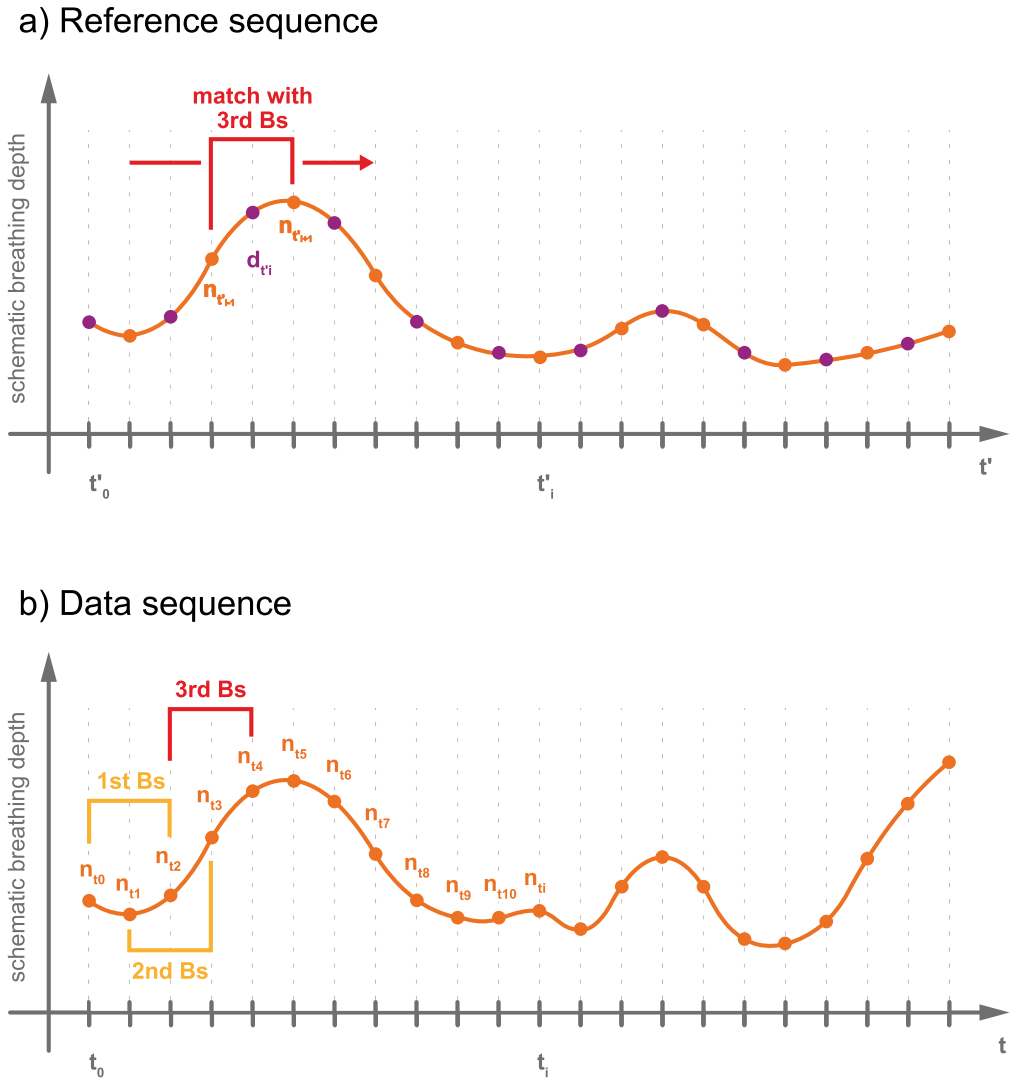

Fig. 5.2: Scheme of matching breathing state of a navigator in the reference sequence a) to the breathing state of a data slices in a data sequence b). On the left hand, the reference sequence is depicted. The red bracket represents the third breathing state and the matching data slice in the data sequence, depicted on the right.

Because the respiratory state of a data frame is determined by its enclosing navigators, the matching criterion compares the tracked vessel positions between the two leading navigators and between the two following navigators. (see brackets in Fig 5.2). Assume a navigator frame n_{t_i} at time point t_i in the reference sequence that shows a reference breathing state BS_r. We want to find a data frame d_{t_j} with the same breathing state as n_{t_i}. To this end, the enclosing navigator frames of both d_{t_j} and n_{t_i} are used. The leading and following navigator frames of d_{t_j} are $n_{t_{j-1}}$ and $n_{t_{j+1}}$ and the leading and following navigator frames of n_{t_i} are $n_{t_{i-1}}$ and $n_{t_{i+1}}$. We compare the tracked vessel positions between $n_{t_{j-1}}$ and $n_{t_{i-1}}$ and compute their displacements. Likewise we compare the tracked vessels in $n_{t_{j+1}}$ and $n_{t_{i+1}}$ to compute their displacements. When the sum of all displacements is under a certain threshold, the frames are assumed to have the same breathing state and be a match. This displacement threshold is a parameter of the method. It determines the maximally allowed displacements for two frames to be counted as a match.

The vessel tracking is realized via template matching using OpenCV (Bradski, 2000) and its similarity measure TM_CCOEFF_NORMED (see equation 5.1).

$$\mathrm{R}(x,y) = \frac{\sum_{x',y'} \left(\mathrm{T}'\left(x',y'\right) \cdot \mathrm{I}'\left(x+x',y+y'\right)\right)}{\sqrt{\sum_{x',y'} \mathrm{T}'\left(x',y'\right)^2 \cdot \sum_{x',y'} \mathrm{I}'\left(x+x',y+y'\right)^2}} \tag{5.1}$$

where

$$\begin{aligned} \mathrm{T}'\left(x',y'\right) &= \mathrm{T}\left(x',y'\right) - 1/(w \cdot h) \cdot \textstyle\sum_{x'',y''} \mathrm{T}\left(x'',y''\right) \\ \mathrm{I}'\left(x+x',y+y'\right) &= \mathrm{I}\left(x+x',y+y'\right) - 1/(w \cdot h) \cdot \textstyle\sum_{x'',y''} \mathrm{I}\left(x+x'',y+y''\right) \end{aligned} \tag{5.2}$$

Here T' is the template T minus its mean pixel intensity, and I' is an image patch with the same size as the template. Its pixel values are also shifted by minus the patches mean pixel intensity. w and h are the width and height of the template and the patch.

R is the resulting image of the template matching. Each entry $\mathrm{R}(x,y)$ contains the similarity value of the template to the source image at position (x,y)

The templates are manually defined for each tracked vessel cross section in the reference sequence. To this end, a user identifies trackable vessels in one slice of the reference sequence prior to the 4D reconstruction, which takes a few seconds. In our case, trackable means that the vessel cross-section or cluster of cross-sections will be visible in most navigator frames throughout the whole navigator sequence

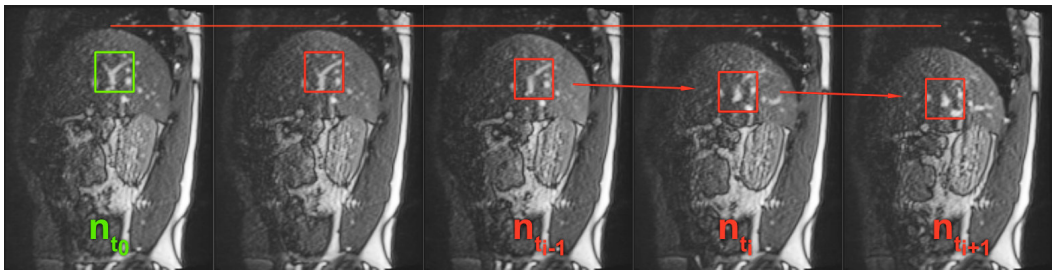

Fig. 5.3: Out-of-plane motion in ROI (green rectangle) in subsequent navigator slices. The vessel cross-section changes its appearance during breathing (red rectangles). For viewing purposes only, the images gradation curve was altered globally to enhance contrast.

and that the cross-section has a high contrast to the surrounding tissue as well as a high signal to noise ratio. This is mostly not the case for small cross sections but true for larger ones. A tool to help facilitate the manual definition of ROIs around the vessel cross section was developed.

5.2.4 Template Updates and Search Regions

One of the challenges for the template matching is the out-of-plane motion of the vessels. That means the vessel cross-sections in the navigator frames are changing, which in turn means, the searched-for regions are changing their appearance throughout breathing. In Fig 5.3, one can see how the appearance of a vessel cross-section can change during breathing. Hence, the approach of von Siebenthal has difficulties to find them, using fixed templates.

To increase robustness against this out-of-plane motion, we propose to apply template updates within the reference sequence. The method starts with the templates that were defined manually on reference frame n_{t_0}. Then, for each following navigator frame n_{t_i} that was captured at time point t_i, the templates get automatically updated, as follows: The positions of all tracked vessels in n_{t_i} are found with subpixel precision using the templates from time point t_{i-1}. Then a new set of templates is cut from n_{t_i} at the matched positions. The template positions are updated with floating-point precision. The updates ensure that changes in the appearance of the tracked vessel are represented in the updated templates. The subpixel precision in the updates avoids template drift during the update.

Another concern of the reconstruction approach is speed. In its naive form, the method matches each template against the whole navigator frame, resulting in a substantial computational burden. We propose to speed up the vessel tracking by

exploiting spatial coherence between temporally adjacent navigator frames. The underlying assumption is that the next searched-for match is in a small spatial neighborhood around the previously found match, which is the case due to fast and continuous acquisition. Therefore, only a small neighborhood around the last matched template position is used as a search area.

5.3 Experimental Design

We compare our method with the state of the art method of von Siebenthal et al. through reconstruction rate and image quality. The reconstruction rate is defined as the percentage of the number of slices in the volume that could be reconstructed by the method. Note that this does not account for false positives or false negatives because the ground truth is not available to us. It is also investigate how the acquisition order of the reference sequence and data sequence influences the method's ability to find matching data frames. False positives are evaluated indirectly using a qualitative assessment of both approaches. The image quality is assessed in a double-blind study with interventional radiologists.

Reconstruction Rate Ablation Study The reconstruction rate of both methods is compared for different parameterizations. This is possible because Siebenthal's method uses the same parameters in its matching criterion. When a subject was imaged multiple times, the reconstruction rates of its respective data sets were averaged for the statistical analysis to avoid possible biases. The parameters shown in Table 5.1 were tested. We tested the displacement threshold, for the values 0.5, 1, and 2. Evaluating different thresholds from a quantitative point-of-view allows us to judge which method will be more suitable for different applications that differ in the wanted trade-off between precision and coverage. With lower (stricter) thresholds, the coverage goes down and the precision increases. With higher thresholds, the coverage increases and the precision decreases.

Two similarity measures from OpenCV were tested, namely TM_CCOEFF_NORMED (defined earlier in equation 5.1) and TM_CCORR_NORMED (see equation 5.3), and the influence of the chosen reference sequence, ref. 1 and ref. 2, was tested, where ref. 1 is acquired before and ref. 2 is acquired after the data sequences.

The TM_CCORR_NORMED is defined as follows:

Tab. 5.1: Tested parameter values

Parameter	Values
Threshold	0.5; 1; 2
Similarity measure	TM_CCORR_NORMED; TM_CCOEFF_NORMED
Reference Sequence	ref. 1; ref. 2

$$\mathrm{R}(x,y) = \frac{\sum_{x',y'} \left(\mathrm{T}\left(x',y'\right) \cdot \mathrm{I}\left(x+x', y+y'\right)\right)}{\sqrt{\sum_{x',y'} \mathrm{T}\left(x',y'\right)^2 \cdot \sum_{x',y'} \mathrm{I}\left(x+x', y+y'\right)^2}} \tag{5.3}$$

where T is the template, I is the image and R is the resulting image with the highest intensity in position (x, y), where the similarity was the highest.

A four-factorial analysis of variance (ANOVA) was conducted to test for the effects of the reconstruction method and the aforementioned factors on the reconstruction rate.

Reconstruction Quality A double-blind study with ten interventional radiologists was conducted to compare the reconstruction quality of both methods and to evaluate whether our method's reconstruction quality improves compared to Siebenthal's method. Participants were recruited from a General Radiology clinic. Their professional experience ranged from 4 months to 20 years (median: 16 months, mean: 62 months).

The interviews were in no way invasive, and no data that would allow for participant identification was included in the analysis. Thus, IRB approval was not requested for the interviews. In all cases oral participation consent was obtained and recorded.

Each radiologist was shown a set of 48 slice image pairs. The images of a single pair were reconstructed from the same subject and breathing state, showing the same anatomical structure and having the same slice position and orientation. One slice in a pair was sampled from a reconstruction from Siebenthal's method. The other was sampled from a reconstruction of our method. The orientation and position of a slice pair was randomly chosen within a range, such that the sampled slice would show the target organ. Slices were sampled either in sagittal, coronal, or axial orientation. Slices of a reconstructed volume are depicted in Fig 5.4. Due to a software error, the number of slices for different planes was slightly imbalanced: Overall, 100 slices were shown for the sagittal and axial orientation each, and 280 slices were shown for the coronal orientation.

Furthermore, in both volumes, sagittal slices were automatically masked out (setting intensity values to black), where either of the methods did not find a matching data frame. Both volumes were made identical in the amount and distribution of black slices. This was done because it is likely that a reduced reconstruction rate for a volume would be detrimental to its perceived reconstruction quality. Sagittal slices were not sampled from masked positions.

The radiologists had to decide which of the images in a pair shows the anatomy of the target organ more faithfully, i.e., with fewer image artifacts. Participants did not see the two slices from each pair simultaneously but could switch back and forth between them as often as they wanted before picking one. Participants were asked to select the slice they considered better. A neutral option was provided. For the evaluation of reconstruction quality, the parameter set was chosen to be 1 px threshold and TM_CCOEFF_NORMED as a similarity measure for both methods. Only those volumes were considered for comparison, for which both methods had a reconstruction rate of at least 80%. For each radiologist, 48 volume pairs were chosen randomly. From these volumes the image pairs were sliced from.

For each of the 480 image pairs shown to participants, it was recorded which method was preferred, if either. For data analysis, the two methods were appointed one 'point' each for every time they had been preferred. For each neutral vote, both methods were appointed a half 'point'. This led to a dichotomous variable that allows for a direct comparison of the two methods' scores. A one-sided binomial test was conducted ($H_0 : p_{our\_method} \leq 0.5$, $H_1 : p_{our\_method} > 0.5$).

5.4 Results

Table 5.2 shows the mean reconstruction rates for all parameter combinations. Our method has a consistently higher reconstruction rate than Siebenthal et al. (about twice as high) for all parameter sets. Fig 5.5 and 5.6 depict the respective distributions of reconstruction rates.

The four-factorial ANOVA showed significant main effects for all four factors and one significant interaction effect for the reconstruction method and the threshold used (Table 5.3). This interaction effect describes that while our method performs better than Siebenthal's method at all threshold levels, it achieves stronger improvements at higher thresholds (see also Fig 5.5 and 5.6).

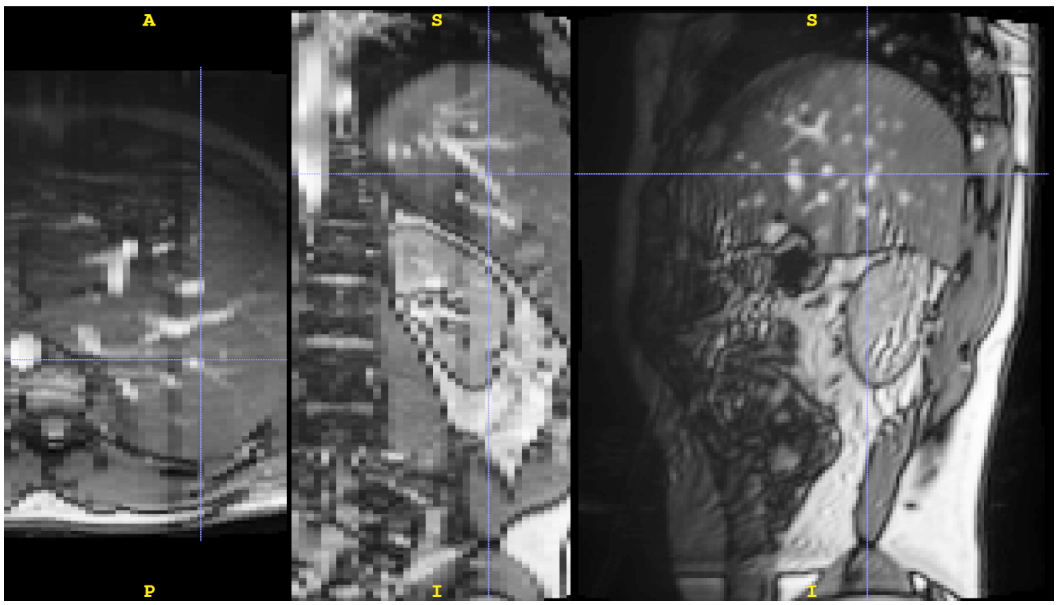

Fig. 5.4: Axial, coronal and sagittal slices of a reconstructed volume. The images gradation curve was altered globally to enhance contrast for better viewing only. In the axial and coronal orientation, one can see that our method is capable of reconstructing smooth and continuous volumes from sagittal slices.

Tab. 5.2: Mean reconstruction rates [%] of our method and baseline Reconstruction rates are given in percent reconstructed of a volume. Bold is the best rates for each parameter set.

		TM_CCORR_NORMED			TM_CCOEFF_NORMED		
threshold		2px	1px	0.5px	2px	1px	0.5px
ref. 1	baseline	24.58	15.95	9.94	41.78	24.10	12.74
	our method	**73.60**	**40.99**	**23.24**	**77.69**	**47.10**	**27.00**
ref. 2	baseline	46.86	31.95	18.75	60.09	40.07	22.92
	our method	**79.67**	**56.89**	**36.78**	**82.18**	**58.53**	**37.34**
avrg.	baseline	35.72	23.95	14.34	50.93	32.08	17.83
	our method	**76.63**	**48.94**	**30.01**	**79.93**	**52.82**	**32.17**

Tab. 5.3: Main results of the ANOVA on the reconstruction rate.

Effect type	Factor	df	F	p
Main effects	*Reconstruction method*	1	134.99	**<0.001**
	Threshold	2	106.56	**<0.001**
	Similarity measure	1	8.33	**0.004**
	Reference sequence	1	37.40	**<0.001**
Interaction effect	*Reconstruction method * Threshold*	2	7.71	**<0.001**
	*Rec method * Similarity measure*	1	1.95	0.164
	*Rec method * Reference sequence*	1	1.41	0.236

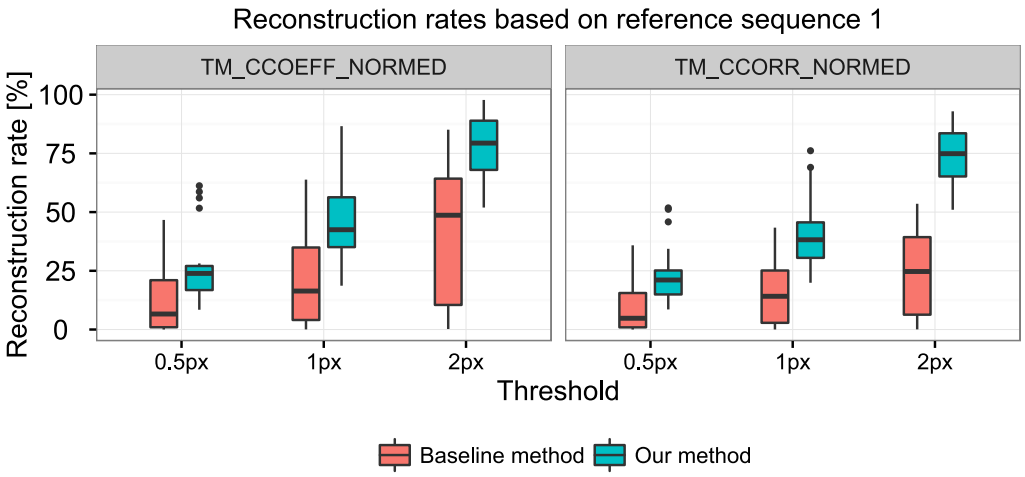

Fig. 5.5: Reconstruction rates of the proposed method compared to the baseline method for reference sequence one. Also compared are the use of two different similarity measures in the template matching.

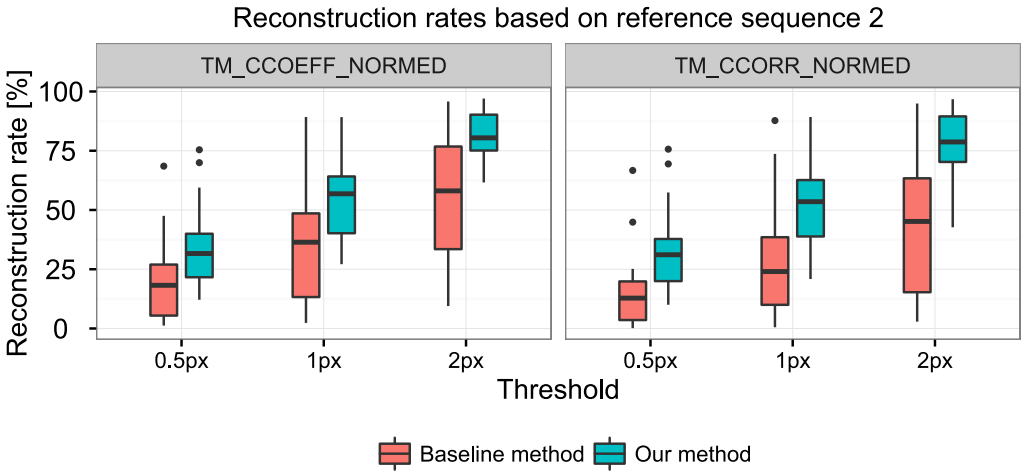

Fig. 5.6: Reconstruction rates of the proposed method compared to the baseline method for reference sequence two. Also compared are the use of two different similarity measures in the template matching.

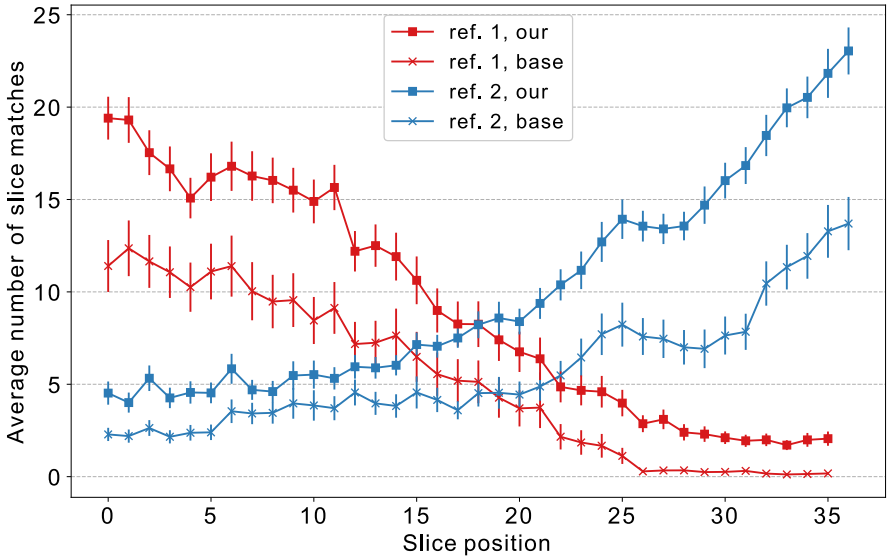

Fig. 5.7: Correlation of slice position and number of slice matches. Red graphs represent the average number of slice matches for the first reference sequence (averaged over all subjects). Blue graphs correspond likewise to the second reference sequence. Graphs with squares represent our method; graphs with crosses represent Siebenthas method. Error bars represent standard deviation and are scaled by 0.1 for better readability.

On the tested data, it was also more robust against the chosen similarity measure used for the template matching and also more robust against whether the reference sequence was acquired in the beginning or at the end of the session. Though, these interaction effects could not be shown to be significant in the ANOVA.

A correlation between acquisition order of the slice positions relative to the reference sequence and the ability of the methods to reconstruct these slice positions can be seen in Fig 5.7. With the increasing temporal distance between the acquisition of an data sequence and the reference sequence, both methods find fewer similar slices for the corresponding slice position. Reference sequence one (red graphs) is acquired before the data sequences. Here both methods find more slices for the earlier slice positions. Reference sequence two (blue graphs) is acquired after all data sequences. Here both methods find more slices for the later slice positions.

The mean reconstruction time of our method is 24.19 seconds, with a standard deviation of 6.82 seconds. The mean reconstruction time of Siebenthals method is 73 seconds, with a standard deviation of 21.81 seconds.

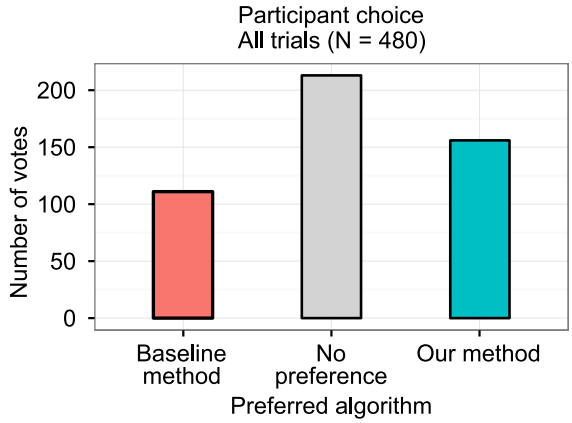

Fig. 5.8: Participant choice. The bars represent the number of times each option was chosen out of 480 trials.

In the double-blind study, overall, participants selected our method in 156 trials, Siebenthal's method in 111 trials, and had no preference in 213 trials (see Fig 5.8). Following our analysis method, this yielded 262.5 'points' for our method and 217.5 'points' for Siebenthal's method (p=0.02).

The study shows that radiologists perceive the reconstruction quality of our method as better than Siebenthal's method, although the effect seems to be small, it was shown to be significant in the one-sided binomial test.

5.5 Discussion

The particular acquisition scheme shows difficulties with changes in breathing patterns that arise over a more extended period, like the typical flattening of the resting breath. Slice positions to the subject's left are imaged only at the end of acquisition time, whereas slices to the right are only imaged at the beginning. As a consequence, if the reference sequence was captured in the beginning, it can show breathing states that do not occur later, when slice positions to the left are imaged. Deep breaths often can not be fully reconstructed since image data of the left slice positions was not acquired for deep breathing states. Generally speaking, the scheme has difficulties with breathing states that are less frequent. This problem can be solved in changing the acquisition scheme. Instead of first acquiring all slices in one position before moving on to the next slice position, it is beneficial to move the slice position after each acquisition while keeping the navigator position

fixed. This rotating acquisition scheme could also be combined with intermediate reference sequences. This would directly counter the problem with flattening breath over time. Furthermore, with the new scheme, it is feasible to give a few commands, so the subject can take a few more deep breaths in the beginning before starting to relax more.

The rotating acquisition scheme was used by Siebenthal et al. on a 1.5T Philips Intera whole-body MRI system (M. v. Siebenthal et al., 2007). However, Siemens MRI machines do not allow this kind of scheme. A solution to the problem that is independent of the scanner used, is, to use external respiratory signals instead of navigator frames. Preiswerk et al. (2017) had correlated 1D MR compatible ultrasound with 2D and multiplanar MRI. This allows for the continuous rotating acquisition of the data slices on any MRI machine. Celicanin et al. (2015) propose a simultaneous multislice (SMS) imaging technique that allows for the simultaneous acquisition of navigator and data frames, increasing the temporal coherence of navigator and data frame. Barth et al. (2016) give a current overview of parallel imaging and SMS imaging techniques. These would integrate well with the rotational acquisition scheme when using body array coils. No body array coil is used in our experiment to ensure a line of sight for external marker tracking. However, when external marker tracking is not needed, a body array coil can readily be used in conjunction with our method to have better image contrast and possible faster imaging with aforementioned SMS techniques applied. When flat, flexible array coils with an opening for operation become available, those benefits, i.e, higher SNR, faster acquisition and line of sight, could be combined. Regarding the acquisition time, the aforementioned changes to the acquisition scheme would half the acquisition time in our case to between 20 and 40 min.

Regarding the reconstruction rate, because of the lack of ground truth, it is not possible to account for false negatives and false positives in the evaluation. Based on this fact, the reconstruction rate of both methods will possibly be higher than measured in this study. This is because, in our test data, for some slice positions there might be no matching image in the data sequence, for a given breathing state, resulting from the acquisition scheme mentioned above.

An open issue arises when vessel cross-sections in the navigator frame are not continually visible. This frequently happens to depend on blood flow. To solve this, one could detect outliers in the template matching step and omit those for the calculation of the summed displacement.

We decided to use MRI data of healthy volunteers for the development and evaluation of the method. For a proof of concept of our method, this eliminates possible adverse

effects of liver diseases on the respiration of the patient, making the evaluation environment more controlled. However, in future work, it has to be evaluated if typical diseases targeted by this method, like liver carcinoma, affect the method. This could be especially the case if the disease impairs the respiration of the patient. If the patient's breathing shows no or few repetitions of patterns, this would be a challenge for the method because whilst allowing for irregular breathing, it assumes that patterns are recurring over time.

In its presented form, our method relies on a manual step in which the ROIs around the vessel cross-sections are defined. In a real clinical setting, this is intended to be done offline after the planning MRI acquisition and before the date of the intervention on a suitable computer, not directly on the MRI machine. Even though this manual interaction is minimal and takes less than a minute to perform, it could and should be automated in future work. This could be solved as a classification problem in image space using the temporal information of the reference sequence as supporting information.

In our evaluation of the visual reconstruction quality, we only compare our method relative to Siebenthal's method. The provided neutral option does not differentiate between equally good and equally bad or unusable, and no absolute data was gathered. Hence, our analysis does not show whether the reconstructions are good enough for a clinical task or not. The analysis only indicates that our method's reconstruction quality improves compared to the other method.

In summary, our results clearly show that template updates are an effective and efficient means to increase reconstruction rates and image quality of the reconstruction result for templatematching-based 4D MRI reconstruction methods. This chapter reported that employing search regions significantly reduces reconstruction time. The results suggest that our method is preferable compared to Siebenthal's method. This is regardless of the application's favorable trade-off between precision and coverage because, in all cases, our reconstruction rates are higher.

5.6 Conclusion

In conclusion, this chapter has presented a method for robust sorting of free-breathing MRI slices to reconstruct 4D MRI. It has been shown that it outperforms the state of the art method of von Siebenthal in terms of reconstruction rate by addressing the problem of ouf-of-plane motion. It also substantially reduces the acquisition time, despite the need for further improvement in that regard. Moreover,

a double-blind study, conducted with radiologists, has shown that the proposed method also produces higher image quality. The method presented in this chapter serves as a baseline for the development of a more advanced method to address still open problems, which are described in the following chapters. The findings from this study highlight the importance of further research in this area and provide a starting point for following investigations to build upon.

6 Predicting 4D Liver MRI for MR-guided Interventions

Synopsis

This chapter reports the development of a Deep Learning based framework for real-time predicted 4D MRI. It is a network-agnostic, end-to-end trainable, deep learning formulation. It can be used in two ways: First, it can reconstruct high quality fast 4D MRI with high resolution $0.6\,\mathrm{s/volume}$ *during an intervention. Second, it can be used for retrospective 4D reconstruction with an even higher temporal resolution of* $0.166\,\mathrm{s/volume}$ *for motion analysis, intervention planning and use in radiation therapy. The mean target registration error (TRE) of* $1.19 \pm 0.74\mathrm{mm}$*, is below voxel size. The results are compared with the baseline described in the previous chapter. Visual evaluation shows comparable quality. Different network architectures are compared within the formulation. Small training sizes with short acquisition times down to* $2\,\mathrm{min}$ *can already achieve promising results and* $24\,\mathrm{min}$ *are sufficient for high quality results.*

About this chapter Parts of this chapter have been published in: Gino Gulamhussene, Anneke Meyer, Marko Rak, Oleksii Bashkanov, Jazan Omari, Maciej Pech, and Christian Hansen (2022). "Predicting 4D Liver MRI for MR-guided Interventions". Computerized Medical Imaging and Graphics, 101, 102122. (Gulamhussene et al., 2022)

6.1 Introduction

As set out in chapter 2 and 5, virtually all earlier proposed 4D MRI methods cannot acquire or reconstruct a 4D MRI in real-time, whether they are prospective or retrospective. Also, the baseline that was presented in the previous chapter, is not real-time capable. To address this limitation, we propose a method that realizes the near real-time prediction of 4D MRI. It uses the concept of a dynamic 2D navigator slice, which is acquired, using a readily available clinical MRI sequences. Our method also has the capability of retrospectively reconstructing 4D MRI.

