

Aus dem Lehrstuhl für Computerunterstützte Klinische Medizin
der Medizinischen Fakultät Mannheim
(Komm. Direktor: Prof. Dr. Ing. Frank G. Zöllner)

Development of Methods for the Registration of Multimodal Medical Images Using Deep Learning

Inauguraldissertation
zur Erlangung des Doctor scientiarum humanarum (Dr. sc. hum.)
der
Medizinischen Fakultät Mannheim
der Ruprecht-Karls-Universität zu Heidelberg

vorgelegt von
Anika Strittmatter, M.Sc.
aus
Mannheim
2024

Dekan: Prof. Dr. med. Sergij Goerdts
Referent: Prof. Dr. Ing. Frank G. Zöllner

Abstract

In the field of medical image processing, multimodal medical image registration is an essential technique due to its fundamental role for integrating complementary information from diverse imaging modalities which is used in various clinical applications, e.g. in the diagnosis and planning and performing of treatments, e.g. for renal and oligometastatic diseases. Innovations in medical image registration increase the accuracy of image fusion, and thus improve patient care. Therefore, the aim of this thesis is to improve and develop new deep learning-based image registration methods for three-dimensional multimodal medical images.

In the first part, 20 different published neural networks were implemented for affine medical image registration and their generalisability to new data was evaluated using two multimodal medical datasets and an in-house developed neural network as benchmark. Only six and nine networks, respectively, significantly improved the pre-registration Dice coefficient and therefore generalized well to the new datasets. This shows that further research is needed to develop medical image registration methods that can be used for a wider range of medical applications.

In the second part, a novel multistage neural network for multimodal medical image registration was presented, addressing large rigid deformations typical in medical imaging. The performance of the novel multistage network, which applies rigid, affine and deformable transformations, was compared with that of a multistage network using affine and deformable transformations. Utilising the novel multistage network resulted in high spatial alignment (Dice up to 68.1 ± 24.6) and medically plausible transformations ($|J| \leq 0 : \leq 1.1\%$) on four multimodal medical datasets.

In the third part, neural networks for deformable multimodal medical image registration were developed that apply multiple steps and resolutions in a coarse-to-fine approach. The potential benefits and limitations of using multiple steps and resolutions were investigated using four multimodal medical datasets. The results demonstrated that the multistep networks achieve high structural similarity (NMI up to 0.33 ± 0.02 , Dice up to 90.8 ± 3.1) and medically plausible registration results ($|J| \leq 0 : < 0.5\%$) and utilising more resolutions (up to 4 in this study) results in increased spatial alignment.

In the fourth study, a groupwise multiresolution network was presented for the deformable registration of DCE-MR images for the assessment of lung perfusion in patients with congenital diaphragmatic hernia. Experimental results showed that the proposed groupwise network outperformed iterative and deep learning-based pairwise and groupwise registration methods and led to increased homogeneity of the perfusion maps, enhancing the accuracy of medical image analysis.

In conclusion, the deep learning-based medical image registration methods developed in this thesis improve the accuracy of medical image registration and may serve as valuable tool for clinical applications of image fusion, such as diagnosis of diseases and treatment planning. By enabling rapid registration, these techniques support real-time decision making and increase the accuracy of interventions, ultimately leading to better patient outcomes.

Zusammenfassung

Im Bereich der medizinischen Bildverarbeitung ist die multimodale medizinische Bildregistrierung aufgrund ihrer grundlegenden Rolle für die Integration komplementärer Informationen aus verschiedenen Bildgebungsmodalitäten eine wesentliche Technik, die in verschiedenen klinischen Anwendungen zum Einsatz kommt, z. B. bei der Diagnose und Planung und Durchführung von Behandlungen, z. B. bei Nieren- und Oligometastasenerkrankungen. Innovationen in der medizinischen Bildregistrierung erhöhen die Genauigkeit der Bildfusion und verbessern so die Patientenversorgung. Ziel dieser Arbeit ist daher, verbesserte und neue, auf Deep Learning basierende Bildregistrierungsmethoden für dreidimensionale multimodale medizinische Bilder zu entwickeln.

Im ersten Teil wurden 20 verschiedene publizierte neuronale Netze für die affine medizinische Bildregistrierung implementiert und ihre Generalisierbarkeit auf neue Daten anhand von zwei multimodalen medizinischen Datensätzen und einem intern entwickelten neuronalen Netz als Benchmark bewertet. Nur sechs bzw. neun Netzwerke verbesserten den Dice-Koeffizienten vor Registrierung signifikant und ließen sich daher gut auf die neuen Datensätze übertragen. Dies zeigt, dass weitere Forschung erforderlich ist, um Methoden zur medizinischen Bildregistrierung zu entwickeln, die für ein breiteres Spektrum medizinischer Anwendungen genutzt werden können.

Im zweiten Teil wurde ein neuartiges mehrstufiges neuronales Netz für die multimodale medizinische Bildregistrierung vorgestellt, das große rigide Verformungen berücksichtigt, die für die medizinische Bildgebung typisch sind. Die Leistung des neuen mehrstufigen Netzwerks, das rigide, affine und deformierbare Transformationen anwendet, wurde mit der Leistung eines mehrstufigen Netzwerks verglichen, das affine und deformierbare Transformationen einsetzt. Die Verwendung des neuen mehrstufigen Netzwerks führte zu einer hohen räumlichen Ausrichtung (Dice bis zu $68,1 \pm 24,6$) und medizinisch plausiblen Transformationen ($|J| \leq 0 : \leq 1,1\%$) bei vier multimodalen medizinischen Datensätzen.

Im dritten Teil wurden neuronale Netze für die deformierbare multimodale medizinische Bildregistrierung entwickelt, die mehrere Schritte und Auflösungen in einem Grob-zu-Fein-Ansatz anwenden. Die potenziellen Vorteile und Einschränkungen der Verwendung mehrerer Schritte und Auflösungen wurden anhand von vier multimodalen medizinischen Datensätzen untersucht. Die Ergebnisse zeigten, dass mehrstufige Netzwerke eine hohe strukturelle Ähnlichkeit (NMI bis zu $0,33 \pm 0,02$, Dice bis zu $90,8 \pm 3,1$) und medizinisch plausible Registrierungsergebnisse ($|J| \leq 0 : < 0.5\%$) erzielen und die Verwendung mehrerer Auflösungen (bis zu 4 in dieser Studie) zu einer erhöhten räumlichen Ausrichtung führt.

In der vierten Studie wurde ein gruppenbasiertes Netzwerk mit mehreren Auflösungen für die deformierbare Registrierung von DCE-MR-Bildern zur Beurteilung der Lungenperfusion bei Patienten mit kongenitaler Zwerchfellhernie vorgestellt. Experimentelle Ergebnisse zeigten, dass das vorgeschlagene gruppenweise Netzwerk iterative und auf Deep Learning basierende paarweise und gruppenweise Registrierungsmethoden übertraf und zu einer gesteigerten Homogenität der Perfusionskarten führte, wodurch die Genauigkeit der medizinischen Bildanalyse verbessert wurde.

Zusammengefasst verbessern die in dieser Arbeit entwickelten, auf Deep Learning basierenden medizinischen Bildregistrierungsverfahren die Genauigkeit der medizinischen Bildregistrierung und können als wertvolles Werkzeug für klinische Anwendungen der Bildfusion dienen, z. B. für die Diagnose von Krankheiten und die Behandlungsplanung. Indem sie eine schnelle Registrierung ermöglichen, unterstützen diese Verfahren die Entscheidungsfindung in Echtzeit und erhöhen die Genauigkeit von Eingriffen, was letztlich zu besseren Resultaten für die Patienten führt.

Preface

This is a cumulative dissertation, several chapters of this dissertation have already been published or are currently submitted for publication. The citations for these chapters are:

Chapter 3: A. Strittmatter, L. R. Schad, F. G. Zöllner. Deep learning-based affine medical image registration for multimodal minimal-invasive image-guided interventions – A comparative study on generalizability. *Zeitschrift für Medizinische Physik*, 34 (2), pp. 291-317, doi: 10.1016/j.zemedi.2023.05.003, 2024.

Chapter 4: A. Strittmatter, A. Caroli, F. G. Zöllner. A Multistage Rigid-Affine-Deformable Network for Three-Dimensional Multimodal Medical Image Registration. *Applied Sciences*, 13(24), 13298, doi: 10.3390/app132413298, 2023.

Chapter 5: A. Strittmatter, F. G. Zöllner. Multistep networks for deformable multimodal medical image registration. *IEEE Access* 12, pp. 82676–82692, doi: 10.1109/ACCESS.2024.3412216, 2024.

Chapter 6: A. Strittmatter, M. Weis, F. G. Zöllner. A Groupwise Multiresolution Network for DCE-MRI Image Registration. *Scientific Reports*. Submitted on 12.09.2024.

Summary of publications included in cumulative dissertations

Name of the doctoral student: Anika Strittmatter

Title of dissertation: Development of Methods for the Registration of Multimodal Medical Images Using Deep Learning

Supervised by: Prof. Dr. Ing. Frank G. Zöllner

☐ I wish to submit a cumulative dissertation, and hereby ask the doctoral committee to examine whether the quantity and quality of the proposed publications are sufficient to meet the requirements for a cumulative dissertation.

☒ The doctoral committee has previously examined whether my publications are suitable for a cumulative dissertation, and this is a final overview of the publications contained in my cumulative thesis.

1. List of peer-reviewed publications included in the cumulative dissertation. For each publication, provide a complete list of authors, title, journal, journal impact factor, and whether the manuscript has been accepted for publication, is in revision after peer review, or has been submitted and is awaiting peer review. Shared first authorships should be clearly indicated. Please also indicate whether the publication is an original research report, a review, or another type of article.

Publication 1: A. Strittmatter, L. R. Schad and F. G. Zöllner. Deep learning-based affine medical image registration for multimodal minimal-invasive image-guided interventions – A comparative study on generalizability. Zeitschrift für Medizinische Physik, 34 (2), pp. 291-317, doi: 10.1016/j.zemedi.2023.05.003, 2024. Original-Forschungsbericht, Published, Impact Factor: 2.4

Publication 2: A. Strittmatter, A. Caroli and F. G. Zöllner. A Multistage Rigid-Affine-Deformable Network for Three-Dimensional Multimodal Medical Image Registration. Applied Sciences, 13(24), 13298, doi: 10.3390/app132413298, 2023. Original-Forschungsbericht, Published, Impact Factor: 2.5

Publication 3: A. Strittmatter and F. G. Zöllner. Multistep Networks for Deformable Multimodal Medical Image Registration. IEEE Access, 12, 82676–82692, doi: 10.1109/ACCESS.2024.3412216, 2024. Original-Forschungsbericht, Published, Impact Factor: 3.4

Publication 4: A. Strittmatter, M. Weis and F. G. Zöllner. A Groupwise Multiresolution Network for DCE-MRI Image Registration. Scientific Reports. Original-Forschungsbericht, Submitted on: 12.09.2024, Impact Factor: 3.8

2. Summary of the doctoral student's contribution to the work reported in each manuscript.

Work steps	Publication 1	Publication 2	Publication 3	Publication 4
Conception (%)	95	95	95	95
Literature search (%)	100	100	100	100
Ethics proposal (%)	0	0	0	0
Animal experimentation proposal (%)	-	-	-	-
Data collection (%)	25	10	10	0
Data analysis (%)	100	100	100	95
Interpretation of results (%)	100	100	100	95
Manuscript writing (%)	95	95	95	95
Revision (%)	100	100	100	To be done

Indicate which figures and tables resulted from your dissertation work.				
---	--	--	--	--

(3) I hereby certify that this is a true representation of the doctoral student's contribution to the publications listed.

A. Stüttgen

Signature of the doctoral candidate

Frank Leut Zöllner

Signature of supervisor

Contents

Abstract	v
List of Figures	xvii
List of Tables	xix
List of Acronyms	xxi
1 Introduction and Outline	1
1.1 Motivation	1
1.2 Outline	2
1.3 Citation of Previous Publications	3
2 Background	5
2.1 Medical Background	5
2.1.1 Lung	5
2.1.2 Liver	7
2.1.3 Kidney	8
2.1.4 Prostate	10
2.2 Medical Imaging	10
2.2.1 Computed Tomography	11
2.2.2 Magnetic Resonance Imaging	13
2.3 Deep Learning	16
2.3.1 Neural Network Structure	17
2.3.2 Training	21
2.3.3 Pipeline	23
2.3.4 Architectures	24
2.4 Medical Image Registration	26
2.4.1 Transformations	27
2.4.2 Conventional Methods	29
2.4.3 Deep Learning-Based Methods	30
2.4.4 Metrics	31
2.4.5 Registration Evaluation	33
3 Publication 1: Generalization of Affine Medical Image Registration Networks	35
3.1 Introduction	35
3.2 Background	36

3.3	Methods	37
3.3.1	Network Architectures	37
3.3.2	Benchmark Neural Network	41
3.3.3	Baseline Method	41
3.3.4	Training Setting	42
3.3.5	Datasets	43
3.3.6	Evaluation Metrics	43
3.4	Results	44
3.4.1	Synthetic Dataset	44
3.4.2	Patient Dataset	46
3.5	Discussion	48
3.5.1	Synthetic Dataset versus Patient Dataset	50
3.5.2	Individual Networks	51
3.5.3	Individual Metrics	52
3.6	Conclusion	53
3.7	Appendix	53
3.7.1	Example Results of an Affine Registration for All Implemented Networks	53
3.7.2	Discussion of the Individual Networks	62
3.7.3	Manual for the Published Source Code	66
4	Publication 2: Multistage Rigid-Affine-Deformable Network for Multimodal Medical Image Registration	67
4.1	Introduction	67
4.2	Background	68
4.3	Methods	69
4.3.1	Network Architectures	69
4.3.2	Benchmark Neural Network	71
4.3.3	Training Setting for Neural Networks	71
4.3.4	Baseline Method	72
4.3.5	Datasets	72
4.3.6	Evaluation Metrics	74
4.3.7	Experiments	75
4.4	Results	76
4.4.1	Experiment I	76
4.4.2	Experiment II	78
4.4.3	Experiment III	83
4.5	Discussion	83
4.5.1	Comparison to Published Multistage Registration Methods	83
4.5.2	Individual networks	85
4.5.3	Individual Datasets	86
4.5.4	Limitations	87
4.6	Conclusions	88
5	Publication 3: Multistep Networks for Deformable Multimodal Medical Image Registration	89
5.1	Introduction	89
5.2	Background	90

5.3	Methods	91
5.3.1	Network Architectures	91
5.3.2	Training Setting for the Neural Networks	93
5.3.3	Baseline Methods	94
5.3.4	Datasets	94
5.3.5	Evaluation Metrics	95
5.3.6	Experiments	96
5.4	Results	97
5.4.1	Experiment I	97
5.4.2	Experiment II	98
5.4.3	Experiment III	101
5.4.4	Experiment IV	103
5.4.5	Experiment V	105
5.5	Discussion	107
5.5.1	Comparison to Published Multistep-Multiresolution Registra- tion Methods	109
5.5.2	Individual Networks	109
5.5.3	Individual Datasets	111
5.5.4	Limitations	112
5.6	Conclusion	113
6	Publication 4: Groupwise Multiresolution Network for DCE-MRI Image Registration	115
6.1	Introduction	115
6.2	Methods	116
6.2.1	Network Architectures	116
6.2.2	Benchmark Networks	117
6.2.3	Training Setting for the Neural Networks	117
6.2.4	Baseline Methods	118
6.2.5	Dataset	118
6.2.6	Evaluation Metrics	119
6.2.7	Experiment	120
6.3	Results	121
6.4	Discussion	124
6.5	Conclusion	126
6.6	Supplementary Information	127
7	Summary of the Publications	129
8	Discussion and Conclusion	133
9	Bibliography	137
10	Publications	153
11	Curriculum Vitae	155
12	Acknowledgements	157

List of Figures

2.1	Hounsfield scale.	12
2.2	X, y and z-axis in MR scanner.	13
2.3	Zeeman levels for spin- $\frac{1}{2}$ system without and with magnetic field B_0 . .	14
2.4	Structure of a biological neuron and artificial neuron.	17
2.5	Structure of a ResNet residual block.	25
2.6	Schematic of an encoder-decoder architecture with latent space representation.	25
2.7	The schematic for conventional iterative image registration.	29
2.8	Visual methods for registration evaluation.	33
3.1	Categories of the implemented architectures.	38
3.2	Architecture of the benchmark CNN for affine registration.	41
3.3	Example results of an affine registration of two volumes from the synthetic dataset.	47
3.4	Example results of an affine registration of two volumes from the patient dataset.	49
3.5	Example results of an affine registration for all implemented networks of two volumes from the synthetic dataset.	54
3.6	Example results of an affine registration for all implemented networks of two volumes from the patient dataset.	58
4.1	Architecture of the rigid and affine subnetworks.	70
4.2	Architecture of the multistage network and the benchmark.	71
4.3	Example results of the registration of two volumes from the first NIDDK dataset.	79
4.4	Example results of the registration of two volumes from the second NIDDK dataset.	80

4.5	Example results of the registration of two volumes from the external kidney dataset.	81
4.6	Example results of the registration of two volumes from the external augmented kidney dataset.	82
4.7	Example results of the registration of two volumes from the M ² OLIE liver dataset.	84
5.1	Architecture of the multistep network variants A, B, C and the Benchmark.	92
5.2	Architecture of the deformable networks, used in the multistep network variants A, B, C and the benchmark.	92
5.3	Example results of the registration of an MRI to CT volume from the M ² OLIE liver dataset.	99
5.4	Example results of the registration of an MRI to CT volume from the synthetic liver dataset.	102
5.5	Example results of the registration of two volumes from the NIDDK dataset.	104
5.6	Example results of the registration of two volumes from the ProstateX dataset.	106
5.7	Example results of the registration of two volumes from the ProstateX dataset of experiment V.	108
6.1	Architecture of the groupwise multistep network.	117
6.2	ROI position and magnitude.	119
6.3	Example of a corrected liver dome ROI magnitude.	122
6.4	Example results of the registration of two volumes from the CDH dataset.	123
6.5	Example of perfusion maps of one representative patient.	127

List of Tables

3.1	Additional information on the implemented networks.	40
3.2	The NMI, Dice coefficient and Hausdorff distance results of the synthetic dataset and the real patient dataset.	45
3.3	Overview of the datasets used in the original papers.	50
4.1	Dataset statistics.	73
4.2	Hyperparameter setting for training the networks with the first NIDDK dataset.	75
4.3	Parameters for data augmentation performed on the external kidney dataset.	76
4.4	The Dice coefficient and Jacobian determinant results of the first NIDDK dataset.	77
4.5	P-values of the significance tests for the Dice coefficient and Jacobian determinant.	77
4.6	The Dice coefficient and Jacobian determinant results of the external kidney dataset without and with augmentation.	78
4.7	The Dice coefficient and Jacobian determinant results of the M ² OLIE liver dataset.	83
5.1	Hyperparameter setting for training the networks.	96
5.2	The NMI, Dice coefficient, Jacobian determinant and registration time results of the M ² OLIE liver dataset.	98
5.3	The NMI, Dice coefficient, Jacobian determinant and registration time results of the synthetic liver dataset.	100
5.4	The NMI, Dice coefficient, Jacobian determinant and registration time results of the NIDDK kidney dataset.	101
5.5	The NMI, SSIM, Jacobian determinant and registration time results of the the ProstateX dataset.	105
5.6	The NMI, SSIM, Jacobian determinant and registration time results of the ProstateX dataset using anti-aliasing filter.	107

5.7	P-values of the significance tests for the NMI and SSIM.	107
6.1	SSIM, registration time and SD liver dome ROI magnitude results of the CDH dataset.	121
6.2	Perfusion map results of the CDH dataset.	124

List of Acronyms

ADPKD	Autosomal Dominant Polycystic Kidney Disease
AIF	Arterial Input Function
ASL	Arterial Spin Labeling
BOLD	Blood Oxygenation Level Dependent
CBCT	Cone-Beam Computed Tomography
CDH	Congenital Diaphragmatic Hernia
CNN	Convolutional Neural Network
CPU	Central Processing Unit
CT	Computed Tomography
DCE	Dynamic Contrast-Enhanced
DWI	Diffusion Weighted Imaging
ESRD	End-stage Renal Disease
FCNN	Fully Convolutional Neural Network
GAN	Generative Adversarial Networks
GPU	Graphics Processing Unit
HD	Hausdorff Distance
HU	Hounsfield Unit
MI	Mutual Information
MLP	Multi-Layer-Perceptron
MR	Magnetic Resonance
MRI	Magnetic Resonance Imaging
MSE	Mean Squared Error
MTT	Mean Transit Time
NMI	Normalized Mutual Information
NMR	Nuclear Magnetic Resonance
PBF	Pulmonary Blood Flow

PBV	Pulmonary Blood Volume
PKD	Polycystic Kidney Disease
ReLU	Rectified Linear Unit
RF	Radio Frequency
RMSE	Root Mean Square Error
ROI	Region Of Interest
SD	Standard Deviation
SGD	Stochastic Gradient Descent
SSIM	Structural Similarity Index Measure
TKV	Total Kidney Volume
TRE	Target Registration Error

1. Introduction and Outline

1.1 Motivation

In the field of medical imaging, multimodal medical image registration is an important technique as it is fundamental for the integration of complementary information from multiple imaging modalities. Before fusing medical images, image registration is required, as deformations occur in the images, e.g. due to different patient positioning in different scanners or acquisitions, due to patient movement in between two sequences or respiration [1]. The integration of complementary information plays a central role in improving various clinical applications and provides the basis for more accurate diagnosis and improved treatment planning, delivery and outcomes for a wide range of diseases [2–4].

One area in which multimodal image registration is of central importance is the diagnosis and monitoring of kidney diseases, in particular [Autosomal Dominant Polycystic Kidney Disease \(ADPKD\)](#). ADPKD is a hereditary chronic disease characterised by progressive kidney enlargement, eventually leading to [End-stage Renal Disease \(ESRD\)](#). ADPKD requires early detection and continuous monitoring to avoid progression to ESRD. Multimodal image registration enables the fusion of data from different imaging modalities, improving the accuracy of ADPKD diagnosis and progression monitoring [5–7]. By integrating information from modalities such as T1- and T2-weighted [Magnetic Resonance \(MR\)](#) images, it is possible to derive important biomarkers such as [Total Kidney Volume \(TKV\)](#) and cyst volume, which are crucial for the assessment of disease progression [8–11].

Furthermore, medical image registration plays a crucial role in optimising treatment strategies for oligometastatic disease, where conventional therapies often deliver sub-optimal results due to lesion heterogeneity. Personalised, minimally invasive, image-guided therapies have demonstrated promise as they offer improved survival rates by addressing tumour heterogeneity. Techniques that use both [Computed Tomography \(CT\)](#) imaging for interventions and [Magnetic Resonance Imaging \(MRI\)](#) to characterise tumour heterogeneity have proven to be effective strategies [12–14]. At the research campus Mannheim Molecular Intervention Environment (M²OLIE) the use of multimodal imaging datasets for oncological interventions is shown, emphasising the crucial role of medical image registration to fuse data from different imaging modalities [1, 10, 11, 15].

Traditionally, conventional methods were used in medical image registration, in which the transformation parameters are iteratively adjusted to optimize the similarity metrics. However, the iterative approaches are time-consuming and require extensive computational resources due to repetitive iterations for each new image pair. To overcome this limitation, deep learning-based registration methods have emerged as a promising alternative. In contrast to conventional approaches, neural networks only require a single training phase. After training, the neural networks can efficiently register new images by processing them in a single forward pass through the network, resulting in significantly faster registration times [16]. A rapid registration then enables its use in time-critical applications, including interventions

such as image fusion for image-guided procedures. By providing fast registration, these techniques support real-time decision making and increase the accuracy of interventions, ultimately leading to better patient outcomes.

Innovations in medical image registration increase the efficiency and accuracy of image fusion, improving image analysis and patient care. Therefore, the aim of this thesis is to improve existing and develop new deep learning-based image registration techniques for three-dimensional multimodal medical images. I focused on lung and abdominal CT and MR data and unsupervised or semi-supervised neural network training. First, I implemented 20 different published neural networks for affine medical image registration. The performance of the networks and their generalizability to new datasets were evaluated on two multimodal liver datasets and compared with my own developed Convolutional Neural Network (CNN) as a benchmark. Subsequently, I developed a multistage network, which includes rigid, affine and deformable transformations. I then developed deformable multiresolution registration networks and investigated the influence of multiple steps and resolutions in a deformable neural network on the registration result of four multimodal datasets. Additionally, I developed a groupwise multiresolution network for the registration of four-dimensional Dynamic Contrast-Enhanced (DCE)-MR images.

1.2 Outline

This thesis is written cumulatively. Chapter 3 to Chapter 6 each represent a self-contained scientific study, each with an introduction, methods, results, discussion and conclusion.

Chapter 2 provides the theoretical background for this thesis. First, the medical background of the organs and diseases included in the studies in this thesis is briefly described. Subsequently, an overview of the medical imaging methods CT and MRI is presented. This is followed by the basics of deep learning and an introduction into medical image registration.

In Chapter 3 20 different published neural networks were implemented for affine registration of medical images. The performance of the networks and their generalizability to new datasets were evaluated using two multimodal datasets - one synthetic and one patient dataset - of three-dimensional CT and MR images of the liver. The results were compared with an in-house developed CNN as a benchmark.

Chapter 4 presents a neural network that incorporates rigid, affine, and deformable transformations in three subnetworks. This addresses the challenge of larger rigid deformations that are common in medical images due to variations in patient positioning in different scanners and rigid anatomical structures. By using an affine-deformable network as a benchmark, the impact of a preceding rigid subnetwork on the registration result of four multimodal medical datasets was investigated.

In Chapter 5 networks for multimodal medical image registration were developed, that apply multiple steps and resolutions, comparable to the coarse-to-fine approach of classical registration methods like SimpleElastix [17] and NiftyReg [18, 19] or published multistep neural networks. By comparing the multistep-multiresolution

networks with up to four resolutions/steps with each other and with a monostep-monoresolution network as a benchmark and the classical registration methods SimpleElastix and NiftyReg as a baseline, the effects of using multiple resolutions on the registration result of four multimodal medical datasets were investigated.

In [Chapter 6](#) a novel groupwise registration network was presented for the registration of four-dimensional DCE-MR images. As underlying network architecture the multistep-multiresolution network was applied, which is introduced in [Chapter 5](#). The groupwise network was developed with two different group sizes and its performance was compared with a pairwise and groupwise registration using the classical registration method SimpleElastix as baselines and published pairwise and groupwise registration networks as benchmarks.

[Chapter 7](#) presents a summary of the thesis and the research presented in [Chapter 3](#) to [Chapter 6](#). And in [Chapter 8](#) the relevance of the studies presented and future research perspectives are discussed.

1.3 Citation of Previous Publications

Several chapters of this dissertation have already been published or are currently submitted for publication. The citations for these chapters are:

Chapter 3: A. Strittmatter, L. R. Schad, F. G. Zöllner. Deep learning-based affine medical image registration for multimodal minimal-invasive image-guided interventions – A comparative study on generalizability. *Zeitschrift für Medizinische Physik*, 34 (2), pp. 291-317, doi: 10.1016/j.zemedi.2023.05.003, 2024.

Chapter 4: A. Strittmatter, A. Caroli, F. G. Zöllner. A Multistage Rigid-Affine-Deformable Network for Three-Dimensional Multimodal Medical Image Registration. *Applied Sciences*, 13(24), 13298, doi: 10.3390/app132413298, 2023.

Chapter 5: A. Strittmatter, F. G. Zöllner. Multistep networks for deformable multimodal medical image registration. *IEEE Access* 12, 82676–82692, doi: 10.1109/ACCESS.2024.3412216, 2024.

Chapter 6: A. Strittmatter, M. Weis, F. G. Zöllner. A Groupwise Multiresolution Network for DCE-MRI Image Registration. *Scientific Reports*. Submitted on 12.09.2024.

2. Background

The following chapter provides the theoretical background for this work. In this thesis, medical image registration methods are developed and evaluated using different [Computed Tomography \(CT\)](#) and [Magnetic Resonance \(MR\)](#) images of lung and abdominal imaging. First, the medical background of the organs included in the studies in this thesis is briefly described. Then, an introduction into the fundamentals of the medical imaging techniques [CT](#) and [Magnetic Resonance Imaging \(MRI\)](#) is given. Subsequently, the basics of deep learning are explained. This is followed by an introduction to medical image registration.

2.1 Medical Background

In this thesis, [CT](#) and [MR](#) images of the lung, liver, kidney and prostate are used. The anatomy and function of these organs and the diseases referred to in the studies in this thesis are briefly explained in the following sections.

2.1.1 Lung

Anatomy

The lungs are the central respiratory organs and are characterised by their paired, cone-shaped structure. The lungs are situated in the thoracic cavity and the mediastinum is located between the two lungs.

The thoracic cavity contains a two-layered serous membrane called pleura. The lungs are covered with the visceral pleura. The parietal pleura lies on the inside of the thoracic wall. The narrow gap, pleural cavity, between the visceral pleura and the parietal pleura is filled with fluid and allows the lungs to move in their thoracic cavity during breathing according to the respiratory muscles and to slide smoothly along the inside of the thoracic wall.

The lungs are divided into several lobes, which are defined by branches of the airways in the lungs, the bronchi. The left lung comprises an upper and a lower lobe and consists of eight sublobular segments. The right lung, on the other hand, contains an upper, middle and lower lobe and consists of nine sublobular segments [20].

There are approximately 300 to 500 million alveoli in the human lungs. These are air sacs, which are located at the end of the bronchioles and are the site of gas exchange. They form a surface measuring 70 to 100 m^2 . The alveoli have thin walls, which therefore only form a small hurdle for gas exchange. Here, oxygen and carbon dioxide are exchanged between the air of the environment and the blood of the lung capillaries [21].

The lung blood system is divided into two different circulatory pathways, the bronchial and the pulmonary circulation. In the bronchial circulation, the lung tissue is supplied with oxygen-rich blood. The pulmonary circulation is used for gas exchange in which deoxygenated blood is transported from the right ventricle to the alveoli and oxygenated blood is transported to the left atrium of the heart [21].

Function

As the main respiratory organs, the lungs have a key role in absorbing oxygen from the air of the environment into the blood and releasing carbon dioxide, which is produced during cellular metabolism, from the blood of the lung capillaries [21].

Physiology of Respiration

During respiration, air passes through the nose or mouth to the throat, then into the trachea and subsequently to the main bronchi of the lungs. The main bronchi branch into two main stems, one located in the right lung and the second one in the left lung. Within the lungs, the main bronchi form a bronchial tree by branching into smaller bronchi and subsequently into bronchioles [20]. The bronchioles end in the alveoli, where gas is exchanged by diffusion [21]. Due to a concentration gradient of the respective gas, oxygen passes through the alveolar walls into the blood, while carbon dioxide diffuses from the blood into the alveoli. Diffusion through the alveolar walls is characterised by Fick's law of diffusion:

$$\frac{dQ}{dt} = Ad \frac{\partial C}{\partial x}, \quad (2.1)$$

The volume transfer of the gas in x direction $\frac{dQ}{dt}$ is dependent on the area of the alveolar wall A, the diffusion coefficient d, and the concentration gradient of the respective gas in x direction $\frac{\partial C}{\partial x}$ [22].

To inhale and exhale air, a difference in pressure is necessary between the surrounding atmosphere and the lung alveoli. To reduce the alveolar pressure, the thoracic volume is increased [23]. The lungs cannot expand on their own; instead, the respiratory muscles, such as the diaphragm and the external intercostal muscles, are responsible for inflating the lungs. Contraction of the respiratory muscles creates negative pressure in the pleural cavity. The external intercostal muscles push the ribs forwards and outwards and the diaphragm moves down into the abdominal cavity, which leads to an increase in thoracic volume [21].

During calm breathing, exhalation occurs passively by relaxing the respiratory muscles, whereas it is actively controlled during intense physical activity or speaking [23, 24]. The respiratory muscles, primarily the abdominal muscles and the inner intercostal muscles, pull the ribs in and down, which reduces the volume of the chest cavity and pushes the air out of the lungs [23]. During exhalation, the lungs do not expel the entire air, instead a residual volume remains, which guarantees continuous gas exchange [21].

The variation in the volume of the thoracic cavity during the respiratory cycle causes movement and deformation of the nearby structures, particularly in the mediastinum and abdomen. The amount of movement and deformation of each organ during the respiratory cycle varies depending on a number of factors including the individual's anatomy and lung capacity. The diaphragm rises or falls about 1 cm within the abdominal cavity when breathing calmly, and can rise or fall up to 10 cm when breathing deeply [24]. Rigid structures such as bones or rigid tumours are mostly moved rigidly (translation and rotation). Flexible structures also experience elastic deformation.

Congenital Diaphragmatic Hernia

Congenital Diaphragmatic Hernia (CDH) is a rare condition that occurs in around 0.8 to 5 cases per 10,000 births [25]. In CDH, the diaphragm does not close completely during foetal development and consequently the foetal abdominal organs are displaced (herniated) into the thoracic cavity [26]. Typically, the opening in the diaphragm occurs on the left side posterolaterally [27]. As the foetal abdominal organs occupy part of the thoracic cavity, the available space for the developing lungs is greatly reduced. This leads to impaired lung development, which can result in pulmonary hypoplasia. In addition, pulmonary hypertension can occur due to cardiac dysfunction [26]. Around a third of affected newborns also show cardiovascular malformations, while other cases may be associated with, e.g., nerve or intestinal anomalies [27]. The extent of cardiovascular and respiratory dysfunction is a major factor responsible for the high incidence of mortality and morbidity in CDH [26].

Prenatal diagnosis of diaphragmatic hernia plays a crucial role and allows early planning for prenatal interventions. Historically, hernia correction was performed through open surgery, whereas today, fetoscopic endotracheal occlusion is preferred [27]. After birth, measures such as intestinal decompression and, if necessary, insertion of an endotracheal tube are taken immediately. Postnatal treatment involves ventilation, monitoring of cardiovascular status, addressing pulmonary hypertension and subsequent surgical intervention. Prenatal ultrasound imaging is used for diagnosis and prognosis by evaluating the lung-to-head ratio and the liver's position [25]. In addition, foetal MRI and echocardiography are usually used to examine the anatomy [26]. Moreover, **Dynamic Contrast-Enhanced (DCE)-MRI** can be utilized to determine lung perfusion [28]. Multiple medical imaging scans can be fused in order to combine the information from different scans or to monitor the progression of the disease. This requires image registration due to the patient's breathing between multiple scans or the different positioning of the patient during multiple acquisitions.

2.1.2 Liver

Anatomy

At 1.3 to 1.7 kg, the liver is the heaviest internal organ and the largest gland in the human body [29]. The liver is located in the right upper abdomen intraperitoneally, below the diaphragm and is supported by the lower ribs. The stomach lies dorsally and the gallbladder is located dorsally between the left and right lobes of the liver. In classical anatomy, the liver is divided into two large lobes - the right (larger) and the left lobe (*lobus hepatis dexter* and *sinister*). The lobes are separated from each other by ligaments. Additionally, models exist for functional segmentation of the liver. The branches of the hepatic artery *A. hepatica propria*, the bile ducts and the portal vein subdivide the liver into the traditional eight-segment scheme [30] or the 1-2-20 concept which is important for liver surgery and transplantation [31].

The liver is composed of about 1-1.5 million functional units called hepatic lobules [30]. The liver is parenchymatous and about 60% of its cells are hepatocytes [29]. The liver is supplied with blood by two vessels: the hepatic artery (*A. hepatica propria*) and the portal vein (*V. portae hepatis*), which supply oxygen and nutrient-rich blood respectively [32].

Function

As the body's central metabolic organ, the liver is involved in carbohydrate, fat and protein metabolism. During protein synthesis, coagulation factors are produced [33]. The liver produces bile for the digestion of fat, which is then delivered to the duodenum via the bile duct.

The liver also has a detoxification function, where harmful substances, drugs and alcohol are converted through chemical modification (biotransformation) or broken down [32]. In addition, various substances such as vitamins and minerals are stored in the liver. And as an endocrine organ, the liver also plays a central role in the activation, synthesis and breakdown of hormones [33].

Oligometastatic Disease

Liver metastases are metastases of tumor cell clusters in the liver. Oligometastatic disease is defined as an intermediate state between localised tumor and disseminated metastatic disease [34]. The diagnosis is primarily made by conventional abdominal sonography, although sensitivity can be increased by using contrast-enhanced sonography or computed tomography.

Liver metastases are treated surgically, preferably with R0 resection, or with chemo-, thermo- or radioablative interventional procedures [35]. However, patients with oligometastatic disease have a low likelihood of survival when standard treatment options are used due to the heterogeneity of local lesions. The genetic and molecular characteristics of cancer cells can differ within a tumor or across various metastatic sites, which leads to some cells showing resistance to particular treatments while others are more susceptible. This may result in the tumor not shrinking completely and potentially developing resistance to treatment. However, the use of tailored techniques, e.g. minimally invasive image-guided interventions that take tumor heterogeneity into account, significantly increases the survival rate [12, 13]. These interventions are usually performed using CT, whereas tumor heterogeneity is determined using MRI [14]. To support oncological interventions and tracking of treatment progress, the images obtained with the various modalities are merged into a multimodal dataset that collects all existing morphological and functional tumor information [1, 10, 11, 15].

Variations in patient positioning between the two scanners, along with the patient's breathing lead to varying deformation of the organs in the corresponding CT and MR images. Therefore, prior to fusion, it is essential to register the CT and MR images.

2.1.3 Kidney

Anatomy

The kidneys are paired organs located along the spine at the level of the eleventh and twelfth ribs. They are bean-shaped and have a reddish-brown colour. The kidneys lie retroperitoneally on the posterior abdominal wall [36]. Due to the liver's position superior to the right kidney, the right kidney is located about 1.5 centimetres lower than the left kidney. Each kidney weighs around 150 grams [37].

A human kidney consists of approximately 1 million functional units, called nephrons, which form the renal parenchyma. Each nephron comprises a glomerulus and a tubule [38].

The kidney is divided into two main anatomical regions, the medulla in the centre and the renal cortex surrounding it, which comprises an outer cortex and a juxtamedullary cortex [36]. The medulla consists of 7-9 cone-shaped segments, called pyramids, whose tips are named papillae [38]. Renal arteries and veins, renal nerves and the ureter, which transports urine from the kidneys to the urinary bladder, enter and/or exit the kidneys at the so-called renal hilum.

The kidneys are supplied with arterial blood by the renal arteries, which branch off the abdominal aorta, and the venous blood is then transported by the renal veins. About 20% of the cardiac output is utilised for the blood supply of the kidneys, whereby 90% is supplied to the renal cortex and 10% to the renal medulla [36].

Function

The kidneys play a central role in the regulation of fluid and electrolyte balance as well as in the removal of metabolic waste products and foreign substances from the blood to detoxify the body. They also produce hormones and enzymes that influence various physiological processes, such as erythropoiesis and the regulation of blood pressure and the acid-base balance [38].

Primary blood filtration is performed in the renal glomeruli, forming primary urine (about 180 litres per day) and removing harmful substances. This urine is then predominantly reabsorbed in the renal tubules, whereby e.g. important electrolytes are recovered. The resulting urine volume is approximately 1.5-2 litres per day [37].

Autosomal Dominant Polycystic Kidney Disease

Polycystic Kidney Disease (*PKD*) is a genetic disease, which leads to the development of fluid-filled cysts in the kidney and can result in an enlargement of the kidney and ultimately to End-stage Renal Disease (*ESRD*) and chronic kidney failure [39]. This disease can be inherited autosomal recessive or autosomal dominant (Autosomal Dominant Polycystic Kidney Disease (*ADPKD*)). Chronic kidney failure leads to an increase in the concentration of urinary substances in the blood, as the kidneys no longer excrete urinary substances sufficiently. In this case, dialysis and kidney transplantation are crucial treatment methods [36].

Early detection and regular monitoring of *ADPKD* play a crucial role in preventing or delaying the development of *ESRD*. Renal morphology is analysed using various imaging techniques such as ultrasound, *CT*, *MRI* and renal scintigraphy. The integration of medical image registration allows data from different imaging modalities to be combined, improving the accuracy of *ADPKD* diagnosis and monitoring of *ADPKD* progression. Thereby, Total Kidney Volume (*TKV*) and the volume and number of cysts serve as important indicators and key biomarkers for assessing the progression of *ADPKD* [8]. In particular, T1-weighted *MR* images are utilised for renal segmentation to determine the *TKV*, whereas T2-weighted *MR* images can be utilised for cyst segmentation to assess the cyst volume and number. By fusing *MR*

modalities, complementary information is merged, which significantly enhances the assessment of the ADPKD biomarkers [9]. In addition, the fusion of images from various time points, e.g. pre-, mid- and post-therapy, allows monitoring of therapy progression. This requires image registration [10, 11].

2.1.4 Prostate

Anatomy

The prostate is a walnut-sized fibromuscular gland, enclosing the prostatic urethra and is located in between the bladder and the urogenital diaphragm. A fibromuscular tissue called the prostatic capsule covers the prostate [40]. The human prostate weighs approximately 15-20 g [41], but can reach up to more than 150 g, e.g. in the case of benign prostatic hyperplasia [21].

Anatomically, the prostate was historically divided into five lobes: Anterior, middle/median, posterior, and two lateral (left and right) lobes. However, the anatomical subdivision into five lobes can only be detected in the foetus before the 20th week of gestation. Afterwards, only the lateral lobes and middle lobe are identifiable. Histologically, the glandular tissue can be divided into three zones, the peripheral, central and transition zones. The peripheral zone makes up around 70% of the total prostate volume [21]. Approximately 70-80% of all prostate cancers occur in this zone [42]. The central zone accounts for about 25% of the prostate volume and encloses the two ejaculatory ducts, that convey the seminal fluid. The remaining 5% of the prostate volume is called the transition zone [21], where approximately 20-30% of all prostate cancers occur [42].

The prostate is mainly supplied with arterial blood by the inferior vesicular artery and additionally by the internal pudendal artery and the middle rectal artery [21], while the prostatic veins comprise the prostatic venous plexus [40].

Function

The prostate is an essential part of the male reproductive system. Its main task is to produce a portion of the seminal fluid that is essential for the transport of sperm. This fluid is released during ejaculation, contains e.g. enzymes, citric acid, acid phosphatase and prostate-specific antigen, and enhances sperm motility and protection to prevent damage [40].

2.2 Medical Imaging

Medical imaging comprises various techniques that visualize the body's anatomy. This is applied for diagnosis, therapy and monitoring of disease progression, as well as evaluation of treatment effectiveness. Different imaging modalities provide complementary information, morphologically as well as functionally, therefore a combination of different imaging techniques is often used, which is called multimodal imaging [1]. In the following sections, the medical imaging techniques CT and MRI are described in more detail, as in the studies of this thesis CT and MR images were used to evaluate the registration methods.

2.2.1 Computed Tomography

CT utilises the principle of attenuation of electromagnetic waves (X-rays) to generate digital volumetric images of the human body. The name computed tomography is derived from the Greek word tomos, meaning section. In CT, digital axially sliced three-dimensional images are created. The z-axis follows the scanning system's axis of rotation, which is roughly parallel to the longitudinal axis of the patient [43].

The CT scanner consists of a scanning unit, also known as a gantry, and the movable patient table. Inside the gantry there are the X-ray tube and the detectors. The X-ray tube contains a cathode and an anode. The cathode is heated to about 2400 K. Electrons from the cathode are then accelerated towards the anode in the electric field present between the two electrodes. For this, an acceleration voltage is applied, which ranges between 25 kV and 150 kV for medical diagnostics and between 10 kV and 300 kV for radiation therapy.

When the electrons hit the anode, they are decelerated. During this process, The majority (approximately 99 %) of the kinetic energy is converted into heat. The remaining 1 % is converted into radiation, so-called Bremsstrahlung. As the deceleration of the electrons at the anode is a deceleration cascade, this produces a continuous Bremsstrahlung spectrum.

The maximum energy of the X-rays E_{max} is calculated using the acceleration voltage U_a , the electron velocity v and the physical constants elementary charge e and Planck's constant h :

$$E_{max} = eU_a = hv_{max}. \quad (2.2)$$

In addition to the continuous Bremsstrahlung, the X-ray spectrum also contains discrete lines. This is characteristic radiation. If a free electron pushes a bound electron out of an inner shell of an atom, the vacancy created on the shell is filled by an electron located on an outer shell. This releases energy in the form of radiation, which corresponds to the energy difference between the two shells of the atom. The characteristic radiation is specific to the anode material used.

The low-energy X-rays are filtered out by metal filters, e.g. aluminium, before they reach the patient, as these do not contribute to imaging but are purely absorbed by the patient's body. Furthermore, a collimator is used to create a beam shape of the X-ray radiation [44].

If the radiation hits a sample, the radiation is attenuated, i.e. the intensity of the radiation is reduced. The intensity and its attenuation can be calculated using the Lambert-Beer law. The intensity at distance d for a monoenergetic X-ray beam and an inhomogeneous sample is given by:

$$I(d) = I(0)e^{-\int_0^d \mu(s)ds} \quad (2.3)$$

with the intensity of the incident X-ray $I(0)$, the attenuation coefficient $\mu(s)$ and the X-ray beam path s [45].

In addition to the intensity attenuation, other tissue interaction effects occur. These include Rayleigh scattering, where no energy is transferred. Additionally, Compton scattering is possible. Here a free photon hits a loosely bound electron. The free photon changes its direction and transfers energy to the second electron (compton electron). This is ejected from the atomic shell. And during the photoelectric effect, a free photon is absorbed. This releases all its energy to an electron, which is ejected from the atomic shell and travels through the material as a free photoelectron [44].

To filter the scattered X-rays, a collimator is positioned in front of the detectors. The X-rays then hit the detector element arrays. A scintillation crystal absorbs these rays and produces photons in the visible spectrum. A photodetector then converts these photons into an electrical signal.

To create a volumetric three-dimensional image of the patient's body, the X-ray tube and detector array are rotated around the patient. In addition, the patient table is moved to acquire a new slice. The three-dimensional CT scan is then calculated from multiple two-dimensional X-ray projections from different angles. For this, a reconstruction algorithm is applied. E.g. back projection, in which the projections are smeared back linearly for each angle. Alternatively, a filtered back projection or an iterative reconstruction can also be utilised [46]. The voxel size of the three-dimensional volume is determined by the in-plane resolution and the slice thickness [43].

The resulting images contain grey values. These CT values are measured in Hounsfield Units (HUs). These values indicate the density of the tissue relative to water:

$$CTvalue = \frac{\mu - \mu_{Water}}{\mu_{Water} - \mu_{Air}} \cdot 1000HU. \quad (2.4)$$

Each tissue is characterised by a specific tissue HU interval, which can be seen in Figure 2.1. Water has a CT value of 0 HU, air of -1000 HU. Fat has negative CT values as it has a lower density and therefore a lower attenuation than water. And soft tissue has positive CT values. Bones have an increased density and attenuation and higher CT values up to 2000 HU.

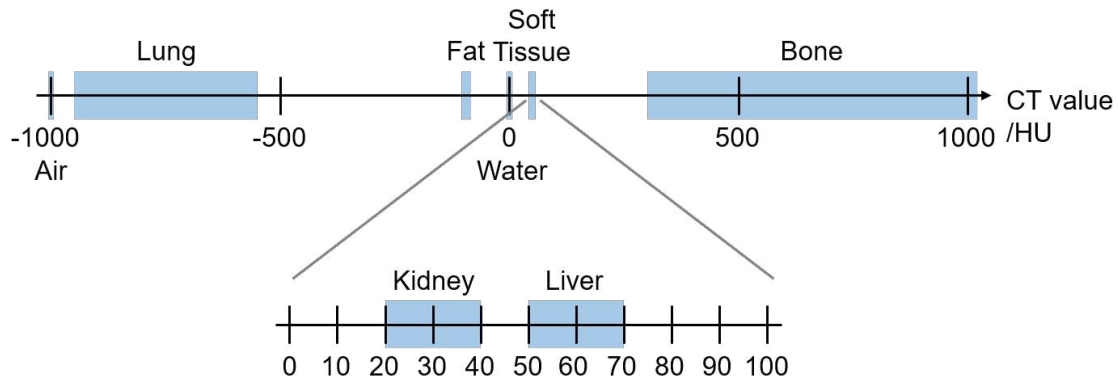


Figure 2.1: Hounsfield scale. Tissue intervals are displayed in blue.

One drawback of CT is that it has a low soft tissue contrast. However, it provides high contrast for structures with high density. This is used e.g. in CT angiography.

To visualise the blood vessels, the patient is injected with a contrast agent that strongly absorbs X-rays [1]. Another drawback of CT is that the patient is subjected to ionising radiation. This can potentially harm the patient's DNA and therefore adequate indication and dose modulation is important to minimise the risks for the patient [47].

2.2.2 Magnetic Resonance Imaging

In contrast to CT, MRI does not utilize ionizing radiation to generate digital volumetric images of the human body. Instead, this non-invasive technique is based on the principle of Nuclear Magnetic Resonance (NMR).

In principle, all nuclei that have a nuclear spin angular momentum, also known as spin, can be used. Since the human body consists to a great extent of water, ^1H nuclei hydrogen (single protons) are used for imaging in conventional MR imaging. ^1H atoms are characterized by a spin of $\frac{1}{2}$, hence it is referred to as a spin- $\frac{1}{2}$ system. The magnetic quantum number can be $m_I = \frac{1}{2}$ or $m_I = -\frac{1}{2}$ for spin- $\frac{1}{2}$ systems. This spin generates a small magnetic field around the nuclei with a small magnetic moment.

If an external strong magnetic field B_0 is applied, the magnetic moment precesses around the magnetic field B_0 with the frequency ω_0 , called Larmor frequency. In MRI, the magnetic field B_0 is applied along the z-direction (Figure 2.2).

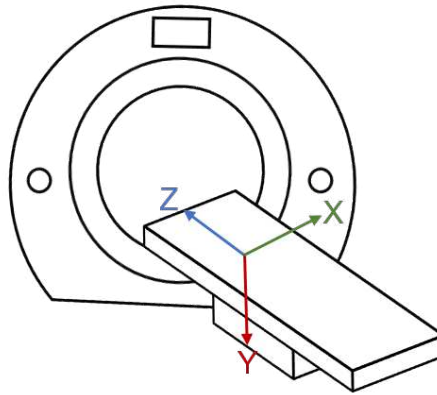


Figure 2.2: X, y and z-axis in MR scanner.

The Larmor frequency depends on the gyromagnetic ratio γ , which is specific for each nucleus, and magnetic field B_0 :

$$\omega_0 = \gamma B_0. \quad (2.5)$$

Spins in a magnetic field B_0 can assume either the spin-up or spin-down state in a spin- $\frac{1}{2}$ system, i.e. either parallel or anti-parallel to B_0 . The spin-up state is the lower energy state compared to the spin-down state. Figure 2.3 shows the splitting of the nuclear energy levels, called Zeemann effect, when an external strong magnetic field B_0 is applied to a spin- $\frac{1}{2}$ system.

The energy difference between the two energy states can be calculated by:

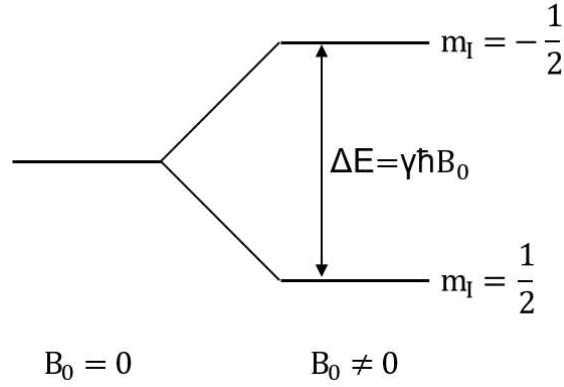


Figure 2.3: Zeeman levels for spin- $\frac{1}{2}$ system without and with magnetic field B_0 .

$$\Delta E = E_{\downarrow} - E_{\uparrow} = \frac{1}{2}\gamma\hbar B_0 - \left(-\frac{1}{2}\gamma\hbar B_0\right) = \gamma\hbar B_0 \quad (2.6)$$

with the spin-up energy $E_{\uparrow}$, the spin-down energy $E_{\downarrow}$ and the reduced Planck's constant $\hbar$.

Due to the energy difference between the two spin states, the spin population in the two spin states is not the same. The ratio of the number of spins in spin-up state $N_{\uparrow}$ to spins in spin-down state $N_{\downarrow}$ is described by the Boltzmann distribution:

$$\frac{N_{\uparrow}}{N_{\downarrow}} = e^{\frac{\Delta E}{kT}} \quad (2.7)$$

with the Boltzmann constant k and the absolute temperature T .

This results in a macroscopic magnetic moment $M_0 > 0$. With a magnetic field strength of $B_0 = 1$ T, this yields a net magnetization which is as large as the magnetic field strength generated by three nuclei. However, since water is a substantial component of the human body, there are about 10^{27} hydrogen nuclei in the human body. For this reason, a macroscopic magnetization is measurable [45].

In addition to the static magnetic field B_0 , an oscillating magnetic field B_1 is applied. This is created by a [Radio Frequency \(RF\)](#) pulse, called an excitation pulse. If the frequency of the [RF](#) pulse corresponds to the Larmor frequency, the precession axis of the protons is altered. By specifying the duration of the [RF](#) pulse, the angle at which the precession axis turns is set. Thus, in addition to the lateral magnetization M_z along B_0 (z-axis), a transverse magnetization $M_{xy} > 0$ is created. By using a so-called 90° pulse, the precession axis is rotated by 90° , i.e. the net magnetization is completely shifted to the xy-plane [47]. The magnetization then returns to the equilibrium state, i.e. $M_{xy} = 0$ and $M_z = M_0$. This process is called relaxation. How quickly the precession axis aligns itself with B_0 is specific to different tissue types.

Two different relaxation times are distinguished. The longitudinal relaxation/spin-lattice relaxation is characterized by the energy exchange between the spin system

and the atomic lattice in its environment. For $t \rightarrow \infty$ the lateral magnetization $M_z \rightarrow M_0$. The longitudinal relaxation, i.e. the formation of the lateral magnetization M_z , is described by:

$$M_z(t) = M_0 - (M_0 - M_z(0)) e^{-\frac{t}{T_1}}. \quad (2.8)$$

The T1 time is the time after the RF pulse that is required for M_z to reach $1 - \frac{1}{e}$ (approx. 63%) of the magnetization M_0 .

The transversal relaxation/spin-spin relaxation is caused by interactions between neighboring atoms. For $t \rightarrow \infty$ the transverse magnetization $M_{xy} \rightarrow 0$. The transversal relaxation, i.e. the decrease of the transversal magnetization M_{xy} is described by:

$$M_{x,y}(t) = M_{x,y}(0) e^{-\frac{t}{T_2}}. \quad (2.9)$$

The T2 time is the time after the RF pulse that it takes for the transverse magnetization M_{xy} to drop to $\frac{1}{e}$ (approx. 37%) of magnetization M_0 [45]. In general $T_2 \leq T_1$. However, due to inhomogeneities in the local magnetic field, the nuclei possess slightly different Larmor frequencies (see Equation 2.5). The magnetization M_{xy} therefore decreases quicker than in theory with a homogeneous magnetic field. Instead of the T2 time, a so-called effective transversal relaxation time T2* is measured [47]. Since T1 and T2 are specific for different tissue types, they can be used to provide a soft tissue contrast for medical imaging.

