UNIVERSITÄT ZU LÜBECK
INSTITUTE OF MEDICAL INFORMATICS

From the Institute of Medical Informatics
of the University of Lübeck
Director: Prof. Dr. rer. nat. habil. Heinz Handels

Architectures and Optimisation for Learning-Based Medical Image Registration

Dissertation
for
Fulfillment of Requirements for the Doctoral Degree
of the University of Lübeck

from the Department of Computer Sciences and Technical Engineering

Submitted by
Hanna Siebert
from Stuttgart

Lübeck, 2023

First referee: Prof. Dr. Mattias Heinrich
Second referee: Prof. Dr. rer. nat. habil. Floris Ernst

Date of oral examination: 28 February 2024

Approved for printing. Lübeck, 18 March 2024

Abstract

Radiological images are of fundamental importance in healthcare and medical research as they enable information gain that supports better medical treatment and research advances. In both research and patient care, there is a demand to compare image data or to fuse the information of images. With image registration techniques, it is possible to align image data that is acquired at different times, with different imaging modalities, or from different patients. This enables direct image comparison or image fusion even though the image data to be considered may initially show inconsistencies.

Deep learning-based medical image registration methods have become the main focus of research in recent years. Their ability to process large datasets with fast inference times is of particular benefit for applications such as image guidance during surgery or processing large image databases for population studies. Yet, there are several challenges that need to be overcome in order to compete with conventional methods or even excel them in terms of robustness, versatile application, and deformation precision and plausibility.

This thesis presents four different medical image registration approaches that are fully or partially deep learning-based. Therefore, a specific focus lies on the aspects supervision, decoupling, and versatility. With regard to the learning procedure, different forms of supervision are utilised and compared. In the course of this thesis, unsupervised and weakly supervised approaches are presented and compared, an inverse consistency constraint for robust deformation field generation is introduced, and a self-supervised learning strategy based on cycle constraints is proposed. Concerning the architecture of image registration methods, forms of decoupling are brought into focus. In particular, for groupwise image registration, a deep learning model that splits up the latent space of an autoencoder architecture to decouple shape and appearance representation is presented. For pairwise image registration, the benefits of a separated feature extraction for moving and fixed images are investigated. Regarding versatility, concepts that aim for transferability to a wide range of image registration approaches and tasks are introduced and a self-configuring multitask medical image registration approach is presented. Overall, each of the introduced methods brings its findings that can contribute to the future development of improved learning-based medical image registration methods.

Zusammenfassung

Im Gesundheitswesen und in der medizinischen Forschung sind radiologische Bilddaten von grundlegender Bedeutung. Sie ermöglichen einen Informationsgewinn, der eine bessere medizinische Behandlung und Fortschritte in der Forschung unterstützt. Sowohl in der Patientenversorgung als auch in der Forschung besteht der Bedarf, Bilddaten zu vergleichen oder Bildinformationen zu fusionieren. Registrierungstechniken ermöglichen es, Bilddaten aufeinander auszurichten, die zu unterschiedlichen Zeiten, mit unterschiedlichen Bildgebungsmodalitäten oder von unterschiedlichen Patienten aufgenommen wurden. Dies ermöglicht direkten Bildvergleich oder Bildfusionierung, selbst wenn die zu berücksichtigenden Bilddaten zunächst Inkonsistenzen aufweisen.

In den letzten Jahren sind Deep Learning basierte Methoden in den Fokus der Forschung zu medizinischer Bildregistrierung gerückt. Ihre Fähigkeit, große Datensätze mit schnellen Inferenzzeiten zu verarbeiten, ist von besonderem Nutzen für Anwendungen wie bildgestützte Operationen oder die Verarbeitung großer Bilddatenbanken für Bevölkerungsstudien. Was Robustheit, Anwendungsvielseitigkeit, Deformationsgenauigkeit und -plausibilität betrifft, gibt es jedoch eine Reihe von Herausforderungen, die bewältigt werden müssen, um mit konventionellen Methoden mitzuhalten oder diese zu übertreffen.

In dieser Arbeit werden vier verschiedene Ansätze zur medizinischen Bildregistrierung vorgestellt, die ganz oder teilweise auf Deep Learning basieren, wobei ein besonderer Schwerpunkt auf den Aspekten Überwachung, Entkopplung und Vielseitigkeit liegt. Im Hinblick auf Lernverfahren werden verschiedene Formen der Überwachung eingesetzt und verglichen. Die Arbeit stellt unüberwachte und schwach überwachte Ansätze vor und vergleicht diese. Darüber hinaus wird eine Methode für Konsistenz von inversen Transformationen zur robusten Deformationsfelderzeugung eingeführt und eine selbstüberwachte Lernstrategie auf Basis von Registrierungszyklen vorgeschlagen. In Bezug auf die Architektur von Bildregistrierungsverfahren stehen Formen der Entkopplung im Fokus. Für gruppenweise Bildregistrierung wird ein Deep-Learning-Modell vorgestellt, das den latenten Raum einer Autoencoder-Architektur derart aufteilt, dass Deformations- und Templatelernen entkoppelt sind. Für paarweise Bildregistrierung erfolgt die Untersuchung der Vorteile einer separierten Merkmalsextraktion für das Referenzbild und das zu deformierende Bild. Hinsichtlich Vielseitigkeit werden Konzepte vorgestellt, die auf die Übertragbarkeit auf ein breites Spektrum von Bildregistrierungsansätzen und -aufgaben abzielen. Des Weiteren wird ein selbstkonfigurierender Multitask-Ansatz für medizinische Bildregistrierung präsentiert. Insgesamt liefert jede der vorgestellten Methoden Erkenntnisse, die zur zukünftigen Entwicklung verbesserter lernbasierter medizinischer Bildregistrierungsmethoden beitragen können.

Contents

Chapter 1

Introduction

Radiological images are of fundamental importance in healthcare and medical research. They allow a look inside the patient and thus enable important information gain that supports better medical treatment and research advances. In both research and patient care, there is a demand to compare image data or to fuse the information of images. For such tasks, image registration plays an important role. With the help of image registration techniques, it is possible to align image data that is acquired at different times, with different imaging modalities, or from different patients. Through the alignment of images, the possibility for direct image comparison or image fusion is provided even though the image data to be considered may initially show inconsistencies due to variations in, for example, acquisition angle, patient positioning, image resolution, or image contrast. There are numerous possible applications of image registration methods in medical context. Aligned images acquired at different times can for example be used for motion analysis of the heart or lungs; aligned images acquired with different imaging modalities can for example be used for treatment planning based on fused image data; and aligned images acquired from different patients can for example be used for population studies.

All fields of application entail certain requirements and challenges for the image registration methods. It is essential that anatomical structures are aligned precisely and that the employed deformation fields are smooth and plausible. Furthermore, for many applications, such as image-guided surgery, it is important that the image registration process can be performed fast to preferably enable real-time applications. The challenges that may be encountered include the possibility of large displacements or movement between images which need to be captured. Another challenging aspect is the size of the image data to be aligned. Modern imaging modalities provide large high-resolution image data that is often three-dimensional. This leads to many degrees of freedom during the registration procedure and an increased computation complexity.

To address these challenges, several survey papers indicate that there are a variety of medical image registration methods that propose to use conventional non-learning-based algorithms [Hill et al., 2001; Sotiras et al., 2013] or deep learning-based approaches [Fu et al., 2020; Zou et al., 2022]. In recent years, deep learning-based methods have become the main focus of research, which is reflected, for example, in the contributions

for the workshop on biomedical image registration 2022 [Hering et al., 2022b] or the submissions for the MICCAI-Learn2Reg challenge [Hering et al., 2022a]. The ability of deep learning-based image registration methods to process large datasets with inference times that outperform most conventional algorithms is of particular benefit for applications such as image guidance during surgery or processing large image databases for population studies.

Yet, there are several challenges that need to be faced in order for learning-based image registration techniques to compete with or excel conventional methods in terms of robustness, versatile application, and deformation precision and plausibility. Deep learning-based image registration methods are often developed in a very task-specific way. The network architectures and training procedures are chosen for a specific application and need to be replaced or modified if transferred to another image registration task. To date, no universally applicable deep learning-based medical image registration method has been published that achieves state-of-the-art registration results for a very wide range of medical image datasets without requiring user intervention. Another challenge is the plausibility of the deformation estimation process, especially when capturing large displacements or movements. Unlike conventional methods, deep learning-based approaches use network architectures in which it is not necessarily obvious which part of the network fulfils which function within the registration process. As for conventional methods, the choice of the objective function plays an important role in the resulting registration performance. Hence, the selection of the objective function and form of supervision for network training is challenging and of particular importance.

This thesis therefore specifically focuses on three aspects of learning-based image registration methods: *supervision*, *decoupling*, and *versatility*. The following section (Sec. 1.1) describes the research objectives with respect to these aspects. In Sec. 1.2, the main contributions and the organisation of the thesis are outlined.

1.1 Objectives

The aim of this thesis is to develop medical image registration methods that are fully or partially deep learning-based. In the course of the individual chapters, there are three main considerations within the context of image registration that are taken up multiple times. With regard to the learning procedure, different forms of *supervision* are utilised and compared. Concerning the architecture of image registration methods, forms of *decoupling* in particular are brought into focus. With respect to the usability of the presented methods for medical image registration tasks, the thesis aims for *versatility* and introduces concepts that strive for transferability to a wide range of image registration approaches and tasks. Further motivation for these three aspects, as well as detailed research objectives, are given subsequently.

Supervision

Supervised learning with dense ground truth deformation fields is only possible for synthetic training data due to the inherent ambiguity in homogeneous areas of real-life deformation fields. Unsupervised image registration approaches operate on image similarities and compare warped moving images to fixed images. Weakly supervised methods make use of auxiliary information such as segmentation labels during the training procedure. The formulation of the loss function used to train deep learning models can therefore vary widely depending on the registration task and the available data. Through the specific formulation of the loss function, the characteristics of the deformation estimated by deep learning-based image registration methods can be adjusted. Especially the smoothness of the deformation fields and the precision of the registration are often taken into account. For parts of this thesis, inverse consistency of deformations is of interest, since registration is often understood as an asymmetric problem with a specific registration direction. Also of relevance for this thesis is a supervision procedure that uses registration cycles and minimises the discrepancy within these cycles. In particular, in terms of supervision, this thesis aims to investigate

- how an inverse consistency constraint for unsupervised registration learning affects the robustness and plausibility of deformation fields estimated by a deep learning model.
- how registration performance differs when comparing weakly supervised and unsupervised registration learning.
- how well a self-supervised learning strategy using cycle constraints is applicable for image registration.

Decoupling

In this thesis, *decoupling*, also denoted as *disentanglement*, refers to the concept of splitting up the neural network architecture used for an image registration method into separate parts. The motivation behind this is to achieve precise and more interpretable registration results, since in deep learning architectures it is often unclear which part of the network is responsible for e.g. feature extraction or deformation estimation. This thesis presents and discusses different architectural design possibilities that use some form of decoupling. In particular, this thesis aims to examine

- how splitting up the latent space of an autoencoder architecture can be used to decouple shape and appearance representation for groupwise image registration.
- how beneficial the use of separated but shared input network blocks for feature extraction for moving and fixed input images might be for pairwise image registration compared to single-stream architectures.

- how separated trainable feature extraction for input image pairs can be used in combination with non-trainable modules within image registration frameworks.

Versatility

Alignment of image data is required for various clinical applications and for research purposes. It is possible to categorise registration of monomodal or multimodal image data, as well as registration of intra-patient or inter-patient data. Especially deep learning-based image registration methods are often developed in an application-specific way and struggle with direct transferability to other tasks. This thesis aims for

- basic concepts for image registration that can easily be transferred to various deep learning-based image registration methods.
- a pairwise image registration approach that can be applied directly to a wide variety of tasks, regardless of whether mono- or multimodal or whether intra- or inter-patient image data are to be processed.

1.2 Contribution and Overview

The thesis presents different approaches for medical image registration. Since 2020, the Learn2Reg challenge [Hering et al., 2022a] provides several medical image datasets and the opportunity to evaluate and compare image registration algorithms. The datasets of the challenge contain images for multiple anatomies, mono- and multimodal image data, and intra- and inter-patient data. All methods presented in this thesis are evaluated on image data related to the Learn2Reg challenge datasets. Overall, the thesis comprises four methodological chapters, each of which introduces another model for medical image registration. Fig. 1.1 depicts the components of the individual methods and indicates the properties of the respective experiments. The organisation of the thesis is as follows:

- **Chapter 2** provides medical and methodological background knowledge for the methods presented in this thesis. The medical background knowledge given relates to how image registration is applied in medical context. As methodological background knowledge, fundamentals of image registration and basic principles of deep learning-based image registration are described.
- **Chapter 3** addresses the challenges of decoupled shape and appearance representation learning and inverse consistency of deformation field generation. It presents a deep learning method for deformable groupwise medical image registration. The end-to-end trainable autoencoder network architecture consists of an encoder which provides a latent vector that is split up and given to two decoupled decoders. The

appearance decoder estimates a template and the shape decoder estimates a deformation field. Image generation is achieved by spatial warping of the estimated template with the estimated deformation. Training is performed in an unsupervised fashion, and an inverse consistency constraint for robust deformation prediction is introduced. Experiments are performed on inter-patient brain MR scans. The method has been published in

[Siebert et al., 2020] Siebert, H. et al. "Deep Groupwise Registration of MRI Using Deforming Autoencoders". In: *Bildverarbeitung für die Medizin 2020 –BVM 2020.* Springer, 2020, pp. 236–241

[Siebert et al., 2021c] Siebert, H. et al. "Learning inverse consistent 3D groupwise registration with deforming autoencoders". In: *Medical Imaging 2021: Image Processing.* Vol. 11596. 2021, pp. 89–95

- **Chapter 4** is dedicated to the comparison of basic architecture design options for U-Net-based deformable registration learning. It addresses the problem that single-stream registration network architectures such as VoxelMorph [Balakrishnan et al., 2019] are not particularly well suited for inter-patient abdominal image data registration with large deformations. A two-stream architecture that includes a partially disentangled feature extraction for pairwise image registration is proposed. The experiments are conducted on inter-patient abdominal CT scans, compare several architecture modifications, and involve a comparison of results for unsupervised and weakly supervised training. This method has been published in

  [Siebert et al., 2021b] Siebert, H. et al. "Architecture Matters: Evaluating Design Choices for Deep Learning Registration Networks". In: *Bildverarbeitung für die Medizin 2021 –BVM 2021.* Springer, 2021, pp. 111–116

- **Chapter 5** introduces a supervision procedure that uses cycle constraints for self-supervised rigid registration learning. The proposed method minimises a discrepancy within cycles that contain a multimodal image pair and an image generated by applying a synthetic random transformation to the moving image. Through this, features suitable for multimodal image registration can be learned without a predefined multimodal similarity metric. The architecture used for image registration in this chapter comprises feature extraction with a CNN with initially separated feature encoding for each modality, a correlation layer, and a least squares fitting procedure for transformation computation. Experiments are conducted on intra-patient abdominal CT and MR scans and include a comparison to metric supervision and label supervision. The approach has been published in

  [Siebert et al., 2022b] Siebert, H. et al. "Learning a Metric for Multimodal Medical Image Registration without Supervision Based on Cycle Constraints". *Sensors* 22 [3], 2022

[Siebert et al., 2021a] Siebert, H. et al. "Learning a Metric without Supervision: Multimodal Registration using Synthetic Cycle Discrepancy". In: *International Conference on Medical Imaging with Deep Learning –Extended Abstract Track –MIDL 2021*. 2021

- **Chapter 6** presents a multitask image registration method that extracts semantic or hand-crafted input image features and uses a coupled convex optimisation followed by Adam-based instance optimisation. The approach makes use of pretrained semantic feature extraction models for the individual datasets and combines them with the optimisation procedure for deformation field computation. An automatic hyperparameter selection technique is proposed that examines various hyperparameter combinations and provides a self-configuring image registration framework. Experiments are conducted on all currently available Learn2Reg challenge datasets. The method presented in this thesis is based on the approach published in

  [Siebert et al., 2022a] Siebert, H. et al. "Fast 3D Registration with Accurate Optimisation and Little Learning for Learn2Reg 2021". In: *Biomedical Image Registration, Domain Generalisation and Out-of-Distribution Analysis –MICCAI 2021 Challenges*. 2022, pp. 174–179

- **Chapter 7** summarises the contributions of the presented methods and discusses the key findings of this thesis with regard to the objectives described in Sec. 1.1. Finally, a brief overview of ongoing research in the field of deep learning-based image registration and an overall conclusion is given.