The novel approach to generate 4D MRI, presented in this chapter, is the first deep learning-based 4D MRI prediction framework. It is fast enough to be practical in the context of medical interventions, while providing high spatial resolution and supporting irregular breathing. It is end-to-end trainable and network-agnostic. By being end-to-end trainable, the proposed approach overcomes the previously necessary explicit sorting of MRI slices. This results in a fast near real-time reconstruction. Our method can be used in two ways. First, it can predict 4D full-liver MRIs in near real-time. This predicted 4D MRI has a high spatial resolution ($209 \times 128 \times 128$ matrix size, isotropic 1.8^3mm^3 voxels) and high temporal resolution (600ms). Second, it can retrospectively reconstruct 4D MRIs. In this case, the reconstructed 4D MRI has an even higher temporal resolution of 116ms with the same spatial resolution. In both cases the method can cope with irregular breathing or arbitrary physiological breathing patterns, extracted from the 2D navigator sequences. Our method is capable of reconstructing 3D liver MRIs even with drastically reduced training data, cutting acquisition times to only a few minutes comparable to breathing phase resolved onboard 4D MRI techniques. Most importantly, it outperforms other methods in the ability of being real-time capable.

6.2 Materials and Methods

This section first discusses the parts of the data base that were used for the development and testing of the deep learning based 4D MRI prediction framework. Then, the concept of the framework is illustrated before finally, a detailed description of the framework is given.

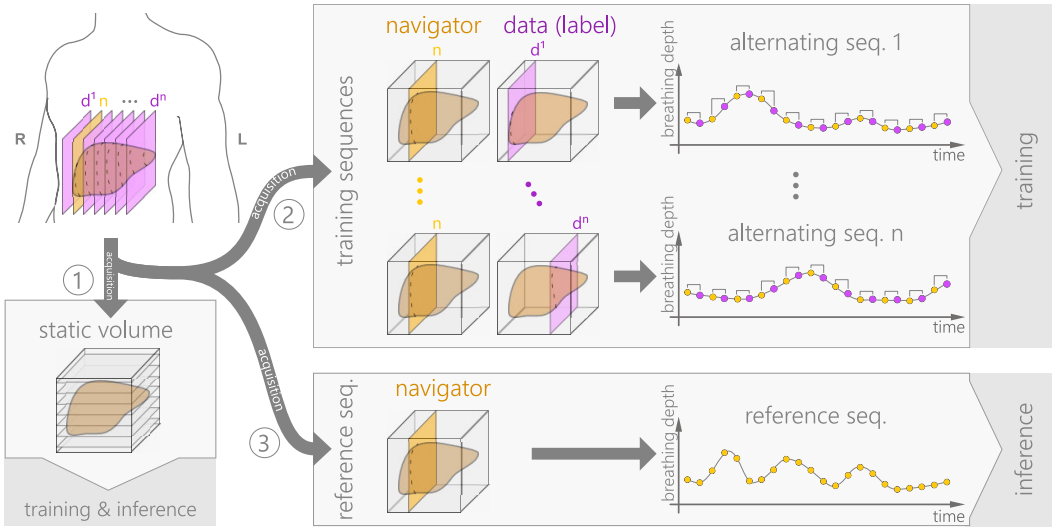

Fig. 6.1: All three parts of the data base are used. Specifically: 1) one static volume, 2) all data sequences, i.e., training sequences; brackets indicate pairs of navigators (orange dots) and data slices (purple dots), and 3) one dynamic reference sequences.

6.2.1 Training Data

The deep learning framework utilizes all three parts of the data base. Namely the data sequences, the breath hold volumes and the reference sequences (see Fig 6.1). In chapter 4 the data parts were described in detail. In the deep learning framework, the data sequences and breath hold volumes are used for training, and the reference sequence is used for inference.

In the last chapter 5, we described the baseline method, which used the data sequences to sort and stack 4D MRI according to the reference sequence. In that context, the data sequences were seen as series of triplets. Furthermore, those triplets were interconnected. That means, the last navigator of one triplet was the first navigator of the next triplet. This was necessary for the sorting part of the method. In contrast to that, in the context of the deep learning base framework, a data sequence is treated as a series of disjunct pairs of training samples, containing navigator and label. In other words, the data sequences are seen as training, validation and test data for the deep learning approach.

Regarding the deep learning task, an important fact is that there are only 2D labels available as ground truth data and no 3D labels. For the loss function there are no full volumes available, but only one slice within each volume as ground truth. How

the framework is formulated, to still be able to train a model to predict 3D volumes, is discussed in section 6.2.2.

In the remainder of the thesis it will be often refer to the amount of used or available training data. Depending on clarity and readability this will be done in one of two ways. First, we will refer to it in the total number of training samples. For example 8431 samples. Second, we will refer to it in the amount of time the acquisition of that number of samples took. Each sample is an image pair and each image takes $166\,\mathrm{ms}$ of acquisition time using the TRUFI MRI sequence. So for example, 8431 samples amount to $46.652\,\mathrm{min}$ or around $47\,\mathrm{min}$ and 361 samples amount to $2\,\mathrm{min}$. We will use the second way of referring to the training data amount more often. It is easier to read and emphasizes the clinical impact or cost of acquiring the training data.

The MRI data is available in the DICOM format. For further processing, it is converted to the nifti format using the dicom2nii command line tool, which conserves all necessary meta data of the image file like scanner coordinates and voxel spacing. Before the images are further converted to image arrays during the training pipeline, the images are re-sampled using scanner coordinates. Re-sampling is done to harmonize the network input. The TRUFI slices are re-sampled to 128×128 voxels with a size of $1.8 \times 1.8\,\mathrm{mm}^2$. The STAR VIBE breath hold volume is re-sampled to 209×128×128 voxels with a size of $1.8 \times 1.8 \times 1.8\,\mathrm{mm}^3$.

To facilitate robustness, we augmented the training data in physiological plausible ranges in-plane with random translation of up to ±10 voxel (±18.18 mm), random rotation of up to ±3° and random scaling within $[0.8, 1.2]$.

$$I_{norm} = \frac{I - \mu}{{\sigma^2}_{adj}}, \qquad \text{and} \qquad {\sigma^2}_{adj} = \max\left(\sigma^2, \frac{1}{\sqrt{\#voxels}}\right), \tag{6.1}$$

where I_{norm} are the whitened (normalized) intensities, μ is the average intensity for all slices of one subject. Likewise, ${\sigma^2}_{adj}$ is the standard deviation, which was adjusted by a reasonable lower bound that depends on the number of voxels $\#voxels$ available for that subject.

6.2.2 Deep Learning based 4D MRI Framework

Lets now illustrate the general idea of the framework. Let's assume we have a real-time interventional 2D sequence with 6 fps during an intervention. From that, we get a dynamic navigator slice from the subjects liver and predict what the full MRI

volume of the abdomen for each navigator would look like, as soon as the navigator is acquired. In other words, the general idea is, to predict a 4D MRI from a series of 2D navigator MRIs, that come from a real-time interventional sequence during an intervention. For the development and testing of the framework, we use the reference sequence, to simulate the real-time stream of interventional navigators.

There are three aspects to the framework. First, the utilization of a deep learning network. Second, the three channel input of the network that is used to encode the slice position and breathing sate, to be predicted. And third, a batching scheme that allows to predict volumes in one forward pass, without the need for 3D training labels.

Lets start with the first aspect: the network. We cannot implement the training in a straightforward way using a 3D network architecture, because we do not have volumetric training data as pointed out in 6.2.1. If we had pairs of navigators and corresponding 3D volumes with the correct breathing state as labels, we could train a 3D network. But we only have pairs of 2D navigator slices and 2D data slices. So instead of training a 3D network to predict a whole volume at once, i.e., EVERY slice position, we train a 2D network to predict ANY slice position that we specifically tell it to predict. In section 6.2.4 we discuss the network architecture in detail. But first lets continue with the second aspect, the three channel input, which becomes important at this point. It is used in a way that we can encode the slice position, the network should predict, and the breathing sate, the slice should have. This encoding, together with, what we call transitive information bridging, is explained in the following section. That does not mean that we have to infer each slice position one after the other. We can infer all slice positions in a single batch, which is the third aspect. For that, we encode each slice of a volume with the same breathing state in another batch entry and infer the volume in one forward pass.

Here, we want to point out a conceptual difference, regarding the breathing state of a data frame, compared to the last chapter. There, it was defined by the two navigators that came before and after the data frame. In contrast to that, in the deep learning framework, the breathing state of a data frame is only defined by the navigator that came before. The implications of that are discussed in 6.5

In the following sections, we describe the three aspects of the framework in more detail. In section 6.2.3 the three input channels are explained and how they form, what we call, a transitive information bridge that allows the network to predict a volume from a 2D slices. Section 6.2.4 will illustrate the network agnostic property of the framework and describe four specific network architectures that were tested

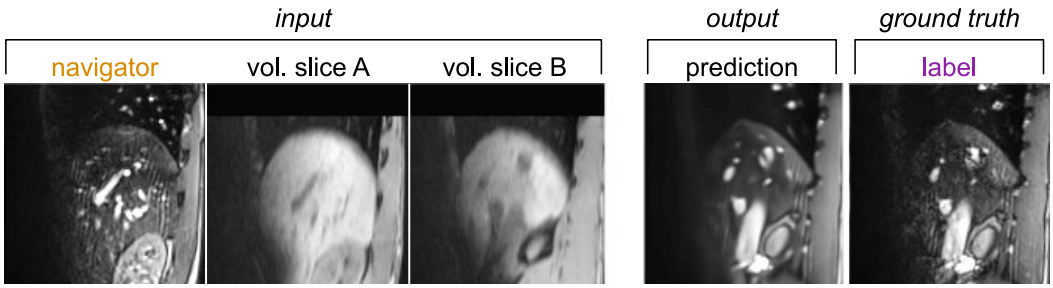

Fig. 6.2: The network input consists of three channels. The first channel receives a navigator slice that tells which breathing state to predict, i.e., the breathing state that follows the navigator. The second channel receives a static volume slice (vol. slice A) at the navigator position, to act as a still reference to the moving navigator. The third channel receives a static volume slice (vol. slice B) that tells the network at which position to predict the new slice.

with the framework. Finally, the actual prediction of a 4D MRI, using the framework is described in section 6.2.6.

6.2.3 Input Channels and Transitive Information Bridges

An integral part of the framework formulation is the three channel input of the network, which is shown in Fig 6.2 together with the prediction and the ground truth label. Lets first see which data is fed to the input channels and than how we can interpret that. The first channel gets the 2D navigator slice. During inference the navigator is taken from the reference sequence, which represents the real-time interventional sequence. During training, however, the navigator slices, together with the training labels are taken from the training samples in the data sequences. Remember the navigator is the first image in the pair, the label is the second image of the pair. The second channel gets a slice, which is sampled from the breath hold volume at the same position and orientation of the navigator. The third input channel gets another slice from the breath hold volume. This one is sampled at the position of the label, again in the same sagittal orientation. Within one batch entry, all three channel inputs as well as the label correspond to the same subject. However, during training, samples of multiple subjects can be used within one training run.

Now we can interpret the three channel input. The idea is that the network can determine the breathing state from the fist two channels. The navigator in the first channel shows the breathing state that the network must predict, but for another slice position. The volume slice in the second channel acts as a still reference to

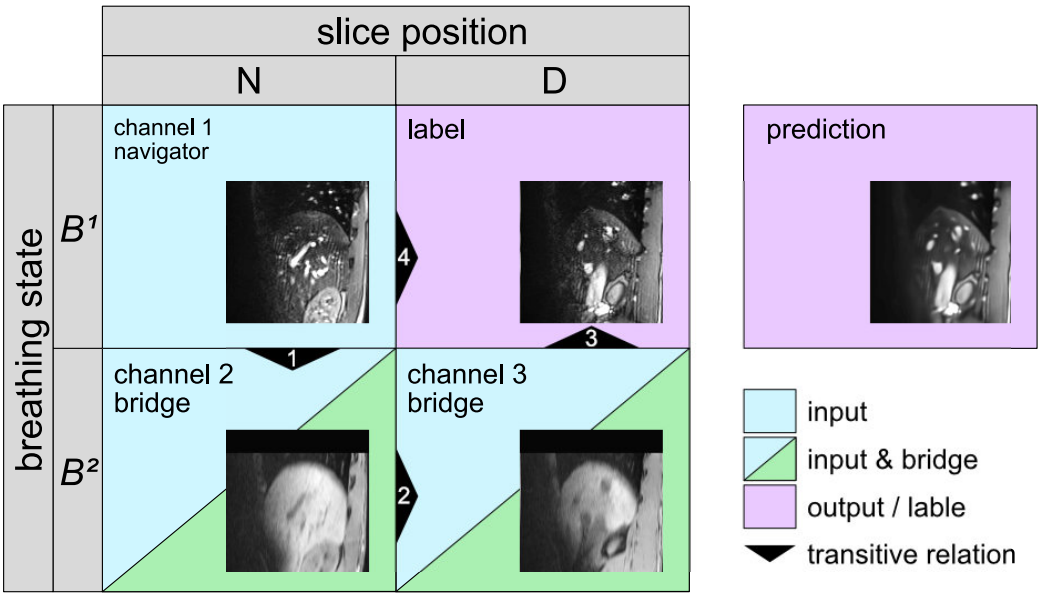

Fig. 6.3: Visualization of the bridge with transitive relations

the navigator. And finally, the volume slice in the third channel encodes the slice position that the network must predict.

In this three channel input, the volume slices in the second and third channel, act as a transitive information bridge. Because this is a central concept within the framework, lets call this simply a bridge and discuss it in more detail. To learn the breathing state from the input and apply it to the output, the network uses the bridge, which is made up by transitive relations between the input channels, the output, and the label. This is visualized in Fig. 6.3, which gives another view on the input and output of the network, focusing on these transitive relations, which are depicted as black arrows. As already discussed, the navigator (channel 1) and the first volume slice (channel 2) are relate, by having the same slice position (N). The two volume slices (channel 2 and 3) are related by sharing the same breathing state (B^2), because they both come from the same breath hold volume. Finally, the second volume slice (channel 3) and the label are related by sharing the same position (D). The bridge works, because of a fourth relation, which is that the navigator and the label share the same breathing state.

Although we said that the network determines the breathing state from the first two channels, actually, it rather learns the difference in breathing states. Let B^1 be the breathing state, which is visible in the navigator in channel one. Although technically, the navigator was acquired $166\,\mathrm{ms}$ before the lable and thus has a slightly different breathing state, we assume it to be the same as the breathing state of the label or at least in a fixed relation. Let ΔB be the difference between the breathing states

B^1 and B^2. The network can learn the effects of ΔB from seeing the two slices with same position but different breathing states in the first two channels (transitive relation 1). Using transitive relation 2, the network applies the inverse of ΔB to the third channel, thus predicting the breathing state B^1 for that slice position. Finally, it learns to apply the MR contrast of the navigator to that slice position, using transitive relation 3 and 4. The result is a predicted slice at the same slice position as the volume slice in channel three and the label and with the same breathing state as the navigator in channel one. In summary, the first two input channels encode the breathing state, while the third channel encodes the slice position to be predicted.

6.2.4 Network Architecture

Fig. 6.4: The four architectures that are evaluated in the deep learning framework. White boxes are convolutions followed by leaky ReLU, grey down arrows denote max or average pooling, gray up arrows denote up sampling via transpose convolution, pluses and dots denote element wise addition and concatenations of feature maps. Number of feature maps are given in brackets in the boxes.

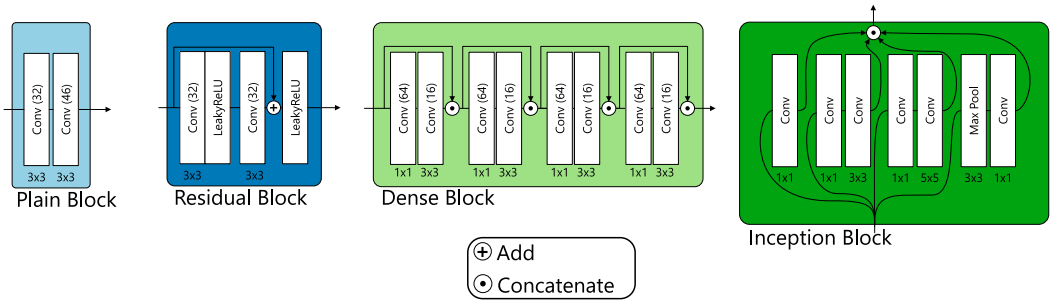

Fig. 6.5: The four block types used to build the different architectures. White boxes are convolutions followed by leaky ReLU, except in residual blocks, where the activation is denoted separately, pluses and dots denote element wise addition and concatenations of feature maps. Kernel size and strides are given below the boxes, number of feature maps are given in brackets in the boxes.

One of the strength of the proposed 4D MRI framework is that it is network agnostic. That means, it does not rely on a specific network architecture. The only requirement is for the network to have a three channel input as described in 6.2.5. The proposed method is evaluated with four different architectures, depicted in Fig. 6.4. The basic structure of the networks is the one of a U-Net (Ronneberger et al., 2015) the used block types are depicted in Fig. 6.5. The input to the networks is processed in an encoding and decoding path. The encoding employs one of the following building blocks, which are also depicted in the figure. To get four different networks, we use four different building blocks. Specifically, these are the plain block as first described in the original U-Net paper (Ronneberger et al., 2015), the residual block (He et al., 2016), the dense block (Huang et al., 2017) and the inception block (Szegedy et al., 2015), which were proposed in the corresponding papers. As with the original U-Net architecture, skip connections are used to forward details from the encoding path to the decoding path. The total number of parameters for each network is presented in Table 6.1. All networks have a $128\times128\times3$ input and 128×128 output. The leaky rectified linear unit (leaky ReLu) with a slope coefficient of 0.1 is used as activation function in all building blocks and networks. Also, all convolutions are padded to keep the size of feature maps. In the following the network architectures are described in more detail. A visual depiction of the blocks are given in Fig. 6.5.

U-Net

The U-Net is constructed from plain blocks. This block consists of two 3×3 convolutions, each followed by an activation. The second convolutional layer in each plain block doubles the number of features, increasing the network's capacity. The

first convolution has the same number of features as the second convolution of the previous block. The first three blocks are followed by a MaxPooling operation. The 128×128×3 input to the network is processed by 32 filters in the first convolutional block and results in 512 filters in the latent feature space. The decoding reconstructs the image from the latent space. To this end, three transposed convolutional blocks up sample the features, each of which consists of two convolutional layers with a dropout layer in between. At each up sampling, the filter size is halved. At the end, a final 1×1 convolution layer outputs the reconstructed image.

Res-UNet

The Res-UNet is constructed from residual blocks. It is similar to the plain block, with the difference being that a residual connection is introduced, which element wise adds the blocks input to the feature map of the second convolution (before activation). To match the number of features after each down or up sampling a 1x1 convolution is used. The number of features is doubled after each second residual block.

D-UNet

The D-UNet is a U-Net constructed from dense blocks, which is similar to the residual block. It consists of eight convolutions that build four groups. Four skip connections concatenate each groups input to its output before its fed into the next group. In the D-UNet the number of features does not increase from block to block, rather a hidden state is build up by concatenating the feature maps of previous blocks.

I-UNet

The I-Unet is a U-Net constructed from inception blocks. This block has four paths from input to output. The fist path contains a 1×1 convolution, the second and third paths begin with a 1×1 convolution, followed by a 3×3 and 5×5 convolution respectively. The fourth path consists of a MaxPooling with stride 3×3, followed by a 1×1 convolution. The number of features within a block is the same for all convolutions. It is doubled from block to block in the encoding path of the I-UNet and results in 512 features in the latent feature space.

	parameters	trainable	sfs	lfs	activ. function	optimizer
U-Net	6.8 Mio.	6.8 Mio.	32	512	leaky ReLU	Adam
Res-UNet	8 Mio.	8 Mio.	32	512	leaky ReLU	Adam
D-UNet	0.77 Mio.	0.77 Mio.	16	64	leaky ReLU	Adam
I-UNet	9.9 Mio.	9.9 Mio.	32	512	leaky ReLU	Adam

Tab. 6.1: Architecture overview, (trainable) parameters, starting feature size (sfs), latent feature size (lfs), activation function and optimizer.

6.2.5 Training

All networks are implemented and trained with Keras (Chollet et al., 2015). For training we used the Adam optimizer (Kingma et al., 2014). The networks were trained for 200 epochs using the mean squared error (MSE) as loss function between predicted slice and label. For that the available samples of each subject were split into 8,811 training and 180 validation samples (roughly 4 validation samples per slice position). Hyper parameters were empirically determined using a Bayesian search using Weights & Biases (Biewald, 2020) across 16 subjects. The following best parameter settings were obtained: learning rate (0.000413), drop out ratio (0.15), data shuffling (true) and batch normalization (false). We also tested the augmentation parameters and found them to improve the reconstruction results for all subjects irrespective of the exact range of any single parameter.

6.2.6 4D MRI Prediction

As mentioned before the method can be utilized in two ways. First, we can use the model to predict 3D liver MRI in near real-time for any real-time navigator image during interventions. Because in the training data, the navigator slice is acquired 166 ms before the label data, the network actually predicts data slices 166 ms ahead of time. Second, we can use the network to reconstruct a 4D liver MRI from a sequence of navigator slices. In both cases, for each time point, an input batch is constructed, where each entry of the batch corresponds with a slice position in the reconstructed volume. This allows us to infer all slices for a 3D volume in a single forward pass. After inference, the predicted 2D slices are concatenated to a volume. The meta information, like scanner coordinates and voxel spacing is copied from the breath hold volume. Note, that within one batch, all inputs have the same navigator slice (first channel), while all third channels show different positions of the static liver volume. This process is repeated for all time points of the navigator sequence

to form a 4D MRI. The computation time for one 3D prediction is $\leq 600\,\mathrm{ms}$ on a GeForce GTX 1080.

6.3 Experimental Design

This section first describes the research questions that will be answered, regarding the deep learning based 4D MRI method. Then the split in training, validation and test data is described. Finally, the experiments are described that were conducted to answer the research questions.

6.3.1 Research Questions and Hypothesis

In this chapter five research questions (RQs) are addressed, which are summarized in Tab. 6.2. The RQs and the corresponding null hypothesis (H_0) and alternative hypothesis (H_A) are described in the following.

RQ 1: Can the deep learning (DL) based 4D MRI framework achieve the same image quality as the baseline method? $\mathbf{H}_0$ **1:** The image quality *is not* the same. $\mathbf{H}_A$ **1:** The image quality *is* the same.

RQ 2: Does the framework work with different network architectures? $\mathbf{H}_0$ **2:** The framework *does not* work with different network architectures. $\mathbf{H}_A$ **2:** the framework *does* work with different network architectures.

RQ 3: How accurate is the position of vessels in the prediction? $\mathbf{H}_0$ **3:** The TRE of the method is *above* voxel size. $\mathbf{H}_A$ **3:** The TRE of the method is *below* voxel size.

RQ 4: What is the minimal amount of training data to achieve the same quality results as the baseline method? There are no hypothesis for RQ 4.

RQ 5: Does the method perform equally good at all slice positions? $\mathbf{H}_0$ **5:** It performs equally good. $\mathbf{H}_A$ **5:** It *does not* perform equally good.

RQ 1:	Can the new method achieve the image quality of the baseline comparable?
RQ 2:	Does the framework work with different network architectures?
RQ 3:	What is the positional accuracy of vessels in the prediction?
RQ 4:	How much training data is required to match the baseline image quality?
RQ 5:	Does the method perform equally good at all slice positions?

Tab. 6.2: Research questions

6.3.2 Training, Validation, Test Split

For evaluation the data was randomly split into training data (16 subjects) and test data (4 subjects). The data sequences of the training subjects was split in training (8,811 samples) and validation (180 samples). The data sequences of the test subjects was split in training (4,496), validation (180), and test (4135).

6.3.3 Experiments

Visual Evaluation We visually assessed prediction results and image quality of our method. For that we trained a model for each test subject and analyzed the predicted 4D liver MRIs. We visually compare the prediction with the reconstruction results of the baseline method for the same test subjects. The baseline method was described in chapter 5.

Architecture Evaluation We evaluate the framework using the four different architectures described in 6.2.4. We compare them with respect to RMSE, MDISP, cosine and prediction time, which were defined in 2.6. To that end, for each network architecture and test subject a model was trained and evaluated for all metrics, following the training described in 6.2.5. The metrics were computed for roughly one breathing cycle per each slice position. The breathing cycles contained ~ 17 consecutive time points. That amounts to $\sim$730 data points per subject. The prediction times were only computed for every second slice position and for ~ 8 time points each. That amounts to $\sim$170 data points per subject. A one-factorial analysis of variance (ANOVA) was performed to test for a main effect of the architecture on the metrics. A post-hoc pair-wise t-test was performed to test for significance of the difference in the means between architectures pairs. The ANOVA and t-test require the distribution of the variables to be normally distributed, which the metric values are not, which is indicated by the Lilliefors-Test and is to be expected, because the values are zero bound. However, the large sample size allows for the use of the t-test. The effect size of differences in means was determined using Cohen's d.

Target Registration Error The best performing architecture in the architecture evaluation was used to evaluate the positional accuracy of vessels in the predicted MRI images. Using this architecture, a model was trained for each test subject. The training and validation parts of the test subjects was used for the training. Training was performed as described in 6.2.5. Then for each subject a subset of four slice positions was chosen at roughly $-3\,\mathrm{cm}$, $-2\,\mathrm{cm}$, $0\,\mathrm{cm}$ and $3\,\mathrm{cm}$ distance to the navigator. Negative distances indicate that the position is left of the navigator and positive distances represent positions right of the navigator position. Then, a time series was predicted for each slice position in the subset using test split of the corresponding data sequences. This corresponds to a total of 16 predicted 2D+t time series (4 subjects $\times$ 4 slice positions). Within each of these the starts and endings of the breathing cycles were manually determined by finding all time points of end-exhale. A breathing cycle contained between 8 to 17 images, from end-exhale to end-exhale. For each time series, one breathing cycle was randomly chosen, on which the TRE was evaluated. To that end, for each selected breathing cycle, 1 to 5 vessel cross sections were manually tracked in all images of the breathing cycle, in both the prediction and the ground truth, i.e., data slices.

That way, we generated a total of 1566 data points in both ground truth and prediction that were used for TRE evaluation. The TRE was computed as the mean distance between corresponding tracked markers in prediction and ground truth. Fig. 6.6 shows an example of a breathing cycle that was chosen for TRE evaluation. It shows both a prediction as well as the ground truth and the manually marked vessel cross sections in prediction and ground truth (arrows). The dashed lines illustrate the breathing motion. The solid red lines mark the end-exhale position of the liver dome.

Ablation Study We evaluated the reconstruction quality of the models depending on the training data size. For that the best performing network from the architecture experiments was used. To simulate the reduction in training data availability, we defined 6 levels. Because each subject has a different number of data sequences, which results in a different total amount of available training data, we define the levels as percentages of the total training data amount of a subject. For the 4 test subjects, the total amount of training data ranged between $44\,\mathrm{min}$ and $50\,\mathrm{min}$. For the levels we chose 98%, 75%, 50%, 25%, 10% and 5% (training + validation data). For example, the number of available training samples at level 98% was 8431 or $47\,\mathrm{min}$ and for 5% it was 430 or $2\,\mathrm{min}$. For this experiment a different training, validation, test split was performed. Independent of the level, the last 2% of the data sequences remained as test data. For each of the test subjects we trained models for all 6 levels. The models were evaluated on the 2% of test data. The evaluation

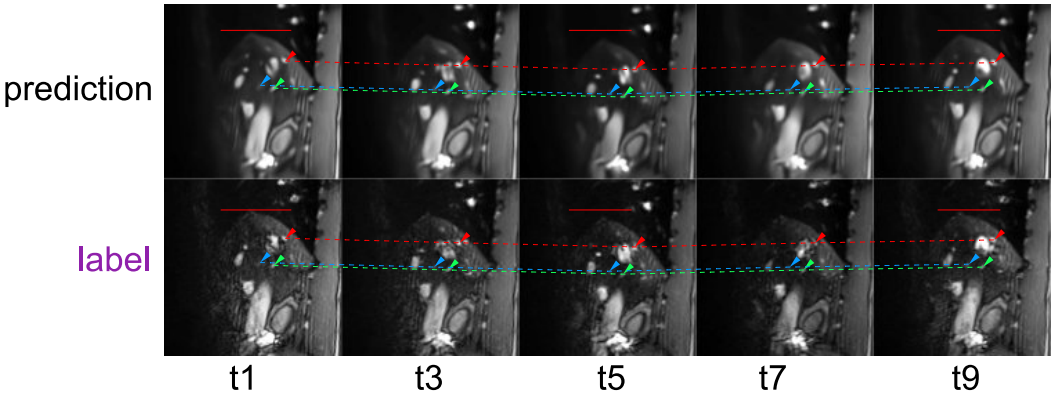

Fig. 6.6: Sample breathing cycle (prediction and ground truth) for the TRE calculation with tracked vessels (arrows) and their traces (dashed lines). Red solid lines serve as reference for the breathing depth. For compactness, only every second time point is shown. Slice position is 3.8 cm left of the navigator.

was performed quantitatively using MSE and qualitatively by visual comparison to the baseline method. Remember, although no 3D ground truth is available, for each predicted volume, the test data contains the ground truth for one of the volume slices when the prediction is performed on navigators from test samples. That means, full volumes were predicted for the visual comparison, but the MSE was only evaluated for the slice positions with existing ground truth. The MSE was avaraged over all slice positions within each test subject.

Dependence on slice position We evaluated how the performance of our method depends on the predicted slice position. For the test we used the training data availability level of 50%. We evaluated the MSE as described before. But instead of averaging the MSE over all slice positions and test subject, we binned the slice positions into 12 mm bins. The MSE was then averaged over all test subjects per bin. We also compare the MSE between the validation data, i.e., the one used during training, and test data, to show potential over fitting in different slice positions.

Comparison to Related Work We compare acquisition and reconstruction times with state of the art methods in Table 6.5. Note that our method can be combined with the last three methods in the table, which would lower the acquisition times further.

6.4 Results

Visual Evaluation All reconstructions were visually plausible when compared with the baseline reconstruction. Fig. 6.7 illustrates a reconstructed end-exhale and end-inhale volume from the baseline method on the left side and for a predicted volume using the framework on the right side. The volumes are presented as three orthogonal slices in axial, sagittal and coronal orientation. At this point remember, both methods, the baseline as well as the deep learning frame work, perform reconstruction and prediction slice-wise using sagittal slices. Therefore, in this figure, special attention should be given to the coronal and axial views, as any out-of-plane discontinuities would be apparent here. For both the exhale and inhale baseline reconstructions (a,b), we observe that blood vessels and liver boundaries are continuous and smooth in axial and coronal views. In the predicted exhale and inhale volumes (c,d), the liver dome and vessels are continuous along all view axes. The major vessels are present and smooth. In-plane details are well reconstructed, however some smaller vessels are missing in the prediction or do not show the correct trajectory in axial orientation. The breathing depths match excellently between baseline and prediction. It should also be noted that, to a limited degree, the deep learning based method is capable of reconstructing regions of the thorax and abdomen on the left and right side that were never seen during training. This is observable in the axial and coronal views. The baseline method, or any other sorting method, cannot reconstruct these regions, as these can only sort and stack available data slices. Interestingly, comparing the prediction with the ground truth (see Fig. 6.6), it can be observed that the network prediction enhances the image quality compared to the label and predicts vessels correctly that are barely visible in the ground truth. This means that **RQ 1** can be answered positively, the image quality of the ædg based frame work is the same as the baseline.