The transversal magnetization M_{xy} is measured by receiver coils. The signals are then displayed on the computer. Spatial encoding is necessary to generate a spatially resolved image. This means that the signals measured can be mapped to a location in the room. For this purpose, gradient fields are used in the x-, y- and z- directions. These are overlaid onto the static magnetic field B_0 . Three methods are used here: slice selection, frequency encoding and phase encoding.

During slice selection, a gradient G_z is applied in the z-direction. Due to the additional magnetic field strength, the Larmor frequency changes; in contrast to Equation 2.5, it now depends on the position along the z-axis:

$$\omega(z) = \omega_0 + \gamma G_z z. \quad (2.10)$$

Depending on the frequency of the RF pulse, a different slice can be selected, as for excitation the frequency of the RF pulse has to match the Larmor frequency. The slice thickness is determined by the bandwidth of the RF pulse.

For frequency encoding, a gradient G_x is applied in the x-direction during readout, thus generating a Larmor frequency, which is a function dependent on the position along the x-axis. In phase encoding, a gradient G_y is applied in the y-direction, which generates a phase shift along the y-axis [45]. The spatial position of the signals can then be determined using a computer.

The signals recorded are plotted in Fourier space, also known as k-space. The low frequencies are located in the center of the k-space; these are responsible for image

contrast and coarse shapes. The higher frequencies are located in the periphery of the k-space and contribute to finer image details. To generate the MR image, an inverse Fourier transformation is then applied.

By using different MR sequences, i.e. pulse sequences, a wide range of contrasts can be created. MRI is a qualitative method. Currently, research is being conducted on quantitative MR methods. In addition, instead of ^1H , other atoms can also be measured; in principle, all atoms that have a spin $\neq 0$ can be used. For instance, ongoing research is investigating the measurement of ^{23}Na (spin- $\frac{3}{2}$ system) in the body for a more precise examination of tissue health, e.g. for the early diagnosis of a stroke [48].

Nowadays, multiparametric MRI is used in medical imaging. In addition to recording morphology (conventional MRI), this also includes the functional examination, for instance diffusion (Diffusion Weighted Imaging (*DWI*)), perfusion (e.g. Arterial Spin Labeling (*ASL*)), oxygenation (Blood Oxygenation Level Dependent (*BOLD*)) and arterial blood flow. The tissue relaxation properties are determined by T1 and T2 mapping.

2.3 Deep Learning

Deep learning is a subcategory within the domain of machine learning, which is itself a subcategory of the broader field of artificial intelligence. It is widely applied in medical image processing, e.g. for image reconstruction, classification, segmentation and registration.

The foundation for deep learning are artificial neurons. These were introduced in 1943 by Warren McCulloch and Walter Pitts [49] and are inspired by cells of the biological nervous system, called neurons [50].

In biological nervous systems, neurons are interconnected. A neuron is depicted in Figure 2.4a. A neuron receives information from the preceding neurons via its dendrites through synapses. At the soma (cell body), these signals are weighted and accumulated. These signals can be activating or inhibiting. If this accumulated signal exceeds a certain threshold value, an electrical pulse, called an action potential, is emitted by the neuron. This follows the all-or-nothing principle. Low-threshold signals have no effect and therefore do not trigger an action potential [51]. The action potential is transmitted along the axon and is passed on to the following nerve cells or muscle fibers via the axon terminal [52]. In the synaptic cleft, the signal is modified by various chemical processes and passed to the following cell. The signal is non-linear [51].

The artificial neuron (Figure 2.4b) is structured in a similar way to a biological neuron. The equivalent to the dendrites is a vectorial input $\vec{x}$ with components x_i . The inputs are then weighted, i.e. multiplied by the weight w_i and accumulated with a propagation function, usually using a weighted sum:

$$\sum_{i=1}^m w_i x_i. \quad (2.11)$$

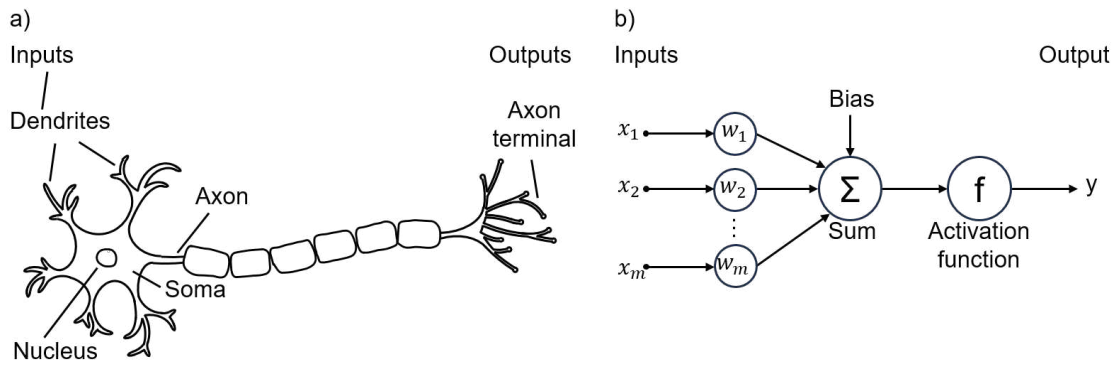


Figure 2.4: Structure of a biological neuron (A) and artificial neuron (B).

Non-linearity is obtained by using an activation function f . The scalar output y can then be calculated by:

$$y = f\left(\sum_{i=1}^m w_i x_i\right). \quad (2.12)$$

Similar to a biological neuron, a threshold value is also applied to an artificial neuron. If the accumulated signal is below this threshold, the output f of the neuron is zero. An additional bias neuron is applied to set the threshold value. In this way, the threshold value is included in the propagation function and not in the activation function itself [51].

Comparable to a biological nervous system, multiple artificial neurons can be interconnected. The perceptron presented by [53] comprises two layers of neurons. These are the input and output layers [52]. Several perceptrons can be connected to form a **Multi-Layer-Perceptron (MLP)**. There are three different layers in a multi-layer neural network. The input layer contains the input data, one or more hidden processing layers and an output layer that includes the output of the network.

In a simple network structure, known as a feed-forward network, each neuron is only connected to neurons in the previous and subsequent layers. The information flow is therefore uni-directional.

Furthermore, shortcut connections can be applied in neural networks, in which one or multiple layers are skipped when connecting neurons [51].

2.3.1 Neural Network Structure

In neural networks, a variety of components and layers are used. The most widely used components in the field of medical image processing are described below.

Fully Connected Layer

A fully connected layer is a layer in which every neuron is connected to every neuron in the previous layer. In contrast, in a sparsely connected layer, the neurons are only connected to some of the neurons in the previous layer.

Convolutional Layer

In the field of image processing, convolutional layers are widely applied as they are specifically developed for data arranged on a cartesian grid, such as two-dimensional images or three-dimensional volumes. Processing image data with convolutional layers is not as computationally expensive and time consuming compared to fully connected layers, as not every voxel is processed separately. Instead, convolutional layers use convolutions with filters. Therefore, the spatial information of surrounding voxels is also taken into account. Using convolutional layers therefore results in a certain degree of translation invariance [50]. A filter kernel is a matrix of weights. This filter is then moved step by step over the entire input of the convolutional layer. At each point, multiplication of the kernel weights with the values of the input is performed element by element. Moving the filter over the entire input creates an output matrix, called a feature map. The two-dimensional convolution of a two-dimensional image f and a filter kernel w is calculated by:

$$h(x,y) = f(x,y) * w(x,y) = \sum_n \sum_m f(n,m)w(x-n,y-m), \quad (2.13)$$

where x and y are the indices for the two-dimensional image f and m and n are determined by the filter kernel size [54].

For two-dimensional inputs, the height and width of the feature map are determined by the height and width of the input, as well as padding and the step size with which the filter of the convolutional layer is moved over the input. This step size is called stride. The stride can therefore be used to reduce the size of the feature map.

For two-dimensional images of size $N \times N$, a filter size of $F \times F$ and stride S , this results in a feature map with the width and height of $O \times O$:

$$O = 1 + \frac{N - F}{S} \quad (2.14)$$

The depth of the feature map is determined by the number of filters. By using convolutional filters, the number of required parameters / weights is drastically reduced compared to fully connected layers. Here, only one parameter per element is used by the filter kernel, instead of one parameter for each input voxel [54].

Padding

The edges of the inputs of a convolutional layer lead to a reduction of the size of the feature map. The reason for this is that the filters can only be applied at points where there is an element in the input for each element of the filter kernel. This can lead to information being lost at the edges. To avoid this, padding can be applied. This involves adding pixels to the edges of the input to increase the size of the inputs of a convolutional layer. In this way, padding can be used to prevent a reduction of the size of the feature map [50].

If padding of size P is applied [54], the Equation 2.14 is modified to:

$$O = 1 + \frac{N + 2P - F}{S} \quad (2.15)$$

Pooling Layer

To downsample a feature map, the stride of the filter of the convolutional layer can be set to a value larger than 1. For downsampling, a pooling layer can also be used, which then increases the translational invariance of the network. For this, the input of the pooling layer is divided into smallgrid regions. For each smallgrid region, an aggregation is performed, e.g. calculation of the maximum (max-pooling) or mean (average pooling) of all values in the region. This reduces the size of the feature maps, i.e. for two-dimensional images the height and width of the feature maps. Downsampling reduces the computation time. It is also used to represent the network inputs in a more compressed form in deeper layers of the neural network. In deeper layers, the size (for two-dimensional network inputs the height and width) is reduced continuously, while the number of filters of the feature maps increases continuously [50].

To increase the size of the feature map again, upsampling is used. The feature map is padded or elements are copied to create a larger feature map.

Drop out Layer

To improve the generalisability of the network, i.e. the performance of the network on new data and applications, drop out can be applied. This involves deactivating random inputs in a dropout layer during training. For each iteration different inputs are chosen. This results in a slightly different network architecture for each iteration, which leads to more robust features being learned and increases the generalisability of the network. During inference, i.e. forward pass through the network with unseen data, all inputs are activated [55].

Activation Function

To generate non-linearity, as in the model of biological neurons, activation functions are used. These process the feature maps and serve as a mathematical ‘gate’ between two layers. The functions are differentiable to enable network training with back-propagation. Various functions can be applied for this purpose, depending on the task, a suitable activation function has to be selected.

For tasks where the neural network calculates a probability, the sigmoid activation function can be applied. The output of this function ranges between 0 and 1 and could therefore be interpreted as a probability between 0 % and 100 %. It is s-shaped and is calculated by:

$$\text{sigmoid}(x) = \frac{x}{1 + e^{-x}}. \quad (2.16)$$

The hyperbolic tangent activation function (tanh) is also characterised by an s-shape. The output values range between -1 and 1. In contrast to sigmoid, negative input values are mapped to negative output values [50]. The hyperbolic tangent activation function can be used, e.g. for the classification of the input data. It is calculated by:

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}. \quad (2.17)$$

For multi-class classification with mutually exclusive classes the softmax activation function can be applied. It is also characterised by an s-shape. Here, k real values inputs are used, the values of all k real values outputs sum up to 1. It is calculated by:

$$\text{softmax}(x_i) = \frac{e^{x_i}}{\sum_{j=1}^k e^{x_j}} \quad \forall i \in \{1, \dots, k\}. \quad (2.18)$$

In the field of medical image processing, the **Rectified Linear Unit (ReLU)** is widely applied. The **ReLU** output value is greater or equal to 0, meaning that negative inputs are mapped to 0. It is calculated by:

$$\text{ReLU}(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ x & \text{otherwise} \end{cases} \quad (2.19)$$

However, the **ReLU** activation function has the drawback that negative inputs have an output value and gradient of zero. This can lead to a **ReLU** activation function always outputting the value 0, regardless of the input, the so-called dying **ReLU** problem. However, as the gradient is 0 for negative input values, the weights will not change during training with gradient descent, making it difficult to overcome the dying **ReLU** problem [56]. The dying **ReLU** problem can be avoided if the gradient of the activation function is always non-zero. The LeakyReLU activation function can be applied for this purpose. Here, negative output values are assigned to negative input values. The LeakyReLU activation function is calculated by:

$$\text{LeakyReLU}(x) = \begin{cases} \alpha \cdot x & \text{if } x \leq 0 \\ x & \text{otherwise} \end{cases} \quad (2.20)$$

with $\alpha \in (0,1)$.

Batch Normalization

When training a network and updating all network weights at the same time, an internal covariate shift may occur, whereby the distribution of network activations is altered. However, if the inputs have a mean value of zero and unit variances, the training process converges more quickly. Batch normalisation can be used to avoid internal covariate shift and stabilise the network training. A batch normalisation layer can be inserted before or after an activation function. First, the input vector x is normalised for each dimension, i.e. the distribution of the current batch is set to a mean of zero and a variance of one:

$$\hat{x}^k = \frac{x^k - E[x^k]}{\sqrt{\text{Var}[x^k]}} \quad (2.21)$$

with the mean value E and variance Var of the current batch.

The two learnable parameters (β and γ) of the batch normalisation layer are used for scaling and shifting the normalized vector $\hat{x}$:

$$y^k = \gamma^k \hat{x}^k + \beta^k. \quad (2.22)$$

Additionally, the mean moving average and variance moving average are stored during training and used as estimation for the mean and variance during inference [57].

2.3.2 Training

By training the neural networks, the network weights w_i are optimized for a specific task. The aim is to minimize an error function, called loss function L . The training can be performed in a supervised, unsupervised or semi-supervised/weakly supervised manner. In supervised training, a ground truth, called label, is used. During training, the network output is then compared with the desired label. An error vector can then be calculated from the label and actual output of the network. In unsupervised training, no label is used; instead, only the input data is used to optimize the network weights. During training, similarities or, e.g., clusters are then calculated using the error function on the input data [51]. In semi-supervised training, auxiliary data is provided in addition to the input data. Examples of semi-supervised training of networks for medical image registration are presented in Section 2.4.3. The auxiliary data is then used in addition to the fixed and moving image to calculate the error function and determine the spatial alignment.

The optimization of network weights is difficult to implement analytically, especially for large networks with multiple layers, due to the high computational effort involved. Instead, iterative processes are applied, such as the iterative gradient decent. This method uses the gradient of the loss function to optimize the weights. This process comprises two steps, which are repeated iteratively until a certain stop criterion is reached. First, in the forward phase, input data is fed into the network, from which the network output is calculated. This output is then used to calculate the value of the loss function and the gradients of the loss function for all network weights. For the weights $w_{i,\Omega}$ for the output neurons Ω , the partial derivatives $\frac{\partial L}{\partial w_{0,\Omega}}, \frac{\partial L}{\partial w_{1,\Omega}}, \dots, \frac{\partial L}{\partial w_{n,\Omega}}$ are calculated.

The chain rule of derivatives is used to track the error through the network, i.e. to determine the share of each individual weight in the total error of the loss function [50]. In the backward phase, the network weights are then updated, i.e. $\Delta w_{i,\Omega}$ is added to the current value of the weights $w_{i,\Omega}$ to the output neurons Ω . $\Delta w_{i,\Omega}$ is calculated by:

$$\Delta w_{i,\Omega} = -\eta \frac{\partial \mathcal{L}}{\partial w_{i,\Omega}}, \quad (2.23)$$

based on the partial derivative of the loss function $\frac{\partial \mathcal{L}}{\partial w_{i,\Omega}}$. The factor η is referred to as the learning rate and defines how much the network weights are changed during an update process [51].

In addition to the gradient descent, there exist other iterative methods, such as the adaptive moment estimation (Adam) optimizer [58] or the Adagrad [59] optimizer. Both methods use adaptive learning rates for the parameters of the network. The Adam optimizer uses the moments of the gradients to determine the learning rate. Whereas with the Adagrad optimizer, a variable learning rate is used for the parameters of the network, depending on whether the features occur frequently (lower learning rate) or less frequently (higher learning rate). The aim is therefore to enable robust training, e.g. for noisy gradients [58].

During network training, a variety of challenges can occur. These include overfitting, where the network performs well on the training dataset but there is a large performance gap to a separate test dataset, meaning that the network has poor generalizability to unseen data. This can be mitigated by using regularization, which is described in more detail below, or by using a validation dataset during training, for example with k-fold cross validation. With k-fold cross validation, the entire dataset is divided into three smaller datasets (training, validation and test dataset). The training dataset is used during training to optimize the network weights, e.g. in the gradient descent. The validation dataset is used during training to evaluate the performance of the network. This enables early-stopping where training is stopped if the performance on the validation dataset does no longer improve and only the performance on the training dataset increases, i.e. the network learns irrelevant details of the training data such as noise. A drop in performance on the validation dataset during training shows that overfitting occurs, and therefore the network does not generalize well to unseen data. The test dataset is used after training to evaluate the performance of the network.

To train the networks, a loss function or a combination of several loss functions is applied. The loss functions are selected specifically for the respective task. The loss functions widely used in the field of medical image registration are explained in more detail in [Section 2.4.4](#). In addition to the loss function, a weight regularization term is usually applied, e.g. L1 or L2 regularization. In L1 regularization, the sum of the absolute value of the weights W ($\sum |W|$) is used as a penalty term with the aim of assigning the value 0 to some weights. As a result, features of the input images are selected and the model is simplified, which also lowers the memory usage. In L2 regularization, the sum of the squares (Euclidean norm, $\sum W^2$) of the weights is calculated, which ensures small network weights and leads to a more stable network [60].

During training, vanishing or exploding gradients may occur. Vanishing gradients become increasingly smaller during network training and converge to 0. As a result, the weights of the first layers are no longer updated during gradient descent. Exploding gradients, on the other hand, become increasingly large during training, which causes the gradient descent results to diverge. Vanishing or exploding gradients can be avoided, e.g. by adjusting the loss function or the learning rate and selecting a suitable weight initialization technique. For instance, the weights can be initialized with random values or certain techniques such as Xavier or Glorot initialization [61] or He initialization [62]. Additionally, local minima exist in the loss function and difficulties in convergence can occur when training is stuck in a local minimum. It

is also not clear where the global minimum of the loss function is, or whether it was reached during training [51].

There are several parameters, known as hyperparameters, that can be set or optimized for network training. These include the learning rate η . The training duration also needs to be set. This is usually specified in the number of epochs. One epoch is defined as the optimization algorithm having used the entire dataset once. The training duration can be set as a fixed period. However, if the training duration is too short, this will lead to underfitting and the network performance will be low for both the training and an unseen test dataset. Another option is to initially set a high training duration and to stop the training during the training using stop criteria, e.g. by using early-stopping if the validation loss does not improve further or deteriorates. The datasets used for training are usually too large to load and process all at once. For this reason, the data is processed in smaller groups, so-called batches. During training, a batch is then processed jointly and the network weights are updated after each batch. The batch size is also a hyperparameter that can be optimized. Furthermore, the optimizer function and the network architecture are hyperparameters that can be optimized [50].

There are various strategies for optimizing the hyperparameters. In a grid search, any combination of different parameter settings is tested. Especially if multiple hyperparameters are optimized and various settings are to be tested, there are numerous possible combinations. In this case, a random search is preferable. This is more efficient and less time consuming, as not all possible combinations but only random combinations of parameter settings are tested. Other methods for optimizing the hyperparameters include the use of a probability function (Bayesian optimization) or the use of early stopping (Hyperband [63]).

2.3.3 Pipeline

For training the networks, and for inference, the following steps are performed in medical image processing.

First, the data is loaded, i.e. the network inputs and the labels or auxiliary data if supervised or semi-supervised training is performed. In preprocessing, the image intensities are normalized in order to input data of a similar intensity range into the network. Intensity normalization can be performed using a window that sets the intensity values outside the window to the maximum or minimum value. An interval from 0 to 1 or from -1 to 1 is usually used. Another method is Z-score normalization, where the intensities are normalized to a specific mean and [Standard Deviation \(SD\)](#) [64]. Subsequently, resampling is performed to resize all input images to the same image size and pixel or voxel spacing.

During training, data augmentation can be applied. This involves enlarging the dataset with images created by adding artificial noise or, e.g., geometric transformations of the image such as translation, rotation, scaling or deformation. The type of augmentation used depends on the specific task. The data is then fed into the network in batches. During training, the network weights are optimized, e.g. by gradient descent using a forward and backward phase. During inference, the network output is calculated by a forward pass through the network.

2.3.4 Architectures

The combination of different layers of neural networks is called architecture. In a **Convolutional Neural Network (CNN)** one or more convolutional layers are used. A CNN can consist of a combination of fully connected layers and convolutional layers. In a **Fully Convolutional Neural Network (FCNN)**, no fully connected layers are utilised. Nevertheless, activation, pooling and upsampling layers can be used in CNNs and FCNNs.

Graphics Processing Unit (GPU) learning was performed for the first time with AlexNet [65]. This significantly speeded up the training process. AlexNet is used for classification and consists of five blocks, each with a convolutional layer and a pooling layer. These five blocks are followed by three fully connected layers.

Following AlexNet, more complex network architectures were developed. However, due to the principle of gradient descent and vanishing gradients, neural networks with many layers are difficult to train. When using many layers, i.e. deep networks, the gradient in the first layers becomes very small and can even become zero.

The ResNet [66] addresses this problem by introducing skip connections in the network. Some layers are skipped and the output of a layer is used as the input of a layer that is not directly following. The skip connections preserve the gradient. This alternative path through the network is used with gradient descent, with skip connections the gradient is passed back unchanged. This results in the gradients for the first layers of the network also assuming larger values. With ResNet, the skip connections are utilised in so-called residual blocks (Figure 2.5). These blocks consist of a stack of multiple convolutional layers, batch normalization and activation functions. The input of the residual block is added to the output of the stack of layers. The output $H(x)$ (before the final activation function) is therefore composed of the result of the input passed through the stack of layers of the residual block $F(x)$ and the identity term x (input residual block) [66]:

$$H(x) = F(x) + x . \quad (2.24)$$

In the field of medical image processing and also in deformable medical image registration, encoder-decoder networks are widely applied. They consist of an encoder, bottleneck and a decoder (Figure 2.6). In the encoder, convolutional layers and downsampling are used, e.g. by strides greater than 1 in the convolutional layers or by pooling layers. In this way, the information of the high-dimensional images and their features are encoded. The input images are represented in compressed form in the bottle neck, the so-called latent space. The decoder uses convolutions and upsampling operations to decode the information of the latent space and reconstruct an image.

The U-Net [67] was developed especially for the segmentation of medical images. It uses an encoder-decoder network architecture. Additionally, skip connections are used as shortcuts from the encoder to the decoder at the same level. This provides the decoder with additional information and spatial context. The information from the skip connections is then combined with the upsampled feature maps in the decoder path using concatenations.

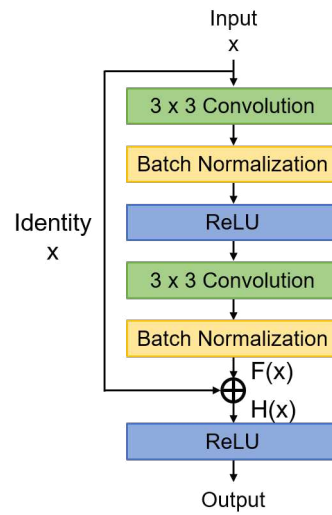


Figure 2.5: Structure of a ResNet residual block. Two-dimensional convolutions with a kernel of 3 x 3 are used.

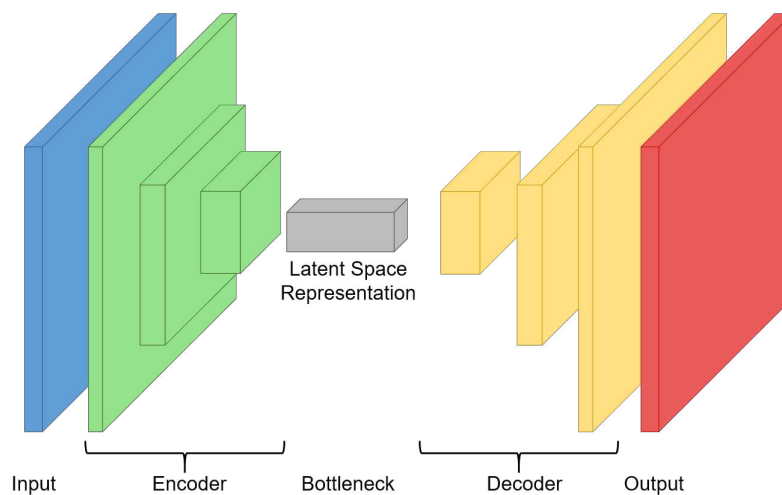


Figure 2.6: Schematic of an encoder-decoder architecture with latent space representation.

In recent years, transformers [68] have been used increasingly. They also use an encoder-decoder structure but process the input data in parallel. Additionally, attention mechanisms are used. The attention layers act as shortcuts between a context vector and the input and can therefore select which part of the input data the model should focus on. In the field of image processing, e.g. the Vision Transformer (ViT) [69] is used. It was developed specifically for image classification. The input image is processed in patches, linear and positions embeddings are applied and the data is then passed to a transformer.

2.4 Medical Image Registration

The aim of medical image registration is to calculate a transformation to optimally align all corresponding anatomical points in two or more images and to transfer the images into the same coordinate system.

Registration methods are divided into pairwise and groupwise registration. In pairwise registration, two images are registered. One image is defined as the reference (fixed) image (I_F). The other image is defined as the source (moving) image (I_M), which is then aligned with the fixed image using a transformation. During image registration, an energy function is optimized:

$$L(T; I_F, I_M) + R(T) \quad (2.25)$$

with the transformation T and the regularization term R , aiming to achieve a smooth and medically plausible transformation.

In a groupwise registration, multiple images are registered. A so-called template image is often selected as fixed image to avoid bias due to the choice of the reference [70]. This template image is calculated from the images to be registered, e.g. by intensity averaging all images. During image registration, a transformation is then calculated for each of the images to be registered, to maximize the image similarity for all images.

Image registration can be divided into additional categories, depending on the task to be performed or input images. Firstly, the dimensionality of the input images is considered. In 2D registration, two-dimensional images are registered, in 3D registration, three-dimensional volumes are registered. In 2D-3D registration, a 2-dimensional image and a three-dimensional volume are spatially aligned. Four-dimensional data, such as time series of three-dimensional volumes, can also be registered (4D registration).

Furthermore, image registration methods can be categorized as monomodal and multimodal methods. In monomodal registration, images of the same modality are registered, while in multimodal registration different modalities are spatially aligned, e.g. a CT and an MR image, or a T1-weighted and a T2-weighted MR image. The choice of modality (mono- or multimodal) has a major influence on the metric to be selected as the energy function to be optimized and for evaluating the registration results. This is discussed in more detail in [Section 2.4.4](#).

Moreover, image registration methods can be divided into intersubject and intrasubject registration. In intersubject registration, a patient's image is usually registered on a template/atlas image. In intrasubject registration, images of the same patient are registered with each other. In co-registration, images are registered, e.g. from different modalities such as CT and MR scan, in order to combine the complementary information of the different modalities. Or, e.g. scans from different acquisitions of the same modality are registered, such as a scan of a preliminary examination and a scan during the image-guided procedure to support the image-guided procedure with the information from the scan of the preliminary examination [71]. Or scans before and after therapy are registered to assess the treatment progress. Additionally, image registration can be applied to correct motion between multiple volumes of a single acquisition, e.g. 4D registration of a four-dimensional scan as a time series of three-dimensional volumes as in DCE MRI.

2.4.1 Transformations

In medical image registration, a three basic transformation types are distinguished: rigid, affine, and deformable transformation. Rigid and affine transformations are categorised as global transformations of the entire image, while deformable transformations represent local transformations of smaller image regions [72].

A rigid transformation T_{rigid} comprises translation (translation vector t) and rotation (rotation matrix R) :

$$T_{rigid}(x) = Rx + t \quad (2.26)$$

of a point x of the moving image. A rigid transformation is described by six parameters for three-dimensional images, the rotation matrix as well as the translation vector each need three parameters.

An affine transformation T_{affine} consists of translation, rotation, scaling and shearing. For three-dimensional images, the affine transformation has 12 parameters:

$$T_{affine} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad (2.27)$$

where the right column of the matrix represents the translations. The affine transformation therefore has a higher degree of freedom compared to the rigid transformation (for three-dimensional images: rigid transformation: six, affine transformation: 12 degrees of freedom).

Deformable transformations are used to correct smaller local deformations in the images that the global rigid and affine transforms are unable to correct due to their limited degrees of freedom [73]. Straightness and parallelism are not preserved in deformable transformations. The deformable transformation $T_{deformable}$ of a point x of the input image is modelled by a deformation field (displacement field) ϕ [1]:

$$T_{deformable}(x) = x + \phi(x). \quad (2.28)$$

For three-dimensional images, the deformation field is four-dimensional. The fourth dimension is 3, which means that the deformation field contains the displacement in all three spatial directions for each point x . The transformation calculated by the registration method is applied to the moving image to obtain the result (moved) image.

Especially in the field of medical image registration, a combination of different transformations, called multistage registration, is often applied. For instance, scans from medical imaging such as CT and MRI usually contain transformations with a large rigid component, e.g., the bones are rigid structures and some tumours are stiff [74] and are therefore mainly deformed rigidly. Furthermore, the varying positioning of the patient in different scanners or multiple acquisitions also results in larger rigid deformations. This can occur e.g. during image-guided biopsy as in [71]. While the patient was lying supine during a preceding CT examination, the patient may be positioned laterally during image-guided biopsy to facilitate needle positioning, which results in large rotations of the patient's body and thus a large rigid component of the total deformation. The rigid part of the transformation can be reduced in multistage registration by an initial rigid transformation. Subsequently, an affine deformation can be used as second stage of the multistage registration to reduce large global deformations and in the final stage a deformable transformation can be applied to correct smaller non-linear deformations.

Additionally, multiresolution approaches are utilised, as for medical image registration in particular, the deformations typically cannot be corrected in a single step using images at full resolution [75]. With multiresolution approaches, the images are processed in multiple differing resolutions either simultaneously in one step or sequentially in several steps. When several steps are applied, multiple cascades are used in sequence, which process the images at increasingly higher resolutions. Each cascade receives and processes the output images from the previous cascade, i.e. the pre-registered images. Each cascade calculates a transformation, e.g. a deformation field. The deformation fields of all k -th cascades can be combined by:

$$\phi_{combined} = \phi_1 \circ \dots \circ \phi_k \quad (2.29)$$

to compute the output (moved) image of the k -th cascade M'_k based on the moving image M [76, 77]:

$$M'_k = M \circ \phi_{combined}. \quad (2.30)$$

In the first cascades at low image resolution, coarse deformations are initially corrected. In the subsequent cascades, with increasingly higher image resolution, increasingly finer details are present in the images so that increasingly smaller deformations can be detected and corrected [78, 79]. The number of cascades used and the image resolution of each cascade are hyperparameters and should be optimized for the specific task.

Furthermore, the images can be pre-aligned before registration using an initial transformation. For instance, the geometric centres or the intensity-based centres of mass of the images to be registered can be aligned. If corresponding points (landmarks) are available in the images to be registered, they can be aligned with a rigid transformation as initialisation. For 2D registration, two landmarks are required, for 3D registration, three landmarks are required in order to clearly determine the rotation and translation and to prevent incorrect flipping in one of the spatial directions. Additionally, a manual pre-registration can also be performed.

2.4.2 Conventional Methods

Conventionally, iterative methods are used for medical image registration, such as SimpleElastix [17] or NiftyReg[18, 19]. The registration process is explained below and in Figure 2.7 using pairwise registration. The process for groupwise registration is performed similarly.

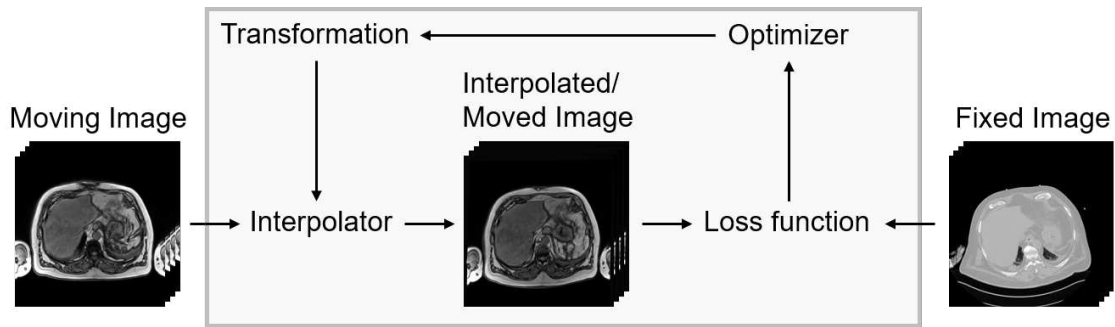


Figure 2.7: The schematic for conventional iterative image registration, based on [1].

Iterative methods utilise an energy function called loss function. This usually comprises a similarity metric that measures the spatial alignment of the images to be registered during the registration process and a penalty function for regularisation.

The loss function is optimised iteratively during the registration process. First, the moving image is transferred to the coordinate system of the fixed image using an interpolator. The voxel intensities of the moving image are resampled, as intensity values can be mapped to sub-pixel/voxel positions when transferred to a different coordinate system or by a transformation. For instance, with the nearest neighbour interpolator, the interpolated output pixel/voxel is assigned the value of the nearest sampling point. This interpolation methods be used, e.g. for the interpolation of binary masks such as organ segmentations. With the linear interpolator, the average of the nearest neighbour pixels/voxels is weighted. With spline interpolation, in addition to the neighbouring pixels, other pixels in the surrounding area are also taken into account, e.g. a cubic spline polynomial function is then used to calculate the value of the output pixel/voxel.

The spatial alignment of the moving and fixed image is then determined using the loss function. The optimizer function then optimizes a transformation based on the loss function. For instance, gradient descent as explained in Section 2.3.2 can be used as the optimizer function. Thus, based on the gradient of the loss function,

the parameters of the transformation are optimized in order to reach the optimum of the loss function.

The number of transformation parameters to be optimized depends on the dimensions of the images to be registered (e.g. 2D or 3D registration) and the transformation type used (rigid, affine, deformable). The calculated transformation is then applied to the moving image. The interpolator is used to transfer the moving image into the new coordinate system. The spatial alignment of the moved image is then compared with the fixed image using the loss function. The optimizer then re-modifies the previously calculated transformation depending on the loss function. This process is performed iteratively until a stop criterion is met, e.g. reaching a previously defined maximum number of iterations.

Due to the iterative process, conventional methods are computationally intensive and time-consuming. For the registration of each new image pair, the iterative process is repeated again. For some methods, GPU implementations are available that speed up the computation. However, this requires the availability of a GPU for each registration of a new image pair [80].

2.4.3 Deep Learning-Based Methods

Deep learning-based methods address this problem and provide a faster approach to image registration. This involves training a neural network to learn a globally optimized function. The registration of a new image pair is then performed by a single forward pass through the network, which is much faster than performing an iterative procedure for each individual image pair [16]. The images to be registered are fed into the network as input. The output of the network is one or multiple transformations, which are then applied to the moving images.

Registration networks can be trained supervised, semi-supervised or unsupervised. The training for pairwise registration is explained below, training for groupwise registration is performed in a similar way. In supervised training, synthetic deformations are usually utilised. In most cases, the fixed image is duplicated and transformed with synthetic deformations. This known deformation is then used as a ground truth/ label. During training, the output of the network is compared with the label and the error is calculated. However, especially in the field of medical image registration, the correspondence of synthetic deformations with real medical deformations is questionable, as the realistic simulation of deformations of the body caused by respiration and movement is crucial for medical plausibility.

For unsupervised training, no ground truth is required. Here, the moved image is compared with the fixed image using a chosen loss function to determine image similarity. In semi-supervised training, auxiliary information is used during training in addition to the fixed and moving images. In medical image registration, e.g. organ segmentations of the fixed and moving image or landmarks, i.e. coordinates of corresponding points in the fixed and moving image, can be used during training. The transformation calculated by the network is then also applied to the segmentation or landmarks of the moving image, resulting in the segmentations or landmarks of the moved image. The loss function that evaluates the spatial alignment based on the fixed and moved images is then extended by another loss function that determines

the spatial alignment based on the auxiliary information. The auxiliary information is only required for training. During inference, only the fixed and moving images are fed into the network.

2.4.4 Metrics

For image registration, various metrics can be used. These can be applied as loss functions for iterative processes, neural network training or for evaluating the registration results. In the following, the most commonly applied metrics are described.

In the field of medical image registration, there is usually no ground truth/label available. If synthetic images or transformations are used, metrics are calculated using this data. The neural network output is then compared with the ground truth. If no ground truth/label is available, the metrics are calculated based on the images to be registered. For pairwise registration, the fixed and moved images are used. For groupwise registration, either the registration results of all images to be registered (all moved images) are compared with each other, or alternatively each moved image is compared with the template image. The following metrics are defined for pairwise registration, i.e. fixed image (F) and moved image (M). The computation for groupwise registration is performed similarly.

Image-based metrics evaluate the spatial alignment based on the images to be registered and are categorised as intensity-based and non-intensity-based metrics. Intensity-based metrics, e.g. the [Mean Squared Error \(MSE\)](#), calculate the image similarity using the intensity of the fixed image (F) and the moved image (M):

$$MSE(F, M) = \frac{1}{n} \sum_{i=1}^n (F_i - M_i)^2. \quad (2.31)$$

While [MSE](#) compares the intensities in the fixed and moved image pixel-wise or voxel-wise, the [Structural Similarity Index Measure \(SSIM\)](#) extends this and compares luminance, contrast and structure of the fixed and moved image:

$$SSIM(F, M) = \frac{(2\mu_F\mu_M + C_1) \cdot (2\sigma_{FM} + C_2)}{(\mu_F^2 + \mu_M^2 + C_1) \cdot (\sigma_F^2 + \sigma_M^2 + C_2)}, \quad (2.32)$$

where the variables $C_1 = (K_1L)^2$ and $C_2 = (K_2L)^2$, the constants $K_1=0.01$ und $K_2=0.03$ and the dynamic range of the intensity values L are used to stabilise the division [81].

Since these two metrics use the image intensities to determine the spatial alignment, these metrics should usually only be used for monomodal registration. Multimodal images have different intensity distributions for the same structures due to using different modalities based on different physical principles. For multimodal medical image registration, entropy-based metrics such as [Mutual Information \(MI\)](#) or [Normalized Mutual Information \(NMI\)](#) can be used. [NMI](#) ranges from 0 to 1, whereby an increase in [NMI](#) through registration means improved spatial alignment. In comparison to [MI](#), [NMI](#) is more robust as it is more independent of the degree of image

overlap [82]. The NMI of the fixed image (F) and the moved image (M) is calculated by:

$$NMI(F, M) = \frac{2 \cdot I(F; M)}{H(F) + H(M)} \quad (2.33)$$

with

$$I(F; M) = H(F) + H(M) - H(F, M) \quad (2.34)$$

where H is the entropy and I is the Mutual Information.

If segmentations of the fixed (S_F) and moving image (S_M) are available, segmentation-based metrics such as the Sørensen-Dice coefficient [83, 84], also called Dice coefficient, can be applied:

$$DICE(S_F, S_M) = 2 \cdot \frac{|S_F \cap (S_M)|}{|S_F| + |S_M|}. \quad (2.35)$$

The Dice coefficient indicates the overlap of the two segmentations and ranges between 0 (no overlap) and 1 (congruent). Furthermore, the Hausdorff Distance (HD) can be applied:

$$HD(S_F, S_M) = \max(h(S_F, S_M), h(S_M, S_F)) \quad (2.36)$$

with

$$h(S_F, S_M) = \max_{s_F \in S_F} \min_{(s_M) \in (S_M)} ||s_F - (s_M)||. \quad (2.37)$$

The Hausdorff Distance represents the longest distance between the nearest points in the corresponding segmentations and is usually given in mm in the field of medical image processing. A smaller Hausdorff Distance indicates a higher spatial alignment of the segmentations.

If landmarks are available, the distance between the corresponding landmarks in the fixed (L_F) and moved image (L_M) can be determined, i.e. the Target Registration Error (TRE) using the Root Mean Square Error (RMSE):

$$TRE(L_F, L_M) = RMSE(L_F, L_M) = \sqrt{\frac{1}{n} \sum_{i=1}^n (L_{F_i} - L_{M_i})^2}. \quad (2.38)$$

Moreover, the computed transformation can be analysed, in the case of deformable transformations, the deformation field ϕ . During the training of the neural networks, the Grad L2 loss is usually applied as a second loss in order to generate a smooth and medically plausible transformation:

$$L_2(\phi) = \sum_{p \in \Omega} ||\nabla \phi(p)||^2. \quad (2.39)$$

The deformation field can also be analysed using the Jacobian determinant. The Jacobian determinant is calculated as follows for each point $P(i,j,k)$ in the deformation field:

$$\det(J(i, j, k)) = \begin{vmatrix} \frac{\partial(i)}{\partial(x)} & \frac{\partial(j)}{\partial(x)} & \frac{\partial(k)}{\partial(x)} \\ \frac{\partial(i)}{\partial(y)} & \frac{\partial(j)}{\partial(y)} & \frac{\partial(k)}{\partial(y)} \\ \frac{\partial(i)}{\partial(z)} & \frac{\partial(j)}{\partial(z)} & \frac{\partial(k)}{\partial(z)} \end{vmatrix}. \quad (2.40)$$

A Jacobian determinant > 1 indicates volume expansion, between 0 and 1 indicates shrinkage of the volume and equal to 1 indicates an unchanged volume. For diffeomorphic transformations, i.e. transformations that preserve the topology of the image [85], the Jacobian determinant of each point of the deformation field is > 0 . A Jacobian determinant equal to 0 results in the volume collapsing to zero and a Jacobian determinant < 0 inverts the orientation [72]. A Jacobian determinant ≤ 0 leads to image folding, which is medically implausible, meaning that the topology of the images is not preserved and should therefore be avoided [86]. The number of Jacobian determinant ≤ 0 in the deformation field can be assessed, a number of 0 is desired.

2.4.5 Registration Evaluation

Especially in the field of medical image registration, ground truth data is only rarely available, thereby complicating the evaluation of registration methods.

If synthetic images or transformations are used, a ground truth exists which can be used as a label for registration evaluation. If no ground truth is available, the image similarity is evaluated between the images being registered. For this, the metrics described in Section 2.4.4 can be applied. Additionally, the registration results can be evaluated visually. For instance, the images being registered can be overlaid colour-coded or fused in a checkerboard (Figure 2.8a). In this way, the spatial alignment before and after registration can be compared and analysed, e.g. the overlap of the target structures and edges in the images.

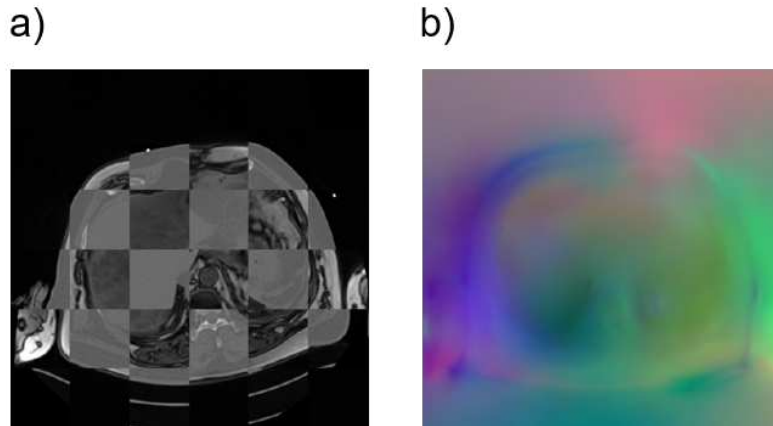


Figure 2.8: Visual methods for registration evaluation (a) checkerboard, (b) deformation field visualised in RGB channels.

Furthermore, the deformation field can be visualised and inspected, e.g. to check whether the transformation is smooth within the organs or to identify the areas where the largest transformations occur. For three-dimensional images and deformable transformations, a four-dimensional deformation field is created. This can then be encoded e.g. in RGB channels to visualise the deformation in all three spatial directions (Figure 2.8b).

Moreover, the developed registration methods are usually compared with a baseline/benchmark. These are usually already published and widely applied registration methods or methods requiring further development.

3. Deep learning-based affine medical image registration for multimodal minimal-invasive image-guided interventions – A comparative study on generalizability

This chapter is based on the following publication: A. Strittmatter, L. R. Schad, F. G. Zöllner. Deep learning-based affine medical image registration for multimodal minimal-invasive image-guided interventions – A comparative study on generalizability. *Zeitschrift für Medizinische Physik*, 34 (2), pp. 291-317, doi: 10.1016/j.zemedi.2023.05.003, 2024.

3.1 Introduction

Patients with oligometastatic disease have a poor survival probability when employing standard therapeutic options due to local lesion heterogeneity. The genetic and molecular characteristics of the cancer cells within a single tumor or among different metastatic lesions can vary, resulting in some cells being resistant to certain treatments while others may be more susceptible. This heterogeneity can lead to incomplete tumor shrinkage and the potential for the development of treatment resistance. However, using tailored techniques, such as minimally invasive image-guided therapies that incorporate tumor heterogeneity, greatly increases survival [12, 13]. Such interventions are commonly performed using [Computed Tomography \(CT\)](#) imaging while tumor heterogeneity can be derived using [Magnetic Resonance Imaging \(MRI\)](#). Also, visibility of certain lesions in [CT](#) is not always given but can be visualized in [MRI](#) [14].

The research campus "Mannheim Molecular Intervention Environment" (M²OLIE) is working on optimizing the treatment of oligo metastatic liver cancer, especially image-guided procedures. For this purpose, three-dimensional [CT](#) and [Magnetic Resonance \(MR\)](#) images are acquired pre- and postoperatively, while three-dimensional [Cone-Beam Computed Tomography \(CBCT\)](#) is applied intraoperatively. To facilitate oncological interventions as well as to monitor the progress of treatment, the images acquired by the different modalities are merged into a multimodal dataset, thus combining all available morphological and functional information regarding the tumour [1]. Therefore the registration of [CT](#) and [MRI](#) images is warranted.

Typically, conventional methods are used for medical image registration. The transformation parameters are iteratively updated to optimise a similarity metric. However, conventional methods are very time-consuming, as the registration of each new image is performed iteratively again. Registration methods based on deep learning address this issue and are used to speed up the registration process. The network is trained once. To register new images, a forward pass through the network is performed, eliminating the need for time-consuming iterations [16].

In recent years, several supervised, semi-supervised or unsupervised neural networks for medical image registration with a wide variety of architectures have been published. These publications are also collected in reviews [2], [3]. However, as different datasets were used to train and evaluate the networks, a fair comparison is difficult to achieve and to the best of our knowledge has yet not been performed. This paper addresses this point. The aim of this paper is to analyse and assess the performance of several neural networks for multimodal affine image registration and to evaluate the generalizability of these networks to new datasets. For this purpose, 20 different neural networks were implemented and compared using two multimodal datasets - a synthetic and a patient dataset - of three-dimensional CT and MR images of the liver. First, the networks were trained semi-supervised with the synthetic dataset. Then the networks' performance was evaluated using the synthetic dataset and the unseen patient dataset. Afterwards, the networks were finetuned with the patient dataset and then evaluated on the patient dataset.

The paper is organised as follows: Section 3.2 outlines the existing image registration methods, Section 3.3 describes the framework used to compare the network architectures and Section 3.4 presents the experimental results.

3.2 Background

In image registration, a transformation is calculated to optimally align two or more images. In this process, the source (moving) image is registered to the reference (fixed) image. The calculated transformation is applied to the moving image, thus creating the moved image. An affine transformation consists of the following individual transformations: translation, rotation, scaling and shearing. The affine transformation is represented for three-dimensional images in a 4 x 4 matrix with 12 parameters (see Equation 3.1). A combination of the single transformations is achieved by matrix multiplication [87]. The right column (parameters a_{14}, a_{24}, a_{34}) of the transformation matrix contains the translation parameters. The remaining nine parameters represent the combined values for rotation, scaling and shearing [1].

$$T_{affine} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (3.1)$$

Conventional registration methods apply a mathematical optimization model that is solved individually for each image pair. In an iterative process, a function is maximized that describes the similarity of fixed and moved images. This is computationally intensive and slow due to the iterative process. Thus, affine registration with state of the art algorithms such as SimpleElastix [17] or NiftyReg [19] takes several seconds to a minute on the Central Processing Unit (CPU). Some of the conventional algorithms also have Graphics Processing Unit (GPU) implementations that shorten the duration of a single registration (a few seconds), but then a GPU is needed for each registration process [80].

When using neural networks, on the other hand, the time-consuming optimization for each new image pair is omitted. A neural network takes two n-dimensional images as input and outputs the transformation parameters. The weights of the neural network are optimized by training. Thus a global function is optimized once, the optimization for each individual test image pair is omitted. For registration, the learned function is evaluated, which saves a lot of time on the CPU as well as on the GPU compared to conventional methods. The networks are trained supervised, unsupervised or semi-supervised (weakly-supervised). Supervised methods require ground truth, which usually does not exist, especially in medical image registration. Instead, images are used as target label, which were registered manually or by conventional methods. This has the drawback that the trained networks are limited by the accuracy of the used target label. The networks can therefore only deliver results with a maximum accuracy that corresponds to the accuracy of the target label. Synthetically generated transformations can also be used. This involves duplicating the fixed image and then transforming it with a synthetically generated transformation to create the moving image. However, it is disputable to what extent these transformations correspond to real transformations. The problem of real image registration between several acquisitions is not fully characterized by the synthetic input data, e.g. the synthetically transformed image (moving image) will contain the same noise as the fixed image. Furthermore, deformation of the patient's body due to breathing as well as patient movement would have to be simulated in a realistic way.

Unsupervised learning, on the other hand, does not require ground truth. During training, the moved image, i.e. the moving image transformed by the computed transformation, is compared with the fixed image. If anatomical segmentations are available for the training data set, the training can be performed semi-supervised (weakly-supervised). In addition to the fixed and moving image, segmentations of the fixed and moving image are also passed to the neural network during training. In addition to the image-based losses, segmentation driven loss functions, such as Dice loss, are used, which can improve the registration result compared to unsupervised learning [16].

3.3 Methods

We implemented 20 neural networks for affine registration of medical images, which we found on IEEE Xplore, PubMed, Web of Science and ScienceDirect with the keywords affine, registration, deep learning, artificial neural network, medical images. Generative Adversarial Networks (GAN) are not taken into account, since an additional loss (adversarial loss) is applied during training by using a second network (discriminator). GAN can therefore not be fairly compared with the other networks, which were trained with only one loss (Dice loss).

3.3.1 Network Architectures

We implemented and compared the architectures for affine registration of 17 papers, one abstract and 2 theses. The architectures can be broadly grouped into different categories, as shown in Figure 3.1. Seventeen architectures can be categorised as Convolutional Neural Networks (CNNs). The CNNs can be further subdivided

into regression networks (16 papers), for example the encoder part of the U-Net followed by dense layers - as well as U-Net-like networks (1 thesis), where the affine parameters are formed at the end of the network by a Global Average Pooling layer. Regression networks can be further subdivided into convolution-based (11 papers) and Siamese network-based (4 papers and 1 abstract). Two papers use transformer, the Swin transformer [88] and an adaptation of the Vision transformer [89]. One network architecture [90] is assigned to the category "other network architectures", a neural network for affine registration is used which contains only one single dense layer with 12 neurons. The architecture category of each network can be found in Table 3.1.

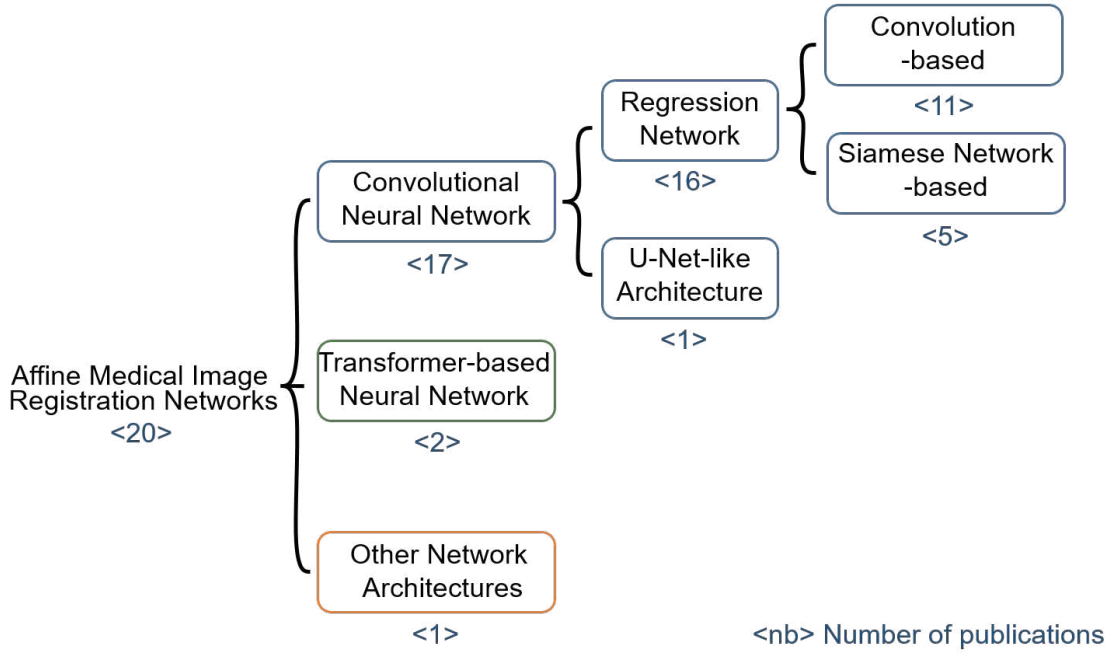


Figure 3.1: Categories of the implemented architectures.

In some papers, parts of the networks' implementation were not disclosed in detail, especially parameter settings were not reported making the reproducibility difficult. For example, in [91] the activation function was not specified and in [92] the kernel size of the convolutions was missing. Assumptions were made for the parameters, which were not specified in the paper or in published source code. Settings were chosen that are commonly used in affine registration network architectures, for example, the activation function LeakyReLU and a kernel size of 3 x 3 x 3. The assumptions that were made during the implementation of the networks can be found in Table 3.1.

In four papers two-dimensional datasets were used. This paper evaluates the networks using three-dimensional data. For this reason, all networks have been adapted to three dimensions. Only the paper [93] uses a two-dimensional affine network to register three-dimensional data. For comparison with the architectures of the other papers, the two-dimensional original network and the three-dimensional adaptation were used.

To allow for a fair comparison of the different algorithms without introducing a bias due to wrong assumptions or errors in the reimplementations of the networks, the

corresponding authors of the original publications were contacted to verify that our implementation of their networks matches those used in the publications (Table 3.1 column "Implementation confirmed by the author of the original publication").