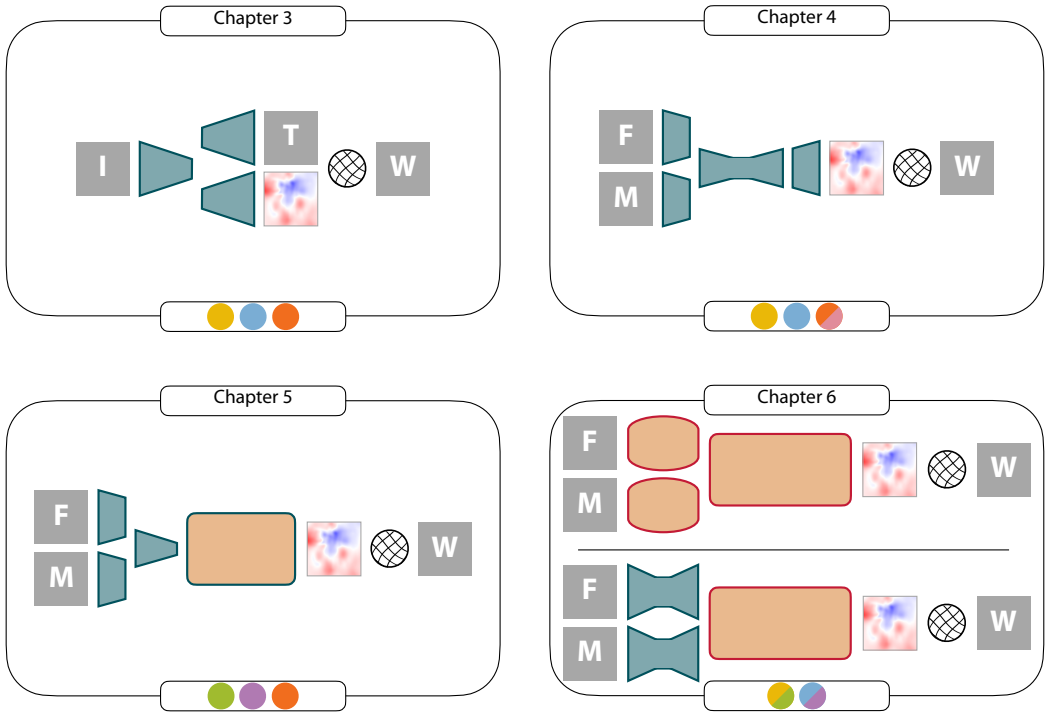


Fig. 1.1: Overview of the medical image registration models presented in this thesis.

Chapter 2

Background

Image registration aims to find spatial correspondences between a pair or a group of images. With the help of a spatial transformation, which maximises the similarity measured by a certain metric and subject to a plausibility constraint, the given input image data is aligned. In a medical context, image registration comprises methods that are, for example, employed for diagnostic tasks, treatment planning, image-guided interventions, or motion analysis. This chapter provides medical (Sec. 2.1) and methodological (Sec. 2.2 and Sec. 2.3) background knowledge for the medical image registration approaches presented in the following chapters of this thesis.

2.1 Medical Background

Medical image registration deals with the anatomically correct alignment of image data. The image data employed for registration is derived from different times, different patients, or different imaging modalities as exemplified in Fig. 2.1. Registration is used, for example, to display different images of one or more patients together and to be able to analyse image data in direct comparison. For medical image datasets, it is mainly needed to compensate for patient positioning differences, imaging modality differences, or patient movements.

Registration of intra-patient image data that has been acquired at different times could be used for motion analysis (like Fig. 2.1 (a), left image pair) or data comparison over time (like Fig. 2.1 (a), right image pair) for research and clinical purposes. The image data used for monomodal intra-patient image registration may be acquired for diagnosis, for treatment planning and monitoring, or for research. The time interval that separates images could be (milli)seconds (e.g. in motion or functional analysis), minutes (e.g. when acquiring pre- and post-contrast images), or weeks (e.g. when monitoring tumour growth) [Hill et al., 2001]. Motion analysis could be, for example, relevant for the analysis of heartbeats or the respiratory movement of the lungs and surrounding anatomical structures. Cardiac motion analysis with the help of image registration methods can be used to support the evaluation of the cardiac function in order to detect dysfunctions [Wiputra et al., 2020]. The analysis of lung motion is of interest in radiotherapy. The respiratory movement of the lungs influences the movement of lung

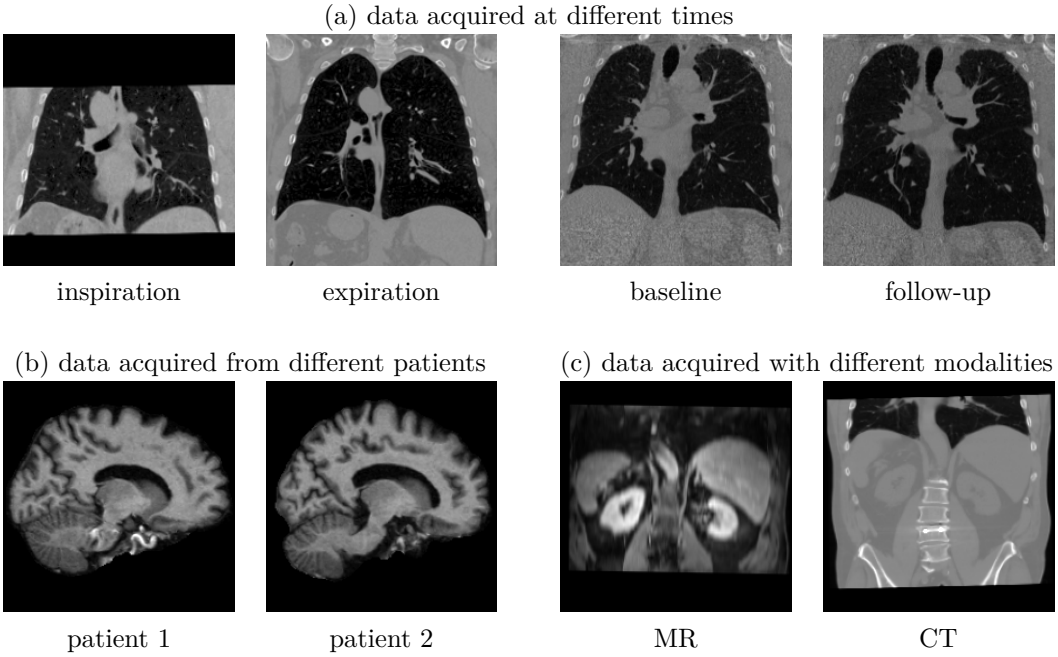


Fig. 2.1: Examples for medical image pairs used for registration: monomodal intra-patient registration of lung CT scans acquired at different times (a), monomodal inter-patient registration of brain MR scans (b), and multimodal intra-patient registration of abdominal MR and CT scans (c).

tumours. Registration methods can be used to gain information about the movement behaviour of co-moving tumours in different lung regions at different respiratory phases. The obtained information can then be used to improve radiotherapeutic treatment, allowing tumours to be radiated more precisely and surrounding tissue to be spared [Handels, 2000]. Registration of scans with a longer time interval between acquisitions enables the analysis of changes in anatomical structures over time for research purposes in longitudinal studies or for disease progression monitoring like tumour growth or degenerative disease monitoring. Monomodal intra-patient lung image data is used, among others, for the approach introduced in Chapter 6.

Registration of inter-patient data (like Fig. 2.1 (b)) can, for example, be useful for population studies, for atlas generation, or for clinical purposes. Especially for monomodal inter-patient image data, not only pairwise image registration, but also groupwise image registration is of interest. Groupwise image registration methods aim to align multiple images to one reference image. This helps to analyse large medical image datasets and to generate atlas images. In general, with inter-patient image registration methods, it is possible to compare image data acquired from multiple individuals and thereby to generate knowledge about anatomical variability across the observed population for research purposes. In clinics, inter-patient registration can be

useful for prognosis or treatment planning. It enables to align patient data to reference images with known information such as segmentation labels, anatomical landmarks, or certain pathological characteristics. In particular, aligning patient scans to atlas images with annotated anatomical structures provides the possibility of atlas-based image segmentation. In Chapters 3, 4, and 6, monomodal inter-patient data is used for the reported experiments.

Registration of intra-patient multimodal medical image data (like Fig. 2.1 (c)) is important for comparison of image data across imaging modalities and for multimodal image fusion. Multimodal image fusion aims to align image data acquired by different imaging modalities such as computed tomography (CT), magnetic resonance imaging (MRI), positron emission tomography (PET), or ultrasound (US). It enables the combined analysis and visualisation of multimodal image data, which is beneficial for diagnosis, treatment planning, and image-guided surgery. This is useful because each imaging modality has its own benefits and characteristics such as good contrast for certain anatomical structures, visualisation of physiological functions, handiness during surgery, slice orientations, fields of view, or image resolution properties. As an example, soft tissue of organs or tumours are particularly well displayed in MR data, while CT data entails good contrast for bone visualisation. In Chapter 5, an image registration approach for rigid alignment of multimodal intra-patient image data is presented. The method presented in Chapter 6 can also be used for deformable multimodal image registration.

2.2 Fundamentals of Image Registration

Pairwise image registration denotes the process of finding spatial correspondences between a pair of images consisting of a fixed image I_F and a moving image I_M. In this thesis, two-dimensional (2D) and three-dimensional (3D) image data is considered. The two images, I_F and I_M, are aligned by applying a spatial transformation φ to I_M that solves the optimisation problem, which is typically described as minimisation of an objective function $\mathcal{E}$ and aims to find the optimal transformation

$$\hat{\varphi} = \underset{\varphi \in \mathcal{T}}{\arg\min}\ \mathcal{E}(I_F, I_M, \varphi) \tag{2.1}$$

within the space of possible transformations $\mathcal{T}$. The objective function

$$\mathcal{E}(I_F, I_M, \varphi) = \mathcal{D}(I_F, I_M \circ \varphi) + \lambda \mathcal{R}(\varphi) \tag{2.2}$$

consists of a metric $\mathcal{D}$ and a regulariser $\mathcal{R}$ weighted by the parameter λ. The (dis)similarity metric $\mathcal{D}$ measures how well the fixed image I_F and transformed moving image $I_M(\varphi(\cdot)) = I_M \circ \varphi$ are aligned. As high-dimensional non-parametric image registration is an ill-posed problem, the transformation is regularised by the regulariser $\mathcal{R}$

to enforce plausible deformations with predefined properties such as a certain smoothness or displacement radius. Therefore, a typical image registration approach consists of three parts: a transformation model (see Sec. 2.2.1), an objective function (Sec. 2.2.2), and an optimiser (Sec. 2.2.3) [Sotiras et al., 2013]. For assessment of image registration methods, measures for quantitative evaluation are required (Sec. 2.2.4).

2.2.1 Transformation Models and Image Warping

Transformation models for image registration can be divided into the mathematical categories of linear and non-linear transformation models. Fig. 2.2 gives an overview of different types of transformations.

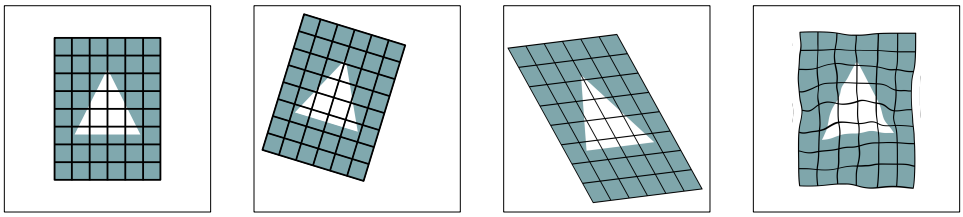


Fig. 2.2: Types of transformations visualised with a 2D example image.

Linear Transformation Models

Linear transformation models, also referred to as global transformation models, are often applied as an initial step in image registration to facilitate the optimisation of subsequent more complex (non-linear) deformable image registration steps [Vos et al., 2019]. A 3D linear transformation $\varphi : \Omega_F \subset \mathbb{R}^3 \mapsto \Omega_M \subset \mathbb{R}^3$ can be expressed in parametric form via the linear mapping

$$\varphi(\boldsymbol{x}) = A\boldsymbol{x}, \quad A = \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ 0 & 0 & 0 & 1 \end{pmatrix}, \tag{2.3}$$

where $\boldsymbol{x}$ denotes the vector of spatial coordinates and the 4×4 transformation matrix A generally includes rotation, translation, shearing, and scaling. A commonly used linear transformation type that preserves the Euclidean distance between every pair of image points is the *rigid* transformation, which describes image rotation and translation. Rigid transformation models entail six degrees of freedom for 3D data (three for rotation and three for translation) and do not change the size or volume of objects. *Affine*

transformations increase the degrees of freedom to twelve, as they additionally include shearing and scaling. Hence, for affine registration the Euclidean distances between image points are not mandatorily preserved, but lines and parallelisms are maintained.