Architecture Evaluation The results of the architecture evaluation and comparison are shown in tabel 6.3. The significance of pair-wise mean differences (pair wise t-test) and the effect sizes (Cohen's d) are shown in Fig. 6.8. Overall, the ANOVA revealed a significant main effect of the architecture on the performance ($p<0.001$), with respect to all tested metrics. While, the effect sizes between the groups were mostly small ($d\leq0.5$) or negligible ($d\leq0.2$) for RMSE, MDISP, and Cosine, the differences in prediction time between the architectures showed large effect sizes. Regarding the image errors, the I-UNet performs best, although the difference to the Res-UNet is not significant, which has the second best prediction time. Regarding the prediction time the UNet outperforms the other architectures. That means

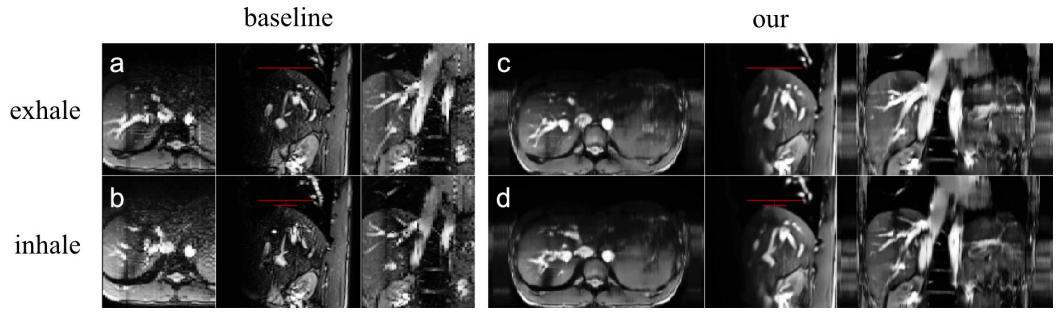

Fig. 6.7: Example reconstruction of baseline and our method, presented as axial, sagittal and coronal slices at identical temporal and spacial position, for an exhale-state (a,c) and inhale-state (b,d). Red lines indicate liver dome position of baseline reconstruction.

RQ 2 can be answered positively, the framework does work with different network architectures.

architecture	RMSE	MDISP	Cosine	Prediction Time
UNet	0.240	0.352	0.861	**0.571**
Res-UNet	0.232	0.347	**0.866**	0.839
D-UNet	0.270	0.397	0.844	1.041
I-UNet	**0.231**	**0.336**	**0.866**	1.270

Tab. 6.3: Architecture comparison

Target Registration Error For the evaluation of the TRE the U-Net was selected as the overall best performing architecture. Time is one of the most valuable resource in a clinical context and here, the U-Net outperformed the other architectures by far and performed similar with regard to the other metrics.

The experimental results for the **TRE** are shown in table 6.4. All TREs are below voxel size, with the only exception being subject S1 at a slice distance of $-3\,\mathrm{cm}$. All subjects have a similar overall TRE. The mean$\pm$standard deviation (std) of the TRE for all test subjects is $0.66 \pm 0.41\mathrm{voxel}$, or expressed in millimeter $1.19 \pm 0.74\mathrm{mm}$. One can also see that, in general, the TRE is smaller near the navigator than further away from the navigator, which will also be seen in the results of the ablation study. Answering **RQ 3**, the TRE of the method is below voxel size.

Ablation Study Fig. 6.10 shows the baseline reconstructions in the leftmost column. The other columns show reconstructions from six networks with decreasing amounts of training data (98% to 5%).

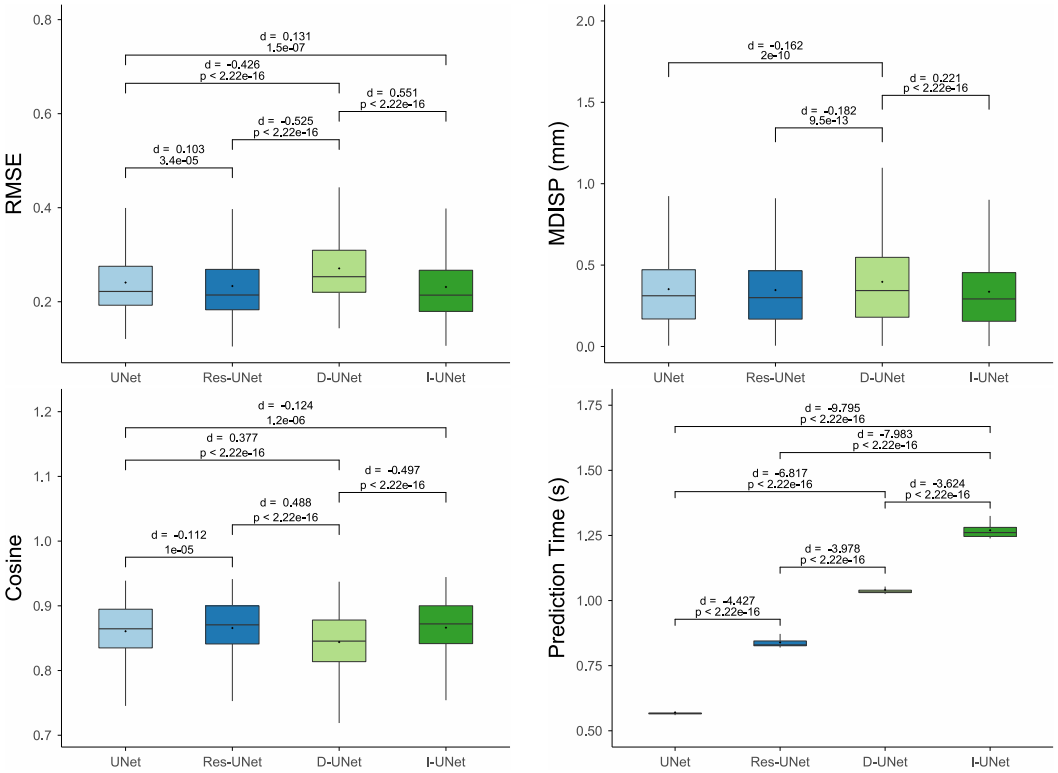

Fig. 6.8: Performance metrics for our method depending on the used architectures. RMSE, MDISP in mm, cosine and prediction time in s are given with effect size (Cohen's d) and significance level from t-test (p).

	slice distance to navigator				
subject	3 cm	0 cm	−2 cm	−3 cm	all positions
S1	0.70 ± 0.38	0.53 ± 0.37	0.92 ± 0.62	1.21 ± 0.65	0.84 ± 0.50
S2	0.66 ± 0.40	0.64 ± 0.47	0.80 ± 0.49	0.80 ± 0.54	0.72 ± 0.47
S3	0.44 ± 0.27	0.45 ± 0.34	0.60 ± 0.30	0.94 ± 0.55	0.61 ± 0.36
S4	0.45 ± 0.27	0.31 ± 0.21	0.53 ± 0.38	0.58 ± 0.33	0.47 ± 0.37
S1-S4	0.56 ± 0.33	0.48 ± 0.35	0.71 ± 0.45	0.88 ± 0.52	0.66 ± 0.41

Tab. 6.4: TREs (in voxel) of predictions using the UNet as the best overall architecture for all test subjects. Columns 2-5 show TREs per slice position, the last column shows mean TREs per subject, i.e., over all four slice positions. The last row shows mean TREs over all test subjects.

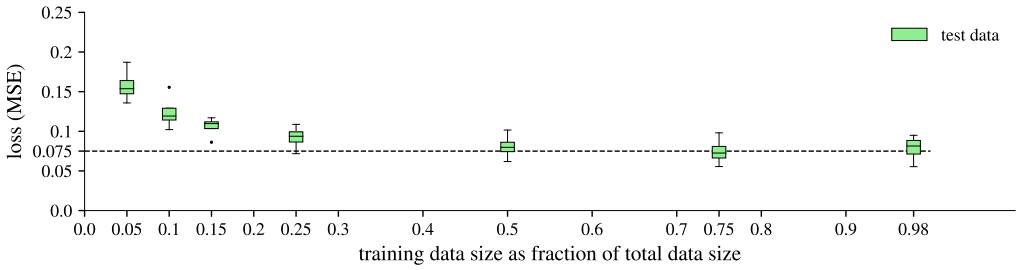

Fig. 6.9: Test data MSE as a function of the amount of available training data.

As can be seen, our method is capable of reconstructing full-liver volumes with different breathing states while capturing major and minor vessels. We observe that 4D MRI, reconstructed from $2\,\mathrm{min}$ of training data has a worse image quality than the baseline, but still looks promising. Compare that with a standard MRI acquisition, it would take roughly $2\,\mathrm{min}$ to capture only one 3D volume with comparable quality. We further observe that 50% of the training data $(24\,\mathrm{min})$ yield comparable results to 98% and the baseline quality.

This observation is also confirmed by the MSE. The mean and standard deviation of the MSE over all slice positions and test subjects as a function of available training data is depicted in Fig. 6.9. Increasing the training data size beyond 50% does not improve the loss further, as the latter plateaus at around 0.075. This indicates that $24\,\mathrm{min}$ of training data would be sufficient for a good reconstruction.

Hence, **RQ 3** can be answered with, the minimal amount of training data to achieve the same quality results as the baseline method is$24\,\mathrm{min}$.

Dependence on slice position Fig. 6.11 shows the loss as a function of the distance of the prediction to the navigator slice. Blue and green boxes represent the validation and test data losses respectively. For visualization, the distances were binned into $12\,\mathrm{mm}$ bins (3 slice positions per bin). The test data loss is comparable with validation data loss. Two effects are visible for both data sets. First, our method performs better on the left of the navigator (subjects' right) and worse on the right (subjects' left). Second, our method performs better when closer to the navigator. This is consistent with the observation made in the analysis of the TRE. Finally, the last research question can be answered with, the method does not perform equally good at all slice positions.

Comparison to Related Work The comparison of our method with the related work shows that it has the shortest acquisition and reconstruction times (see table 6.5).

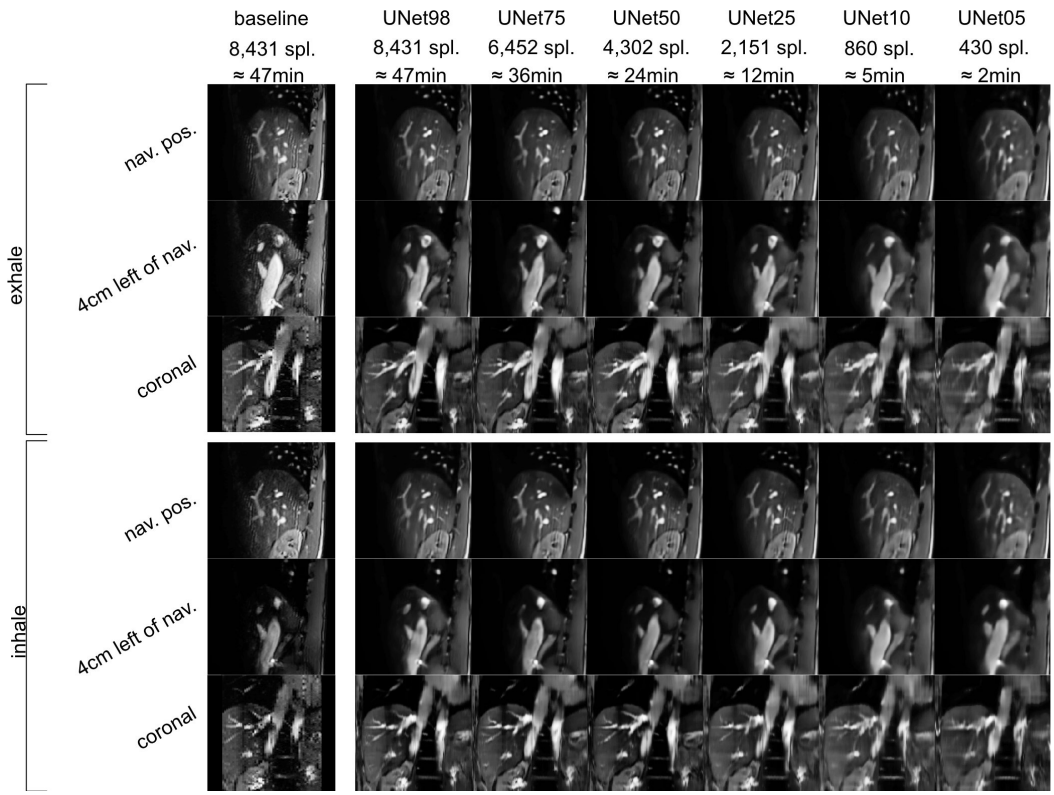

Fig. 6.10: Reconstruction results depending on training data size. Reconstructed are an inhale and exhale state. The training data size in samples (spl.) is depicted at the top. For each volume reconstruction, three slices are presented: two sagittal slices, one at navigator position (nav. pos.), one 4 cm left of the navigator and one coronal slice. Predictions are cropped to the same FOV as the baseline.

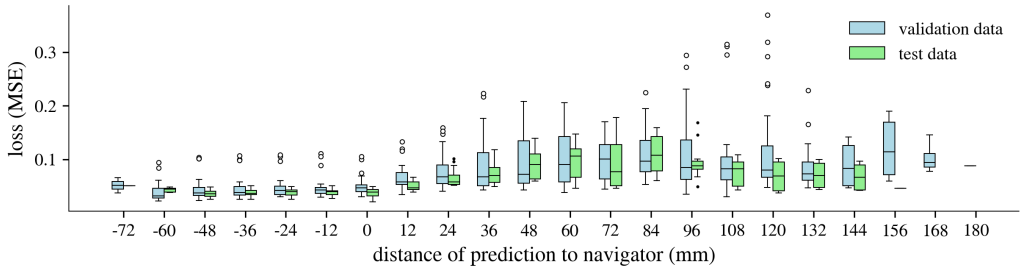

Fig. 6.11: Loss as function of distance between prediction and navigator position.

method	acq. time	recon. time
ours	**2 to** 24 **min**	**0.571** s
Yuan et al. (2019)	*	0.615 s
M. v. Siebenthal et al. (2007)	15 to 60 min	73 s
Gulamhussene et al. (2020)	15 to 60 min	24 s
Tanner et al. (2014)	**9 to** 12 **min**	-
Celicanin et al. (2015)	1/2	-
Zhang et al. (2018)	1/4	1/2 **

Tab. 6.5: Reported acquisition time (of training/stacking data) and reconstruction time (per time point). * reconstruction during acquisition, ** for sorting approaches.

6.5 Discussion

6.5.1 Interpretation of Result

Regarding the visual comparison to the baseline one can say, the predictions of the deep learning based 4D MRI framework are anatomically correct and have the same image quality as the baseline. The comparison of different architectures within the framework revealed, regarding the image quality, that the I-UNet and Res-Unet are the best architectures. However the Res-UNet must be favored over I-UNet because of its significant shorter prediction time. Sill, the UNet outperforms the other architectures by far regarding the prediction time. Because this is the most valuable clinical recourse and the difference in the image quality is small compared to the other architectures, the UNet is the overall best architecture tested in our framework.

The mean TRE is well below voxel resolution 0.66 ± 0.41voxel (1.19 ± 0.74mm), which is medically sufficient and renders the approach very promising. However, the methods accuracy is dependent on the distance of predicted slice to the navigator, especially further away from the navigator position the prediction quality decreases. Regarding the needed amount of training data one can say, that $24\,\mathrm{min}$ and more are sufficient to yield excellent image quality on a par with the baseline. However, such a lengthy acquisition time is likely to be impractical for clinical use and needs further reduction to make the approach feasible.

6.5.2 General Discussion

The three key strengths of the deep learning based 4D MRI framework, from a medical point of view, are high reconstruction accuracy, high image quality and resolution and high speed in both acquisition and reconstruction. The main contribution of the approach described in this chapter, to achieve these key strengths, is the proposition of an end-to-end trainable deep learning formulation of the 4D MRI reconstruction problem.

The chapter showed that, with our method, the acquisition time could be halved from $47\,\mathrm{min}$ (8431 training samples) to $24\,\mathrm{min}$ (4302 samples) without losing reconstruction quality. It can even be reduced to $2\,\mathrm{min}$ (430 samples) while losing some image quality. Some of the earlier proposed methods report acquisition times between 9 and $60\,\mathrm{min}$, while others report acquisition time reductions between 1/4 and 1/2. Because our new method complements these methods, it can be used in conjunction to multiply the reduction effects. Thus, a combined acquisition time reduction of up to 3/4 without loss of reconstruction quality seems achievable. In practice, reconstructing breathing sequences of arbitrary length would mean acquisition times of around $6\,\mathrm{min}$, which is a reasonable time in clinical practice. Our reconstruction quality is comparable to the state-of-the-art and robust with respect to the network architecture. Only the prediction time is highly dependent on the chosen network architecture.

In contrast to the last chapter, the breathing sate was defined in this chapter by only one navigator. This has two implications. On the one hand, breathing states are no longer interpolated between two navigators in time, but predicted, i.e., extrapolated. This means the new method is more potent than the baseline. On the other hand, it is possible that the information about whether the breathing state is in the inhalation or exhalation phase could be obscure or even lost, making it arguably more difficult to predict the next breathing state. This is a possible risk, but also makes the method as fast as possible, because it needs just one navigator slice not two.

The transitive information bridge are just an interpretation and informed the design of the model input. To get a better understanding of how the model internally actually works, more specific experiments must be conducted, using techniques from the research field regarding the explain-ability of deep learning models.

The t-test showed high significance levels even for the metrics where the effect sizes were small. This is likely due to the large sample size.

6.5.3 Future Work

For the future work, there are possibilities to further improve the framework. The primary limitation of the method, which currently impedes its clinical feasibility, lies in the substantial amount of training data required, resulting in a long prior acquisition time. In future research, it would be valuable to explore methods such as transfer learning and domain adaptation to address this limitation effectively. These approaches have the potential to reduce the dependency on extensive training data and mitigate the need for prolonged prior acquisition, making the method more viable for clinical applications. The combination of transfer learning with the deep learning base 4D MRI framework is investigated in the next chapter.

Another direction to investigate is the used architecture dimensionality. Because our formulation works with a 2D architecture, the network cannot acquire full knowledge of 3D relations between navigator and data slices. The further away the data slice is from the navigator, the looser the 3D relations become, and the poorer the reconstruction quality ends up. To mitigate this effect, one could potentially divide the volume into distance ranges and train one network for each range, thus reinforcing knowledge for 3D relations over larger distances. We expect that an ensemble of such networks will provide a considerable gain in quality for a fixed level of training data or constant quality for less training data. In general an ensemble strategy is likely to improve image quality further without needing more training data.

Additionally, in our method, one model is trained for each subject. The possibility of having only one base model that abstracts not only beyond seen breathing states, but also beyond seen subjects, or adapts quickly to new subjects is also investigated in the next chapter using a transfer learning strategy which, in turn, will further reduce the amount of necessary training data.

6.6 Conclusion

This chapter presented a novel end-to-end trainable, network agnostic, deep learning formulation of the 4D MRI reconstruction problem that predicts high quality, fast, 4D full-liver MRI. It shows that predicted real-time 4D MRI techniques are possible and provides a solution to reduce the acquisition time and effort for retrospective reconstruction approaches. Nevertheless, the method still necessitates a significant number of pre-acquisition scans for training, which remains the primary obstacle to its clinical feasibility. In conclusion, the deep learning-based 4D MRI framework

has exhibited promising results, providing motivation for further research in the field. This thesis addresses some of the ongoing research in the following chapters, presenting the findings and advancements achieved.

7

Transfer-Learning is a Key Ingredient to Fast Deep Learning-Based 4D Liver MRI Reconstruction

Synopsis

The last chapter showed that deep learning based real-time 4D MRI prediction methods are promising. However, the prior acquisition time of the proposed method was still to long to be clinically feasible. This chapter describes how a transfer learning (TL) strategy can be used to drastically reduce prior acquisition time and an ensembling strategy is proposed to realize an uncertainty estimation and improve image quality of the prediction. Models trained from scratch on target domain data are compared with models fine-tuned from a pre-trained base model. Significant improvements (P<.001) of the root mean squared error (RMSE) of up to 12% *(effect size d=-0.5),the mean displacement (MDISP) of up to* 12.5% *(d=-0.263), and the deformation-normalized RMSE (DN_RMSE) of up to* 15% *(d=-0.679) are reported. It is shown that the prior acquisition time can be significantly shortened down to* $2\,\mathrm{min}$ *and still preserving a high level of image detail. This shows that TL significantly reduces beforehand acquisition time and improves reconstruction quality, rendering it a key component in making 4D MRI clinically feasible.*

About this chapter Parts of this chapter have been published in: Gino Gulamhussene, Marko Rak, Oleksii Bashkanov, Fabian Joeres, Jazan Omari, Maciej Pech, and Chistian Hansen (2023). "Transfer-learning is a key ingredient to fast deep learning-based 4D liver MRI reconstruction". Scientific Reports, 13(1), 11227. (Gulamhussene et al., 2023b)

7.1 Introduction

Current 4D reconstruction techniques are unsuitable for most interventional settings because they are limited to specific breathing phases, lack temporal/spatial resolution, and have long prior acquisitions or reconstruction times. As set out in the last chapter, deep learning-based 4D MRI approaches promise to overcome these shortcomings. However, acquiring real-time 4D MRIs of a large target region during an intervention is currently not feasible due to the need for a significant amount of reference data beforehand. Although the framework presented in the last chapter demonstrated promising results using only $24\,\mathrm{min}$ of training data, this timeframe is still impractical for routine clinical settings where time is crucial. Additionally, there are limits to the specific absorption rate (sar) allowed during MRI imaging, and these limits are likely to be exceeded during prolonged imaging. On the other hand, reducing the prior acquisition time and training data, domain shift becomes a considerable problem, which will be shown in this chapter. Consequently, the effective application of 4D MRI in the intervention room remains challenging.

In essence, using previous works, one had to choose between long acquisition times or limited breathing phase support, i.e., no irregular breathing, none of which is clinically acceptable.

In the work presented in this chapter, the shortcoming of our previously proposed methods' long prior acquisition time is addressed. First, domain shift is identified as a major issue for DL-based 4D MRI prediction, which gets more severe the smaller the amount of available target domain data is, which fits into the observations of a recent 2021 survey of Guan et al. (2021). Second, it is showed that the beforehand acquisition time can be substantially reduced (from $24\,\mathrm{min}$ to $2\,\mathrm{min}$) by using transfer learning (TL) techniques without losing the support for irregular breathing. Third, combining multiple models in an ensemble strategy, mitigates the negative impact of reduced training data and improves the accuracy and reliability of the predictions.

7.2 Materials and Methods

This section first gives an overview of the parts of the data base that were used for the development and evaluation of the TL and ensembling approach that is used to reduce the prior acquisition time of the prediction framework. Then, a short recap of the framework is given, lastly, the TL and ensembling method is described.

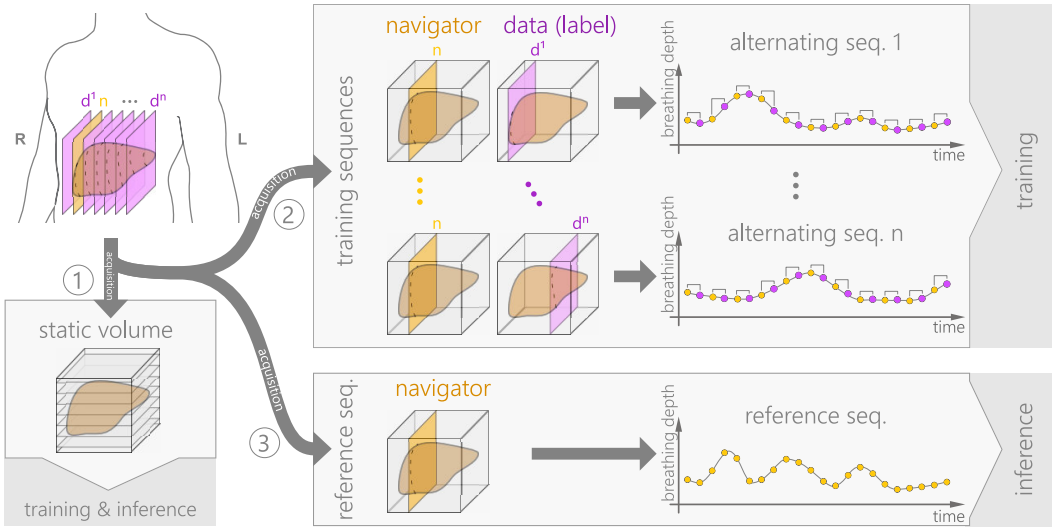

Fig. 7.1: 1) a static volume, 2) several alternating dynamic sequences (brackets indicate pairs of navigators and data slices), and 3) a dynamic reference sequence.

7.2.1 Training Data

In this chapter all three parts of the data set (see gray boxes in Fig. 7.1) are used. They are briefly described in the following. The detailed description is found in chapter 4.

Static volume

The static 3D liver volume is used as an anatomical reference during training and inference. It was acquired with a STAR VIBE MR Sequence (320×320×72-88 matrix size, 3 mm slice thickness, $1.19 \times 1.19\,\mathrm{mm}^2$ in-plane resolution).

Training sequences

The training sequences are several dynamic 2D sequences that were acquired during free breathing. In these sequences navigator slices alternate with data slices. Navigators and data slices form pairs and are used as training samples. While the navigator slice position is fixed in the right liver lobe, the data slice position is unique for all sequences, equidistantly sampling the liver from right to left. The navigator serves as a respiratory motion signal. Each training sequence consists of 175 navigators and 175 data slices. For each subject, the number of training

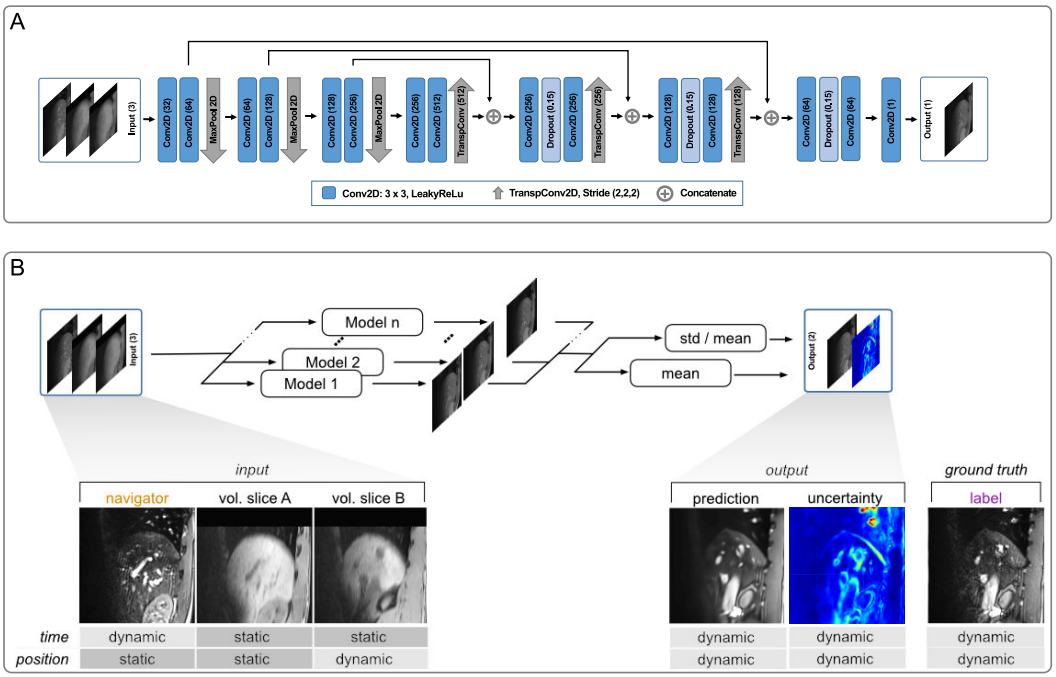

Fig. 7.2: A) U-Net architecture, with three-channel input. Blue boxes are convolutions; grey arrows are max pooling or upsampling, and pluses are feature map concatenations. B) Ensembling of n models and generation of uncertainty map.

sequences ranged between 38 and 57, depending on the size of the subjects' livers. Overall the acquisition time ranged between 40 and $80\,\mathrm{min}$ per subject.

Reference sequence

Reference sequences are dynamic 2D sequence of navigator slices only that were also acquired during free breathing. The navigator position is the same fixed position as in the training sequences. The reference sequence contains a natural succession of different breathing cycles/patterns. It is used for inference as a respiratory reference, i.e., a breathing signal. The reference sequence comprises 513 time points in our data, covering 85 seconds (typically about 20 breathing cycles).

Both, training as well as reference sequence were acquired using the TRUFI MR sequence (14×176 matrix size, $1.8 \times 1.8\,\mathrm{mm}^2$ in-plane resolution, $4\,\mathrm{mm}$ out of plane resolution, $255 \times 320\,\mathrm{mm}^2$ FOV). The acquisition time was $166\,\mathrm{ms/slice}$.

7.2.2 Deep learning prediction of 4D MRI

Deep learning formulation

To recap the deep learning base 4D MRI framework. A deep network with three 2D input channels is trained using training sequences together with slices of the static volume. Each training input corresponds to a specific subject. However, samples from different subjects can be used. A training input consists of three channels (see Fig. 7.2). Pairs of navigator and data slices are taken from the training sequences of a subject. The navigator is fed to the first channel. The data slice serves as ground truth (label). Two slices are sampled from the static volume (from the same subject): one slice at the navigator position is fed to the second channel, and one at the ground truth position (the slice to be predicted) is fed to the third and last channel.

The navigator (first channel) is dynamic in time and static in its position. It determines (shows) the breathing state. The volume slice at the navigator position (second channel) is static in both time and position and acts as a still reference to the dynamic navigator. It contains information on the relationship between the two different MR contrasts of the TRUFI and STAR VIBE MR images. The volume slice at the label position (third channel) is static in time but dynamic in position and acts as a still reference to the dynamic output we seek to predict and expresses the position that should be predicted.

During inference, the first channel determines the breathing state of the slice that is to be predicted, and the third channel determines its position. That way any current breathing state (time domain) is reconstructed by providing a real-time navigator (acquired during the intervention) at any position (space domain) by choosing the proper position from the static volume (acquired before the intervention). By simply inferring all positions in one batch (in one forward pass), a total 3D volume for a time point is reconstructed. On a GeForce GTX 1080, this takes $\leq 600\,\mathrm{ms}$, yielding real-time 4D MRI. Furthermore, if reconstruction is done retrospectively - and graphics card memory permits - a whole 2D+T series at a fixed slice position or even a full 4D reconstruction can be performed in one batch.

Network architecture and training

The three-channel input is processed in standard U-Net encoding and decoding paths. A leaky rectified linear unit (slope=0.1) follows each convolutional layer. The

convolutions are padded to keep the size of feature maps and input constant. The second convolutional layer in each block doubles the number of features, increasing the network's capacity. A MaxPooling operation follows the first three blocks. In the first convolutional layer, 32 filters process the 128×128×3 input to the network. Following the architecture results in 512 feature maps in the latent feature space. The decoding reconstructs the image from the latent space. To this end, three blocks of two transposed convolutional layers are employed that up-sample the features. Between every two transposed convolutional layers, a dropout layer is used. With each up-sampling, the filter size is halved. A final 1×1 convolutional layer outputs the reconstructed MR image.

The network was implemented with Keras (Chollet et al., 2015). The total of 6.8 million parameters are trained by an Adam optimizer (Kingma et al., 2014) (learning rate = 0.0004). In the training run, a total of 200 epochs were performed using the mean squared error (MSE) as the loss function. The training was conducted with a batch size of 64. Checkpoints were employed and the model with the best validation loss was used. Z-score normalization was performed, also known as whitening, to the image intensities of each subject. This normalization process ensured that the intensities had a zero mean and unit variance. It is important to note that this normalization was reversed after the prediction stage and specifically before the uncertainty map generation processes in the case of ensembling. The training data was augmented in a physiologically plausible range as described in the last chapter to facilitate robustness. Random augmentation was seeded for reproducibility. To simplify the processing, all images were re-sampled to $1.8\,\mathrm{mm}^3$ voxels.