Table 3.1: Additional information on the implemented networks

References	Category according to Figure 3.1	Implementation confirmed by the author of the original publication	Assumptions	Normalization technique	Learning rate	Optimizer function
Benchmark (Our CNN)	Convolution-based	-	-	[0, 1]	1.00E-03	Adam
Hu et al., 2018 [94]	Convolution-based	Code available in NiftyNet, DeepReg, MONAI	-	[0, 1]	1.00E-04	SGD
Gu et al., 2019 [95]	Convolution-based	No contact details found	-	[0, 1]	1.00E-02	SGD
Shen et al. 2019 [96]	Convolution-based	Did not respond	-	[0, 1]	1.00E-04	Adam
Zhao et al., 2020 [91]	Convolution-based	Did not respond	Activation function: LeakyReLU	[0, 1]	1.00E-02	Adagrad
Luo et al., 2020 [97]	Convolution-based	Did not respond	-	[0, 1]	1.00E-04	SGD
Tang et al., 2020 [98]	Convolution-based	Did not respond	-	[0, 1]	1.00E-04	Adam
Zeng et al., 2020 [93] 2D	Convolution-based	Did not respond	-	[0, 1]	1.00E-04	Adagrad
Zeng et al., 2020 [93] 3D	Convolution-based	Did not respond	-	[0, 1]	1.00E-04	Adagrad
Chen et al., 2021b [99]	Convolution-based	Did not respond	-	mean=0, std=1	1.00E-02	Adam
Gao et al., 2021 [100]	Convolution-based	Did not respond	-	[0, 1]	1.00E-04	Adam
Roelofs, 2021 [101]	Convolution-based	No contact details found	-	[0, 1]	1.00E-04	Adam
Zhu et al., 2021 [102]	Convolution-based	Initial contact, no further response	-	[0, 1]	1.00E-04	Adam
Chee and Wu, 2018 [103]	Siamese Network-based	Did not respond	Number of filters in convolutional layers: [16, 8, 8, 8, 8, 8, 8]	[0, 1]	1.00E-04	Adam
de Vos et al., 2019 [86]	Siamese Network-based	Did not respond	Number of filters in convolutional layers: [16, 32, 64, 128, 256] Number of neurons in first dense layer: 32 Kernel size: 3x3x3 Drop out rate: 0.2	[0, 1]	1.00E-04	Adam
de Silva et al., 2020 [92]	Siamese Network-based	Did not respond	-	[0, 1]	1.00E-04	Adam
Shao et al., 2021 [104]	Convolution-based	Initial contact, no further response	-	[0, 1]	1.00E-04	Adam
Venkata et al., 2022 [105]	Siamese Network-based	Initial contact, no further response	Number of filters in convolutional layers: [16, 8, 8, 8, 8, 8, 8]	[0, 1]	1.00E-04	Adam
Waldkirch, 2020 [1]	U-Net-like Architecture	Received original source code	-	[0, 1]	1.00E-03	Adam
Chen et al., 2022 [88]	Transformer-based	Did not respond	-	mean=0, std=1	1.00E-02	Adam
Mok and Chung, 2022 [89]	Transformer-based	Initial contact, no further response	-	mean=0, std=1	1.00E-02	Adam
Hasenstab et al., 2019 [90]	Other Network Architecture	Did not respond	-	[0, 1]	1.00E-04	SGD

3.3.2 Benchmark Neural Network

As benchmark, we used our own developed CNN (Figure 3.2), a ResNet-18 encoder and a U-Net-like decoder with one convolution per convolutional block without skip connections. The decoder consists of six three-dimensional convolutions with a kernel size of $3 \times 3 \times 3$, a stride of 1 and padding same. The first five convolutions and activation functions (LeakyReLU) of the decoder are followed by upsampling. Global Average Pooling is then applied to generate the 12 parameters needed for a three-dimensional affine transformation. To calculate the moved image, the affine transformation is then applied to the moving image. This can be done by matrix multiplication or a spatial transformer [106]. We applied a spatial transformer as this was also used for affine transformation in numerous publications, including [103], [88], [89], [98], [105], [91] and [102], and it facilitates an adaptation of the entire network structure for deformable or multi-stage registration as a future work.

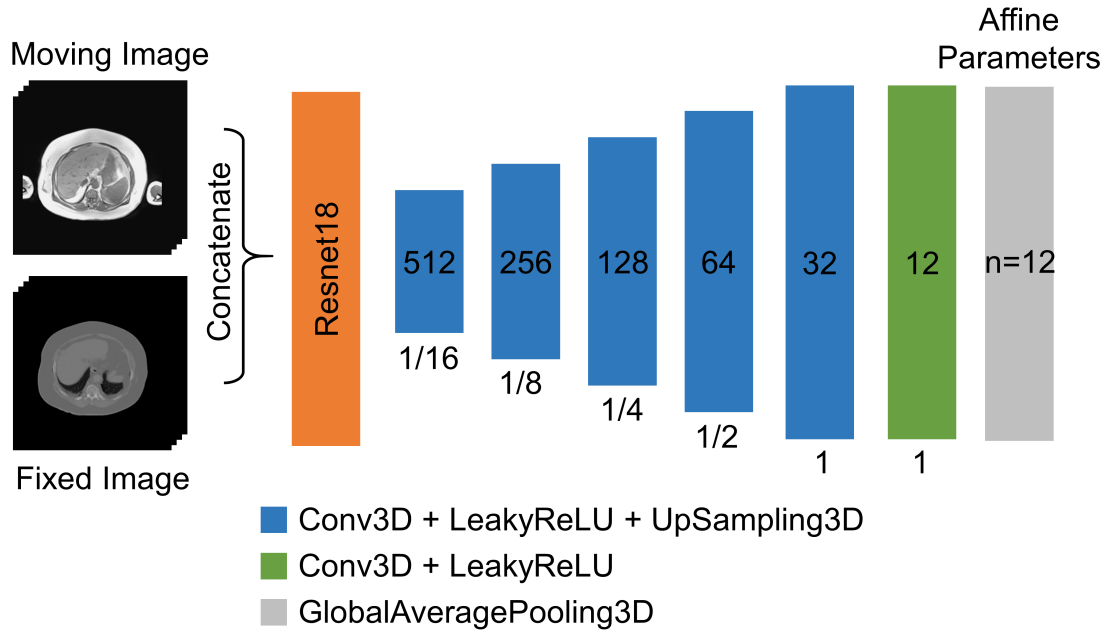


Figure 3.2: Architecture of the benchmark CNN for affine registration

The benchmark CNN was specifically designed for the synthetic dataset. In an optimization process, different architectures were examined, e.g. for the encoder part, different variants of the ResNet, VGGNet or a U-Net encoder. In addition, different numbers of filters were investigated for the convolutional layers in the decoder. The architecture that yielded the highest Dice coefficient on the validation data of the synthetic dataset was then chosen as the benchmark CNN for our experiments.

3.3.3 Baseline Method

As baseline, i.e. as state-of-the-art, we applied the Elastix affine registration algorithm implemented in the SimpleElastix Toolbox using default parameters for an affine transformation. Mutual Information (*MI*) is used as objective function.

3.3.4 Training Setting

The training of the implemented neural networks is performed with semi-supervised learning according to the VoxelMorph framework [16]. During training, in addition to the fixed and moving images, the segmentations of the fixed and moving images are given as input. The neural network calculates the 12 parameters required for an affine transformation of 3-dimensional images using the fixed image and the moving image. This transformation is then applied to the segmentation of the moving image. The result is the moved image segmentation. Our goal when performing affine registration was to register the target structure (liver) as accurately as possible in both images, i.e. to maximize the overlap of the segmentations of the livers in the fixed and moved image. For this reason, dice loss was chosen as the loss function. During training, the overlap of the segmentation of the moved image and the segmentation of the fixed image is calculated. We performed fivefold cross-validation and trained the networks for 100 epochs with a batch size of one, stopping after five epochs with no improvement in validation loss. The network's weights with the highest accuracy on the validation data are saved.

For each of the proposed models, the normalisation technique, learning rate and optimizer function were optimized using the Keras Hyperband Tuner [63]. The learning rate was optimized using 1e-2, 1e-3, 1e-4 and the optimizer functions Adam, *Stochastic Gradient Descent (SGD)* and Adagrad were used. As image normalisation techniques, the scaling of the image intensities between [0,1] and a z-score normalisation (mean is 0 and standard deviation is 1) were tested. The parameter configuration that yielded the highest Dice coefficient was selected as the optimal parameter, as shown in Table 3.1. The networks were then trained with this parameter configuration. We resized the input of the network, i.e. the fixed image, the moving image and the segmentation masks to 256x256x64 voxels. The images were used in half resolution due to memory limitations of the GPUs. For evaluation, the resulting affine transformation was applied to the full resolution images. No augmentation was applied.

The networks were first trained with the synthetic dataset. Afterwards, the networks were finetuned with the patient dataset. Therefore, the weights of the networks were initialised with the weights of the networks that were trained on the synthetic dataset. The networks were then finetuned using the training settings described in Table 3.1 (normalisation technique, learning rate, optimizer function). Finetuning was performed using Dice loss.

All experiments were performed in Python using Tensorflow [107] on a workstation equipped with a NVIDIA RTX A5000 with 24 GB GPU memory and a NVIDIA RTX A6000 with 48 GB GPU memory. The networks with a convolution-based network architecture or other network architecture (Figure 3.1 and Table 3.1) were trained on the NVIDIA RTX A5000. The transformer-based neural networks were trained on the NVIDIA RTX A6000, as they required more than 24 GB GPU memory for training.

Our implementations are available at https://github.com/Computer-Assisted-Clinical-Medicine/Networks_for_Affine_Medical_Image_Registration. A manual for our source code can be found in Section 3.7.3.

3.3.5 Datasets

To investigate the performance of the neural networks and the networks' generalizability to new datasets we used two datasets.

The networks were trained and evaluated with a synthetic dataset of 112 **CT** and **MR** T1-weighted three-dimensional images of the liver. These volumes were generated from images of the XCAT [108] phantom using CycleGAN [109]. Each **CT** volume is reconstructed to a 512 x 512 matrix with 82-124 slices and a resolution of 1 x 1 x 2 mm³. The **MR** volumes were created with a matrix of 330 x 450 and 54-83 slices to match a resolution of 1 x 1 x 3 mm³. This dataset contains anatomical segmentation masks for the liver. The **CT** and **MR** T1-weighted images are available in two breathing states: inhaled and exhaled. In the scope of this paper, the **MR** images are registered affinely to the **CT** images, the inhaled state to the exhaled state and vice versa. The synthetic dataset is available at <https://heidata.uni-heidelberg.de/dataset.xhtml?persistentId=doi:10.11588/data/ICSFUS>.

The second dataset consists of **CT** and **MR** T1-weighted patient images of the liver. These were acquired on 39 patients as part of the M²OLIE project. This dataset contains 47 **CT** images and 73 **MR** T1-weighted images (in-phase and opposed-phase) in transversal orientation. The in-plane resolution of the **CT** scans is 0.78 x 0.78 (± 0.12) mm² and the slice thickness is 2.04 (± 0.76) mm. The **CT** volumes were made with a 512 x 512 matrix and 25-132 slices. The **MR** scans have an in-plane resolution of 1.27 x 1.27 (± 0.17) mm² and slice thickness of 3.77 (± 0.46) mm. The matrices used for the **MR** scans range from 190-440 x 288-640 voxels with 52-120 slices. Anatomical segmentation masks of the liver were created using two neural networks - separately for the **CT** images and for the **MR** images. Three-dimensional U-Nets were used. To perform liver segmentation on **CT** data, a network was trained with the 3D-IRCADb-01 dataset [110], for liver segmentation on **MR** data, a network was trained with the **MR** images of the Combined Healthy Abdominal Organ Segmentation (CHAOS) dataset [111]. Afterwards, the networks were finetuned with 20 segmentations of the **CT** images and 20 segmentations of the **MR** images of the M²OLIE dataset, which were created manually in the Medical Imaging Interaction Toolkit (MITK) [112]. To assess the performance of the two U-Nets for the segmentation of the patient dataset, the segmentations generated by the U-Nets were checked visually in order to verify that the segmentations were all reasonable. In the scope of this paper, the **MR** images are registered affinely to the **CT** images.

For both datasets, we used trilinear interpolation to resample all volumes to an identical voxel spacing of 1 x 1 x 2 mm³. The volumes were then cropped and zero-padded to a size of 512 x 512 x 128 voxels.

3.3.6 Evaluation Metrics

We evaluated the registration's performance using three metrics. **Normalized Mutual Information (NMI)** was applied to quantify structural differences of the fixed image (F) and the moved image (M). The **NMI** is defined as:

$$NMI(F, M) = \frac{2 \cdot I(F; M)}{H(F) + H(M)} \quad (3.2)$$

with

$$I(F; M) = H(F) + H(M) - H(F, M) \quad (3.3)$$

where I is the Mutual Information and H is the entropy.

In addition, we measured the overlap of the segmentations of the moved image (S_M) and the fixed image (S_F) using the Dice coefficient and the Hausdorff distance [Hausdorff Distance \(HD\)](#).

The Dice coefficient is given by:

$$DICE(S_F, S_M) = 2 \cdot \frac{|S_F \cap S_M|}{|S_F| + |S_M|} \quad (3.4)$$

The Hausdorff distance is defined as:

$$H(S_F, S_M) = \max(h(S_F, S_M), h(S_M, S_F)) \quad (3.5)$$

with

$$h(S_F, S_M) = \max_{s_F \in S_F} \min_{s_M \in S_M} ||s_F - s_M|| \quad (3.6)$$

We compared the metrics' mean and standard deviation ([Standard Deviation \(SD\)](#)) across the neural networks described in the papers, using our [CNN](#) as a benchmark and the conventional affine registration method as a baseline ([Table 3.2](#)).

In addition, we used a paired t-test (if normally distributed) or a Wilcoxon signed-rank test (if not normally distributed) to examine the significance of the improvement in [NMI](#), Dice coefficient, and Hausdorff distance by registration compared with the unregistered data. The null hypothesis states that registration by the different implemented methods does not improve the three specified evaluation metrics. This hypothesis is rejected if $p < 0.05$.

3.4 Results

3.4.1 Synthetic Dataset

We trained and evaluated the networks with the synthetic dataset (column "Synthetic Dataset" in [Table 3.2](#)).

By using the conventional registration method SimpleElastix, the [NMI](#), Dice coefficient as well as the Hausdorff distance could be improved compared to the unregistered images.

In contrast, no implemented neural network could increase the [NMI](#) of the unregistered images (pre-registration [NMI](#)). Six out of 20 networks implemented from papers improved the Dice coefficient significantly (p -value < 0.05) [[1](#), [86](#), [88](#), [91](#), [95](#), [104](#)] but only one network [[88](#)] was able to reduce the Hausdorff distance significantly. The network of [[88](#)] improved the two segmentation-based metrics Dice coefficient and Hausdorff distance significantly simultaneously.

Table 3.2: The Normalized Mutual Information (NMI), Dice coefficient (%) and Hausdorff distance (HD, in mm) results (mean $\pm$ SD) on synthetic Dataset and real patient Dataset for the implemented networks, the benchmark (our CNN) and the baseline (SimpleElastix affine). NMI: higher value is better (maximum is 1), Dice coefficient: higher value is better (maximum is 100%), Hausdorff distance: lower value is better (minimum is 0). Marked in green: improvement compared to unregistered data, with the bold numbers indicating significant improvement (p -value < 0.05) to the unregistered data

Method	Synthetic Dataset			Patient Dataset evaluated			Patient Dataset finetuned		
	NMI	DICE (%)	HD (mm)	NMI	DICE (%)	HD (mm)	NMI	DICE (%)	HD (mm)
Before registration	0.246 $\pm$ 0.024	66 $\pm$ 4	23 $\pm$ 3	0.128 $\pm$ 0.042	50 $\pm$ 18	50 $\pm$ 27	0.128 $\pm$ 0.042	50 $\pm$ 18	50 $\pm$ 27
SimpleElastix affine	0.271 $\pm$ 0.021	73 $\pm$ 5	18 $\pm$ 3	0.225 $\pm$ 0.067	60 $\pm$ 19	36 $\pm$ 18	0.225 $\pm$ 0.067	60 $\pm$ 19	36 $\pm$ 18
Benchmark (Our CNN)	0.192 $\pm$ 0.021	89 $\pm$ 5	18 $\pm$ 9	0.093 $\pm$ 0.034	49 $\pm$ 19	51 $\pm$ 27	0.124 $\pm$ 0.035	66 $\pm$ 17	43 $\pm$ 25
Hu et al., 2018 [94]	0.190 $\pm$ 0.021	62 $\pm$ 13	35 $\pm$ 7				0.109 $\pm$ 0.035	50 $\pm$ 16	51 $\pm$ 26
Gu et al., 2019 [95]	0.186 $\pm$ 0.015	87 $\pm$ 5	23 $\pm$ 5	0.102 $\pm$ 0.037	51 $\pm$ 18	50 $\pm$ 27	0.125 $\pm$ 0.034	70 $\pm$ 16	46 $\pm$ 27
Shen et al. 2019 [96]	0.210 $\pm$ 0.021	66 $\pm$ 8	26 $\pm$ 3				0.121 $\pm$ 0.042	47 $\pm$ 19	51 $\pm$ 27
Zhao et al., 2020 [91]	0.197 $\pm$ 0.020	85 $\pm$ 6	24 $\pm$ 5	0.102 $\pm$ 0.037	50 $\pm$ 18	51 $\pm$ 28	0.125 $\pm$ 0.036	69 $\pm$ 16	46 $\pm$ 27
Luo et al., 2020 [97]	0.172 $\pm$ 0.027	63 $\pm$ 12	37 $\pm$ 11				0.108 $\pm$ 0.036	52 $\pm$ 15	53 $\pm$ 27
Tang et al., 2020 [98]	0.109 $\pm$ 0.031	44 $\pm$ 14	83 $\pm$ 45				0.095 $\pm$ 0.042	25 $\pm$ 20	82 $\pm$ 35
Zeng et al., 2020 [93] 2D	0.088 $\pm$ 0.041	25 $\pm$ 13	120 $\pm$ 63				0.052 $\pm$ 0.027	18 $\pm$ 11	116 $\pm$ 53
Zeng et al., 2020 [93] 3D	0.080 $\pm$ 0.042	20 $\pm$ 15	123 $\pm$ 60				0.049 $\pm$ 0.030	17 $\pm$ 11	123 $\pm$ 67
Chen et al., 2021b [99]	0.219 $\pm$ 0.038	65 $\pm$ 5	27 $\pm$ 8				0.106 $\pm$ 0.037	50 $\pm$ 18	50 $\pm$ 27
Gao et al., 2021 [100]	0.157 $\pm$ 0.024	54 $\pm$ 11	42 $\pm$ 9				0.125 $\pm$ 0.047	45 $\pm$ 19	54 $\pm$ 27
Roelofs, 2021 [101]	0.230 $\pm$ 0.023	68 $\pm$ 7	28 $\pm$ 4	0.127 $\pm$ 0.041	50 $\pm$ 18	50 $\pm$ 28	0.137 $\pm$ 0.050	51 $\pm$ 18	50 $\pm$ 28
Zhu et al., 2021 [102]	0.210 $\pm$ 0.025	64 $\pm$ 5	29 $\pm$ 4				0.132 $\pm$ 0.044	50 $\pm$ 18	53 $\pm$ 27
Chee and Wu, 2018 [103]	0.174 $\pm$ 0.050	51 $\pm$ 13	57 $\pm$ 23				0.125 $\pm$ 0.041	45 $\pm$ 17	56 $\pm$ 26
de Vos et al., 2019 [86]	0.220 $\pm$ 0.019	78 $\pm$ 4	23 $\pm$ 6	0.126 $\pm$ 0.042	51 $\pm$ 18	51 $\pm$ 27	0.140 $\pm$ 0.040	60 $\pm$ 17	48 $\pm$ 30
de Silva et al., 2020 [92]	0.244 $\pm$ 0.023	66 $\pm$ 4	24 $\pm$ 3				0.134 $\pm$ 0.044	52 $\pm$ 17	49 $\pm$ 28
Shao et al., 2021 [104]	0.179 $\pm$ 0.040	75 $\pm$ 10	27 $\pm$ 11	0.118 $\pm$ 0.039	32 $\pm$ 13	68 $\pm$ 25	0.119 $\pm$ 0.039	52 $\pm$ 18	53 $\pm$ 25
Venkata et al., 2022 [105]	0.193 $\pm$ 0.031	51 $\pm$ 14	51 $\pm$ 16				0.123 $\pm$ 0.042	45 $\pm$ 16	57 $\pm$ 27
Waldkirch, 2020 [1]	0.189 $\pm$ 0.020	85 $\pm$ 7	25 $\pm$ 7	0.099 $\pm$ 0.037	49 $\pm$ 19	51 $\pm$ 28	0.123 $\pm$ 0.039	56 $\pm$ 17	50 $\pm$ 17
Chen et al., 2022 [88]	0.040 $\pm$ 0.196	77 $\pm$ 21	19 $\pm$ 13	0.111 $\pm$ 0.040	47 $\pm$ 20	50 $\pm$ 27	0.115 $\pm$ 0.038	52 $\pm$ 16	47 $\pm$ 26
Mok and Chung, 2022 [89]	0.208 $\pm$ 0.018	67 $\pm$ 9	35 $\pm$ 5	0.103 $\pm$ 0.035	48 $\pm$ 18	53 $\pm$ 26	0.107 $\pm$ 0.036	50 $\pm$ 18	50 $\pm$ 27
Hasenstab et al., 2019 [90]	0.016 $\pm$ 0.017	2 $\pm$ 4	145 $\pm$ 42				0.039 $\pm$ 0.021	9 $\pm$ 7	132 $\pm$ 52

The best results in terms of [NMI](#) were obtained with the neural network of [\[92\]](#), but registration with this network did not increase the [NMI](#) compared to the unregistered data. The network also did not succeed in improving the segmentation-based metrics Dice coefficient and Hausdorff distance.

The neural network of [\[95\]](#) achieved the highest Dice coefficient. Our benchmark network solely yielded a higher Dice coefficient. The networks of [\[91\]](#) and [\[1\]](#) also achieved good Dice coefficient values, which were above 80%.

The worst results for all three metrics were obtained by the neural network of [\[90\]](#). The network failed to register the data at all.

Examples of registration results can be seen in [Figure 3.3](#). Before registration, the abdomen in the [CT](#) and [MR](#) images is not well aligned. Registration with the baseline SimpleElastix (second row in [Figure 3.3](#)) improved the structural similarity and the overlap of liver segmentations. After registration with the neural network of [\[98\]](#) (a model that performed poorly in the experiment), the structural similarity and overlap of the anatomical segmentations has worsened ([Figure 3.3](#), third row). An inaccurate affine transformation matrix was predicted by the neural network, so the moved image was rotated incorrectly.

An example of a good registration is shown in [Figure 3.3](#) in the fourth row. This is the registration result of the best performing network from papers in terms of the Dice coefficient [\[95\]](#). The overlap of the liver segmentations improved but the abdomen of the moved image and the fixed image align worse than before registration. The registration result from our own developed CNN ([Figure 3.3](#) last row) further increased the overlap of the anatomical segmentations compared to the best performing network from the papers. As visible in the composite image ([Figure 3.3](#), fourth column), the benchmark CNN also yielded a lower structural similarity than before the registration. A similar figure with examples of registration results for all 20 networks from papers and the baseline and benchmark can be found in [Figure 3.5](#) in [Section 3.7.1](#).

3.4.2 Patient Dataset

We evaluated the networks trained on the synthetic dataset with the patient dataset (column "Patient Dataset evaluated" in [Table 3.2](#)). Only those models are listed in the column "Patient Dataset evaluated" that were able to improve the Dice coefficient of the synthetic dataset.

No implemented network was able to increase the pre-registration [NMI](#). Only three networks could improve the pre-registration Dice coefficient ([\[86, 91, 95\]](#)), but none of them significantly ($p\text{-value} < 0.05$) and no network was able to reduce the Hausdorff distance significantly. Comparable to the synthetic dataset, the network of [\[95\]](#) yielded the highest Dice coefficient for the patient dataset.

We then finetuned the networks trained on the synthetic dataset with the patient dataset (column "Patient Dataset finetuned" in [Table 3.2](#)). Nine finetuned networks could increase the pre-registration Dice coefficient of the patient dataset significantly ($p\text{-value} < 0.05$) [\[1, 86, 88, 91, 92, 95, 97, 101, 104\]](#) and five networks improved the

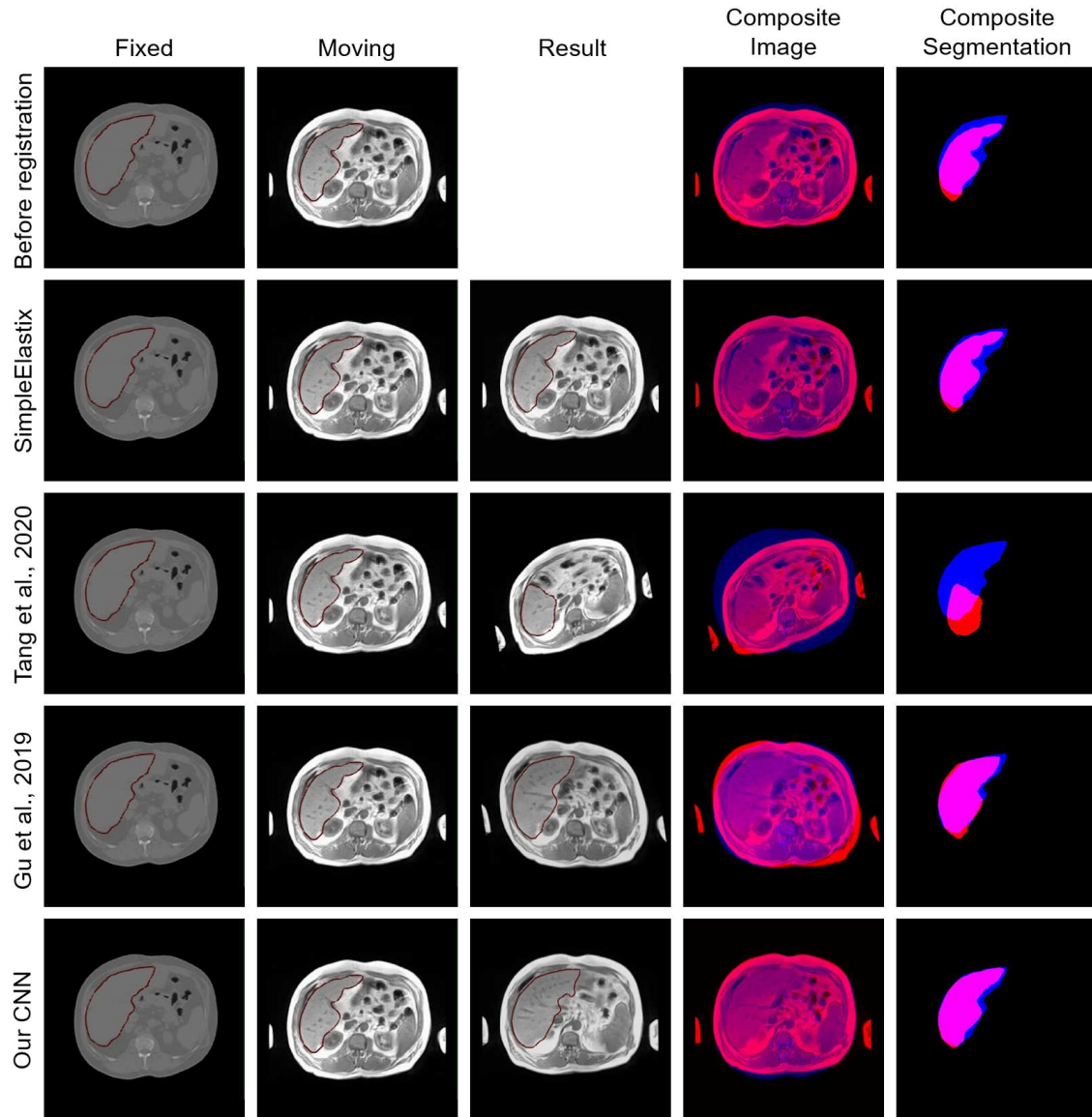


Figure 3.3: Example results of an affine registration of two volumes from the synthetic dataset. The images show central slices of the axial plane. The fixed images, moving images and result (moved) images are overlaid by the liver segmentations (column 1 - 3). The resulting composites for image and segmentation show the fixed data in blue and the moving/moved data in red. The third row illustrates an example of a poor registration result. Row four and five represent good registration results.

pre-registration Hausdorff distance significantly [86, 88, 91, 92, 95]. Once again, the network of [95] yielded the highest Dice coefficient.

Examples of registration results of the patient dataset are shown in Figure 3.4. Before registration (first row in Figure 3.4), the abdomen is not well aligned in the CT and MR images. This can be observed distinctly in the composite image and composite segmentation (second to last and last column). Registration with the baseline SimpleElastix (second row in Figure 3.4) improved the structural similarity and the overlap of liver segmentations. After registration with the neural network of [98] (a model that performed poorly in the experiment), the structural similarity and the overlap of the anatomical segmentations decreased (Figure 3.4, third row). As already observed for the network of [98] trained with the synthetic dataset, the network of [98] finetuned with the patient dataset computed a wrong affine transformation, so that the moved image was rotated incorrectly. The NMI, the Dice coefficient and the Hausdorff distance deteriorated compared to the unregistered images (see Table 3.2, row "Tang et al., 2020").

The fourth row in Figure 3.4 shows the registration result of the network with the best performance in terms of the Dice coefficient [95]. The overlap of the liver segmentations improved compared to the unregistered data. The registration result of our CNN (Figure 3.4 last row) is worse than the registration result of the network of [95]. Our own developed CNN could only slightly improve the the structural similarity and the overlap of the segmentation of fixed and moved image compared to the unregistered images. A similar figure with examples of registration results for all 20 networks from papers and the baseline and benchmark can be found in Figure 3.6 in Section 3.7.1.

3.5 Discussion

We implemented 20 different machine learning-based image registration methods and evaluated them with respect to an affine multimodal registration of CT and MR images of the liver. Furthermore, we implemented a framework to test and evaluate these approaches systematically. Emphasising that all 20 networks performed well for the experiments conducted in the original publications, we implemented these networks and investigated their performance on our datasets. Our results show that although each network was trained on the synthetic data and patient data only few could generalize, i.e. adapt to the new data and produce good registration results compared to a conventional affine registration algorithm and an inhouse developed deep learning approach.

Table 3.3 lists the datasets that were used in the original papers. The networks were developed specifically for these datasets. In our experiments, the networks were trained and evaluated with two new datasets that differ in modality from the datasets used in the original papers. For instance, [96] used monomodal T1-weighted MR images, in our experiments multimodal (T1-weighted MR and CT) images were registered instead. Differences in the performance of the networks in the original paper and in our experiments may therefore also be caused by the different modalities of the datasets used.

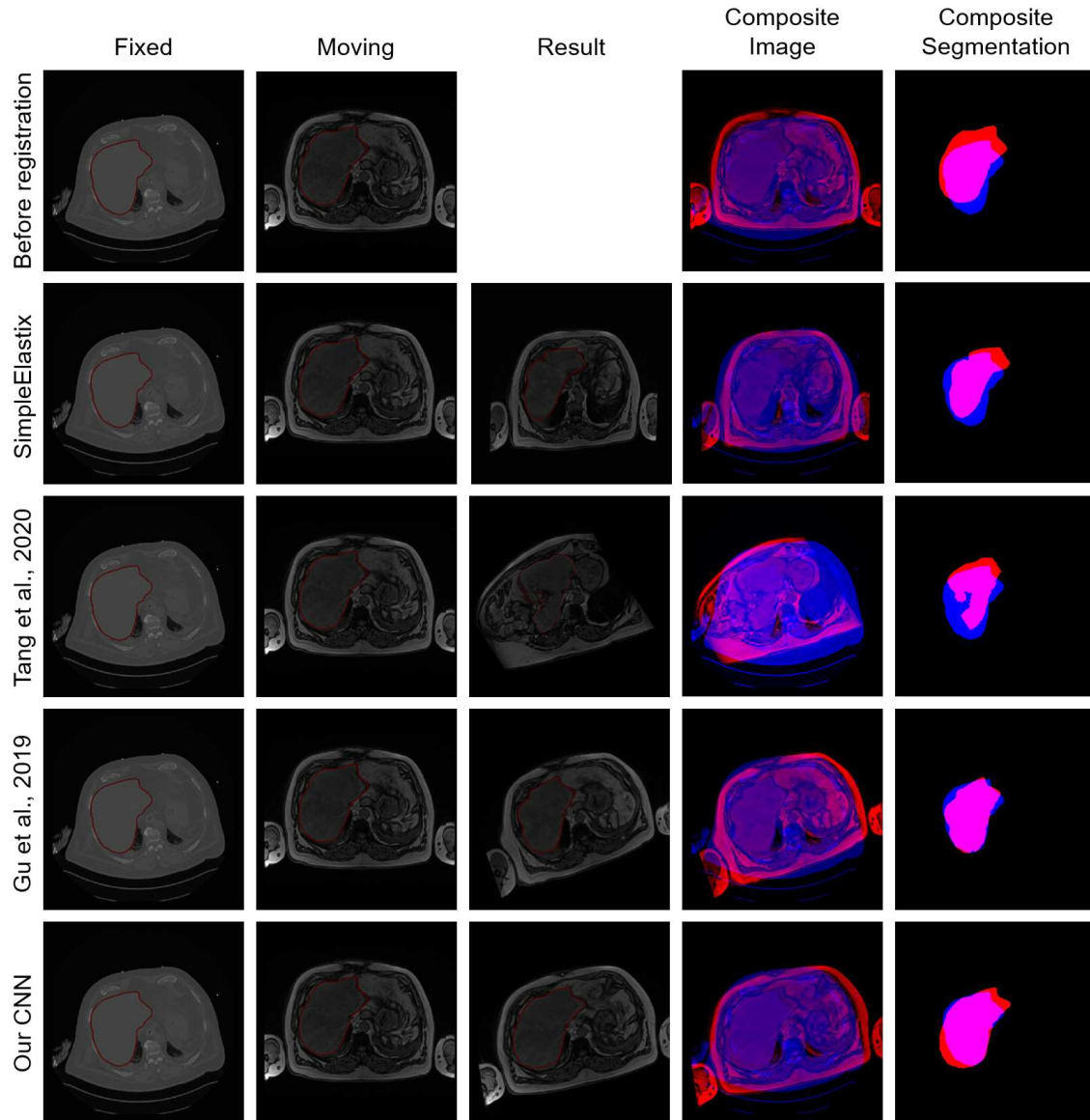


Figure 3.4: Example results of an affine registration of two volumes from the patient dataset. The images show central slices of the axial plane. The fixed images, moving images and result (moved) images are overlaid by the liver segmentations (column 1 - 3). The resulting composites for image and segmentation show the fixed data in blue and the moving/moved data in red. The third row illustrates an example of a poor registration result. Row four and five represent good registration results.

Table 3.3: Overview of the datasets used in the original papers

References	Modality	Mono-/Multimodal	Dimension	Type	Region of Interest
Hu et al., 2018 [94]	T2-weighted MR - TRUS (Transrectal ultrasound)	Multimodal	3D	in-vivo	prostate
Gu et al., 2019 [95]	T1-weighted MR - T1-weighted MR	Monomodal	3D	in-vivo	brain
Shen et al., 2019 [96]	MR - MR	Monomodal	3D	in-vivo	knee
Zhao et al., 2020 [91]	CT - CT	Monomodal	3D	in-vivo	liver
	MR - MR	Monomodal	3D	in-vivo	brain
Luo et al., 2020 [97]	ICGA (Indocyanine green angiography) - MCSL (multi-color scanning laser)	Multimodal	3D	in-vivo	eye
Tang et al., 2020 [98]	T1-weighted MR - T1-weighted MR	Monomodal	3D	in-vivo	brain
Zeng et al., 2020 [93]	MR - TRUS (Transrectal Ultrasound)	Multimodal	3D	in-vivo	prostate
Chen et al., 2021b [99]	X-ray - X-ray	Monomodal	2D	in-vivo	chest
	X-ray - X-ray	Monomodal	2D	in-vivo	hand
	CT - CT	Monomodal	3D	in-vivo	thorax
Gao et al., 2021 [100]	CT - CT	Monomodal	3D	cadaver	femur
Roelefs, 2021 [101]	CT - CT	Monomodal	3D	in-vivo	pelvic region
Zhu et al., 2021 [102]	MR - MR	Monomodal	3D	in-vivo	brain
Chee and Wu, 2018 [103]	T1-weighted MR - T1-weighted MR	Monomodal	3D	in-vivo	brain
	T1-weighted MR - T2-weighted MR	Multimodal	3D	in-vivo	brain
de Vos et al., 2019 [86]	MR - MR	Monomodal	3D	in-vivo	heart
	CT - CT	Monomodal	3D	in-vivo	chest
de Silva et al., 2020 [92]	FAF - FAF (Fundus autofluorescence)	Monomodal	2D	in-vivo	eye
	CFP - CFP (Color fundus photography)	Monomodal	2D	in-vivo	eye
	IR - IR (Infrared reflectance)	Monomodal	2D	in-vivo	eye
	CFP - FAF	Multimodal	2D	in-vivo	eye
	IR - FAF	Multimodal	2D	in-vivo	eye
	CFP - IR	Multimodal	2D	in-vivo	eye
Shao et al., 2021 [104]	MR - MR	Monomodal	2D	synthetic	prostate
	MR - Histopathology	Multimodal	2D	in-vivo	prostate
Venkata et al., 2022 [105]	T1 pre-contrast MR - T1-CE MR	Multimodal	3D	in-vivo	brain
Waldkirch, 2020 [1]	T1-weighted MR - T1-weighted MR	Monomodal	3D	in-vivo	brain
	CT - CT	Monomodal	3D	in-vivo	abdomen
	T1-weighted MR - T2-weighted MR	Multimodal	3D	in-vivo	brain
	T1-weighted MR - CT	Multimodal	3D	synthetic	abdomen
	CBCT - CT	Multimodal	3D	synthetic	abdomen
	T1-weighted MR - T1-weighted MR	Monomodal	3D	in-vivo	brain
Chen et al., 2022 [88]	T1-weighted MR - Atlas MR	Monomodal	3D	in-vivo	brain
	XCAT (extended cardiac torso) - CT	Multimodal	3D	synthetic, in-vivo	chest abdomen pelvis
Mok and Chung, 2022 [89]	T1-weighted MR - T1-weighted MR	Monomodal	3D	in-vivo	brain
	MR - MR	Monomodal	3D	in-vivo	brain
Hasenstab et al., 2019 [90]	T1-weighted MR - T1-weighted MR	Monomodal	3D	in-vivo	Liver

To allow a fair comparison, we tried to reproduce the algorithms as best as possible, however, in some papers, not all parameters for the neural networks' implementation were described. To mitigate this we also contacted the corresponding authors for feedback to our implementation or assumption made on the missing parameters, unfortunately, only little feedback was received. This is detrimental to the reproducibility and therefore, assumptions were made, that may limit possible better registration results. The performance of the models could thus potentially be influenced not only by their generalizability, but also by the reproducibility of the proposed models. However, it is difficult to separate the influence of these two aspects on the performance of the models. A first step in this direction could be to quantify the degree of reproducibility of machine learning-based methods, as described in [113].

The use of augmentation, as well as testing different normalisation techniques, learning rates and loss functions could improve the registration results. Furthermore, the batch size influences the training and affects the accuracy of the network. In the experiments in this paper, a batch size of one was used due to the memory limitation of the GPUs.

3.5.1 Synthetic Dataset versus Patient Dataset

Six networks trained with the synthetic dataset could improve the pre-registration Dice coefficient of the synthetic dataset significantly (p -value < 0.05). When evaluated with the patient dataset, no network could improve the pre-registration Dice

coefficient significantly. This is due to the fact that the networks could not generalize to the patient data. The synthetic data simulates breathing, but does not consider the fact that in real patient data the patient was positioned differently in the CT and MR scanner. The use of augmentations on the synthetic dataset could be an area of future work in order to also be able to handle positioning-related deformations. In the patient dataset, there is not only a deformation of the organs due to respiration, but also due to a different positioning of the patient in the two scanners. Finetuning of the networks with the real patient dataset was necessary here in order to also take the positioning-related deformation into account. Nine finetuned networks then improved the pre-registration Dice coefficient of the patient dataset significantly (p -value < 0.05).

Compared with the registration results of the synthetic dataset, the registration results of the patient dataset had a higher standard deviation (Table 3.2). This is due to the fact that in the synthetic dataset the respiration - i.e. the deformation - was modeled similarly in all models. There are no outliers here. However, in the patient dataset, the respiration and the additional deformation due to the different positioning of the patient in the CT and MR scanner was not similar for the different patients. For example, there were seven cases in which the patients were not lying on their backs, but rather in a more sideways position in the CT scanner, whereas they were lying on their backs in the MR scanner. For these CT-MR image pairs, the registration result was thus worse than for image pairs in which the patient was positioned similarly in the CT and MR scanner.

The synthetic dataset contains anatomical segmentation masks for the liver, whereas the segmentations for the patient dataset were created by neural networks. The performance of the networks, and the differences in the performance of the CT and MR segmentation could have a significant impact on the segmentation-based metrics, therefore also influencing the training with Dice coefficient and the evaluation with Dice coefficient and Hausdorff distance.

3.5.2 Individual Networks

For the synthetic dataset, our own developed CNN yielded better results than the best performing network from papers in terms of the Dice coefficient [95]. For the patient dataset our own developed CNN - finetuned on the patient dataset - improved the pre-registration Dice coefficient significantly, but the network did not give as good results as the best performing network from papers in terms of the Dice coefficient [95]. We developed our own CNN specifically for the synthetic dataset. Hence, this network architecture delivered good results for the synthetic dataset. If the network was then trained with other datasets (e.g. the patient dataset), the accuracy of the result decreased depending on the generalizability of the network.

While all 20 networks performed well in the experiments in the original publications and achieved good registration results, only few networks were able to generalize, i.e., adapt to the new data in our experiments. Therefore, for the networks that did not perform well in our experiments, adjustments to the network architecture, such as the number of convolutional layers and filters, are needed to improve the generalizability and performance on new data.

Our results show that larger networks with many parameters to learn do not necessarily lead to better registration results, especially when there is not enough training data to support the complexity of the network, and that networks with fewer parameters may perform better on small datasets. Several papers [92, 98, 99] also highlight the importance of preventing overfitting, especially on small datasets, by using dropout layers. However, a balance has to be found, as too much regularization can cause the network to lose information and perform poorly. A detailed discussion of the individual networks can be found in [Section 3.7.2](#).

3.5.3 Individual Metrics

Some networks improved only one of the two segmentation-based metrics (Dice coefficient or Hausdorff distance), not both at the same time. The reason for this may be that the Hausdorff distance is more sensitive to outliers than the Dice coefficient.

It is also important to note that the value of the evaluation metric is biased towards the loss function used during the optimisation. As baseline SimpleElastix was applied using the default settings for an affine transformation. The similarity metric Mutual Information was optimised, as there is no Dice coefficient similarity metric in SimpleElastix. The Normalized Mutual Information of the data registered with SimpleElastix was therefore relatively high. The neural networks, on the other hand, were trained with the Dice loss. The Dice coefficient of the data registered with a neural network trained with Dice loss was higher than if the neural networks were trained with the mutual information loss. The registration results of the baseline and the neural networks therefore cannot be compared directly in terms of value of the evaluation metrics.

By optical evaluation, we verified that the structural similarity of a network improved as the Dice coefficient increased through registration. However, this was not always reflected by the [NMI](#), which decreased for some networks after registration. One reason for this could be that the [NMI](#) is calculated globally, i.e. over the entire image. It is therefore not able to take spatial information into account. The intensity distribution is calculated over the entire image, the distance of the voxels with the same intensities is not considered. A locally calculated [NMI](#) would therefore be suitable here. This way, only the voxels that are close to each other would be taken into account [114].

As demonstrated with the synthetic dataset in [Figure 3.3](#), structural similarity could improve, i.e. overlap of the liver increased, but the overlap of the abdomens in total deteriorated. In the experiments of this paper, only affine registration was used, which is limited in its transformation. Due to respiration, movement of the patient and a different positioning of the patient in the [CT](#) and [MR](#) scanner, the abdomen is transformed deformable. In this case, an affine registration is not sufficient to spatially align both the entire abdomen in total and the liver at the same time. For this purpose, a deformable transformation should be used to further improve the spatial alignment of the images. Future work will focus on further transformations, e.g. deformable registration.

3.6 Conclusion

Although various machine learning algorithms for affine multimodal image registration are proposed, we could demonstrate that few can generalize to the new data and application in our experiments. More work should be invested in developing more general applicable image registration techniques to be able to be used in minimal-invasive image guided interventions to support precision medicine in cancer treatment.

3.7 Appendix

3.7.1 Example Results of an Affine Registration for All Implemented Networks

See [Figure 3.5](#), [Figure 3.6](#).

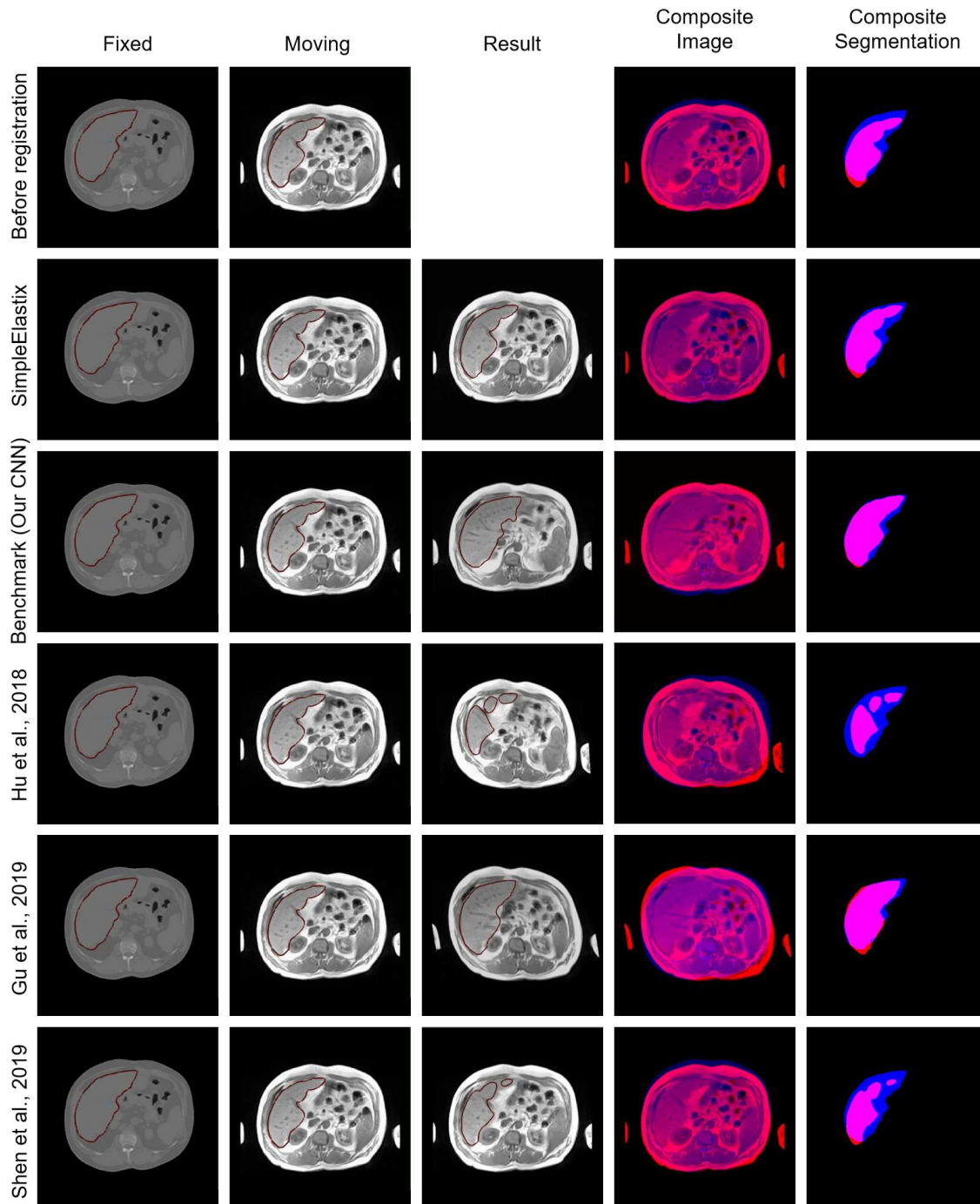


Figure 3.5: Example results of an affine registration of two volumes from the synthetic dataset. The images show central slices of the axial plane. The fixed images, moving images and result (moved) images are overlaid by the liver segmentations (column 1 - 3). The resulting composites for image and segmentation show the fixed data in blue and the moving/moved data in red.

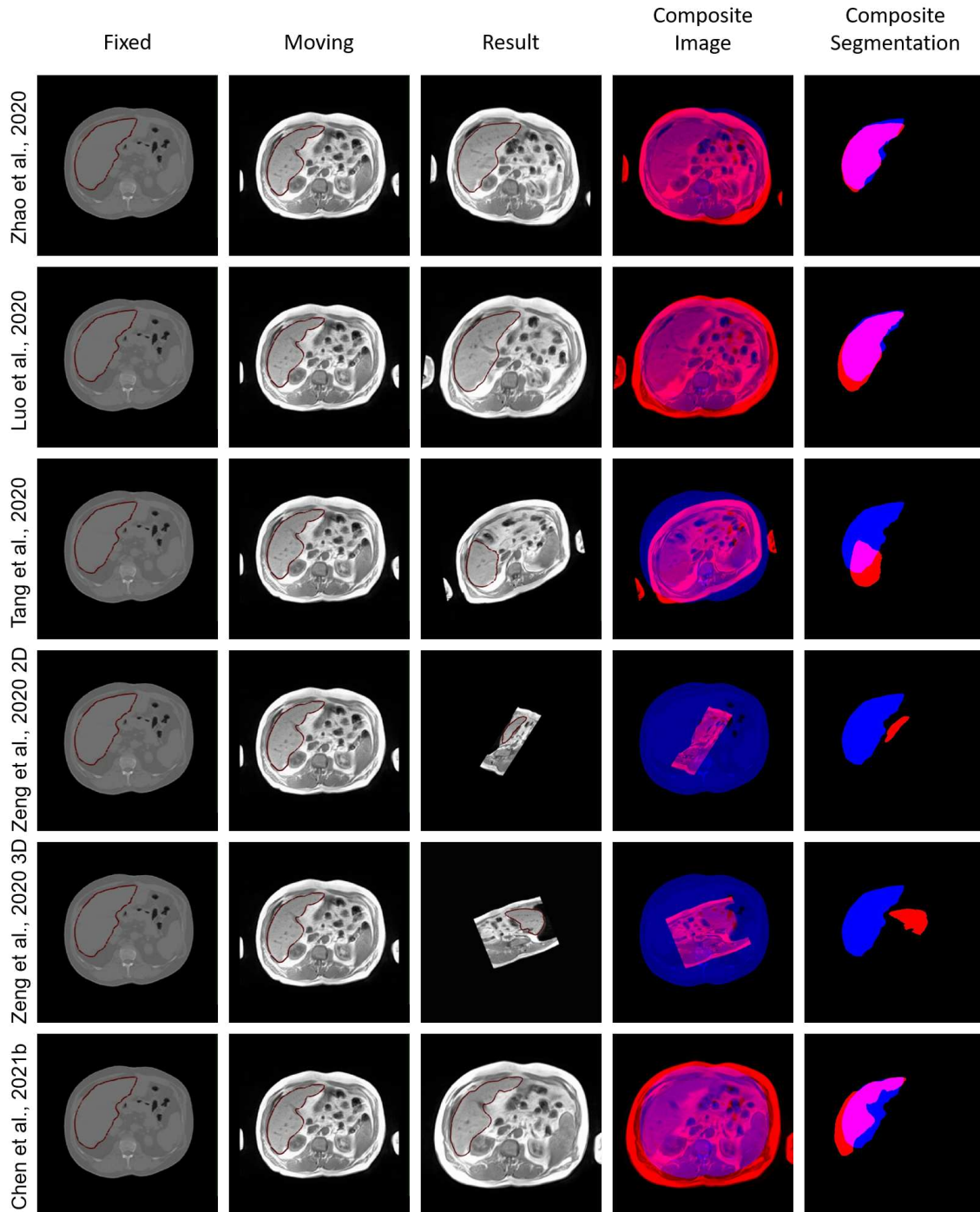


Figure 3.5: Example results of an affine registration of two volumes from the synthetic dataset. The images show central slices of the axial plane. The fixed images, moving images and result (moved) images are overlaid by the liver segmentations (column 1 - 3). The resulting composites for image and segmentation show the fixed data in blue and the moving/moved data in red.

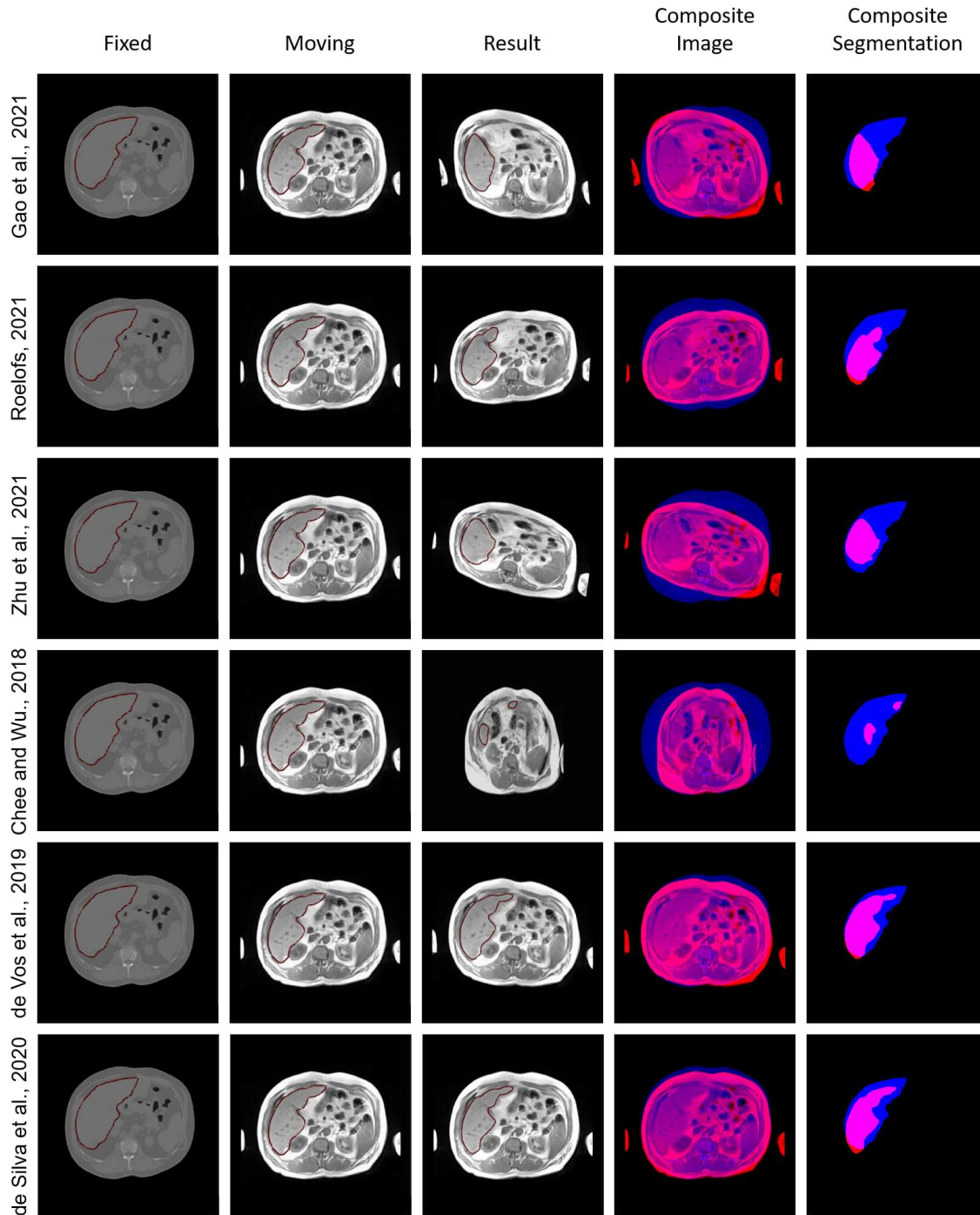


Figure 3.5: Example results of an affine registration of two volumes from the synthetic dataset. The images show central slices of the axial plane. The fixed images, moving images and result (moved) images are overlaid by the liver segmentations (column 1 - 3). The resulting composites for image and segmentation show the fixed data in blue and the moving/moved data in red.