Non-linear Transformation Models

Non-linear transformation models which are also designated as deformable or local transformation models include parametric models, such as free-form deformations based on B-splines with a number of control points that is lower than the number of voxels [Rueckert et al., 1999], and non-parametric models with dense displacement fields. For non-parametric models, the transformation for every position $\boldsymbol{x}$ in the image domain Ω is defined by a deformation field

$$\varphi(\boldsymbol{x}) = \boldsymbol{x} + \boldsymbol{u}(\boldsymbol{x}), \tag{2.4}$$

where $\boldsymbol{u} \in \mathbb{R}^d$ denotes the displacement vector field for d-dimensional image registration. Here, the number of degrees of freedom is defined by the product of image dimensionality and the number of image points.

Image Warping

A transformation can be applied to an image by forward or backward mapping. Forward mapping with $\psi : \Omega_M \mapsto \Omega_F$ and $\boldsymbol{y} \in \Omega_M$, designated as Lagrange approach, maps an intensity value at a position $\boldsymbol{y}$ of a moving image to a new position $\psi(\boldsymbol{y})$. The position $\psi(\boldsymbol{y})$ is then interpolated to regular image grid points of the fixed image, which entails the risk of holes in the resulting warped image. Backward mapping is mostly used for image registration. It prevents holes in the warped image and uses transformations defined on the fixed image domain Ω_F. For backward mapping with $\varphi : \Omega_F \mapsto \Omega_M$ and $\boldsymbol{x} \in \Omega_M$, referred to as Euler approach, the inverse transformation $\varphi = \psi^{-1}$ must exist. The intensity value of a fixed image at the position $\boldsymbol{x}$ is set to the position $\varphi(\boldsymbol{x})$ of the moving image. The moving image is transformed by backward mapping into the coordinate system of the fixed image. When applying a spatial transformation φ, the resulting position normally does not hit a voxel grid position. Therefore, it becomes necessary to use interpolation techniques to determine the intensity values of the warped image. The most commonly used interpolation methods are nearest neighbour interpolation, linear interpolation, and spline interpolation. Learning-based registration methods as described in Sec. 2.3 require differentiable interpolation techniques such as bilinear interpolation [Jaderberg et al., 2015].

2.2.2 Objective Function

The objective function that is optimised during the process of image registration includes a (dis)similarity metric $\mathcal{D}$ and a regulariser $\mathcal{R}$ to be selected as suitable for the underlying registration task. The similarity metrics and regularisation methods relevant to this work are described in the following.

Similarity Metrics

Similarity metrics could be feature-based or intensity-based, and their choice often depends on whether monomodal or multimodal image data is aligned. For monomodal image registration, the *sum of squared differences (SSD)*

$$\mathcal{D}_{\mathrm{SSD}}(I_F, I_M, \varphi) = \sum_{\boldsymbol{x}\in\Omega} \Big(I_F(\boldsymbol{x}) - I_M(\varphi(\boldsymbol{x}))\Big)^2 \tag{2.5}$$

is a similarity metric that assumes identity relationship between the intensities of fixed image I_F and moving image I_M in image domain Ω.

The *mean squared error (MSE)*

$$\mathcal{D}_{\mathrm{MSE}}(I_F, I_M, \varphi) = \frac{1}{|\Omega|} \sum_{\boldsymbol{x}\in\Omega} \Big(I_F(\boldsymbol{x}) - I_M(\varphi(\boldsymbol{x}))\Big)^2 \tag{2.6}$$

additionally averages over the number of positions in image domain Ω.

Cross-correlation similarity metrics assume a linear relation between the image intensities, and therefore are more robust to noise and brightness differences than $\mathcal{D}_{\mathrm{SSD}}$. The most commonly used cross-correlation metric is *normalised cross-correlation (NCC)*

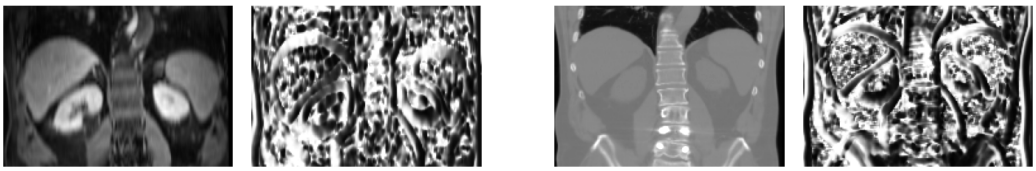
$$\mathcal{D}_{\mathrm{NCC}}(I_F, I_M, \varphi) = \frac{\sum_\Omega I_F(\boldsymbol{x}) \cdot I_M(\varphi(\boldsymbol{x})) - \sum_\Omega \mathrm{E}(I_F(\boldsymbol{x}))\mathrm{E}(I_M(\varphi(\boldsymbol{x})))}{|\Omega| \cdot \sum_\Omega \mathrm{Var}(I_F(\boldsymbol{x}))\mathrm{Var}(I_M(\varphi(\boldsymbol{x})))}, \tag{2.7}$$

where E is the expectation value (or mean value), Var is the variance of the respective image and Ω can be chosen as global image domain or local region.

In multimodal image data, corresponding structures may have different intensity values. Therefore, it is important to choose similarity metrics for multimodal registration that can deal with different image intensities and contrasts. *Mutual information (MI)*

$$\mathcal{D}_{\mathrm{MI}}(I_F, I_M, \varphi) = H\Big(I_F(\boldsymbol{x})\Big) + H\Big(I_M(\varphi(\boldsymbol{x}))\Big) - H\Big(I_F(\boldsymbol{x}), I_M(\varphi(\boldsymbol{x}))\Big) \tag{2.8}$$

with H denoting the (joint) entropy is a similarity metric that assumes a statistical relationship between image intensities and expresses how well one image is explained by the other [Maes et al., 1997; Viola et al., 1995].



MR image and a corresponding MIND feature channel

CT image and a corresponding MIND feature channel

Fig. 2.3: Visualisation of original images from different modalities together with the first channel of their corresponding feature maps generated using MIND with SSC.

The *modality independent neighbourhood descriptor (MIND)* proposed in [Heinrich et al., 2012] is based on the concept of image self-similarities. It extracts distinctive structures in local neighbourhoods, which are preserved across modalities. This leads to similar descriptors for the same geometric structures captured with different imaging modalities. MIND for an image I is defined by

$$\mathrm{MIND}(I, \boldsymbol{x}, \boldsymbol{r}) = \frac{1}{n} \exp\Big(-\frac{\mathcal{D}_p(I, \boldsymbol{x}, \boldsymbol{x}+\boldsymbol{r})}{\mathrm{Var}(I, \boldsymbol{x})}\Big) \tag{2.9}$$

with a patch-based distance $\mathcal{D}_p$ based on SSD, a variance estimate Var, a spatial search region R with positions $\boldsymbol{r} \in R$, and a normalisation constant n which sets the maximum value to 1. An image is represented by MIND as vectors with a vector size of $|R|$ at each image position $\boldsymbol{x}$. As similarity metric for image registration

$$\mathcal{D}_{\mathrm{MIND}}(I_F, I_M, \varphi, \boldsymbol{r}) = \mathcal{D}_{\mathrm{SSD}}\Big(\mathrm{MIND}(I_F, \boldsymbol{x}, \boldsymbol{r}), \mathrm{MIND}(I_M \circ \varphi, \boldsymbol{x}, \boldsymbol{r}), \varphi\Big) \tag{2.10}$$

based on SSD dissimilarity can be used. In this thesis, MIND is used in combination with self-similarity context [Heinrich et al., 2013], yielding a robust feature descriptor for multimodal image data. The MIND-SSC descriptor extracts similarity among neighbouring patches around the voxel of interest without taking the centre patch into account. It is therefore capable of finding a good representation for the context within its neighbourhood. A visualisation of original images and corresponding MIND feature channels is given in Fig. 2.3.

Regularisation

Regularisation is required to achieve plausible deformation fields. It constrains a non-parametric transformation model with a large number of degrees of freedom in order to obtain a well-posed optimisation problem. Constraining spatial derivatives is a commonly used regularisation that operates on the deformation field φ to induce

deformation smoothness. *Diffusion regularisation*, which is often used in this thesis for regularisation, is defined by

$$\mathcal{R}_{\text{diffusion}}(\varphi) = \frac{1}{|\Omega|} \sum_{\boldsymbol{x} \in \Omega} \sum_{i=1}^{d} \|\nabla \varphi_i(\boldsymbol{x})\|_2^2 \tag{2.11}$$

and penalises changes in the d-dimensional deformation field by taking the first derivatives into account. It is also possible to use total variation or bending energy regularisation, which operates on the second derivatives of the deformation field. Curvature regularisation can be used if invariance to linear transformation such as rotation and scaling is of importance. To penalise volume changes and thus to induce topology preservation, regularisation based on the Jacobian matrix can be applied. For parametric image registration, regularisation that directly operates on the utilised parameters is possible. Using diffeomorphic transformations also ensures smooth deformation fields and topology preservation, as diffeomorphic transformations are invertible and differentiable and both the function and its inverse are differentiable.

2.2.3 Optimiser

The optimisers used for image registration aim to find the transformation that best aligns the input images according to the objective function. Optimisation methods could be generally classified with regard to the underlying search space, leading to the category of discrete optimisers and the category of continuous optimisers. Discrete optimisation methods comprise graph-based methods, message passing methods, and Linear-Programming (LP) approaches. They require discretisation of the image data, as well as a discretised space of transformations and discrete objective functions. Continuous optimisation methods are used for registration problems with differentiable objective functions and where the variables take real values. Continuous image registration methods include gradient descent, conjugate gradient, Powell's conjugate directions, Quasi-Newton, Levenberg-Marquardt, and stochastic gradient descent. [Klein et al., 2007; Sotiras et al., 2013]

Of particular importance for this work is *gradient descent*-based optimisation. Gradient descent can be used to minimise an objective function $\mathcal{E}(\theta)$ parametrised by the parameters $\theta \in \mathbb{R}^d$. It takes steps in the direction of the negative gradient of the objective function by following the update rule

$$\theta_{t+1} = \theta_t - \alpha_t \nabla_{\theta_t} \mathcal{E}(\theta_t) \tag{2.12}$$

with step length $\alpha \in \mathbb{R}^+$ and iteration index t. For *stochastic gradient descent*, the gradient is replaced by an estimate calculated from a randomly selected subset of the data. *Adam* [Kingma et al., 2014] is a modification of stochastic gradient descent that updates exponential moving averages of the gradient and the squared gradient.

2.2.4 Evaluation Metrics

Quantitative evaluation of the performance of a registration method requires that ground truth deformation fields or surrogate measures are available. Due to the inherent ambiguity in homogeneous regions of real-life deformation fields, evaluation of image registration methods based on ground truth is only possible for synthetic deformations.

If anatomical landmarks or corresponding keypoints are specified for the input images, the *target registration error (TRE)*

$$\mathrm{TRE}(L_F, L_M, \varphi) = \frac{1}{N} \sum_{k=1}^{N} \| L_F^k(\boldsymbol{x}) - L_M^k(\varphi(\boldsymbol{x})) \|_2 \tag{2.13}$$

with N landmarks L_F and L_M s for I_F and I_M can be calculated.

In case segmentation label data is provided, registration accuracy can be approximated with the *Dice coefficient* as a measure for segmentation overlap of fixed and warped moving image. With S_F and S_M as segmentations for I_F and I_M with N label classes, the Dice coefficient is defined as

$$\mathrm{Dice}(S_F, S_M, \varphi) = \frac{1}{N} \sum_{k=1}^{N} \frac{2 \sum_{\boldsymbol{x} \in \Omega} S_F^k(\boldsymbol{x}) S_M^k(\varphi(\boldsymbol{x}))}{\sum_{\boldsymbol{x} \in \Omega} \Big(S_F^k(\boldsymbol{x}) \Big)^2 + \sum_{\boldsymbol{x} \in \Omega} \Big(S_M^k(\varphi(\boldsymbol{x})) \Big)^2}. \tag{2.14}$$

However, as pointed out in [Rohlfing, 2011], the validity of evaluation using tissue overlap scores such as the Dice coefficient is affected by the near-monotonic relationship between image intensities and tissue labels.

Another measure that can be calculated if segmentation labels are available is the *Hausdorff distance (HD)* as a metric for surface distance. It uses contour labels and computes the maximum distance between any point on one contour and its nearest point on the second contour. For evaluation of image registration methods, it can be expressed as

$$\mathrm{HD}(C_F, C_M, \varphi) = \max\Big(\mathrm{d_H}(C_F, C_M), \mathrm{d_H}(C_M(\varphi(\boldsymbol{x})), C_F) \Big) \tag{2.15}$$

with C_F and C_M being the contour labels on the fixed and the moving image and

$$\mathrm{d_H}(C_F, C_M) = \max_{a \in C_F} \Big(\min_{b \in C_M} \|a - b\| \Big) \tag{2.16}$$

measuring the surface distance. As a robust score, HD95 can be employed that considers the 95th percentile instead of the maximum distance.

To evaluate the topology of deformation fields, the *Jacobian determinant*

$$\det\Big(J(i,j,k)\Big) = \begin{vmatrix} \frac{\partial i}{\partial x} & \frac{\partial j}{\partial x} & \frac{\partial k}{\partial x} \\ \frac{\partial i}{\partial y} & \frac{\partial j}{\partial y} & \frac{\partial k}{\partial y} \\ \frac{\partial i}{\partial z} & \frac{\partial j}{\partial z} & \frac{\partial k}{\partial z} \end{vmatrix} \tag{2.17}$$

can be calculated for every coordinate (i, j, k) in the deformation field of 3D image data by computation of the first-order partial derivatives with respect to the image dimensions (x, y, z). The values of the Jacobian determinant indicate the presence of volume changes. A value higher than 1 indicates volume expansion, a value between 0 and 1 indicates shrinkage, and a value of ≤ 0 indicates a singularity which means that a folding has occurred. Plausible deformation fields and smoothness are thus indicated by a small number or no values below 0 and a small standard deviation of the Jacobian determinant.

Another way to evaluate registration performance is based on the *inverse consistency* of a registration method [Christensen et al., 2001, 2006]. Inverse consistency means that the obtained results are consistent in terms of forward and reverse (backward) registration direction and addresses the problem that image registration is often only considered as an asymmetric problem with a specific registration direction. The inverse consistency metric as defined in [Christensen et al., 2006] measures the voxel-wise inverse consistency error

$$\mathrm{IC}(\varphi) = \mathcal{D}_{\mathrm{SSD}}\Big(\varphi^{-1}(\varphi(\boldsymbol{x})), \mathrm{Id}(\boldsymbol{x}), \varphi\Big) \tag{2.18}$$

between a forward and reverse transformation for an image pair by means of the squared differences. Ideally, the composition of these transformations should yield identity mapping Id.