Transfer learning

As will be shown in the next section, domain shift is a problem in MRI liver data and results in a discrepancy in model performance. This issue was addressed by fine-tuning a pre-trained model to a new target subject, because fine-tuning is a simple to use and effective technique. Its practicality and effectiveness make it particularly advantageous in a clinical context. Let **S** be the source domain and $\mathbf{s} \in \mathbf{S}$ be the subjects of the source domain. Likewise let **T** be the target domain and $\mathbf{t} \in \mathbf{T}$ be the subjects of the target domain. Transfer learning in the form of fine-tuning is used to reduce the discrepancy in model performance in **S** and **T**. Specifically, let $\mathbf{M}^{j}_{pre}$ be a pre-trained model that was trained on data from all N source domain subjects $[\mathbf{s}_1, \mathbf{s}_N] \in \mathbf{S}$, where j denotes the minutes of training data per subject **s**. $\mathbf{M}^{j}_{pre}$ is then fine-tuned with i minutes of training samples from a new subject $\mathbf{t} \in \mathbf{T}$ using the same training parameters as were used for the training of

the pre-trained model (200 epochs, MSE loss, 64 batch size, data augmentation), resulting in the fine-tuned model $\mathbf{M}^{i}_{pre+TL}$.

Ensembling and uncertainty map

We propose to combine an ensembling strategy together with the transfer learning strategy with our 4D MRI framework. This is illustrated in Fig.7.2 B). While fine-tuning does enhance prediction quality, when limited training samples are available, it may not completely mitigate the decrease in prediction quality caused by the smaller training data set. Ensembling plays an important role in addressing this issue. By combining multiple models, ensembling significantly improves the overall prediction quality and helps to mitigate the negative impact of the reduced training data set. To employ the ensembling strategy, N models were pre-trained, each starting from a random parameter initialisation. These N models were fine-tuned to a new subject following the training as described before. To form the final 4D MRI the predictions of the individual models in the ensemble are averaged. An uncertainty map is generated by computing the Coefficient of variation between the predictions. For that, after the normalization was reversed, the voxel wise standard deviation is dividing by the voxel wise mean.

7.3 Experimental Design

7.3.1 Research Questions and Hypothesis

In this chapter four RQs are addressed. They are summarized in Tab. 7.1. The RQs and the corresponding null hypothesis (H_0) and alternative hypothesis (H_A) are described in the following.

RQ 1: Is domain shift a significant problem? $\mathbf{H}_0$ **1:** Domain shift *is not* a significant problem. $\mathbf{H}_A$ **1:** Domain shift *is* a significant problem.

RQ 2: Can TL reduce training sample size and improve image quality of a fine-tuned model compared to a pre-trained model? $\mathbf{H}_0$ **2:** TL *can not* reduce training sample size and improve image quality. $\mathbf{H}_A$ **2:** TL *can* reduce training sample size and improve image quality.

RQ 3: Can TL reduce training sample size and improve image quality of a fine-tuned model compared to a directly trained model? $\mathbf{H}_0$ **3:** TL *can not* reduce training

RQ 1:	Is domain shift a significant problem?
RQ 2:	Does TL have an effect compared to pre-trained models?
RQ 3:	Does TL have an effect compared to directly trained models?
RQ 4:	Does ensembling improve image quality?

Tab. 7.1: Research questions

sample size and improve image quality. $\mathbf{H}_A$ **3:** TL *can* reduce training sample size and improve image quality.

RQ 4: Does ensembling improve the image quality? $\mathbf{H}_0$ **4:** Ensembling *does not* improve the image quality. $\mathbf{H}_A$ **4:** Ensembling *does* improve the image quality.

7.3.2 Training, Validation, Test Split

The 20 subjects were divided into a source domain **S**, containing 16 subjects, and a target domain **T**, containing 4 Subjects. In both **S** and **T**, the first half of each training sequence was used as training data and the second half as validation data.

7.3.3 Experiments

To quantitatively assess model performance and for statistical analysis, the RMSE, MDISP, and DN_RMSE were used as error measures that express the dissimilarity of predicted MR slice and ground truth. All three error measures were described in 2.6.

Experiment 1: Test for domain shift In this study the term domain shift is used in a general way, where it refers to the situation that the data distribution in the training set is different from the test set. And that this leads to a decrease in model performance. In clinical settings, the quantity of available training data is limited, and there is a high likelihood that a new subject may not be adequately represented by the training set distribution. The inadequate representation of the new subject by the training set can be considered as domain shift. In our case, a small training distribution does not faithfully represent the following variations: liver shape and size, body height, abdominal girth (and, consequently, signal-to-noise ratio), body fat, sex, and age. This list might not be exhaustive.

Remember, $\mathbf{M}_{pre}^{24}$ is a model pre-trained on all 16 Subjects from the source domain **S**, using $24\,\mathrm{min}$ worth of training samples per subject. Of course, it would be best

if it could be applied to a new subject **t** $\in$ **T** directly and without any adaptation. However, this requires that there is no domain shift present between **S** and **T**. To test this, the domains are compared in two ways. First, the performance of $\mathbf{M}_{pre}^{24}$ is compared between validation data (from **S**) and test data (**T**) using the MDISP and DN_RMSE. To that end, 50% of test samples were randomly chosen from the first 10 seconds of the second half of each training sequence, i.e., for each subject (in **S** and **T**) and slice position. Then both similarity measures were computed for all predictions of the test samples. Second, the anatomical variance was assessed visually using the navigator frames.

Experiment 2: Pre-trained vs. TL and influence of source domain data availability Because domain shift is a challenge in deep learning-based 4D MRI prediction, we propose to employ TL. The effect of TL on our models is evaluated by comparing $\mathbf{M}_{pre}^{j}$ ($j \in [2, 5, 12, 24]$) with $\mathbf{M}_{pre+TL}^{2}$ regarding their performance in **T**. Where $\mathbf{M}_{pre+TL}^{2}$ is the result of fine-tuning $\mathbf{M}_{pre}^{j}$ with 2 minutes of samples from **T** (720 samples = $2\,\mathrm{min}$ acquisition time). By that, it is also analyzed how the source data amount j influences the effect of TL. For comparison, the RMSE, MDISP, and DN_RMSE are used.

Experiment 3: Direct learning vs. TL and the influence of target domain data availability It is evaluated whether TL is beneficial compared to directly learning a model from scratch in the target domain. Moreover, it is evaluated how the target sample availability influences that effect regarding the effect size. To that end, models were trained directly from scratch on samples from **T**. They were compared with fine-tuned models. Let $\mathbf{M}_{direct}^{i}$ be a directly learned model and let $\mathbf{M}_{pre+TL}^{i}$ be a model fine-tuned from $\mathbf{M}_{pre}^{2}$, where $i \in [1, 2, 5, 12, 24, 47]$. $\mathbf{M}_{pre}^{2}$ was chosen as the base model because j showed virtually no influence on model performance in **T**. Furthermore, acquiring only a few samples to train a base model in a real-world scenario would be more economical. The model performance was tested dependent on the availability of target domain samples from 1 minute to 47 minutes (see the bottom row in Fig. 7.4). For each target data availability level i and target subject t, one model was trained directly and one with TL (in total, 48 models).

Experiment 4: TL vs. TL+Ens This experiment evaluates whether the combination of transfer learning with the ensembling strategy (TL+Ens) enhances the model performance. For that, ensembles of fine-tuned models of different ensemble sizes were compared with regard to RMSE, MDISP, and DN_RMSE. Where the ensemble size N=1 represents only TL, i.e. no ensembling. A one-factorial ANOVA (Analysis of variance) was performed to test for a primary effect of the ensemble size, which reveled a significant effect. A post-hoc pair-wise Tukey's test was performed for the

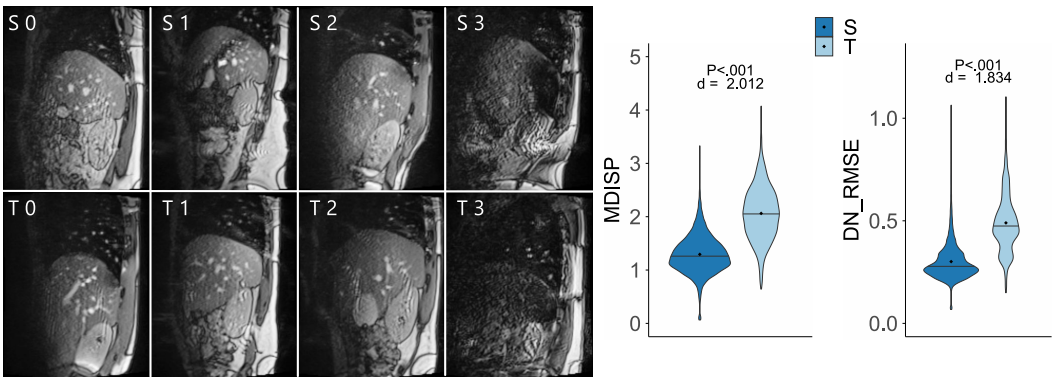

Fig. 7.3: Navigators show considerable variance in anatomy, as illustrated by four source domain subjects (top row) and four target domain subjects (bottom row). The violin plot (right) shows the prediction error of a pre-trained model in the source domain (S) and the target domain (T).

RMSE, MDISP, and DN_RMSE independently using p-adjustment. The pair-wise effect size was computed, using Cohen's d.

7.4 Results

7.4.1 Results of experiment 1: Domain shift

The MDISP and DN_RMSE distributions are visualised in a violin plot (see Fig. 7.3). The violin plots show non-normal distributions with different mean. Because a Shapiro-Wilk Test (n = 4000) and Kolmogorov-Smirnov test also showed that the distributions are not normally distributed (p<.001). The Wilcoxon rank sum test (m = 3040, n = 12352) was used to test for significance of the distribution shift. The null hypothesis $\mathbf{H}_0$ **1** of no shift in error distribution was rejected in favour of the alternative hypothesis $\mathbf{H}_A$ **1** at a significance level of p<.001. The mean of MDISP and DN_RMSE are 0.30 and 1.29 in **S** and 0.49 and 2.06 in **T**. The effect size is quantified with Cohen's d (n=3040, m=12352). The effect size is large with d = 2.01 and 1.834. The visual comparison of the navigators shows variability in liver anatomy across subjects concerning the superior-inferior extent of the liver and the number and arrangement of vessels. Research question **RQ 1** can be answered with yes, domain shift is a significant problem in deep learning based 4D MRI.

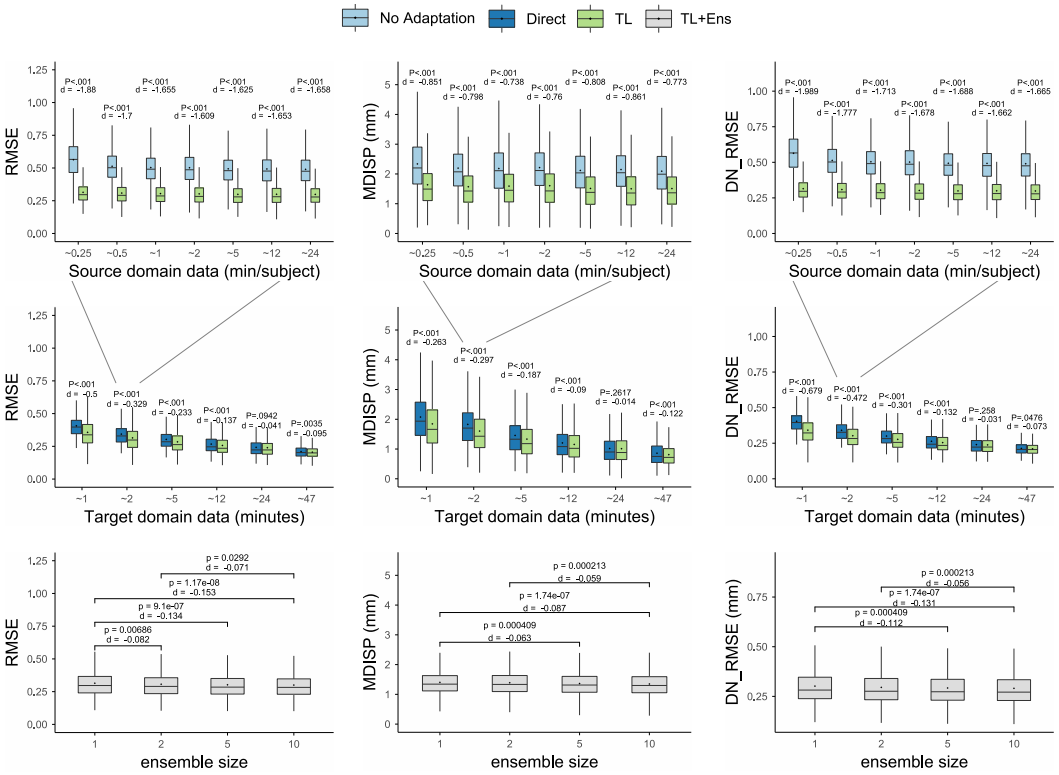

Fig. 7.4: Top: Comparison of no adaptation and TL at different levels of source domain data. Middle: Comparison of Direct learning and TL at different levels of target domain data. Bottom: Comparison of ensemble sizes.

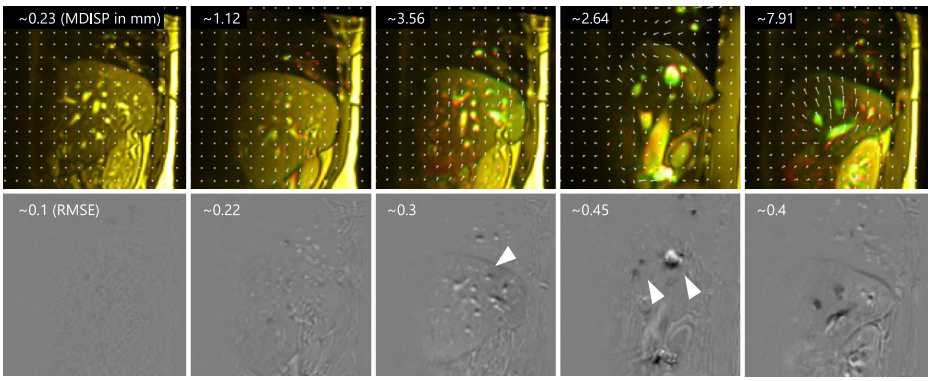

Fig. 7.5: Top row: displacement fields with a composite of (red) labels and (green) predictions as reference. Bottom row: intensity differences images.

7.4.2 Results of experiment 2: Pre-trained vs. TL and influence of source domain data availability

The top row of box plots in Fig. 7.4 shows the results of this experiment. Two observations can be made. First, transfer learning improves the model performance in the target domain for all tested measures. All tested measures show significant differences between TL and pre-trained models, with a significance level of p<.001. For example, at $2\,\mathrm{min}$ source domain data the mean RMSE is improved from 0.51 to 0.31 (p<0.001, d=-1.609), the mean MDISP is improved from 2.22 to 1.61 (p<0.001, d=-0.76), and the mean DN_RMSE is improved from 0.5 to 0.3 (p<0.001, d=-1.678). Significances were computed using the Wilcoxon rank sum test (m = 3040, n = 12352) after confirming none normal distributions using the Shapiro-Wilk test (n = 3040) and Kolmogorov-Smirnov test. High effect sizes can be observed with $|\mathrm{d}| > 1.6$ for RMSE and DN_RMSE and medium effect sizes with $|\mathrm{d}| > 0.7$ for MDISP.

That means, the null hypothesis $\mathbf{H}_0$ **2** must be rejected in favour of the alternative hypothesis $\mathbf{H}_A$ **2** and the research question **RQ 2** can be answered with yes, TL has an effect on the model performance compared to pre-trained models. Second, the amount of source domain data (beyond $\sim 1\,\mathrm{min}\,/\mathrm{subject}$) has little to no influence on the effect size d. It also does not affect the performance of either $\mathbf{M}^{j}_{pre}$ or $\mathbf{M}^{2}_{pre+TL}$ in $\mathbf{T}$. In table 7.2 means and 95th percentiles are reported.

	mean													
	15s		30s		1min		2min		5min		12min		24min	
	no A	TL	no A	TL	no A	TL	no A	TL	no A	TL	no A	TL	no A	TL
RMSE	0.57	0.33	0.52	0.32	0.51	0.32	0.51	0.31	0.5	0.31	0.5	0.31	0.5	0.3
MDISP	2.65	1.64	2.21	1.57	2.19	1.59	2.22	1.61	2.12	1.51	2.15	1.5	2.1	1.51
DN_RMSE	0.56	0.31	0.51	0.31	0.5	0.3	0.5	0.3	0.49	0.3	0.49	0.3	0.49	0.3
	95th percentile													
RMSE	0.83	0.89	0.77	0.48	0.76	0.48	0.76	0.48	0.75	0.48	0.75	0.47	0.74	0.46
MDISP	4.42	2.98	3.87	2.93	3.77	3	3.82	3.11	3.65	2.81	3.65	2.8	3.63	2.79
DN_RMSE	0.82	0.46	0.76	0.45	0.76	0.45	0.76	0.46	0.75	0.45	0.75	0.45	0.74	0.45

Tab. 7.2: Comparison of our method with no adaptation (no A) and with TL and different availability of source domain data.

7.4.3 Results of experiment 3: TL vs. Direct learning and the influence of target domain data availability

For target data availability between 1 and 12 minutes, significant improvements ($p<.001$) are observed when using TL concerning RMSE, MDISP, and DN_RMSE, and visual assessment reveals detail gain (see Fig. 7.6). For example, at $2\,\mathrm{min}$ target domain data the mean RMSE is improved from 0.34 to 0.31 ($p<0.001$, d=-0.329), the mean MDISP is improved from 1.83 to 1.61 ($p<0.001$, d=-0.297), and the mean DN_RMSE is improved from 0.34 to 0.3 ($p<0.001$, d=-0.472). That means, the null hypothesis $\mathbf{H}_0$ **3** must be rejected in favour of the alternative hypothesis $\mathbf{H}_A$ **3** and the research question **RQ 3** can be answered with yes, TL has an effect on the model performance compared to directly trained models.

Beyond the level of $12\,\mathrm{min}$, improvements are not significant. Effect sizes are largest (small to medium) between 1 and 12 minutes when few target samples are available. The effect size becomes negligible when large amounts of target samples are available. The Wilcoxon rank sum test (m = 3040, n = 3040) was used to test for significance after checking that the distributions are not normally distributed using the Shapiro-Wilk test (n = 3040) and Kolmogorov-Smirnov test. Effect sizes are reported as Cohen's d. In table 7.3 means and 95th percentiles are reported. Figure 7.5 illustrates the image quality and displacement fields of predictions for increasing MDISP and RMSE values. 4D visualizations are presented in this video: https://youtu.be/w1CAzOr2XEY.

	mean											
	1min		2min		5min		12min		24min		47min	
	Direct	TL	Direct	TL	Direct	TL	Direct	TL	Direct	TL	Direct	TL
RMSE	0.41	0.36	0.34	0.31	0.3	0.28	0.27	0.26	0.24	0.24	0.21	0.2
MDISP	2.08	1.85	1.83	1.61	1.46	1.33	1.2	1.15	1.01	1.01	0.86	0.81
DN_RMSE	0.4	0.34	0.34	0.3	0.3	0.28	0.26	0.25	0.24	0.24	0.21	0.21
	95th percentile											
RMSE	0.59	055	0.49	0.48	0.44	0.44	0.4	0.4	0.38	0.37	0.29	0.28
MDISP	3.67	3.57	3.29	3.11	2.67	2.65	2.23	2.27	1.97	1.98	1.64	1.54
DN_RMSE	0.57	0.51	0.48	046	0.42	0.43	0.39	0.39	0.36	0.36	0.29	0.28

Tab. 7.3: Comparison of our method with direct learning and with TL. Availability of target domain data given in minutes.

mean (95th percentile)				
	N = 1	2	5	10
RMSE	0.31 (0.49)	0.31 (0.47)	0.3 (0.47)	0.3 (0.46)
MDISP	1.58 (2.98)	1.56 (3.04)	1.53 (3.04)	1.51 (2.98)
DN_RMSE	0.3 (0.46)	0.3 (0.45)	0.29 (0.44)	0.29 (0.44)

Tab. 7.4: Comparison of ensemble sizes N.

7.4.4 Results of experiment 4: TL vs. TL+Ens

One can see that ensembles (TL+Ens) of size N=5 and 10 perform significantly better than N=1 (TL) in all tested metrics. For N=10 the mean RMSE is improved from 0.31 to 0.3 ($p<0.001$, d=-0.153), the mean MDISP is improved from 1.58 to 1.51 ($p<0.001$, d=-0.087), and the mean DN_RMSE is improved from 0.3 to 0.29 ($p<0.001$, d=-0.131). Although ensembling provides some benefits, the effect size is relatively small, suggesting that our TL strategy has reached a saturation point in terms of quantitative result quality. However, based on a subjective perspective, in an interview, senior radiologists with extensive experience consistently preferred the results of the TL+Ens approach over the TL-only results in all tested cases.

Together with the significant differences in mean RMSE, MDISP, and DN_RMSE, that means, the null hypothesis $\mathbf{H}_0$ **4** must be rejected in favour of the alternative hypothesis $\mathbf{H}_A$ **4** and the research question **RQ 4** can be answered with yes, ensembling does improve prediction and image quality. The boxplots and all pairwise significances and Cohen's d are presented in Fig. 7.4. The mean and 95th percentile are reported in Table 7.4.

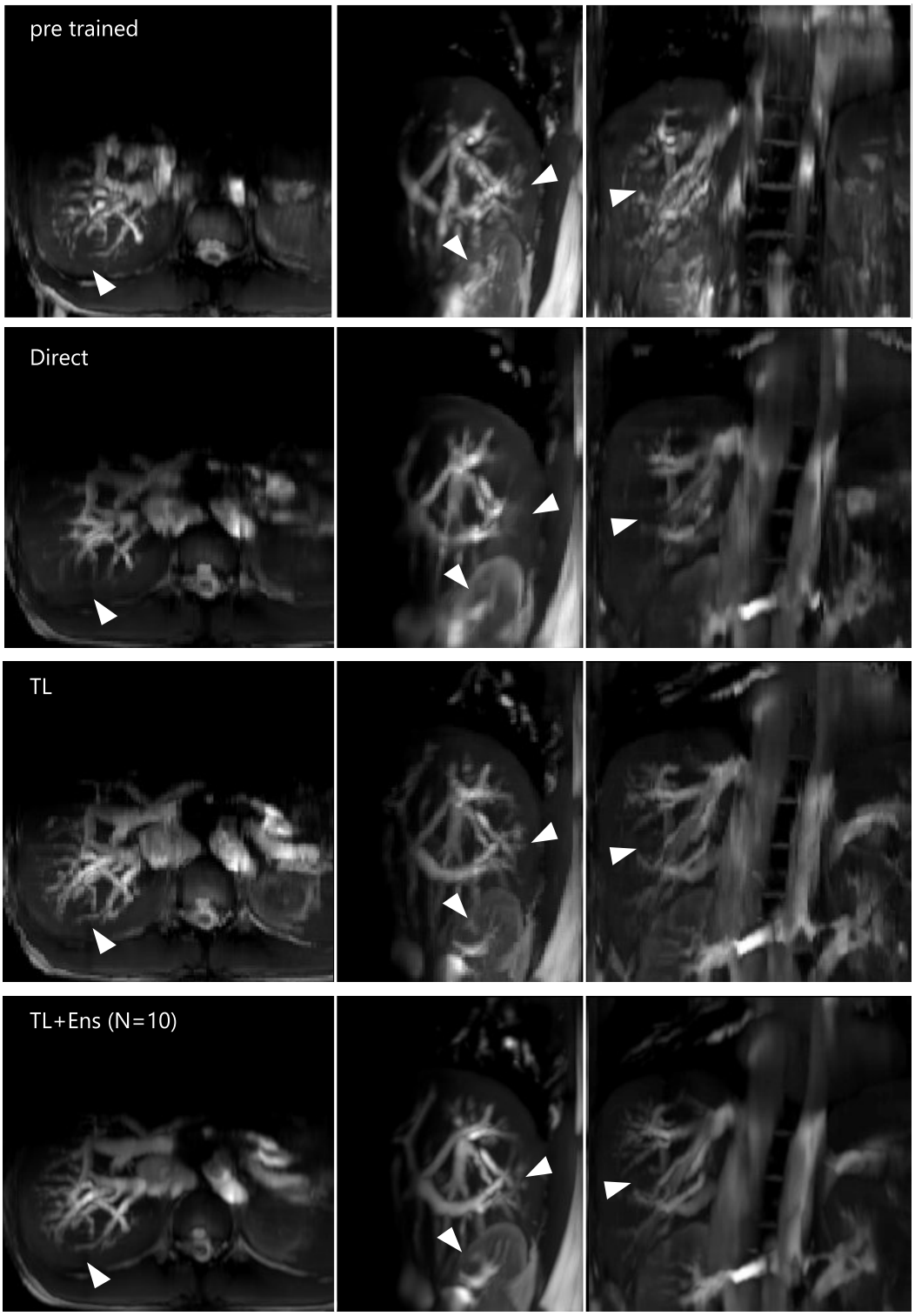

Fig. 7.6: From top to bottom predictions of: $\mathbf{M}^2_{pre}$, $\mathbf{M}^2_{direct}$, $\mathbf{M}^2_{pre+TL}$, ensemble of $10 \times \mathbf{M}^2_{pre+TL}$. Arrows indicate places of varying detail and image quality.

7.5 Discussion

7.5.1 Interpretation of Result

The results of experiment 1 show, small training data sets especially when they contain few subjects are not representative of the population causing a domain shift between training data and unseen subjects. This makes it hard for a deep learning based 4D MRI method to generalize to unseen subjects. The results of experiment 2 show that in those cases where few training samples are available fine tuned models outperform pre-trained models significantly. The advantage of TL over a pre-trained model is greatest when few training samples are available but becomes negligible for training sample sizes of $24\,\mathrm{min}$ and beyond. However, this matches with the clinical need for short acquisition times. One can say that in cases where only few subjects are available for training a base model, transfer learning is a key component of the method.

The results from experiment 4 indicate that ensembling also leads to a significant improvement in image quality. However, the magnitude of this improvement is not as substantial as that achieved through TL. This could suggest two possibilities. Firstly, it could imply that TL is more effective than ensembling in enhancing image quality. Alternatively, it could indicate that the image quality has reached a point of saturation, where further improvements are minimal. Based on the experimental results alone, it is not possible to draw a definitive conclusion. To ascertain this, ensembling should be tested in isolation, without the inclusion of TL, in order to evaluate its independent impact on image quality. At $2\,\mathrm{min}$ worth of training samples, our method requires a fraction of beforehand acquisitions compared to the related work.

7.5.2 General Discussion

The main advantage of utilizing TL and ensembling in our DL-based 4D MRI method is that it dramatically reduces the effect of domain shift. Moreover, the amount of target domain samples can be halved without hampering the model's performance compared to direct learning. From a clinical perspective, TL makes our method more economical because less beforehand acquisition and, therefore, less patient time in the MRI machine is needed. This is where our method stands out the most from the related work. It enables short pre-imaging times while achieving high prediction quality concerning RMSE, MDISP, and DN_RMSE comparable with the related

work. We evaluated our method with different amounts of training data for fine-tuning and believe 2 minutes yield a good balance between short acquisition time and high prediction quality. With 2 minutes, our method achieves a mean MDISP below voxel size with the 95th percentile below two voxels. Unlike the related work, our method is an extrapolation technique fast enough to predict real-time 4D MRI during an intervention, which is another unique strength. It should be noted that comparing the related work with our method regarding MDISP is a bit unfair because interpolation, where the used temporal context can extend into the future, is easier than extrapolation. Nonetheless, our method can also be used retrospectively and still be competitive.

It should be noted that although most tests showed high significance for our experiments, this is not the main point, especially where the effect size is small. In these cases, the high significance levels are caused by the large statistical sample size. Overall the effect size is of greater relevance. We have shown that the effect of TL is greatest when few training samples are available but becomes negligible for training sample sizes of $24\,\mathrm{min}$ and beyond. However, this matches with the clinical need for short acquisition times.

We received positive feedback from two senior radiologists with extensive experience in image-guided liver interventions, who confirmed that the presented results would offer significant benefits if implemented in clinical practice. They preferred the TL+ensemble. Specifically, the translation of our work to the clinic could yield significant advantages in interventional planning and simulation. This would only be possible because of the very short pre-acquisition time. The significant reduction in pre-acquisition time is crucial for two reasons. Firstly, time is a critical clinical resource. Reducing the time required for pre-acquisition allows for more efficient and streamlined imaging procedures. Secondly, there are strict limits on the specific absorption rate (SAR), which measures the amount of energy absorbed by the patient during the MRI scan. Prolonged acquisition times could potentially exceed these limits and pose safety risks. Therefore, the ability to shorten the pre-acquisition time is not only advantageous for time management but also for ensuring compliance with SAR regulations.

7.5.3 Future Work

The data set used in this study contains only healthy subjects. New studies are needed to conclude how well the 4D MRI models generalize to patient data from image guided liver interventions and other clinical settings.

Fine-tuning was chosen as a simple yet effective way of transfer learning to exemplify the novel combination of transfer learning with the deep learning based 4D MRI method. Of course, more advanced techniques could help to gain additional quality, which should be investigated in the future.

At $2\,\mathrm{min}$ worth of training samples, our method requires a fraction of beforehand acquisitions compared to the related work but has a larger MDISP. It would be interesting to quantify the benefit of improving sub-millimeter precision in the context of medical imaging, where voxel sizes typically range from $1\,\mathrm{mm}$ to $2\,\mathrm{mm}$ and whether a mean displacement of < 1 voxel might be sufficient. We see a few avenues to improve our method for future work. First, in the case of retrospective use, it would be interesting to increase the amount of training data by incorporating navigator interpolation (Zhang et al., 2018; Karani et al., 2018), and data interpolation (Tanner et al., 2014) to double the temporal resolution to $83\,\mathrm{ms}$ to increasing prediction quality. Second, it would be interesting to investigate the use of coordConv layers (R. Liu et al., 2018) in place of normal convolutions to improve prediction quality. This seems very promising because the spatial component of the learning task is dominant. Lastly, a 3D architecture instead of a 2D one might make it easier to learn the 3D spatial relations of the liver motion. In that case, the training task could also be reformulated to directly predict the 3D motion field, which would be beneficial for use in radiation therapy.

We received positive feedback from two senior radiologists with extensive experience in image-guided liver interventions, who confirmed that the presented results would offer significant benefits if implemented in clinical practice. They preferred the TL+ensemble. Specifically, the translation of our work to the clinic could yield significant advantages in interventional planning and simulation. This would only be possible because of the very short pre-acquisition time. The significant reduction in pre-acquisition time is crucial for two reasons. Firstly, time is a critical clinical resource. Reducing the time required for pre-acquisition allows for more efficient and streamlined imaging procedures. Secondly, there are strict limits on the specific absorption rate (SAR), which measures the amount of energy absorbed by the patient during the MRI scan. Prolonged acquisition times could potentially exceed these limits and pose safety risks. Therefore, the ability to shorten the pre-acquisition time is not only advantageous for time management but also for ensuring compliance with SAR regulations.

The senior radiologists suggested, that, in future research, it would be intriguing to adapt our method to simulate the breathing motion of planning data from patients.

7.6 Conclusion

This chapter presented the utilization of TL and an ensembling strategy to substantially reduce beforehand acquisition time and improve the prediction quality of a DL-based 4D MRI prediction model. The approach uses only a few training samples for each new subject. Both TL and ensembling can be combined with the 4D MRI method in both use cases: the real-time use, predicting 4D MRI during image-guided interventions and the retrospective use, i.e., creating a 4D MRI as a precursor for a respiratory motion model for intervention planning or radiotherapy. We believe DL-based real-time 4D MRI with high spatial and temporal resolution has the potential to impact image-guided interventions and radiation therapy because it can help to solve the problem of organ motion without interfering with the clinical workflow. Reducing the required training data to a minimum while maintaining the prediction quality is a crucial step in advancing towards that goal.