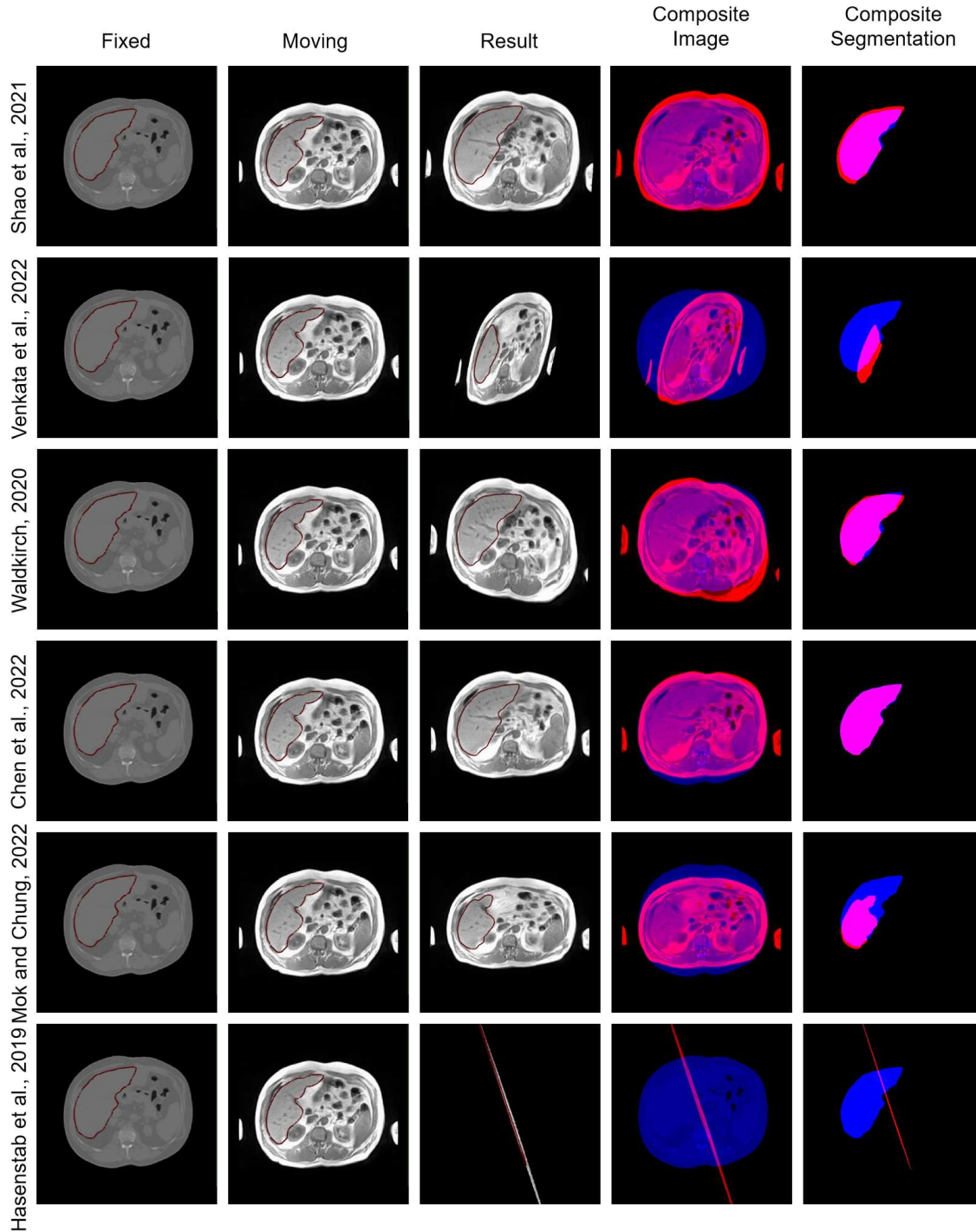


Figure 3.5: Example results of an affine registration of two volumes from the synthetic dataset. The images show central slices of the axial plane. The fixed images, moving images and result (moved) images are overlaid by the liver segmentations (column 1 - 3). The resulting composites for image and segmentation show the fixed data in blue and the moving/moved data in red.

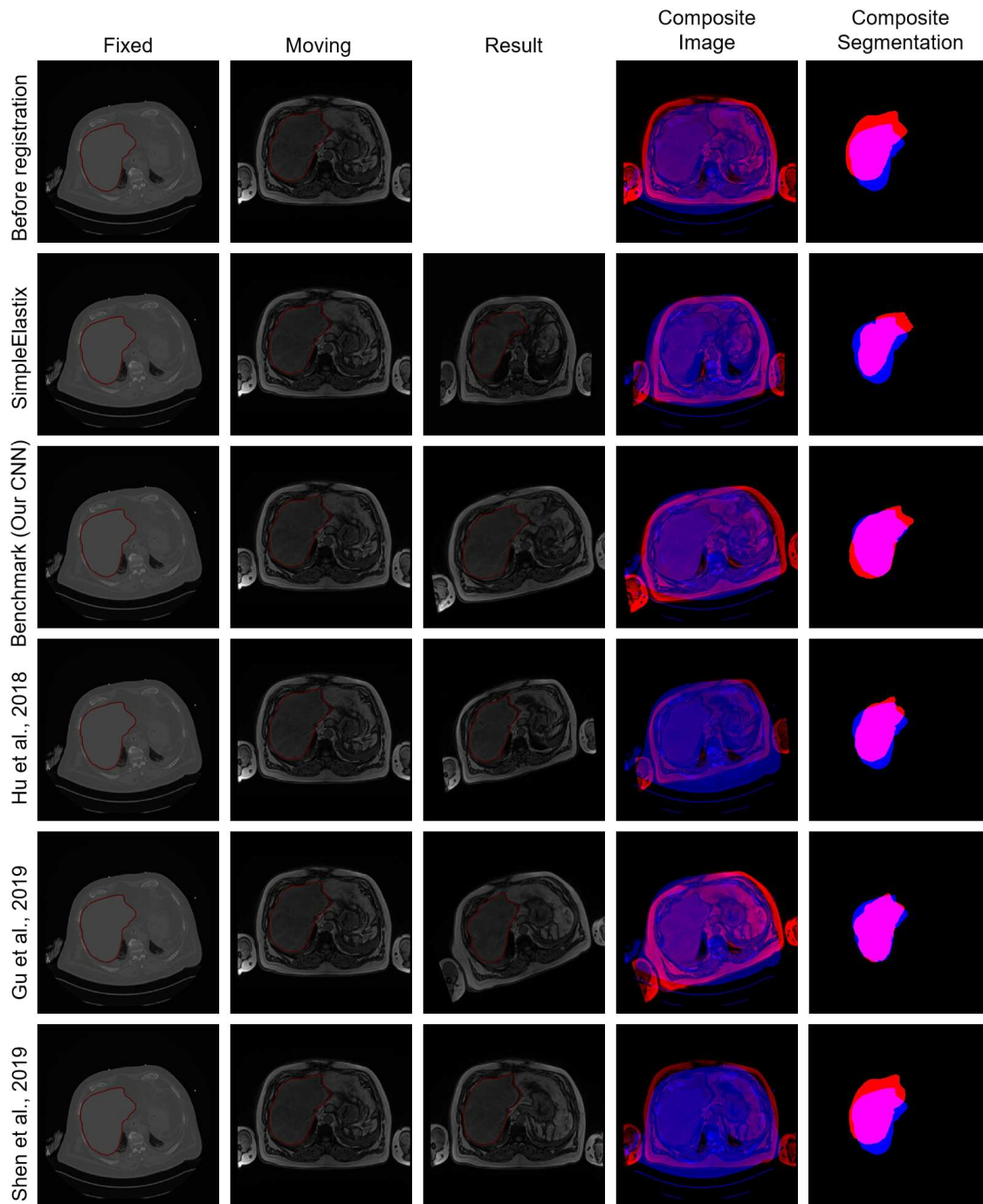


Figure 3.6: Example results of an affine registration of two volumes from the patient dataset. The images show central slices of the axial plane. The fixed images, moving images and result (moved) images are overlaid by the liver segmentations (column 1 - 3). The resulting composites for image and segmentation show the fixed data in blue and the moving/moved data in red.

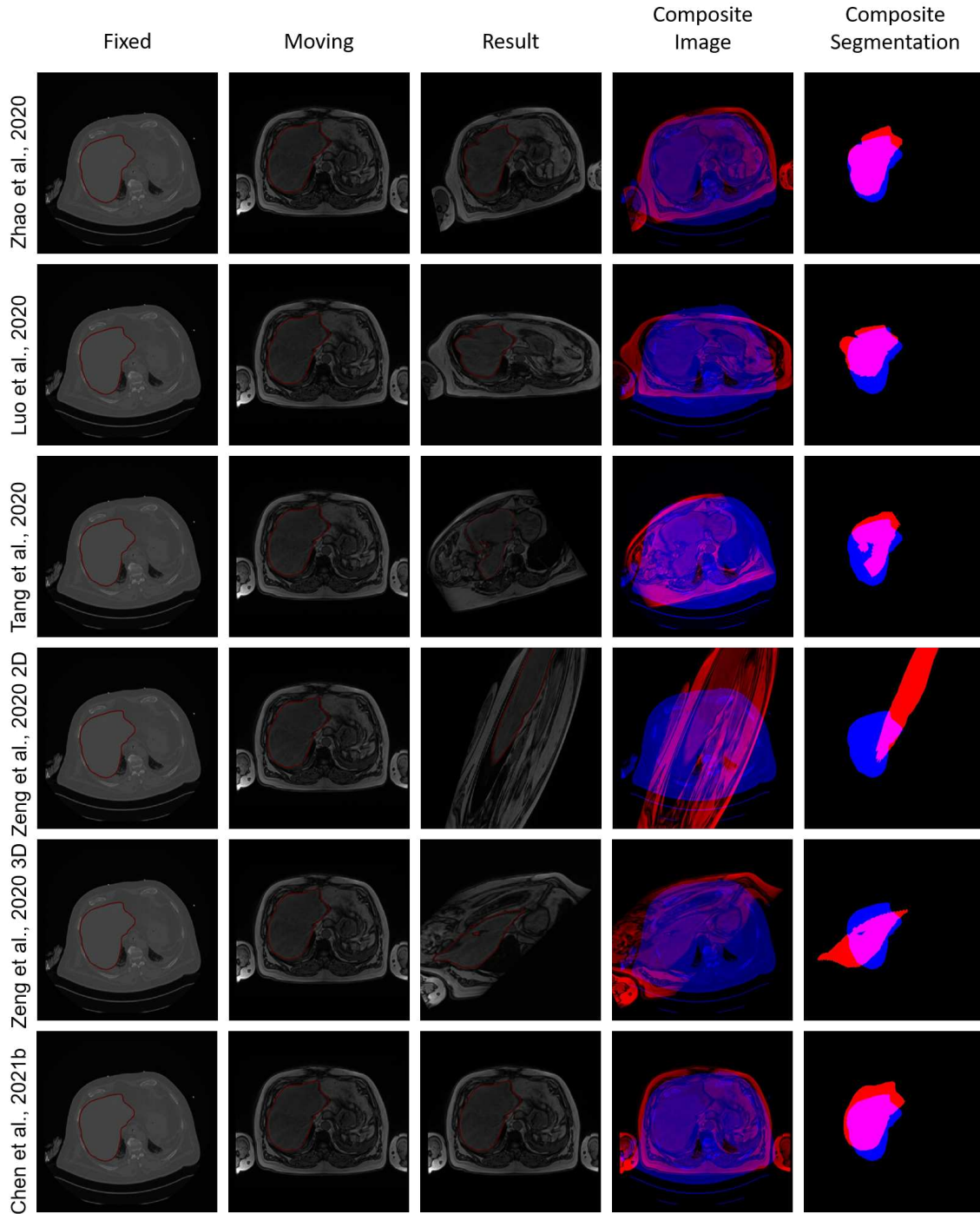


Figure 3.6: Example results of an affine registration of two volumes from the patient dataset. The images show central slices of the axial plane. The fixed images, moving images and result (moved) images are overlaid by the liver segmentations (column 1 - 3). The resulting composites for image and segmentation show the fixed data in blue and the moving/moved data in red.

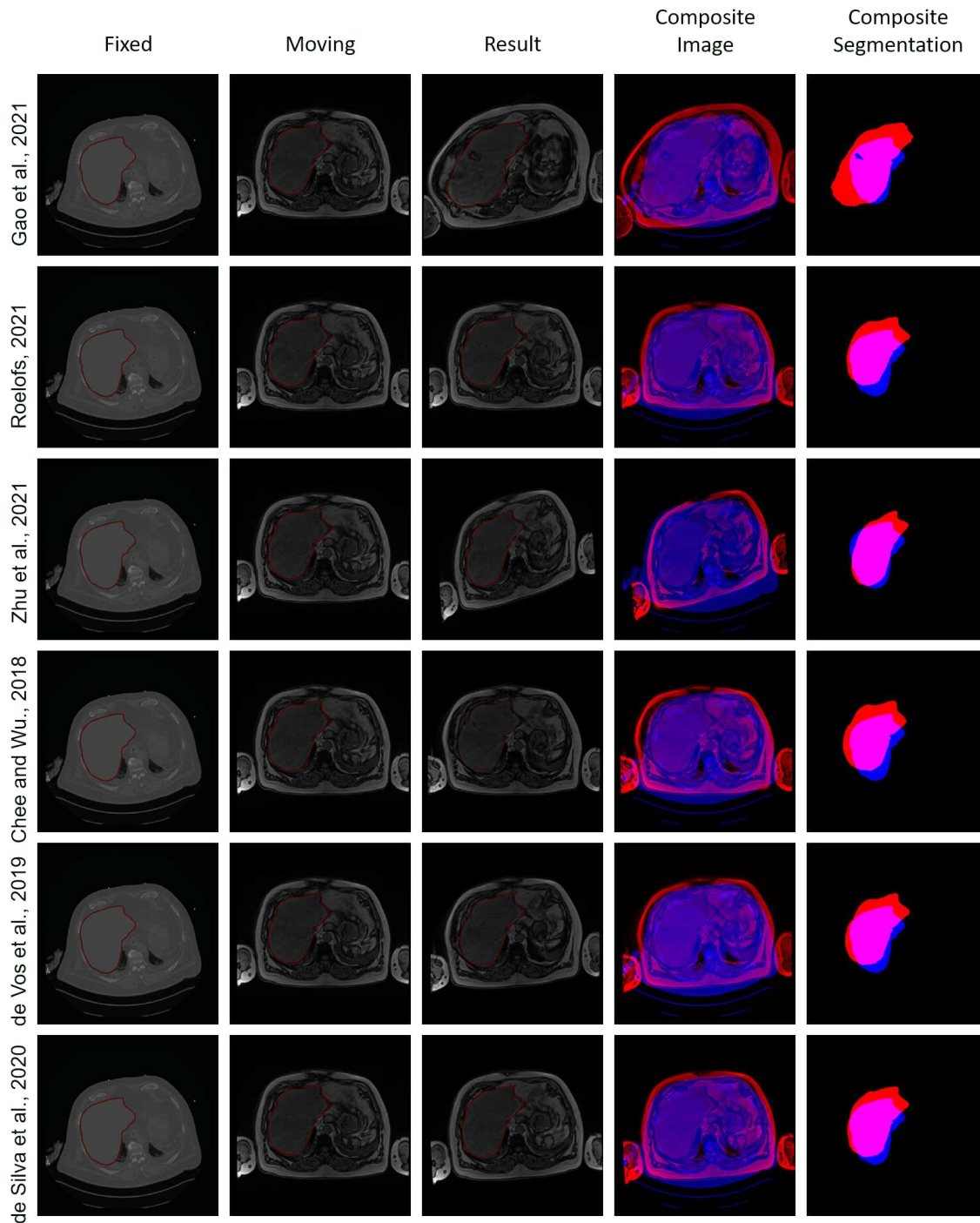


Figure 3.6: Example results of an affine registration of two volumes from the patient dataset. The images show central slices of the axial plane. The fixed images, moving images and result (moved) images are overlaid by the liver segmentations (column 1 - 3). The resulting composites for image and segmentation show the fixed data in blue and the moving/moved data in red.

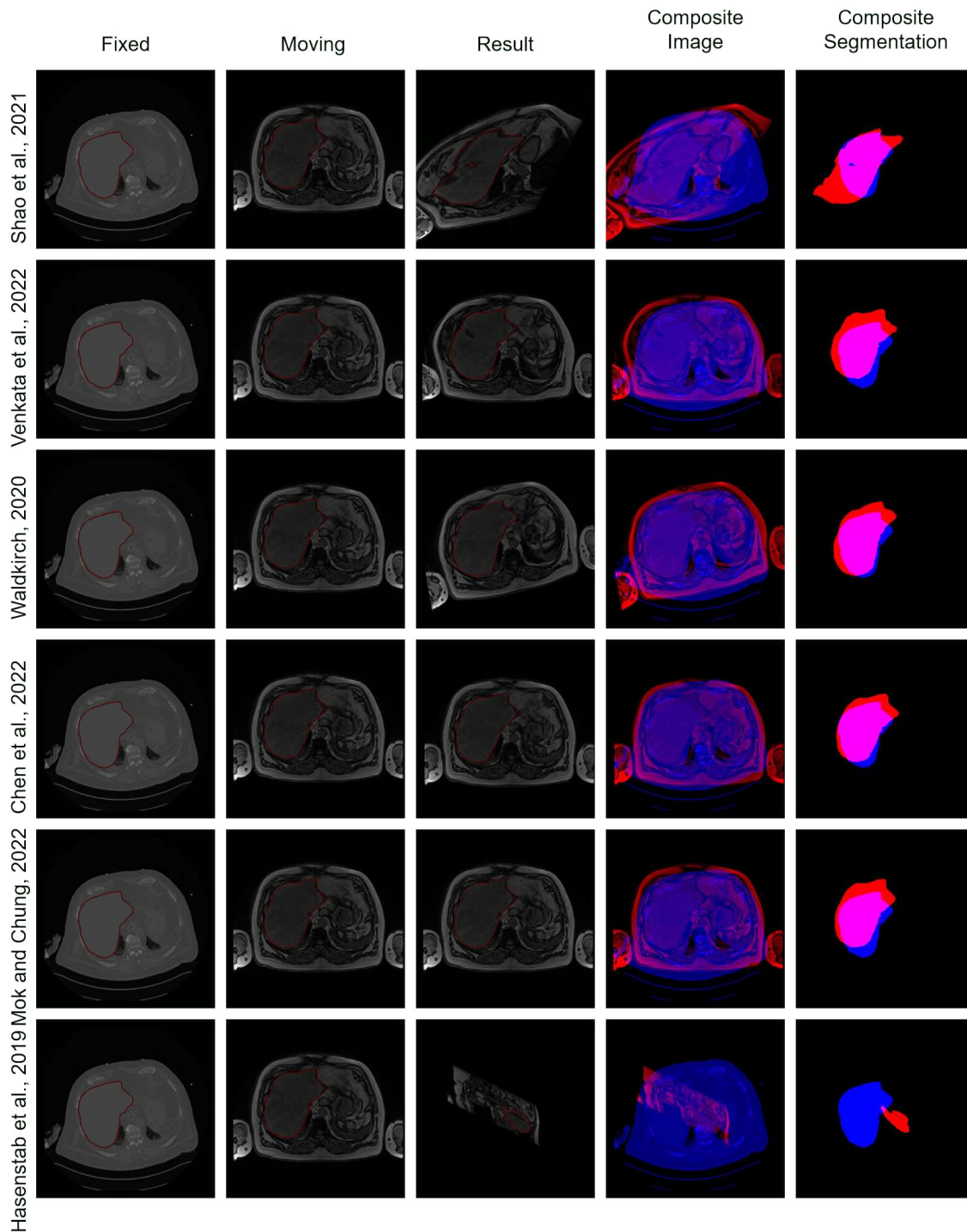


Figure 3.6: Example results of an affine registration of two volumes from the patient dataset. The images show central slices of the axial plane. The fixed images, moving images and result (moved) images are overlaid by the liver segmentations (column 1 - 3). The resulting composites for image and segmentation show the fixed data in blue and the moving/moved data in red.

3.7.2 Discussion of the Individual Networks

In [94] DownBlocks consisting of convolutional layers and residual blocks are applied. The convolution-based network achieved good registration results when applied to the dataset used in [94] (T2-weighted MR volumes and transrectal ultrasound images of prostate cancer patients). However, the results presented in this paper (as seen in Table 3.2) indicate that the network developed by [94] lacks generalizability when applied to new datasets. Therefore, further adjustments to the network architecture, such as the number of DownBlocks, are necessary in order to achieve good registration results on new datasets as well.

The network of [95] uses a U-Net-like encoder part and a fully-connected layer to create the 12 parameters for the affine transformation. The network architecture can be classified as convolution-based network. The network of [95] yielded the best registration results, i.e. the best Dice coefficient, in all 3 cases: 1) network trained and evaluated on synthetic dataset; 2) network trained on synthetic dataset and evaluated with patient dataset; 3) network trained on synthetic dataset and finetuned and evaluated with patient dataset. This shows that the network generalizes to the synthetic and patient dataset well. A high degree of generalizability was already ensured by [95] during the development of the network by testing and evaluating the developed framework using four publicly available in-vivo datasets [95].

An affine convolution-based network structure with multiple steps is applied in [96]. Comparable to a recurrent network, the parameters of the affine network from all steps are identical. The number of steps of the network influences the performance of the network and could lead to poor generalizability of the entire network. In addition to modifying the individual layers in the network, the number of steps could also be adjusted in order to improve the generalizability of the network and its performance on new datasets.

[91] used a convolution-based network architecture similar to the one used by [95], but with a different number of filters for each convolutional layer and a different number of convolutional layers per convolutional block. This parameter variation can change the representational power of the network, enabling it to learn more complex features. On the other hand, this can also increase the risk of overfitting. [91], [95] found the right balance between these factors and their networks thus achieved good registration results for the datasets used in our experiments.

In [97] a coarse and a fine registration stage are used. The coarse scale calculates anatomical segmentations and a coarse transformation (translation and scaling), which is used to reduce large global image displacements. The fine scale corrects smaller affine image displacements. However, no implementation details of the U-Net were given in [97]. Only the fine convolution-based stage was described completely and could be implemented. If the coarse stage is added, the registration result for the synthetic dataset and the patient dataset would probably be improved.

The network of [98] can be categorised as convolution-based. [98] used dropout layers as final layers of their neural network. Dropout layers are applied to prevent overfitting, which is particularly important for small datasets such as our synthetic

dataset and patient dataset. On the other hand, the use of dropout layers or a too high dropout rate can cause the network to lose information. That occurred while training the network of [98] with the synthetic dataset and the patient dataset and has led to a poor registration result.

[93] also developed a convolution-based network. The adapted three-dimensional network of [93] yielded better registration results than the original two-dimensional version. However, both networks worsened the NMI, Dice coefficient and Hausdorff distance compared to the unregistered images, despite using an architecture resembling the best performing network in terms of Dice coefficient [95]. This could be due to the number of filters for each convolutional layer being too high. The first convolution in the network of [93] already contains 128 filters. The network of [95] starts with only 4 filters in the first convolutional layer. Especially for small datasets, it might happen that for large networks with many parameters to be learned, there is not enough training data available to support the complexity of the network and a model with fewer parameters performs better.

Cross-stitch units are applied in the convolution-based network of [99] to learn the best combination of the outputs of the previous layers. However, the registration results in Table 3.2 show that the cross-stitch units might be inadequate for small datasets and a simpler architecture without cross-stitch units could achieve better registration results. To overcome this problem, caused by small training datasets, [99] used 6,000 pairs of synthetic CT scans (2,400 pairs for training) as initialisation of the three-dimensional network before training the network on a small dataset [115, 116] consisting of 60 three-dimensional CT volumes [99].

[100] also applied a convolution-based network with an architecture similar to the best performing network in terms of Dice coefficient [95], using a different number of filters for each convolutional layer and a different number of convolutional blocks. However, these parameters are set in the network of [100] in such a way that the network yielded good registration results for the dataset used in their work, but did not generalize well and achieved poor registration results for the datasets used in this paper. In order to achieve a higher generalizability, the parameters aforementioned have to be adjusted.

In his master's thesis, [101] evaluated different ResNet architectures for affine registration of medical images. The best results were obtained with ResNet-18, therefore ResNet-18 was examined in this paper. The skip connections used in ResNets overcome the vanishing gradient problem and considerably improve the overall performance of neural networks comprising a substantial number of layers. For this reason, the convolution-based network used in [101] (ResNet-18) was able to improve the Dice coefficient of the patient dataset (network finetuned with the patient dataset).

In [102] a joint affine and deformable network is used. In the implementation of the network for this work, only the affine part was used, which can be categorised as convolution-based network. However, the affine and deformable subnetworks are not independent of each other but are linked through parameter sharing of the affine network and the downsample part of the deformable network. Training the joint affine and deformable network therefore also affects the predicted affine transformation matrix and could improve the affine registration result.

[103] used an Siamese network-based encoder which, in addition to convolutional layers and pooling layers, also uses dense blocks. In dense blocks, skip connections are applied between convolutional layers to combine features from different levels. [103] evaluated their network on a single dataset (three-dimensional MR images of the human brain). For this dataset, the network achieved good registration results. However, poor results were obtained for new datasets due to weak generalizability. The network architecture, for example the number of dense blocks, could be further adjusted to improve generalizability to new datasets.

As first part of their network, [86] used two distinct pipelines for fixed and moving image to enable inputting image pairs into the network that do not have the same volume size without prior resizing. Weight sharing is applied between the two pipelines. The network architecture can therefore be categorised as a Siamese network-based. Weight sharing results in a less complex model with a reduced number of parameters and enhances its generalizability. The network also achieved good registration results for new datasets and increased the Dice coefficient significantly for the synthetic dataset and the patient dataset (Table 3.2).

The affine network of [92] consists of feature extraction, matching and affine regression network. A correlation matrix was used as matching part, which has the advantage over concatenating or subtracting the CNN features that it improves the generalizability and performance of the network [117]. The affine network of [92] can be categorised as Siamese network-based. The network provided good registration results for the data used in [92] (fundus autofluorescence, color fundus photography and infrared reflectance images, see Table 3.3). However, the entire network architecture only generalized well to the patient dataset (network trained and evaluated with the patient dataset) and generalized poorly to the synthetic dataset and thus produced poor registration results for the synthetic dataset. To improve generalizability for the synthetic dataset, the architecture of the feature extraction network and the affine regression network could be further modified.

In [104] a similar network structure is applied as in [92], i.e. feature extraction, matching and regression network. The network architectures of [104] and [92] are classified as Siamese network-based due to the Siamese structure of their feature extractors. Both networks are based on the network architecture of [117]. But [104] applied a ResNet-101 as feature extractor, while [92] used a VGG-16 network. This difference in the feature extractor might have resulted in the network of [104] achieving a good registration result for the synthetic and patient dataset, while the network of [92] could not improve the registration of the synthetic dataset.

[105] used a modified Siamese network-based architecture of [103]. In [103], an encoder is used whose first part includes convolutional layers with two-dimensional filters and the second part of the encoder applies three-dimensional filters. This structure is used to handle that the volume of a three-dimensional medical image is usually not cube-shaped, but has a lesser depth than height and width [103]. [105], on the other hand, used three-dimensional filters for all convolutional layers. As shown in Table 3.2, this change in the network architecture had no significant effect on the Dice coefficient of the registration result for the synthetic dataset and the patient dataset.

[1] used a U-Net architecture and a global average pooling layer. The U-Net architecture has demonstrated high accuracy in biomedical image segmentation [67]. The U-Net’s high generalizability also allows it to be applied to other image-related tasks, e.g. it is already often used for deformable image registration, and achieved good affine registration results for the datasets used in this paper.

In [88] the swin transformer is used, which combines transformer and convolutions. Transformers provide the advantage over convolutions that the receptive field of the transformer is as large as the image, i.e. it is not restricted to a small area as in convolutions. Spatial relationships in the two images can therefore be better investigated and understood [88]. The network of [88] (trained on the synthetic dataset as well as finetuned with the patient dataset) was therefore able to improve the spatial alignment (based on the segmentation-based metrics) of the synthetic dataset and the patient dataset significantly.

The C2FViT transformer of [89] is a modification of the Vision Transformer [118]. Vision Transformer already yielded good results in the area of image classification, but has the drawback that training with a large dataset is necessary [118]. However, our in-house synthetic dataset and patient dataset are relatively small. This led to a poor performance of the network in the experiments. For larger datasets, the network will probably achieve better registration results.

The neural network of [90] achieved the worst registration results. This is due to the fact that the network consists of only one single dense layer with 12 parameters. The network is therefore classified as "Other Network Architecture" in Table 3.1, column "Category according to Figure 3.1". Convolutions or transformers were used for all other implemented networks. Convolutions have the advantage over dense layers in that they have fewer parameters to be learned during training, and the kernels are translationally-invariant and can recognise local patterns. Applying convolutions will lead to an improvement of the efficiency and performance of the network [119].

3.7.3 Manual for the Published Source Code

Our source code is available at https://github.com/Computer-Assisted-Clinical-Medicine/Networks_for_Affine_Medical_Image_Registration.

Usage:

1. In config.py:
 - (a) Add the paths to your data: path to your fixed and moving images and segmentations (lines 10-13).
 - (b) Enter the path where the results should be saved (line 14) and the name of the output folder (line 17).
 - (c) If necessary, also specify the path to the pre-trained weights (line 15).
 - (d) Specify the path to the csv files containing the filenames of the fixed and moving images and segmentations (line 22, 27, 32, 37).
2. Run main.py with the following settings
 - (a) To train a network from scratch: set `is_training = True` (line 384). After training, the weights of each fold of the five-fold cross-validation are stored in subfolders in the output folder (numbered from 0 to 4).
 - (b) To finetune a network: set `is_pretrained = True` (line 385). In config.py, specify the path to the pre-trained weights (line 15). This folder should contain a separate folder for each of the individual folds of the five-fold cross-validation (numbered from 0 to 4). These subfolders contain the pre-trained weights for the respective fold in a file named "weights.h5". After finetuning, the weights of each fold of the five-fold cross-validation are stored in subfolders in the output folder (numbered from 0 to 4).
 - (c) For inference and evaluation, set `is_apply = True` and `is_evaluate = True` (line 386 and 387). In the output folder, the subfolders "predict" and "eval" are created. The predict folder contains the registered moved images and segmentations. The evaluation folder contains the Normalized Mutual Information, Dice coefficient and Hausdorff distance values for each registered image pair. Additionally, boxplots for the three metrics and individual folds of the five-fold cross-validation are stored in "eval/plots".
 - (d) The network architecture can be selected in line 401. The architectures can be found in `NetworkBasis.networks.py`.

4. A Multistage Rigid-Affine-Deformable Network for Three-Dimensional Multimodal Medical Image Registration

This chapter is based on the following publication: A. Strittmatter, A. Caroli, F. G. Zöllner. A Multistage Rigid-Affine-Deformable Network for Three-Dimensional Multimodal Medical Image Registration. Applied Sciences, 13(24), 13298, doi: 10.3390/app132413298, 2023.

4.1 Introduction

Multimodal image registration is a vital research field in medical image processing as it allows the integration of complementary information from various imaging modalities. This fusion can aid numerous clinical applications, such as improved diagnosis and treatment planning [2–4].

For instance, medical image registration is of great importance for the diagnosis and monitoring of renal diseases [5–7]. Among them, autosomal dominant polycystic kidney disease ([Autosomal Dominant Polycystic Kidney Disease \(ADPKD\)](#)) is a chronic hereditary disorder that slowly leads to renal enlargement and eventually to end-stage renal disease ([End-stage Renal Disease \(ESRD\)](#)). Early detection and continuous monitoring of [ADPKD](#) progression are crucial for preventing or delaying [ESRD](#). Medical image registration can be used to combine information from multiple imaging modalities, which can enhance the accuracy of [ADPKD](#) diagnosis and monitoring. Here, the total kidney volume ([Total Kidney Volume \(TKV\)](#)) and the volumes and numbers of cysts are important biomarkers used to evaluate the progression of [ADPKD](#) [8]. Specifically, T1-weighted [MR](#) images are used to generate kidney segmentations to calculate the total kidney volume while T2-weighted [MR](#) images are used to create cyst segmentations to determine the cyst volume and the number of cysts. By fusing the images of the two [MR](#) modalities and thereby combining complementary information, the determination of the two biomarkers can be significantly improved [9]. Furthermore, images from multiple time points can be fused, e.g., images before, during and after a therapy, to monitor the therapy’s progress. For this, registration of the images is necessary.

Another example of a clinical application that benefits from image registration is the treatment of oligometastases disease, which has poor outcomes with standard therapies due to the local lesion heterogeneity. Personalized image-guided therapies that consider tumor heterogeneity have shown potential in improving survival rates [12, 13]. Typically, [CT](#) imaging is used for these interventions, while [MRI](#) is utilized to characterize tumor heterogeneity, which is often invisible in [CT](#) scans [14]. Differences in patient positioning in the two scanners, as well as the patient’s respiration, result in the organs being deformed differently in the corresponding [CT](#)

and MR images. Therefore, prior to fusion, the registration of CT and MR images is inevitable.

Traditionally, medical image registration is performed using conventional methods that involve iteratively updating the transformation parameters to optimize the similarity metrics. However, this approach is often time-consuming as each new image must be re-registered iteratively. To solve this problem, deep learning-based registration methods have been developed to speed up the registration process. In these methods, the network is trained once and subsequent registrations can be performed by passing the new images through the network, eliminating the need for time-consuming iterative methods [16].

In recent years, several neural networks for medical image registration have been published with a wide variety of architectures. These include, e.g., the use of multi-stage networks, in which different subnetworks are used [76, 86, 91, 93–96, 98, 100–102, 104, 120–123]. In these approaches, an initial affine transformation is combined with a subsequent deformable transformation within a neural network in order to first reduce large global image displacements by the affine transformation subnetwork and then to correct smaller local image displacements by the deformable transformation subnetwork.

In medical image registration, image deformation also consists of larger rigid parts, for example, due to different patient positioning in different scanners or acquisitions. For instance when planning and performing an image-guided biopsy as in [71], large rotations occur if the patient is supine during a preliminary examination but lateral during the image-guided biopsy due to needle placement. In addition, rigid structures such as bones or tumors that behave rigidly [74] are only or primarily deformed rigidly. Initially, a rigid transformation could be used to roughly reduce the large rigid displacements before correcting affine and deformable deformations. For instance, the NiftyReg algorithm `reg_aladin` [19] first applies rigid and then affine registration. In combination with the algorithm NiftyReg `reg_f3d` [18], a multistage registration with first rigid, then affine and subsequently deformable transformation is obtained.

However, to the best of our knowledge, no multistage network has yet been published that combines the three transformations rigid, affine and deformable. For this reason, we developed a multistage network for jointly rigid, affine and deformable transformation. This network can be trained end-to-end.

4.2 Background

The aim of image registration is to compute a transformation to spatially align two images, where the source (moving) image is registered to the reference (fixed) image. The computed transformation is then applied to the moving image to create the moved image. In medical image registration, three basic transformation types are applied: rigid, affine, and deformable. A rigid transformation includes translation and rotation and is represented by:

$$T_{rigid}(x) = Rx + t \quad (4.1)$$

where R describes the rotation matrix, x a point and t the translation vector. For three-dimensional images, the rotation matrix and the translation vector require three parameters each. Thus, six parameters have to be calculated for a rigid transformation for three-dimensional images.

The affine transformation is composed of translation, rotation, scaling, and shearing. Matrix multiplication is used to merge the single transformations, as described by [87]. For three-dimensional images, the affine transformation is represented by a 4×4 matrix that contains 12 parameters, including the translation parameters in the right column (parameters a_{14}, a_{24}, a_{34}), while the other nine parameters represent the combined values of rotation, scaling, and shearing.

$$T_{affine} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (4.2)$$

Nonlinear (deformable) transformations are used to model local deformations in images that rigid and affine transformations cannot capture. These transformations are complex and do not preserve straightness or parallelism. The transformation is represented by a deformation (displacement) field ϕ :

$$T_{deformable}(x) = x + \phi(x) \quad (4.3)$$

Since deformable image registration is an ill-posed problem, regularization is necessary to ensure a smooth and plausible deformation field, e.g. by using penalty terms during the optimization of the transformation (iterative methods) or training of the registration network (deep learning-based) [1].

Conventional registration methods such as SimpleElastix [17] or NiftyReg [18, 19] rely on solving a mathematical optimization model separately for each pair of images, which can be a slow and computationally intensive process. This involves iteratively maximizing a function that measures the similarity between the fixed and moved images. While some conventional methods can be accelerated using GPU implementations, this still requires a GPU for each registration process [80].

In contrast, neural networks offer a faster and more efficient approach to image registration. Instead of solving an optimization problem for each new pair of images, a neural network takes two n -dimensional images as input and outputs the transformation in a single forward pass through the neural network [16]. The weights of the neural network are optimized through training, resulting in a globally optimized function that can be used for the registration of any test image pair without requiring an iterative optimization process. This approach saves time both on the CPU and GPU compared to conventional methods [16].

4.3 Methods

4.3.1 Network Architectures

We developed a multistage rigid-affine-deformable neural network that comprises three subnetworks, namely a rigid, affine, and deformable registration network.

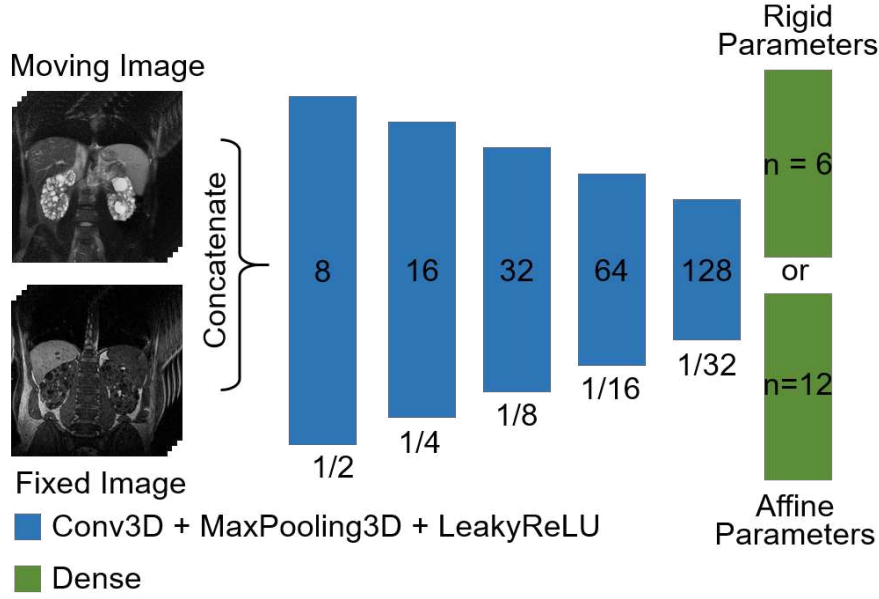


Figure 4.1: Architecture of the rigid and affine subnetworks. Rigid subnetwork: 6 neurons compute 6 parameters for rigid transformation. Affine subnetwork: dense layer computes 12 parameters for affine transformation

We used a modified variant of the affine network proposed by [95], since it achieved the best registration results in our datasets as demonstrated previously [15]. The network architecture was modified to allow for a rigid transformation in addition to the original affine transformation. The rigid and affine subnetwork is composed of five convolutional blocks (Figure 4.1). Each contains a three-dimensional convolutional layer with a kernel size of $3 \times 3 \times 3$, a stride of 1 and padding same, a maxpooling layer with a factor of 2 and a LeakyReLU activation layer. A dense layer is then used to calculate the 6 or 12 parameters required for a three-dimensional rigid or affine transformation.

The third subnetwork network (Figure 4.2a) performs a deformable transformation. Here, we used the VoxelMorph-2 network [16, 124].

To form the aforementioned rigid-affine-deformable multistage network (see Figure 4.2a), the individual subnetworks are connected in series. The rigid subnetwork calculates a rigid transformation, which is then applied to the moving image with a spatial transformer [106]. The result image is used as input for the affine transformation subnetwork. The calculated affine transformation is then combined with the rigid transformation. The result transformation is then applied to the original moving image and the result image then serves as an input for the deformable subnetwork. The deformation field calculated by the deformable subnetwork is then combined with the rigid and affine transformation. The result transformation is applied to the moving image using the spatial transformer. This result image is the final moved image in our proposed multistage rigid-affine-deformable network (Figure 4.2a).

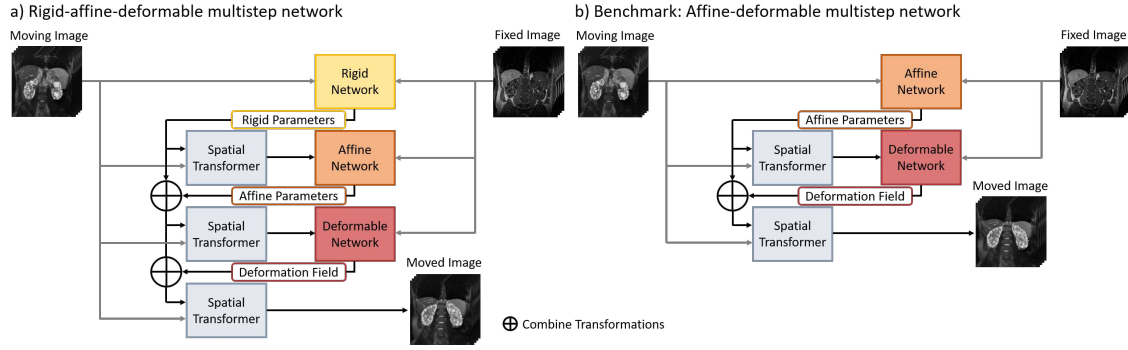


Figure 4.2: (a) Architecture of the multistage network with rigid, affine and deformable subnetworks. The individual subnetworks are connected in series. The rigid, affine and deformable transformation is combined and then applied to the moving image by a spatial transformer [106]. This result image is the final moved image. (b) Architecture of the benchmark with affine and deformable subnetworks

To facilitate the combination of the different single transformations, we calculated the deformation field (vectors) from the rigid and affine transformations. Two transformations can then be combined by adding their deformation fields.

4.3.2 Benchmark Neural Network

As a benchmark, we used a neural network with two subnetworks (Figure 4.2b). The first subnetwork performs an affine transformation, while the second subnetwork a deformable transformation. For the affine and deformable subnetworks, we used the same architectures as for the affine and deformable subnetworks from our multistage rigid-affine-deformable network. The difference between our proposed multistage rigid-affine-deformable network and the benchmark network therefore lies in a preceding rigid subnetwork. Thus, in our experiments, we examined the influence of a preceding rigid subnetwork on the registration result.

4.3.3 Training Setting for Neural Networks

We trained each subnetwork separately while keeping the weights of the preceding stages fixed. For our proposed multistage network, we first trained the rigid subnetwork and then added the affine subnetwork. We froze the layers of the rigid subnetwork so that only the weights of the affine subnetworks were updated during the training of the multistage rigid-affine network. After training, we added the deformable subnetwork and trained this subnetwork while preserving the weight of the preceding stages. We then unfroze the layers of all subnetworks and finetuned all subnetworks in the multistage network jointly. Similarly, for the benchmark CNN, we first trained the affine subnetwork, then froze its layers, added the deformable stage and then trained the network. We then unfroze the layers of the affine subnetwork and finetuned the two subnetworks jointly. In the experiments, we trained the networks unsupervised with the Mutual Information loss of the fixed image (F) and the predicted moved image (M):

$$MI(F, M) = I(F; M) = H(F) + H(M) - H(F, M) \quad (4.4)$$

with the entropy H . Additionally, we used the Gradient L2 loss as a regularization of the deformation field ϕ :

$$L_2(\phi) = \sum_{p \in \Omega} \|\nabla \phi(p)\|^2 \quad (4.5)$$

The image similarity loss (Mutual Information, L_{MI}) and the regularization loss of the deformation field (Gradient L2, L_2) are then combined by the lambda-parameter λ :

$$L(F, M) = L_{MI}(F, M) + \lambda L_2(\phi) \quad (4.6)$$

We used the Keras Hyperband Tuner [63] for optimizing the learning rate, optimisation function and λ -parameter for each of the proposed models. The parameter configuration that resulted in the highest Mutual Information was selected, the networks were trained with this configuration. For all experiments with the datasets, intensity normalization to the range $[0, 1]$ was used. We conducted five-fold cross-validation and trained the networks for 200 epochs using a batch size of one. We applied early stopping of the training if the validation loss did not improve after five epochs. We saved the weights of the network that achieved the highest accuracy (highest Mutual Information) on the validation data. The fixed and moving images were resampled to $256 \times 256 \times 64$ voxels and a voxel spacing of $2 \times 2 \times 4 \text{ mm}^3$ using trilinear interpolation. The result transformation was resized and applied to the full resolution images. The experiments were performed in Python using Tensorflow [107] on a workstation equipped with an NVIDIA RTX A5000 with 24 GB GPU memory and an NVIDIA RTX A6000 with 48 GB GPU memory (NVIDIA, Santa Clara, CA, USA).

4.3.4 Baseline Method

As a baseline, we applied the classical registration method NiftyReg. First, the moving image was registered rigidly and then affinely with the algorithm `reg_aladin` [19]. The result image was then transformed deformably with the algorithm `reg_f3d` [18]. In addition to the fixed and moving image, both algorithms used the segmentation of the fixed and moving image to determine the structure of interest [19]. Additionally, we transformed the segmentation of the moving image with the transformation calculated by NiftyReg. The result is the segmentation of the moved image.

4.3.5 Datasets

We assessed the performance of the proposed multistage network using four multimodal three-dimensional in-vivo datasets. The imaging parameters of the datasets are listed in Table 4.1.

NIDDK

As the first dataset, we used in vivo images from the Consortium for Radiologic Imaging Studies of Polycystic Kidney Disease (CRISP) study, provided by the National Institute of Diabetes and Digestive and Kidney Disease (NIDDK), National

Table 4.1: Dataset statistics

Dataset	Modality	Volumes	Volume size	Resolution (mm ³)
First NIDDK dataset	T1-weighted MR scans	100	256 x 256 x [30, 80]	1.41 x 1.41 x 3.06
	T2-weighted MR scans	100	256 x 256 x [12, 30]	1.39 x 1.39 x 9.02
Second NIDDK dataset	T1-weighted MR scans	100	256 x 256 x [30, 80]	1.41 x 1.41 x 3.06
	T2-weighted MR scans	250	256 x 256 x [10, 51]	1.38 x 1.38 x 3.00
External kidney dataset	T1-weighted MR scans	41	512 x 512 x [40, 66]	0.74 x 0.74 x 4.00
	T2-weighted MR scans	41	256 x 256 x [30, 70]	1.46 x 1.46 x 4.02
M ² OLIE dataset	CT scans	47	512 x 512 x [25, 132]	0.78 x 0.78 x 2.04
	T1-weighted MR scans	73	[190, 440] x [288, 640] x [52, 120]	1.27 x 1.27 x 3.77

Institutes of Health, USA [125]. We utilized 100 datasets (i.e., a total of 100 T1-weighted and 100 T2-weighted scans) of patients with different stages of ADPKD (for details see [126]) obtained from the NIDDK database.

Manual segmentation of the kidneys was performed by two physicians on the T1- and T2-weighted coronal MR scans using an in-house developed annotation tool based on MeVisLab SDK (MeVis Medical Solutions, Inc., Bremen, Germany). Furthermore, the annotation tool was used to evaluate the inter-user agreement of the kidney segmentations. This resulted in a mean inter-user agreement of 0.91 ± 0.06 (Dice) for the kidney segmentations of the 100 T1-weighted MR scans and a mean inter-user agreement of 0.92 ± 0.02 (Dice) for the kidney segmentations of the 100 T2-weighted MR scans.

The second dataset also consists of T1- and T2-weighted MR scans taken from the NIDDK database. The second dataset contains the 100 T1-weighted MR scans of our first dataset. As a second modality, we used 250 T2-weighted MR scans with a slice thickness of 3 mm.

External kidney dataset

The third dataset contains 41 in-vivo T1-weighted and T2-weighted renal MR scans from patients with ADPKD, acquired at the Azienda Socio-Sanitaria Territoriale (ASST) Papa Giovanni XXIII, Bergamo, Italy in the context of the observational EuroCYST study (ClinicalTrials.gov Identifier: NCT02187432). The T1- and T2-weighted MR scans were acquired in coronal orientation. For the T1-weighted and T2-weighted MR scans, segmentation was performed manually. For the T1-weighted MR scans, kidney segmentations were obtained to quantify the total kidney volume. For the T2-weighted scans, kidney cyst segmentations were created.

In our experiments, in the first, second and third dataset, the T2-weighted MR scans were registered to the T1-weighted MR scans.

M²OLIE

The fourth dataset comprises CT and T1-weighted MR liver volumes of 39 patients. These were acquired as part of the "Mannheim Molecular Intervention Environment" (M²OLIE) project. The dataset includes 47 CT scans and 73 T1-weighted MR scans with in-phase and opposed-phase in transversal orientation. Liver segmentations were created using two separate segmentation networks for the CT and MR scans

(three-dimensional U-Nets [67]). The networks were trained on the 3D-IRCADb-01 [110] and Combined Healthy Abdominal Organ Segmentation (CHAOS) [111] datasets, respectively. We then finetuned both networks using 20 segmentations of the CT scans and 20 segmentations of the MR scans of the M²OLIE dataset, which we created manually in the Medical Imaging Interaction Toolkit (MITK) [112]. The segmentations generated by the networks were visually inspected to ensure that they were all acceptable. In our experiments, the MR scans were registered to the CT scans.

4.3.6 Evaluation Metrics

The performance of the proposed multistage network was evaluated by conducting experiments on the multimodal datasets. The evaluation was based on volumes (three-dimensional images and segmentations). The degree of overlap between the anatomical segmentations of the moved image (S_M) and the fixed image (S_F) was determined using the Dice coefficient.

Additionally, we evaluated the level of image folding due to the deformable transformation by calculating the Jacobian determinant ($\det(J)$ or $|J|$) of each point of the deformation field. In cases where the Jacobian determinant is equal to 1, the volume remains unaltered, if it exceeds 1, there is volumetric expansion, and if it falls between 0 and 1, there is volumetric shrinkage. A Jacobian determinant of ≤ 0 means that folding has occurred here. Image folding is anatomically impossible in medical images [86]. Therefore, the aim is a registration with a deformation field where the Jacobian determinant for each point is > 0 . For rigid and affine transformations, no folding can occur. The number of Jacobian determinants ≤ 0 is therefore 0.

We applied a paired t-test (for normal distribution) or a Wilcoxon signed-rank test (for non-normal distribution) to determine the significance of the improvement in the Dice coefficient and number of Jacobian determinants ≤ 0 after registration compared to the data before applying the registration (sub)network. The null hypothesis, which states that registration by the neural networks does not improve the two metrics, is rejected if $p < 0.05$.

For visual evaluation of the spatial alignment before and after registration, the fixed, moving and moved image were examined. Additionally, the moving/moved and fixed images were color overlaid to form a "Composite Image". If available, the segmentations of the moving/moved and fixed images were also color overlaid to form a "Composite Segmentation". In addition, the deformation field was displayed in RGB colors. The three color channels represent the deformation in the three spatial directions of three-dimensional images and transformations. The color red corresponds to a displacement on the y-axis (height), green on the x-axis (width) and blue on the z-axis (slice direction). This enables the identification of areas with strong deformation and a check on whether the deformation is smooth within the individual organs and structures. For areas that are displayed in gray values, the same deformation is applied in all three spatial directions.

4.3.7 Experiments

Experiment I

First, we trained our proposed multistage rigid-affine-deformable network with the first NIDDK dataset and compared its registration results with those of the benchmark network. This comparison aimed to evaluate the effect of using a preceding rigid subnetwork on the registration result.

We trained our proposed multistage network (Figure 4.2 a)) as described in Section 4.3.3 and Table 4.2. First, all subnetworks were trained separately, i.e. the weights of all previous stages were frozen. We then finetuned the whole multistage network jointly. We trained the benchmark network similarly.

Table 4.2: Hyperparameter setting for the training of the different subnetworks and finetuning of the whole benchmark network and proposed multistage-network (row 4 and 9), first NIDDK dataset used

Network		Trained Subnetwork	Learning rate	Optimizer function	λ -parameter
Benchmark	Affine	Affine	3×10^{-4}	Adam	0.8
	Affine-Deformable	Deformable	2×10^{-4}	Adam	0.7
	Affine-Deformable	All	3×10^{-4}	Adam	1
	Finetuned				
Proposed Multistage Network	Rigid	Rigid	3×10^{-4}	Adam	0.6
	Rigid-Affine	Affine	5×10^{-4}	Adam	0.7
	Rigid-Affine-Deformable	Deformable	5×10^{-4}	Adam	1
	Rigid-Affine-Deformable	All	3×10^{-4}	Adam	0.8
	Finetuned				

We then applied the trained networks on the second NIDDK dataset (T1-weighted and T2-weighted MR images with a slice thickness of 3 mm).

Experiment II

In addition, we applied the networks trained on the first NIDDK dataset to the external kidney dataset to determine whether the proposed multistage network also produces good registration results on a before unseen dataset.

To further evaluate the performance of the proposed multistage network, we augmented the external kidney dataset, i.e. applied artificial transformations and investigated if these transformations could be corrected with the proposed multistage network. We created an augmented dataset comprising 410 T1-weighted and T2-weighted MR volumes. This dataset contains the original scans and additionally to each T1-weighted or T2-weighted MR scan 9 augmented volumes. These volumes were augmented with the transformations described in Table 4.3. Translation, rotation, scaling and deformation were applied with a probability of 40% each. For the deformation, the Vomentations toolbox [127] was applied with the deformation limits (0, 0.25). The segmentations were transformed accordingly.

We finetuned the networks (benchmark and proposed multistage network) trained on the first NIDDK dataset on this augmented external kidney dataset. First, we

Table 4.3: Parameters for data augmentation performed on the T1-weighted and T2-weighted MR scans of the external kidney dataset

Parameter	Value
Translation (fraction of image size)	(-10, 10)
Rotation (degrees)	(-10, 10)
Scaling (factor)	(0.9, 1.1)
Deformation (parameter deformation_limits)	(0, 0.25)

froze the weight of the deformable subnetwork of the benchmark and the multistage network. So, we finetuned the rigid or rigid and affine subnetworks (benchmark: learning rate of 1×10^{-4} , Adam optimizer and a λ -parameter of 0.9, multistage network: learning rate of 2×10^{-4} , Adam optimizer and a λ -parameter of 1). Then, we unfroze the weights of the deformable subnetwork and jointly finetuned all subnetworks (benchmark: learning rate of 1×10^{-4} , Adam optimizer and a λ -parameter of 10, multistage network: learning rate of 2×10^{-4} , Adam optimizer and a λ -parameter of 10).