2.3 Deep Learning-based Image Registration

Deep learning-based image registration models learn a general representation of image registration from a large dataset in order to align unseen image data. In contrast to conventional non-learning-based image registration, optimisation is performed over many training samples instead of over a single image pair. Once trained, a deep learning-based image registration model is able to align image data in a runtime that is generally much faster than registration with conventional methods.

Deep learning methods use deep neural networks (DNNs) which include a large number of network weights that are modified during the training process. For image

registration, a trained network f_ω parametrised by the weights ω encodes the spatial transformation φ to align input data. During inference

$$\varphi = f_\omega(I_F, I_M) \tag{2.19}$$

is evaluated for pairwise image registration. The network is trained with respect to a loss function denoting the objective function that is minimised to receive the optimal network parameters.

2.3.1 Fundamentals of Convolutional Neural Network Architectures

The most prominent type of artificial deep neural networks are convolutional networks (CNNs) [LeCun et al., 1989] as they have proven to be particularly advantageous for solving various tasks. All deep learning models for image registration implemented for this thesis are of this type. The following is a brief outline of the basics for CNN architectures. An extensive introduction to this field can be found in [Goodfellow et al., 2016].

As implied by their name, CNNs use convolution operations to extract feature maps instead of plain matrix multiplication. Conventional (non-convolutional) neural networks use multiplications with a matrix of separate parameters to describe the relationships between all input and output units. This results in every input unit being connected to every output unit. CNNs, however, typically use sparse connections, which is achieved by choosing the convolution kernel to be smaller than the input. This means fewer parameters and a smaller number of operations to compute the output. Another characteristic of CNNs is parameter sharing. A convolutional layer only requires one parameter per element of a filter kernel. This ensures translation invariance.

The typical layer of a CNN can be divided into three parts. First, several convolution operations are carried out in parallel. This is followed by the application of non-linear activation functions, e.g. the rectified linear unit (ReLU) activation. Afterwards, pooling functions are used. A pooling function replaces the output at a certain position with a statistical summary of the surrounding outputs, which increases the receptive field and reduces the resolution of the feature maps. Another type of downsampling is the use of strides, where the spacing of the sampled positions in each output direction is increased.

For image processing, encoder-decoder network architectures are very common. They consist of an encoder which embeds image information in a low-dimensional feature representation called latent space and a decoder which decodes the latent space information. The encoder typically compresses the image information with sequences of convolution and pooling steps. By this, the receptive field is successively increased, which enables the network to not only learn local features, but also global context and semantic information for larger image regions. The decoder typically consists of

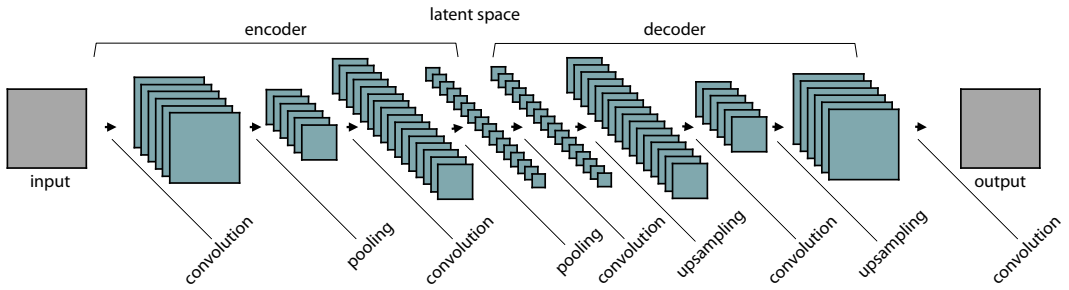


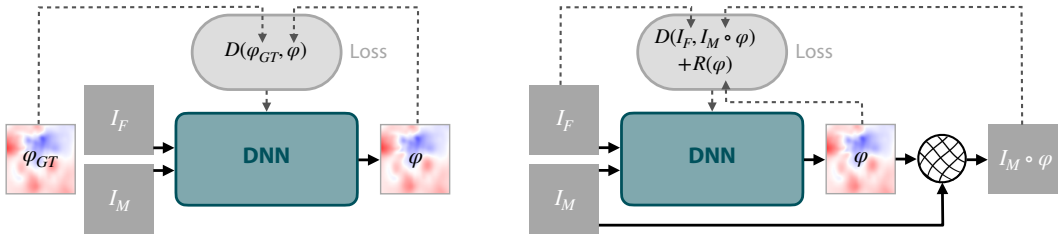
Fig. 2.4: Example for a typical structure of an encoder-decoder CNN architecture with convolution, pooling, and upsampling operations.

sequences of convolution and upsampling steps and aims to recover spatial information. An example for a typical structure of a convolutional encoder-decoder network is given in Fig. 2.4. The main idea for using encoder-decoder architectures is that by embedding the input in a small intermediate feature representation, the network has to learn a suitable representation of the input data to ensure that the decoder is able to reconstruct it correctly.

An encoder-decoder architecture that has gained particular prominence is the *U-Net* architecture [Ronneberger et al., 2015]. A characteristic of this architecture, originally introduced for image segmentation, is that it consists of a contracting part (encoder) for capturing contextual information and an expanding part (decoder) for precise localisation. The architecture is symmetric and doubles the number of feature maps in every encoder layer and halves the number likewise in every decoder layer. Skip connections from the encoder to the decoder at corresponding feature map resolution levels enable to preserve local information within the decoding part. Numerous works on image registration, e.g. [Balakrishnan et al., 2018], adopt the U-Net architecture for estimating deformation fields, some of which use multi-level or multi-warping procedures [Eppenhof et al., 2020; Hering et al., 2019; Mok et al., 2020a] (see Sec. 2.3.3 and Sec. 4.1.1 in Chapter 4 for more details).

2.3.2 Training Supervision for Registration Learning

Depending on the selected loss function, the spatial transformation for image registration is learned either in supervised, unsupervised, or weakly supervised fashion. Fig. 2.5 and the following subsections give an overview of learning procedures for image registration categorised by these different forms of supervision.



(a) supervised (b) unsupervised

Fig. 2.5: Supervised and unsupervised end-to-end trainable image registration models. Supervised models (a) optimise a loss function that measures the dissimilarity $\mathcal{D}(\varphi_{\text{GT}}, \varphi)$ between the transformation φ estimated by a DNN and the ground truth transformation φ_{GT}. Unsupervised models (b) do not use ground truth deformation fields and optimise a loss function that is similar to conventional image registration methods with dissimilarity metric $\mathcal{D}(I_F, I_M \circ \varphi)$ for fixed and warped moving image and regularisation $\mathcal{R}(\varphi)$ of the deformation field.

Supervised Registration Learning

Supervised approaches (see Fig. 2.5 (a)) use ground truth deformation fields φ_{GT} that are either produced synthetically or generated by conventional registration approaches. During training, they optimise a loss function

$$\mathcal{L} = \mathcal{D}_{\text{def}}(\varphi_{\text{GT}}, \varphi) \tag{2.20}$$

that measures the dissimilarity between the transformation generated by a DNN and the provided ground truth transformation.

The concept of using deep learning methods for deformation estimation with ground truth transformation supervision has originated in the field of computer vision for optical flow computation in non-medical datasets. *FlowNet* [Dosovitskiy et al., 2015] is an end-to-end supervised trainable CNN architecture that estimates deformation fields between pairs of input images from a large synthetically generated dataset. To generate ground truth data, the authors randomly applied affine transformations to the foreground and the background of the image data. In Sec. 2.3.3, more details about the architecture of FlowNet and its different versions is given.

For medical image data, this has been done in a similar fashion to learn deformable image registration with ground truth deformation supervision. To generate synthetic training data, several approaches have been proposed, like random transformations based on Gaussian kernels [Sokooti et al., 2017], random B-Spline transformations [Eppenhof et al., 2020], or locality-based shape and appearance models [Uzunova et al., 2017]. Other supervised registration learning approaches use existing non-learning-

based algorithms for the generation of ground truth deformation fields [Cao et al., 2017; Rohé et al., 2017; Yang et al., 2017].

The limitations of supervised registration learning are mainly due to the ground truth deformation fields that are required in sufficient amount to train a deep learning model. If synthetically generated ground truth data is used, it is difficult to generate ground truth deformations in such a way that they are realistic, especially for medical image data and dense deformable deformation learning. In case of ground truth data computed by conventional image registration algorithms, the performance of the underlying algorithm can be a limiting factor.

Unsupervised and Weakly Supervised Registration Learning

Unsupervised (see Fig. 2.5 (b)) and weakly supervised approaches do not use ground truth deformation fields for registration learning. The loss function

$$\mathcal{L} = \mathcal{D}_{\text{img}}(I_F, I_M \circ \varphi) + \mathcal{R}(\varphi) \tag{2.21}$$

that is minimised during training of unsupervised registration models resembles objective functions of conventional registration methods with usually a dissimilarity metric $\mathcal{D}_{\text{img}}(I_F, I_M \circ \varphi)$ for fixed and warped moving image and regularisation $\mathcal{R}(\varphi)$ of the deformation field. The precondition for this type of learning is a differentiable warping functionality as introduced in [Jaderberg et al., 2015] that allows for end-to-end network training (see Sec. 2.3.3 for details).

Most of the weakly supervised registration methods extend the loss function formulated for unsupervised methods by a metric $\mathcal{D}_{\text{seg}}(S_F, S_M \circ \varphi)$ that measures the dissimilarity between segmentations for the fixed input image and warped segmentations for the moving input image. This leads to

$$\mathcal{L} = \mathcal{D}_{\text{img}}(I_F, I_M \circ \varphi) + \mathcal{R}(\varphi) + \mathcal{D}_{\text{seg}}(S_F, S_M \circ \varphi) \tag{2.22}$$

and requires that segmentation labels are provided for the training data. If available, other auxiliary information on the training images such as keypoints or landmarks may be used instead of or in addition to segmentation information for weak supervision. It is also possible that a method only learns with help of the auxiliary information and omits image similarity during training. However, it is important to be aware of the potential problem of label bias for deformable image registration, especially if labels with large volumes or few labelled structures are employed.

2.3.3 Popular Methods in the Context of Image Registration

This section describes four pioneering and popular methods in the field of deep learning-based image registration: FlowNet [Dosovitskiy et al., 2015] is one of the first supervised

end-to-end trainable CNN architectures for optical flow estimation. In [Jaderberg et al., 2015], the spatial transformer module is introduced, which provides a differentiable warping functionality that is of importance for unsupervised registration learning. VoxelMorph [Balakrishnan et al., 2019] is one of the first popular deep learning-based medical image registration methods and makes use of differentiable warping for unsupervised registration learning. LapIRN [Mok et al., 2020b] is a more recently introduced CNN architecture that shows very promising registration performance with a coarse-to-fine multi-resolution registration learning approach.

FlowNet

In [Dosovitskiy et al., 2015], *FlowNet1.0* is introduced for optical flow learning with CNNs. The authors propose two different architectures: *FlowNetSimple (FlowNetS)* and *FlowNetCorr (FlowNetC)*. Both versions contain a contracting part and an expanding part and are end-to-end trainable.

The contracting part of FlowNetS concatenates both input images and only uses convolutional layers to extract the motion information. FlowNetC includes two separate processing streams for the input images and combines them in a correlation layer to find correspondences. Thus, FlowNetS decides itself how to process the input image pair, whereas FlowNetC is constrained to first generate meaningful representations of both images separately and then combine them at a deeper level.

The correlation layer of FlowNetC performs multiplicative patch comparisons between two feature maps $\boldsymbol{f}_1$ and $\boldsymbol{f}_2$ with C, H, W being their number of channels, height, and width. The correlation of two patches from $\boldsymbol{f}_1$ centred at $\boldsymbol{x}_1$ and $\boldsymbol{f}_2$ centred at $\boldsymbol{x}_2$ is defined as

$$\text{correlation}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sum_{\boldsymbol{o} \in [-k,k] \times [-k,k]} \langle \boldsymbol{f}_1(\boldsymbol{x}_1 + \boldsymbol{o}), \boldsymbol{f}_2(\boldsymbol{x}_2 + \boldsymbol{o}) \rangle \tag{2.23}$$

for square patches of size $K := 2k + 1$. For computational reasons, the maximum displacement for each position $\boldsymbol{x}_1$ is limited to d and correlations are only computed in a neighbourhood of size $D := 2d + 1$ by limiting the range of $\boldsymbol{x}_2$. The result of the correlation is organised in such a way that the relative displacements are given as channels yielding an output of size $H \times W \times D^2$. Then, the feature map, which is extracted from $\boldsymbol{f}_1$ with a convolutional layer, is concatenated with the output. Afterwards, the result is further processed by the contracting and the expanding network part.

The expanding part is used for both, FlowNetS and FlowNetC, to refine the estimated flow to high resolution with the help of up-convolutions (upsampling of feature maps followed by convolution). Refinement is performed by applying up-convolutions to feature maps and concatenating the output with corresponding feature maps from the contractive parts as well as, if available, an upsampled coarser flow prediction. This

preserves high-level information from coarser feature maps and fine local information from lower layer feature maps.

FlowNet is trained in a supervised manner using the Euclidean distance between the optical flow predicted by the network and synthetic ground truth data. Ground truth data is generated by randomly applying affine transformations to the foreground and the background of image data.

In [Ilg et al., 2017], an extension of FlowNet, *FlowNet2.0*, is presented with a learning schedule and stacking of multiple FlowNets to learn large displacements.

Spatial Transformer Networks

The *spatial transformer module* introduced in [Jaderberg et al., 2015] allows the spatial manipulation of data within a neural network. It is a differentiable module which can be integrated in a CNN architecture and allows for end-to-end training. Spatial Transformer Networks (STN) learn how to spatially transform an input image or feature map in order to enhance the geometric invariance of a deep learning model. The spatial transformer module predicts transformation parameters and uses them to create a sampling grid which consists of a set of points where the input feature map is then sampled to generate the warped output. It is composed of three parts: a localisation network, a grid generator, and a sampler.

The localisation network f_{loc} is a regular CNN which regresses the transformation parameters. It takes an input feature map $U \in \mathbb{R}^{C \times H \times W}$ with C channels, the height H, and the width W, and outputs $\theta = f_{loc}(U)$, the parameters of the transformation $\mathcal{T}_\theta$ to be applied on the input feature map. The transformation $\mathcal{T}_\theta$ can be of any parametrised form, given that it is differentiable with respect to the parameters.

The grid generator generates a grid of coordinates in the input feature map U corresponding to each pixel from the output feature map $V \in \mathbb{R}^{C \times H' \times W'}$. The output pixels are defined to lie on a regular grid $G = \{G_i\}$ of elements $G_i = (x_i^t, y_i^t)$, where (x_i^t, y_i^t) are the target coordinates. The pointwise transformation is then defined by $(x_i^s, y_i^s)^T = \mathcal{T}_\theta(G_i)$ with (x_i^s, y_i^s) being the source coordinates in the input feature map that define the sample points.