8

Using Training Samples as Transitive Information Bridges in Predicted 4D MRI

Synopsis

The last chapter showed that the prior acquisition time of the 4D MRI prediction method can be drastically reduced by TL. The present chapter concerns the regain of prediction quality that was in part compromised for shorter acquisition times in the last chapter. The existence of multiple MR contrasts in the network input is identified as a factor that makes it challenging for the model to accurately predict the inner structures of the liver when only a small number of training samples are available. To overcome this problem, we propose to re-utilize 2D training samples as a secondary input for construction of transitive information bridges between the navigator slice primary input and the data slice prediction. We thus equalize the MR contrasts at the input and remove the need for a separate 3D breath-hold MRI with different MR contrast as the secondary input. Results show that this construction leads to improved prediction quality, with a significant decrease in median RMSE from 0.3 to 0.27 ($p < 0.001$, $d = -0.19$). Additionally, removing 3D imaging reduces prior acquisition time from $3\,\mathrm{min}$ to $2\,\mathrm{min}$.

About this chapter Parts of this chapter have been published in: Gino Gulamhussene, Oleksii Bashkanov, Jazan Omari, Maciej Pech, Christian Hansen, and Marko Rak (2023). "Using Training Samples as Transitive Information Bridges in Predicted 4D MRI". In Workshop on Medical Image Learning with Limited and Noisy Data (pp. 237-245). Cham: Springer Nature Switzerland. (Gulamhussene et al., 2023a)

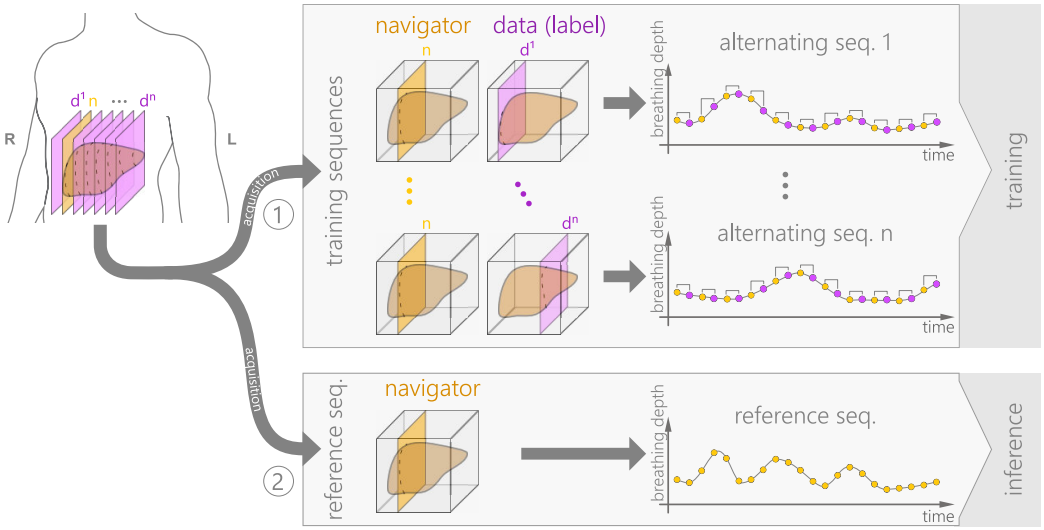

Fig. 8.1: For the work presented in this chapter, only the training sequences and the reference sequence were used.

8.1 Introduction

In the last chapter the issue of long prior acquisition times of the 4D MRI prediction framework was addressed. The proposed solution, however, could not preserve the image quality of the prediction fully. In this chapter, we present an updated form of our 4D MRI prediction framework that involves reusing 2D training slices for transitive information bridging, which leads to a reduction in acquisition time and an improvement in image quality compared to the method presented in the last chapter.

8.2 Materials and Methods

8.2.1 Data

In chapter 4 we described three parts that make up the data set. In this chapter, the method that is discussed, uses only two parts of the data. The two parts are depicted in the gray boxes in Fig. 8.1. Remember, the data sequences alternate between navigator and data slices, forming navigator-data pairs that are used as training samples in the 4D MRI prediction framework. While the navigator has a fixed position and serves as a respiratory motion surrogate during training, the data

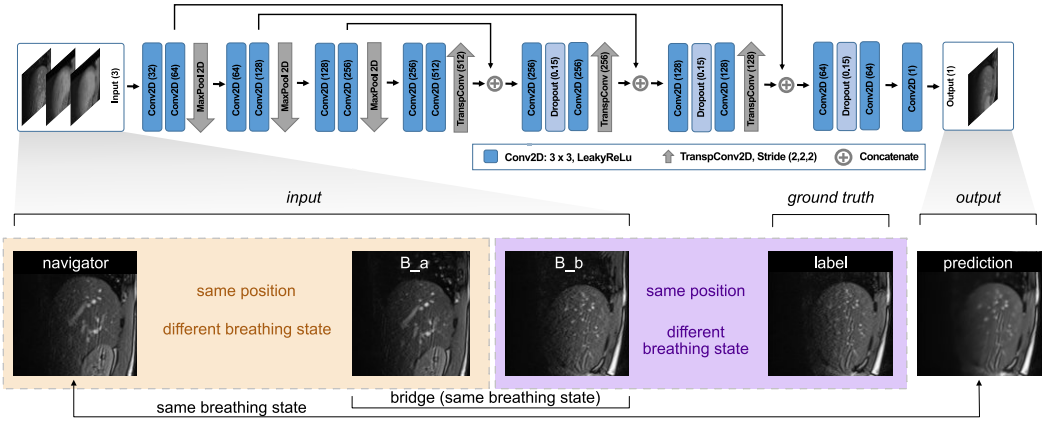

Fig. 8.2: The Network architecture is that of an U-Net. Also shown is the implicit transitive information bridging between input and output.

slice position changes for every sequence to sample the liver equidistantly in 4 mm steps. For each slice position, 175 navigator-data pairs are available. The total number of slice positions ranges from 38 to 57, depending on the liver size. The acquisition of a one slice took 166 ms.

Reference sequences contain navigator slices only and show the succession of breathing states. They are used as breathing surrogate during inference. Such a sequence comprises 513 MRIs images. It represents a total time duration of 85 s and shows around 20 breathing cycles.

8.2.2 Improved Transitive Information Bridging

The updated form of our deep learning-based 4D MRI framework is depicted in Fig. 8.2. Before we discuss the changes to the formulation, lets first remember the original formulation, presented in chapter 6. There, we proposed to train a Net with a three channel input, using the data sequences together with slices of the 3D breath-hold volume. The trained net is able to perform a 4D MRI prediction based on a sequence of navigator frames. The first channel is fed by a navigator. During training the navigator comes from a training pair. During inference it comes from the reference sequence. The second channel is fed by a slice of the 3D volume at navigator position, the third channel is fed by another slice of the 3D volume at label position, which is the position we seek to predict. In this formulation, the second and third channel are used to build, what we call, a transitive information bridge, which we described in more detail in 6.2.3. The term transitive information bridge is quite long and will be often used in the chapter, so we will use the term bridge as a shorthand. Likewise we will call transitive information bridging just bridging.

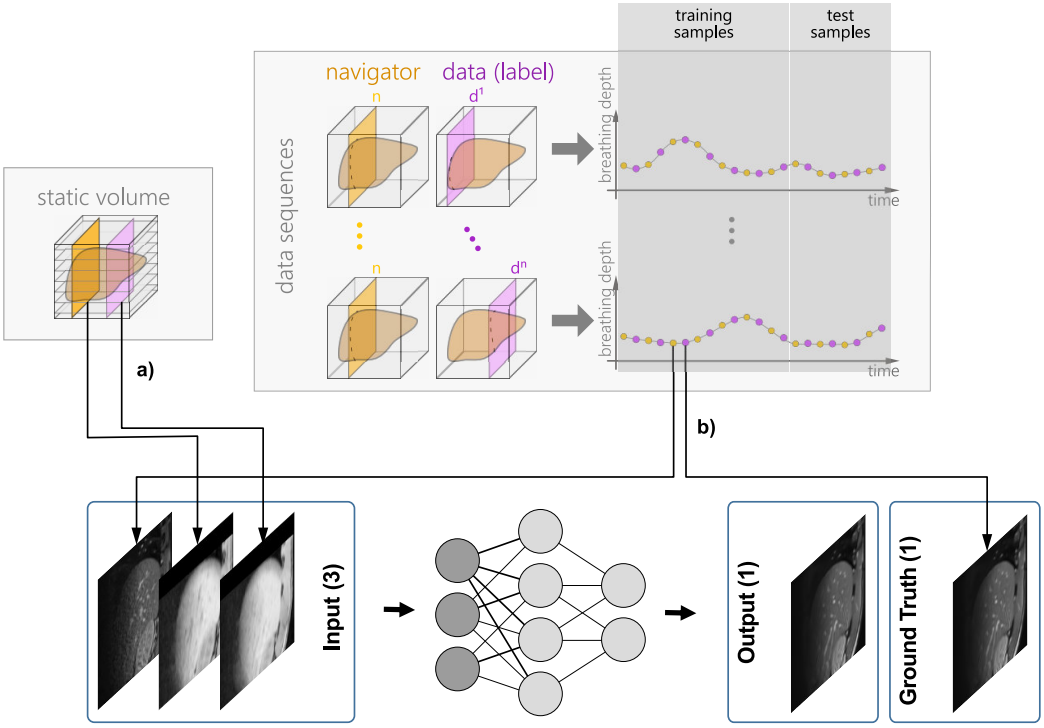

Fig. 8.3: Volume bridging: a) transitive information bridges are resampled from the static STARE-VIBE volume in sagittal orientation and at the same position as the navigator and ground truth, respectively. b) the main input and ground truth are taken from the training split of the training sequences. Note the different MR contrast of the bridges compared to main input, output and ground truth.

8.2.3 Volume Bridges

Now, central to this chapter, is the bridging part, and the fact that the breath-hold volume, used for it until this point, has a different MR contrast than the data sequences. The breath hold volume are acquired using the STAR VIBE MRI sequence and the reference and training sequences are acquired using the TRUFI MRI sequence. While in a STAR VIBE volume the liver appears bright with dark vessels, in the TRUFI the liver tissue is dark and the vessels are bright. A direct comparison is given in Fig. 8.4. Both MRI sequences were also described in 4.3. Let us call the type of bridging that is based on the breath-hold volume *volume bridging* (VB). In this formulation, the learning task does not only involved learning ΔB, which was the difference in breathing states (see 6.2.3) but also learning the apparent difference in MR contrast and different appearance of vessels between the breath-hold volume and the label.

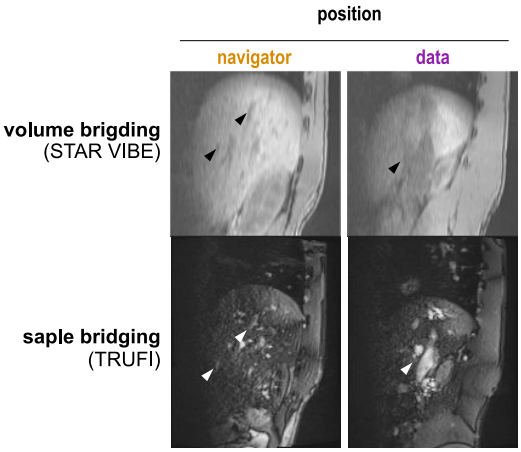

Fig. 8.4: Comparison of bridges sampled from a STAR-VIBE volume vs. bridges taken from a TRUFI training sequence. Note, that the sample bridges have the same contrast and vessel appearance as the main input and output of the network, whereas the volume bridges have a different contrast and vessel appearance.

8.2.4 Sample Bridges

In this chapter, we propose to reuse the training samples from the data sequences as bridges instead of requiring a separate breath-hold volume with a different MR contrast. Lets call this type of bridging sample bridging (SB). In the case of SB, channel two and three are fed by a second navigator-data slice pair. That means within one batch entry two navigator-data slice pairs are used. One as navigator and label in the first channel and the output and another one for the bridging in the second and third channel. The second pair, however, is taken from the same data sequence as the first one. Fig. 8.2 shows how the two navigator-data slice pair are used during training. The two navigators (orange box) share the same slice position but have different breathing states. The two data slices (purple box) also have the same slice position and different breathing state. In this way it resembles the original formulation. However, there are notable differences between VB and SB. First, in SB, channels two and three have a know time offset of $166\,\mathrm{ms}$ due to the acquisition time. We hypothesize that the effect is negligible or compensated by the network. Second, in VB, the bridges not only have a different MR contrast but also lack image detail compared to the label. Whereas in SB, all inputs and outputs share the same MR contrast and level of image detail. We hypothesize that this simplifies the learning task. Third, SB avoids the need for acquiring a 3D breath-hold volume, which reduces the beforehand acquisition time significantly by $1\,\mathrm{min}$. Within the framework, we use the same network architecture and training setup as described in chapter 6, for comparability.

We further differentiate SB into two types or flavors. First, fixed sample bridging (FSB), and second, random sample bridging (RSB). In FSB for each slice position one navigator-data pair is chosen as a bridge for the training. The same bridges are then also used during inference. In RSB the bridges for each slice position are randomly chosen for each forward pass during training. In the following two sections both FSB and RSB will be described.

8.2.5 Fixed Sample Bridges

Remember each data sequence corresponds to one slice position. We need a sample bridge (SB) for each slice position, which is a training sample, selected from the corresponding data sequence. In our tests we used the last sample of the training part of the data sequence, as shown in Fig. 8.5. This sample, however, was not used as training sample but is regarded as part of another split, seperate from the training and test split. Once, the bridges are selected before the training starts, they are kept the same for training and inference. Hence, the name fixed sample bridges. The SBs for all slice positions are cached to reduce the latency caused by the reading access to the hard drive.

8.2.6 Random Sample Bridges

In the case of RSB we want to randomize the SBs. So instead of selecting one separate training sample as SB for one slice position and keep it the same, we use the available training samples as a pool to randomly select a SB during training. That means, for each training iteration, the bridges were randomly chosen for all slice positions. When the batch size was greater than one, then SBs were independently selected for each batch entry. It was ensured, that the training sample used as SB was not the same as the training sample used for channel one and the label. If per chance, the same sample was selected, then a new random selection was performed. In the experiments we varied the available amount of training samples. So for RSB the pool of potential SBs also varies.

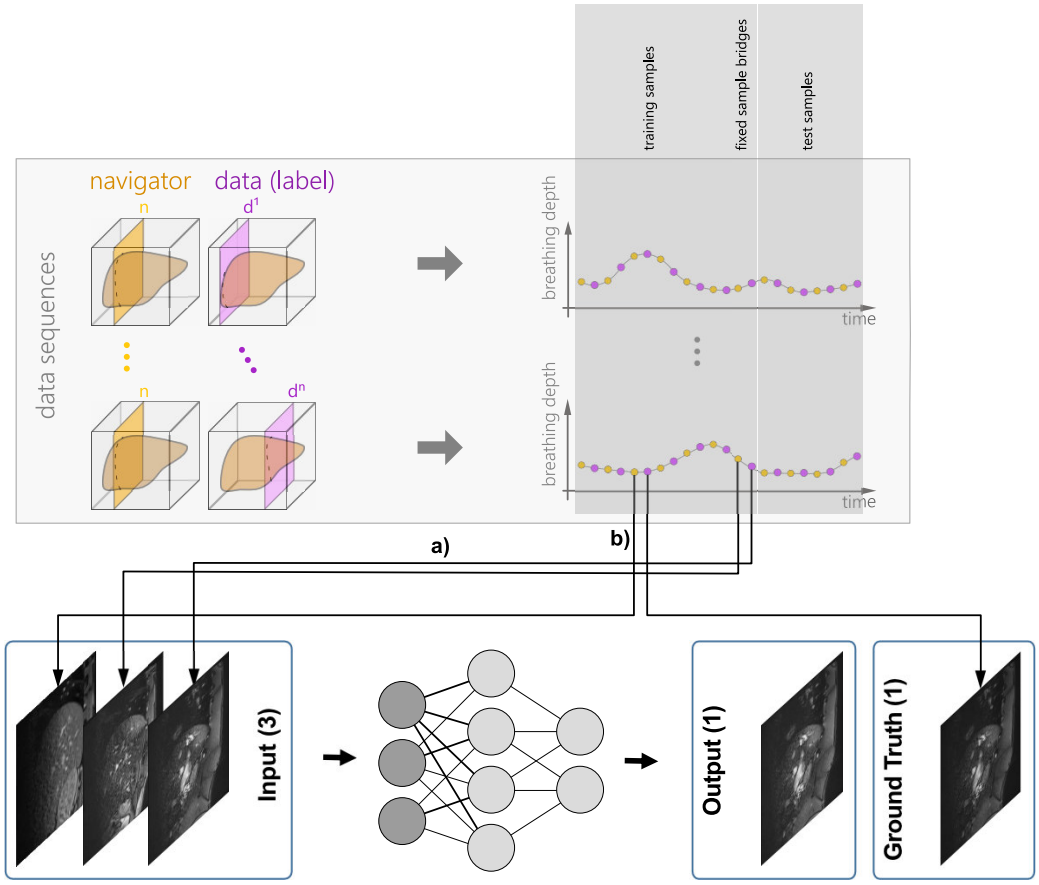

Fig. 8.5: Fixed sample bridging: a) transitive information bridges are taken from a first split, reserved for bridges, of the training sequences. Bridges are chosen to have the same positions as the navigator and ground truth, respectively. b) the main input and ground truth are taken from the training split of the training sequences. Note the MR contrast of the bridges, main input, output and ground truth is the same.

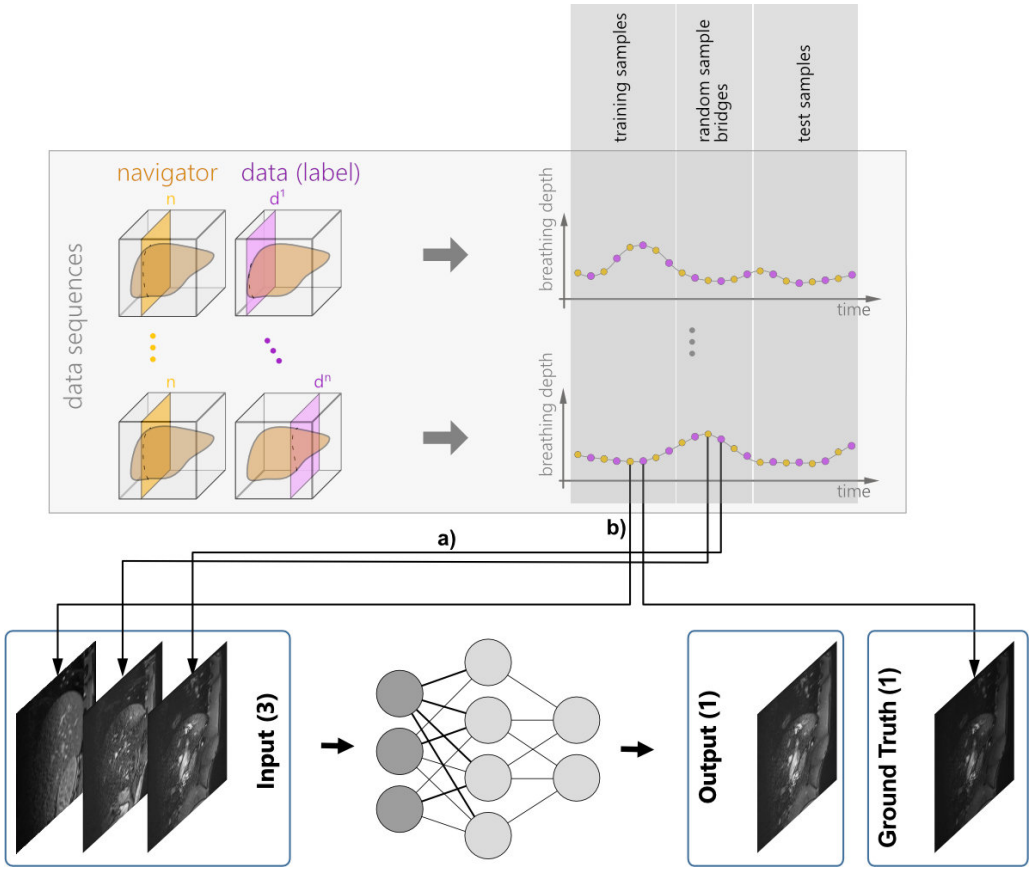

Fig. 8.6: Random sample bridging: a) transitive information bridges are randomly chosen from the training samples. Bridges are chosen to have the same positions as the navigator and ground truth, respectively. b) the main input and ground truth are taken from the training split of the training sequences.

8.3 Experimental Design

8.3.1 Research Questions and Hypothesis

This chapter addresses seven RQs. A summary is listed in Tab. 8.2. The RQs regard three aspects under which the framework is evaluated. The first aspect is the use of data from the source domain **S** and the target domain **T**. The split into **S** and **T** will be described in 8.3.2. The second aspect regards the type of bridging that is used, i.e., VB, FSB, RSB, and RFSB. The third aspect regards the amount of training data that is available, i.e., used, in **S** and **T**, respectively.

Combining the first two aspects results in a matrix of possible configurations. The matrix overview of the configurations is depicted in table 8.1. We denote these configurations as XX+YYY (for example TL+FSB), where XX+ corresponds to the first aspect and can be either one of TL+, DL+ (direct learning), or empty. It indicates whether source domain data or target domain data or both kinds of data were used. Specifically, TL+ indicates transfer learning was utilized, and means a base model was trained in **S** and fine-tuned in **T**. So, respective models have seen samples from both **S** and **T**. DL+ indicates direct learning, and means a model was directly trained in the target domain **T** without starting from a base model that is fine-tuned. So, respective models only have seen training samples from **T**. Note, only in the context of the configurations, do we use the term DL to denote direct learning. Configurations missing the XX+ part indicate that a model was trained only in **S** and that its out-of-domain performance was evaluated in **T**. So, respective models have only seen samples from **S**. In this context, out-of-domain performance always refers to the performance of a model that was trained in the source domain **S**, when it is tested in the target domain **T** without fine-tuning.

The YYY part of the configuration denotes the bridge type. It can be either one of the three earlier described bridging types: VB, RSB, or FSB with one additional special case of RFSB. The special case means RSB was used in the base model and FSB was used during fine-tuning and testing. That means it only appears in conjunction with TL. Following this notation we can note three groups of configurations that will be evaluated. First, VB, RSB, and FSB, second, DL+VB, DL+RSB, and DL+FSB, and third, TL+VB, TL+RSB, TL+FSB, and TL+RFSB.

The third aspect is the amount of available data in **S** and **T**. To address this aspect, we define discrete levels. The data availability level in **S** we call L_S and the data availability level in **T** we call L_T. For L_S, seven levels are defined at the following steps: $24\,\mathrm{min}$, $12\,\mathrm{min}$, $5\,\mathrm{min}$, $2\,\mathrm{min}$, $1\,\mathrm{min}$, $0.5\,\mathrm{min}$, and $0.25\,\mathrm{min}$. Note, models that

Usage of **S** and **T**	Bridge Type			
	VB	RSB	FSB	RFSB
only **S**	VB	RSB	FSB	
only **T** (DL)	DL+VB	DL+RSB	DL+FSB	
S and **T** (TL)	TL+VB	TL+RSB	TL+FSB	TL+RFSB

Tab. 8.1: Evaluated configurations.

are trained in the source domain **S** are always trained with samples of all 16 source subjects. So, for example, a L_S of $2\,\mathrm{min}$ means, that from each source subject, there were $2\,\mathrm{min}$ worth of training samples used. So, at that level, a total of $16\times2\,\mathrm{min}$ of training samples was used to train the respective model.

For L_T, six levels were defined at the following steps: $47\,\mathrm{min}$, $24\,\mathrm{min}$, $12\,\mathrm{min}$, $5\,\mathrm{min}$, $2\,\mathrm{min}$, and $1\,\mathrm{min}$. Models that are trained directly or are fine-tuned in **T** are always trained exclusively on the training samples of a single target subject. So unlike in **S**, L_T directly represents the training sample amount used for training or fine-tuning a respective model.

Using the ten configurations and levels of data availability we formulate seven RQs and the corresponding null hypothesis (H_0) and alternative hypothesis (H_A).

RQ 1: Does SB, i.e., FSB, and RSB, improve the out-of-domain performance over VB?
H_0 1: SB *does not* improve the out-of-domain performance over VB.
H_A 1: SB *improves* the out-of-domain performance over VB.

RQ 2: If SB, does improve the out-of-domain performance, is there an interaction effect between the bridge type and L_S?
H_0 2: There *is no* interaction effect between bridge type and L_S.
H_A 2: There *is an* interaction effect between bridge type and L_S.

RQ 3: In combination with TL, i.e., in fine-tuned models, does SB, i.e., TL+FSB, TL+RSB, and TL+RFSB, improve model performance over TL+VB?
H_0 3: For fine-tuned models, SB *does not* improve model performance over TL+VB.
H_A 3: For fine-tuned models, SB *does* improve model performance over TL+VB.

RQ 4: For fine-tuned models, is there an interaction effect between the bridge type, and L_S, used for the training of the base model?
H_0 4: For fine-tuned models, there *is no* interaction effect between the bridge type and L_S, used for base model training.

RQ 1:	Does SB improve out-of-domain performance over VB?
RQ 2:	If so, does this effect depend on L_S?
RQ 3:	Does SB improve performance over VB in fine-tuned models?
RQ 4:	If so, does the effect depend on L_S?
RQ 5:	And does it depend on L_T for fine-tuning?
RQ 6:	Does FSB and RSB improve models directly learned in **T**?
RQ 7:	If so, is the effect dependent on L_T?

Tab. 8.2: Summary of the research questions in this chapter

H$_A$ 4: For fine-tuned models, there *is an* interaction effect between the bridge type and L_S, used for base model training.

RQ 5: For fine-tuned models, is there an interaction effect between the bridge type and L_T, used for fine-tuning?
H$_0$ 5: For fine-tuned models, there *is no* interaction effect between the bridge type and L_T, used for fine-tuning.
H$_A$ 5: For fine-tuned models, there *is an* interaction effect between the bridge type and L_T, used for fine-tuning.

RQ 6: For directly learned models in **T**, does SB, i.e., DL+FSB and DL+RSB improve the model performance over DL+VB?
H$_0$ 6: For directly learned models in **T**, SB *does not* improve the model performance over DL+VB.
H$_A$ 6: For directly learned models in **T**, SB *does* improve the model performance over DL+VB.

RQ 7: For directly learned models in **T**, is there an interaction effect between the bridge type and L_T?
H$_0$ 7: For directly learned models in **T**, there *is no* interaction effect between the bridge type and L_T.
H$_A$ 7: For directly learned models in **T**, there *is an* interaction effect between the bridge type and L_T.

8.3.2 Training, Validation, Test Split

For the experiments the data is randomly split into two domains, the source domain **S** that contains 16 subjects and the target domain **T** that contains 4 subjects. The experimental design requires a second split. The data of each subject from both **S** and **T** are split into training, validation and test sets. For that, each data sequence

is split into: training (50%), validation (2%), and test (48%) sets. The second split is performed in sequence, i.e., in order. That means the training set contains the first 50% of the data sequence, the validation set the following 2% and the test set the last 48%.

8.3.3 Experiments

To address the RQs that we formulated in 8.3.1, four experiments were conducted. They are described in the following. An overview of the experiments and the respective configurations that were compared is given in 8.3. All quantitative experiments are conducted for each slice position separately. That means, the performance of the models or configurations is computed as RMSE on slice predictions not on volume predictions, because the ground truth is not available for whole volumes, but only for slice positions, as set out in chapter 4 section 4.2.3. However, the statistical tests were perform over all slice positions combined. Furthermore, a visual comparison of all combinations based on 3D MRI predictions was performed.

Exp 1 (addressing RQ 1 and 2): For every subject in **S**, three sets of seven models each were trained. Each of the three sets corresponds to one of the bridge types, utilizing VB, RSB, or FSB respectively. And each of the seven models per set was trained at a specific level of training data amount. As described earlier, the seven levels were defined at the following distinct steps: $24\,\mathrm{min}$, $12\,\mathrm{min}$, $5\,\mathrm{min}$, $2\,\mathrm{min}$, $1\,\mathrm{min}$, $0.5\,\mathrm{min}$, and $0.25\,\mathrm{min}$. The out-of-domain performance of all models was evaluated separately for all four target subjects in **T** on the respective test sets. For that the RMSE was computed between slice predictions and ground truth labels, which came from the test split of the data secquences. That resulted in a sample size for statistical analysis of 21280 samples for each of the three tested configurations, VB, RSB, and FSB, or 3040 sample per group, which is a combination of the three bridge types and the seven data availability levels. We call the data availability levels in **S** L_S.

We want to perform a two-way ANOVA, to test not only for the main effect of the bridge type (**RQ 1**) but also for an possible interaction effect between bridge type and training data availability level (**RQ 2**). We know the data is not normally distributed, because the RMSE is zero bounded. Also, the Lilliefors test confirms none-normality. However, the ANOVA is known to be robust against none-normally distributed variables when the sample size is large. In this case the sample size is very large, with 3040 samples per group. Also, the ratio of the largest variance and smallest variance between the groups was smaller then four to one, which is a widely used rule

of thumb for accepting the variances as being equal, which is another assumption of the test. After verification of the assumptions, a two-factorial ANOVA was performed to test for main effects and possibly interaction effects of the bridge type and data availability level on the RMSE. A post-hoc Tukey's HSD test was performed to determine which pairs of group means are significantly different from each other. Finally, the Cohen's d was computed to quantify the effect sizes.

Furthermore and to support the ANOVA results in the presence of none-normal distributed data, a second line of statistical tests using none-parametric tests was performed. The one-way Kruskal-Wallis-Test was performed to test for effects of the bridge types on the RMSE. This was done for each data availability level. After a significant effect was observed a post-hoc pairwise Wilcoxon Rank Sum Tests was performed to determine which pairs of group means are significantly different. To correct for multiple testing the Benjamini-Hochberg method was used. The significance level was set to 0.05 for all statistical tests.

Exp 2 (addressing RQ 3 and 4): All pre-trained models of Exp 1 were fine-tuned for each subject in **T** separately. For that $2\,\mathrm{min}$ of training data were used. This resulted in three model sets corresponding to the configurations TL+VB, TL+RSB, and TL+FSB in which the same bridge types are used as in the base models. A fourth set corresponds to the configuration TL+RFSB, where the base models were trained using RSB and fine-tuned using FSB. All models were evaluated separately on the test data of the same subject used for fine-tuning. That resulted in a sample size for statistical analysis of $21\,280$ samples for each of the four tested configurations, TL+VB, TL+RSB, TL+FSB, and TL+RFSB, or 3040 sample per group, which is a combination of the four bridge types and the seven levels of L_S.

The statistical analysis was performed in the same way as described in Exp 1. First an ANOVA was performed, testing for the main effect of the bridge type (**RQ 3**) and a possible interaction effect between the bridge type and data availability level on the RMSE (**RQ 4**) for models that were fine-tuned. A post-hoc Tukey's HSD test was performed, identifying group pairs with significantly different mean RMSE. The effect sizes were computed with Cohen's d. Also a none-parametric second line of testing was performed. Using the Kruskal-Wallis-Test and a post-hoc pairwise Wilcoxon Rank Sum Tests, with correction for multiple testing.

Exp 3 (addressing RQ 5): Those pre-trained models from Exp 1 that were trained at the training data amount level of $2\,\mathrm{min}$, were used as base models. Each base model was fine-tuned separately for each subject in **T**. Specifically, each base model was fine-tuned once at each of the six training data amount levels that are defined in the following. The levels were defined at decreasing training data amounts at

Exp	configurations				data avalability levels
Exp 1	VB	RSB	FSB		7 levels in **S**
Exp 2	TL+VB	TL+RSB	TL+FSB		
Exp 3	TL+VB	TL+RSB	TL+FSB	TL+RFSB	6 levels in **T**
Exp 4	DL+VB	DL+RSB	DL+FSB		

Tab. 8.3: Overview of the experiments

the following steps: $47\,\mathrm{min}$, $24\,\mathrm{min}$, $12\,\mathrm{min}$, $5\,\mathrm{min}$, $2\,\mathrm{min}$, and $1\,\mathrm{min}$. This resulted in four model sets of six models each, where the sets are corresponding to the configurations TL+VB, TL+RSB, TL+FSB, and TL+RFSB. The performance of the models was evaluated on the test data of the subject used for fine-tuning. That resulted in a sample size for statistical analysis of $18\,240$ samples for each of the four tested configurations, TL+VB, TL+RSB, TL+FSB, and TL+RFSB, or 3040 sample per group, which is a combination of the four bridge types and the six data availability levels. We call the data availability levels in **T** L_T.The statistical analysis was performed in the same way as in Exp 1 and 2.