Experiment III

We finetuned our proposed multistage network (trained on the first NIDDK dataset) with the M²OLIE liver dataset (CT and MR scans) to demonstrate that the multistage network is generalisable and can be used for further applications, such as registration of CT and MR scans of the liver.

We finetuned all subnetworks jointly without freezing weights of individual subnetworks. We finetuned the benchmark with a learning rate of 2×10^{-4} , Adam optimizer, and a lambda-parameter of 4.5. The multistage network was finetuned with a learning rate of 3×10^{-4} , Adam optimizer and a lambda-parameter of 4.5.

4.4 Results

We demonstrate the effectiveness of our proposed multistage model through extensive experiments using the multimodal three-dimensional datasets described in Section 4.3.5. The experiments are described in Section 4.3.7.

4.4.1 Experiment I

Table 4.4 shows the registration results. The first stage of the multistage network (rigid subnetwork) and the first stage of the benchmark (affine subnetwork) yielded similar Dice coefficients. The rigid-affine subnetwork of the multistage network produced registration results with a significantly ($p < 0.05$) higher Dice coefficient than the affine subnetwork of the benchmark and a similar Dice coefficient as the baseline NiftyReg. The multistage rigid-affine-deformable network (separately trained subnetworks, row eight in Table 4.4) and the finetuned variant (row nine in Table 4.4) yielded similar Dice coefficients as the affine-deformable benchmark networks (separately trained and jointly finetuned, row four and five in Table 4.4). All stages of

the benchmark and the proposed multistage network (Table 4.4 row three to row nine) yielded registration results with a significantly higher Dice coefficient than before registration (Table 4.4 row one). The proposed multistage networks computed transformations with a significant ($p < 0.05$) lower level of image folding than the benchmark network ($|J| \leq 0$ in Table 2, row eight and nine vs row four and five). By jointly finetuning the multistage network, the number of $|J| \leq 0$ was significantly reduced (row eight vs row nine in Table 4.4). The results of the significance tests for the benchmark and proposed multistage network are listed in Table 4.5.

Table 4.4: The Dice coefficient and Jacobian determinant ($|J| \leq 0$) results (mean $\pm$ SD) of the first NIDDK dataset of the baseline and each stage of the benchmark and our proposed multistage network. Dice coefficient: higher value is better (maximum is 100%), $|J| \leq 0$: lower value is better (minimum is 0%).

Network		DICE (%)	$ J \leq 0$ (%)
Before Registration		67.6 ± 18.1	-
Baseline NiftyReg		70.9 ± 24.5	0.2 ± 0.6
Benchmark	Affine	69.1 ± 16.5	0 ± 0
	Affine-Deformable	76.4 ± 12.4	0.7 ± 0.8
	Affine-Deformable Finetuned	76.4 ± 12.7	0.7 ± 0.8
Proposed Multistage Network	Rigid	68.7 ± 17.3	0 ± 0
	Rigid-Affine	70.9 ± 14.2	0 ± 0
	Rigid-Affine-Deformable	76.7 ± 11.4	0.6 ± 0.7
	Rigid-Affine-Deformable Finetuned	76.7 ± 12.5	0.5 ± 0.7

Table 4.5: P-values of the significance tests (paired t-test or a Wilcoxon signed-rank test, $p < 0.05$) for the Dice coefficient and Jacobian determinant ($|J| \leq 0$) of the benchmark and proposed multistage network. Metrics with a significant difference ($p < 0.05$) are indicated in bold.

Proposed Multistage Network	Benchmark					
	Affine		Affine-Deformable		Affine-Deformable Finetuned	
	DICE	$ J \leq 0$	Dice	$ J \leq 0$	Dice	$ J \leq 0$
Rigid	0.0784	-				
Rigid-Affine	0.0048	-				
Rigid-Affine-Deformable			0.3125	<.00001		
Rigid-Affine-Deformable Finetuned					0.43675	<.00001

Examples of registration results can be seen in Figure 4.3. Before registration, the abdomen in the T1-weighted (fixed) and T2-weighted (moving) MR images is not well aligned. Registration with the baseline NiftyReg (second row in Figure 4.3) improved the structural similarity and the overlap of kidney segmentations. The rigid subnetwork (row six in Figure 4.3) significantly improved the spatial alignment compared to before registration. When adding the affine subnetwork, the overlap of the kidney segmentations is further improved compared to the rigid subnetwork (row six and seven in Figure 4.3). However, other structures, such as the liver, do not overlap well in the fixed and moved image, as visible in the composite image. By

adding the deformable subnetwork (row eight in Figure 4.3), the structural similarity of all structures is improved, especially visible in the liver in the composite image. Jointly finetuning the multistage network further increased the overlap of all structures (clearly visible in the liver). Comparing the registration result of the jointly finetuned multistage network (row nine) with the jointly finetuned benchmark (row five), demonstrates improved spatial alignment of all corresponding structures, especially visible in the liver, using the proposed multistage network. The deformations within the organs are smooth (sixth column in Figure 4.3). The largest deformations are located at the organ boundaries, especially at the liver and spleen.

We applied the trained networks on the second NIDDK dataset. The proposed multistage network yielded transformations with significantly ($p < 0.05$) less image folding than the benchmark ($|J| \leq 0$ (mean $\pm$ std): multistage network: 4.8 ± 2.1 , benchmark: 6.1 ± 2.1).

Examples of registration results of the second NIDDK dataset can be seen in Figure 4.4. Before the registration, especially the liver is not well aligned in the fixed and moving image (white arrow in Figure 4.4). Registration with the baseline did not improve the spatial alignment the corresponding structures. The overlap of the liver in the fixed and moved image did not increase. The benchmark and multistage CNN produced registration results with good alignment. Within the organs, the deformations were smooth.

4.4.2 Experiment II

Table 4.6 shows the registration results. Registration with the baseline NiftyReg improved the Dice coefficient. The benchmark and the proposed multistage network yielded similar Dice coefficients as before registration. Additionally, the benchmark and the proposed multistage network produced a similar level of image folding ($|J| \leq 0$).

Table 4.6: The Dice coefficient and Jacobian determinant ($|J| \leq 0$ results (mean $\pm$ SD) of the external kidney dataset without and with augmentation of the baseline, benchmark and proposed multistage network. Dice coefficient: higher value is better (maximum is 100%), $|J| \leq 0$: lower value is better (minimum is 0%).

Network	External kidney dataset without augmentation		External kidney dataset with augmentation	
	DICE (%)	$ J < 0$ (%)	DICE (%)	$ J < 0$ (%)
Before Registration	63.9 ± 11.3	-	45.1 ± 16.9	-
Baseline NiftyReg	67.3 ± 13.9	0.1 ± 0.2	59.5 ± 22.5	0.2 ± 1.0
Benchmark	62.4 ± 13.7	0.9 ± 0.4	62.5 ± 15.6	0.0 ± 0.0
Proposed Multistage Network	61.1 ± 14.0	1.1 ± 0.6	64.8 ± 16.2	0.0 ± 0.0

Examples of registration results can be seen in Figure 4.5. The fixed and moving image already show a reasonably good spatial alignment before registration, only fine structures in the kidney need to be corrected. The registration with the Baseline

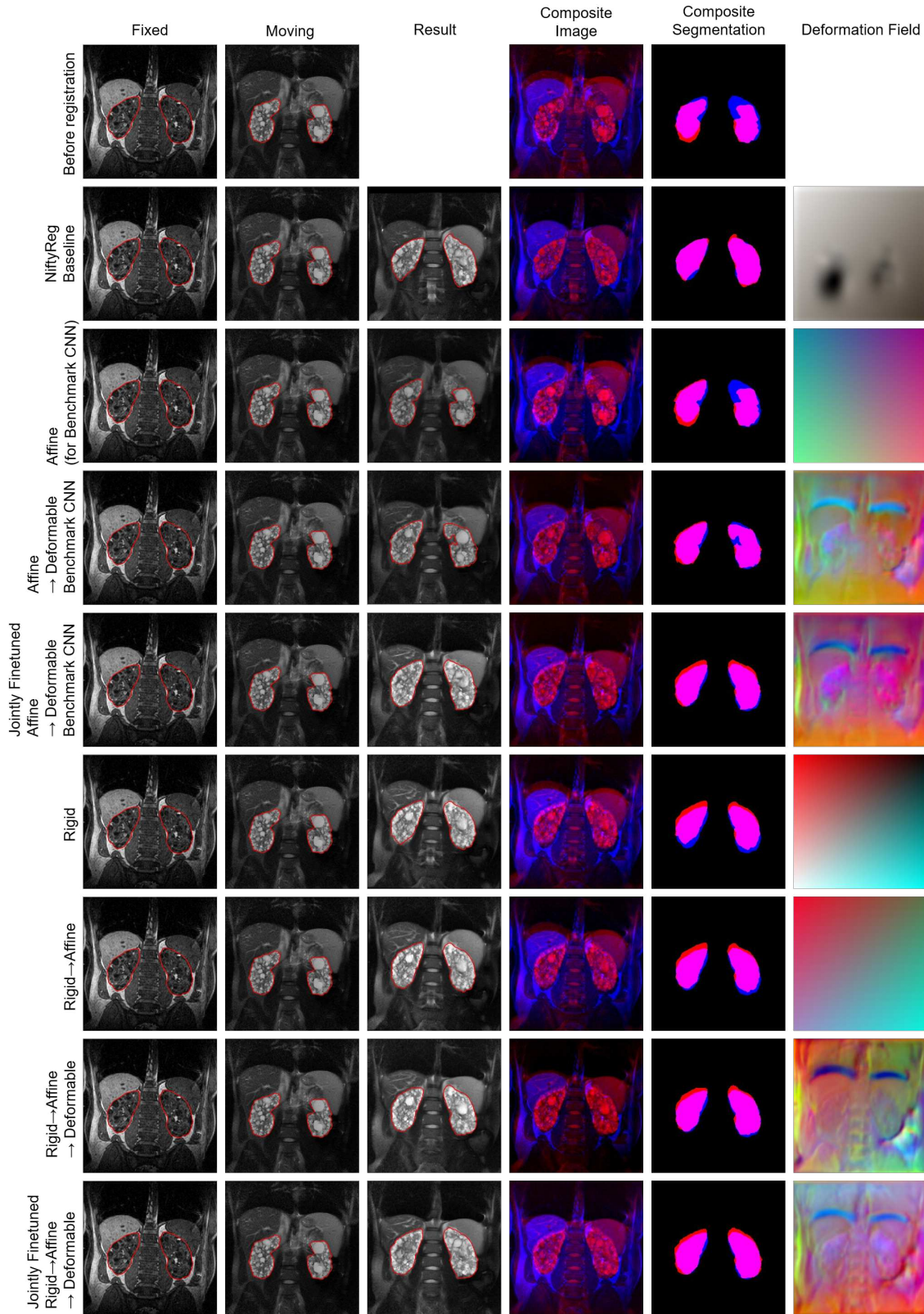


Figure 4.3: Example results of the registration of two volumes from the first NIDDK dataset. The images show central slices of the coronal plane. The fixed images, moving images and result (moved) images are overlaid by the kidney segmentations (column 1 - 3). The resulting composites for image and segmentation show the fixed data in blue and the moving/moved data in red. The deformation field is visualized in RGB channels.

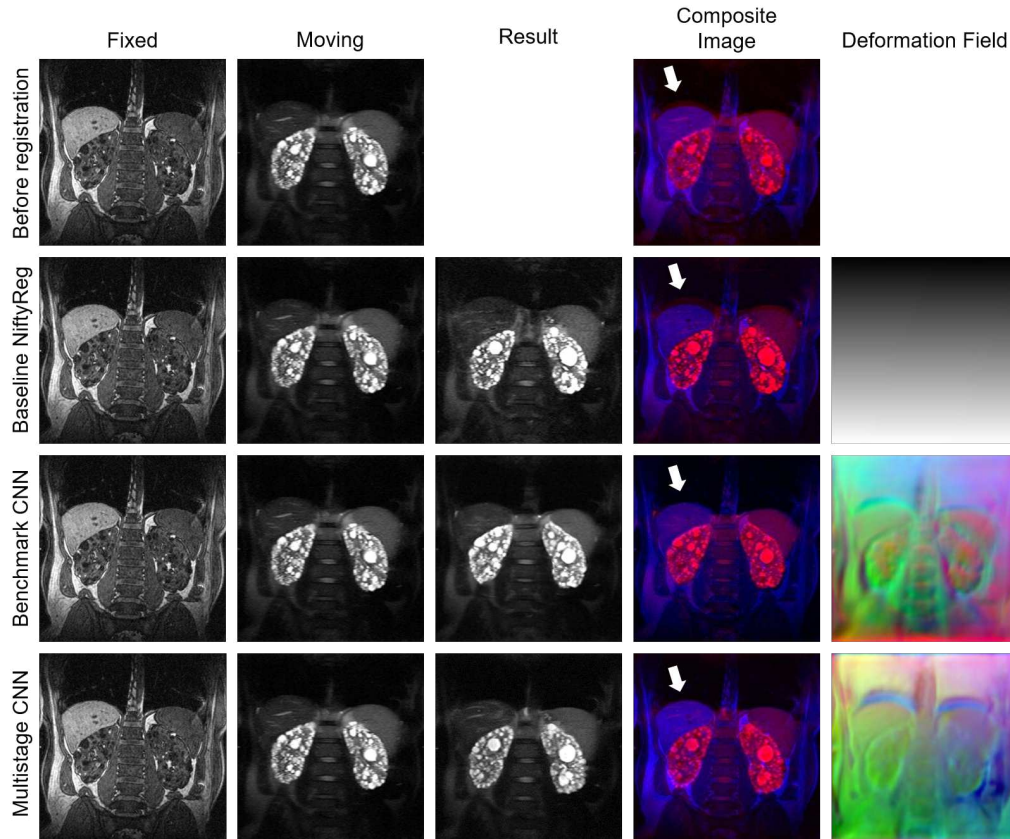


Figure 4.4: Example results of the registration of two volumes from the second NIDDK dataset. The images show central slices of the coronal plane. The resulting composites for image and segmentation show the fixed data in blue and the moving/moved data in red. The arrow in the composite image indicates the area with misalignment between the fixed and moving image. The deformation field is visualized in RGB channels.

NiftyReg has further improved the overlap of the kidneys. The benchmark and the proposed multistage network yielded similar results as before the registration. The deformation field of the multistage network shows smooth deformations.

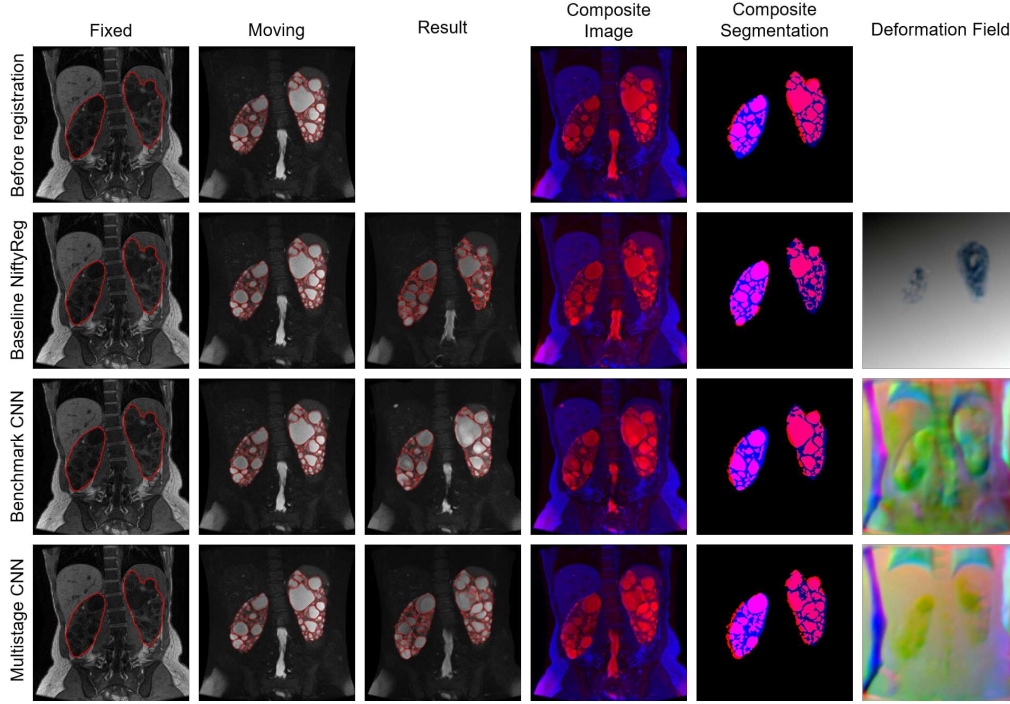


Figure 4.5: Example results of the registration of two volumes from the external kidney dataset. The images show central slices of the coronal plane. The fixed images are overlaid by the kidney segmentations and the moving images and result (moved) images are overlaid by the kidney cyst segmentations (column 2 and 3). The resulting composites for image and segmentation show the fixed data in blue and the moving/moved data in red. The deformation field is visualized in RGB channels.

Table 4.6 shows the registration results of the neural networks, which were finetuned and evaluated on the augmented external kidney dataset. The benchmark and the proposed multistage network yielded a significantly ($p < 0.05$) higher Dice coefficient than the baseline and registration with the proposed multistage network resulted in a significantly higher Dice coefficient than the benchmark. In addition, for the benchmark and the proposed multistage network, the level of image folding ($|J| < 0$) is significantly lower than for the baseline. For the benchmark and the proposed multistage network, the $|J| < 0$ is smaller than 0.05.

Examples of the registration results can be seen in Figure 4.6. Before registration, the corresponding images are not well aligned, especially visible at the spleen in the column "Composite Image" and at the segmentations in the column "Composite Segmentations". The baseline, benchmark, and multistage network produced registration results with good spatial alignment of all corresponding structures. And the networks produced smooth deformations.

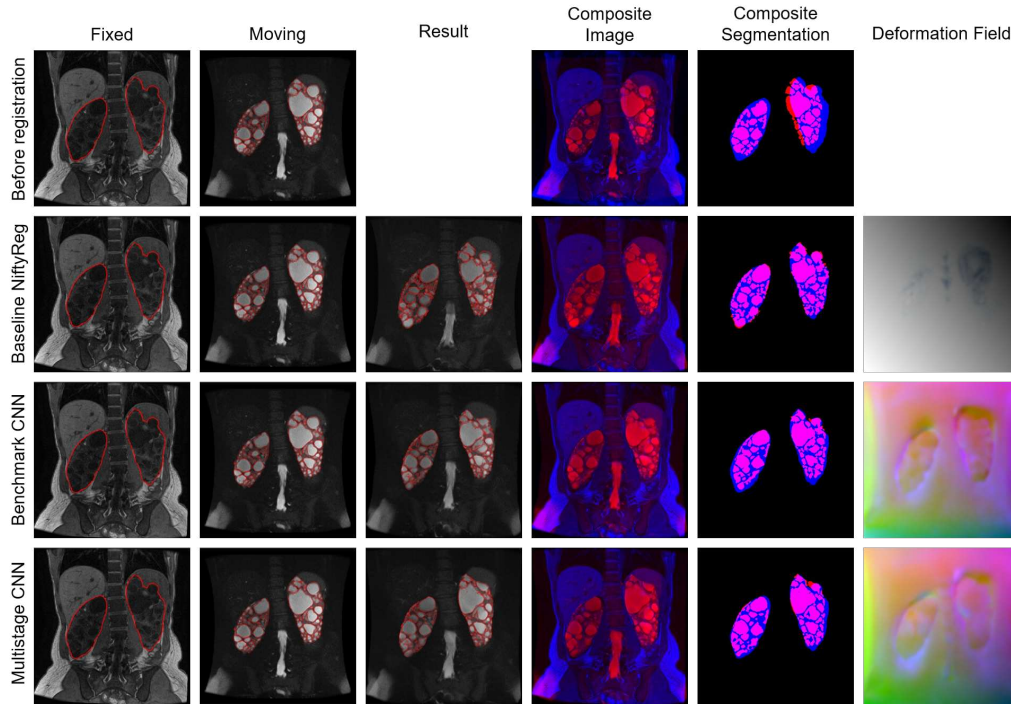


Figure 4.6: Example results of the registration of two volumes from the external augmented kidney dataset. The images show central slices of the coronal plane. The fixed images are overlaid by the kidney segmentations and the moving images and result (moved) images are overlaid by the kidney cyst segmentations (column 2 and 3). The resulting composites for image and segmentation show the fixed data in blue and the moving/moved data in red. The deformation field is visualized in RGB channels.

4.4.3 Experiment III

The results of the registration are shown in Table 4.7. The benchmark and our proposed multistage network yielded significantly ($p < 0.05$) higher Dice coefficients than the baseline. And registration with the proposed multistage network resulted in a significantly higher Dice coefficient than the benchmark. The level of image folding ($|J| < 0$) is similar for the benchmark and the proposed multistage network.

Table 4.7: The Dice coefficient and Jacobian determinant ($|J| \leq 0$) results (mean $\pm$ SD) of the M²OLIE liver dataset of the baseline, the benchmark and our proposed multistage network. Dice coefficient: higher value is better (maximum is 100%), $|J| \leq 0$: lower value is better (minimum is 0%).

Network	DICE (%)	$ J \leq 0$ (%)
Before Registration	54.4 $\pm$ 19.9	-
Baseline NiftyReg	63.3 $\pm$ 25.5	0.0 $\pm$ 0.1
Benchmark	66.6 $\pm$ 24.0	0.1 $\pm$ 0.1
Proposed Multistage Network	68.1 $\pm$ 24.6	0.1 $\pm$ 0.1

Example registration results are presented in Figure 4.7. A large rigid part of the deformation is visible the images before registration, as the fixed image is shifted downwards compared to the moving image. Registration with the baseline improved the overlap of the liver segmentations, however, the moved image is rotated incorrectly. Here the rigid or affine algorithm of the baseline computed an incorrect transformation. Registration with the benchmark improved the spatial alignment and overlap of the liver segmentations. However, in the moved image, the torso on the right side and parts of the right arm were deformed in a medically implausible way. This area is marked by a white arrow. The registration with the proposed multistage network also produced a medically implausible deformation on the right side of the torso (white arrow in result image in the fourth row). Apart from this area, the overlap of the corresponding structures improved by the registration. The area of the medical implausible deformation is also visible in the deformation field (white arrow in column six in Figure 4.7).

4.5 Discussion

We proposed a multistage end-to-end image registration neural network comprising rigid, affine, and deformable transformation. By using an affine-deformable network as a benchmark in our experiments, we investigated the influence of a preceding rigid subnetwork on the registration result.

4.5.1 Comparison to Published Multistage Registration Methods

In classical (iterative) methods, approaches to multistage registration (affine and deformable or rigid-affine-deformable) exist. For instance, the SimpleElastix [17] documentation describes a non-rigid registration. Here, the affine algorithm is applied first and the bspline (deformable) algorithm afterwards. In the iterative registration method NiftyReg [18, 19] the algorithm `reg_aladin` applies a rigid and affine

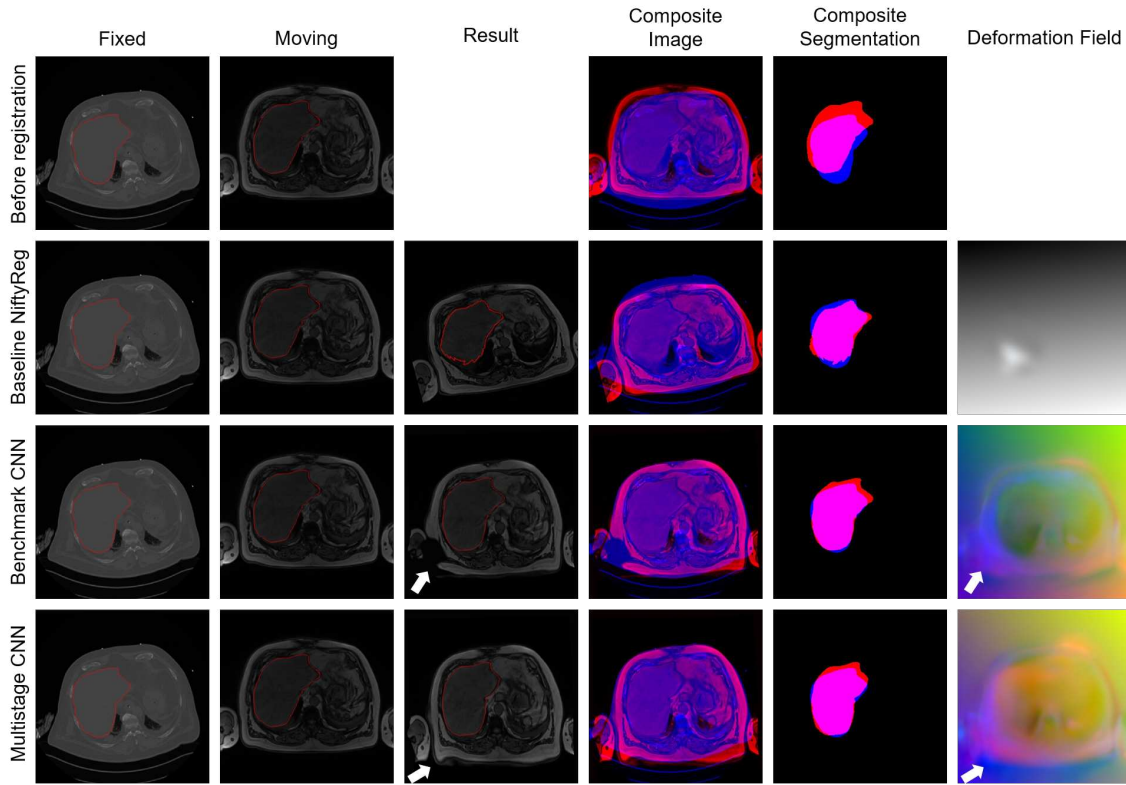


Figure 4.7: Example results of the registration of two volumes from the M²OLIE liver dataset. The images show central slices of the coronal plane. The fixed images, moving images and result (moved) images are overlaid by the liver segmentations (column 1 - 3). The resulting composites for image and segmentation show the fixed data in blue and the moving/moved data in red. The arrow in the result image indicates the area with medical implausible deformation. The deformation field is visualized in RGB channels.

registration by default. Thus, by connecting `reg_aladin` (rigid and affine) and the deformable registration with `reg_f3d`, a multistage rigid-affine-deformable registration is obtained. The ANTS (Advanced Neuroimaging Tools) [128] toolbox provides an elastic registration that uses affine and deformable transformations. In addition, a variant of the symmetric image normalization method (SyN) combines rigid, affine and deformable transformations (SyNRA).

However, these classical methods are often time-consuming as each new image pair requires iterative registration. Thus, the registration of a single image pair takes several minutes, depending on the registration method and the hardware used (CPU or GPU). With neural networks, the registration process speeds up. In comparison, our proposed multistage network takes under a second to register an image pair on CPU or GPU.

Several neural networks have been published with multistage approaches for medical image registration [76, 86, 91, 93–96, 98, 100–102, 104, 120–123]. [93–96, 98, 100–102, 104, 120, 121] apply neural networks with two stages. In the first stage/subnetwork an affine transformation is calculated, whereas in the second stage/subnetwork a deformable transformation is calculated. [76, 86, 91, 122] also utilize an affine

network as the first subnetwork, but this is followed by a cascade of multiple deformable subnetworks. And [123] applies Generative Adversarial Networks (GAN) for affine-deformable transformation. The generator consists of a modality translator (for multimodal image registration), an affine-transformation regressor and a nonlinear-deformation regressor.

However, only affine and deformable transformations were applied in these papers. To the best of our knowledge no multistage network has yet been published that combines the three transformations rigid, affine, and deformable. Therefore, our proposed multistage network represents a new approach to medical image registration and addresses the challenge of larger rigid deformations beyond the typical capture ranges of affine registration approaches [1, 71] that might occur in medical images.

A fair comparison of the results of our proposed rigid-affine-deformable multistage network with results in papers from published multistage approaches is difficult to achieve. In the published papers, different datasets, image modalities and target structures were used than in our experiments. The differences in the performance of the published multistage methods and our proposed multistage network could therefore also result from the different datasets and modalities used. Moreover, different target structures such as femur [100], brain [76, 91, 95, 98, 102, 120–123], liver or kidney (used in our experiments) contain different degrees of rigid, affine or deformable transformations, which may affect the performance of the multistage methods. A fair comparison is therefore only possible if all multistage approaches are evaluated on the same datasets, e.g. by also evaluating the published multistage methods on the datasets used in our experiments. The performance of the published networks on our datasets then also depends on the generalizability of the networks to new data. A comparative study on affine medical image registration networks [15] has shown that only a few networks are generalizable to new data and applications.

4.5.2 Individual networks

Our results demonstrate that the rigid subnetwork could correct rigid transformations that could not be corrected by the affine subnetwork alone. Using a rigid-affine network yielded better registration results than using an affine subnetwork without a preceding rigid subnetwork (cf. Table 4.4).

The results for Dice coefficient and level of image folding were mixed when comparing our approach to a benchmark network. For the first NIDDK dataset, the Dice coefficient of the benchmark and the proposed network were similar. However, upon further visual evaluation, the proposed multistage network showed a higher overlap of all corresponding structures. For the second NIDDK dataset, the spatial alignment between the benchmark and the proposed network was similar, indicating comparable performance. Both models also achieved a comparable performance (similar Dice coefficients) on the external kidney dataset. However, for the augmented external kidney dataset and the M²OLIE liver dataset, the proposed multistage network significantly outperformed the benchmark and yielded significantly ($p < 0.00001$ and $p = 0.00152$) higher Dice coefficients. For the first and second NIDDK dataset, our proposed multistage networks produced registration results with significantly less image folding compared to the benchmark, resulting in medical plausible transformations. For the external kidney dataset, the augmented external kidney dataset

and the M²OLIE liver dataset, the proposed network and the benchmark yielded similar levels of image folding. These results show that the proposed multistage network is promising for improving registration of certain datasets while competing with the benchmark for other datasets.

The level of image folding is low for all registration approaches used. It should also be noted that the number of Jacobian determinants ≤ 0 is a global metric. Spatial information, i.e., where image folding occurs, is not represented in this metric. For the neural networks and the baseline, the majority of image folding occurred in areas that are not relevant for the clinical analysis, e.g. in the image background.

The visual evaluation of the deformation fields demonstrated that the deformation within the individual organs and structures was smooth. The deformation fields of the baseline contained large areas in gray. Only in the area of the segmentations or in their surroundings colors are they visible in the deformation field. This is due to the fact that the segmentation of the fixed and moving image was used to determine the structure of interest for the baseline. Outside of these structures, deformations were generated that were identical in all three spatial directions.

4.5.3 Individual Datasets

In the first NIDDK dataset, the T2-weighted MR scans had a slice thickness of 9 mm, three times that of the T1-weighted MR scans. The T2-weighted slice in Figure 4.3 is therefore interpolated from multiple slices. The multistage network therefore not only performs registration of the T2-weighted scans on the corresponding T1-weighted scans, but also reconstructs thinner (3 mm) T2-weighted slices using T1-weighted MR scans.

We applied the networks trained on the first NIDDK dataset to the second NIDDK dataset, without finetuning the networks with the second dataset. This yielded registration results with a high spatial alignment, which indicates that generalization to new data was successful.

In the external kidney dataset, kidney segmentations were created for the T1-weighted scans, whereas kidney cyst segmentations were created for the T2-weighted MR scans. A Dice coefficient of 100 % cannot be achieved even with perfect registration and alignment of the kidneys as the two types of segmentation focus on different aspects. Kidney segmentations cover the entire kidney, including cysts and the renal parenchyma, whereas kidney cyst segmentations exclusively mark the cysts themselves. Moreover, a significant increase in the Dice coefficient does not necessarily indicate better spatial alignment. Hence, a visual evaluation of the registration results is important.

We also applied the networks trained on the first NIDDK dataset to the external kidney dataset. Good registration results were achieved without needing to finetune the networks. Compared to this, the multistage network yielded registration results with a higher Dice coefficient on the augmented external kidney dataset (last row in Table 4.6). This is due to the fact that the multistage network trained on the first NIDDK dataset was applied to the external kidney dataset without finetuning, whereas the multistage network was finetuned on the augmented external kidney

dataset before evaluation on the augmented external kidney dataset. Without finetuning, the multistage network performed poorly on the augmented dataset. The reason for this is that the augmentation generated transformations, such as rotation or translation, which do not occur to this extent in the training dataset (first NIDDK dataset). Hence, finetuning of the multistage network with the augmented external kidney dataset was necessary in order to be able to correct these deformations as well.

We augmented the external kidney dataset, i.e., performed artificial transformations similar to those used to augment training datasets, i.e., rotation, translation, scaling and deformation. In further work, the deformation of the abdomen due to respiration could also be simulated in order to investigate to what extent this deformation can be corrected with the proposed multistage network.

For jointly finetuning all subnetworks of the benchmark and multistage network with the augmented external kidney dataset, we used a large value for the lambda parameter (10) to regularize the deformation field. If the networks were trained with smaller values for the lambda parameter, the networks produced deformations of the moving image that were not medically plausible. Using a large value for the lambda parameter then resulted in a very low level of image folding, which leads to a medically plausible transformation.

Similarly, for the M²OLIE liver dataset, a large lambda parameter (4.5) was chosen to regularize the deformation field. However, medically implausible deformations were visible in the moved image at one area. These can be prevented by increasing the lambda parameter when finetuning the benchmark and proposed multistage network. However, this then leads to a worse registration result, i.e., lower overlap of all corresponding structures, compared to finetuning with a lambda parameter of 4.5. For the finetuning of the networks, a compromise had to be chosen between a smooth deformation field, which leads to medically plausible transformations (large lambda parameter) and a high deformation correction, i.e., improved spatial alignment through registration (lower lambda parameter, e.g., 4.5).

The segmentations of the kidney datasets were created manually. In contrast, the liver segmentations of the M²OLIE liver dataset were generated by neural networks. The performance of the segmentation networks and the differences in the performance of the **CT** and **MR** segmentations could have a significant impact on segmentation-based metrics, thus on the evaluation with the Dice coefficient.

For the registration of new datasets, the proposed multistage network trained on the first NIDDK dataset can be applied. Finetuning the network with the new dataset can improve the registration result. For this, different stages of the proposed multistage network can be finetuned separately (see finetuning with the augmented external kidney dataset in [Section 4.4.2](#)) or all stages can be finetuned jointly (see finetuning with the M²OLIE liver dataset, [Section 4.4.3](#)).

4.5.4 Limitations

Evaluating image registration is a challenging task, mainly since ground truth data is usually not available [7]. We applied a combination of different approaches for

evaluation. If segmentations of organs or structures are available in the fixed and moving image, segmentation-based metrics such as the Dice coefficient can be used to quantify the overlap between the segmentations in the two images before and after registration. Structural differences can be determined visually by color overlay of the fixed and moved images and segmentations (if available). In addition, the deformation field was examined. By calculating the Jacobian determinant of each point in the deformation field, the image folding can be determined. And optical evaluation of the deformation field visualizes the regions with smooth or large deformations.

Training the multistage network with three subnetworks requires a lot of GPU memory. Training the multistage network was only possible with a maximum image size of $256 \times 256 \times 64$ (resolution $2 \times 2 \times 4 \text{ mm}^3$) on an NVIDIA RTX A6000 GPU with 48 GB of GPU memory. This could limit the registration result, especially in the deformable network, as the higher the resolution of the images, the smaller the deformations that can be corrected. As future work, a multistage network could be developed that combines rigid, affine and deformable transformation within a single network without multiple subnetworks (for example, a U-Net-like network) which could be trained with a larger image size (e.g. $512 \times 512 \times 128$, resolution $1 \times 1 \times 2 \text{ mm}^3$) due to the smaller network size.

4.6 Conclusions

We proposed a multistage neural network for three-dimensional multimodal medical image registration which combines rigid, affine and deformable transformations. Our results show that incorporating rigid, affine, and deformable transformations within a multistage network yields registration results with a high structural similarity, overlap of all corresponding structures and a low level of image folding, resulting in a medically plausible transformation.

5. Multistep Networks for Deformable Multimodal Medical Image Registration

This chapter is based on the following publication: A. Strittmatter, F. G. Zöllner. Multistep networks for deformable multimodal medical image registration. IEEE Access 12, 82676–82692, doi: 10.1109/ACCESS.2024.3412216, 2024.

5.1 Introduction

Patients with oligometastatic disease often have poor survival rates due to the different genetic and molecular characteristics of the cancer cells in the tumours that may resist standard treatments. Tailored techniques, such as minimally invasive image-guided therapies that take into account the heterogeneity of the tumour, greatly improve survival rates [12, 13]. These techniques use MR imaging in addition to CT to assess tumour heterogeneity, which allows better visibility of lesions compared to CT alone [14]. The Mannheim Molecular Intervention Environment (M²OLIE) research campus focuses on optimising the treatment of oligo-metastatic liver cancer, particularly through image-guided procedures. Pre- and postoperative CT and MR imaging, as well as intraoperative CBCT are used. The images are combined into a multimodal dataset containing morphological and functional tumour information [1]. Differences in patient positioning between the different scanners and variations in patients' respiration lead to differing organ deformations in the corresponding multimodal images. Consequently, image registration is an essential step before fusing the images.

Medical image registration also plays an important role in the diagnosis and monitoring of kidney disease [5–7], particularly in autosomal dominant polycystic kidney disease (ADPKD), a hereditary disease leading to kidney enlargement and end-stage renal disease (ESRD). Detecting ADPKD at an early stage and consistently monitoring its progression are crucial for preventing or delaying the onset of ESRD. Medical image registration enables the fusion of images from different MR modalities and the combination of complementary information, significantly improving the accuracy of ADPKD diagnosis and monitoring [9]. In addition, the registration of images at different time points, e.g., before, during, and after therapy, is essential for monitoring treatment progress.

In recent years, several medical image registration methods have been developed that use multiple steps with different resolutions. E.g., the conventional registration method SimpleElastix [17] applies four resolutions, whereas NiftyReg [18, 19] uses three resolutions. Furthermore, several published neural networks utilize multiple subnetworks with different resolutions in a coarse-to-fine approach [86, 129–140]. However, to the best of our knowledge, the impact of multiple cascaded subnetworks with multiple resolutions on the registration result has not been investigated and compared to the performance of a monoresolution/monostep network. Therefore, we implemented multistep networks with different numbers of resolutions (four, three,

and two steps/resolutions) and compared their performance to each other and to the performance of a monostep (one resolution) network. Thus, we investigated the potential benefits and trade-offs of cascaded multiresolution registration networks with up to four steps/subnetworks.

5.2 Background

In image registration, a transformation is calculated to optimally align all corresponding structures in two or more images. This process registers a moving image (M) to a fixed (reference) image (F) by optimizing the energy function L :

$$L(T; F, M) + R(T) \quad (5.1)$$

where T is the transformation and R is a regularization term to ensure a smooth and (medically) plausible transform. The moved image is then created by applying the computed transformation T to the moving image.

In non-linear (deformable) registration, the transformation is modeled by a deformation field (displacement field) ϕ . The registered (moved) image M' is then created by warping the moving image M with the deformation field ϕ :

$$M' = M \circ \phi \quad (5.2)$$

Especially in the field of medical image registration, the deformations are usually not correctable in one single step at full resolution [75]. Instead, multiresolution approaches are often applied to improve the accuracy of the image registration. This involves processing the input images in different resolutions, which can be performed in parallel (in a single step) or in multiple steps sequentially.

In a sequential multiresolution process (called multistep multiresolution registration), multiple cascades are used. In the first cascades, the images are processed, i.e. registered, at low resolution. Thus, initially coarse deformations are corrected. The inputs of the first cascade are the downsampled fixed and moving images F and M . The first cascade outputs the moved image M'_1 and the deformation field ϕ_1 . The moved image M' and the fixed image F are then subsequently processed in higher resolution, so that for increasing resolutions increasingly finer deformations are corrected [78, 79]. All subsequent cascades therefore receive the moved image of the previous cascade M'_k as input in addition to the fixed image F . The moved image after the k -th cascade M'_k is therefore calculated from the moving image M , warped by the combined deformation fields of all k cascades $\phi_{combined}$ [76, 77]:

$$M'_k = M \circ \phi_{combined} \quad (5.3)$$

with

$$\phi_{combined} = \phi_1 \circ \dots \circ \phi_k \quad (5.4)$$

For multiresolution approaches, Gaussian pyramids are commonly used. The individual pyramid levels contain the input images at different resolutions [141, 142]. At

the base level, the images are processed in full resolution with a size of $N_x \times N_y \times N_z$. At the other pyramid levels, the images then have a size of $\frac{N_x}{k_i} \times \frac{N_y}{k_i} \times \frac{N_z}{k_i}$, where k_i is the downsampling factor of the individual level. The number of levels and the resolutions/downsampling factors used should be selected depending on the specific task to be performed [75]. The image resolution can be reduced using various methods. For instance, a low-pass filter can be applied prior to downsampling to avoid alias effects, which could affect the performance of the registration method.

5.3 Methods

5.3.1 Network Architectures

We developed three neural network variants for multistep deformable registration (Figure 5.1a-c). In multistep network variant A, four steps/resolutions are applied, in variant B, three and in variant C, two steps/resolutions are used. The different resolutions are processed in different subnetworks ("deformable subnetwork" in Figure 5.1). The subnetworks are connected in series. Each subnetwork receives as input the fixed image and the moving image (first subnetwork of the series) or the moving image, which was transformed with the transformations that were calculated by the preceding subnetworks. The transformations generated by the subnetworks are represented as deformation fields. A combination of the different deformation fields is achieved by matrix addition. A spatial transformer [106] is used to apply the transformation to the moving image. The spatial transformer receives the moving image and the deformation field as input and then performs a transformation (warping) of the moving image according to (Equation 5.2), thus creating the moved image.

The first subnetwork in the series processes the images in the lowest resolution/image size. The following subnetworks process the images in increasingly higher resolution. For this, the moving and fixed images are downsampled with the following factor k :

$$k = 2^{r-s} \quad (5.5)$$

with r as the total number of resolutions in the multiresolution network and s as the subnetwork number in the series (first subnetwork $s=1$, second subnetwork $s=2$, etc.).

All subnetworks use the same architecture (Figure 5.2). Here, an adaptation of the VoxelMorph-2 network is applied. Due to the small image size (especially the number of slices) of the first subnetwork in multistep network variant A, only three U-Net encoder layers/downsampling steps can be applied. Thus, the features $[[16, 32, 32], [32, 32, 32, 16, 16]]$ are used in all subnetworks instead of $[[16, 32, 32, 32], [32, 32, 32, 32, 16, 16]]$ of VoxelMorph-2 [16, 124]. We applied convolutional layers with a kernel size of $3 \times 3 \times 3$, Leaky ReLU activation layers and $2 \times 2 \times 2$ MaxPooling and UpSampling layers.

As a benchmark, a monostep network is applied (Figure 5.1d). Only one (the highest) resolution and only one subnetwork is used. This is the modified VoxelMorph-2

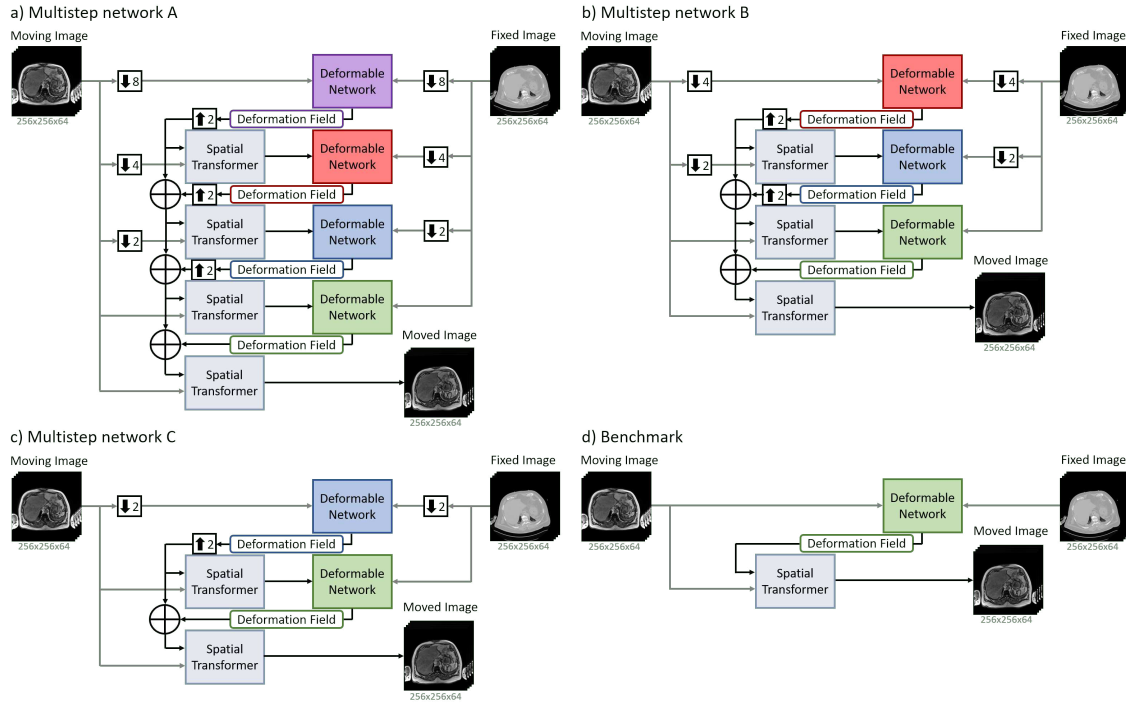


Figure 5.1: Architecture of the multistep network variants A, B, C and the Benchmark (monostep network). As subnetworks, an adaption of VoxelMorph-2 [16, 124] is used.

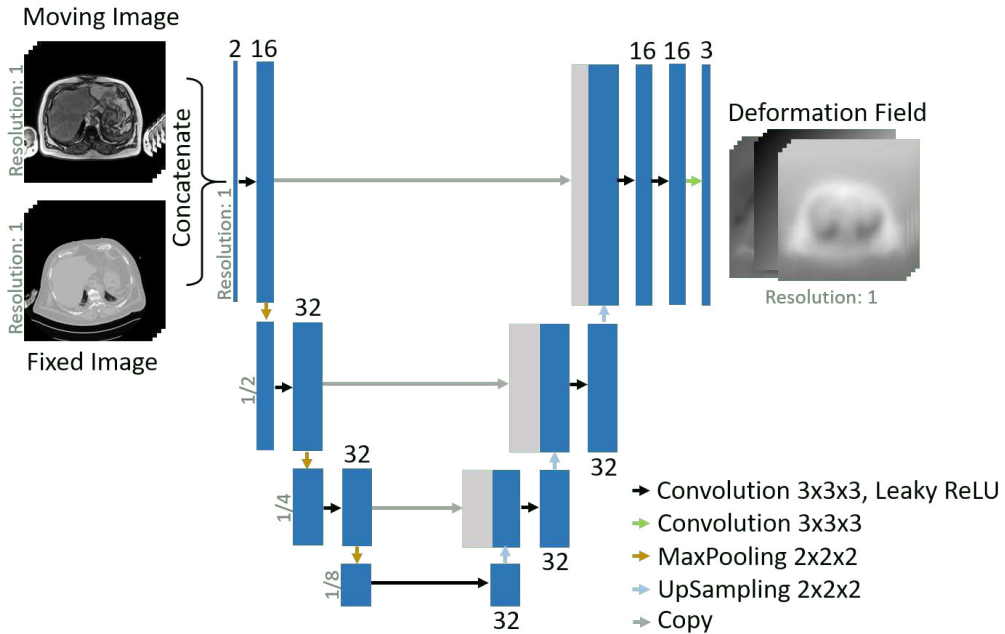


Figure 5.2: Architecture of the deformable networks, used in the multistep network variants A, B, C and the benchmark (monostage network). Adapted from VoxelMorph-2 [16, 124]. The resolutions of the volumes in relation to the resolution of the inputs are shown in gray.

architecture, as with the multistep variants A-C. Additionally, we used the unmodified VoxelMorph-2 network. Furthermore, we compared the networks with LapIRN [143]. This network applies a coarse-to-fine monostep-multiresolution approach using Laplacian pyramids.

5.3.2 Training Setting for the Neural Networks

We trained the multistep networks jointly, i.e., the weights of all subnetworks of a multistep network were optimized within one training process.

Since we use multimodal datasets, we applied Mutual Information (MI) of the fixed image (F) and the moved image (M) as similarity loss, as it is based on entropies (H) instead of a voxel-wise comparison of the image intensities:

$$L_{MI}(F, M) = I(F; M) = H(F) + H(M) - H(F, M) \quad (5.6)$$

To regularize and smoothen the deformation field ϕ , we used Gradient L_2 loss:

$$L_2(\phi) = \sum_{p \in \Omega} \|\nabla u(p)\|^2 \quad (5.7)$$

Mutual Information loss (L_{MI}) and Gradient L2 loss (L_2) are then linked by a lambda parameter:

$$L(F, M) = L_{MI}(F, M) + \lambda L_2(\phi) \quad (5.8)$$

We applied five-fold cross-validation and trained the networks for 200 epochs with a batch size of 1. Early stopping was applied if the validation loss did not show improvement over five consecutive epochs. We saved the model weights with the highest validation accuracy.

Hyperparameter tuning involved optimizing the learning rate, the optimization function, and the λ -parameter through a grid search. We selected the parameter configuration that yielded the highest Mutual Information on the validation dataset and used this setting to train the multistep networks. The image intensities were normalized to the range [0, 1].

To overcome GPU memory constraints, we resized the network inputs (fixed and moving images) to 256 x 256 x 64 voxels with a spatial resolution of 2 x 2 x 4 mm³. During inference, we applied the transformations computed by the networks to the full-resolution images.

We conducted our experiments using Python with the Tensorflow framework [107] on a workstation equipped with a NVIDIA RTX A6000 GPU with 48 GB of GPU memory.

Our source code is available at https://github.com/Computer-Assisted-Clinical-Medicine/Multistep_Networks_for_Deformable_Medical_Image_Registration.

5.3.3 Baseline Methods

As a first baseline, we used a conventional deformable registration with SimpleElastix [17]. We applied the default parameters for a bspline transformation. The "FinalGridSpacingInPhysicalUnits" was set to [100, 100, 100] to achieve smooth and medically plausible transformations. By default, four resolutions are used with a grid spacing downsampling factor of [2.80322, 1.9881, 1.41, 1].

As a second baseline, we applied the deformable algorithm `reg_f3d` of the conventional registration method NiftyReg [18] with default parameters. Three resolutions are applied with a downsampling factor of [4, 2, 1]. The algorithm uses the segmentation of the fixed and moving image to identify the structure of interest [19].

5.3.4 Datasets

In-vivo M²OLIE liver dataset

The first dataset comprises in-vivo CT and MR T1-weighted liver volumes, acquired as part of the "Mannheim Molecular Intervention Environment" (M²OLIE) study. The dataset consists of 73 MR T1-weighted scans in both in-phase and opposed-phase and 47 CT scans. The CT scans have a matrix size of 512 x 512 voxels with 25-132 slices, an in-plane resolution of 0.78 x 0.78 (± 0.12) mm, and a slice thickness of 2.04 (± 0.76) mm. The MR images were obtained using a matrix of 190-440 x 288-640 voxels with 52-120 slices. The in-plane resolution of the MR scans is 1.27 x 1.27 (± 0.17) mm, and the slice thickness is 3.77 (± 0.46) mm. To perform liver segmentation on CT and MR scans, two different segmentation networks were used (for details, see [15]). In our experiments, we registered the MR scans to the CT scans.

Synthetic liver dataset

The second dataset comprises 112 synthetic CT and MR T1-weighted liver volumes. These synthetic volumes were created using a CycleGAN as described in [109] from images of the XCAT phantom [108]. In the XCAT phantom, 56 models were available. For each model, the inhaled and exhaled states were simulated. The resolution of the CT volumes is 1 x 1 x 2 mm³, and the matrix size is 512 x 512 with 82-124 slices. MR volumes have a resolution of 1 x 1 x 3 mm³ and a matrix of 330 x 450 and 54-83 slices. Additionally, this dataset contains anatomical segmentation masks for the liver, which were generated by the XCAT phantom.

In our experiments, we performed multimodal image registration of the MR images to the CT images, aligning the inhaled state with the exhaled state and vice versa.

The synthetic liver dataset is available at <https://heidata.uni-heidelberg.de/dataset.xhtml?persistentId=doi:10.11588/data/ICSFUS>.

NIDDK kidney dataset

As third dataset, we used MR scans from the Consortium for Radiologic Imaging Studies of Polycystic Kidney Disease (CRISP) study. This study was conducted by the National Institute of Diabetes and Digestive and Kidney Disease (NIDDK),

National Institutes of Health, USA [125]. The NIDDK dataset comprises T1- and T2-weighted kidney MR scans, capturing various stages of ADPKD (for details, see [126]). For our study, we used 100 datasets (100 T1-weighted and 100 T2-weighted scans) from the NIDDK database. The T1- and T2-weighted MR scans were acquired in coronal orientation. The T1-weighted MR volumes were acquired with a 256 x 256 matrix and 30-80 slices. The in-plane resolution is 1.41 x 1.41 (± 0.13) mm², and the slice thickness is 3.06 (± 0.29) mm. The T2-weighted MR scans have an in-plane resolution of 1.39 x 1.39 (± 0.10) mm², a slice thickness of 9.02 (± 0.14) mm, and a matrix of 256 x 256 voxels with 12-30 slices.

Two physicians manually segmented the kidneys on the coronal T1- and T2-weighted MR scans. The segmentation was performed with an in-house developed annotation tool using MeVisLab SDK (MeVis Medical Solutions, Inc., Bremen, Germany). The inter-user agreement for kidney segmentation in the T1-weighted and T2-weighted MR scans was 0.91 ± 0.06 and 0.92 ± 0.02 (Dice) with a coefficient of variation of 0.07 and 0.03, respectively.

In our experiments, we registered the T2-weighted MR scans to the T1-weighted MR scans.

ProstateX dataset

As fourth dataset, we applied the ProstateX dataset[144–146]. The dataset comprises 349 prostate MR studies, including prostate MR scans of 346 patients. For each study, T2-weighted, proton density-weighted, dynamic contrast-enhanced, and diffusion-weighted (DW) MR scans with the b-values 50, 400, and 800 were acquired in axial orientation (for details, see [145]). For our study, we used the T2-weighted and diffusion-weighted (DW) MR scans (349 T2-weighted and 1047 Diffusion Weighted Imaging (DWI) MR scans). The T2-weighted MR volumes were acquired with a 256-640 x 256-640 matrix and 18-29 slices. The in-plane resolution of the T2-weighted MR volumes is 0.51 x 0.51 (± 0.04) mm², the slice thickness is 3.04 (± 0.19) mm. The DWI scans have a matrix of 84-128 x 106-128 voxels with 19-25 slices, an in-plane resolution of 2.00 x 2.00 mm² and a slice thickness of 3.11 (± 0.31) mm. We cropped the T2-weighted and DWI MR scans so that corresponding images have the same field of view. In our experiments, we registered the DWI (b-values 50, 400, and 800) MR scans to the T2-weighted MR scans.