The sampler uses the input feature map U and the sampling points $\mathcal{T}_\theta(G)$ and produces the output feature map V. Given a generic sampling kernel k which defines the image interpolation (e.g. bilinear) and Φ_x and Φ_y as kernel parameters, the values at a particular position in the output feature map V are provided by

$$V_i^c = \sum_n^H \sum_m^W U_{nm}^c k(x_i^s - m; \Phi_x) k(y_i^s - n; \Phi_y) \quad \forall i \in [1, ..., H'W'] \quad \forall c \in [1, ..., C] \quad (2.24)$$

where U_{nm}^c is the value at location (n, m) in channel c of the input and V_i^c is the output value for pixel i at location (x_i^t, y_i^t) in channel c.

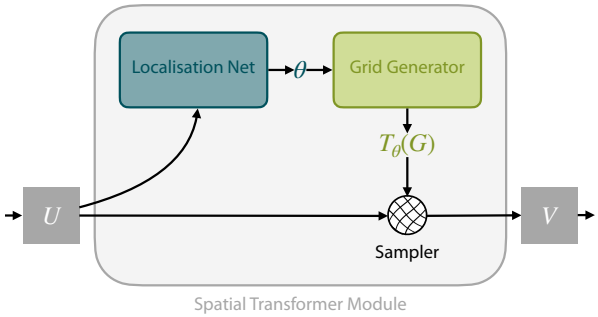


Fig. 2.6: The Spatial Transformer Module [Jaderberg et al., 2015] is composed of three parts: a localisation network, a grid generator, and a sampler. It predicts transformation parameters and uses them to create a sampling grid which consists of a set of points where the input feature map is then sampled to generate the warped output.

The combination of the localisation network, the grid generator, and the sampler forms the spatial transformer module as visualised in Fig. 2.6. It can be integrated into a CNN architecture at any point and in any number, and allows the network to learn how to transform the feature maps for a minimised overall objective function. In image registration networks, the module can be used to warp a moving image to align it with a fixed image.

VoxelMorph

One of the first popular and promising deep learning-based end-to-end trainable medical image registration methods is *VoxelMorph* [Balakrishnan et al., 2018, 2019]. The framework is used for deformable image registration and its performance is comparable to conventional non-learning-based image registration methods while being significantly faster. It uses a fully convolutional architecture with a spatial transformer module to estimate deformation fields directly from input image pairs. In Fig. 2.7, an overview of VoxelMorph is visualised.

The network architecture is similar to U-Net [Ronneberger et al., 2015] and consists of an encoder-decoder structure with skip connections. The input given to the network is formed by concatenating a moving and a fixed image and processed by 3D convolutions with a kernel size of $3\times3\times3$ and leaky ReLU activations. To reduce the spatial dimension in the encoder, strided convolutions are used until the layer with the smallest resolution is reached. The authors remark that the receptive fields of the convolutional kernels of the smallest resolution layer should be at least as large as the maximum expected displacement length. In the decoder part, upsampling operations and concatenating skip connections are employed.

The network is trained to model the deformation field φ that minimises the differences between fixed image I_F and warped moving image $I_M(\varphi(\boldsymbol{x}))$ computed with a

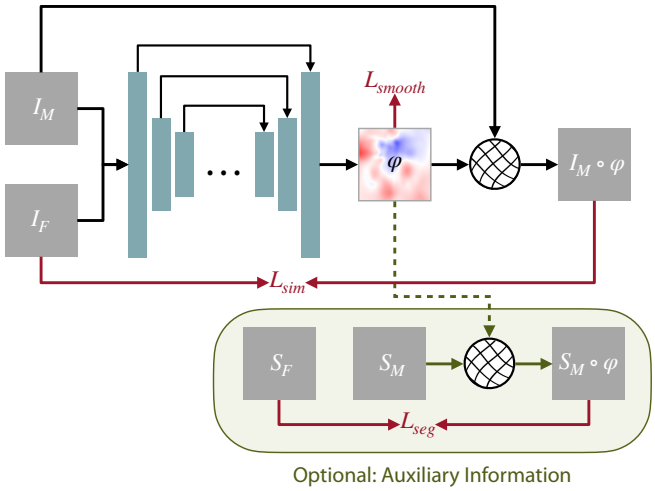


Fig. 2.7: VoxelMorph [Balakrishnan et al., 2019] concatenates an input image pair (I_F, I_M) that is processed by a U-Net architecture to predict a deformation field φ. The network is trained in unsupervised way based on an image similarity loss term $\mathcal{L}_{\text{sim}}$ and a smoothness loss term $\mathcal{L}_{\text{smooth}}$. Alternatively, a weakly supervised training with auxiliary information such as segmentations (S_F, S_M) and additional segmentation similarity loss term $\mathcal{L}_{\text{seg}}$ is proposed.

differentiable spatial transformer module [Jaderberg et al., 2015]. The authors propose an unsupervised training procedure with a loss function

$$\mathcal{L}(I_F, I_M, \varphi) = \mathcal{L}_{\text{sim}}(I_F, I_M \circ \varphi) + \lambda \mathcal{L}_{\text{smooth}}(\varphi) \tag{2.25}$$

including a loss term $\mathcal{L}_{\text{sim}}$ based on image similarities measured by the mean squared error function or local cross-correlations, combined with a loss term $\mathcal{L}_{\text{smooth}}$ for diffusion regularisation weighted by the parameter λ. Moreover, they propose a weakly supervised training procedure that additionally includes a loss term $\mathcal{L}_{\text{seg}}$ measuring the Dice similarity of anatomical segmentation labels.

The concept of VoxelMorph is taken up and modified in Chapter 4. Chapter 4 also gives an overview of several other deep learning-based image registration methods related to VoxelMorph.

Laplacian Pyramid Image Registration Network

In [Mok et al., 2020b], the *Laplacian Pyramid Image Registration Network (LapIRN)* is presented. It is a deep learning-based image registration method that shows the potential to outperform conventional methods in terms of registration accuracy and speed as demonstrated in [Hering et al., 2022a]. LapIRN performs image registration in a coarse-to-fine multiresolution fashion within a diffeomorphic setting.

Therefore, an input image pyramid is created by downsampling the input images I_F and I_M. CNN-based registration networks with an encoder part, residual blocks, and a decoder part are used to solve the optimisation problem for each pyramid level. At the first pyramid level, the non-linear misalignment from the concatenated input images I_F^1 and I_M^1 with the coarsest resolution are captured and a dense vector field $\boldsymbol{v}_1$ and deformation field φ_1 are generated. The following pyramid levels $i > 1$ use the upsampled output deformation field from the previous pyramid level $\hat{\varphi}_{i-1}$ to warp the moving image I_M^i with the resolution corresponding to the current pyramid level, yielding $I_M^i(\hat{\varphi}_{i-1})$. Additionally, the output velocity field from the previous level is upsampled to $\hat{\boldsymbol{v}}_{i-1}$ and concatenated with the input scans to form a 5-channel input for the network at level i. The output velocity field from level i is added to $\hat{\boldsymbol{v}}_{i-1}$ to obtain the velocity field $\boldsymbol{v}_i$ that is integrated to produce the final deformation field φ_i. A skip connection from a lower to the next higher pyramid level increases the receptive field and a skip connection between encoding and decoding part within a pyramid layer maintains low-level features. Diffeomorphism is obtained by applying scaling and squaring to the velocity fields.

For model training, the authors propose to first train the network for the coarsest resolution level alone. Then, the networks of the subsequent levels are progressively added while trainable network parameters of the pretrained networks are frozen for a certain number of iterations whenever a new network is added to the training. As the similarity metric for network training, a similarity pyramid

$$\mathcal{S}^K(I_F, I_M) = \sum_{i=1,\dots,K} -\frac{1}{2^{(K-i)}} \mathrm{NCC}_w(I_F^i, I_M^i) \tag{2.26}$$

is proposed to capture the similarity in a multi-resolution fashion. Here, $\mathcal{S}^K(\cdot,\cdot)$ denotes the similarity pyramid with K levels and NCC_w the local normalised cross correlation with window size w^3 and $w = 1 + 2i$. To avoid the similarity from lower resolution to dominate, a lower weight is assigned. Within the coarser resolution levels, the similarity metric is smoother and less sensitive to noise. Therefore, integrating the similarity metric from a lower level helps to prevent local minima during optimisation of higher resolution levels. The loss function is formulated as

$$\mathcal{L}_p(I_F, I_M \circ \varphi, \boldsymbol{v}) = \mathcal{S}^p(I_F, I_M(\varphi)) + \frac{\lambda}{2^{(L-p)}} \|\nabla \boldsymbol{v}\|_2^2 \tag{2.27}$$

with $p \in [1, ..., L]$ as current pyramid level. The second term is the smoothness regularisation on the velocity field $\boldsymbol{v}$ with the regularisation parameter λ.

In [Mok et al., 2021], a conditional deformable image registration method is proposed that extends LapIRN. The authors replace the residual blocks within the registration network with conditional image registration modules. These modules comprise hidden layers that are directly conditioned on the smoothness regularisation hyperpa-

rameter. Learning of conditional features that are correlated with the regularisation hyperparameter yields a single CNN that can capture optimal solutions with arbitrary hyperparameters.

Chapter 3

Deforming Autoencoders for Groupwise Registration

This chapter introduces a method for groupwise registration of medical image data published in [Siebert et al., 2020] and [Siebert et al., 2021c]. The proposed method performs image registration by using deformable templates in a framework that decouples image shape and image appearance with an autoencoder architecture. Here, it is applied for registration of two- and three-dimensional inter-patient brain MR scans and an inverse consistency constraint is added to an unsupervised learning procedure.

3.1 Introduction

With advancing technological progress, the amount of acquired image data in clinical routine is increasing more and more. This entails a demand for methods to process large image datasets in order to make use of them for research or clinical purposes. Groupwise image registration facilitates the analysis of large medical image datasets, as they can be used to align multiple images within a single optimisation procedure while taking into account all available image information. Methods performing groupwise image registration are applied for the creation of anatomical atlases, the building of shape models, motion tracking along time series, and population studies [Geng et al., 2009]. Especially for the processing of large image datasets, techniques are often required that enable fast registration. Therefore, deep learning-based methods can be particularly helpful.

For groupwise registration or the analysis of anatomical shape variation, often deformable templates or atlases are used that aim to represent the variability of an image dataset. It is essential to employ meaningful appearance and deformation models, which represent the underlying variability of a medical dataset appropriately. With the help of deformable templates, shape can be delineated as a geometric deformation between a template and an input image. Therefore, most template-based groupwise registration methods consist of several pairwise image registration steps between a certain template as reference image and input images that are aligned to the template image. The templates used for groupwise image registration in medical context could

either be generated or selected from a given image dataset. Non-learning-based methods often either choose one 'typical' of the observed images as a template, compute an average image, or use a statistical formulation of template generation, the last two of which often use elaborate iterative algorithms [Allassonnière et al., 2007]. Recently, deep learning-based models have been used to generate deformable templates as will be described in the following Sec. 3.1.1. Due to anatomical differences across a population, e.g. the topological variations in brain anatomy with differing numbers of cortical folds, there is no perfect template that fits all subjects. For this reason, a compromise between the sharpness of the atlas and the variability with respect to anatomical variants has to be found.

3.1.1 Related Work

Groupwise registration methods have been established in recent times as being faster and more accurate at aligning groups of images than pairwise registration methods for the same registration tasks [Che et al., 2019a; Che et al., 2019b; Haase et al., 2020]. In [Che et al., 2019a], deep learning is applied to simultaneously align groups of multispectral images. They use an unbiased template image based on principal component analysis (PCA), which is updated iteratively. A joint optimisation for an unsupervised learning algorithm is employed while using a U-Net architecture [Ronneberger et al., 2015] and measuring the similarity of groups of images with respect to the current template image with mutual information. The authors continue their work in [Che et al., 2019b] aiming for more robustness by using a network architecture that fuses residual units like in ResNet [He et al., 2016] into the encoder part of a U-Net-based architecture. They also combine internal smoothing and external correlation of different deformation fields to guide the correct calculation of deformations.

While the methods of Che et al. create templates based on PCA, there are various approaches that use deep learning techniques to generate templates. In [Kang et al., 2018], adaptive template generation using supervised neural networks is investigated. The authors compare templates generated by a convolutional autoencoder, templates generated by a generative adversarial network, and average templates. The results of their experiments on PET brain scans show the superiority of the deep learning approaches in terms of correlation and bias properties compared to the investigated average templates. In their studies, the anatomical details in the templates generated by the convolutional autoencoder and the generative adversarial network are not significantly different, whereas the noise level is lower in their autoencoder-based template generation approach.

A probabilistic model for template generation is proposed in [Dalca et al., 2019b] that helps to determine conditional templates together with a neural network model for image registration. The authors present a probabilistic spatial deformation model based on diffeomorphisms and provide an end-to-end CNN-based framework to jointly

synthesise templates and generating the deformation field for any new input image. Their model can be used to study the population variation with respect to certain attributes, such as age, sex, and disease state. In Dalca et al.'s work, the proposed approach takes an image and an optional attribute vector, e.g. consisting of age and gender attributes, as input data. A template generation network outputs a conditional template, which is then passed to a second network that outputs a deformation field to align the template to the input image. The unsupervised model is trained with a loss function that minimises the negative maximum likelihood of the data and deformations and controls the prior over the deformation fields encouraging smooth and unbiased deformation properties. Diffeomorphism is induced by *scaling and squaring*.

Another approach to generate age-conditioned templates is published in [Mouches et al., 2021]. The authors use an autoencoder that disentangles age-related and subject-specific variations. The disentanglement is achieved with an invertible latent space disentanglement module to transform input datasets into a low dimensional latent space, where the age-related and age-unrelated image information is separated with a subspace projection method based on [Li et al., 2020].

In general, the latent space of autoencoders can be used to reduce image analysis problems that are complicated in image space to simpler problems in latent space. In the representation learning approach presented in [Hagenah et al., 2019], it is demonstrated that complex shape deformation can be transferred to a simple vector translation in the latent space. The authors use this for mapping from pathological samples to healthy samples by latent space manipulation. This is possible because the applied variational autoencoder is trained to encode disentangled representations for healthy and pathological samples yielding two representation clusters in latent space. Reconstruction can then be performed after translation from an encoded pathological sample towards the cluster of healthy samples.