Exp 4 (addressing RQ 6 and 7): For all four subjects in **T**, three sets of six models each were trained. The models were trained directly and separately on the target subjects. The three sets per subject correspond to the three configurations DL+VB, DL+RSB, and DL+FSB. Each set contains six models that correspond to the six levels of available trained data amount defined in Exp 3. The performance of all models was evaluated on the test data part of the subject used for training. That resulted in a sample size for statistical analysis of $18\,240$ samples for each of the three tested configurations, DL+VB, DL+RSB, and DL+FSB, which is a combination of the three bridge types and the six of L_T. The statistical analysis was performed in the same way as in Exp 1, 2 and 3.

8.4 Results

8.4.1 Results of Exp 1

The distribution of RMSEs is presented as box plot in Fig. 8.7. The median and 95th percentile are presented in Tab. 8.4. The main results of the ANOVA are presented in Tab. 8.5. The ANOVA shows significant main effects for both L_S and the bridge type and a significant interaction effect between the two factors. The first main effect, regarding L_S was already shown in chapter 7 and is confirmed in this experiments.

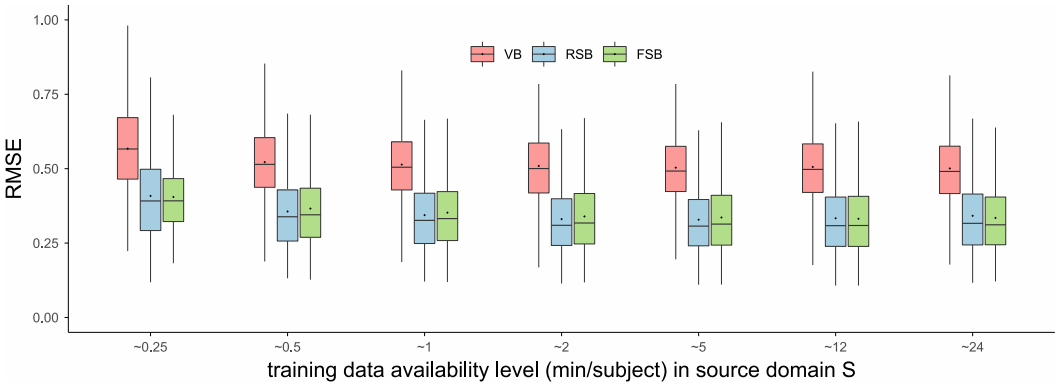

Fig. 8.7: **Exp 1** RMSE box plots grouped by configuration and L_S.

	L_S (min/subject)						
	0.25	0.5	1	2	5	12	24
VB	0.57 (0.83)	0.51 (0.78)	0.5 (0.77)	0.5 (0.77)	0.49 (0.76)	0.5 (0.76)	0.49 (0.75)
RSB	0.39 (0.68)	**0.34 (0.58)**	**0.33 (0.55)**	**0.31 (0.55)**	**0.31 (0.55)**	0.31 (0.57)	0.32 (0.58)
FSB	**0.39 (0.61)**	0.35 (0.59)	0.33 (0.57)	0.32 (0.55)	**0.31 (0.55)**	**0.31 (0.55)**	**0.31 (0.55)**

Tab. 8.4: **Exp 1** RMSE Median and 95th percentile in brackets.

We can interpret the second main effect, regarding the bridge type, with the result of the post-hoc Tukey's HSD test (Tab. 8.6). Here we see that SB, i.e., both RSB and FSB, reduce the RMSE in the target domain compared to VB. For example, at the data availability level of $2\,\mathrm{min}$ RSB reduces the median RMSE from 0.50 to 0.31 (see Tab. 8.4). The Cohen's d reveals a large negative effect of SB, reducing the RMSE in all tested data availability levels. This effect is significant with $p < 0.05$ (see Tab. 8.8). That means, we must reject the null hypothesis $\mathbf{H}_0$ **1** in favour of the alternative hypothesis $\mathbf{H}_A$ **1** and can answer the research question **RQ 1** with yes, SB improves the out-of-domain performance over VB. This is solidified by the Kruskall-Wallis test that also confirms that, in general, the bridge type has a

significant effect on the RMSE in all tested data availability level L_S (see Tab. 8.7). In Exp 1, the difference between RSB and FSB is not significant (Tab. 8.6).

On the other hand, the found interaction effect describes that, while at all levels of L_S, SB performs better than VB, it achieves greater improvements at higher L_S. This can be seen in Tab. 8.8 that shows that the effect size d increases with increasing L_S. That means, we must reject the null hypothesis $\mathbf{H}_0$ **2** in favour of the alternative hypothesis $\mathbf{H}_A$ **2** and can answer the research question **RQ 2** with yes, the effect of bridge type depends on L_S. The effect size also shows that the difference between RSB and FSB is small and, at some levels, not significant.

Fig. 8.8 depicts a qualitative comparison of the out-of-domain performance of all three related configurations, using the same amount of training data $L_S = 2\,\mathrm{min}$. One can see that RSB and FSB yield more image detail than VB. For example, in the sagittal orientation, we see the vessel tree clearly when SB is used, which is not the case for VB.

Effect type	Factor	Df	F	p
Main effects	L_S	6	342.133	**<0.001**
	bridge type	2	11411.1	**<0.001**
Interaction effects	L_S * bridge type	12	4.20652	**<0.001**

Tab. 8.5: **Exp 1** Main results of the ANOVA on the RMSE.

comparison	Estimate	Std. Error	p
RSB - VB	-0.159	0.003	**<0.001**
FSB - VB	-0.163	0.003	**<0.001**
FSB - RSB	-0.004	0.003	0.522

Tab. 8.6: **Exp 1** Main results of the Tukey's HSD test.

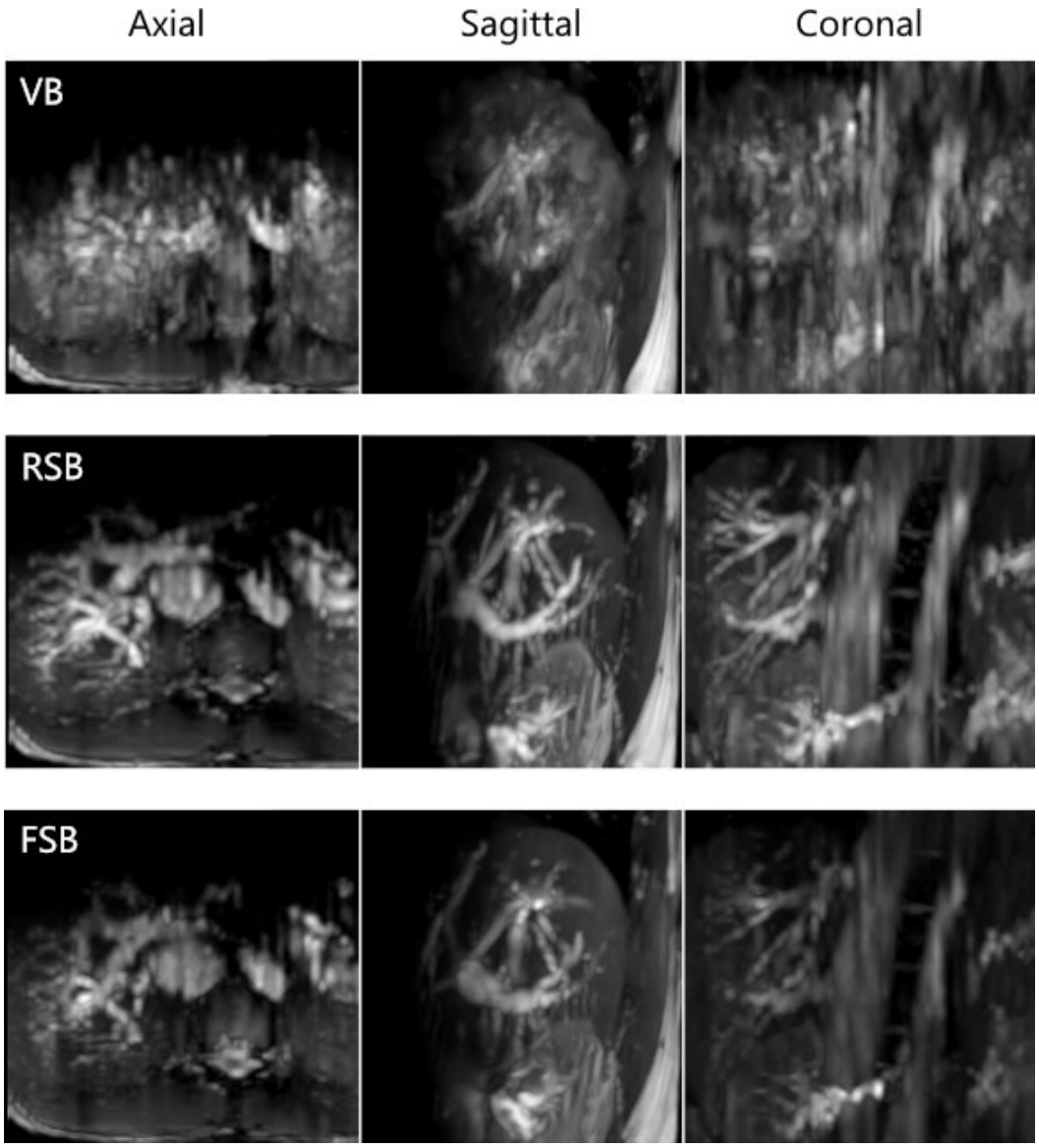

Fig. 8.8: **Exp 1** Out-of-Domain predictions of VB, RSB, and FSB for $L_S = 2\,\mathrm{min}$.

L_S	χ^2	p
0.25	1992.231	**<0.001**
0.5	2420.277	**<0.001**
1	2526.175	**<0.001**
2	2688.222	**<0.001**
5	2699.339	**<0.001**
12	2683.511	**<0.001**
24	2509.224	**<0.001**

Tab. 8.7: **Exp 1** main results of the Kruskall-Wallis test on the RMSE, showing the effect of bridge types within data availability level l in minutes.

L_S	comparison	d	p
0.25	RSB vs FSB	0.028	0.209
	RSB vs VB	-1.028	**<0.001**
	FSB vs VB	**-1.171**	**<0.001**
0.5	RSB vs FSB	-0.080	**<0.001**
	RSB vs VB	**-1.238**	**<0.001**
	FSB vs VB	-1.180	**<0.001**
1	RSB vs FSB	-0.068	**0.009**
	RSB vs VB	**-1.280**	**<0.001**
	FSB vs VB	-1.213	**<0.001**
2	RSB vs FSB	-0.075	**0.002**
	RSB vs VB	**-1.342**	**<0.001**
	FSB vs VB	-1.262	**<0.001**
5	RSB vs FSB	-0.062	**0.013**
	RSB vs VB	**-1.333**	**<0.001**
	FSB vs VB	-1.264	**<0.001**
12	RSB vs FSB	0.015	0.908
	RSB vs VB	-1.274	**<0.001**
	FSB vs VB	**-1.307**	**<0.001**
24	RSB vs FSB	0.058	0.161
	RSB vs VB	-1.184	**<0.001**
	FSB vs VB	**-1.281**	**<0.001**

Tab. 8.8: **Exp 1** Main results of Cohen's d and the pairwise Wilcoxon rank sum test (p) on the RMSE.

8.4.2 Results of Exp 2

The distribution of RMSEs is presented as box plot in Fig. 8.9. The median and 95th percentile are presented in Tab. 8.9. The main results of the ANOVA are presented in Tab. 8.10. The ANOVA shows significant main effects for both L_S and the bridge type and a significant interaction effect between the two factors. The first main effect, regarding L_S was already shown in chapter 7 and is confirmed in this experiments.

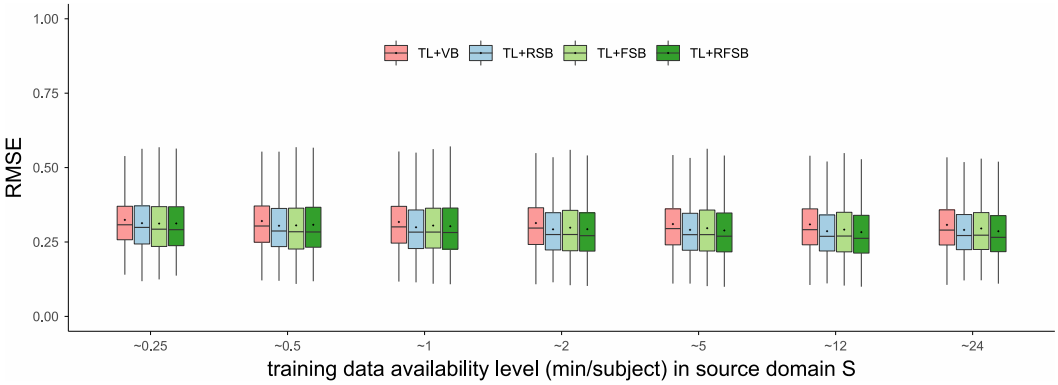

Fig. 8.9: **Exp 2** RMSE box plot.

	L_S (min/subject)						
	0.25	0.5	1	2	5	12	24
TL+VB	0.31 (0.48)	0.3 (0.49)	0.3 (0.48)	0.3 (0.48)	0.29 (0.47)	0.29 (0.46)	0.29 (0.46)
TL+RSB	0.3 (0.48)	0.29 (0.47)	**0.28 (0.47)**	0.28 (0.46)	0.27 (0.46)	0.27 (0.45)	**0.27 (0.45)**
TL+FSB	**0.29 (0.5)**	**0.28 (0.5)**	0.28 (0.49)	0.28 (0.49)	0.27 (0.48)	0.27 (0.48)	0.27 (0.47)
TL+RFSB	0.29 (0.51)	**0.28 (0.5)**	0.28 (0.49)	**0.27 (0.48)**	**0.27 (0.47)**	**0.26 (0.45)**	**0.27 (0.45)**

Tab. 8.9: **Exp 2** RMSE Median and 95th percentile.

We can interpret the second main effect, regarding the bridge type, again, with the result of the post-hoc Tukey's HSD test (Tab. 8.11). Here we see that SB, i.e., all three combinations TL+RSB, TL+FSB, and TL+RFSB, compared to TL+VB, reduce the RMSE in the target domain. For example in Tab. 8.9, we see at $L_S = 2\,\text{min}$, TL+RFSB reduces the median RMSE from 0.3 to 0.27 ($d = -0.213$ see Tab. 8.13). The Cohen's d (see Tab. 8.13) reveals a small negative effect of SB compared to VB, reducing the RMSE in all tested data availability levels. This effect is also significant. This means, we must reject the null hypothesis $\mathbf{H}_0$ **3** in favour of the alternative hypothesis $\mathbf{H}_A$ **3** and can answer the research question **RQ 3** with yes, SB compared to VB improves the performance of fine-tuned models. This is supported by the Kruskall-Wallis test that also confirms that, in general, the bridge type has a significant effect on the RMSE in all data availability level L_S (see Tab. 8.12). In Exp 2, the Tukey's HSD test reveals that the differences between TL+RSB, TL+FSB, and TL+RFSB, are not significant (Tab. 8.11). This is supported by the Cohen's d, which

shows very small effect sizes for the differences between the three configurations and the pairwise Wilcoxon rank sum test that shows the differences are not significant at most levels of L_S.

On the other hand, the found interaction effect describes that, while at all levels of L_S, TL+SB performs better than TL+VB, it achieves greater improvements at higher L_S. This can be seen in Tab. 8.13 that shows that the effect size d increases with increasing L_S. This means, we must reject the null hypothesis $\mathbf{H}_0$ **4** in favour of the alternative hypothesis $\mathbf{H}_A$ **4** and can answer the research question **RQ 4** with yes, the effect of the bridge type depends on L_S.

Fig. 8.10 depicts a qualitative comparison of the performance of all four configurations tested in Exp 2. The examples correspond to data availability levels of $L_S = 2\,\mathrm{min}$ and $L_T = 2\,\mathrm{min}$. One can see that TL+RSB, TL+FSB, and TL+RFSB contain slightly more image detail than TL+VB. For example, in the coronal view, at the upper right liver dome, marked by the arrows, we see small vessels that are not visible for TL+VB.

Effect type	Factor	Df	F	p
Main effects	L_S	6	92.288	**<0.001**
	bridge type	3	163.755	**<0.001**
Interaction effects	$L_S * $ bridge type	18	2.255	**0.002**

Tab. 8.10: **Exp 2** Main results of the ANOVA on the RMSE.

comparison	Estimate	Std. Error	p
TL+RSB - TL+VB	-0.011	0.003	**<0.001**
TL+FSB - TL+VB	-0.012	0.003	**<0.001**
TL+RFSB - TL+VB	-0.012	0.003	**<0.001**
TL+FSB - TL+RSB	-0.001	0.003	0.946
TL+RFSB - TL+RSB	-0.001	0.003	0.982
TL+RFSB - TL+FSB	0.000	0.003	0.998

Tab. 8.11: **Exp 2** Main results of the Tukey's HSD test.

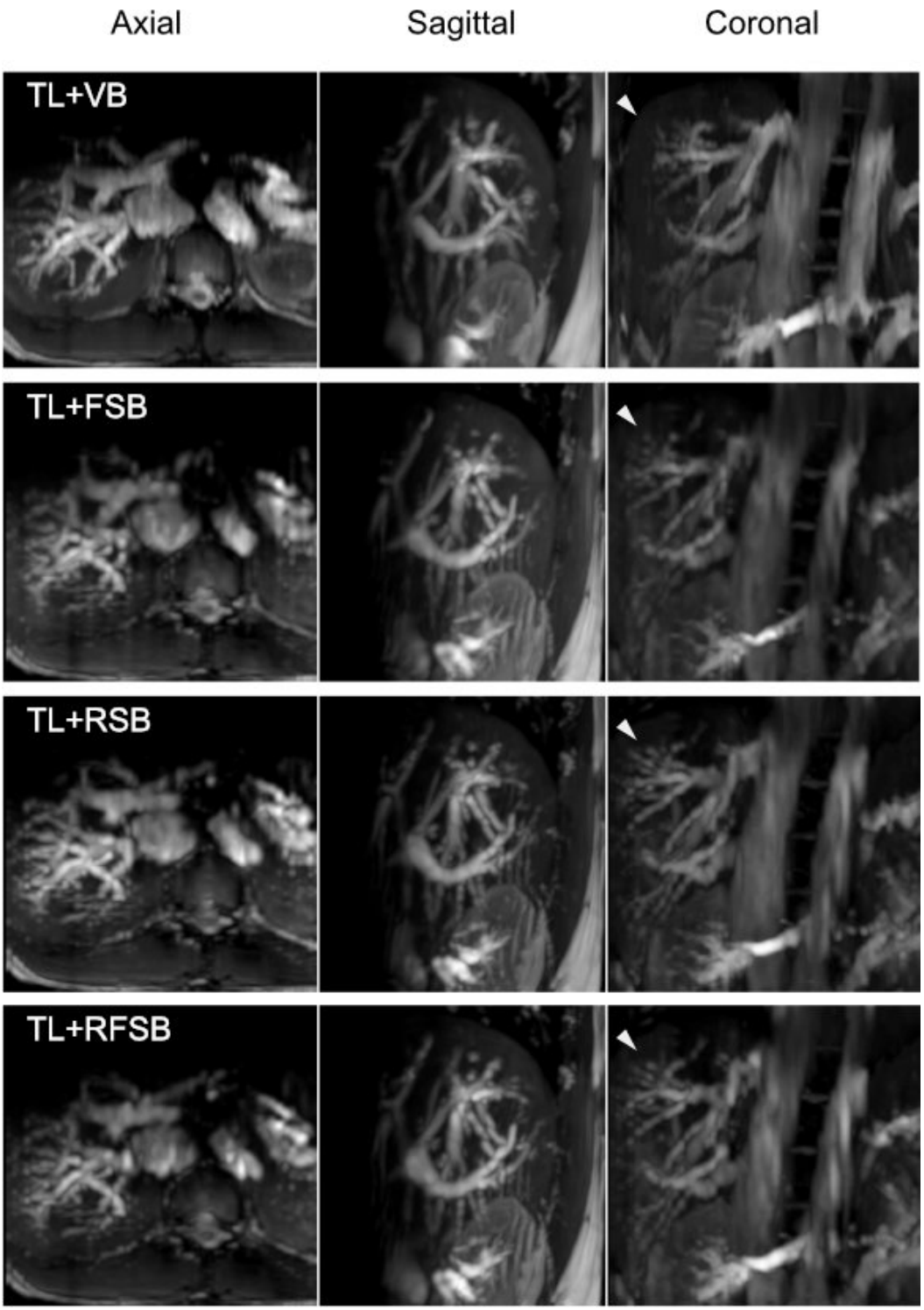

Fig. 8.10: **Exp 2** Predictions of TL+VB, TL+RSB, TL+FSB, and TL+RFSB, for $L_S = 2\,\mathrm{min}$ and $L_T = 2\,\mathrm{min}$.

L_S	χ^2	p
0.25	80.022	**<0.001**
0.5	93.933	**<0.001**
1	103.240	**<0.001**
2	149.895	**<0.001**
5	139.973	**<0.001**
12	211.332	**<0.001**
24	146.695	**<0.001**

Tab. 8.12: **Exp 2** main results of the Kruskall-Wallis test on the RMSE, showing the effect of bridge types within data availability level l in minutes.

L_S	comparison	d	p
0.25	TL+RSB vs. TL+VB	-0.118	**<0.001**
	TL+FSB vs. TL+VB	**-0.129**	**<0.001**
	TL+RFSB vs. TL+VB	-0.125	**<0.001**
	TL+FSB vs. TL+RSB	-0.014	0.079
	TL+RFSB vs. TL+RSB	-0.009	0.107
	TL+RFSB vs. TL+FSB	0.004	0.76
0.5	**TL+RSB vs. TL+VB**	**-0.167**	**<0.001**
	TL+FSB vs. TL+VB	-0.146	**<0.001**
	TL+RFSB vs. TL+VB	-0.127	**<0.001**
	TL+FSB vs. TL+RSB	0.013	0.324
	TL+RFSB vs. TL+RSB	0.032	0.945
	TL+RFSB vs. TL+FSB	0.018	0.324
1	**TL+RSB vs. TL+VB**	**-0.191**	**<0.001**
	TL+FSB vs. TL+VB	-0.125	**<0.001**
	TL+RFSB vs. TL+VB	-0.146	**<0.001**
	TL+FSB vs. TL+RSB	0.058	0.252
	TL+RFSB vs. TL+RSB	0.034	0.984
	TL+RFSB vs. TL+FSB	-0.022	0.252
2	**TL+RSB vs. TL+VB**	**-0.224**	**<0.001**
	TL+FSB vs. TL+VB	-0.157	**<0.001**
	TL+RFSB vs. TL+VB	-0.213	**<0.001**
	TL+FSB vs. TL+RSB	0.053	0.558
	TL+RFSB vs. TL+RSB	0.005	0.421
	TL+RFSB vs. TL+FSB	-0.047	0.227
5	**TL+RSB vs. TL+VB**	**-0.208**	**<0.001**
	TL+FSB vs. TL+VB	-0.142	**<0.001**
	TL+RFSB vs. TL+VB	-0.225	**<0.001**
	TL+FSB vs. TL+RSB	0.056	0.248
	TL+RFSB vs. TL+RSB	-0.021	0.098
	TL+RFSB vs. TL+FSB	-0.075	**0.008**
12	TL+RSB vs. TL+VB	-0.248	**<0.001**
	TL+FSB vs. TL+VB	-0.182	**<0.001**
	TL+RFSB vs. TL+VB	**-0.277**	**<0.001**
	TL+FSB vs. TL+RSB	0.052	0.519
	TL+RFSB vs. TL+RSB	-0.034	**0.014**
	TL+RFSB vs. TL+FSB	-0.084	**0.003**
24	TL+RSB vs. TL+VB	-0.186	**<0.001**
	TL+FSB vs. TL+VB	-0.130	**<0.001**
	TL+RFSB vs. TL+VB	**-0.233**	**<0.001**
	TL+FSB vs. TL+RSB	0.051	0.23
	TL+RFSB vs. TL+RSB	-0.048	**0.011**
	TL+RFSB vs. TL+FSB	-0.097	**<0.001**

Tab. 8.13: **Exp 2** Main results of Cohen's d and the pairwise Wilcoxon rank sum test (p) on the RMSE.

8.4.3 Results of Exp 3

The distribution of RMSEs is presented as box plot in Fig. 8.11. The median and 95th percentile are presented in Tab. 8.14. The main results of the ANOVA are presented in Tab. 8.15. The ANOVA shows significant main effects for both L_T and the bridge type and a significant interaction effect between the two factors. The first main effect, regarding L_S was already shown in chapter 7 and is confirmed in this experiments.

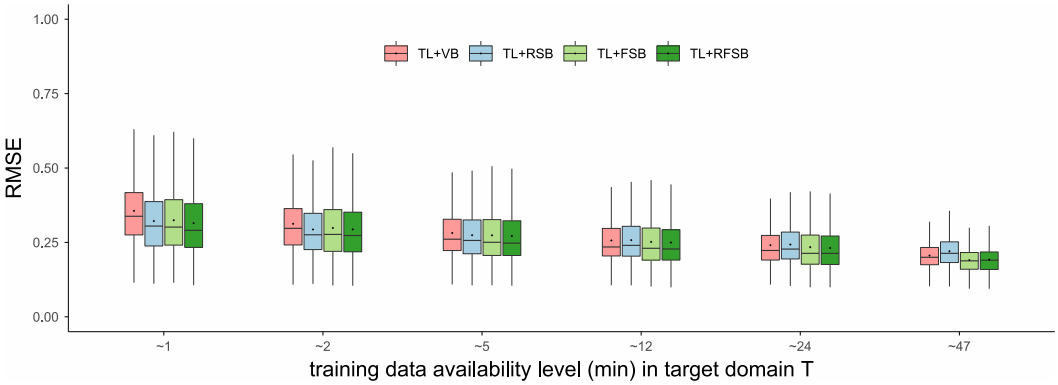

Fig. 8.11: **Exp 3** RMSE box plot.

	L_T (min)					
	1	2	5	12	24	47
TL+VB	0.34 (0.55)	0.3 (0.47)	0.26 (0.43)	0.23 (0.4)	0.22 (0.38)	0.2 (0.28)
TL+RSB	0.3 (0.52)	0.28 (0.46)	0.26 (0.43)	0.24 (0.4)	0.23 (0.37)	0.21 (0.32)
TL+FSB	0.3 (0.52)	0.28 (0.48)	0.25 (0.46)	0.23 (0.42)	0.21 (0.39)	**0.19 (0.27)**
TL+RFSB	**0.29 (0.51)**	**0.27 (0.48)**	**0.25 (0.45)**	**0.23 (0.41)**	**0.21 (0.38)**	**0.19 (0.27)**

Tab. 8.14: **Exp 3** RMSE Median and 95th percentile.

The second main effect, describes that the bridge type has a significant effect on the RMSE and the result of the Tukey's HSD test (Tab. 8.16), indicates that all three combinations TL+RSB, TL+FSB, and TL+RFSB, compared to TL+VB, reduce the RMSE in the target domain and that this effect is significant. The combination TL+RSB constitutes an exception. It inverts its effect for $L_T \geq 12$. The Cohen's d (see Tab. 8.18) reveals a medium negative effect of TL+SB compared to TL+VB, reducing the RMSE in all tested data availability levels, again, with the exception of TL+RSB. The significance of the pairwise differences is confirmed by the Wilcoxon rank sum test. This is supported by the Kruskall-Wallis test that also confirms that, in general, the bridge type has a significant effect on the RMSE in all data availability level L_T (see Tab. 8.17).

The found interaction effect between the bridge type and L_T is of complex nature. We use the Cohen's d and the results of the pairwise Wilcoxon rank sum test (Tab. 8.18) for the interpretation of that interaction effect. First, the negative effect of TL+RSB decreases with increasing L_T and even inverts for $L_T \geq 12$. Second, the negative effect of TL+FSB starts to decrease with increasing L_T, but then increases again for large $L_T \geq 24$, having the greatest effect for $L_T = 47$. Third, the same behavior can be observed for TL+RFSB. And fourth, the difference between TL+RSB on the one hand and TL+FSB and TL+RFSB on the other, shows a very small effect and mostly no significance for $L_T \leq 5$, however the difference increases to a medium effect for $L_T \geq 12$ and becomes significant. This means, we must reject the null hypothesis $\mathbf{H}_0$ **5** in favour of the alternative hypothesis $\mathbf{H}_A$ **5** and can answer the research question **RQ 5** with yes, the effect of the bridge type depends on L_T.

Effect type	Factor	Df	F	p
Main effects	L_T	5	3256.196	**<0.001**
	bridge type	3	129.065	**<0.001**
Interaction effects	$L_T *$ bridge type	15	27.493	**<0.001**

Tab. 8.15: **Exp 3** Main results of the ANOVA on the RMSE.

comparison	Estimate	Std. Error	p
TL+RSB - TL+VB	-0.034	0.002	**<0.001**
TL+FSB - TL+VB	0.031	0.002	**<0.001**
TL+RFSB - TL+VB	-0.041	0.002	**<0.001**
TL+FSB - TL+RSB	0.003	0.002	0.5856
TL+RFSB - TL+RSB	-0.007	0.002	**0.0113**
TL+RFSB - TL+FSB	-0.010	0.002	**<0.001**

Tab. 8.16: **Exp 3** Main results of the Tukey's HSD test.

L_T	χ^2	p
1	343.870	**<0.001**
2	126.981	**<0.001**
5	59.710	**<0.001**
12	61.782	**<0.001**
24	118.162	**<0.001**
47	657.174	**<0.001**

Tab. 8.17: **Exp 3** main results of the Kruskall-Wallis test on the RMSE, showing the effect of bridge types within data availability level l in minutes.

L_T	comparison	d	p
1	TL+RSB vs. TL+VB	-0.317	**<0.001**
	TL+FSB vs. TL+VB	-0.291	**<0.001**
	TL+RFSB vs. TL+VB	**-0.384**	**<0.001**
	TL+FSB vs. TL+RSB	0.025	0.429
	TL+RFSB vs. TL+RSB	-0.062	**0.008**
	TL+RFSB vs. TL+FSB	-0.087	**<0.001**
2	**TL+RSB vs. TL+VB**	**-0.203**	**<0.001**
	TL+FSB vs. TL+VB	-0.141	**<0.001**
	TL+RFSB vs. TL+VB	-0.192	**<0.001**
	TL+FSB vs. TL+RSB	0.051	0.527
	TL+RFSB vs. TL+RSB	0.004	0.343
	TL+RFSB vs. TL+FSB	-0.046	0.155
5	TL+RSB vs. TL+VB	-0.085	**<0.001**
	TL+FSB vs. TL+VB	-0.085	**<0.001**
	TL+RFSB vs. TL+VB	**-0.114**	**<0.001**
	TL+FSB vs. TL+RSB	-0.004	0.063
	TL+RFSB vs. TL+RSB	-0.030	**0.005**
	TL+RFSB vs. TL+FSB	-0.025	0.361
12	TL+RSB vs. TL+VB	0.014	0.434
	TL+FSB vs. TL+VB	-0.059	**<0.001**
	TL+RFSB vs. TL+VB	**-0.083**	**<0.001**
	TL+FSB vs. TL+RSB	-0.069	**<0.001**
	TL+RFSB vs. TL+RSB	-0.094	**<0.001**
	TL+RFSB vs. TL+FSB	-0.022	0.434
24	TL+RSB vs. TL+VB	0.029	**0.049**
	TL+FSB vs. TL+VB	-0.084	**<0.001**
	TL+RFSB vs. TL+VB	**-0.124**	**<0.001**
	TL+FSB vs. TL+RSB	-0.110	**<0.001**
	TL+RFSB vs. TL+RSB	-0.151	**<0.001**
	TL+RFSB vs. TL+FSB	-0.036	0.309
47	TL+RSB vs. TL+VB	0.287	**<0.001**
	TL+FSB vs. TL+VB	**-0.353**	**<0.001**
	TL+RFSB vs. TL+VB	-0.312	**<0.001**
	TL+FSB vs. TL+RSB	-0.595	**<0.001**
	TL+RFSB vs. TL+RSB	-0.556	**<0.001**
	TL+RFSB vs. TL+FSB	0.034	0.393

Tab. 8.18: **Exp 3** Main results of Cohen's d and the pairwise Wilcoxon rank sum test (p) on the RMSE.