5.3.5 Evaluation Metrics

We assessed the network’s performance using several quantitative metrics.

We evaluated the structural similarity of the fixed image and the moved image using Normalized Mutual Information (NMI).

We evaluated the overlap of the segmentations of the fixed image and the moved image using the Dice coefficient. For the fourth dataset, no segmentations were available. Therefore, instead of the Dice coefficient, we applied a different metric. Here we calculated the Structural Similarity Index (Structural Similarity Index Measure (SSIM)) of the fixed and moved image.

For all datasets, we evaluated the degree of image folding (anatomically impossible deformations) by calculating the number of Jacobian determinants ≤ 0 in the deformation field. Additionally, we measured the time needed to register a new image pair.

To determine the statistical significance of the improvements in the Dice coefficient and in the number of Jacobian determinants ≤ 0 , we performed statistical hypothesis tests. We compared the registered data with the unregistered data and compared our multistep networks with each other and with the benchmark and baselines. For normally distributed data, we applied a paired t-test, whereas for non-normally distributed data, we applied a Wilcoxon signed-rank test. Our null hypothesis states that the registration performed by our multistep networks did not lead to improvements in either metric. The null hypothesis is rejected if $p < 0.05$.

To visually evaluate the registration results, we displayed fixed, moving, and moved images of an example result. A "Composite Image" and "Composite Segmentation" are created by overlaying moving/moved (in red) and fixed (in blue) image or segmentation. In addition, the deformation field is inspected visually to identify areas of strong or smooth deformation. The deformation field is visualized with RGB colors, where each channel corresponds to the deformation in one of the three spatial directions of the three-dimensional images (red: displacement on the y-axis/height, green: x-axis/width, blue: z-axis/slice direction).

5.3.6 Experiments

Experiment I

First, we trained the multistep networks, the benchmark, the unmodified VoxelMorph-2 network and LapIRN on the in-vivo M²OLIE liver dataset with the hyperparameters in Table 5.1. We trained all subnetworks jointly.

Table 5.1: Hyperparameter setting for the training of the multistep networks, the benchmark, VoxelMorph-2 and LapIRN

Architecture	Number of parameters	M ² OLIE dataset		Synthetic dataset		NIDDK dataset		ProstateX dataset	
		Learning rate	λ -parameter	Learning rate	λ -parameter	Learning rate	λ -parameter	Learning rate	λ -parameter
Multistep A	977,292	1.00E-04	10	5.00E-04	8	2.00E-04	10	5.00E-04	10
Multistep B	732,969	5.00E-04	8	1.00E-04	8	4.00E-04	10	2.00E-04	8
Multistep C	488,646	2.00E-04	12	2.00E-04	10	2.00E-04	8	1.00E-04	10
Benchmark	244,323	6.00E-04	10	4.00E-04	6	1.00E-04	12	4.00E-04	10
VoxelMorph-2	327,331	2.00E-04	8	1.00E-04	10	5.00E-04	10	1.00E-04	12
LapIRN	964,575	1.00E-04	4	2.00E-04	2	1.00E-04	6	5.00E-04	10

Experiment II

As an additional test dataset, the synthetic dataset was used. First, the networks that were trained on the in-vivo M²OLIE liver dataset were evaluated with the synthetic dataset. Afterwards, the neural networks (trained on the in-vivo M²OLIE dataset) were finetuned with the synthetic dataset with the hyperparameters in Table 5.1. All subnetworks were finetuned jointly.

Experiment III

We trained the multistep networks, the benchmark, the unmodified VoxelMorph-2 network and LapIRN with the in-vivo NIDDK kidney dataset and the hyperparameter settings listed in Table 5.1. All subnetworks were trained jointly. In this experiment, all networks were trained from scratch, i.e., without the use of pretrained weights.

Experiment IV

To evaluate the performance of the multistage networks on additional imaging modalities, we used the in-vivo ProstateX dataset. The original MR DWI scans (b50, b400 and b500) and MR T2-weighted scans are already spatially aligned well. Therefore, we added artificial transformations to the dataset. We created an additional augmented volume for each original scan. This augmented volume was created by deforming the original scan with the VoxelMorph toolbox [127] with the deformation limits (0, 0.25). We then registered the undeformed/original MR DWI scans (b50, b400 and b500) to the deformed MR T2-weighted scans and the deformed MR DWI scans to the undeformed/original MR T2-weighted scans. We trained all networks from scratch using the hyperparameters in Table 5.1.

Experiment V

Since the training of the networks required a lot of GPU memory due to the large networks and three-dimensional input images, downsampling the input images to $256 \times 256 \times 64$ (with a resolution of $2 \times 2 \times 4 \text{ mm}^3$) was required. However, downsampling can lead to alias effects and thus affect performance. Therefore, in an additional experiment, we used a decimation filter to downsample the input images. Here, an anti-aliasing filter (order 8 Chebyshev type I filter) is applied first before the image is downsampled. The neural networks were then trained with these downsampled images (in-vivo ProstateX dataset), using the hyperparameters in Table 5.1. The performance of these networks was then compared to that of the networks that did not apply a decimation filter (experiment IV).

5.4 Results

5.4.1 Experiment I

Table 5.2 shows the registration results.

All methods except the baseline NiftyReg were able to improve the NMI significantly ($p < 0.00001$) compared to the unregistered images. Multistep variant B achieved the highest NMI and VoxelMorph-2 yielded a similar NMI as the benchmark. The baseline SimpleElastix significantly ($p < 0.00001$) improved the Dice coefficient compared to the unregistered images. Using multistep variant A resulted in a similar Dice coefficient ($p = 0.74171$). When only three steps/resolutions (multistep variant B) were used instead of four (variant A), the Dice coefficient of the registration result worsened significantly ($p < 0.00001$). And with a further reduction in the number of steps (variant C), the Dice coefficient continued to deteriorate

significantly ($p < 0.00001$). The benchmark and VoxelMorph-2 yielded Dice coefficients that were significantly worse than the Dice coefficient of multistep variant C (benchmark vs multistep variant C: $p = 0.00053$) and the baseline SimpleElastix ($p < 0.00001$), but still significantly ($p < 0.00001$) higher than the Dice coefficient of the unregistered images. LapIRN yielded a Dice coefficient that is significantly lower ($p < 0.00001$) than the Dice coefficient of the benchmark. Registration results with the lowest Dice coefficient were achieved with the baseline NiftyReg.

Registration with the baselines, multistep networks, and the benchmark all led to a very low number of image folding (Jacobian determinant ≤ 0). Registration of each image pair took 18.01 ± 4.16 seconds with SimpleElastix and 9.40 ± 11.17 seconds with NiftyReg. By using neural networks, the registration time was significantly reduced. The registration process speeded up the fewer steps/subnetworks were used, however, it was less than half a second for all neural networks.

Table 5.2: The Normalized Mutual Information (NMI), Dice coefficient, Jacobian determinant ($|J| \leq 0$) and registration time results (mean $\pm$ SD) of the M²OLIE liver dataset of the baselines, multistep networks, benchmark, VoxelMorph-2 and LapIRN. NMI: higher value is better (maximum is 1), Dice coefficient: higher value is better (maximum is 100%), $|J| \leq 0$: lower value is better (minimum is 0%).

Method/Network	NMI	DICE (%)	$ J \leq 0$ (%)	Registration Time (s)
Before registration	0.12 ± 0.04	54.4 ± 19.9	-	-
Baseline SimpleElastix	0.17 ± 0.03	71.5 ± 16.9	0.0 ± 0.0	18.01 ± 4.16
Baseline NiftyReg	0.12 ± 0.04	54.9 ± 18.3	0.0 ± 0.0	9.40 ± 1.17
Multistep A	0.14 ± 0.04	72.0 ± 18.9	0.0 ± 0.1	0.36 ± 0.04
Multistep B	0.15 ± 0.04	69.1 ± 18.7	0.0 ± 0.0	0.35 ± 0.03
Multistep C	0.14 ± 0.04	64.2 ± 18.3	0.0 ± 0.0	0.32 ± 0.05
Benchmark	0.13 ± 0.03	61.0 ± 18.0	0.0 ± 0.0	0.25 ± 0.04
VoxelMorph-2	0.13 ± 0.03	61.3 ± 18.6	0.0 ± 0.0	0.26 ± 0.04
LapIRN	0.13 ± 0.03	57.1 ± 17.8	0.2 ± 0.3	0.30 ± 0.06

Figure 5.3 shows examples of registration results. Before registration (row one), the abdomen and liver is not well aligned in the CT and MR images. By registration with the baseline SimpleElastix (row two) a high spatial alignment of the abdomen and the liver was achieved, especially visible in the "Composite image" and the "Composite segmentation". The baseline NiftyReg could not improve the spatial alignment compared to the unregistered images. Using multistep variant A resulted in a high spatial alignment of the abdomen and the liver. By reducing the number of steps (multistep variant B $\rightarrow$ multistep variant C $\rightarrow$ benchmark), the spatial alignment of the corresponding structures and the overlap of the livers deteriorates increasingly. LapIRN could not achieve a good spatial alignment. Deformations in the organs are smooth for all registration methods.

5.4.2 Experiment II

Table 5.3 shows the registration results of the networks that were trained on the M²OLIE liver dataset and evaluated by the synthetic dataset (column "Trained on M²OLIE dataset").

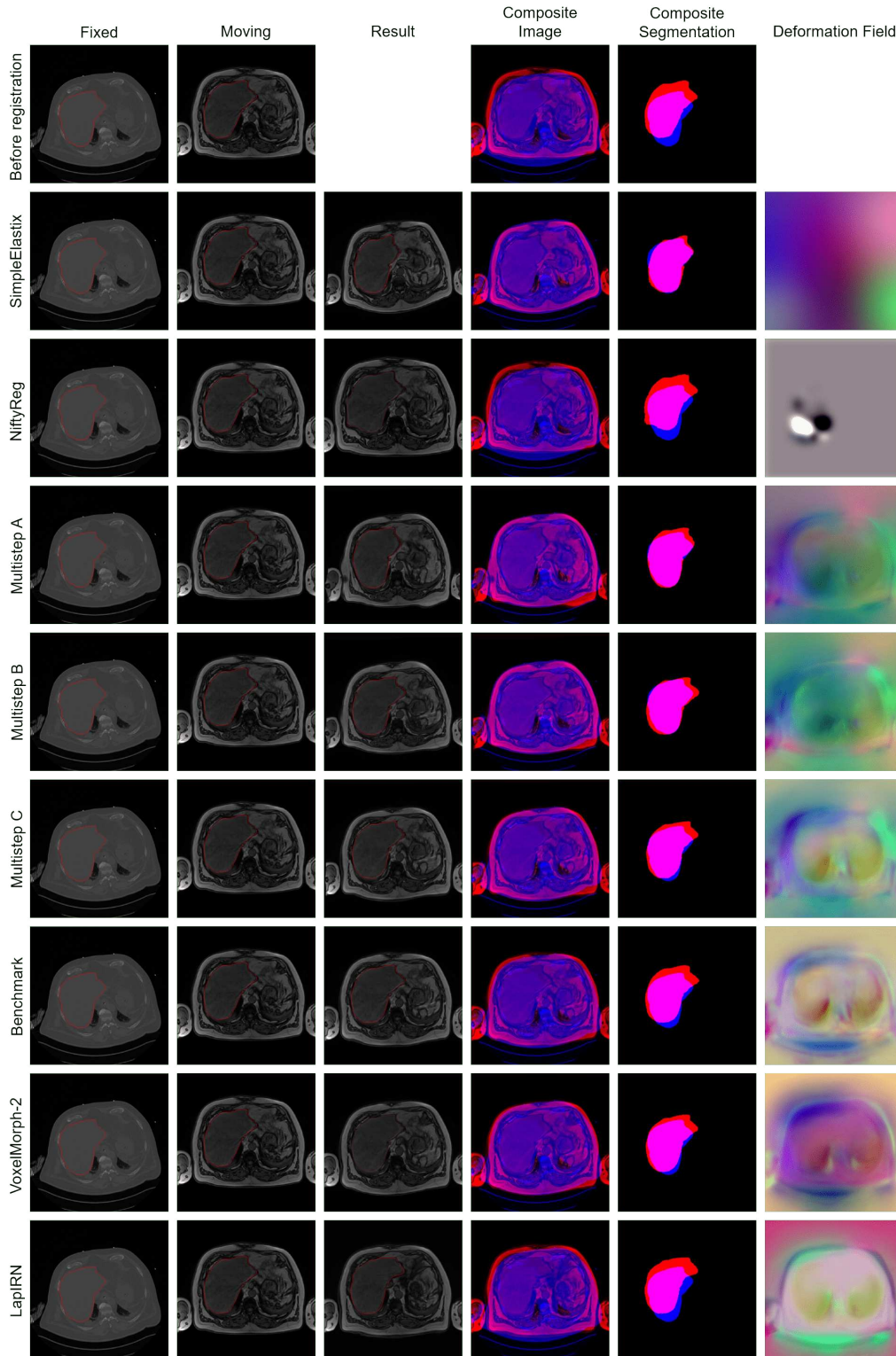


Figure 5.3: Example results of the registration of an MRI to CT volume from the M²OLIE liver dataset. The images show central slices of the axial plane. The fixed images, moving images and result (moved) images are overlaid by the liver segmentations (column 1 - 3). The resulting composites for image and segmentation show the fixed data in blue and the moving/moved data in red. The deformation field is visualized in RGB channels.

All methods significantly ($p < 0.00001$) improved the **NMI** compared to the unregistered images. The baseline SimpleElastix produced a significantly ($p < 0.00001$) higher **NMI** than NiftyReg. Multistep variant C and LapIRN yielded the highest **NMI** and VoxelMorph-2 achieved a significantly ($p < 0.00001$) higher **NMI** than the benchmark.

The highest Dice coefficient was achieved with the baseline SimpleElastix. Multistep variant A resulted in the highest Dice coefficient compared to all neural networks, however, being significantly ($p < 0.00001$) lower than the Dice coefficient of the baseline SimpleElastix. When the number of steps was reduced, the Dice coefficient decreased significantly (variant A vs. B: $p = 0.03337$, variant B vs. C: $p = 0.04444$), comparable to the registration results of the M²OLIE dataset. The benchmark yielded a Dice coefficient significantly lower than the Dice coefficients of the multistep methods (vs. variant C: $p = 0.0019$, vs. baseline SimpleElastix: $p < 0.00001$) but significantly ($p < 0.00001$) higher than before registration. The baseline NiftyReg achieved registration results with a similar Dice coefficient ($p = 0.71138$) as the benchmark. VoxelMorph-2 achieved a similar ($p = 0.90881$) Dice coefficient to Multistep variant C. LapIRN yielded the lowest Dice coefficient. All registration methods achieved a very low number of image folding (Jacobian determinant ≤ 0). Registration of an image pair with SimpleElastix took 26.20 ± 5.93 seconds and with NiftyReg 56.72 ± 9.47 seconds on the synthetic data, significantly longer than using the neural networks, which took less than half a second (last column in Table 5.3).

Table 5.3: The Normalized Mutual Information (**NMI**), Dice coefficient, Jacobian determinant ($|J| \leq 0$) and registration time results (mean $\pm$ SD) of the synthetic liver dataset of the baselines, multistep networks, benchmark, VoxelMorph-2 and LapIRN. **NMI**: higher value is better (maximum is 1), Dice coefficient: higher value is better (maximum is 100%), $|J| \leq 0$: lower value is better (minimum is 0%).

Method/Network	Trained on M ² OLIE dataset			Finetuned on sythetic dataset			Registration Time (s)
	NMI	DICE (%)	$ J \leq 0$ (%)	NMI	DICE (%)	$ J \leq 0$ (%)	
Before registration	0.21 $\pm$ 0.03	66.5 $\pm$ 3.7	-	0.21 $\pm$ 0.03	66.5 $\pm$ 3.7	-	-
Baseline SimpleElastix	0.26 $\pm$ 0.02	89.7 $\pm$ 10.6	0.0 $\pm$ 0.0	0.26 $\pm$ 0.02	89.7 $\pm$ 10.6	0.0 $\pm$ 0.0	26.20 $\pm$ 5.93
Baseline NiftyReg	0.22 $\pm$ 0.03	73.2 $\pm$ 7.2	0.0 $\pm$ 0.0	0.22 $\pm$ 0.03	73.2 $\pm$ 7.2	0.0 $\pm$ 0.0	56.72 $\pm$ 9.47
Multistep A	0.22 $\pm$ 0.02	79.2 $\pm$ 10.8	0.0 $\pm$ 0.0	0.33 $\pm$ 0.02	90.8 $\pm$ 3.1	0.4 $\pm$ 0.2	0.38 $\pm$ 0.03
Multistep B	0.23 $\pm$ 0.02	76.8 $\pm$ 6.8	0.0 $\pm$ 0.0	0.33 $\pm$ 0.02	90.9 $\pm$ 2.9	0.3 $\pm$ 0.2	0.35 $\pm$ 0.06
Multistep C	0.24 $\pm$ 0.02	74.9 $\pm$ 9.4	0.0 $\pm$ 0.0	0.31 $\pm$ 0.02	89.5 $\pm$ 3.1	0.1 $\pm$ 0.1	0.31 $\pm$ 0.05
Benchmark	0.21 $\pm$ 0.02	73.4 $\pm$ 10.6	0.0 $\pm$ 0.0	0.30 $\pm$ 0.02	84.3 $\pm$ 3.8	0.1 $\pm$ 0.2	0.27 $\pm$ 0.06
VoxelMorph-2	0.22 $\pm$ 0.02	74.8 $\pm$ 10.2	0.0 $\pm$ 0.0	0.29 $\pm$ 0.02	88.1 $\pm$ 3.4	0.0 $\pm$ 0.0	0.28 $\pm$ 0.05
LapIRN	0.24 $\pm$ 0.02	69.8 $\pm$ 8.0	0.2 $\pm$ 0.3	0.29 $\pm$ 0.02	86.3 $\pm$ 3.9	0.2 $\pm$ 0.1	0.31 $\pm$ 0.06

The registration results of the neural networks that were finetuned and evaluated with the synthetic dataset can be seen in Table 5.3 column "Finetuned on sythetic dataset".

All networks yielded an **NMI** that was significantly ($p < 0.00001$) higher than the **NMI** of the baselines. The highest **NMI** was achieved with multistep variant A. Decreasing the number of subnetworks significantly ($p < 0.00001$) decreased the **NMI**. Voxelmorph-2 yielded a **NMI** significantly ($p < 0.00001$) lower than the benchmark but significantly higher ($p = 0.01778$) than LapIRN.

The baseline SimpleElastix and multistep variants A and B achieved similar Dice coefficients (variant A vs. baseline SimpleElastix: $p = 0.25428$, variant A vs. B:

$p = 0.4009$, variant B vs. baseline SimpleElastix: $p = 0.23014$). Multistep variant C yielded a significant lower Dice coefficient ($p < 0.00001$). The benchmark achieved registration results with a Dice coefficient significantly lower than all other registration networks and the baseline SimpleElastix (vs. variant C: $p < 0.00001$, vs. baseline SimpleElastix: $p < 0.00001$), but still significantly higher than before registration ($p < 0.00001$). Voxelmorph-2 and LapIRN yielded a significantly ($p < 0.00001$) higher Dice coefficient than the benchmark, but significantly ($p < 0.00001$) lower than the Dice coefficient of multistep variant C. The baseline NiftyReg yielded the lowest Dice coefficient. The largest degree of image folding was obtained with multistep variant A, but still very small (below 0.5%).

Figure 5.4 shows examples of registration results. Before registration, corresponding structures are not well aligned. This is especially visible in the overlap in the abdomen ("Composite Image" in Figure 5.4) and the liver ("Composite Segmentation" in Figure 5.4). Inhalation was simulated in the CT image, whereas exhalation was simulated in the MR image. Registration with the baseline SimpleElastix, multistep variants A, B and C, VoxelMorph-2 and LapIRN resulted in a high spatial alignment. In contrast, the benchmark and the baseline NiftyReg achieved poor registration results and overlap of all corresponding structures. In the results of all neural networks, the largest deformations are located on the arms (sixth column in Figure 5.4). The liver is deformed smoothly.

5.4.3 Experiment III

Table 5.4 shows the results of the registration methods that were trained and evaluated on the in-vivo NIDDK dataset.

The baseline SimpleElastix, the multistep networks A, B and C and LapIRN achieved a significantly ($p < 0.00001$) higher NMI than the unregistered images. The other methods significantly ($p < 0.00001$) worsened the NMI.

Table 5.4: The Normalized Mutual Information (NMI), Dice coefficient, Jacobian determinant ($|J| \leq 0$ and registration time results (mean $\pm$ SD) of the NIDDK kidney dataset of the baselines, multistep networks, benchmark, VoxelMorph-2 and LapIRN. NMI: higher value is better (maximum is 1), Dice coefficient: higher value is better (maximum is 100%), $|J| \leq 0$: lower value is better (minimum is 0%).

Method/Network	NMI	DICE (%)	$ J \leq 0$ (%)	Registration Time (s)
Before registration	0.08 ± 0.04	67.6 ± 18.1	-	-
Baseline SimpleElastix	0.09 ± 0.04	62.1 ± 23.7	0.0 ± 0.0	8.12 ± 0.49
Baseline NiftyReg	0.07 ± 0.04	63.5 ± 19.0	0.0 ± 0.0	1.88 ± 0.14
Multistep A	0.09 ± 0.03	74.7 ± 10.6	0.0 ± 0.1	0.37 ± 0.04
Multistep B	0.08 ± 0.03	69.3 ± 13.9	0.0 ± 0.0	0.34 ± 0.06
Multistep C	0.08 ± 0.03	69.1 ± 14.0	0.0 ± 0.0	0.32 ± 0.04
Benchmark	0.07 ± 0.03	68.5 ± 16.6	0.0 ± 0.0	0.26 ± 0.05
VoxelMorph-2	0.07 ± 0.03	67.7 ± 16.6	0.0 ± 0.0	0.27 ± 0.05
LapIRN	0.08 ± 0.04	68.3 ± 17.4	0.1 ± 0.2	0.31 ± 0.06

The highest Dice coefficient was achieved by the multistep variant A. We observed that if the number of steps/resolutions is reduced, the Dice coefficient decreased

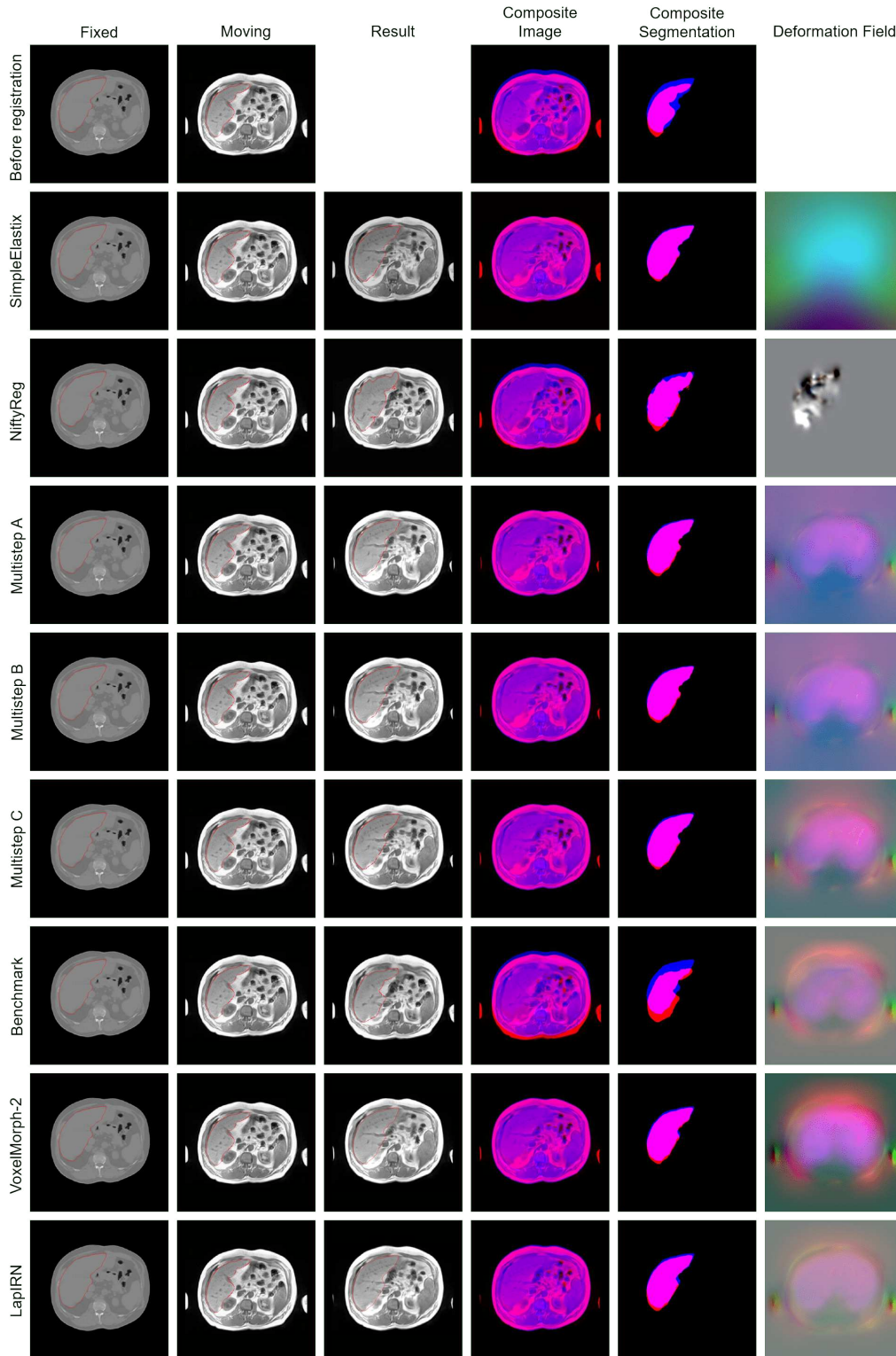


Figure 5.4: Example results of the registration of an MRI to CT volume from the synthetic liver dataset. The images show central slices of the axial plane. The fixed images, moving images and result (moved) images are overlaid by the liver segmentations (column 1 - 3). The resulting composites for image and segmentation show the fixed data in blue and the moving/moved data in red. The deformation field is visualized in RGB channels.

significantly ($p < 0.00001$). Multistep variants B and C achieved similar Dice coefficients ($p = 0.34722$). All neural networks except VoxelMorph-2 yielded Dice coefficients that were significantly higher than before registration ($p < 0.00001$). The baseline SimpleElastix and NiftyReg achieved poor registration results compared to before registration. This can also be observed visually (cf. Figure 5.5).

Before registration, the abdomen is not well aligned in the fixed and moved image (Figure 5.5 first row). This is especially visible in the "Composite Image" at the liver and in the "Composite Segmentation" at the overlap of the kidneys. Although the overlap of the liver has improved through registration with the baseline SimpleElastix, the overlap of the kidneys and spleen has worsened. The baseline NiftyReg could not improve the spatial alignment. The highest spatial alignment was achieved with multistep variant A. As already observed in Table 5.4 in the Dice coefficient, the overlap of the corresponding structures, such as the liver and kidneys, decreased when the steps/resolutions were reduced. LapIRN was able to increase the overlap of the kidneys, but did not improve the overlap of the liver.

5.4.4 Experiment IV

Table 5.5 shows the results of the registration methods that were trained and evaluated on the in-vivo ProstateX dataset. The baselines significantly ($p < 0.00001$) improved the NMI compared to the unregistered images. In contrast, the neural networks yielded a similar NMI as before registration. All registration methods were able to significantly ($p < 0.00001$) improve the SSIM. The highest SSIM was achieved with multistep variant B. Reducing the number of subnetworks (multistep variant C or monostep networks instead of multistep variant B) significantly reduced the SSIM (multistep variant B vs. C and multistep variant C vs. benchmark: $p < 0.00001$). Multistep network A yielded a similar SSIM as multistep variant C ($p = 0.4009$). VoxelMorph-2 achieved a significantly higher ($p < 0.00001$) SSIM than the benchmark, but still significantly lower ($p = 0.03752$) than the SSIM of multistep variant C. LapIRN yielded an NMI and SSIM that were significantly lower ($p < 0.00001$) than the benchmark.

All registration methods yielded a very low number of image folding. The registration of a ProstateX image pair took 8.84 ± 0.54 seconds with SimpleElastix and 6.11 ± 0.45 seconds with NiftyReg on the ProstateX dataset, and thus significantly longer than with the neural networks. These needed less than half a second to register an image pair.

Figure 5.6 shows examples of registration results. Before registration, corresponding structures such as the prostate, bladder, rectum and tissue are not well aligned (white arrow in "Composite Image" in the first row in Figure 5.6). The prostate is stretched vertically in the moving image compared to the fixed image, and the bladder is significantly larger in the moving image than in the fixed image. Registration with SimpleElastix achieved a high deformation correction, i.e. a high spatial alignment of the corresponding structures. In contrast, NiftyReg generated medical implausible anatomy. Multistep variant A achieved similar results as multistep variant B. Registration with these networks resulted in a high spatial alignment of the prostate. The overlap of the bladder, rectum and corresponding tissue (white

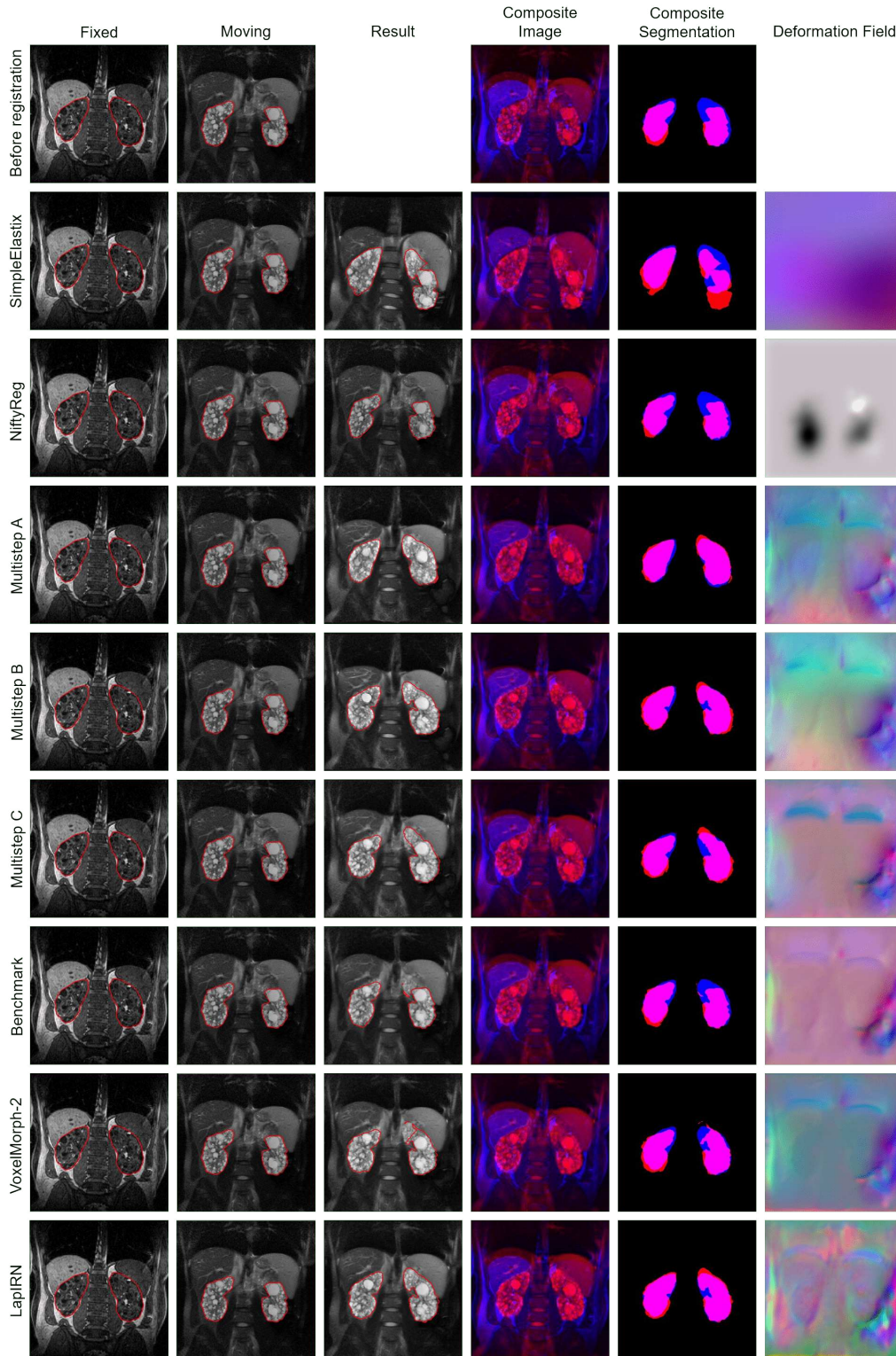


Figure 5.5: Example results of the registration of two volumes from the NIDDK dataset. The images show central slices of the coronal plane. The fixed images, moving images and result (moved) images are overlaid by the kidney segmentations (column 1 - 3). The resulting composites for image and segmentation show the fixed data in blue and the moving/moved data in red. The deformation field is visualized in RGB channels.

Table 5.5: Results of applying the ProstateX dataset to the baselines, multi-step networks, benchmark, VoxelMorph-2 and LapIRN evaluated by the Normalized Mutual Information (**NMI**), Structural Similarity Index (**SSIM**) and Jacobian determinant ($|J| \leq 0$ and registration time results (mean $\pm$ **SD**). **NMI** and **SSIM**: higher value is better (maximum is 1), $|J| \leq 0$: lower value is better (minimum is 0%).

Method/Network	NMI	SSIM	$ J \leq 0$ (%)	Registration Time (s)
Before registration	0.02 ± 0.01	0.03 ± 0.03	-	-
Baseline SimpleElastix	0.04 ± 0.02	0.03 ± 0.04	0.0 ± 0.0	8.84 ± 0.54
Baseline NiftyReg	0.04 ± 0.02	0.03 ± 0.03	0.0 ± 0.0	6.11 ± 0.45
Multistep A	0.02 ± 0.01	0.04 ± 0.02	0.0 ± 0.0	0.37 ± 0.05
Multistep B	0.02 ± 0.01	0.05 ± 0.02	0.0 ± 0.0	0.34 ± 0.05
Multistep C	0.02 ± 0.01	0.04 ± 0.02	0.0 ± 0.0	0.32 ± 0.04
Benchmark	0.02 ± 0.01	0.04 ± 0.02	0.0 ± 0.0	0.27 ± 0.04
VoxelMorph-2	0.02 ± 0.01	0.04 ± 0.02	0.0 ± 0.0	0.27 ± 0.04
LapIRN	0.02 ± 0.01	0.04 ± 0.02	0.0 ± 0.0	0.31 ± 0.06

arrow in column "Composite Image") was also improved. However, the bladder is not spatially aligned well. The multistep variant C yielded similar results as the monostep networks (benchmark and VoxelMorph-2), however, with a lower spatial alignment of the prostate than multistep variants A and B.

The deformation field of the baseline SimpleElastix is smooth. The deformation field of the baseline NiftyReg shows the deformations that led to the medically implausible anatomy in the registration result. The deformations of the neural networks mainly focus on the structures with high intensity values on the DWI scan (bladder and prostate). Nevertheless, the registration result of multistep variant A, for instance, demonstrates that tissue with lower intensity was also spatially aligned well (white arrow in column "Composite Image" in Figure 5.6).

5.4.5 Experiment V

Table 5.6 shows the results of the registration methods that were trained and evaluated on the in-vivo ProstateX dataset, where an anti-aliasing filter (order 8 Chebyshev filter type I) was applied to the images before downsampling. Comparable to experiment IV, the neural networks yielded a similar **NMI** as before registration. All registration methods were able to improve the **SSIM** significantly ($p < 0.00001$). In contrast to experiment IV, in experiment V multistep variants A and B achieved a similar **SSIM** ($p = 0.12852$). The **SSIM** decreased significantly (multistep variant B vs. C and multistep variant C vs. benchmark: $p < 0.00001$) when the number of sub-networks was reduced. VoxelMorph-2 achieved a significantly higher ($p < 0.00001$) **SSIM** than the benchmark, however still significantly lower ($p < 0.00001$) than multistep variant C. And LapIRN yielded an **NMI** and **SSIM** that were significantly lower ($p < 0.00001$) than the benchmark. All registration methods yielded a very low number of image folding.

The registration results of experiment IV and V (**NMI** and **SSIM**) were compared with each other. Table 5.7 indicates that the networks in experiment V yielded either

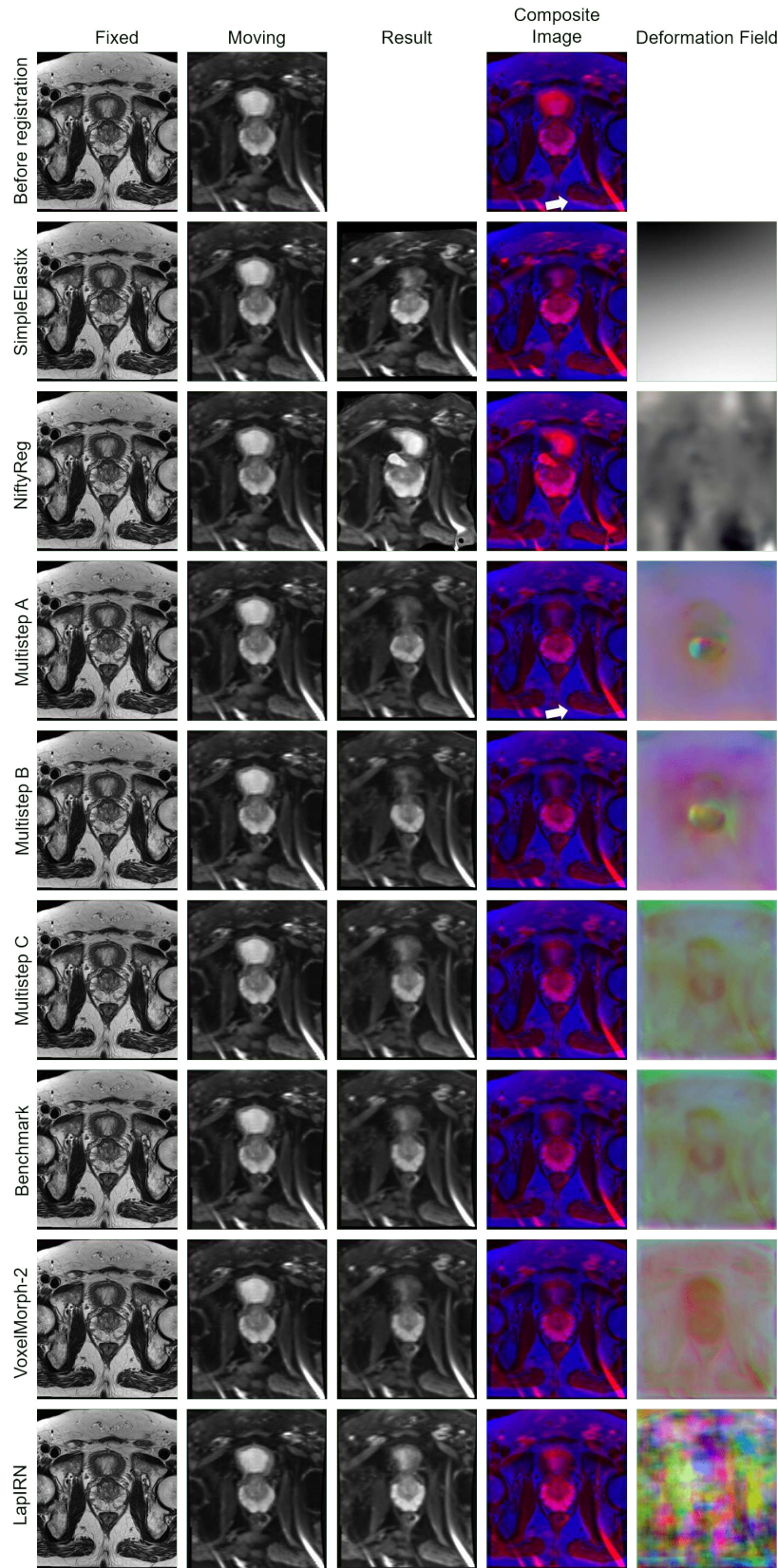


Figure 5.6: Example results of the registration of two volumes from the ProstateX dataset. The images show central slices of the coronal plane. The resulting composites show the fixed data in blue and the moving/moved data in red. The deformation field is visualized in RGB channels.

Table 5.6: Results of applying an anti-aliasing filter (order 8 Chebyshev type I filter) before downsampling the images of the ProstateX dataset (experiment V). The Normalized Mutual Information (**NMI**), Structural Similarity Index (**SSIM**) and Jacobian determinant ($|J| \leq 0$ and registration time results (mean $\pm$ **SD**) of the baselines, multistep networks, benchmark, VoxelMorph-2 and LapIRN are given. **NMI** and **SSIM**: higher value is better (maximum is 1), $|J| \leq 0$: lower value is better (minimum is 0%).

Method/Network	NMI	SSIM	$ J \leq 0$ (%)	Registration Time (s)
Before registration	0.02 ± 0.01	0.03 ± 0.03	-	-
Baseline SimpleElastix	0.04 ± 0.02	0.03 ± 0.04	0.0 ± 0.0	8.84 ± 0.54
Baseline NiftyReg	0.04 ± 0.02	0.03 ± 0.03	0.0 ± 0.0	6.11 ± 0.45
Multistep A	0.02 ± 0.01	0.04 ± 0.02	0.0 ± 0.0	0.37 ± 0.05
Multistep B	0.02 ± 0.01	0.04 ± 0.02	0.0 ± 0.0	0.34 ± 0.05
Multistep C	0.02 ± 0.01	0.04 ± 0.02	0.0 ± 0.0	0.32 ± 0.04
Benchmark	0.02 ± 0.01	0.04 ± 0.02	0.0 ± 0.0	0.27 ± 0.04
VoxelMorph-2	0.02 ± 0.01	0.04 ± 0.02	0.0 ± 0.0	0.27 ± 0.04
LapIRN	0.02 ± 0.01	0.04 ± 0.02	0.0 ± 0.0	0.31 ± 0.06

significantly worse or similar results than the corresponding networks in experiment IV. Using an anti-alias filter did not improve the performance of the networks on the ProstateX dataset.

Table 5.7: P-values of the significance tests (paired t-test or a Wilcoxon signed-rank test, $p < 0.05$) for the Normalized Mutual Information (**NMI**) and Structural Similarity Index (**SSIM**) of the proposed multistage networks: experiment IV vs. experiment V. Metrics with a significant difference ($p < 0.05$) are indicated in bold.

Method/Network	NMI	SSIM
Multistep A	<0.00001	<0.00001
Multistep B	<0.00001	0.41794
Multistep C	<0.00001	0.33204
Benchmark	<0.00001	0.01684
VoxelMorph-2	<0.00001	0.93624
LapIRN	0.65994	0.00880

Figure 5.7 shows examples of registration results. This shows that the spatial alignment and especially the overlap of the bladder and prostate deteriorated slightly for all networks compared to experiment IV. The deformations in the organs are smooth for all registration methods.

5.5 Discussion

We proposed neural networks for deformable, multimodal medical image registration that apply multiple steps and resolutions. By comparing them with each other and with a monostep/ monoresolution network as a benchmark, we investigated the

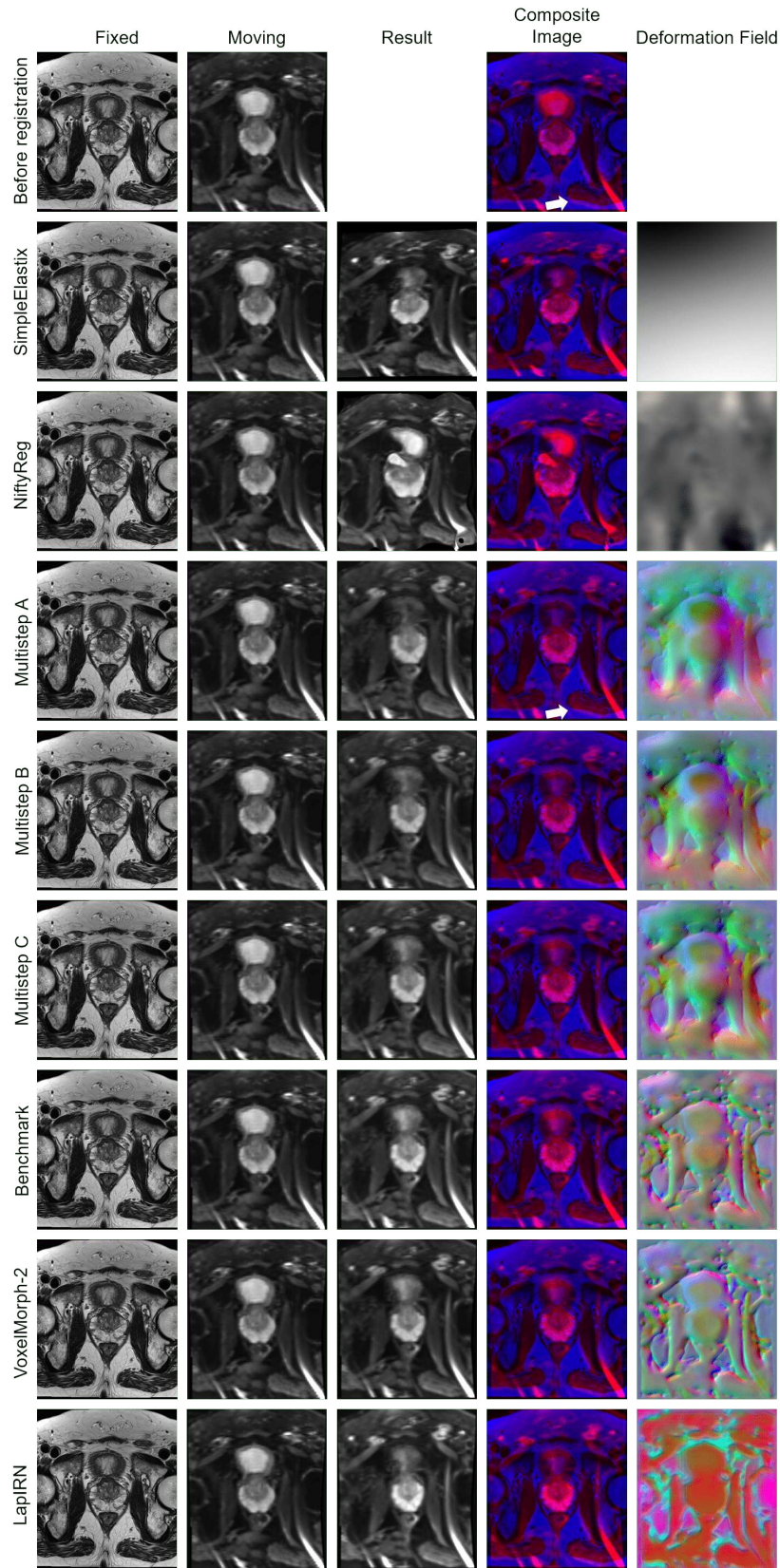


Figure 5.7: Example results of the registration of two volumes from the ProstateX dataset of experiment V. The images show central slices of the coronal plane. The resulting composites show the fixed data in blue and the moving/-moved data in red. The deformation field is visualized in RGB channels.

impact of using multiple resolutions on the registration result and potential benefits and trade-offs of cascaded multiresolution registration networks with up to four steps/ subnetworks.

5.5.1 Comparison to Published Multistep-Multiresolution Registration Methods

Classical (iterative) methods such as SimpleElastix [17] or NiftyReg [18, 19] apply multistep-multiresolution methods using a multi-resolution Gaussian pyramid. SimpleElastix applies four resolutions with a downsampling factor of [2.80322, 1.9881, 1.41, 1]. In contrast, NiftyReg uses 3 steps/resolutions, with a different downsampling factor ([4, 2, 1]). As our experiments have shown, registration with the classical methods SimpleElastix and NiftyReg is time consuming, since an iterative registration process is performed for each new image pair.

Applying deep learning decreases the registration time. In recent years, several neural networks have been developed that use a coarse-to-fine approach [86, 129–140]. This involves cascading multiple subnetworks with different resolutions. However, in the published multistep papers, the cascaded subnetworks are only used in combination.

We therefore implemented multistep networks with varying numbers of resolutions (four, three and two steps/resolutions) and compared their performance with each other and with the performance of a monostage/ monoresolution network. As the architecture of the subnetworks, we chose to adapt the VoxelMorph-2 network, as this is a widely used architecture/ benchmark. In general, any network architecture can be used as the architecture of the subnetworks. Future work could therefore include using additional subnetwork architectures, for example from published multistep papers [86, 129–140].

5.5.2 Individual Networks

In our experiments, we compared the performance of our multistep networks to a benchmark and two baselines using four different datasets and several metrics.

When comparing NMI and Dice coefficient in Table 5.2, it is noticeable that the change in NMI and Dice coefficient by registration was not always uniform. For instance, the multistep network variant A in experiment I achieved the highest Dice coefficient of all neural networks, but not the highest NMI. The reason for this is that the NMI, i.e. the intensity distribution, is calculated globally without taking spatial information into account. Future work could apply a local metric, e.g. a local NMI [114]. Therefore, as a second metric, the Dice coefficient was used to evaluate the overlap of the target structures. In this research, the focus for evaluating the performance of the registration methods therefore lies on the Dice coefficient.

The impact of introducing multiple resolutions into a neural network on the registration result varied. For the M²OLIE dataset, multistep variant A achieved a similar Dice coefficient as the baseline SimpleElastix. When evaluating these networks on the synthetic dataset, the baseline SimpleElastix achieved a significantly higher Dice coefficient than multistep variant A. This was due to the fact that the

multistep networks were only trained on the M²OLIE dataset, but not on the unseen synthetic dataset. After finetuning the networks with the synthetic dataset, multistep network variants A and B achieved similar Dice coefficients as the baseline SimpleElastix. For the NIDDK dataset, multistep variant A also achieved the highest spatial alignment, which decreased significantly when the number of steps/resolutions was reduced. Multistep variants B and C achieved similar Dice coefficients. In contrast to the liver datasets, the baseline SimpleElastix achieved poor registration results here. Various parameter settings for SimpleElastix were tested, however, no good spatial alignment could be achieved. For all datasets, NiftyReg could not achieve a good deformation correction, despite the additional information (segmentation of the fixed and moving image; for the liver and kidney datasets) and testing different parameter settings. When the number of steps/resolutions of the neural networks was reduced in experiment I-III, the Dice coefficient deteriorated significantly. In these experiments, the monostep/monoresolution networks (benchmark and VoxelMorph-2) achieved the worst registration results out of all neural networks (except LapIRN). The monoresolution approaches were not sufficient to achieve a good deformation correction and a high spatial alignment.

For the ProstateX dataset in experiment IV, however, multistep variant B delivered the best spatial alignment (based on SSIM). Using an additional subnetwork (multistep variant A instead of B) did not deliver better results. The spatial alignment decreased with the use of four subnetworks and is similar to the SSIM of multistep C with only two subnetworks. The reason for this could be that adding subnetworks increases the number of parameters to be optimised during training (see Table 5.1). This results in a more complex network that is more difficult to train. In contrast, in experiment V, multistep variants A and B yielded a similar SSIM. The SSIM decreased significantly when the number of steps/resolutions was reduced.

For all datasets, the LapIRN monostep-multiresolution network achieved registration results with a lower spatial alignment (based on Dice coefficient and SSIM, as well as in optical evaluation) than the multistep networks presented in this paper. This indicates that a multistep-multiresolution approach may be preferable to the monostep-multiresolution network for the tasks and datasets in this study.

For all datasets, all registration methods (except NiftyReg on the ProstateX dataset) achieved smooth deformations within the organs and a low degree of image folding, leading to a medical plausible transformation. In the deformation fields of the baseline SimpleElastix, in contrast to the deformation fields of the neural networks, no structures such as abdominal or organ contours could be recognized. The reason for this is that for the baseline SimpleElastix, the final grid spacing (parameter "FinalGridSpacingInPhysicalUnits") was set to a high value (100 x 100 x 100 mm³) to achieve smooth and medically plausible transformations. Lower grid spacing achieved medically implausible deformations. Whereas for the neural networks, the grid spacing of the final deformation field was 2 x 2 x 4 mm³ (size of the fixed and moving image). The deformation fields of the baseline NiftyReg are predominantly uniformly medium grey, deformations are only visible in the segmentation area. The reason for this is that, in addition to the fixed and moving image, the segmentations of the fixed and moving image were also used for the registration. Thus, the compu-

tation of the deformation mainly focuses on the segmentation area. Gray values in the deformation field indicate a similar deformation for all three spatial directions.

Registration with the baselines took significantly longer than with the neural networks. The reason for this is that SimpleElastix and NiftyReg apply an iterative algorithm instead of deep learning. GPU implementations [80] or approaches that use local similarity estimations such as point similarity measures [147] significantly reduce the registration time. However, to employ a baseline for this work, SimpleElastix and NiftyReg were chosen as they are widely accepted and apply multiple resolutions during the registration process. Differing registration times were also measured among the different neuronal networks. The fewer steps were used, the faster the registration of a new image pair. However, the registration always took less than half a second. The difference in registration time between the different neural networks is very small and therefore not a crucial reason for choosing a network with a lower number of steps.

5.5.3 Individual Datasets

By using the four datasets, we investigated the impact of using multiple resolutions on the registration result on both in-vivo and synthetic data. The registration results of the synthetic liver dataset had a lower standard deviation compared with the registration results of the patient liver dataset. This is due to the fact that in the synthetic dataset, only respiration was modeled. In the patient dataset, in addition to deformation due to respiration, deformation due to differences in patient positioning in the CT and MR scanners were present, with large differences in patient positioning in some cases. For instance, there were seven cases in which patients were not supine in the CT scanner but in a more lateral position, whereas they were supine in the MR scanner. For these image pairs, the spatial alignment after the registration was worse than for image pairs where the patient’s positioning in the CT and MR scanners was similar. Differences in patient positioning in the corresponding images could be reduced by adding additional pre-processing steps. For example, as in [86, 95, 120] an affine network or, as in [10] a rigid and affine network could be used before the deformable registration to correct large global deformations, e.g. rotation and translation. The succeeding deformable subnetworks then correct smaller local deformations.

The liver segmentations of the synthetic dataset were directly generated by the XCAT phantom. In contrast, the liver segmentations of the M²OLIE dataset were created by neural networks. The results of these segmentation networks and variations in the quality of the CT and MR segmentations may greatly influence segmentation-based metrics, and consequently, the evaluation based on the Dice coefficient.

In the NIDDK dataset, the slice thickness of the T1-weighted MR scans was 3 mm, whereas it was three times as much (9 mm) for the T2-weighted MR scans. Therefore, the T2-weighted slice shown in Figure 5.5 is generated by interpolation from multiple slices. The multistage network registers the T2-weighted MR scans on the corresponding T1-weighted MR scans, and also utilizes the T1-weighted MR scans to reconstruct thinner (3 mm) T2-weighted slices.