The idea of disentanglement of image information is also taken up in other works. Various publications disentangle shape and appearance with generative adversarial networks [Chen et al., 2016; Tewari et al., 2022] or variational autoencoders [Higgins et al., 2017; Kumar et al., 2017]. In [Xing et al., 2020], a probabilistic framework is introduced to disentangle appearance and geometric information in an unsupervised manner with two generator networks which are combined by a warping function. A method that goes in the same direction, but in an non-probabilistic way, is the approach of deforming autoencoders presented by [Shu et al., 2018]. This method follows the deformable template paradigm and models image generation through appearance synthesis and a spatial deformation that warps the generated appearance. For this purpose, they introduce a network architecture that breaks the latent code provided by a decoder network into two parts that are fed into separate decoder networks for appearance and deformation estimation. The method is enforced to model shape variability through the deformation decoder by keeping the latent vector for appearance very small. After the spatial warping step, the reconstructed image is fed into the loss

function that uses a reconstruction loss for unsupervised training. Shu et al. report experiments on two-dimensional image data showing digits, hands, and faces and demonstrate that once learned, the latent representations can be used for shape and appearance interpolation between images. Experiments with a related method have been performed by [Bône et al., 2020] on medical image data. In their work, they present metamorphic autoencoders which involve Bayesian generative statistics, metamorphoses-based shape analysis, and variational autoencoding. This approach aims to implicitly learn a single global template, which is an approach that has been adopted in [Uzunova et al., 2021]. In the work of Uzunova et al., a guided filter smoothing for appearance regularisation [He et al., 2012] is included in another method that is related to deforming autoencoders and is successfully applied on brain MR scans.

3.1.2 Contribution

The works of [Dalca et al., 2019b; Kang et al., 2018; Mouches et al., 2021] have focused their experiments on neuroimaging and outlined the benefits of deformable templates for this medical application. In this chapter, a method is proposed that extends and adapts the concept of deforming autoencoders from [Shu et al., 2018] for medical application. The purpose of the method is to decouple shape and appearance in an unsupervised learning setting to represent anatomical variability in a robust and plausible manner. The method is used for groupwise image registration and template generation for 2D and 3D inter-patient brain MR scans. Registration robustness is gained by a loss function that includes an invertibility loss term to achieve backward consistency of image registration.

3.2 Methods

The proposed autoencoder model for image alignment interprets image generation as a composition of *shape* and *appearance*. Here, *shape* depicts geometric information and *appearance* the texture or intensity information. Image registration is thus interpreted as a synthesis of an intensity template (*appearance*) followed by a subsequent deformation (*shape*). The warping step induces the observed shape variability. The idea of the proposed approach is inspired by the architecture of deforming autoencoders presented by [Shu et al., 2018] and the model for learning anatomical templates presented by [Dalca et al., 2019b]. The presented objective function restricts the model to generate good reconstruction results, smooth and realistic deformation fields, and backward consistency of registration. In Fig. 3.1, an overview of the proposed network architecture and the deployed objective function terms is given. In Sec. 3.2.1, the architecture is described in detail and in Sec. 3.2.2, the objective function is explained.

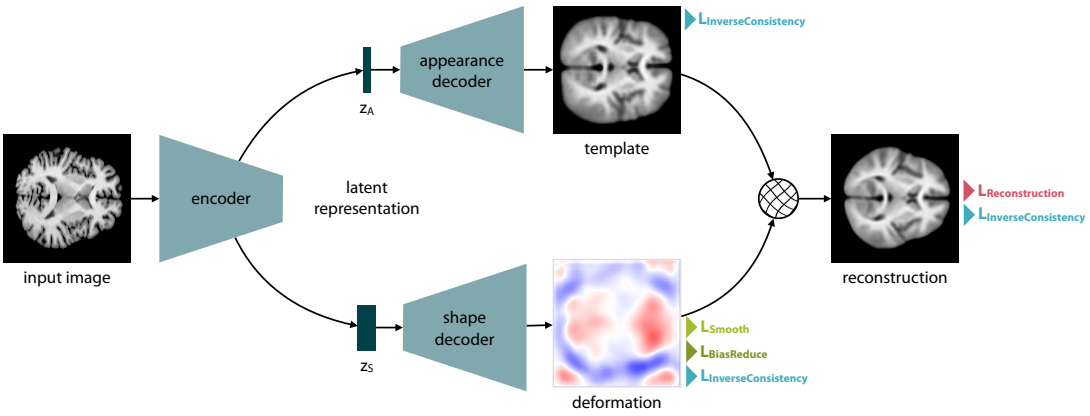


Fig. 3.1: Overview of the proposed autoencoder architecture (here: for 2D input data): A joint encoder network predicts latent representations of appearance and shape, which are fed separately into the appearance and shape decoder networks. Image generation is achieved by spatial warping of the estimated template with the estimated deformation. The applied objective function restricts the model to predict good reconstruction results and realistic deformations with help of four different loss terms.

3.2.1 Network Architecture

The basic design of the proposed autoencoder architecture consists of a joint encoder block that provides representations for both appearance and shape, followed by two separate decoder blocks. A differentiable warping step [Jaderberg et al., 2015] takes the output of the two decoder blocks and generates the reconstructed image. The structure of the architecture for processing 3D image data differs from the architecture for processing 2D image data only in the fact that the corresponding 3D operations are used. In addition, the encoder and decoder blocks of the 3D architecture are one resolution layer less deep in order to keep the number of trainable parameters limited.

The encoder and the two decoder blocks consist of several modules. Every encoder module comprises two sequences of convolution, instance normalisation, and leaky ReLU as activation. Starting from eight feature channels, the number of channels is doubled in every module until a number of 256 (architecture for 2D input data) or 128 (architecture for 3D input data) feature channels is reached. Every encoder module halves the spatial resolution of its input in each image dimension using strided convolutions. The output of the joint encoder is used as input for the two separate decoders, whereby the latent representation, which serves as input for the appearance decoder, is first reduced to a size of $z_A = 1$ using additional convolution layers followed by a linear layer and a sigmoid function as activation. Following the idea of [Shu et al.,

2018], keeping the latent vector for appearance small should enforce the autoencoder to model shape through the shape decoder with a latent vector that is with a size of $z_A = 256$ (2D architecture) or $z_A = 128$ (3D architecture) considerably larger.

The structure of the decoder modules are similar to those of the encoder, only that now the spatial resolution is doubled and the number of feature channels is halved. Except for the last decoder module, transposed convolutions are used for spatial upsampling. The last upsampling step in both of the decoder paths is performed with bilinear (2D architecture) or trilinear (3D architecture) interpolations to avoid checkerboard artifacts. The output of the appearance decoder path is single channel and produces a template image, while the output of the shape decoder path comprises one channel for each image dimension. Here, each channel represents the deformation between the input image and the generated template image in the direction of one image dimension. To obtain the reconstruction, a spatial transformer module [Jaderberg et al., 2015] is used subsequently.

3.2.2 Objective function

The objective function used for training of the proposed autoencoder model is a modification of the loss function used in [Shu et al., 2018]. It restricts the model in such a way that smooth and realistic deformations are ensured, and compared to [Shu et al., 2018] it includes an inverse consistency constraint that is presented in detail in Sec. 3.2.2.1. Altogether, the objective function consists of four terms and can be expressed as

$$\mathcal{L} = \mathcal{L}_{\text{Reconstruction}} + \mathcal{L}_{\text{InverseConsistency}} + \mathcal{L}_{\text{Smooth}} + \mathcal{L}_{\text{BiasReduce}}. \tag{3.1}$$

The smoothing loss term $\mathcal{L}_{\text{Smooth}}$ and bias reducing loss term $\mathcal{L}_{\text{BiasReduce}}$ operate directly on the deformation field output of the shape decoder and the reconstruction loss term $\mathcal{L}_{\text{Reconstruction}}$ operates on the reconstructed image generated by the warping step.

Within the reconstruction loss term

$$\mathcal{L}_{\text{Reconstruction}} = \|I_{\text{recon}} - I_{\text{in}}\|_1 \tag{3.2}$$

image similarity is penalised by means of the ℓ_1 norm between the input image I_{in} and the reconstructed output image I_{recon} [Shu et al., 2018]. The smoothing loss term

$$\mathcal{L}_{\text{Smooth}} = \lambda_{\text{grad}}(\|\nabla W_x(x,y)\|_1 + \|\nabla W_y(x,y)\|_1) \tag{3.3}$$

penalises quick spatial changes in the displacements encoded by the warping field W by using the total variation norm of the warping fields in x- and y-dimension [Shu et al.,

2018] weighted by $\lambda_{\text{grad}} = 1 \times 10^{-4}$. For the 3D application, the term is extended by the z-dimension. With

$$\mathcal{L}_{\text{BiasReduce}} = \lambda_{\text{br}} \|\bar{W} - W_0\|^2 \tag{3.4}$$

and the weighting parameter $\lambda_{\text{br}} = 1 \times 10^{-3}$, the mean squared error between the average deformation grid $\bar{W}$ and an identity mapping grid W_0 is calculated as a further regularisation of the warping field [Shu et al., 2018].

3.2.2.1 Inverse Consistency Constraint

Particularly for image registration in medical applications, robustness is highly relevant. To obtain accurate and plausible deformation fields, several approaches have been introduced in the medical and non-medical fields. In [Zhu et al., 2017], a cycle consistency loss has been introduced for image-to-image translation. For pairwise image registration, the work of [Kim et al., 2019] introduces a cycle-consistency constraint that provides accurate registration for image data with severe deformation in an unsupervised deep learning-based approach. Inverse consistency for image registration has also been addressed before in [Avants et al., 2008].

Registration is often considered as an asymmetric problem with a specific direction. By adding an inverse consistency constraint $\mathcal{L}_{\text{InverseConsistency}}$ the proposed method is made more robust in terms of plausibility of the estimated deformation field. The presented extension of the objective function is novel in this context and operates on the interaction of the prediction for the template image I_{template}, the predicted deformation φ, and the generated reconstructed image I_{recon}. The basic concept is that warping the reconstructed image I_{recon} with the flow field $\varphi_{\text{t}\to\text{r}}$ generated by the shape decoder should ideally result in the template image I_{template} generated by the appearance decoder. Vice versa, warping the template image I_{template} with the reversed flow field $\varphi_{\text{t}\to\text{r}}$ should produce the reconstructed image I_{recon}.

Similar to the invertibility loss used in [Zhao et al., 2019b], any deviation from the reconstructed image is penalised when warping the generated reconstruction with the generated flow field $\varphi_{\text{r}\to\text{t}}$ and then warping the result with the reversed flow field $\varphi_{\text{t}\to\text{r}}$ again. Furthermore, warping the template image generated by the appearance decoder with the reversed flow field $\varphi_{\text{t}\to\text{r}}$ followed by warping with the forwards flow field $\varphi_{\text{r}\to\text{t}}$ should recover the template image. This results in the loss term

$$\mathcal{L}_{\text{InverseConsistency}} = \lambda_{\text{inv}}\Big(\text{MSE}(I^*_{\text{recon}}, I_{\text{recon}}) + \text{MSE}(I^*_{\text{template}}, I_{\text{template}})\Big) \tag{3.5}$$

summing the mean squared errors (MSE) between reconstructed image I_{recon} and recovered reconstruction image $I^*_{\text{recon}} = I_{\text{template}} \circ \varphi_{\text{t}\to\text{r}}$ and between template image I_{template} and recovered template image $I^*_{\text{template}} = I_{\text{recon}} \circ \varphi_{\text{r}\to\text{t}}$. Additional weighting with the paramter $\lambda_{\text{inv}} = 1 \times 10^{-1}$ is applied for this term.

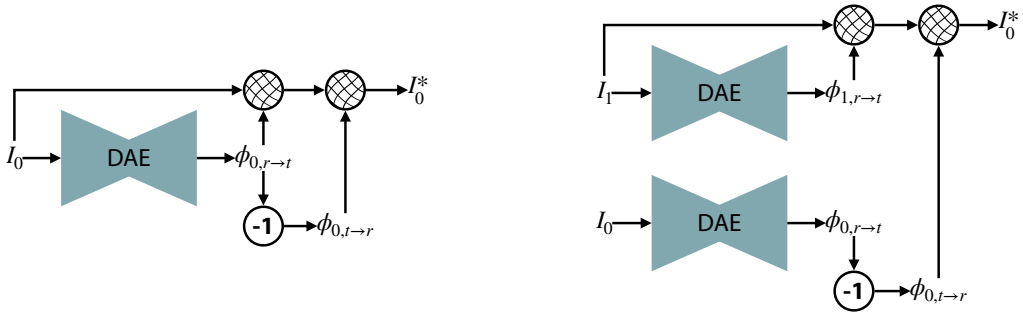
Theoretically, diffeomorphic transformations ensure inverse consistency for image registration problems. Nevertheless, they do not guarantee the symmetry required for cyclic registration to achieve $\mathcal{L}_{\text{InverseConsistency}} = 0$. Therefore, the proposed model is trained including the introduced inverse consistency constraint for deforming autoencoders.

3.3 Experiments and Results

The experiments are performed on a subset of 21 volumes from the MindBoggle101 dataset [Klein et al., 2012] containing labeled T1-weighted brain MRI volumes acquired from healthy persons. For evaluation of the methods, eight different label classes are considered (lateral ventricle, brain stem, hippocampus, cerebellum exterior, cerebellum white matter, thalamus proper, putamen, caudate).

The experiments with 2D input data are performed using 315 middle slices (240 for training, 75 for testing) in coronal orientation, which are resampled to a size of 256×256 voxels. The experiments with 3D input data are performed using 21 volumes (16 for training, 5 for testing) which are resampled to a size of $128 \times 128 \times 128$ voxels. Data augmentation is performed by random affine deformations. The models are trained for $5,000$ epochs using the Adam optimiser and an initial learning rate of 2×10^{-4}.

To investigate the benefits of the inverse consistency constraint, the models are trained with the introduced loss term $\mathcal{L}_{\text{InverseConsistency}}$ and without this loss term. In order to evaluate the pairwise registration performance, in total $10,000$ image pairs with random affine augmentations from the test dataset are randomly chosen and a pairwise registration using the generated warping fields is performed. Additionally, experiments for cyclic registration are performed.



(a) cyclic registration (b) pairwise registration

Fig. 3.2: Experimental setup for image registration using the proposed deforming autoencoder models: (a) cyclic registration of an input image I_0 and (b) pairwise registration with a fixed image I_0 and a moving image I_1.

The experimental setup for cyclic and pairwise image registration is visualised in Fig. 3.2. For cyclic registration, the results display the comparison of the input image and the input image first warped with the forward flow field $\varphi_{0,\mathrm{r}\to\mathrm{t}}$ and then warped with the reversed flow field $\varphi_{0,\mathrm{t}\to\mathrm{r}}$ [Chen et al., 2008]. The results for pairwise registration compare the fixed input image I_0 with the moving image I_1 first warped with $\varphi_{1,\mathrm{r}\to\mathrm{t}}$ followed by warping by $\varphi_{0,\mathrm{t}\to\mathrm{r}}$. Here, $\varphi_{1,\mathrm{r}\to\mathrm{t}}$ is the deformation generated by the model when using the moving image I_1 as input data and $\varphi_{0,\mathrm{t}\to\mathrm{r}}$ is the reversed deformation of $\varphi_{0,\mathrm{r}\to\mathrm{t}}$ generated by the model when using the fixed image I_0 as input data.