8.4.4 Results of Exp 4

The RMSEs distribution is presented as box plot in Fig. 8.12. The median and 95th percentile are presented in Tab. 8.19. The main results of the ANOVA are presented in Tab. 8.20. The ANOVA shows significant main effects for both L_T and the bridge type and a significant interaction effect between the two factors.

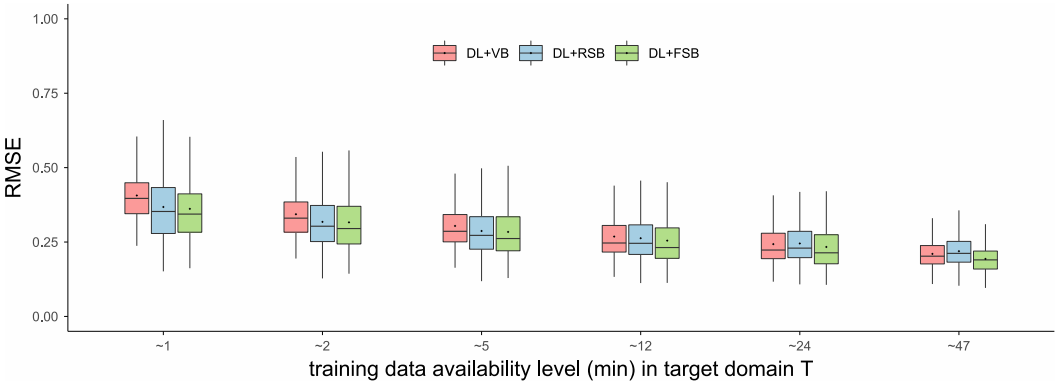

Fig. 8.12: **Exp 4** RMSE box plot.

	L_T (min)					
	1	2	5	12	24	47
DL+VB	0.4 (0.58)	0.33 (0.49)	0.29 (0.45)	0.25 (0.41)	0.22 (0.38)	0.2 (0.29)
DL+RSB	0.35 (0.57)	**0.3 (0.49)**	0.27 (0.44)	0.25 (0.4)	0.23 (0.37)	0.21 (0.31)
DL+FSB	**0.34 (0.55)**	0.3 (0.51)	**0.26 (0.45)**	**0.23 (0.42)**	**0.21 (0.39)**	**0.19 (0.28)**

Tab. 8.19: **Exp 4** RMSE Median and 95th percentile.

The first main effect describes that an increase in L_T leads to a decrease of the RMSE, which is a confirmation of the expected behavior that was already described in chapter 7.

The second main effect, describes that the bridge type has a significant effect on the RMSE. At first, the result of the Tukey's HSD test (Tab. 8.21) suggests that both combinations DL+RSB and DL+FSB, compared to DL+VB, reduce the RMSE in the target domain and that this effect is significant. However, taking L_T into account, the combination DL+RSB, again, constitutes an exception. It inverts its effect for $L_T \geq 24$. The Cohen's d (see Tab. 8.23) reveals a medium negative effect of DL+SB compared to DL+VB, reducing the RMSE in all tested data availability levels, again, with the exception of DL+RSB. The significance of the pairwise differences is confirmed by the Wilcoxon rank sum test, which is supported by the Kruskall-Wallis test that also confirms that, in general, the bridge type has a significant effect on the RMSE for directly learned models in all data availability level L_T (see Tab. 8.22).

This means, we can not reject the null hypothesis $\mathbf{H}_0$ **6** in favour of the alternative hypothesis $\mathbf{H}_A$ **6**. The answer to research question **RQ 6** is inconclusive. DL+SB is different from DL+VB, but it does not always improve the performance of directly learned models.

The reason is the interaction effect between the bridge type and L_T that was also found by the ANOVA. Again, this interaction is complex. Using the Cohen's d and the results of the pairwise Wilcoxon rank sum test (Tab. 8.23) that interaction effect can be interpreted. First, the negative effect of DL+RSB decreases with increasing L_T and even inverts for $L_T \geq 24$. Second, the negative effect of DL+FSB starts to decrease with increasing L_T, but then increases again for $L_T = 47$, having almost the same effect size as for $L_T = 1$. Third, the difference between TL+RSB and TL+FSB shows a very small effect and no or weak significance for $L_T \leq 5$, however the difference increases to a notable effect for $L_T \geq 12$ and becomes significant. This means, we must reject the null hypothesis $\mathbf{H}_0$ **7** in favour of the alternative hypothesis $\mathbf{H}_A$ **5** and can answer the research question **RQ 5** with yes, the effect of the bridge type depends on L_T.

Fig. 8.13 depicts a qualitative comparison of all 3 configurations with direct learning using the same amount of training data $L_T = 2\,\mathrm{min}$. We see that the predictions with DL+RSB, DL+FSB, and DL+FSB contain more image detail than the on with DL+VB. This can be seen in the sagittal view were the vessel tree is almost not reconstructed in the case of DL+VB, but is reconstructed in the case of DL+RSB and DL+FSB.

Effect type	Factor	Df	F	p
Main effects	L_T	5	4895.921	**<0.001**
	bridge type	2	315.889	**<0.001**
Interaction effects	L_T * bridge type	10	40.319	**<0.001**

Tab. 8.20: **Exp 4** Main results of the ANOVA on the RMSE.

comparison	Estimate	Std. Error	p
DL+RSB - DL+VB	-0.039	0.002	**<0.001**
DL+FSB - DL+VB	-0.045	0.002	**<0.001**
DL+FSB - DL+RSB	-0.006	0.002	**0.011**

Tab. 8.21: **Exp 4** Main results of the Tukey's HSD test.

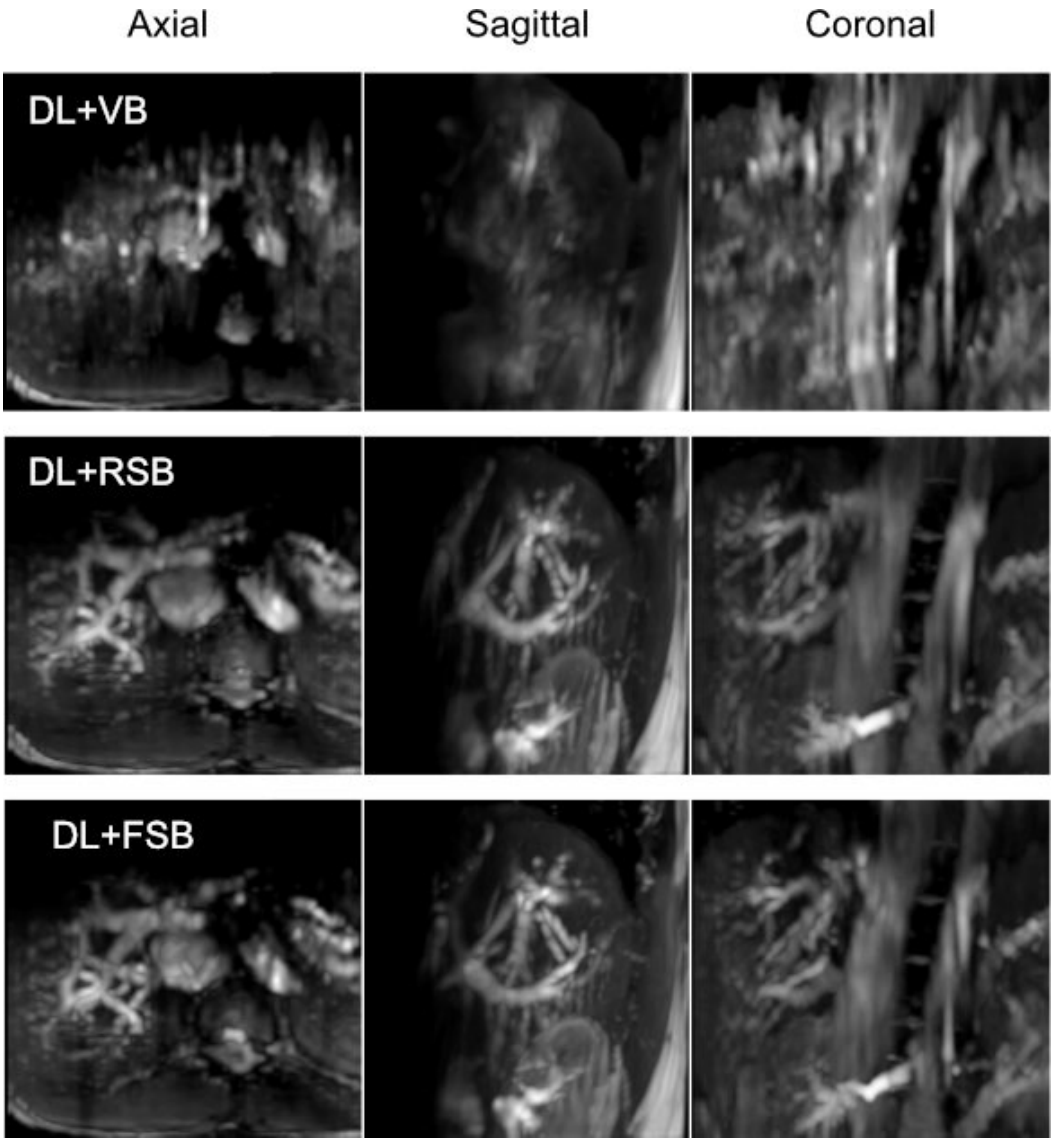

Fig. 8.13: **Exp 4** Predictions of DL+VB, DL+RSB, and DL+FSB for $L_T = 2\,\mathrm{min}$.

L_T	χ^2	p
1	459.419	**<0.001**
2	278.035	**<0.001**
5	187.276	**<0.001**
12	104.279	**<0.001**
24	101.419	**<0.001**
47	403.109	**<0.001**

Tab. 8.22: **Exp 4** main results of the Kruskall-Wallis test on the RMSE, showing the effect of bridge types within data availability level l in minutes.

L_T	comparison	d	p
1	DL+RSB vs. DL+VB	-0.366	**<0.001**
	DL+FSB vs. DL+VB	-0.449	**<0.001**
	DL+FSB vs. DL+RSB	-0.057	0.076
2	DL+RSB vs. DL+VB	-0.290	**<0.001**
	DL+FSB vs. DL+VB	-0.294	**<0.001**
	DL+FSB vs. DL+RSB	-0.016	**0.018**
5	DL+RSB vs. DL+VB	-0.205	**<0.001**
	DL+FSB vs. DL+VB	-0.233	**<0.001**
	DL+FSB vs. DL+RSB	-0.035	**0.001**
12	DL+RSB vs. DL+VB	-0.070	**0.006**
	DL+FSB vs. DL+VB	-0.162	**<0.001**
	DL+FSB vs. DL+RSB	-0.095	**<0.001**
24	DL+RSB vs. DL+VB	0.030	**0.016**
	DL+FSB vs. DL+VB	-0.123	**<0.001**
	DL+FSB vs. DL+RSB	-0.151	**<0.001**
47	DL+RSB vs. DL+VB	0.179	**<0.001**
	DL+FSB vs. DL+VB	-0.359	**<0.001**
	DL+FSB vs. DL+RSB	-0.512	**<0.001**

Tab. 8.23: **Exp 4** Main results of Cohen's d and the pairwise Wilcoxon rank sum test (p) on the RMSE.

8.5 Discussion

8.5.1 Interpretation of Results

Interpretation of Exp 1: SB has a profound effect on the prediction quality and image detail of an out-of-domain prediction and it out-performs VB by far. in its strength, the effect is comparable with the effect of transfer learning. It is large in all levels of L_S, regardless of the interaction effect found by the ANOVA. In experiment 1 RSB and FSB perform equally good, showing almost no significant differences. There seams, however, to be a tendency of RSB to slightly perform better. From a qualitative point of view, SB makes the difference between an almost unrecognisable image (VB) and a clearly visible vessel structure within the liver (RSB and FSB). This means it is strongly advisable to utilize SB to substantially improve the prediction quality, when no data of the target domain is available.

Interpretation of Exp 2: When target domain data is available and fine-tuning is an option, then TL+SB also improves the performance compared to TL+VB. The effect is smaller than the one found in Exp 1, which regards the out-of-domain performance, but it is significant. The Tukey's HSD test did not reveal a significant

difference between the three SB configurations TL+RSB, TL+FSB, and TL+RFSB. Although the box plots as well as the Cohen's d indicate that TL+RSB and TL+RFSB perform slightly better than TL+FSB at $L_T = 2\,\mathrm{min}$, this difference is not significant. It is also evident, that the effect of SB is greater at higher levels of L_S, which is part of the interaction effect. However, the effect size becomes nearly constant for $L_S \geq 2\,\mathrm{min}$. From a qualitative stand point, SB adds image detail, although the effect is far more subtle than it is in the case of out-of-domain predictions. This means, that SB should be utilized in conjunction with fine-tuning to further improve the prediction result. This is also a strong recommendation, because even if the improvements are small, they come at no additional cost. In fact, SB reduce the cost in form of prior acquisition time.

Interpretation of Exp 3: When only $1\,\mathrm{min}$ of target data and a base line model are available, TL+RFSB is the option that yields the best results, significantly outperforming TL+VB, TL+RSB, and TL+FSB. When $2\,\mathrm{min}$ of target data and a base line model are available, TL+RSB yielded the best results. Outperforming TL+VB more then the other two alternatives. However, the difference to the alternatives TL+FSB and TL+RFSB were not significant. Also TL+RSB is sensitive to L_T. It loses its power to improve the prediction in conjunction with fine-tuning when $L_T \geq 12\,\mathrm{min}$ and is most effective for $L_T = 2\,\mathrm{min}$, but does not significantly outperform the other SB options. In this experiment TL+RFSB performed the best over all. So we find that TL+RFSB is be the best option of configurations, when the level L_T varies. It either performs better than the other options or equally good.

Interpretation of Exp 4: When no base model and no source domain data is available, then DL+SB improves prediction performance over DL+VB, however, DL+RSB has to be viewed with care, as its positive effect inverts for $L_T \geq 12\,\mathrm{min}$. We argue that $L_S = 2\,\mathrm{min}$ and $L_T = 2\,\mathrm{min}$ yields a good trade-off between prediction quality and acquisition time. At this level DL+RSB outperforms DL+VB and it has also a slightly better 95th percentile then DL+FSB, however this effect is negligible. Together with the results of Exp 3 we can reason that both DL+RSB and TL+RSB, in general, have problems with large L_T because the randomization of the bridge makes the learning task harder, when the pool of possible bridges is large. This means, that first, DL+FSB is the best option when no source domain data is available, as it significantly outperforms the other two options DL+VB and DL+RSB. And second, TL+RSB is unsuitable for large L_T.

8.5.2 General Discussion

Comparing the results of all experiments TL+RFSB yields the best over all performance. Especially at $L_S = 2\,\mathrm{min}$ and $L_T = 2\,\mathrm{min}$, TL+RFSB provides the best balance of efficiency and performance. If fine-tuning is not possible, FSB still greatly improves prediction performance, almost as much as TL alone. An advantage of all FSB configurations and TL+RFSB is that it also works with extremely few data. That is not the case with the RSB configurations that need a minimum pool size of training samples to chose from as bridge. SB comes at no additional cost, in fact, it reduces the prior acquisition time and at the same time improves the prediction quality. So it can be recommended to use SB in all cases.

As expected, the data that was statistically analysed is not normally distributed, because the error values are bound by zero. Of course, the Lilliefors test confirmed that. A transformation of the data was not possible. Even after a log transform and mean shift the distribution diverted from the Gaussian bell and did not pass the Lilliefors test. Because the ANOVA is known to be robust against none-normal data and the sample size was large the ANOVA was used for statistical analysis but was supported by a one-way Kruskal-Wallis-Test, to confirm found effects.

A concern of the proposed method is that the out-of-plane resolution of the 3D predictions is limited by the slice thickness and distance of the training samples. However, it is easy to handle this. One must acquire more slice positions with thinner slices and to keep the prior acquisition time unchanged, fewer slices per position are taken. As an illustration, collecting 2 samples for every $1\,\mathrm{mm}$ slice instead of 8 samples for each $4\,\mathrm{mm}$ slice, improves the out-of-plane resolution from $4\,\mathrm{mm}$ to $1\,\mathrm{mm}$ while maintaining image quality and acquisition time.

8.5.3 Future Work

In future work it would be interesting to add new subjects with more slice positions of thinner slices to the public data base to evaluate whether it is possible to increase the out-of-plane resolution of the method without hampering the strength of it. Then it should also be investigated, whether the samples for SB should also be used as navigator-label pairs, when there are only 2 samples acquired per slice position. With the updated database, another possible approach to explore would be to use training samples for SB only at intervals of every $4\,\mathrm{mm}$. During training and inference, the SB for the intermediate slices could be interpolated using the techniques proposed by Karani et al. (2018) or Zhang et al. (2018). In future research, the most important

findings of this chapter should be further evaluated using additional metrics like the SSIM and TRE and cross-validation techniques. Finally, a very promising possibility to investigate is to utilize SB for ensembling. Predicting the same slice position multiple times, using different SB would serve two goals. First, the avarege of the prediction could improve the prediction quality. Second, the standart deviation of the predictions would serve as an uncertainty map. This map could be a valuable secondary input to the intervention support system that relies on the 4D MRI data or the derived 4D motion data. This can increase safety because the navigation system can decide whether the movement information is reliable or not.

8.6 Conclusion

This chapter presented a new approach for utilizing training samples in the input of the 4D MRI model, resulting in a significant improvement in prediction quality and shortened acquisition time by reducing the complexity of the learning task and eliminating the need for a prior 3D scan.

9 Conclusion

Organ motion poses a challenge in image-guided interventions like radiation therapy, biopsies or tumor ablation. The research field of time-resolved volumetric magnetic resonance imaging or has evolved in the pursuit of solving this challenge. However, for most interventional settings current techniques are falling short in providing a integration of sufficient temporal and spatial resolution, large field of view, as well as short acquisition and reconstruction times.

In this work a new data set for the development and testing of 4D MRI methods was established and made publicly available and a new deep learning based framework for the generation of 4D liver MRI was proposed. The framework is an end-to-end trainable solution to the 4D MRI problem, achieving sub-second reconstruction times. Further a transfer learning approach and the reusing of training samples as transitive information bridges was proposed to reduce prior acquisition times for training data and the improvement of prediction quality. Finally an ensemble strategy was proposed to facilitate the generation of a uncertainty map giving insight on the certainty of the prediction.

9.1 Contribution

To conclude this thesis we revisit the research gap that was set out in the introduction. It can be summarized as the four following points:

1. There exists no 4D MRI method that combines the following characteristics:
    a) Large FOV
    b) High resolution
    c) Real-time imaging
    d) Time-resolved
    e) Short prior acquisition time
    f) High image quality

2. There is no uncertainty estimation for 4D MRI
3. No public data set and common benchmark for development and comparison of 4D MRI methods

Furthermore, the following research questions were defined in chapter 1 and addressed in chapters 6 to 8.

1. "Can deep learning methods be used to generate real-time 4D MRI with high spatiotemporal resolution, base on a real-time 2D MRI sequence?"
2. "Can the training data requirement for the training of such a deep learning based 4D MRI method be limited to below $3\,\mathrm{min}$, while achieving high prediction quality?"

In the following the contribution of this thesis will be summarized and a tabular comparison of the methods proposed in this thesis with the related work is given in Tab. 9.1.

Public Data Set

A data set of 20 healthy subjects for the development and testing of 4D MRI methods was created. The data set comprising 291GB of image and meta data, was made publicly available to be used by fellow researchers. This addresses the problem of *no public data set and common benchmark*. The dataset can be used for both classical sorting based 4D MRI methods, using a variety of different image based breathing surrogates as well as for fully deep learning based approaches.

Real-Time 4D MRI Framework

This thesis introduced a novel deep learning-based framework for 4D MRI that addressed the first outlined problem. Specifically, the proposed framework is the first to exhibit all of the five previously outlined characteristics. First, the framework works with large FOV 2D MRI slices, sampled over the full width of the liver and predicts a *Large FOV* comprising the entire liver. Second, the framework works with high resolution 2D MRI slices and predicts *high resolution* dynamic 3D volumes. Third, the framework is based entirely on deep learning. It tackles the entire task of predicting a 3D volume from a single 2D MRI slice in a single network inference. Because dedicated hardware can readily be used for inference, this is very fast and works in sub-seconds, i.e., in *real-time*. Fourth, the framework reconstructs 3D

volumes with respect to a time-resolved, physiological, multi-dimensional breathing signal, i.e., the whole dynamic navigator sequence. Consequently, it can produce *time-resolved* 4D æmri images capturing multiple distinct breathing cycles, rather than averaging phases of a single cycle. Fifth, the framework suffices with little subject specific training data, equivalent to $2\,\mathrm{min}$ of acquisition time, hence facilitating a *short prior acquisition time*. And lastly, the framework is able to carry the high image detail of the 2D slices over to the 3D volumes and by that provides *high image quality*. So *research question 1* can be answered with yes. Because the framework reconstructs 4D MRI with high resolution in real-time, based on 2D slices.

Demonstration of Framework Capabilities

Chapter 7 highlighted the versatility of the proposed framework, illustrating the seamless integration with transfer learning and ensemble strategies. Chapter 8 added to this, by demonstrating an effective method for reusing training samples within the framework, leading to enhanced prediction quality and decreased prior acquisition time. The seamless integration of these strategies into the framework addressed two of the method requirements. First, the use of transfer learning in the framework is integral for reducing the prior acquisition time from $47\,\mathrm{min}$ to $2\,\mathrm{min}$ to achieve *short prior acquisition times*. Second, the use of transfer learning and reusing of training samples are central for the *high image quality* of the prediction.

Also the problem of *uncertainty estimation* was addressed by using an ensembling strategy in the framework, facilitating the estimation of uncertainty of the prediction on the image level. With that, *research question 2* can be answered with yes, because the framework is capable of generating high quality 4D MRI with only $2\,\mathrm{min}$ of training data per subject.

	TR	P/R	Matrix size	Resolution in mm^3	Breath. cycle smpl.		vps		Recon. time in s/vol.	befAcq in min	RMSE median (95%)
					P	R	P	R	-		
Tokuda et al. (2008)	no	R	256x128x24	-	-	5	-	-	-	18	-
Cai et al. (2011)	no	R	256x166	1.5x1.5x5	-	4	-	-	-	-	-
Hu et al. (2012)	no	P	-	-	10	-	-	-	-	-	-
Yanle Hu et al. (2013)	no	R	250x176x32	1.5x1.5x5	-	4	-	-	-	3	-
Tryggestad et al. (2013)	no	R	175x190x9	2x2x5	-	10	-	-	-	13	-
Y. Liu et al. (2014)	no	R	256x166	2.5x2.5x5	-	10	-	-	-	-	-
Paganelli et al. (2015)	no	R	256x224x20	1.28x1.28x5	-	8	-	-	-	1.2	-
Deng et al. (2016)	no	R	-	-	-	10	-	-	-	8	-
Han et al. (2017)	no	R	416x250x125	**1.2x1.2x1.6**	-	8	-	-	75	5	-
Rank et al. (2017)	no	-	256x256x60	1.5x1.5x5	-	20	-	-	22.5	0.7	-
Lindt et al. (2018)	no	R	138x208x30	2x2x5	-	10	-	-	30	5	-
Harris et al. (2018)	no	P	-	1.67x1.67x1.67	10	-	-	-	-	-	-
Meschini et al. (2019)	no	R	256x224x20	1.28x1.28x5	-	8	-	-	262	1.2	-
Kavaluus et al. (2020)	no	R	-	1.33x1.33x3	-	8	-	-	-	15	-
Richter et al. (2020)	no	R	224x224x144	2.24x2.23x2.23	-	8	-	-	11	10	6.51
Navest et al. (2020)	no	R	-	-	-	10	-	-	-	-	-
Yang et al. (2020)	no	R	-	1.67^1.67x5	-	10	-	-	-	-	-
Eldeniz et al. (2021)	no	R	318x318x96	1.13x1.13x3	-	10	-	-	2.7	5	-
Keijnemans et al. (2022)	no	R	52x240x136	6.7x1.9x1.9	-	8	-	-	-	3	-
Li et al. (2022)	-	-	256x256x25	1.37x1.37x4	-	-	-	-	-	-	-
Wachinger et al. (2012)	no	R	-	-	-	-	-	-	-	-	-
M. v. Siebenthal et al. (2007)	**yes**	R	192x192x25	1.8x1.8x4	-	36	-	5	73	60	-
Tanner et al. (2014)	**yes**	R	224x224x53	1.3x1.3x5	-	36	-	4.4	-	10	-
Celicanin et al. (2015)	**yes**	R	120x128	1.87x1.87x6	-	20	-	3.33	-	-	-
Zhang et al. (2018)	**yes**	R	-	1.33x1.33x5	-	36	-	2.4	36.5	30	10.23 (13.74)
Karani et al. (2018)	**yes**	R	-	1.33x1.33x5	-	36	-	2.4	-	20	4.09 (6.81)
Romaguera et al. (2019)	**yes**	R	-	1.7x1.7x3	-	**46**	-	**7.5**	-	28	-
Qiu et al. (2019)	**yes**	R	256x256x53	1.34x1.56x4	-	-	-	1.6	-	-	-
Yuan et al. (2019)	**yes**	R	128x128x56	2.7x2.7x4	-	9.78	-	1.63	20	**0.33**	-
TU (Ch. 5)	**yes**	R	140x176x47	1.82x1.82x4	-	36	-	6	27	60	-
DL Framework (Ch. 6)	**yes**	**P/R**	209x128x128	1.8x1.8x1.8	**10.5**	36	**1.75**	6	**0.57**	6	0.29 (0.45)
TL (Ch. 7)	**yes**	**P/R**	209x128x128	1.8x1.8x1.8	**10.5**	36	**1.75**	6	**0.57**	3	0.31 (0.47)
TL+RFSB (Ch. 8)	**yes**	**P/R**	50x128x128	4x1.8x1.8	**10.5**	36	**1.75**	6	**0.57**	2	**0.27 (0.48)**

Tab. 9.1: Comparison of the methods proposed in this thesis with the related work regarding whether its time-resolved (TR), whether reconstruction is done pro-/retrospectively (P/R), matrix size, voxel resolution, how many phases of a breathing cycle can be resolved (breath. cycle smpl.) based on a 6 s breathing cycle, volumes per second (vps) in pro- and retrospective reconstruction (P/R), prior acquisition time (befAcq), reconstruction time, and RMSE. Values taken from respective publications. Best values bold.

9.2 Limitations

9.2.1 Technical Limitations

3D Relationships

The proposed 4D MRI framework employs a 2D network, which limits its ability to fully comprehend the 3D relationships between navigator and data slices. As the distance between a data slice and the navigator increases, the 3D relationships weaken, resulting in poorer reconstruction quality. To address this issue, one potential approach is to divide the volume into distance ranges and train a separate network for each range. This would reinforce knowledge of 3D relationships across larger distances. We anticipate that an ensemble of such networks would significantly improve quality with a fixed level of training data or maintain consistent quality with less training data. Alternatively, a 3D architecture like the 3D U-Net (Çiçek et al., 2016) instead of a 2D one might make it easier to learn the 3D spatial relations of the liver motion. In that case, the training task might also be reformulated to directly predict the 3D motion field, which could be beneficial for use in radiation therapy or intervention assistance systems.

Model Re-Use

Although the application of transfer learning strategies, diminishes the required amount of training data within the framework, it is still a separate model trained for each subject. Data of more different subjects and patients is needed to to explore the feasibility of employing a single model that strongly and sufficiently generalizes across observed subjects.

Basic Deep Learning Methods

In the presented work a prove of concept is shown. Basic architectures, like the UNet and normal convolutional layers as well as basic deep learning methods, like fine tuning and transfer learning were utilized to show a solution to a complex problem, i.e., 4D æmri reconstruction. In future work the utilization of more advanced techniques should be investigated. This includes exploring state-of-the-art architectures, leveraging advanced optimization algorithms, incorporating attention mechanisms,

and exploring novel regularization techniques. By embracing these advanced methods, the method could achieve even more accurate and robust results. In the case of retrospective use of the presented framework, it would be interesting to virtually increase the amount of training data by incorporating navigator interpolation (Karani et al., 2018; Zhang et al., 2018) , and data interpolation (Tanner et al., 2014) to double the temporal resolution to 83 ms to increase the prediction quality. It would also be interesting to investigate the use of coordConv layers (R. Liu et al., 2018), which artificially introduce positional context to each convolution, in place of normal convolutions to improve prediction quality. This seems very promising because the spatial component of the learning task is dominant.

9.2.2 Methodological Limitations

Compatibility with other Organs

It's important to note that the presented method has not yet been tested on other organs, which are a place of interventions and subject to breathing motion, like the lung. While the current focus has been on demonstrating feasibility and efficacy in the context of the liver, future investigations should involve extending the data basis to encompass diverse anatomical structures such as the lungs and other organs. This expansion will be crucial for validating the generalizability and applicability of the approach across a broader range of imaging guided interventions.

Small Data Basis and Biases

It's important to acknowledge the limitations regarding subject selection underlining the presented work. The subjects included in the dataset are not be representative of the global population, as they predominantly consist of healthy individuals from European descent in their 20s and 30s, with a limited representation of women. This skewed sample population raises concerns about the potential presence of biases within the model or framework. It is imperative to investigate whether the trained model exhibits biases related to factors such as race, gender, or other demographic characteristics.

For instance, it's known that liver shape can vary among different ethnic groups, with Europeans having distinct liver shapes compared to Asians. Therefore, applying the proposed method to individuals from diverse ethnic backgrounds may result in varying degrees of quality degradation, highlighting the importance of assessing the

robustness and generalizability of the model across different populations. Also the data set used in this work contains only healthy subjects. New studies are needed to conclude how well the 4D MRI models generalize to patient data from image guided liver interventions and other clinical settings.

To mitigate these concerns, continuous fine-tuning of the model with each new subject and intervention will be necessary. This iterative refinement process will help address potential biases and improve the model's adaptability to a wider range of demographic and anatomical variations. However, thorough investigation and validation are essential to ensure the reliability and fairness of the model across diverse populations.

Ground Truth

As described in previous chapters, ground truth data is only available for a single slice position within each predicted volume. Albeit different slices for different volumes, in total encompassing the full liver. Still, for a single predicted volume, there is always just one slice for which the prediction error can be measured. This limitation is inherent to the problem itself. If it were possible to acquire ground truth of the full liver for a specific breathing state within 200 ms the problem of 4D MRI would be solved already. Consequently, the evaluation metrics are performed on a slice level and further primarily encompass the entire image rather than focusing solely on the liver region. While this approach might bias metrics towards larger regions within the image, it provides insights into the model's ability to capture and reconstruct anatomical structures beyond the liver, which is valuable for assessing its generalizability and robustness. However, it's worth noting that evaluating the entire image may obscure the specific performance of the model within the liver region. Thus, future investigations may consider refining the evaluation metrics to focus specifically on liver-related metrics, e.g., using automatic liver segmentation to asses metrics only within the liver, for a more nuanced assessment of the model's efficacy. Despite these limitations, the current evaluation approach provides valuable insights into the model's performance and lays the groundwork for further refinement and optimization in future research endeavors.