For the ProstateX dataset, no segmentations were available, so we could not calculate a Dice coefficient. However, we checked the overlap of corresponding structures

using a visual evaluation. The registration results of the ProstateX dataset have a lower spatial alignment compared to the registration results of the other three datasets. Especially the bladder could not be spatially aligned well. The reason for this is that DWI scans should be registered in the ProstateX dataset. DWI scans contain less morphology compared to T1-weighted or T2-weighted MRI scans or CT images. This complicates the registration of corresponding structures, which was also demonstrated during the training of the networks, in which the Mutual Information only improved slightly. In further work, segmentations of different organs or target structures such as prostate, rectum and bladder could be created for this dataset. These could then be used for semi-supervised training. That is, in addition to the fixed and moving image, the segmentations of the fixed and moving image are also input as additional information for the training of the registration networks. The networks could then be trained with Mutual Information (of the fixed and moving image) and additionally with a segmentation-based metric such as Dice coefficient (of the segmentation of the fixed and moving image) in order to further improve the spatial alignment and the overlap of the target structures.

5.5.4 Limitations

The evaluation of image registration is challenging due to the lack of ground truth data [7]. Therefore, we utilized different methods to evaluate our registration results. We evaluated the structural similarity of fixed and moved image using NMI. For the ProstateX dataset, we also calculated the SSIM of fixed and moved image. This metric evaluates luminance, contrast and structure. In three of the four datasets used, organ segmentations were available in the fixed and moving images. Thus, we applied the Dice coefficient to assess the overlap of the target structures before and after registration. The spatial alignment was examined visually by color overlaying images or segmentations. Furthermore, the evaluation of the deformation field included the calculation of the Jacobian determinant for each point, which revealed image folding. Additionally, the deformation field was evaluated visually, which highlighted the areas with smooth or larger deformations.

In our study, the training of the multistep network with up to four subnetworks was limited by the available amount of GPU memory (NVIDIA RTX A6000 GPU with 48 GB of GPU memory). Thus, the size of the fixed and moving image was limited to 256 x 256 x 64 (with a resolution of 2 x 2 x 4 mm³). Nevertheless, we could demonstrate the impact of multistep/ resolution networks to multi-modal image registration.

Downsampling, i.e. using input images at a lower resolution, could limit the registration result, as very small deformations may not be represented in the downsampled image and therefore cannot be corrected. Downsampling can also lead to smaller vessels or finer structures, such as in retinal images, being lost or no longer sufficiently identifiable in the downsampled images, especially if they are close to the resolution limit of the system. However, it is important to note that the actual effects of downsampling may depend on various factors, including the particular application and its image resolution requirements. In some cases, important image features may be preserved even at greatly reduced resolution, while in other cases a higher resolution may be required to perform an accurate analysis.

Downsampling can also lead to alias effects and thus impact performance. We investigated this in experiment V by adding an anti-aliasing filter before downsampling the image. In our study (experiment IV vs V, ProstateX dataset), however, using an anti-alias filter did not improve the performance of the networks. The impact of an anti-alias filter on the performance of the networks could depend on the network architectures and datasets used. This could be investigated in more detail in a follow-up study.

The aim of future work would therefore be to process the images at full resolution. For example, the multistep networks could be modified by adding a sliding window. This enables the processing of several layers of the full-resolution images at once (e.g. with a sliding window of $512 \times 512 \times 8$, 8 layers are processed of a $512 \times 512 \times 128$ image) and the calculation of a deformation field of the same size as the sliding window. The sliding window is then moved across the image. All deformation fields are then accumulated to form the final deformation field/ final transformation.

5.6 Conclusion

We proposed neural networks for deformable multimodal medical image registration that use multiple steps and varying resolutions. Our results demonstrate that incorporating multiple steps and resolutions within the neural network framework leads to registration results with high structural similarity and minimal image folding, resulting in a medically plausible transformation, while maintaining a low registration time of less than half a second. Furthermore, our results show that the spatial alignment improves with increasing number of resolutions/steps (up to a maximum of 4 in this study) and multistep networks outperform monostage/monoresolution networks.

6. A Groupwise Multiresolution Network for DCE-MRI Image Registration

This chapter is based on the following manuscript submitted to Scientific Reports on 12.09.2024: A. Strittmatter, M. Weis, F. G. Zöllner. A Groupwise Multiresolution Network for DCE-MRI Image Registration.

6.1 Introduction

In image registration, corresponding points or structures of two or more images are spatially aligned. For instance, in co-registration, multiple images of different acquisitions or modalities can be spatially aligned, which enables the fusion of the images and their (complementary) information [109], e.g. the fusion of CT and MR scans for multimodal minimal-invasive image-guided interventions [15]. In addition, image registration can be used to correct movement between different scans of the same acquisition. For instance, in four-dimensional time series such as Dynamic Contrast-Enhanced (DCE) MRI, the motion between individual time steps due to the patient's breathing or movement can be corrected. Therefore image registration is an important step in medical image analysis and helps to improve various applications.

One medical application that benefits from image registration is the assessment of lung perfusion in patients with Congenital Diaphragmatic Hernia (CDH). Briefly, CDH is a birth defect in which the diaphragm did not close completely during the foetal development [26]. The foetal abdominal organs are partially located in the thoracic cavity, which then leads to impaired lung development and can also result in cardiovascular malformations. Prenatal diagnosis of diaphragmatic hernia plays a crucial role and allows early planning for prenatal interventions [27]. In addition to prenatal ultrasound imaging, foetal MRI is also used to examine the pulmonary function as well as the lung anatomy. DCE-MRI can be used to determine quantitative parameters of lung perfusion [28]. This involves acquiring a time series of three-dimensional MR scans to monitor the distribution of the administered contrast agent. The four-dimensional data can then be used to calculate perfusion maps, e.g. mean transit time, pulmonary blood volume or pulmonary blood flow [7].

In patients with CDH, DCE-MRI is performed in free-breathing. Due to the patient's respiration between the individual time step scans, the patient is at a different point in the respiratory cycle in each time step scan. This results in a varying deformation of the organs, such as the lungs, heart and liver in each time step. The varying deformation of the corresponding structures in the individual time steps leads to an inaccuracy in the resulting perfusion map. By spatially aligning the corresponding structures before calculating the perfusion maps, the accuracy of the resulting perfusion map can be improved.

In recent years, some methods have been published for the registration of DCE-MR images [7, 148–151]. Ruthotto et al. [149] and Cai et al. [150] apply pairwise

registration, whereby one image is defined as a reference/fixed image and the other images are defined as moving images. The moving images are then transformed with the transformation calculated by the registration process to create the moved images. For four-dimensional images like DCE-MR scans, groupwise registration can be applied, where more than two images are spatially aligned [148, 151, 152]. Setting a fixed image is usually omitted, thus avoiding the selection bias of a fixed image [70]. Instead, a template image is calculated from the images to be registered, e.g. the average intensity image. A transformation is calculated for all images to align them with the template image. In recent years, several groupwise medical image registration methods have been published, that apply iterative/classical methods [17, 70, 153–158] or deep learning-based approaches [159–167].

Particularly in the field of medical image registration, the deformation cannot usually be corrected in a single step with a single resolution [75]. Instead, methods that utilise multiple resolutions are applied, such as the classical registration methods, e.g. SimpleElastix [17], or deep learning-based registration methods [11, 86, 129, 131, 132, 138, 139].

To achieve accurate four-dimensional image registration, specifically on DCE-MR scans, we propose a novel groupwise deformable registration framework that utilises multiple resolutions. The groupwise registration network can be implemented with an arbitrary group size. We performed extensive experiments and evaluated the groupwise registration network with two different group sizes: groups of 5 and groups of 10 three-dimensional volumes from the four-dimensional DCE-MR dataset. We analyse the effect of groupwise registration by comparing it with pairwise registration using the same network architecture. Moreover, we compared the performance of the proposed groupwise registration network with a pairwise and groupwise registration using the classical registration method SimpleElastix as a baseline and the published groupwise network GroupRegNet [164] to show the superiority of the proposed registration network.

6.2 Methods

6.2.1 Network Architectures

Figure 6.1 shows the architecture of the groupwise deformable registration network. The group size n , i.e. the number of images to be registered, can be chosen arbitrarily. The images to be registered are concatenated and form the input of the groupwise registration network.

As underlying network architecture, we utilised a multiresolution network [11], as this demonstrated good performance for various multimodal medical imaging datasets and applications. Four subnetworks are used, which process the images in increasingly higher resolution. Each subnetwork calculates n deformation fields, i.e. one deformation field per image to be registered. Subsequent subnetworks then process the previously transformed image using the deformation fields of the preceding subnetworks. The aim is to first reduce coarser deformations in the first subnetworks and then to correct increasingly finer deformations in the subsequent subnetworks with increasingly higher resolution. The calculated transformation is applied to the images using a spatial transformer [106], resulting in the moved images.

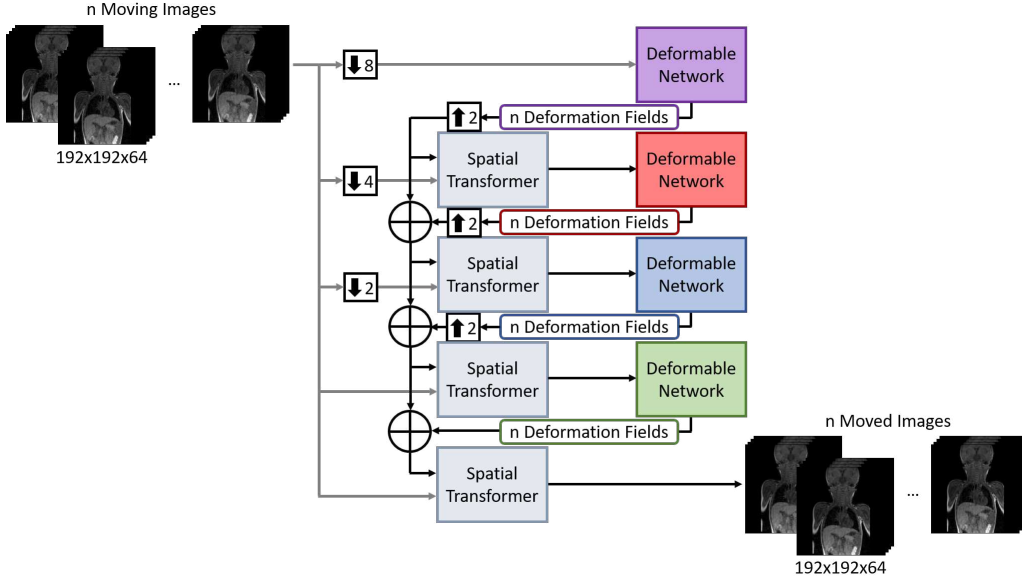


Figure 6.1: Architecture of the groupwise multistep network. As subnetworks, we adapted VoxelMorph-2 [11].

6.2.2 Benchmark Networks

As a benchmark network, we applied a pairwise deformable registration [11] with the same network architecture as the groupwise registration. Here, two images, the fixed (time step 1) and the moving image (remaining time steps), are spatially aligned. A transformation is calculated, which is then applied to the moving image using the spatial transformer. Furthermore, we utilised GroupRegNet [164] as an additional benchmark network.

6.2.3 Training Setting for the Neural Networks

We trained the networks unsupervised using MI and Gradient L_2 as loss functions.

In groupwise registration, the Mutual Information is calculated between each moved image M and the template image T based on the entropy H :

$$L_{MI}(T, M) = I(T; M) = H(T) + H(M) - H(T, M) \quad (6.1)$$

For each group of input images, the template image T is calculated anew by using the mean intensity of the input images. For pairwise registration, the Mutual Information is calculated between the fixed and moved image.

The Gradient L_2 loss is calculated for groupwise and pairwise registration using the deformation fields ϕ :

$$L_2(\phi) = \sum_{p \in \Omega} \|\nabla u(p)\|^2, \quad (6.2)$$

and is utilised to regularise and smoothen the deformation field.

The Mutual Information and Gradient L_2 loss are combined using a λ parameter. For groupwise registration, the Gradient L_2 loss of the n deformation fields is weighted with λ . The combined loss function is therefore calculated by:

$$L(T, M, \phi) = \sum_{i=1}^n L_{MI}(T_i, M_i) + \lambda L_2(\phi_i) \quad (6.3)$$

We used five-fold cross-validation. The networks were trained for 200 epochs with early stopping, where training was ended if the validation loss has not improved in five consecutive epochs. The model weights with the highest validation accuracy were selected and saved. We used the Adam optimiser and a batch size of 1. We tuned the learning rate and the λ -parameter by a grid hyperparameter search. The parameter configuration that yielded the highest Mutual Information on the validation dataset was used to train the networks. We resized the network input images to 192 x 192 x 64 voxels with a spatial resolution of 2 x 2 x 2 mm³ and normalised the intensities of the input images to the range [0, 1].

Our experiments were performed using Python with the Tensorflow framework [107]. The networks were trained on a NVIDIA A100 GPU with 80 GB of GPU memory from the bwForCluster Helix (<https://wiki.bwhpc.de/e/Helix>). Our source code is available at https://github.com/Computer-Assisted-Clinical-Medicine/Groupwise_Multiresolution_Deformable_Medical_Image_Registration.

6.2.4 Baseline Methods

As baseline, we used the classical iterative registration method SimpleElastix. We applied the pairwise bspline algorithm with default parameter settings. As a second and third baseline method, we used the groupwise registration with SimpleElastix with group sizes 5 and 10 and default parameter settings. To achieve medically plausible registration results, we adjusted the parameter "FinalGridSpacingInPhysicalUnits" to ['50', '50', '50'] for SimpleElastix pairwise and groupwise registration.

6.2.5 Dataset

The dataset of this retrospective study comprises 30 four-dimensional DCE-MR scans of 2-year-old patients after CDH repair following a standardized protocol as described by Weidner et al. [28, 168]. Briefly, each patient underwent a DCE-MR scan using a three-dimensional TWIST sequence for DCE-MRI lung perfusion imaging at a 3T scanner. The three-dimensional scans were obtained using a matrix of [126-162] x 192 voxels with 52-56 slices reconstructed to an isotropic voxel size of 1.97 mm³. In total, 50 time steps, i.e. volumes, were acquired with a temporal resolution of 1.5 s.

Contrast agent (Dotarem, Guerbet, France) was injected at a flow rate of 1 mL/s after the fifth time step at a dose of approximately 1 mL (0.05 mmol/kg body weight) of the contrast agent mixed with an equal volume of sodium chloride. The contrast agent injection was immediately succeeded by a 10 mL bolus of sodium chloride [169]. All acquisitions were acquired in free breathing, the patients received sedation.

For the first three-dimensional volume (first time step), lung segmentations were created semi-automatically using a segmentation tool [169] implemented in the DICOM Workstation OsiriX [170].

6.2.6 Evaluation Metrics

We evaluated the registration results using several metrics. We calculated the **SSIM** between the moved images and the template images of the groupwise registration with a group size of 5 three-dimensional volumes (called **SSIM 5**) and a group size of 10 three-dimensional volumes (called **SSIM 10**). Additionally, we analysed the degree of image folding as in Strittmatter et al. [11]. Image folding leads to medically implausible deformations and is caused by a Jacobian determinant ≤ 0 in the deformation field. We calculated the relative value of Jacobian determinants ≤ 0 in the deformation field. Furthermore, we measured the time required to register a whole four-dimensional **DCE-MR** scan with 50 three-dimensional volumes / time steps.

To analyse the deformation of the organs over time / time steps we placed a rectangular **Region Of Interest (ROI)** on the liver dome (Figure 6.2a). We calculated the **ROI** magnitude and its standard deviation over the time steps. The offset of the **ROI** magnitude was excluded, as this only depends on the intensity of the tissue and contrast agent administration and spread in the body and does not reflect the deformation of the target structures. For this purpose, a second-order polynomial was fitted to the uncorrected **ROI** magnitude curve (Figure 6.2b). This polynomial was subtracted from the **ROI** magnitude. The corrected **ROI** magnitude therefore only contains components that are dependent on the deformation of the target structures.

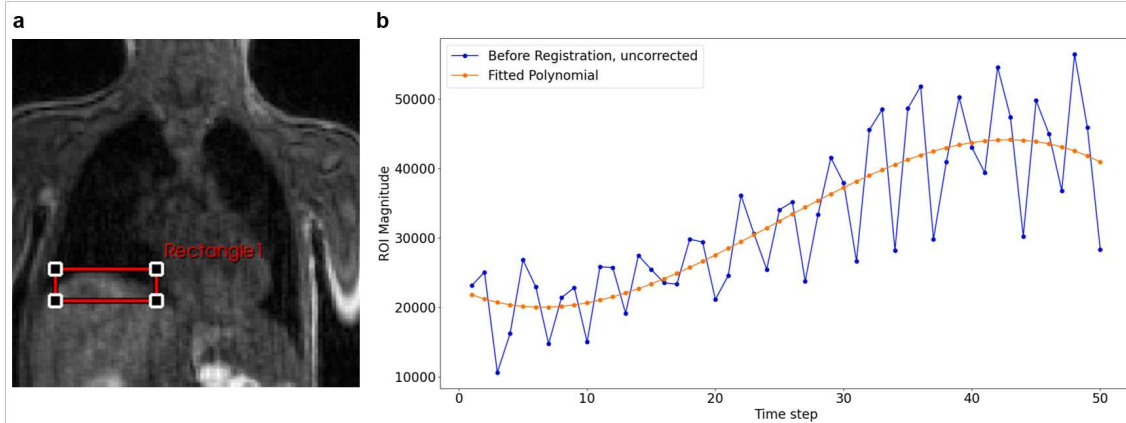


Figure 6.2: a: **ROI** position b: Uncorrected **ROI** magnitude of the time series before registration plotted in blue and fitted polynomial plotted in orange.

Additionally, we assessed the registration results visually. For this purpose, we displayed the scans of two time steps of a **DCE-MR** scan. A central coronal slice was selected from the three-dimensional volumes, cropped and overlaid with the lung segmentation. In addition, we colour-coded and overlaid the scans of the two time steps in a composite image. For this, the scan of the first time step was displayed in blue and the scan of the other time step in red. Moreover, the deformation field was analysed visually to identify the regions with larger deformations. For

the three-dimensional volumes of individual time steps, the registration methods calculate four-dimensional deformation fields. The last dimension has the value 3, i.e. the displacement is calculated in each of the three spatial directions of the three-dimensional volumes. The deformation field is displayed in RGB colours. The red channel describes the displacement on the y-axis, i.e. the image's height, green the displacement on the x-axis/width and blue the displacement on the z-axis/slice direction.

Moreover, for this four-dimensional DCE-MR scan, we calculated the perfusion parameters using the OsiriX plugin UMMPerfusion [171, 172]. UMMPerfusion calculates voxel-wise maps of the pulmonary perfusion (Pulmonary Blood Flow (*PBF*)), Pulmonary Blood Volume (*PBV*), and the Mean Transit Time (*MTT*). For the calculation an Arterial Input Function (*AIF*) has been carefully selected by placing a ROI in the right pulmonary artery in the unregistered scans for each patient. The ROI was then transferred to all registration results. In case that the ROI of the AIF did not match the lumen of the pulmonary artery any more (due to the registration), it was replaced to avoid partial volume effects.

As in Weis et al. [173], to analyse the parametric maps, we then manually positioned three circular ROIs on each lung (ipsilateral and contralateral) apically, centrally and basally. These ROIs were placed through five neighbouring slices to obtain cylindrical ROIs. The ROIs were positioned ventrally, on central slices and dorsally. The mean of the ROIs and the mean standard deviation within the ROIs were calculated to determine whether registration of the DCE-MR scans subsequently resulted in more homogeneous perfusion maps, i.e. a reduction in the standard deviation of the ROIs.

We performed statistical hypothesis tests to assess the statistical significance of the improvements in the SSIM and standard deviation of the ROIs in the perfusion maps. We compared the unregistered data with the registered data of all methods, and the results of the registered methods with each other. We conducted a paired t-test for normally distributed data and a Wilcoxon signed-rank test for non-normally distributed data. Our null hypothesis states that image registration did not lead to improvements in the SSIM and standard deviation of the ROIs in the perfusion maps. Our null hypothesis is rejected if $p < 0.05$. We performed Bonferroni correction.

6.2.7 Experiment

We evaluated the performance of the proposed groupwise registration network on the four-dimensional CDH dataset using two group sizes. We used groups of 5, called Groupwise 5 network in the following, and groups of 10 three-dimensional volumes / time steps, called Groupwise 10 network in the following. The Groupwise 5 network was trained with a learning rate of 1e-4 and a λ -parameter of 2. For training the Groupwise 10 network, we chose a learning rate of 1e-4 and a λ -parameter of 1. The benchmark pairwise network was trained with a learning rate of 5e-4 and a λ -parameter of 20. As additional benchmark, the GroupRegNet was applied with a group size of 5 and 10, called GroupRegNet 5 and GroupRegNet 10. The GroupRegNet 5 was trained with a learning rate of 1e-5 and a λ -parameter of 5. The GroupRegNet 10 was trained with a learning rate of 1e-4 and a λ -parameter of 2.

6.3 Results

Table 6.1 shows the registration results. The pairwise and groupwise registration with SimpleElastix (groups of 5 and 10) reduced the SSIM 5 and SSIM 10 significantly ($p < 0.00001$) compared to before registration. The proposed groupwise networks were able to increase the SSIM 5 and SSIM 10 significantly ($p < 0.00001$). The Groupwise 5 network produced a significantly higher SSIM 5 ($p < 0.00001$) but significantly lower SSIM 10 ($p < 0.00001$) than the Groupwise 10 network. The pairwise network yielded a similar SSIM 5 ($p = 0.24635$) as before registration but a significantly higher SSIM 10 ($p < 0.00001$) compared to before registration. The SSIM 5 and SSIM 10 of the GroupRegNet 5 network are significantly higher than before registration ($p < 0.00001$), the SSIM 5 is significantly lower ($p < 0.00001$) than the Groupwise 5 network but the SSIM 10 is significantly higher ($p < 0.00001$) than the Groupwise 5 network. Registration with the GroupRegNet 10 network significantly reduced the SSIM 5 ($p < 0.00001$) but significantly ($p < 0.00001$) improved the SSIM 10.

Registration with all neural networks (except GroupRegNet 5 on SSIM 10) resulted in a reduction in the standard deviation of the SSIM 5 and SSIM 10.

All registration methods yielded registration results with a very low degree of image folding ($|J| \leq 0: 0.0 \pm 0.0 \%$). The Baseline SimpleElastix pairwise required more than 13 minutes, SimpleElastix groupwise 5 about 14 minutes and SimpleElastix groupwise 10 about 16 minutes to register a four-dimensional DCE-MR scan (50 three-dimensional volumes / time steps), significantly longer than the neural networks, which required less than 10 seconds.

Table 6.1: The Structural Similarity Index (SSIM), registration time and standard deviation of the corrected liver dome ROI magnitude (mean $\pm$ SD) results of the CDH dataset of the baselines SimpleElastix, groupwise networks and the benchmark networks. SSIM: higher value is better (maximum is 1), SD ROI: lower value is better (minimum is 0)

Method/Network	SSIM 5	SSIM 10	Registration Time (s)	SD ROI liver dome
Before registration	0.935 $\pm$ 0.104	0.911 $\pm$ 0.032	-	2485 $\pm$ 1112
Baseline SimpleElastix pairwise	0.927 $\pm$ 0.036	0.909 $\pm$ 0.039	804.07 $\pm$ 153	1628 $\pm$ 590
Baseline SimpleElastix groupwise 5	0.894 $\pm$ 0.094	0.873 $\pm$ 0.093	892.42 $\pm$ 85.28	1723 $\pm$ 703
Baseline SimpleElastix groupwise 10	0.841 $\pm$ 0.147	0.824 $\pm$ 0.143	977.53 $\pm$ 76.98	1783 $\pm$ 684
Groupwise 5	0.953 $\pm$ 0.025	0.935 $\pm$ 0.031	8.55 $\pm$ 0.51	1516 $\pm$ 665
Groupwise 10	0.951 $\pm$ 0.021	0.936 $\pm$ 0.028	9.07 $\pm$ 0.56	1018 $\pm$ 1660
Benchmark pairwise	0.937 $\pm$ 0.021	0.924 $\pm$ 0.027	9.81 $\pm$ 0.60	2037 $\pm$ 801
Benchmark GroupRegNet 5	0.951 $\pm$ 0.023	0.942 $\pm$ 0.036	2.26 $\pm$ 0.17	1728 $\pm$ 775
Benchmark GroupRegNet 10	0.934 $\pm$ 0.022	0.930 $\pm$ 0.027	3.14 $\pm$ 0.06	2377 $\pm$ 1141

Table 6.1 column "SD ROI liver dome" shows the standard deviation of the corrected liver dome ROI magnitudes over the time steps of the whole CDH dataset before registration and after registration with the different registration methods. All registration methods except the benchmark pairwise network were able to significantly reduce the standard deviation of the corrected liver dome ROI magnitude ($p < 0.00001$). The Groupwise 10 network yielded the lowest standard deviation of the corrected liver dome ROI magnitude.

Figure 6.3 shows examples of the corrected liver dome ROI magnitude of one patient before and after registration plotted over the timesteps. To ensure clarity, we do not show the ROI magnitude of all methods in this graph, instead we display the ROI magnitude before registration and after registration with our Groupwise 5 and Groupwise 10 networks. The curve before registration contains large breathing artefacts due to non-uniform deformation of the target structures in the ROI across the time steps. The ROI magnitude curve is smoothed by registration, i.e. by spatial alignment of the corresponding structures in the images of the different time steps. The best results, i.e. curve with the lowest standard deviation, was achieved with the Groupwise 5 network.

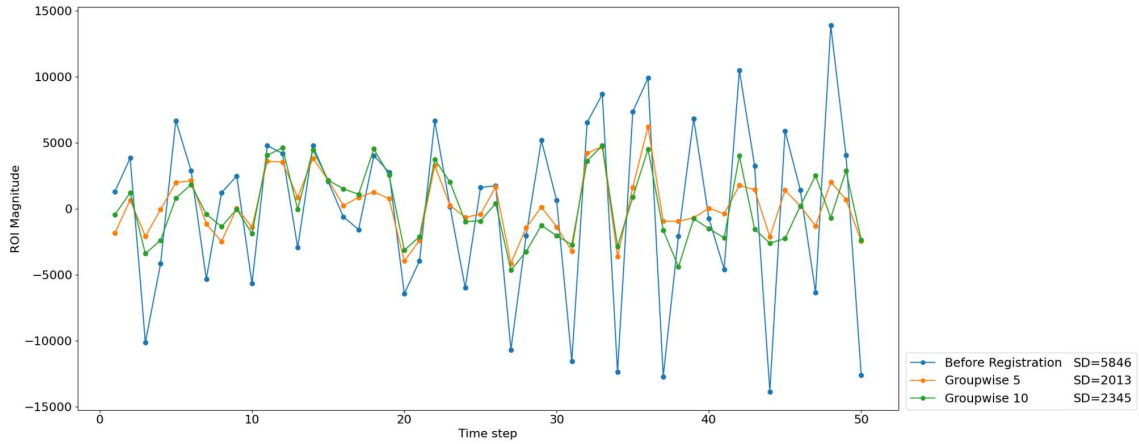


Figure 6.3: An example of a corrected liver dome ROI magnitude (arbitrary units) of one patient’s four-dimensional DCE-MR scan over the time steps before and after registration with the Groupwise 5 and Groupwise 10 networks. The standard deviation (SD) of each signal for this one patient is displayed.

Figure 6.4 shows examples of registration results. Here, we chose the same DCE-MR scan for which we also plotted the ROI magnitude (Figure 6.3). In Figure 6.4, before registration, especially the liver dome and smaller lung structures are not spatially aligned well. Registration with all registration methods except GroupReg-Net10 yielded a high overlap of the corresponding structures. The deformation fields of the Baselines SimpleElastix pairwise and groupwise are smooth, the largest deformations are located in the area of the right abdomen. The deformation fields of the neural networks indicate that the largest deformation of corresponding structures between the different time steps occurred at the liver dome. Additionally, the spatial alignment was corrected for finer lung structures.

The perfusion maps (Table 6.2) were calculated for the DCE-MR scan for which we also plotted the ROI magnitude and evaluated the registration results visually in Figure 6.4. Our proposed groupwise networks (Groupwise 5 and Groupwise 10) significantly reduced the standard deviation of the ROIs on the maps of the mean transit time, pulmonary blood volume and pulmonary blood flow. All registration networks significantly reduced the standard deviation of the ROI on the maps of the pulmonary blood volume and pulmonary blood flow. For the registration results of the baseline SimpleElastix groupwise 5, the pulmonary blood volume and pulmonary blood flow were incorrectly overestimated.

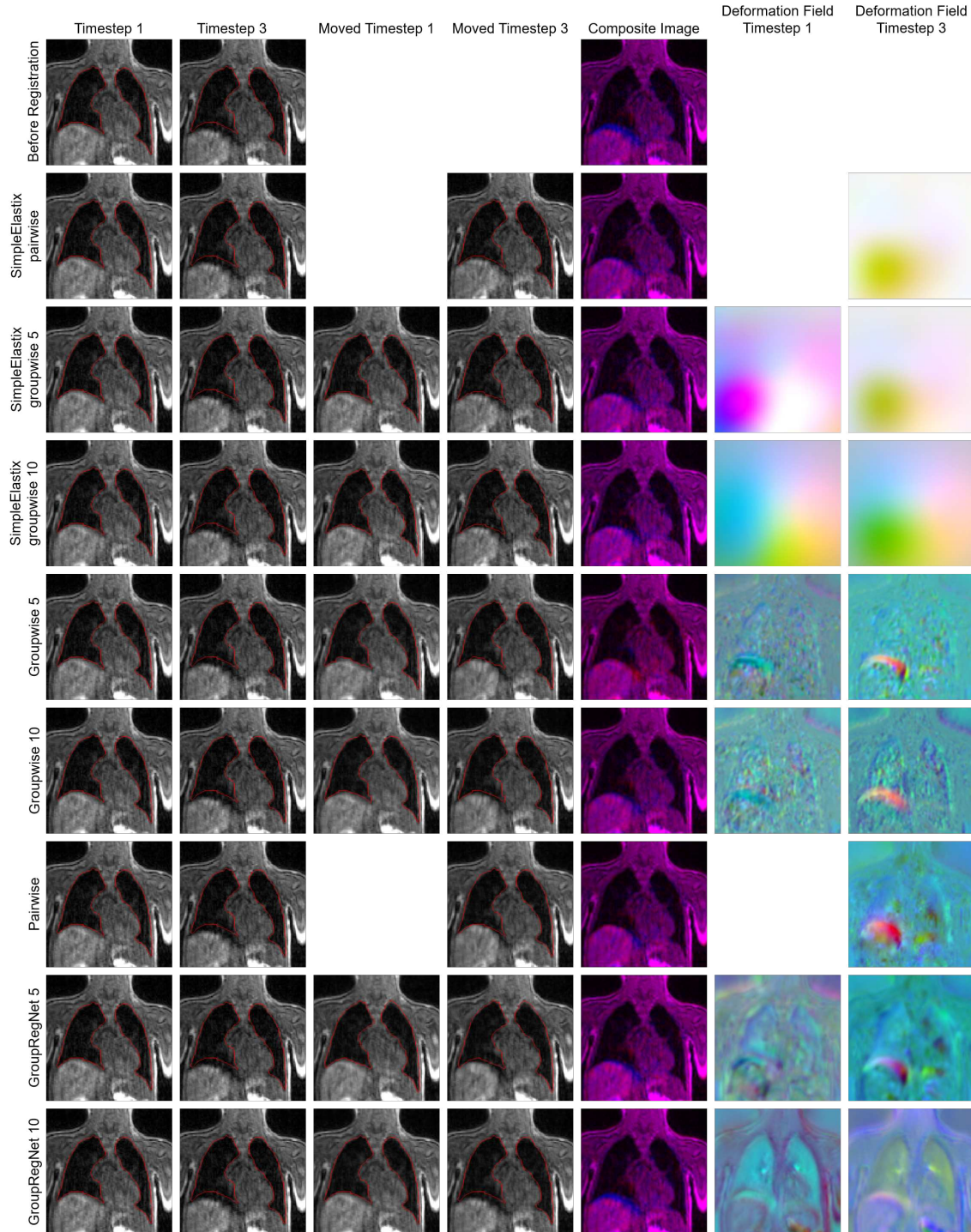


Figure 6.4: Example results of the registration of two volumes/time steps from the CDH dataset. The cropped images show central slices of the coronal plane. The MR images are overlaid by the lung segmentations (column 1 - 4). The resulting composites show the data of the first time step in blue and the data of the third time step in red. The deformation field is visualized in RGB channels.

Table 6.2: Perfusion maps: mean transit time (MTT), pulmonary blood volume (PBV) and pulmonary blood flow (PBF) (mean $\pm$ SD) calculated on the images before registration and the results of the baselines SimpleElastix, groupwise networks and the benchmark networks. With the bold numbers indicating significant improvement (p-value < 0.05) to the unregistered data (first row).

Method/Network	MTT (s)	PBV (ml/100ml)	PBF (ml/100ml/min)
Before registration	5.72 $\pm$ 1.96	6.36 $\pm$ 2.74	70.16 $\pm$ 26.28
Baseline SimpleElastix pairwise	5.67 $\pm$ 2.08	5.59 $\pm$ 2.24	64.41 $\pm$ 23.67
Baseline SimpleElastix groupwise 5	6.16 $\pm$ 2.19	9.46 $\pm$ 3.31	104.49 $\pm$ 32.19
Baseline SimpleElastix groupwise 10	5.22 $\pm$ 1.75	7.76 $\pm$ 3.26	88.97 $\pm$ 28.04
Groupwise 5	5.71 $\pm$ 1.73	6.35 $\pm$ 2.43	70.76 $\pm$ 21.78
Groupwise 10	6.40 $\pm$ 1.62	6.76 $\pm$ 2.38	67.99 $\pm$ 20.07
Benchmark pairwise	5.91 $\pm$ 1.73	6.60 $\pm$ 2.17	72.22 $\pm$ 22.20
Benchmark GroupRegNet 5	5.94 $\pm$ 1.80	6.04 $\pm$ 2.24	65.65 $\pm$ 22.10
Benchmark GroupRegNet 10	5.91 $\pm$ 1.82	6.12 $\pm$ 2.30	67.79 $\pm$ 23.77

6.4 Discussion

We proposed groupwise multiresolution deep learning networks for the deformable registration of DCE-MR time series for the assessment of lung perfusion in patients with CDH. By comparing the groupwise networks with a pairwise registration network as benchmark and a pairwise registration with SimpleElastix as baseline, we investigated the impact of using groupwise vs pairwise registration on the registration result. Additionally, we compared the proposed groupwise registration network to an iterative groupwise registration method SimpleElastix and a published groupwise registration network to show the superiority of the proposed groupwise registration network. For our proposed groupwise network, we chose a network architecture that uses multiple resolutions, thus combining the advantages of groupwise and multiresolution registration.

Recently published DCE-MRI image registration methods address the challenge of multimodal image registration. For instance, Cai et al. [150] apply an image-to-image translation network to convert the multimodal image registration into a monomodal image registration task. Sun et al. [151] also focus on reducing the intensity changes induced by contrast agent administration and, additionally, applies a coarse-to-fine image registration approach in which pairwise registration is used to compute a coarse transform and groupwise registration is used to refine the transformation. In our study, we focused on using groupwise multiresolution registration. In order to address the challenge of multimodal image registration, we used the entropy-based loss function Mutual Information to train the networks as well as a basic network architecture that has already demonstrated high performance on four multimodal medical datasets [11].

In recent years, several groupwise medical image registration methods have been published, these include iterative/classical methods [17, 153–158] and deep learning-based approaches [159–167]. Only one published neural network for groupwise registration also applies a multiresolution approach [166], and one network uses a coarse and a fine network to first reduce larger image displacements and then correct finer

deformations [162]. As a benchmark network, we used the published groupwise registration network GroupRegNet [164], as it was developed and evaluated on two publicly available four-dimensional CT thorax datasets and outperformed several published medical image registration methods.

Assessing the networks' performance is challenging due to the lack of ground truth data [7]. To address this problem, we applied different methods to evaluate the registration results. We used the SSIM to compare the structural similarity between the images, focusing on luminance, contrast and structure. We analyzed the deformation field by calculating the Jacobian determinant at each point to detect any image folding. Furthermore, we calculated the standard deviation of the liver dome ROI magnitude to track changes in the deformation of the liver dome over the time steps. Additionally, a visual inspection of the deformation field was performed to identify areas with major deformation and to check whether the deformation was smooth within the individual organs. Moreover, we analyzed the homogeneity of the perfusion maps mean transit time, pulmonary blood volume and pulmonary blood flow.

Pairwise and groupwise registration with SimpleElastix resulted in a significant reduction of the SSIM. In some three-dimensional volumes, larger circular areas with uniform intensity (standard pixel value of the registration algorithm) appear in the first and last two slices due to the large grid spacing of the registration methods. Therefore, in medical image analysis, the perfusion cannot be evaluated in these slices. For SimpleElastix groupwise 5, the ROI placement for the UMMPerfusion algorithm was challenging, therefore affecting the perfusion calculation.

We evaluated the groupwise registration network using two group sizes. In our proposed network, using groups of five volumes performed better than groups of ten volumes in SSIM 5, but worse in SSIM 10. The standard deviation of the liver dome ROI magnitude and the standard deviation of the ROIs of the perfusion maps are significantly lower after registration with the Groupwise 10 network than after registration with the Groupwise 5 network. The visual evaluation demonstrated that the spatial alignment was considerably improved for both group sizes compared to before registration.

The proposed groupwise networks achieved significantly better results than the pairwise network, based on SSIM and standard deviation of liver dome ROI magnitude, and resulted in a low degree of image folding (number of $|J| \leq 0$). For the standard deviation of the ROIs of the perfusion maps the results were mixed. For the pairwise network, a comparatively high lambda parameter had to be set in order to achieve medically plausible registration results for the scans of all time steps. With less regularization of the displacement field, good registration results could still be achieved for the first 5-10 timesteps. In subsequent time steps, corresponding structures in the fixed and moving image differed in intensity due to the contrast agent administration, which made registration difficult and led to medically implausible deformation of the liver and kidneys. Here we had to compromise between a high deformation correction, that is, improved spatial alignment through registration (low lambda-parameter) and a smooth deformation field that leads to medically plausible transformations (high lambda-parameter).

The proposed Groupwise network outperformed the benchmark GroupRegNet network (Groupwise 5 vs GroupRegNet 5 and Groupwise 10 vs GroupRegNet 10, except for SSIM 10 and the standard deviation of the ROIs of the pulmonary blood volume map). The reason for this could be that the monoresolution approach of the GroupRegNet is not able to correct the varying degrees of deformation, especially in the liver and lungs. Whereas the proposed groupwise network utilises multiple resolutions to correct initially coarser and then increasingly finer deformations.

Different from the deformation fields generated by neural networks, the deformation fields generated by the baselines SimpleElastix pairwise and groupwise did not show any recognizable structures such as organ contours. This is due to the high value of the final grid spacing parameter ("FinalGridSpacingInPhysicalUnits") of $50 \times 50 \times 50 \text{ mm}^3$, which ensures smooth and medically plausible transformations. Using a smaller grid spacing led to unrealistic deformations. For the neural networks, the grid spacing was set to $2 \times 2 \times 2 \text{ mm}^3$, with the degree of deformation being controlled by the lambda parameter. For groupwise registration with SimpleElastix, no regularisation loss can be added as of now, so the smoothness of the deformation field cannot be controlled via a lambda parameter.

The registration process with the baseline took significantly longer than with the neural networks. This difference results from SimpleElastix using an iterative algorithm in contrast to deep learning-based methods. While GPU implementations [80] or techniques that involve local similarity estimations [147], can drastically reduce the time required for registration, SimpleElastix was chosen for this study due to its wide acceptance and use in a wide variety of medical image registration applications. The registration with the proposed groupwise networks took significantly longer than with the benchmark GroupRegNet, as in the proposed groupwise network several networks were cascaded for a multiresolution approach, whereas in the benchmark GroupRegNet a single U-Net [67] with only one resolution was used.

In our study, the network training was limited by the available GPU memory (NVIDIA A100 GPU with 80 GB), limiting the group size for the groupwise registration networks to 10. Nevertheless, we were able to demonstrate the effectiveness of groupwise registration networks for DCE-MR image registration.

6.5 Conclusion

We developed neural networks for groupwise deformable registration of medical images, especially for DCE-MRI image registration. Our experiments demonstrate that registration with the groupwise networks results in high spatial alignment, medically plausible transformations, and fast registration of less than 10 seconds for a DCE-MR scan. Additionally, our results show that image registration with our proposed groupwise network results in more homogeneous perfusion maps, thereby improving the accuracy of medical image analysis.

6.6 Supplementary Information

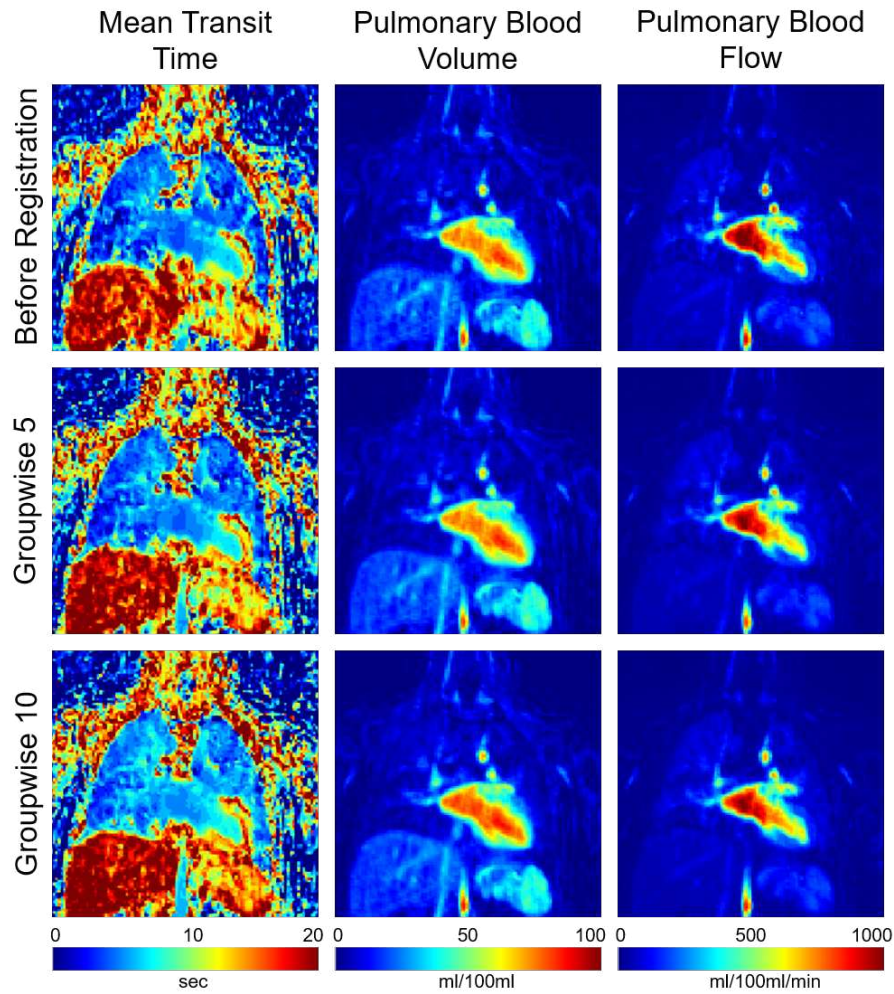


Figure 6.5: Example of perfusion maps generated by UMMPerfusion of one representative patient. From left to right mean transit time (sec), pulmonary blood volume (ml/100ml), and pulmonary blood flow (ml/100ml/min) are shown. Top row represents before registration, second and third row represents results from our proposed groupwise deep learning network.

7. Summary of the Publications

In medical image processing, the use of multimodal image registration is crucial as it allows the combination of complementary data from multiple imaging modalities. This integration improves clinical applications such as more precise diagnosis and better treatment planning and delivery and outcomes for a variety of diseases, such as kidney or oligometastatic disease, by combining the strengths and information of each modality. Advances in medical image registration techniques contribute to improving image fusion, thereby enhancing patient care. Thus, the aim of this thesis is to further develop and innovate deep learning-based methods for the registration of three-dimensional multimodal medical images.

In our studies, we addressed various aspects in the field of medical image registration and thus cover a wide range of tasks and applications of medical image registration. In all studies we performed multimodal image registration as fusing multimodal images is an important aspect in the field of medical image analysis. In our studies, we registered several different target regions: abdominal organs (liver, kidney and prostate) and the lungs. Additionally, we performed the tasks of co-registration (studies 1-3) and motion correction between multiple scans of a single four-dimensional image (study 4).

In [Chapter 3](#), 20 different published neural networks for affine medical image registration were implemented and their generalizability to new datasets was assessed using two multimodal liver datasets and an in-house developed CNN as a benchmark. In [Chapter 4](#), a novel multistage network was presented, which combines rigid, affine and deformable transformations, addressing large rigid deformations typical in medical imaging. In [Chapter 5](#) potential benefits and limitations of using multiple steps and resolutions were investigated in a deformable neural network on the registration result of four multimodal datasets. And in [Chapter 6](#) we developed a groupwise multiresolution network for the registration of four-dimensional DCE MR images.

The following paragraphs comprehensively summarise the individual studies.

Deep learning-based affine medical image registration for multimodal minimal-invasive image-guided interventions – A comparative study on generalizability
Z Med Phys, doi: 10.1016/j.zemedi.2023.05.003

In [Chapter 3](#) we assessed the generalizability of published deep learning-based methods for affine medical image registration. For this purpose, we conducted an extensive literature search on IEEE Xplore, PubMed, Web of Science and ScienceDirect to collect published neural networks for affine registration of medical images. Our search yielded 20 published networks, which we subsequently implemented. In some cases, reproducing the published networks was challenging due to insufficient details disclosed about their implementation. So we contacted corresponding authors to verify our implementations. When no response was received, we made assumptions.

We developed a comprehensive framework for systematic training and evaluation of the networks, the code is available online. The 20 implemented networks were

trained and evaluated using a synthetic and an in-vivo patient dataset of CT and MR liver scans. The performance of the networks was compared with each other and with an iterative affine registration using SimpleElastix as a baseline. Additionally, we developed our own neural network (CNN) specifically for the synthetic dataset to serve as a benchmark. The networks were trained semi-supervised. In addition to the fixed and moving image, the segmentations of the fixed and moving images were used during training. We applied the Dice loss, which measures the overlap between the segmentation of the fixed and moved image. For this purpose, liver segmentations were created for the patient dataset using U-Nets, while segmentations for the synthetic dataset are available from the XCAT phantom. In the first experiment the networks were trained on the synthetic dataset and evaluated on both the synthetic and unseen patient datasets. Subsequently, the networks were finetuned and evaluated on the patient dataset.

The results showed that only six and nine networks significantly improved the spatial alignment (based on Dice coefficient) of the synthetic dataset and the patient dataset, respectively, and therefore generalize well to the datasets applied in our study. This demonstrates that only a few of the published neural networks for affine medical image registration are capable of generalizing to new data. Consequently, further research is needed to develop methods for medical image registration suitable for a wider range of data and applications.

A Multistage Rigid-Affine-Deformable Network for Three-Dimensional Multimodal Medical Image Registration

Appl. Sci., doi: 10.3390/app132413298

Especially in the field of medical image registration, large rigid deformations often occur, e.g. as a result of differing patient positioning across scanners or acquisitions, or caused by rigid anatomical structures like bones. In Chapter 4, we address this problem by presenting a novel multistage deep learning-based approach for the registration of three-dimensional multimodal medical images. In our novel network, we integrate rigid, affine and deformable transformations using three cascaded subnetworks. In the first stage (rigid subnetwork), the larger rigid deformations are corrected. The subsequent affine subnetwork then reduces smaller global deformations and the following deformable subnetwork corrects smaller local deformations.

We evaluated our multistage network against a rigid-affine-deformable registration approach using the iterative registration method NiftyReg as a baseline. In addition, we used a multistage network that combines affine and deformable transformations as a benchmark to assess the impact of a preceding rigid subnetwork on the registration result. The performance of the proposed multistage network was evaluated on four three-dimensional in-vivo multimodal medical datasets: three kidney MR datasets, comprising T1- and T2-weighted MR scans, and one liver dataset, containing CT and T1-weighted MR scans. We trained the networks unsupervised using Mutual Information and Gradient L2 loss.

The experimental results showed that our multistage approach, integrating rigid, affine, and deformable transformations, led to registration results with high spatial alignment (Dice: 76.7 ± 12.5 , 61.1 ± 14.0 , 64.8 ± 16.2 , 68.1 ± 24.6 for our datasets) and medically plausible transformations ($|J| \leq 0: \leq 1.1\%$).

Multistep Networks for Deformable Multimodal Medical Image Registration *IEEE Access, doi: 10.1109/ACCESS.2024.3412216*

In [Chapter 5](#), we investigated the potential advantages and trade-offs of employing multiple resolutions for multimodal medical image registration. For this purpose, we developed neural networks for deformable registration of three-dimensional multimodal medical images utilizing multiple steps and differing resolutions. This approach is based on the coarse-to-fine strategy of iterative registration methods such as SimpleElastix and NiftyReg or published multistep neural networks.

Our neural networks function through a series of cascading subnetworks, each applying a different resolution. The first subnetwork processes the fixed and moving image at a low resolution, correcting only the coarse deformations. Subsequent subnetworks process the images registered by the previous subnetworks, using progressively higher resolutions to address finer deformations. The final subnetwork utilizes the images at full resolution.

We evaluated the multistep-multiresolution networks with up to four resolutions/steps with each other and with a monostep-monoresolution network as benchmark. As baselines, we utilized the iterative registration methods SimpleElastix and NiftyReg. Additionally, the performance of the multistep networks was compared with the performance of two widely used neural networks for medical image registration, the monostep-monoresolution network VoxelMorph-2 and the monostep-multiresolution network LapIRN. In the study, we utilized four multimodal medical datasets: a synthetic and a patient dataset comprising [CT](#) and T1-weighted liver [MR](#) scans, an in-vivo renal [MR](#) dataset containing T1-weighted and T2-weighted [MR](#) scans and a publicly available in-vivo prostate [MR](#) dataset comprising T2-weighted and [DWI MR](#) scans. The networks were jointly trained using unsupervised learning with Mutual Information and Gradient L2 loss. Experimental results showed that deep learning-based registration approaches applying multiple steps and resolutions yield registration results with high spatial alignment (NMI up to 0.33 ± 0.02 , Dice up to 90.8 ± 3.1) and medically plausible registration results ($|J| \leq 0$: $< 0.5\%$), while keeping the registration time low (< 0.5 s). Furthermore, our results show that increasing the number of resolutions/steps (with a maximum of four resolutions/steps in this study) improved the structural similarity and that multistep networks show better performance than monostep-monoresolution networks.

A Groupwise Multiresolution Network for DCE-MRI Image Registration *submitted to Scientific Reports, 12.09.2024*

In [Chapter 6](#), we addressed the groupwise registration of four-dimensional MR scans, i.e. motion correction of individual three-dimensional volumes within a single sequence. In this study, [DCE-MR](#) images were registered to evaluate lung perfusion in patients with [CDH](#). For this purpose, we developed a groupwise multiresolution network for deformable four-dimensional medical image registration.

As underlying network architecture, the multistep-multiresolution network presented in [Chapter 5](#) was used. The groupwise network was implemented with two different group sizes (5 and 10). The networks were trained unsupervised with Mutual Information and Gradient L2 loss. We compared our groupwise networks with the

pairwise registration network of [Chapter 5](#) and a published groupwise registration network as benchmarks and the classical registration methods SimpleElastix pairwise and groupwise as baselines using four-dimensional DCE-MR scans of patients after [CDH](#) repair. In our experiments, we applied several methods to evaluate the registration results. In addition to calculating [SSIM](#), the Jacobian determinant and the registration time, we also utilised [ROIs](#). To analyse the movement of the liver dome over the time steps, we placed a rectangular [ROI](#) on the liver dome, and we placed cylindrical [ROIs](#) on the calculated perfusion maps mean transit time, pulmonary blood volume and pulmonary blood flow to investigate whether the homogeneity of the perfusion maps is increased by registration of the four-dimensional [DCE-MRI](#) scans.

Experimental results showed that our proposed groupwise network performed better than iterative and deep learning-based pairwise and groupwise registration methods on the [DCE-MRI](#) dataset. Our results demonstrate that registration with our groupwise networks results in high spatial alignment, medically plausible registration results ($|J| \leq 0$: 0.0 ± 0.0 %) and fast registration of less than 10 seconds for a four-dimensional DCE-MR scan. Moreover, our results show that registration with our groupwise networks led to increased homogeneity of the perfusion maps, enhancing the accuracy of medical image analysis.

8. Discussion and Conclusion

The networks developed in [Chapter 3](#) to [Chapter 6](#) outperformed existing registration methods and thus improve the accuracy of image registration. Additionally, the conducted studies provide in-depth insights into various aspects of deep learning-based medical image registration.

[Chapter 3](#) provided valuable insights into the generalisability of published neural networks for affine medical image registration and showed that only a few networks are generalisable, i.e. performed well on new datasets. The generalisability to new datasets should therefore already be considered during the development of the neural networks in order to ensure that the networks can also be used for other applications. The best results of published neural networks were achieved by a network that had already been developed using four publicly available datasets. Thus, we suggest using multiple different datasets with varying modalities or target structures for training and evaluation when developing deep learning-based approaches, rather than just a single dataset. We applied this concept for the development of our neural networks in [Chapter 4](#) and [Chapter 5](#). Here we developed, trained and evaluated the networks using four multimodal medical datasets with varying modalities and target structures in order to ensure the generalisability of the networks for a wide range of applications. Moreover, the study in [Chapter 3](#) also highlighted difficulties in reproducing the published neural networks. The code for all frameworks and image registration networks developed in this thesis is available online. This enhances the reproducibility of our research and enables others to apply and test our methods for their own data and applications.

The novel neural network presented in [Chapter 4](#), incorporating rigid, affine and deformable transformations, addresses the challenge of correcting large rigid deformations that commonly occur in medical images due to variations in patient positioning in different scanners and rigid anatomical structures. The novel multistage network outperforms a network that does not include rigid registration, thus providing a valuable new image registration method to improve the accuracy of image registration and image fusion.

In recent years, several multistep medical image registration networks have been published, but to the best of our knowledge, the influence of multiple steps and resolutions on the registration result has not been studied. Our study in [Chapter 5](#) therefore provides valuable insight into the benefits and limitations of using multiple steps and resolutions. The multistep networks demonstrated good performance on four different multimodal medical imaging datasets and are therefore applicable to a wide variety of data and applications.

Finally, the network in [Chapter 6](#) combines the advantages of groupwise and multiresolution registration methods and, with its high accuracy and short registration time, is a valuable tool for the four-dimensional registration and motion correction. Utilising this network leads to enhanced accuracy of medical image analysis, e.g. when calculating perfusion parameters.

During the development of our medical image registration networks and while conducting the studies in this thesis, several challenges occurred. In [Chapter 3](#) to [Chap-](#)

ter 6, network training was limited by the size of the GPU memory. In Chapter 3, neural networks with a large number of parameters and GPU memory-intensive transformer networks were used. In Chapter 4 and Chapter 5, neural networks were developed that comprise several subnetworks. Due to their large network size (1,211,141 parameters for multistage rigid-affine-deformable network and 977,292 parameters for multistep network variant A) and having two three-dimensional input volumes, they require a lot of GPU memory for training. Therefore, the input images were downsampled in Chapter 3 to Chapter 5. However, downsampling potentially reduces the registration accuracy, as small deformations cannot be displayed in the downsampled image and therefore cannot be corrected. In the experiments in Chapter 6, network training was also limited by the size of the GPU memory, as the multistep network variant A from Chapter 5 was used and groupwise registration was performed with multiple three-dimensional input volumes. As a result, the proposed groupwise multiresolution network could only be implemented and trained with a maximum group size of 10. In the future, with the development of GPUs with larger GPU memory, full resolution images can be used for training (Chapter 3 to Chapter 5) and the group size of the proposed groupwise multiresolution network can be increased (Chapter 6).