Table 3.1: Results for 2D model: Registration and reconstruction results for 2D trained with $\mathcal{L}_{\text{InverseConsistency}}$ (w/ IC) and trained without $\mathcal{L}_{\text{InverseConsistency}}$ (w/o IC). MSE scores are calculated for image intensities scaled between 0 and 255. Dice scores are averaged over all considered eight label classes (background excluded). The scores given by *initial* describe the average values for image pairs before registration. The deformation field is evaluated by examination of the Jacobian determinant. The inference time refers to pairwise registration of 10,000 image pairs on GPU.

		initial	w/o IC	w/ IC
deformation	std $\det(J)$	-	1.555	1.083
	$\det(J) < 0$ [%]	-	13.3	7.0
reconstruction	MSE	-	155.023 ±56.786	187.133 ±86.644
cyclic registration	MSE	-	201.416 ±49.945	118.192 ±27.503
	Dice [%]	-	88.4 ±5.2	90.4 ±4.3
pairwise registration (10,000 pairs)	MSE	1997.696 ±554.075	904.445 ±170.163	525.960 ±116.493
	Dice [%]	14.3 ±7.9	23.8 ±5.3	**24.1 ±5.5**
	inference time	-	70 s (7 ms/image pair)	

In Table 3.1 and Table 3.2, the results for cyclic and pairwise registration are shown for 2D and 3D input data. Furthermore, the reconstruction errors averaged over the entire test datasets and properties of the Jacobian determinant of the deformation fields estimated by the shape decoder are reported. Small standard deviations indicate smooth deformation fields and values below 0 indicate singularities, i.e. foldings. The

Table 3.2: Results for 3D model: Registration and reconstruction results for 3D trained with $\mathcal{L}_{\text{InverseConsistency}}$ (w/ IC) and trained without $\mathcal{L}_{\text{InverseConsistency}}$ (w/o IC). MSE scores are calculated for image intensities scaled between 0 and 255. Dice scores are averaged over all considered eight label classes (background excluded). The scores given by *initial* describe the average values for image pairs before registration. The deformation field estimated by the shape decoder is evaluated by examination of the Jacobian determinant. The inference time refers pairwise registration of $10,000$ image pairs on GPU.

		initial	w/o IC	w/ IC
deformation	std $\det(J)$	-	1.560	0.878
	$\det(J) < 0$ [%]	-	9.4	2.2
reconstruction	MSE	-	306.150 ±102.883	261.997 ±43.555
cyclic registration	MSE	-	106.413 ±35.135	29.498 ±19.722
	Dice [%]	-	92.9 ±1.1	96.1 ±1.1
pairwise registration ($10,000$ pairs)	MSE	1763.016 ±447.623	797.198 ±162.127	519.322 ±120.342
	Dice [%]	18.6 ±11.1	51.8 ±9.3	**57.5 ±10.3**
	inference time	-	1414 s (141.4 ms/image pair)	

inference times given for pairwise image registration refer to the duration for pairwise registration of $10,000$ image pairs on GPU (Quadro P6000).

For the 2D and the 3D experiments, applying the inverse consistency penalty leads to smoother deformation fields and a decreased number of locations with occurring foldings, as indicated by the decreasing standard deviation and number of values < 0 of the Jacobian determinant. While for the 2D model the mean squared error of the reconstruction output is slightly increased with the inverse consistency constraint, the reconstruction performance of the 3D model benefits. The results for cyclic registration show the benefit of training with $\mathcal{L}_{\text{InverseConsistency}}$ for both, the 2D model and the 3D model. Considering the results for pairwise image registration, adding the inverse consistency penalty to the objective function leads to a similar average Dice overlap for the 2D model and an improvement of 5.7 % points for the 3D model.

In Fig. 3.3, results for the 3D model trained with $\mathcal{L}_{\text{InverseConsistency}}$ are visualised. It shows the reconstruction output of the model, as well as the template generated by the appearance decoder path and the deformation fields provided by the shape decoder path. In Fig. 3.4, a result comparison of the 3D model with labelled structures for cyclic registration after training with and without the introduced inverse consistency constraint is shown.

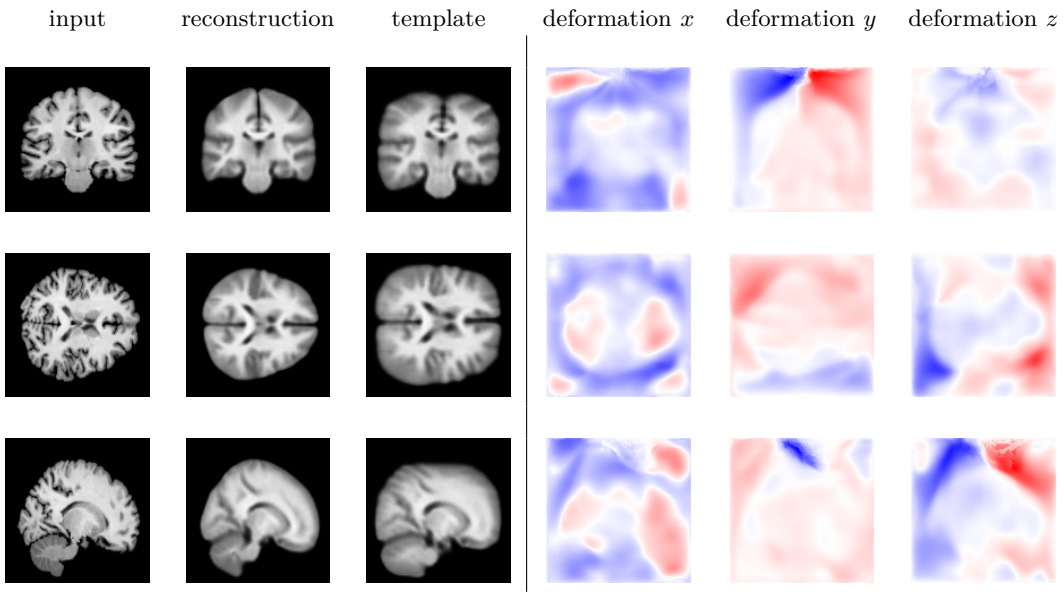


Fig. 3.3: Result visualisation (coronal/axial/sagittal; middle slice displayed respectively) of the 3D model trained with $\mathcal{L}_{\text{InverseConsistency}}$ for one test sample. Left: input image, reconstructed output, and generated template. Right: decoded deformations in all spatial directions.

3.4 Discussion

The concept of deforming autoencoders is applicable for 2D and 3D medical image processing as demonstrated by the presented results regarding the reconstruction and registration performances. Whereas the reconstruction performance is only evaluated by the MSE of the input image and the reconstructed image, the registration performance, which is the focus of this chapter, is also evaluated by the Dice score of semantic segmentations. The Dice scores for cyclic registration, that ideally should achieve a value of 100 %, yield values of 88 % (2D model) and 93 % (3D model) when training is performed without the inverse consistency loss term. Those results indicate that the presented models are applicable for image registration, but the gap to the score of

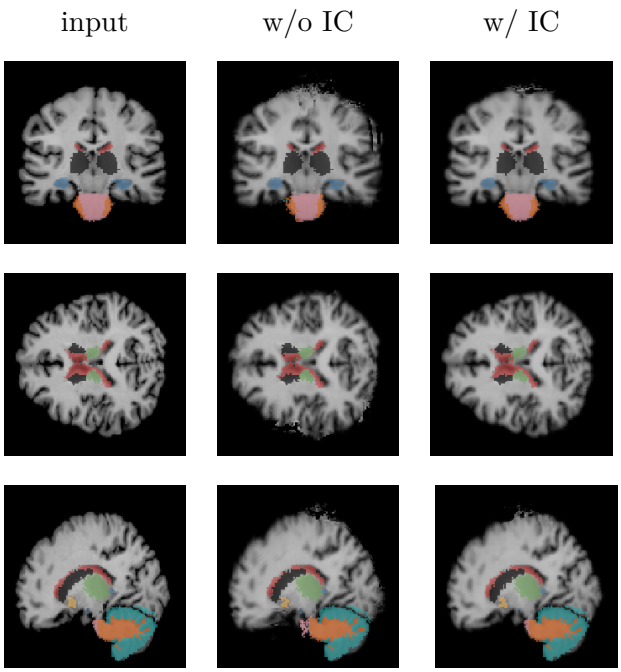


Fig. 3.4: Comparison of results (coronal/axial/sagittal) for cyclic registration of labelled data using the 3D model: input image (left column), recovered input image (successive warping of the input image with $\varphi_{\mathrm{r}\to\mathrm{t}}$ and $\varphi_{\mathrm{t}\to\mathrm{r}}$) for the model trained without $\mathcal{L}_{\mathrm{InverseConsistency}}$ (middle column), and recovered input image for the model trained with $\mathcal{L}_{\mathrm{InverseConsistency}}$ (right column). The visualisation includes the label classes considered for evaluation (lateral ventricle, brain stem, hippocampus, cerebellum exterior, cerebellum white matter, thalamus proper, putamen, caudate).

100 % suggests that the generated deformations are not yet ideally plausible. This is also underlined by the observed properties of the Jacobian determinant and motivates the necessity of the inverse consistency constraint that is introduced in this chapter.

In fact, training with the inverse consistency loss term leads to improved results for cyclic and pairwise registration, as well as improved values for the measured properties of the Jacobian determinant that indicate smoother and more plausible deformations. Nevertheless, deformation smoothness remains improvable for the presented approach. Empirical studies, however, have shown that a higher weighting of the inverse consistency loss term leads to an imbalanced training that results in substantially worse reconstruction and registration performance. In general, the search for a suitable weighting of the individual loss terms is an elaborate issue that for the presented method has been performed empirically through observation of the balance of the

registration performance, the reconstruction performance, the smoothness of the deformation fields, and the inverse consistency of the deformations. Too much weighting of the loss terms $\mathcal{L}_{\text{InverseConsistency}}$, $\mathcal{L}_{\text{Smooth}}$, and $\mathcal{L}_{\text{BiasReduce}}$ inducing smoother and more plausible deformations quickly leads to poor registration and reconstruction results. On the other hand, too high a weighting of $\mathcal{L}_{\text{Reconstruction}}$ leads to better reconstruction and registration results, but also leads to implausible deformations. The findings obtained from observing this balance lead to the choice of weightings as described in Sec. 3.2.2.

Quantitative and qualitative results imply a network architecture and an objective function that are capable of image reconstruction, image registration, and generation of reasonably plausible deformations. The proposed method is especially beneficial for groupwise image registration of large image datasets. In contrast to pairwise image registration methods, this method offers the possibility to align images in all pairwise combinations with a reduced number of operations. Once the deformation field between each image and the template image is generated by the trained model and the reversed deformation fields are computed, it is possible to perform image registration for any image pair within the dataset.

3.5 Conclusion

In this chapter, a model is proposed that learns to reconstruct an input image, to generate a canonical deformable template, and to estimate realistic deformation fields in an unsupervised learning framework. The method successfully decouples appearance and shape to represent anatomical variability and can be used for 2D or 3D groupwise image registration. The registration performance of the model is outlined by using it for pairwise image registration of labelled image data. It is able to perform registration of $10,000$ image pairs within 70 s for 2D 256×256 data and 1414 s for 3D $128 \times 128 \times 128$ data. Adding an inverse consistency constraint to the objective function used for training leads to more precise and more plausible registration results. The experiments that compare the registration performance for the 3D models trained with and without the proposed inverse consistency constraint point out that integrating the constraint into the objective function leads an increase of 11 % of Dice overlap and a reduction of foldings within the predicted deformation fields by 77 %.

Chapter 4

Design Choices for Pairwise Image Registration Network Architectures

This chapter focuses on design considerations for deep learning-based image registration networks, as published in [Siebert et al., 2021b]. In contrast to the previous chapter, which presents a network architecture whose main focus is groupwise image registration with a disentanglement of the latent space, this chapter introduces a method for pairwise image registration with initially disentangled feature extraction network parts. Overall, this chapter is dedicated to U-Net-based architectures for pairwise deformable image registration and, based on the findings therein, proposes a partially decoupled feature extraction for the two input images. In the further course of this chapter, several design options for deformation learning based on U-Net architectures are compared. The proposed two-stream architecture concatenates features extracted by separate encoder blocks for moving and fixed image before further processing. For this image registration architecture, unsupervised learning and label supervised learning are compared. Experiments are conducted on inter-patient abdominal CT scans.

4.1 Introduction

Pairwise deformable image registration aims to align two images or image volumes by predicting non-linear transformations that optimise an appearance or shape-based metric. It plays an important role in clinical practice, including diagnostic tasks, image-guided interventions, and motion tracking [Hill et al., 2001]. As outlined in Chapter 2.1, pairwise registration also helps to analyse medical image datasets for research purposes. Recent deep learning-based image registration methods, such as [Balakrishnan et al., 2019; Chen et al., 2022; Eppenhof et al., 2019, 2020; Heinrich, 2019; Mok et al., 2020b] show the potential to outperform conventional methods in terms of improved registration speed and accuracy. However, the estimation of large deformations is still considered challenging. Meanwhile, there is a large variety of publications presenting different deep learning networks for image registration, each offering suggestions for the design of network architectures.

As outlined in the following Sec. 4.1.1, the structure of numerous registration network architectures is based on a U-Net [Ronneberger et al., 2015]. The U-Net has proven to be not only suitable for image segmentation tasks, but also for application in image registration tasks [Balakrishnan et al., 2018]. Its use for pairwise image registration with two input images rather than a single input image allows for a variety of design details concerning the different architectural modules. This chapter will examine whether and how certain design details affect the image registration performance.

4.1.1 Related Work

A deep learning-based method for medical image registration called VoxelMorph is presented in [Balakrishnan et al., 2018] and in an extended version in [Balakrishnan et al., 2019]. The authors propose a framework with an included U-Net architecture that takes a fixed and a moving image as input data and outputs a deformation field, which is used to warp the moving image with a spatial transformer function. In their work, the authors employ VoxelMorph for deformable image registration of 3D brain MR scans, but emphasise that their method is applicable to a broad range of other registration tasks. It can be trained in an unsupervised manner or in a setting with auxiliary segmentation labels as weak supervision. The loss function used for unsupervised training consists of two components: a similarity inducing component and a smoothness inducing component. The similarity component penalises differences in appearance either by minimising the mean squared error of image intensities or by maximising local cross correlations. The smoothness component regularises the deformation field by penalising local spatial variations. If anatomical segmentation masks are available, an auxiliary data loss term based on the Dice score can be added to the loss function [Sudre et al., 2017]. In [Dalca et al., 2019a], an extension of VoxelMorph is presented that introduces a probabilistic model for diffeomorphic image registration by integrating stationary velocity fields that are estimated by a U-Net architecture through scaling and squaring as differentiable network operations.