Quantitative Comparison

In this thesis, a quantitative comparison was conducted with several other works. However, it was limited to metrics reported in the respective papers. While this

comparison provides valuable insights, it's crucial to recognize that the metrics used for comparison may vary across different studies, which can influence the comprehensiveness of the evaluation. Only one method from the literature was reimplemented and tested within the same test setup as the proposed method in this work.

Furthermore, it's important to consider that the prediction time score is dependent of the hardware used for evaluation, even though the experiments were conducted on the same hardware throughout the study. While maintaining hardware consistency ensures reliability within the study, it's essential to recognize that the prediction time may vary across different hardware configurations of other studies.

Moreover, enhancing the evaluation process by incorporating additional metrics and employing cross-validation techniques can further improve the accuracy and reliability of the results. For example, evaluating performance using metrics beyond those reported in the literature can provide a more comprehensive understanding of the framework's capabilities.

9.2.3 Medical Limitations

The medical limitations regard mainly the clinical transfer that has not yet been done. It is divided into three aspects. First, the clinical workflow, second, the clinical-technical limitations, and lastly the clinical evaluation.

Clinical Workflow

Workflow: The workflow outlined in this thesis remains in a prototypical research state and has not yet been tailored for potential end-users, lacking consideration for aspects such as clinical workflow or user experience. The setup of the required MRI sequences on the MRI machine remains a task for experts. It requires reading the papers that this thesis is based on and involves significant complexity, particularly with Siemens MRI machines where each slice position must be adjusted separately. Additionally, the mental workload associated with this setup process has yet to been investigated in the future and needs to be addressed by streamlining it.

Pipeline: Moreover, the framework currently has no continuous pipeline in place. Pre-acquisition tasks, such as setting up sequences and selecting navigator positions, as well as post-acquisition tasks like transferring MR images to the training machine and initiating model finetuning with new subject data, are not seamlessly integrated. Similarly, prior to and during interventions, processes such as detecting the correct navigator slice position and transferring interventional images to the model for automatic inference are not streamlined. Furthermore, there is a lack of direct visualization tools for radiologists and a lack of an interface to assistance systems that want to make use of the real-time information about the breathing motion. This lack of a continuous pipeline could be addressed in the future using existing interfaces like Access-I. Also, in the case of assistance systems, part of the user experience has to be addressed by the assistance system itself.

Finally, there has been no systematic survey conducted on various current interventional workflows, and it's possible that some workflows may struggle to integrate this framework effectively. Appropriate studies still need to be carried out.

Clinical-Technical Limitations

From a clinical-technical perspective, several limitations merit consideration.

Needle Insertion: The impact of needle insertion has not been investigated within the scope of this work. This omission raises questions about the robustness and efficacy of the proposed method in scenarios involving needle insertion, which is a common aspect of many medical interventions. To address this concern, one could investigate the possibility of angulated navigator slices that always contain the needle, By incorporating data from the needle into the navigator slice, the model could gain insights into the altered motion and deformation of the organ caused by the needle insertion. It's important to note that integrating such data into the training database would be necessary to ensure the model's ability to effectively learn and adapt to these changes.

Vendor Compatibility: The framework has not been tested across all MRI vendors. Variations in MRI machine specifications and software implementations may introduce inconsistencies or limitations in the applicability of the method across different platforms. Thus, broader vendor compatibility testing is needed to ensure the framework's applicability and reliability across diverse MRI systems.

Sequence Compatibility: Similarly, the framework has not been evaluated with a wide range of interventional MRI sequences. Different sequences offer distinct advantages and may be preferred for specific interventional scenarios. Assessing the performance of the framework with various sequences is essential in future work to ascertain its suitability and effectiveness across different imaging protocols.

Overall, these clinical-technical limitations underscore the need for further research to address potential challenges and ensure the robustness and applicability of the proposed method in clinical settings.

Clinical Evaluation

Data Privacy Concerns: The potential implications of data privacy have not yet been fully addressed. Given the sensitive nature of medical data, it is essential to implement robust data privacy measures to protect patient confidentiality and comply with different national regulatory requirements. Failure to address data privacy concerns could hinder the adoption of the proposed method in clinical settings and undermine patient trust.

Clinical Validation: While initial validation studies have been conducted, further clinical validation involving collaboration with medical experts and real interventional scenarios is imperative. This will provide valuable insights into the method's clinical utility, safety, and effectiveness in real-world settings.

Addressing these clinical considerations is essential to ensure the successful translation of the proposed method from research to clinical practice. By prioritizing data privacy, engaging medical experts in the validation process, and conducting rigorous clinical evaluations, researchers can establish the method's credibility, foster trust among healthcare professionals, and ultimately enhance patient care outcomes.

9.3 Future Work

Overall, a multidisciplinary approach that combines clinical expertise, technical innovation, and regulatory compliance is essential to advance the method and facilitate its translation from research to clinical practice.

9.3.1 Further Technical Development

Continued technical development is necessary to enhance the method's performance, scalability, and robustness. This will involve optimizing hardware and software integration and refining the deep learning methods, such as other transfer learning techniques, ensemble methods, or attention mechanisms. As well as investigating other slice orientations to follow a potential needle with the navigator.

9.3.2 Clinical Validation and Integration

Conducting comprehensive clinical validation studies involving medical experts and real interventional scenarios is essential. This would involve testing the method in diverse clinical settings to assess its efficacy, safety, and usability. Integration into existing clinical workflows and systems would also need to be explored to ensure seamless adoption and integration into routine clinical practice.

According to our medical partners, the next step in a clinical research scenario would be expanding our method to simulate the breathing motion of planning data from patients. This adaptation has the potential to revolutionize treatment planning in various medical disciplines, particularly in radiation therapy and image-guided interventions. Simulating patient-specific breathing motion would enable clinicians to anticipate and account for respiratory motion during treatment planning, leading to more precise and effective treatment delivery. For example, in radiation therapy, accurately modeling breathing motion can help optimize treatment plans to minimize radiation exposure to healthy tissues while maximizing dose delivery to target areas. Similarly, in image-guided interventions, simulating patient-specific breathing patterns can enhance the path planning by accounting for organ motion and deformation and help the radiologist mentally prepare for the patient's specific breathing movement patterns.

9.3.3 Data Privacy and Regulatory Compliance

Addressing data privacy concerns and ensuring compliance with regulatory requirements is crucial and needs to be done in the future, implementing robust data privacy measures and protocols to protect patient confidentiality and comply with regulatory standards.

Bibliography

Akinyemiju, Tomi, Semaw Abera, Muktar Ahmed, et al. (2017). “The burden of primary liver cancer and underlying etiologies from 1990 to 2015 at the global, regional, and national level: results from the global burden of disease study 2015”. In: *JAMA oncology* 3.12, pp. 1683–1691 (cit. on p. 10).

Alpers, Julian (2023). *Improving thermal cancer treatment with 2D to 3D heat map reconstruction* (cit. on p. 59).

Anatomy and physiology of the liver – Canadian Cancer Society (2015). https://web.archive.org/web/20150626110554/http://www.cancer.ca/en/cancer-information/cancer-type/liver/anatomy-and-physiology/?region=on. Accessed: 2024-02-13 (cit. on p. 10).

Bale, Reto, Gerlig Widmann, and DI Rudolf Stoffner (2010). “Stereotaxy: breaking the limits of current radiofrequency ablation techniques”. In: *European Journal of Radiology* 75.1, pp. 32–36 (cit. on p. 14).

Barth, Markus, Felix Breuer, Peter J Koopmans, David G Norris, and Benedikt A Poser (2016). “Simultaneous multislice (SMS) imaging techniques”. In: *Magnetic Resonance in Medicine* 75.1, pp. 63–81 (cit. on p. 77).

Bernstein, Matt A, Kevin F King, and Xiaohong Joe Zhou (2004). *Handbook of MRI pulse sequences*. Elsevier (cit. on p. 22).

Biewald, Lukas (2020). *Experiment Tracking with Weights and Biases*. Software available from wandb.com (cit. on p. 92).

Bled, Emilie, Wadie Ben Hassen, Line Pourtau, et al. (2011). “Real-time 3D MRI of contrast agents in whole living mice”. In: *Contrast Media & Molecular Imaging* 6.4, pp. 275–281 (cit. on p. 39).

Bradski, G. (2000). “The OpenCV Library”. In: *Dr. Dobb’s Journal of Software Tools* (cit. on p. 68).

Bray, Freddie, Jacques Ferlay, Isabelle Soerjomataram, et al. (2018). “Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries”. In: *CA: A Cancer Journal for Clinicians* 68.6, pp. 394–424 (cit. on p. 10).

Brown, Robert W, Y-C Norman Cheng, E Mark Haacke, Michael R Thompson, and Ramesh Venkatesan (2014). *Magnetic resonance imaging: physical principles and sequence design*. John Wiley & Sons (cit. on pp. 21, 22).

Cai, Jing, Zheng Chang, Zhiheng Wang, William Paul Segars, and Fang-Fang Yin (2011). "Four-dimensional magnetic resonance imaging (4D-MRI) using image-based respiratory surrogate: a feasibility study". In: *Medical Physics* 38.12, pp. 6384–6394 (cit. on pp. 40, 47, 162).

Celicanin, Zarko, Oliver Bieri, Frank Preiswerk, et al. (2015). "Simultaneous acquisition of image and navigator slices using CAIPIRINHA for 4D MRI". In: *Magnetic Resonance in Medicine* 73.2, pp. 669–676 (cit. on pp. 44, 47, 77, 102, 162).

Chollet, Francois et al. (2015). *Keras*. https://github.com/fchollet/keras (cit. on pp. 92, 112).

Chu, Katrina F and Damian E Dupuy (2014). "Thermal ablation of tumours: biological mechanisms and advances in therapy". In: *Nature Reviews Cancer* 14.3, pp. 199–208 (cit. on p. 15).

Çiçek, Özgün, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger (2016). "3D U-Net: learning dense volumetric segmentation from sparse annotation". In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*. Springer, pp. 424–432 (cit. on p. 163).

Cleary, Kevin and Terry M Peters (2010). "Image-guided interventions: technology review and clinical applications". In: *Annual Review of Biomedical Engineering* 12, pp. 119–142 (cit. on p. 1).

Colvill, Emma, Jeremy Booth, Simeon Nill, et al. (2016). "A dosimetric comparison of real-time adaptive and non-adaptive radiotherapy: a multi-institutional study encompassing robotic, gimbaled, multileaf collimator and couch tracking". In: *Radiotherapy and Oncology* 119.1, pp. 159–165 (cit. on p. 1).

Deng, Zixin, Jianing Pang, Wensha Yang, et al. (2016). "Four-dimensional MRI using three-dimensional radial sampling with respiratory self-gating to characterize temporal phase-resolved respiratory motion in the abdomen". In: *Magnetic Resonance in Medicine* 75.4, pp. 1574–1585 (cit. on pp. 40, 47, 162).

Dinkel, Julien, Christian Hintze, Ralf Tetzlaff, et al. (2009). "4D-MRI analysis of lung tumor motion in patients with hemidiaphragmatic paralysis". In: *Radiotherapy and Oncology* 91.3, pp. 449–454 (cit. on p. 39).

Eldeniz, Cihat, Weijie Gan, Sihao Chen, et al. (2021). "Phase2Phase: respiratory motion-resolved reconstruction of free-breathing magnetic resonance imaging using deep learning without a ground truth for improved liver imaging". In: *Investigative Radiology* 56.12, pp. 809–819 (cit. on pp. 42, 43, 47, 162).

Fischbach, Frank, Jürgen Bunke, Markus Thormann, et al. (2011). "MR-guided freehand biopsy of liver lesions with fast continuous imaging using a 1.0-T open MRI scanner: experience in 50 patients". In: *Cardiovascular and Interventional Radiology* 34, pp. 188–192 (cit. on p. 18).

Fritz, J, N Tzaribachev, C Thomas, et al. (2011). "Evaluation of MR imaging guided steroid injection of the sacroiliac joints for the treatment of children with refractory enthesitis-related arthritis". In: *European Radiology* 21, pp. 1050–1057 (cit. on p. 18).

Galle, Peter R, Alejandro Forner, Josep M Llovet, et al. (2018). “EASL clinical practice guidelines: management of hepatocellular carcinoma”. In: *Journal of Hepatology* 69.1, pp. 182–236 (cit. on pp. 10, 14, 15).

Guan, Hao and Mingxia Liu (2021). “Domain adaptation for medical image analysis: a survey”. In: *IEEE Transactions on Biomedical Engineering* 69.3, pp. 1173–1185 (cit. on p. 108).

Gueulette, John, Hans Blattmann, Eros Pedroni, et al. (2005). “Relative biologic effectiveness determination in mouse intestine for scanning proton beam at Paul Scherrer Institute, Switzerland. Influence of motion”. In: *International Journal of Radiation Oncology* Biology* Physics* 62.3, pp. 838–845 (cit. on p. 1).

Gulamhussene, Gino, Oleksii Bashkanov, Jazan Omari, et al. (2023a). “Using Training Samples as Transitive Information Bridges in Predicted 4D MRI”. In: *Workshop on Medical Image Learning with Limited and Noisy Data*. Springer, pp. 237–245 (cit. on p. 127).

Gulamhussene, Gino, Fabian Joeres, Marko Rak, Maciej Pech, and Christian Hansen (2020). “4D MRI: Robust sorting of free breathing MRI slices for use in interventional settings”. In: *PloS one* 15.6, e0235175 (cit. on pp. 51, 63, 102).

Gulamhussene, Gino, Anneke Meyer, Marko Rak, et al. (2022). “Predicting 4D liver MRI for MR-guided interventions”. In: *Computerized Medical Imaging and Graphics* 101, p. 102122 (cit. on p. 81).

Gulamhussene, Gino, Marko Rak, Oleksii Bashkanov, et al. (2023b). “Transfer-learning is a key ingredient to fast deep learning-based 4D liver MRI reconstruction”. In: *Scientific Reports* 13.1, p. 11227 (cit. on p. 107).

Ha, In Young, Matthias Wilms, Heinz Handels, and Mattias P Heinrich (2018). “Model-based sparse-to-dense image registration for realtime respiratory motion estimation in image-guided interventions”. In: *IEEE Transactions on Biomedical Engineering* 66.2, pp. 302–310 (cit. on pp. 1, 2).

Han, Fei, Ziwu Zhou, Minsong Cao, et al. (2017). “Respiratory motion-resolved, self-gated 4D-MRI using rotating cartesian k-space (ROCK)”. In: *Medical Physics* 44.4, pp. 1359–1368 (cit. on pp. 40, 47, 162).

Han, Fei, Ziwu Zhou, Dongsu Du, et al. (2018). “Respiratory motion-resolved, self-gated 4D-MRI using Rotating Cartesian K-space (ROCK): Initial clinical experience on an MRI-guided radiotherapy system”. In: *Radiotherapy and Oncology* 127.3, pp. 467–473 (cit. on p. 1).

Harris, Wendy, Chunhao Wang, Fang-Fang Yin, Jing Cai, and Lei Ren (2018). “A Novel method to generate on-board 4D MRI using prior 4D MRI and on-board kV projections from a conventional LINAC for target localization in liver SBRT”. In: *Medical Physics* 45.7, pp. 3238–3245 (cit. on pp. 41, 47, 162).

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun (2016). “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778 (cit. on p. 90).

Heinrich, Florian, Fabian Joeres, Kai Lawonn, and Christian Hansen (2019). “Comparison of projective augmented reality concepts to support medical needle insertion”. In: *IEEE Transactions on Visualization and Computer Graphics* 25.6, pp. 2157–2167 (cit. on p. 2).

Hu, Y, S Caruthers, D Low, P Parikh, and S Mutic (2012). “WE-G-213CD-04: A Triggering System to Guide 4DMRI Image Acquisition”. In: *Medical Physics* 39.6Part28, pp. 3971–3971 (cit. on pp. 40, 47, 162).

Hu, Yanle, Shelton D Caruthers, Daniel A Low, Parag J Parikh, and Sasa Mutic (2013). “Respiratory amplitude guided 4-dimensional magnetic resonance imaging”. In: *International Journal of Radiation Oncology* Biology* Physics* 86.1, pp. 198–204 (cit. on pp. 40, 47, 162).

Huang, Gao, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger (2017). “Densely connected convolutional networks”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708 (cit. on p. 90).

Ioffe, Sergey and Christian Szegedy (2015). “Batch normalization: Accelerating deep network training by reducing internal covariate shift”. In: *International conference on machine learning*. pmlr, pp. 448–456 (cit. on p. 28).

Kägebein, Urte (2018). “MRT-geführte Ablation mit Hilfe des optischen Moiré Phase Trackingsystems”. PhD thesis. Otto-von-Guericke-Universität Magdeburg (cit. on p. 21).

Kägebein, Urte, Frank Godenschweger, Brian SR Armstrong, et al. (2018a). “Percutaneous MR-guided interventions using an optical Moire Phase tracking system: Initial results”. In: *Plos one* 13.10, e0205394 (cit. on p. 19).

Kägebein, Urte, Oliver Speck, Frank Wacker, and Bennet Hensen (2018b). “Motion correction in proton resonance frequency–based thermometry in the liver”. In: *Topics in Magnetic Resonance Imaging* 27.1, pp. 53–61 (cit. on p. 17).

Karani, Neerav, Christine Tanner, Sebastian Kozerke, and Ender Konukoglu (2018). “Reducing navigators in free-breathing abdominal MRI via temporal interpolation using convolutional neural networks”. In: *IEEE Transactions on Medical Imaging* 37.10, pp. 2333–2343 (cit. on pp. 44, 47, 124, 157, 162, 164).

Kavaluus, Henna, Tiina Seppälä, Lauri Koivula, et al. (2020). “Retrospective four-dimensional magnetic resonance imaging of liver: Method development”. In: *Journal of Applied Clinical Medical Physics* 21.12, pp. 304–313 (cit. on pp. 41, 47, 162).

Keijnemans, Katrinus, Pim TS Borman, Prescilla Uijtewaal, et al. (2022). “A hybrid 2D/4D-MRI methodology using simultaneous multislice imaging for radiotherapy guidance”. In: *Medical Physics* 49.9, pp. 6068–6081 (cit. on pp. 42, 47, 162).

Kim, Yoon-Chul, R Marc Lebel, Ziyue Wu, et al. (2014). “Real-time 3D magnetic resonance imaging of the pharyngeal airway in sleep apnea”. In: *Magnetic Resonance in Medicine* 71.4, pp. 1501–1510 (cit. on p. 39).

Kingma, Diederik P and Jimmy Ba (2014). “Adam: A method for stochastic optimization”. In: *arXiv preprint arXiv:1412.6980* (cit. on pp. 92, 112).

Koenig, Claudius W, Stephan H Duda, Jochen Truebenbach, et al. (2001). “MR-guided biopsy of musculoskeletal lesions in a low-field system”. In: *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine* 13.5, pp. 761–768 (cit. on p. 18).

König, Claudius W, Philippe L Pereira, Jochen Trübenbach, et al. (2003). “MR Imaging–Guided Adrenal Biopsy Using an Open Low-Field-Strength Scanner and MR Fluoroscopy”. In: *American Journal of Roentgenology* 180.6, pp. 1567–1570 (cit. on p. 18).

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton (2017). “ImageNet classification with deep convolutional neural networks”. In: *Communications of the ACM* 60.6, pp. 84–90 (cit. on p. 25).

Lambert, J, N Suchowerska, DR McKenzie, and M Jackson (2005). “Intrafractional motion during proton beam scanning”. In: *Physics in Medicine & Biology* 50.20, p. 4853 (cit. on p. 1).

LeCun, Yann, Bernhard Boser, John S Denker, et al. (1989). “Backpropagation applied to handwritten zip code recognition”. In: *Neural Computation* 1.4, pp. 541–551 (cit. on p. 25).

LeCun, Yann, Corinna Cortes, Chris Burges, et al. (2010). *MNIST handwritten digit database* (cit. on p. 29).

Lewin, Jonathan S, Sherif Gamal Nour, and Jeffrey L Duerk (2000). “Magnetic resonance image-guided biopsy and aspiration”. In: *Topics in Magnetic Resonance Imaging* 11.3, pp. 173–183 (cit. on p. 18).

Li, Peng, Jialei Chen, Dong Nan, et al. (2022). “Motion-aligned 4D-MRI reconstruction using higher degree total variation and locally low-rank regularization”. In: *Magnetic Resonance Imaging* 93, pp. 97–107 (cit. on pp. 42, 47, 162).

Lindt, Tessa van de, Jan-Jakob Sonke, Marlies Nowee, et al. (2018). “A self-sorting coronal 4D-MRI method for daily image guidance of liver lesions on an MR-LINAC”. In: *International Journal of Radiation Oncology* Biology* Physics* 102.4, pp. 875–884 (cit. on pp. 41, 47, 162).

Liu, Rosanne, Joel Lehman, Piero Molino, et al. (2018). “An intriguing failing of convolutional neural networks and the coordconv solution”. In: *Advances in Neural Information Processing Systems* 31 (cit. on pp. 124, 164).

Liu, Yilin, Fang-Fang Yin, Zheng Chang, et al. (2014). “Investigation of sagittal image acquisition for 4D-MRI with body area as respiratory surrogate”. In: *Medical Physics* 41.10, p. 101902 (cit. on pp. 40, 47, 162).

Lowekamp, Bradley Christopher, David T Chen, Luis Ibáñez, and Daniel Blezek (2013). “The design of SimpleITK”. In: *Frontiers in Neuroinformatics* 7, p. 45 (cit. on p. 32).

Malek, Nisar P, Sebastian Schmidt, Petra Huber, Michael P Manns, and Tim F Greten (2014). “The diagnosis and treatment of hepatocellular carcinoma”. In: *Deutsches Ärzteblatt International* 111.7, p. 101 (cit. on p. 14).

Merchavy, Shlomo, Judith Luckman, Michal Guindy, Yoram Segev, and Avi Khafif (2016). “4D MRI for the localization of parathyroid adenoma: a novel method in evolution”. In: *Otolaryngology–Head and Neck Surgery* 154.3, pp. 446–448 (cit. on p. 1).

Meschini, Giorgia, Chiara Paganelli, Chiara Gianoli, et al. (2019). “A clustering approach to 4D MRI retrospective sorting for the investigation of different surrogates”. In: *Physica Medica* 58, pp. 107–113 (cit. on pp. 41, 47, 162).

Mewes, André (2019). “Projector-based augmented reality and touchless interaction to support MRI-guided interventions”. PhD thesis. Otto-von-Guericke-Universität Magdeburg (cit. on pp. 15, 16, 18, 19).

Mewes, André, Florian Heinrich, Urte Kägebein, et al. (2019). “Projector-based augmented reality system for interventional visualization inside MRI scanners”. In: *The International Journal of Medical Robotics and Computer Assisted Surgery* 15.1, e1950 (cit. on p. 2).

Mickevicius, Nikolai J and Eric S Paulson (2017). “Investigation of undersampling and reconstruction algorithm dependence on respiratory correlated 4D-MRI for online MR-guided radiation therapy”. In: *Physics in Medicine & Biology* 62.8, p. 2910 (cit. on p. 42).

Navest, RJM, S Mandija, T Bruijnen, et al. (2020). “The noise navigator: a surrogate for respiratory-correlated 4D-MRI for motion characterization in radiotherapy”. In: *Physics in Medicine & Biology* 65.1, 01NT02 (cit. on pp. 41, 47, 162).

Paganelli, Chiara, Paul Summers, Massimo Bellomi, Guido Baroni, and Marco Riboldi (2015). “Liver 4DMRI: a retrospective image-based sorting method”. In: *Medical Physics* 42.8, pp. 4814–4821 (cit. on pp. 40, 47, 162).

Pang, J, W Yang, X Bi, et al. (2016). “4D-MRI with Iterative Motion Correction and Averaging Improves Image SNR and Reduces Streaking Artifacts without Compromising Tumor Motion Trajectory”. In: *International Journal of Radiation Oncology, Biology, Physics* 96.2, S62–S63 (cit. on p. 42).

Pereira, Philippe L (2007). “Actual role of radiofrequency ablation of liver metastases”. In: *European Radiology* 17, pp. 2062–2070 (cit. on pp. 10, 15).

Pooley, Robert A (2005). “Fundamental physics of MR imaging”. In: *Radiographics* 25.4, pp. 1087–1099 (cit. on pp. 20, 21).

Preiswerk, Frank, Matthew Toews, Cheng-Chieh Cheng, et al. (2017). “Hybrid MRI-Ultrasound acquisitions, and scannerless real-time imaging”. In: *Magnetic Resonance in Medicine* 78.3, pp. 897–908 (cit. on p. 77).

Qiu, Wenyuan, Dongxiao Li, Xinyu Jin, et al. (2019). “Sliding motion compensated low-rank plus sparse (SMC-LS) reconstruction for high spatiotemporal free-breathing liver 4D DCE-MRI”. In: *Magnetic Resonance Imaging* 58, pp. 56–66 (cit. on pp. 44, 47, 162).

Rank, Christopher M, Thorsten Heußer, Maria TA Buzan, et al. (2017). “4D respiratory motion-compensated image reconstruction of free-breathing radial MR data with very high undersampling”. In: *Magnetic Resonance in Medicine* 77.3, pp. 1170–1183 (cit. on pp. 41, 47, 162).

Richter, Julian AJ, Tobias Wech, Andreas M Weng, et al. (2020). “Free-breathing self-gated 4D lung MRI using wave-CAIPI”. In: *Magnetic Resonance in Medicine* 84.6, pp. 3223–3233 (cit. on pp. 41, 47, 162).

Romaguera, Liset Vázquez, Nils Olofsson, Rosalie Plantefève, et al. (2019). “Automatic self-gated 4D-MRI construction from free-breathing 2D acquisitions applied on liver images”. In: *International Journal of Computer Assisted Radiology and Surgery* 14.6, pp. 933–944 (cit. on pp. 45, 47, 162).

Ronneberger, Olaf, Philipp Fischer, and Thomas Brox (2015). “U-net: convolutional networks for biomedical image segmentation”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241 (cit. on pp. 30, 90).

Rothgang, Eva, Wesley D Gilson, Frank Wacker, et al. (2013). “Rapid freehand MR-guided percutaneous needle interventions: An image-based approach to improve workflow and feasibility”. In: *Journal of Magnetic Resonance Imaging* 37.5, pp. 1202–1212 (cit. on pp. 18, 19).

Roujol, Sébastien, Mario Ries, Bruno Quesson, Chrit Moonen, and Baudouin Denis de Senneville (2010). “Real-time MR-thermometry and dosimetry for interventional guidance on abdominal organs”. In: *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine* 63.4, pp. 1080–1087 (cit. on p. 17).

Rumelhart, David E, Geoffrey E Hinton, and Ronald J Williams (1986). “Learning representations by back-propagating errors”. In: *nature* 323.6088, pp. 533–536 (cit. on p. 26).

Schreiter, Josefine, Tonia Mielke, Danny Schott, et al. (2023). “A multimodal user interface for touchless control of robotic ultrasound”. In: *International Journal of Computer Assisted Radiology and Surgery* 18.8, pp. 1429–1436 (cit. on p. 19).

Shorten, Connor and Taghi M Khoshgoftaar (2019). “A survey on image data augmentation for deep learning”. In: *Journal of big data* 6.1, pp. 1–48 (cit. on pp. 27, 28).

Sibulesky, Lena (2013). “Normal liver anatomy”. In: *Clinical Liver Disease* 2.Suppl 1, S1–S3 (cit. on p. 10).

Siebenthal, M von, Ph Cattin, Urs Gamper, Antony Lomax, and Gábor Székely (2005). “4D MR imaging using internal respiratory gating”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 336–343 (cit. on p. 64).

Siebenthal, Martin von, Gabor Szekely, Urs Gamper, et al. (2007). “4D MR imaging of respiratory organ motion and its variability”. In: *Physics in Medicine & Biology* 52.6, p. 1547 (cit. on pp. 43, 47, 59, 64, 77, 102, 162).

Srivastava, Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov (2014). “Dropout: a simple way to prevent neural networks from overfitting”. In: *The journal of machine learning research* 15.1, pp. 1929–1958 (cit. on p. 28).

Sung, Hyuna, Jacques Ferlay, Rebecca L Siegel, et al. (2021). “Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries”. In: *CA: A Cancer Journal for Clinicians* 71.3, pp. 209–249 (cit. on p. 10).

Szegedy, Christian, Wei Liu, Yangqing Jia, et al. (2015). “Going deeper with convolutions”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9 (cit. on p. 90).

Tanner, Christine, Dirk Boye, Golnoosh Samei, and Gabor Szekely (2012). “Review on 4D models for organ motion compensation”. In: *Critical Reviews™ in Biomedical Engineering* 40.2, pp. 135–154 (cit. on p. 2).

Tanner, Christine, Golnoosh Samei, and Gábor Székely (2014). “Improved reconstruction of 4D-MR images by motion predictions”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 146–153 (cit. on pp. 44, 47, 102, 124, 162, 164).

Tokuda, Junichi, Shigehiro Morikawa, Hasnine A Haque, et al. (2008). “Adaptive 4D MR imaging using navigator-based respiratory signal for MRI-guided therapy”. In: *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine* 59.5, pp. 1051–1061 (cit. on pp. 40, 47, 162).

Tryggestad, Erik, Aaron Flammang, Sarah Han-Oh, et al. (2013). “Respiration-based sorting of dynamic MRI to derive representative 4D-MRI for radiotherapy planning”. In: *Medical Physics* 40.5, p. 051909 (cit. on pp. 40, 47, 162).

Tsekos, Nikolaos V, Alpay Özcan, and Eftychios Christoforou (2005). “A prototype manipulator for magnetic resonance-guided interventions inside standard cylindrical magnetic resonance imaging scanners”. In: *Journal of Biomechanical Engineering* 127.6, pp. 972–980 (cit. on p. 19).

Wachinger, Christian, Mehmet Yigitsoy, Erik-Jan Rijkhorst, and Nassir Navab (2012). “Manifold learning for image-based breathing gating in ultrasound and MRI”. In: *Medical Image Analysis* 16.4, pp. 806–818 (cit. on pp. 43, 47, 162).

Xing, Lei, Brian Thorndyke, Eduard Schreibmann, et al. (2006). “Overview of image-guided radiation therapy”. In: *Medical Dosimetry* 31.2, pp. 91–112 (cit. on pp. 1, 2).

Yang, Zi, Lei Ren, Fang-Fang Yin, Xiao Liang, and Jing Cai (2020). “Motion robust 4D-MRI sorting based on anatomic feature matching: A digital phantom simulation study”. In: *Radiation Medicine and Protection* 1.1, pp. 41–47 (cit. on pp. 42, 47, 162).

Yuan, Jing, Oi Lei Wong, Yihang Zhou, Kin Yin Chueng, and Siu Ki Yu (2019). “A fast volumetric 4D-MRI with sub-second frame rate for abdominal motion monitoring and characterization in MRI-guided radiotherapy”. In: *Quantitative Imaging in Medicine and Surgery* 9.7, p. 1303 (cit. on pp. 45, 47, 102, 162).

Zhang, Lin, Neerav Karani, Christine Tanner, and Ender Konukoglu (2018). “Temporal interpolation via motion field prediction”. In: *arXiv preprint arXiv:1804.04440* (cit. on pp. 44, 47, 102, 124, 157, 162, 164).

Declaration

I hereby certify that I have prepared this thesis without the unauthorized assistance of third parties and without the use of resources other than those indicated; external and my own sources used are identified as such. In particular, I have not used the help of a commercial doctoral advisor. Third parties have neither directly nor indirectly received monetary benefits from me for work related to the content of the submitted dissertation. In particular, I did not knowingly:

- invent any results or conceal contradictory results,
- intentionally misuse statistical procedures to interpret data in an unjustified manner,
- plagiarize other people's results or publications,
- distort the results of other research.

I am aware that violations of copyright law may give rise to injunctive relief and claims for damages by the author as well as criminal prosecution by the law enforcing authorities. In this dissertation, AI systems, e.g., large language models, were employed for editing and improving the grammar and wording in parts of the text. The author verified that the statement of the initial text parts remained unchanged. The work has not yet been submitted as a dissertation in the same or a similar form either in Germany or abroad and has not yet been published as a whole.

Magdeburg, 07. Oktober 2024

Gino Gulamhussene