As future work, different approaches could be implemented to reduce the GPU memory required for training. For instance, the size of the multistage rigid-affine-deformable network could be decreased by combining the three networks into a single U-Net-like network with three outputs for the three transformations. Alternatively, the input images of the networks of all four studies could be processed patch-based. The networks therefore output transformations for patches of the images. The final transformation of the full image volume is then composed of the individual transformation patches. Alternatively, several slices (e.g. eight slices) could be processed together in a sliding window approach, and subsequently the window is moved further. The final transformation is calculated from the individual transformations. The patch-based and sliding window approach, as well as the medically plausible merging of the individual transformations into the final transformation are objective of future work.

Neural network training for medical image registration is impeded by the lack of ground truth data. The use of manual registration results or gold standard methods as labels for supervised learning is not an ideal solution, as the neural networks can only achieve at most the accuracy of the label registration method. Some published neural networks apply supervised training with synthetic deformations. It is however questionable to what extent the synthetic deformations reflect medical reality. In the future, realistic transformations such as medically plausible respiration could be investigated and, e.g., modeled using deep learning. These synthetic deformations could then be used to train neural networks for supervised medical image registration.

Furthermore, evaluating the registration results is challenging, when no ground truth is available. In the studies of this thesis, various evaluation methods were applied in combination. Image-based metrics such as NMI or the SSIM and segmentation-based metrics such as Dice coefficient and Hausdorff distance were used. In addition, the deformation field was investigated by calculating the Jacobian determinant to

determine the degree of image folding. Furthermore, the registration results, i.e. moved image, segmentation if available and deformation field were evaluated optically. In Chapter 6, a rectangular ROI was placed on the liver dome to analyse the deformation of the liver dome over the individual time steps by examining the standard deviation of the ROI magnitude before and after registration. Additionally, perfusion maps were calculated and cylindrical ROIs were utilised to investigate whether registering the DCE MR scans subsequently results in more homogeneous perfusion maps. The objective of future work could be to apply additional methods for the evaluation of registration results. For example, calculate the Target Registration Error (*TRE*), if corresponding landmarks are available in the fixed and moving image, or apply metrics that are calculated locally, such as local Mutual Information, and thus include spatial information.

Training the networks is computationally intensive and took several hours in our studies. On the other hand, the inference process with our developed neural networks was very fast, so the registration of a new three-dimensional image pair took less than half a second (Chapter 3 to Chapter 5) and the registration and motion correction of a DCE-MR scan took less than 10 seconds (Chapter 6). This makes it possible to use the registration methods for time-critical applications, such as image-guided interventions. This requires the integration of the registration methods into the clinical infrastructure. In future work, the integration of the developed networks into radiology systems and medical image processing software used by radiologists, such as Medical Imaging Interaction Toolkit (MITK) [112], is envisioned.

In conclusion, the deep learning-based medical image registration methods developed in this thesis improve the accuracy of medical image registration for a wide variety of different tasks and applications. The developed methods serve as valuable tool for clinical applications of image fusion, e.g. for performing minimally invasive image-guided interventions, e.g. for oligometastatic disease (study 1-3), for diagnosing and monitoring diseases such as ADPKD (study 2 and 3) and for monitoring the progress of a therapy, e.g. CDH repair (study 4). By enabling rapid registration, these techniques support real-time decision making and increase the accuracy of interventions as well as medical image analysis, ultimately leading to better patient outcomes.

9. Bibliography

- [1] B. I. Waldkirch, *Methods for three-dimensional Registration of Multimodal Abdominal Image Data*. Doctoral dissertation, Ruprecht Karl University of Heidelberg, 2020.
- [2] X. Chen, A. Dians-Pinto, N. Ravikumar, and A. F. Frangi, “Deep learning in medical image registration,” *Prog. Biomed. Eng.*, vol. 3, p. 012003, 2021.
- [3] Y. Fu, Y. Lei, T. Wang, W. J. Curran, T. Liu, and X. Yang, “Deep learning in medical image registration: a review,” *Phys. Med. Biol.*, vol. 65, p. 20TR01, 2020.
- [4] H. Xiao, X. Teng, C. Liu, T. Li, G. Ren, R. Yang, D. Shen, and J. Cai, “A review of deep learning-based three-dimensional medical image registration methods,” *Quantitative Imaging in Medicine and Surgery*, vol. 11, no. 12, 2021.
- [5] A. Anderlik, A. Z. Munthe-Kaas, O. K. Oye, E. Eikefjord, J. Rorvik, D. M. Ulvang, F. G. Zöllner, and A. Lundervold, “Quantitative assessment of kidney function using dynamic contrast enhanced mri - steps towards an integrated software prototype,” in *2009 Proceedings of 6th International Symposium on Image and Signal Processing and Analysis*, pp. 575–581, 2009.
- [6] E. Hodneland, E. Keilegavlen, E. A. Hanson, E. Andersen, J. A. Monssen, J. Rørvik, S. Leh, H.-P. Marti, A. Lundervold, E. Svarstad, and J. M. Nordbotten, “In vivo detection of chronic kidney disease using tissue deformation fields from dynamic mr imaging,” *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 6, pp. 1779–1790, 2019.
- [7] F. G. Zöllner, A. Šerifović-Trbalić, G. Kabelitz, M. Kociński, A. Materka, and P. Rogelj, “Image registration in dynamic renal MRI—current status and prospects,” *Magnetic Resonance Materials in Physics, Biology and Medicine*, vol. 33, pp. 33–48, Oct. 2019.
- [8] A. Caroli and T. L. Kline, “Abdominal imaging in adpkd: Beyond total kidney volume,” *Journal of Clinical Medicine*, vol. 12, no. 15, 2023.
- [9] S. Riyahi, H. Dev, J. D. Blumenfeld, H. Rennert, X. Yin, H. Attari, I. Barash, I. Chicos, W. Bobb, S. Donahue, and M. R. Prince, “Hemorrhagic cysts and other mr biomarkers for predicting renal dysfunction progression in autosomal dominant polycystic kidney disease,” *Journal of Magnetic Resonance Imaging*, vol. 53, no. 2, pp. 564–576, 2021.
- [10] A. Strittmatter, A. Caroli, and F. G. Zöllner, “A multistage rigid-affine-deformable network for three-dimensional multimodal medical image registration,” *Applied Sciences*, vol. 13, p. 13298, Dec. 2023.
- [11] A. Strittmatter and F. G. Zöllner, “Multistep networks for deformable multimodal medical image registration,” *IEEE Access*, vol. 12, pp. 82676–82692, 2024.

- [12] H. Qiu, A. W. Katz, and M. T. Milano, “Oligometastases to the liver: predicting outcomes based upon radiation sensitivity,” *J Thorac Dis*, vol. (10):E1384-E1386, 10 2016.
- [13] T. Ruers, F. Van Coevorden, C. Punt, J. Pierie, I. Borel-Rinkes, J. Ledermann, G. Poston, W. Bechstein, M. Lentz, M. Mauer, G. Folprecht, E. Van Cutsem, M. Ducreux, and B. Nordlinger, “Local Treatment of Unresectable Colorectal Liver Metastases: Results of a Randomized Phase II Trial,” *J Natl Cancer Inst*, vol. 109(9):dix015, 9 2017.
- [14] D. F. Bauer, J. Rosenkranz, A.-K. Golla, C. Tönnies, I. Hermann, T. Russ, G. Kabelitz, A. J. Rothfuss, L. R. Schad, J. L. Stallkamp, and F. G. Zöllner, “Development of an abdominal phantom for the validation of an oligometastatic disease diagnosis workflow,” *Medical Physics*, vol. 49, no. 7, pp. 4445–4454, 2022.
- [15] A. Strittmatter, L. R. Schad, and F. G. Zöllner, “Deep learning-based affine medical image registration for multimodal minimal-invasive image-guided interventions – a comparative study on generalizability,” *Zeitschrift für Medizinische Physik*, vol. 34, no. 2, pp. 291–317, 2024. Special Issue: Artificial Intelligence in Medical Physics.
- [16] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, and A. V. Dalca, “VoxelMorph: A Learning Framework for Deformable Medical Image Registration,” *IEEE Transactions on Medical Imaging*, vol. 38, no. 8, pp. 1788–1800, 2019.
- [17] K. Marstal, F. Berendsen, M. Staring, and S. Klein, “SimpleElastix: A User-Friendly, Multi-lingual Library for Medical Image Registration,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 574–582, 2016.
- [18] M. Modat, G. R. Ridgway, Z. A. Taylor, M. Lehmann, J. Barnes, D. J. Hawkes, N. C. Fox, and S. Ourselin, “Fast free-form deformation using graphics processing units,” *Computer Methods and Programs in Biomedicine*, vol. 98, no. 3, pp. 278–284, 2010. HP-MICCAI 2008.
- [19] M. Modat, D. Cash, P. Daga, G. Winston, J. Duncan, and S. Ourselin, “Global image registration using a symmetric block-matching approach,” *Journal of medical imaging (Bellingham, Wash.)*, vol. 1, p. 024003, 07 2014.
- [20] C. Ionescu, *The Human Respiratory System: An Analysis of the Interplay between Anatomy, Structure, Breathing and Fractal Dynamics*. Series in Bio-Engineering, London, England: Springer, Aug. 2013.
- [21] S. Standring, ed., *Gray’s Anatomy: The Anatomical Basis of Clinical Practice*. Gray’s Anatomy Series, Philadelphia, United States: Elsevier Limited, 41 ed., Sept. 2016.
- [22] V. Kulish, ed., *Human Respiration: Anatomy and Physiology, Mathematical Modeling, Numerical Simulation and Applications*. Advances in bioengineering, Southampton, England: WIT Press, May 2006.

-
- [23] M. Levitzky, *Pulmonary Physiology, Ninth Edition*. McGraw-Hill Education, 9 ed., Jan. 2018.
 - [24] M. Cloutier, *Respiratory Physiology: Mosby Physiology Series*. Mosby's Physiology Monograph, Philadelphia, United States: Elsevier, 2 ed., Sept. 2018.
 - [25] P. K. Chandrasekharan, M. Rawat, R. Madappa, D. H. Rothstein, and S. Lakshminrusimha, "Congenital diaphragmatic hernia – a review," *Maternal Health, Neonatology and Perinatology*, vol. 3, Mar. 2017.
 - [26] A. Zani, W. K. Chung, J. Deprest, M. T. Harting, T. Jancelewicz, S. M. Kunisaki, N. Patel, L. Antounians, P. S. Puligandla, and R. Keijzer, "Congenital diaphragmatic hernia," *Nature Reviews Disease Primers*, vol. 8, June 2022.
 - [27] J. A. Tovar, "Congenital diaphragmatic hernia," *Orphanet Journal of Rare Diseases*, vol. 7, no. 1, p. 1, 2012.
 - [28] M. Weidner, F. G. Zöllner, C. Hagelstein, K. Zahn, T. Schaible, S. O. Schoenberg, L. R. Schad, and K. W. Neff, "High temporal versus high spatial resolution in mr quantitative pulmonary perfusion imaging of two-year old children after congenital diaphragmatic hernia repair," *European Radiology*, vol. 24, p. 2427–2434, July 2014.
 - [29] A. Tannapfel and G. Klöppel, eds., *Pathologie: Leber, Gallenwege und Pankreas*. Springer-Verlag, 3 ed., July 2020.
 - [30] "Leber (Hepar)." <https://www.kenhub.com/de/library/anatomie/leber-hepar>. Accessed: 2024-01-15.
 - [31] P. Majno, G. Mentha, C. Toso, P. Morel, H. O. Peitgen, and J. H. Fasel, "Anatomy of the liver: An outline with three levels of complexity – a further step towards tailored territorial liver resections," *Journal of Hepatology*, vol. 60, p. 654–662, Mar. 2014.
 - [32] I. M. Arias, H. J. Alter, J. L. Boyer, D. E. Cohen, D. A. Shafritz, S. S. Thorgeirsson, and A. W. Wolkoff, eds., *The Liver: Biology and Pathobiology*. Hoboken, NJ: Wiley-Blackwell, 6 ed., Feb. 2020.
 - [33] S. Sargent, ed., *Liver Diseases: An Essential Guide for Nurses and Health Care Professionals*. Chichester, England: Wiley-Blackwell, May 2009.
 - [34] T. Izmailov, S. Ryzhkin, G. Borshchev, and S. Boichuk, "Oligometastatic disease (omd): The classification and practical review of prospective trials," *Cancers*, vol. 15, p. 5234, Oct. 2023.
 - [35] M. P. Manns and S. Schneidewind, eds., *Praxis der Hepatologie*. Berlin, Germany: Springer, 1 ed., Jan. 2016.
 - [36] D. Hansel, C. Kane, G. Paner, and S. Chang, eds., *The Kidney: A Comprehensive Guide to Pathologic Diagnosis and Management*. New York, United States: Springer, 1 ed., Dec. 2015.

- [37] M. Klingele and D. Brodmann, eds., *Einführung in die Nephrologie und Nierenersatzverfahren*. Berlin, Germany: Springer, 1 ed., July 2017.
- [38] S. Gilbert, D. Weiner, D. Gipson, M. Perazella, M. Tonelli, and N. K. Foundation, eds., *National Kidney Foundation Primer on Kidney Diseases*. ClinicalKey 2012, Philadelphia, United States: Elsevier, 6 ed., 2014.
- [39] N. Nobakht, R. M. Hanna, M. Al-Baghdadi, K. M. Ameen, F. Arman, E. Nobakht, M. Kamgar, and A. Rastogi, “Advances in autosomal dominant polycystic kidney disease: A clinical review,” *Kidney Medicine*, vol. 2, no. 2, pp. 196–208, 2020.
- [40] R. Snell, *Clinical Anatomy by Regions*. Philadelphia, United States: Lippincott Williams & Wilkins, 9 ed., Nov. 2011.
- [41] M. Ittmann, “Anatomy and histology of the human and murine prostate,” *Cold Spring Harbor Perspectives in Medicine*, vol. 8, p. a030346, Oct. 2017.
- [42] J. A. Sinnott, J. R. Rider, J. Carlsson, T. Gerke, S. Tyekucheva, K. L. Penney, H. D. Sesso, M. Loda, K. Fall, M. J. Stampfer, L. A. Mucci, Y. Pawitan, S.-O. Andersson, and O. Andrén, “Molecular differences in transition zone and peripheral zone prostate tumors,” *Carcinogenesis*, vol. 36, pp. 632–638, 04 2015.
- [43] W. Kalender, *Computed Tomography: Fundamentals, System Technology, Image Quality, Applications*. Erlangen, Germany: Publicis Publishing, 3 ed., July 2011.
- [44] T. Buzug, *Computed Tomography: From Photon Statistics to Modern Cone-Beam CT*. Berlin, Germany: Springer, June 2008.
- [45] A. Maier, S. Steidl, V. Christlein, and J. Hornegger, eds., *Medical Imaging Systems: An Introductory Guide*. Lecture Notes in Computer Science, Basel, Switzerland: Springer International Publishing, 1 ed., Aug. 2018.
- [46] E. Seeram, *Computed Tomography: Physical Principles, Clinical Applications, and Quality Control*. Elsevier Health Sciences, 4 ed., Sept. 2015.
- [47] J. Haaga and D. Boll, *Computed Tomography & Magnetic Resonance Imaging Of The Whole Body E-Book*. Elsevier Health Sciences, 6 ed., June 2016.
- [48] A. Adlung, C. Licht, S. Reichert, S. Özdemir, S. A. Mohamed, M. Samartzi, M. Fatar, A. Gass, E. N. Prost, and L. R. Schad, “Quantification of tissue sodium concentration in the ischemic stroke: A comparison between external and internal references for ^{23}Na mri,” *Journal of Neuroscience Methods*, vol. 382, p. 109721, 2022.
- [49] W. S. McCulloch and W. Pitts, “A logical calculus of the ideas immanent in nervous activity,” *The Bulletin of Mathematical Biophysics*, vol. 5, p. 115–133, Dec. 1943.

-
- [50] C. Aggarwal, *Neural Networks and Deep Learning: A Textbook*. Basel, Switzerland: Springer International Publishing, 1 ed., Aug. 2018.
 - [51] D. Kriesel, “A brief introduction to neural networks.” <http://www.dkriesel.com>, 2007.
 - [52] B. Müller, J. Reinhardt, and M. Strickland, *Neural Networks: An Introduction*. Physics of Neural Networks, Berlin, Germany: Springer, 2 ed., Dec. 2012.
 - [53] F. Rosenblatt, “The perceptron: A probabilistic model for information storage and organization in the brain.,” *Psychological Review*, vol. 65, no. 6, p. 386–408, 1958.
 - [54] S. Albawi, T. A. Mohammed, and S. Al-Zawi, “Understanding of a convolutional neural network,” in *2017 International Conference on Engineering and Technology (ICET)*, pp. 1–6, 2017.
 - [55] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A simple way to prevent neural networks from overfitting,” *Journal of Machine Learning Research*, vol. 15, no. 56, pp. 1929–1958, 2014.
 - [56] L. Lu, Y. Shin, Y. Su, and G. Em Karniadakis, “Dying relu and initialization: Theory and numerical examples,” *Communications in Computational Physics*, vol. 28, no. 5, pp. 1671–1706, 2020.
 - [57] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” *CoRR*, vol. abs/1502.03167, 2015.
 - [58] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings* (Y. Bengio and Y. LeCun, eds.), 2015.
 - [59] J. Duchi, E. Hazan, and Y. Singer, “Adaptive subgradient methods for online learning and stochastic optimization,” *Journal of Machine Learning Research*, vol. 12, no. 61, pp. 2121–2159, 2011.
 - [60] H. Izadkhah, *Deep Learning in Bioinformatics: Techniques and Applications in Practice*. San Diego, United States: Elsevier, Jan. 2022.
 - [61] X. Glorot and Y. Bengio, “Understanding the difficulty of training deep feed-forward neural networks,” in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics* (Y. W. Teh and M. Titterton, eds.), vol. 9 of *Proceedings of Machine Learning Research*, (Chia Laguna Resort, Sardinia, Italy), pp. 249–256, PMLR, 13–15 May 2010.
 - [62] K. He, X. Zhang, S. Ren, and J. Sun, “Delving deep into rectifiers: Surpassing human-level performance on imagenet classification,” *CoRR*, vol. abs/1502.01852, 2015.

- [63] L. Li, K. Jamieson, G. DeSalvo, A. Rostamizadeh, and A. Talwalkar, “Hyperband: A novel bandit-based approach to hyperparameter optimization,” *Journal of Machine Learning Research*, vol. 18, no. 185, pp. 1–52, 2018.
- [64] K. Priddy and P. Keller, *Artificial Neural Networks: An Introduction*. Tutorial Texts in Optical Engineering, SPIE Publications, 2005.
- [65] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in Neural Information Processing Systems* (F. Pereira, C. Burges, L. Bottou, and K. Weinberger, eds.), vol. 25, Curran Associates, Inc., 2012.
- [66] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, June 2016.
- [67] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation,” in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, eds.), (Cham), pp. 234–241, Springer International Publishing, 2015.
- [68] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017.
- [69] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An image is worth 16x16 words: Transformers for image recognition at scale,” 2021.
- [70] X. Geng, G. E. Christensen, H. Gu, T. J. Ross, and Y. Yang, “Implicit reference-based group-wise image registration and its application to structural and functional mri,” *NeuroImage*, vol. 47, no. 4, pp. 1341–1351, 2009.
- [71] A. Strittmatter, A. Hertel, S. Diehl, M. F. Froelich, S. O. Schoenberg, S. Loges, T. Boch, D. Nowak, A. Streuer, L. R. Schad, and F. G. Zöllner, “A multistage registration of ct and biopsy ct images of lung tumors,” in *6th Conference on Image-Guided Interventions*, pp. 17–18, 2023.
- [72] M. Pilia, “Groupwise whole-body MR image registration guided by zero-average volume changes,” Master’s thesis, Uppsala University, Uppsala, Sweden, May 2018.
- [73] H. Xiao, G. Ren, and J. Cai, “A review on 3d deformable image registration and its application in dose warping,” *Radiation Medicine and Protection*, vol. 1, no. 4, pp. 171–178, 2020.

-
- [74] B. Eiben, J. Bertholet, M. Menten, S. Nill, U. Oelfke, and J. McClelland, “Consistent and invertible deformation vector fields for a breathing anthropomorphic phantom: a post-processing framework for the xcat phantom,” *Physics in Medicine & Biology*, vol. 65, Apr. 2020.
 - [75] A. Joshi and V. Bhaskar, “Multi-resolution approach for medical image registration using genetic algorithm,” in *2013 International Conference on Communication and Signal Processing*, pp. 612–616, 2013.
 - [76] S. Zhao, Y. Dong, E. Chang, and Y. Xu, “Recursive cascaded networks for unsupervised medical image registration,” in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, Oct. 2019.
 - [77] L. Wang, H. Shao, and X. Deng, “An unsupervised end-to-end recursive cascaded parallel network for image registration,” *Neural Processing Letters*, vol. 55, p. 8255–8268, June 2023.
 - [78] M. A. Oghabian and A. Todd-Pokropek, “Registration of brain images by a multi-resolution sequential method,” in *Information Processing in Medical Imaging* (A. C. F. Colchester and D. J. Hawkes, eds.), (Berlin, Heidelberg), pp. 165–174, Springer Berlin Heidelberg, 1991.
 - [79] M. Corvi and G. Nicchiotti, “Multiresolution image registration,” in *Proceedings., International Conference on Image Processing*, vol. 3, pp. 224–227 vol.3, 1995.
 - [80] O. Fluck, C. Vetter, W. Wein, A. Kamen, B. Preim, and R. Westermann, “A survey of medical image registration on graphics hardware,” *Computer Methods and Programs in Biomedicine*, vol. 104, no. 3, pp. e45–e57, 2011.
 - [81] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, “Image quality assessment: from error visibility to structural similarity,” *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.
 - [82] A. Bardera, M. Feixas, and I. Boada, “Normalized similarity measures for medical image registration,” in *Medical Imaging 2004: Image Processing* (J. M. Fitzpatrick and M. Sonka, eds.), vol. 5370, pp. 108 – 118, International Society for Optics and Photonics, SPIE, 2004.
 - [83] L. R. Dice, “Measures of the amount of ecologic association between species,” *Ecology*, vol. 26, p. 297–302, July 1945.
 - [84] T. Sørensen, *A Method of Establishing Groups of Equal Amplitude in Plant Sociology Based on Similarity of Species Content and Its Application to Analyses of the Vegetation on Danish Commons*. Biologiske skrifter, Copenhagen, Denmark: Munksgaard in Komm., 1948.
 - [85] J. Ashburner, “A fast diffeomorphic image registration algorithm,” *NeuroImage*, vol. 38, no. 1, pp. 95–113, 2007.

- [86] B. D. de Vos, F. F. Berendsen, M. A. Viergever, H. Sokooti, M. Staring, and I. Išgum, “A deep learning framework for unsupervised affine and deformable image registration,” *Medical Image Analysis*, vol. 52, pp. 128–143, 2019.
- [87] N. Chumchob and K. Chen, “A robust affine image registration method,” *International Journal of Numerical Analysis and Modeling*, vol. 6, pp. 311–334, 01 2009.
- [88] J. Chen, E. C. Frey, Y. He, W. P. Segars, Y. Li, and Y. Du, “TransMorph: Transformer for unsupervised medical image registration,” *Medical Image Analysis*, vol. 82, p. 102615, 2022.
- [89] T. C. W. Mok and A. C. S. Chung, “Affine Medical Image Registration with Coarse-to-Fine Vision Transformer,” *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20803–20812, 2022.
- [90] K. A. Hasenstab, G. M. Cunha, A. Higaki, S. Ichikawa, K. Wang, T. Delgado, R. L. Brunsing, A. Schlein, L. K. Bittencourt, A. Schwartzman, K. J. Fowler, A. Hsiao, and C. B. Sirlin, “Fully automated convolutional neural network-based affine algorithm improves liver registration and lesion co-localization on hepatobiliary phase T1-weighted MR images,” *Eur Radiol Exp.*, vol. 3(1):43, 2019.
- [91] S. Zhao, T. Lau, J. Luo, E. I.-C. Chang, and Y. Xu, “Unsupervised 3D End-to-End Medical Image Registration With Volume Tweening Network,” *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 5, pp. 1394–1404, 2020.
- [92] T. de Silva, E. Y. Chew, N. Hotaling, and C. A. Cukras, “Deep-Learning based Multi-Modal Retinal Image Registration for Longitudinal Analysis of Patients with Age-related Macular Degeneration,” *Biomedical Optics Express*, vol. 12, 10 2020.
- [93] Q. Zeng, Y. Fu, Z. Tian, Y. Lei, Y. Zhang, T. Wang, H. Mao, T. Liu, W. Curran, A. Jani, P. Patel, and X. Yang, “Label-driven MRI-US registration using weakly-supervised learning for MRI-guided prostate radiotherapy,” *Physics in Medicine & Biology*, vol. 65, 04 2020.
- [94] Y. Hu, M. Modat, E. Gibson, W. Li, N. Ghavami, E. Bonmati, E. Wang, S. Bandula, C. M. Moore, M. Emberton, S. Ourselin, J. A. Noble, D. C. Barratt, and T. Vercauteren, “Weakly-supervised convolutional neural networks for multimodal image registration,” *Medical Image Analysis*, vol. 49, pp. 1–13, 2018.
- [95] D. Gu, G. Liu, J. Tian, and Q. Zhan, “Two-Stage Unsupervised Learning Method for Affine and Deformable Medical Image Registration,” in *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 1332–1336, 2019.
- [96] Z. Shen, X. Han, Z. Xu, and M. Niethammer, “Networks for Joint Affine and Non-Parametric Image Registration,” in *2019 IEEE/CVF Conference on*

-
- Computer Vision and Pattern Recognition (CVPR)*, (Los Alamitos, CA, USA), pp. 4219–4228, IEEE Computer Society, jun 2019.
- [97] G. Luo, X. Chen, F. Shi, Y. Peng, D. Xiang, Q. Chen, X. Xu, W. Zhu, and Y. Fan, “Multimodal affine registration for ICGA and MCSL fundus images of high myopia,” *Biomedical Optics Express*, vol. 11, 07 2020.
 - [98] K. Tang, Z. Li, L. Tian, L. Wang, and Y. Zhu, “ADMIR–Affine and Deformable Medical Image Registration for Drug-Addicted Brain Images,” *IEEE Access*, vol. 8, pp. 70960–70968, 2020.
 - [99] X. Chen, Y. Meng, Y. Zhao, R. Williams, S. R. Vallabhaneni, and Y. Zheng, *Learning Unsupervised Parameter-Specific Affine Transformation for Medical Images Registration*, pp. 24–34. Springer Cham, 09 2021.
 - [100] X. Gao, J. Van Houtte, Z. Chen, and G. Zheng, “DeepASDM: a Deep Learning Framework for Affine and Deformable Image Registration Incorporating a Statistical Deformation Model,” in *2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, pp. 1–4, 2021.
 - [101] T. J. T. Roelofs, “Deep Learning-Based Affine and Deformable 3D Medical Image Registration,” Master’s thesis, Aalto University, Espoo, Finland, Sep 2021.
 - [102] Z. Zhu, Y. Cao, Q. Chenchen, Y. Rao, L. Di, Q. Dou, D. Ni, and Y. Wang, “Joint affine and deformable three-dimensional networks for brain MRI registration,” *Medical Physics*, vol. 48, 2021.
 - [103] E. Chee and Z. Wu, “AIRNet: Self-Supervised Affine Registration for 3D Medical Images using Neural Networks,” *CoRR*, vol. abs/1810.02583, 2018.
 - [104] W. Shao, L. Banh, C. A. Kunder, R. E. Fan, S. J. Soerensen, J. B. Wang, N. C. Teslovich, N. Madhuripan, A. Jawahar, P. Ghanouni, J. D. Brooks, G. A. Sonn, and M. Rusu, “ProsRegNet: A deep learning framework for registration of MRI and histopathology images of the prostate,” *Medical Image Analysis*, vol. 68, p. 101919, 2021.
 - [105] S. P. Venkata, B. A. Duffy, and K. Datta, “An unsupervised deep learning method for affine registration of multi-contrast brain MR images,” in *ISMRM 2022*, (Concord, CA, USA), International Society for Magnetic Resonance in Medicine, jun 2022.
 - [106] M. Jaderberg, K. Simonyan, A. Zisserman, and k. kavukcuoglu, “Spatial Transformer Networks,” in *Advances in Neural Information Processing Systems* (C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, eds.), vol. 28, Curran Associates, Inc., 2015.
 - [107] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. J. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Józefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. G. Murray, C. Olah, M. Schuster, J. Shlens,

- B. Steiner, I. Sutskever, K. Talwar, P. A. Tucker, V. Vanhoucke, V. Vasudevan, F. B. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, “TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems,” *CoRR*, vol. abs/1603.04467, 2016.
- [108] W. Segars, G. Sturgeon, S. Mendonca, J. Grimes, and B. Tsui, “4D XCAT phantom for multimodality imaging research,” *Medical physics*, vol. 37, pp. 4902–15, 09 2010.
- [109] D. F. Bauer, T. Russ, B. I. Waldkirch, C. Tönnies, W. P. Segars, L. R. Schad, F. G. Zöllner, and A.-K. Golla, “Generation of annotated multimodal ground truth datasets for abdominal medical image registration,” *International Journal of Computer Assisted Radiology and Surgery*, vol. 16, pp. 1277–1285, may 2021.
- [110] L. Soler, A. Hostettler, V. Agnus, A. Charnoz, J. Fasquel, J. Moreau, A. Osswald, M. Bouhadjar, and J. Marescaux, “3D image reconstruction for comparison of algorithm database: A patient specific anatomical and medical image database,” *IRCAD, Strasbourg, France, Tech. Rep*, 2010.
- [111] A. E. Kavur, M. A. Selver, O. Dicle, M. Barış, and N. S. Gezer, “CHAOS - Combined (CT-MR) Healthy Abdominal Organ Segmentation Challenge Data,” Apr 2019.
- [112] German Cancer Research Center (DKFZ) Division of Medical Image Computing, “Medical Imaging Interaction Toolkit (MITK).” v2021.02.
- [113] E. Raff, “A step toward quantifying independently reproducible machine learning research,” *CoRR*, vol. abs/1909.06674, 2019.
- [114] C. K. Guo, “Multi-Modal Image Registration with Unsupervised Deep Learning,” Master’s thesis, Massachusetts Institute of Technology, United States, May 2019.
- [115] A. Hering, K. Murphy, and B. van Ginneken, “Learn2Reg Challenge: CT Lung Registration - Training Data,” May 2020.
- [116] A. Hering, K. Murphy, and B. van Ginneken, “Learn2Reg Challenge: CT Lung Registration - Test Data,” September 2020.
- [117] I. Rocco, R. Arandjelović, and J. Sivic, “Convolutional neural network architecture for geometric matching,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 41, pp. 2553–2567, 2018.
- [118] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” *CoRR*, vol. abs/2010.11929, 2020.
- [119] R. Yamashita, M. Nishio, R. Do, and K. Togashi, “Convolutional neural networks: an overview and application in radiology,” *Insights into Imaging*, vol. 9, 06 2018.

-
- [120] J. Chen, E. C. Frey, Y. He, W. P. Segars, Y. Li, and Y. Du, “TransMorph: Transformer for unsupervised medical image registration,” *Medical Image Analysis*, vol. 82, p. 102615, 2022.
 - [121] W. Huang, H. Yang, X. Liu, C. Li, I. Zhang, R. Wang, and S. Wang, “A coarse-to-fine deformable transformation framework for unsupervised multi-contrast mr image registration with dual consistency constraint,” *IEEE Transactions on Medical Imaging*, vol. PP, pp. 1–1, Feb. 2021.
 - [122] Z. Zheng, W. Cao, Z. He, and Y. Luo, “Progressive anatomically constrained deep neural network for 3d deformable medical image registration,” *Neurocomputing*, vol. 465, pp. 417–427, 2021.
 - [123] Y. Zheng, X. Sui, Y. Jiang, T. Che, S. Zhang, J. Yang, and H. Li, “Symreg-gan: Symmetric image registration with generative adversarial networks,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 9, pp. 5631–5646, 2022.
 - [124] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. V. Guttag, and A. V. Dalca, “An unsupervised learning model for deformable medical image registration,” *CoRR*, vol. abs/1802.02604, 2018.
 - [125] A. B. Chapman, L. M. Guay-Woodford, J. J. Grantham, V. E. Torres, K. T. Bae, D. A. Baumgarten, P. J. Kenney, B. F. King, J. F. Glockner, L. H. Wetzel, M. E. Brummer, W. Charles O’Neill, M. L. Robbin, W. M. Bennett, S. Klahr, G. H. Hirschman, P. L. Kimmel, P. A. Thompson, and J. Philip Miller, “Renal structure in early autosomal-dominant polycystic kidney disease (adpkd): The consortium for radiologic imaging studies of polycystic kidney disease (crisp) cohort1,” *Kidney International*, vol. 64, no. 3, pp. 1035–1045, 2003.
 - [126] A. Raj, F. Tollens, A. Caroli, D. Nörenberg, and F. G. Zöllner, “Automated prognosis of renal function decline in adpkd patients using deep learning,” *Zeitschrift für Medizinische Physik*, vol. 34, no. 2, pp. 330–342, 2024. Special Issue: Artificial Intelligence in Medical Physics.
 - [127] R. Solovyev, A. A. Kalinin, and T. Gabruseva, “3d convolutional neural networks for stalled brain capillary detection,” *Computers in Biology and Medicine*, vol. 141, p. 105089, 2022.
 - [128] B. Avants, C. Epstein, M. Grossman, and J. Gee, “Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain,” *Medical Image Analysis*, vol. 12, pp. 26–41, Feb. 2008.
 - [129] A. Hering, B. van Ginneken, and S. Heldmann, “mlVIRNET: Multilevel variational image registration network,” in *Lecture Notes in Computer Science*, pp. 257–265, Springer International Publishing, 2019.
 - [130] H. Sokooti, B. de Vos, F. Berendsen, M. Ghafoorian, S. Yousefi, B. P. F. Lelieveldt, I. Isgum, and M. Staring, “3d convolutional neural networks image

- registration based on efficient supervised learning from artificial deformations,” 2019.
- [131] Z. Jiang, F.-F. Yin, Y. Ge, and L. Ren, “A multi-scale framework with unsupervised joint training of convolutional neural networks for pulmonary deformable image registration,” *Physics in Medicine & Biology*, vol. 65, p. 015011, Jan. 2020.
 - [132] L. Zhang, L. Zhou, R. Li, X. Wang, B. Han, and H. Liao, “Cascaded feature warping network for unsupervised medical image registration,” *CoRR*, vol. abs/2103.08213, 2021.
 - [133] R. Sun, J. Wu, Y. Miao, L. Ouyang, and L. Qu, “Progressive 3d biomedical image registration network based on deep self-calibration,” *Frontiers in Neuroinformatics*, vol. 16, Sept. 2022.
 - [134] J. Zhang, Y. Wang, J. Dai, M. Cavichini, D.-U. G. Bartsch, W. R. Freeman, T. Q. Nguyen, and C. An, “Two-step registration on multi-modal retinal images via deep neural networks,” *IEEE Transactions on Image Processing*, vol. 31, pp. 823–838, 2022.
 - [135] M. G. French, G. D. M. Talou, T. P. B. Gamage, M. P. Nash, P. M. Nielsen, A. J. Doyle, J. E. Iglesias, Y. Balbastre, and S. I. Young, “Diffeomorphic multi-resolution deep learning registration for applications in breast mri,” 2023.
 - [136] X. Ma, J. He, X. Liu, Q. Liu, G. Chen, B. Yuan, C. Li, and Y. Xia, “Hierarchical cumulative network for unsupervised medical image registration,” *Computers in Biology and Medicine*, vol. 167, p. 107598, 2023.
 - [137] W. Yin, J.-J. Sonke, and E. Gavves, “Pc-reg: A pyramidal prediction–correction approach for large deformation image registration,” *Medical Image Analysis*, vol. 90, p. 102978, 2023.
 - [138] Q. Chang and Y. Wang, “Structure-aware independently trained multi-scale registration network for cardiac images,” *Medical & Biological Engineering & Computing*, Feb. 2024.
 - [139] J. Honkamaa and P. Marttinen, “Sitreg: Multi-resolution architecture for symmetric, inverse consistent, and topology preserving image registration,” 2024.
 - [140] W. Ji and F. Yang, “Affine medical image registration with fusion feature mapping in local and global,” *Physics in Medicine & Biology*, 2024.
 - [141] P. Burt and E. Adelson, “The laplacian pyramid as a compact image code,” *IEEE Transactions on Communications*, vol. 31, no. 4, pp. 532–540, 1983.
 - [142] R. Gan, J. Wu, A. C. S. Chung, S. C. H. Yu, and W. M. Wells, “Multiresolution image registration based on kullback-leibler distance,” in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2004* (C. Barillot, D. R. Haynor, and P. Hellier, eds.), (Berlin, Heidelberg), pp. 599–606, Springer Berlin Heidelberg, 2004.

-
- [143] T. C. W. Mok and A. C. S. Chung, “Large deformation diffeomorphic image registration with laplacian pyramid networks,” 2020.
 - [144] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, L. Tarbox, and F. Prior, “The cancer imaging archive (tcia): Maintaining and operating a public information repository,” *Journal of Digital Imaging*, vol. 26, p. 1045–1057, July 2013.
 - [145] G. Litjens, O. Debats, J. Barentsz, N. Karssemeijer, and H. Huisman, “Computer-aided detection of prostate cancer in mri,” *IEEE Transactions on Medical Imaging*, vol. 33, p. 1083–1092, May 2014.
 - [146] G. Litjens, O. Debats, J. Barentsz, N. Karssemeijer, and H. Huisman, “Spie-aapm prostatex challenge data,” 2017.
 - [147] P. Rogelj, S. Kovačič, and J. C. Gee, “Point similarity measures for non-rigid registration of multi-modal data,” *Computer Vision and Image Understanding*, vol. 92, no. 1, pp. 112–140, 2003.
 - [148] A. Melbourne, D. Atkinson, M. J. White, D. Collins, M. Leach, and D. Hawkes, “Registration of dynamic contrast-enhanced mri using a progressive principal component registration (ppcr),” *Physics in Medicine and Biology*, vol. 52, p. 5147–5156, Aug. 2007.
 - [149] L. Ruthotto, E. Hodneland, and J. Modersitzki, “Registration of dynamic contrast enhanced mri with local rigidity constraint,” in *Biomedical Image Registration* (B. M. Dawant, G. E. Christensen, J. M. Fitzpatrick, and D. Rueckert, eds.), (Berlin, Heidelberg), pp. 190–198, Springer Berlin Heidelberg, 2012.
 - [150] N. Cai, H. Chen, Y. Li, Y. Peng, and L. Guo, “Registration on dce-mri images via multi-domain image-to-image translation,” *Computerized Medical Imaging and Graphics*, vol. 104, p. 102169, 2023.
 - [151] Y. Sun, Y. Gu, F. Shi, J. Liu, G. Li, Q. Feng, and D. Shen, “Coarse-to-fine registration and time-intensity curves constraint for liver dce-mri synthesis,” *Computerized Medical Imaging and Graphics*, vol. 111, p. 102319, 2024.
 - [152] C. Mattusch, U. Bick, and F. Michallek, “Development and validation of a four-dimensional registration technique for dce breast mri,” *Insights into Imaging*, vol. 14, Jan. 2023.
 - [153] G. Wu, H. Jia, Q. Wang, and D. Shen, “Sharpmean: Groupwise registration guided by sharp mean image and tree-based registration,” *NeuroImage*, vol. 56, p. 1968–1981, June 2011.
 - [154] S. Ying, G. Wu, Q. Wang, and D. Shen, “Groupwise registration via graph shrinkage on the image manifold,” in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2323–2330, 2013.
 - [155] A. A. Joshi, R. M. Leahy, R. D. Badawi, and A. J. Chaudhari, “Registration-based morphometry for shape analysis of the bones of the human wrist,” *IEEE Transactions on Medical Imaging*, vol. 35, p. 416–426, Feb. 2016.

- [156] J.-M. Guyader, W. Huizinga, D. H. J. Poot, M. van Kranenburg, A. Uitterdijk, W. J. Niessen, and S. Klein, “Groupwise image registration based on a total correlation dissimilarity measure for quantitative mri and dynamic imaging data,” *Scientific Reports*, vol. 8, Aug. 2018.
- [157] B. W. Papież, D. R. McGowan, M. Skwarski, G. S. Higgins, J. A. Schnabel, and M. Brady, *Fast Groupwise 4D Deformable Image Registration for Irregular Breathing Motion Estimation*, p. 37–46. Springer International Publishing, 2018.
- [158] R. Haase, S. Heldmann, and J. Lellmann, “Deformable groupwise image registration using low-rank and sparse decomposition,” *Journal of Mathematical Imaging and Vision*, vol. 64, p. 194–211, Jan. 2022.
- [159] T. Sentker, F. Madesta, and R. Werner, “Gdl-fire4d: Deep learning-based fast 4d ct image registration,” in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018* (A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, and G. Fichtinger, eds.), (Cham), pp. 765–773, Springer International Publishing, 2018.
- [160] S. Ahmad, J. Fan, P. Dong, X. Cao, P.-T. Yap, and D. Shen, “Deep learning deformation initialization for rapid groupwise registration of inhomogeneous image populations,” *Frontiers in Neuroinformatics*, vol. 13, May 2019.
- [161] T. Che, Y. Zheng, J. Cong, Y. Jiang, Y. Niu, W. Jiao, B. Zhao, and Y. Ding, “Deep group-wise registration for multi-spectral images from fundus images,” *IEEE Access*, vol. 7, pp. 27650–27661, 2019.
- [162] Y. Fu, Y. Lei, T. Wang, K. Higgins, J. D. Bradley, W. J. Curran, T. Liu, and X. Yang, “Lungregnet: An unsupervised deformable image registration method for 4d-ct lung,” *Medical Physics*, vol. 47, p. 1763–1774, Feb. 2020.
- [163] E. Martín-González, T. Sevilla, A. Revilla-Orodea, P. Casaseca-de-la Higuera, and C. Alberola-López, “Groupwise non-rigid registration with deep learning: An affordable solution applied to 2d cardiac cine mri reconstruction,” *Entropy*, vol. 22, p. 687, June 2020.
- [164] Y. Zhang, X. Wu, H. M. Gach, H. Li, and D. Yang, “Groupregnet: a group-wise one-shot deep learning-based 4d image registration method,” *Physics in Medicine & Biology*, vol. 66, p. 045030, feb 2021.
- [165] J. Yang, T. Küstner, P. Hu, P. Liò, and H. Qi, “End-to-end deep learning of non-rigid groupwise registration and reconstruction of dynamic mri,” *Frontiers in Cardiovascular Medicine*, vol. 9, Apr. 2022.
- [166] C. C. Hammecher, K. van Garderen, M. Smits, P. Wesseling, B. Westerman, P. French, M. Kouwenhoven, R. Verhaak, F. Vos, E. Bron, and B. Li, “Deep learning-based group-wise registration for longitudinal mri analysis in glioma,” 2023. arXiv.

-
- [167] X. Luo, X. Wang, L. Shapiro, C. Yuan, J. Feng, and X. Zhuang, “Bayesian intrinsic groupwise image registration: Unsupervised disentanglement of anatomy and geometry,” 2024. arXiv.
 - [168] F. G. Zöllner, K. Zahn, T. Schaible, S. O. Schoenberg, L. R. Schad, and K. W. Neff, “Quantitative pulmonary perfusion imaging at 3.0 t of 2-year-old children after congenital diaphragmatic hernia repair: initial results,” *European Radiology*, vol. 22, p. 2743–2749, June 2012.
 - [169] F. G. Zöllner, M. Daab, M. Weidner, V. Sommer, K. Zahn, T. Schaible, G. Weisser, S. O. Schoenberg, K. W. Neff, and L. R. Schad, “Semi-automatic lung segmentation of dce-mri data sets of 2-year old children after congenital diaphragmatic hernia repair: Initial results,” *Magnetic Resonance Imaging*, vol. 33, p. 1345–1349, Dec. 2015.
 - [170] A. Rosset, L. Spadola, and O. Ratib, “Osirix: An open-source software for navigating in multidimensional dicom images,” *Journal of Digital Imaging*, vol. 17, p. 205–216, June 2004.
 - [171] F. G. Zöllner, G. Weisser, M. Reich, S. Kaiser, S. O. Schoenberg, S. P. Sourbron, and L. R. Schad, “Ummperfusion: an open source software tool towards quantitative mri perfusion analysis in clinical routine,” *Journal of Digital Imaging*, vol. 26, p. 344–352, July 2012.
 - [172] F. G. Zöllner, M. Daab, S. P. Sourbron, L. R. Schad, S. O. Schoenberg, and G. Weisser, “An open source software for analysis of dynamic contrast enhanced magnetic resonance images: Ummperfusion revisited,” *BMC Medical Imaging*, vol. 16, Jan. 2016.
 - [173] M. Weis, V. Sommer, F. G. Zöllner, C. Hagelstein, K. Zahn, T. Schaible, S. O. Schoenberg, and K. W. Neff, “Region of interest-based versus whole-lung segmentation-based approach for mr lung perfusion quantification in 2-year-old children after congenital diaphragmatic hernia repair,” *European Radiology*, vol. 26, p. 4231–4238, Apr. 2016.

10. Publications

Journal Papers

- **A. Strittmatter**, M. Weis, F. G. Zöllner. A Groupwise Multiresolution Network for DCE-MRI Image Registration. *Scientific Reports*. Submitted on 12.09.2024.
- **A. Strittmatter**, F. G. Zöllner. Multistep networks for deformable multimodal medical image registration. *IEEE Access* 12, pp. 82676–82692, doi: 10.1109/ACCESS.2024.3412216, 2024.
- **A. Strittmatter**, L. R. Schad, F. G. Zöllner. Deep learning-based affine medical image registration for multimodal minimal-invasive image-guided interventions – A comparative study on generalizability. *Zeitschrift für Medizinische Physik*, 34 (2), pp. 291-317, doi: 10.1016/j.zemedi.2023.05.003, 2024.
- **A. Strittmatter**, A. Caroli, F. G. Zöllner. A Multistage Rigid-Affine-Deformable Network for Three-Dimensional Multimodal Medical Image Registration. *Applied Sciences*, 13(24), 13298, doi: 10.3390/app132413298, 2023.

Conference Abstracts

- A. Streuer, A. Hertel, S. J. Diehl, T. Boch, **A. Strittmatter**, F. G. Zöllner, V. Nowak, N. Weimer, J. Obländer, I. Palme, M. Göl, S. O. Schoenberg, W. K. Hofmann, S. Loges, D. Nowak, M. F. Froelich. Integration of radiologic features and genomic profiles in lung cancer to investigate intratumor heterogeneity and clonal dynamics. *Proc. DGHO Congress, Basel, Switzerland, 2024*.
- **A. Strittmatter**, F. Zöllner. Deep Learning-Based Multistep Deformable Medical Image Registration for Multimodal Minimal-Invasive Image-Guided Intervention. *Proc. STMM Summer Symposium, Mannheim, 2024*.
- **A. Strittmatter**, L. Schad, and F. Zöllner. Deep Learning-Based Multistep Deformable Medical Image Registration for Multimodal Minimal-Invasive Image-Guided Intervention. *Proc. ISMRM Congress, Singapore, 2024*.
- **A. Strittmatter**, A. Hertel, S. Diehl, M. F. Froelich, S. O. Schoenberg, S. Loges, T. Boch, D. Nowak, A. Streuer, L. R. Schad, and F. G. Zöllner. A Multistage Registration of CT and Biopsy CT Images of Lung Tumors. *German Workshop on Medical Image Computing - BVM 2024, Erlangen, Germany, 2024*.
- **A. Strittmatter**, A. Hertel, S. Diehl, M. F. Froelich, S. O. Schoenberg, S. Loges, T. Boch, D. Nowak, A. Streuer, L. R. Schad, and F. G. Zöllner. A multistage registration of CT and biopsy CT images of lung tumors. *Proc. 6th Conference on Image-Guided Interventions, Mannheim, Germany, 2023*.
- S. Pasini, T. Vendelboe, A. Raj, L. Garcia-Ruiz, **A. Strittmatter**, R. Echeverria-Chasco, G. Villa, P. Brambilla, E. S. S. Hansen, S. Ringgaard, L. Pierotti, M. Renzulli, F. G. Zöllner, M. Fernandez-Seara, S. Francis, C. Laustsen, and A. Caroli. Multi-center and multi-vendor reproducibility of T1, T2 and ADC phantom data on 1.5T and 3T MRI scanners. *Proc. 5th International Meeting on Renal Imaging, Ghent, Belgium, 2023*.
- **A. Strittmatter**, L. R. Schad, and F. G. Zöllner. A Multistage Rigid-Affine-Deformable Network for Three-Dimensional Multimodal Medical Image Regis-

- tration. Proc. 5th International Meeting on Renal Imaging, Ghent, Belgium, 2023.
- A. Adlung, F. Tollens, **A. Strittmatter**, N. Westhoff, D. Hausmann, S. Schoenberg, L. Schad, D. Nörenberg, and F. Zöllner. Primary prostate cancer characterisation using internal references for ^{23}Na MRI quantification. Proc. ISMRM Congress, Toronto, Canada, 2023.
 - A. Caroli, S. Pasini, T. Vandelboe, A. Raj, L. Garcia-Ruiz, **A. Strittmatter**, R. Echeverria-Chasco, G. Villa, P. Brambilla, E. S. S. Hansen, S. Ringgaard, F. Zöllner, M. Fernandez-Seara, S. Francis, and C. Laustsen. Multicenter and multi-vendor reproducibility of T1, T2 and ADC phantom data on 1.5T and 3T MRI scanners. Proc. ISMRM Congress, Toronto, Canada, 2023.
 - **A. Strittmatter**, L. Schad, and F. Zöllner. Deep Learning-Based Affine Medical Image Registration - A Review and Comparative Study on Generalizability. Proc. ISMRM Congress, Toronto, Canada, 2023.
 - F. Tollens, A. Adlung, **A. Strittmatter**, N. Westhoff, D. Hausmann, S. O. Schönberg, L. R. Schad, F. G. Zöllner, and D. Nörenberg. Charakterisierung von Prostatakarzinomen in der ^{23}Na MRT mittels interner Referenzen. Proc. RÖKO Congress, Wiesbaden, Germany, 2023.
 - A. Raj, F. Tollens, **A. Strittmatter**, L. Hansen, D. Nörenberg, and F. G. Zöllner. Deep Learning based Total Kidney Volume Segmentation in Autosomal Dominant Polycystic Kidney Disease. Proc. ISMRM Congress, London, United Kingdom, 2022.

Supervised Theses

- **Master's Thesis** "Improving multi-modal synthetic medical image generation using unpaired image-to-image translation", Nektarios Winter, 2022.
- **Master's Thesis** "Development of a multistep neural network for deformable 3D medical image registration", Reeth Nalamitha, 2023.

11. Curriculum Vitae

Personal Data

Name **Anika Strittmatter.**

Date of Birth **13. May 1997.**

Place of Birth **Mannheim.**

Nationality **German.**

Education

24 November **Doctoral Candidate: Dr. sc. hum.,**
2021 – today *Topic: Development of Methods for the Registration of Multimodal Medical Images Using Deep Learning,*
Ruprecht Karl University of Heidelberg,
Supervisor: Prof. Dr. Frank Gerrit Zöllner.

09.2019 – **Master of Science: Biomedical Engineering,**
04.2021 *Master's thesis: "Development of a measuring system for determining the pulse transit time using optical and piezoelectric sensors for blood pressure measurement on the upper arm",*
Mannheim University of Applied Sciences, Grade: 1.0.

03.2016 – **Bachelor of Science: Biomedical Engineering,**
08.2019 *Bachelor's thesis: "Determination of the charge injection capacity of microelectrodes based on long-term measurements",*
Mannheim University of Applied Sciences, Grade: 1.1.

15 June 2015 **Abitur,**
Grade: 1.9.

2007 – 2015 **Gymnasium,**
Privatgymnasium St. Leon-Rot.

Experience

15 June 2021 – **Research Assistant,**
today *Computer Assisted Clinical Medicine, Medical Faculty Mannheim, Ruprecht Karl University of Heidelberg,*
since 09.2021: Lecturer Medical students,
since 10.2021: Lecturer Biomedical Engineering.

07.2020 – **Working Student,**
08.2021 *AFS ASS Field Service GmbH, Service-Backoffice for in-vitro diagnostic devices of Roche Diagnostics Deutschland GmbH.*

- 08.2018 – **Working Student**,
06.2020 *MTP Medizintechnik Potsdam GmbH and PFS-Walldorf GmbH*, Field Service Engineer and Service-Backoffice for in-vitro diagnostic devices of Roche Diagnostics Deutschland GmbH.
- 03.2018 – **Practical Semester**,
08.2018 *Roche Diagnostics Deutschland GmbH*, Field Service Engineer in Customer Support, Medical Laboratory Diagnostics.
- 10.2017 – **Research Assistant**,
01.2018 *Mannheim University of Applied Sciences, Institute for Software Engineering and Data Communication, Professor Dr. Damm*.
- 03.2017 – **Tutor Computer Science**,
02.2021 *Mannheim University of Applied Sciences, Institute for Software Engineering and Data Communication*, Supervision of programming exercises in C.

Awards & Honors

- 19 July 2024 **STMM Summer Symposium Scientific Poster Prize**,
Poster prize for scientific merit, Deep Learning-Based Multi-step Deformable Medical Image Registration for Multimodal Minimal-Invasive Image-Guided Intervention.
- 2021 **Karl-Völker-Preis, Award for an outstanding Master's thesis.**
- 09.2017 – **Deutschlandstipendium (scholarship)**,
08.2018 *sponsored by Roche Diagnostics GmbH.*

12. Acknowledgements

With the finalisation of this thesis, I would like to take the opportunity to thank all the people who have supported me on this journey.

First of all, I would like to express my sincere gratitude to Prof. Dr. Lothar Schad for offering me the opportunity to start my doctoral studies in his research group.

I would also like to thank Prof. Dr. Frank Zöllner for supervising this thesis and for always taking time to discuss ideas. His expertise and feedback have been invaluable in shaping this work.

Special thanks are also due to Professor Dr. Stefan Heger for his assistance in finding a doctoral position and for the valuable advice and personal experiences he has shared.

I am grateful to all my CKM colleagues for their valuable contributions through thought-provoking discussions that greatly influenced the development of this thesis.

Many thanks to my colleagues from the RESPECT project for their valuable collaboration. I would also like to thank the Federal Ministry of Education and Research (BMBF) for the financial support provided as part of this project.

Finally, I am deeply thankful to my family and friends for their constant support and encouragement during my studies, which has always been a source of motivation. Above all, I would like to thank my boyfriend Mick for his unwavering patience, encouragement and understanding, as well as for his invaluable support during the stressful phases of this work.