In [Hu et al., 2019], the authors modify VoxelMorph by using a dual-stream network architecture and a pyramid registration module. The dual-stream U-Net architecture generates two feature pyramids separately for the fixed and the moving input image, which allows for meaningful convolutional features. Whereas VoxelMorph estimates a single deformation field, the proposed pyramid registration module of Hu et al. predicts multiple deformation fields with different resolutions. With this approach, the authors were able to achieve improved results compared to VoxelMorph for brain MRI registration.

In [Zhao et al., 2019a], the authors indicate that a cascaded version of VoxelMorph with multiple warping steps increases the registration performance. To this end, the authors present recursive cascaded networks for deformable image registration. It can be built on VoxelMorph or on any other base registration network. In an unsupervised

end-to-end training, several cascades learn to perform a progressive deformation for the current warped image. Thus, the flow field is decomposed, such that the top cascades rather learn a global alignment, while the bottom cascades take over the role of refinement. When using the network architecture that is introduced in [Zhao et al., 2019b] as base network, the registration results for the investigated liver CT and brain MR datasets could be improved compared to using VoxelMorph as base network for the cascades.

For deformable pulmonary CT registration, a supervised learning method based on a U-Net architecture is proposed in [Eppenhof et al., 2019]. The network is trained to estimate transformations between image pairs from a set of images that are synthetically deformed. The loss function that is used during training of the end-to-end U-Net architecture minimises the ℓ_1 norm of the estimated and simulated ground truth vector field. In [Eppenhof et al., 2020], the same network architecture is used for a progressive training during which the U-Net is gradually expanded to include higher resolution layers. Thereby larger deformations are learned in lower-resolution layers and finer deformations are learned in higher-resolution layers.

A symmetric image registration method that integrates a U-Net architecture is presented in [Mok et al., 2020a]. It maximises the similarity of a fixed and a warped moving image within the space of diffeomorphic maps and estimates forward and inverse transformation simultaneously. The authors propose a local orientation-consistency loss term that leverages the Jacobian determinant for diffeomorphic deformation properties. The target velocity fields are learned in unsupervised manner by a five-level U-Net and then passed to scaling and squaring layers, as well as differentiable spatial transformers.

In [Heinrich et al., 2022], a heatmap prediction network head is appended to VoxelMorph and a multichannel instance optimisation is proposed for fine-tuning the feedforward displacement field predictions. The heatmap enables a discretised integral regression of displacements for keypoint supervision or for a non-local unsupervised metric loss based on MIND features. Ablation experiments on lung CT scan registration demonstrate the benefit of the proposed methods for non-linear alignment.

4.1.2 Contribution

The design options that are considered in this chapter take up the idea of several registration networks, which include a U-Net architecture to learn deformations (see Sec. 4.1.1). A comparison of basic architectural settings, i.e. the number of feature channels and the number of convolutions is performed. Moreover, the idea of not directly concatenating a fixed and a moving image for feature extraction is examined. Since many single-stream registration network architectures such as VoxelMorph struggle with the alignment of inter-patient abdominal image data with large deformations, this chapter aims to address this problem by proposing a two-stream registration architecture. The presented approach employs separate feature extraction modules for

both input images in the first network blocks. The following sections aim to provide a detailed investigation regarding the impact of different architectural design ideas on the registration performance in order to propose a pairwise image registration architecture for inter-patient abdominal CT scans that can be trained both in unsupervised manner or with label supervision.

4.2 Methods

The methodological focus of this chapter is on the comparison of U-Net-based image registration network architecture modifications. The considered design options are described in detail in Sec. 4.2.1. In Sec. 4.2.2, the unsupervised training procedure relying on MIND feature similarity [Heinrich et al., 2013] and diffusion regularisation, as well as the supervised training procedure that adds label supervision are described.

4.2.1 Network Architecture Modifications

Since U-Net architectures play a key role in numerous pairwise medical image registration frameworks, four basic designs options are examined in this chapter. The investigations are carried out in such a way that a simple architecture acts as a starting point, which is then gradually extended and further modified. All considered architectures are visualised in Fig. 4.1.

The studies start with a registration model that is from its basic structure similar to VoxelMorph [Balakrishnan et al., 2019] with a configuration that contains two skip connections. The model concatenates the fixed and moving images at the beginning and uses sequences of convolution followed by instance normalisation and leaky ReLU. In total, this first variant of an image registration U-Net comprises approximately $750,000$ parameters. Within the network architecture, the resolution of the spatial dimensions is first successively decreased by strided convolutions to $\frac{1}{24}$ of the input image dimensions and then increased again by up-convolutions. The central part of the architecture consists of a U-Net like part using two skip connections [Ronneberger et al., 2015]. Conversely to the reduction in resolution, the number of feature channels from the convolution layers is firstly increased up to 64 and then decreased until the output yields three feature channels that correspond to the three displacement dimensions.

In detail, as illustrated in Fig. 4.1 (a), the first network layer outputs features with a resolution of half of the input image dimensions and two feature channels. Within the second layer, the number of feature channels is increased to eight and the output resolution is decreased to a third of the input image dimensions. The following layer increases the number of feature channels to 32 while keeping the current spatial resolution, whereas the layer hereafter keeps the number of feature channels while decreasing the spatial resolution to a sixth of the initial image dimensions. Then, the number of feature channels is doubled to 64 and the resolution is again halved to $\frac{1}{12}$

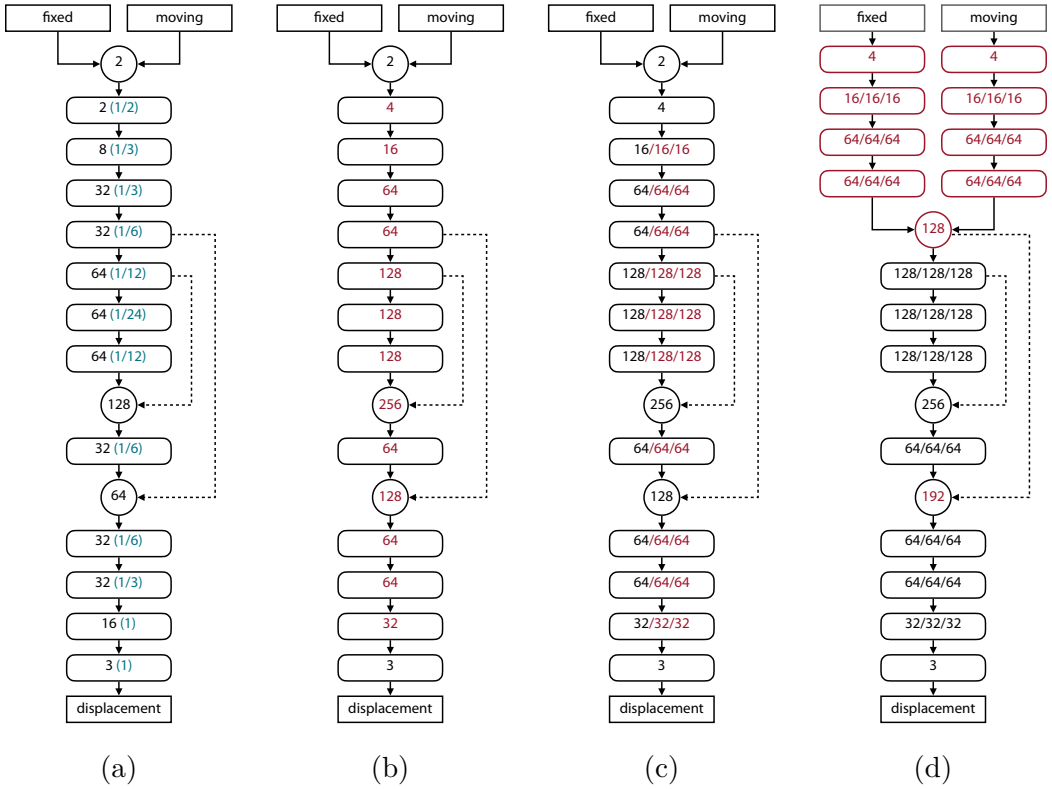


Fig. 4.1: Overview of the considered modifications with given number of feature channels being output from the convolutions (rounded rectangles) and concatenations (circles). The initial architecture (a) is modified so that first the number of feature channels (b) and then the number of convolutions (c) is increased. The last modification (d) results in a two-stream architecture that starts with separate encoder blocks for the fixed and the moving image. Modifications to the previous architecture are marked in red, respectively. The corresponding output resolutions are indicated in blue within the visualisation of model (a) and apply to all of the models.

before the lowest resolution of $\frac{1}{24}$ is reached within the next layer, which constitutes the last layer of the network's encoding part. The decoding part starts with increasing the feature dimensions to $\frac{1}{12}$ of the original image dimensions, and then concatenating the output with 64 feature channels with the 64 feature channel output of the same resolution from the encoding part using a skip connection. This again is passed through the next network layer, which increases the feature resolution to a sixth and yields a 32 channel output that is then concatenated by the other skip connection with the encoding part's 32-channel output with the same resolution. The resulting 64-channel feature maps are given to the next network layer, which halves the number of feature channels to 32 before the next layer increases the spatial resolution to a third of the initial image dimensions. Within the following layer, a 16-channel output with the initial spatial

image dimensions is reached. Finally, the number of channels is reduced to three, corresponding to the three displacement dimensions. As the proposed implementation operates within a coordinate system normalised on the underlying image dimensions, the displacements are considered to be within the range of $[-1, 1]$. Therefore, the last convolution layer is followed by a tanh activation function and the obtained output is used for warping with the moving input image.

The first modification that is made to this initial architecture is to double the number of feature channels of all convolution layers of the network as outlined in Fig. 4.1 (b). This adjustment quadruples the number of parameters. Specifically, this modification entails that the first convolution layer outputs four feature channels and the convolutional layer in the lowest resolution outputs 128 feature channels. Only the final layer remains unchanged as the number of desired output channels is three.

A further modification extends the number of Convolution-InstanceNorm-ReLU sequences per resolution level to three (Fig. 4.1 (c)). Exceptions to this modification are the first resolution layer directly after concatenation of the input images and the last layer in which the three final feature maps are provided.

Finally, a two-stream architecture (Fig. 4.1 (d)) with separate encoder blocks for fixed and moving images is proposed. Their concatenated output is used as input for the U-Net part of the architecture. In contrast to the work of [Hu et al., 2019], which introduces a continuous dual-stream architecture, the two streams are concatenated at one fixed location within the encoder part of the network at a spatial resolution of $\frac{1}{6}$ of the input dimensions. This implies that in each of the streams, the first layer generates an output with four feature channels corresponding to the respective fixed or moving input image. The following three encoder layers match the corresponding encoder layers from the architecture described previously and successively increase the number of feature channels to 64 and decrease the feature resolution to $\frac{1}{6}$. As monomodal data is used for the experiments, which means that the fixed and moving images are interchangeable, the weights are shared between the two encoders of this two-stream architecture. The obtained 64-channel features for both input images are then concatenated yielding a 128-channel feature map that is passed to the remaining and not further modified network layers, consisting of the central part of the architecture with the U-Net like part and the final decoder layers.

4.2.2 Training Supervision

The models in all examined modifications are trained using a loss function

$$\mathcal{L} = \mathcal{L}_{\text{Similarity}} + \lambda_{\text{dr}} \mathcal{L}_{\text{Regularisation}} \tag{4.1}$$

with a regularisation weighting parameter λ_{dr}. The loss function ensures similarity of the fixed and warped moving images and smooth deformation fields. Modality independent

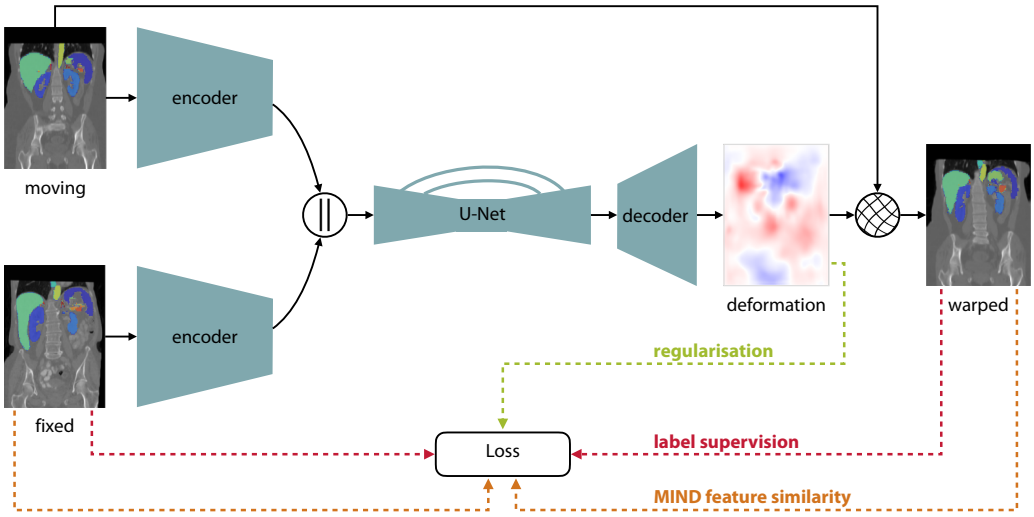


Fig. 4.2: The model for pairwise image registration with label supervision: Fixed and moving images are given into separate encoder blocks for the extraction of features that are then concatenated and passed to a U-Net and following decoder block for the estimation of displacements. The obtained displacement fields are used to warp the moving image. The loss function is designed in such a way that the warped moving image and labels resemble the fixed image and labels (similarity of MIND features and label supervision) and that furthermore the deformation fields are smooth (diffusion regularisation).

neighbourhood descriptors (MIND) with self-similar context (SSC) [Heinrich et al., 2013] (see Sec. 2.9) are extracted from the fixed image I_F and $I_W = I_M \circ \varphi$ denoting the moving image I_M warped with the displacement field φ. The mean squared error (MSE) is calculated yielding

$$\mathcal{L}_{\text{Similarity}} = \text{MSE}\Big(\text{MIND}(I_F), \text{MIND}(I_W)\Big) \tag{4.2}$$

with MIND features computed as defined in Eq. 2.9. Additionally, diffusion regularisation

$$\mathcal{L}_{\text{Regularisation}} = |\nabla\varphi_x|^2 + |\nabla\varphi_y|^2 + |\nabla\varphi_z|^2 \tag{4.3}$$

is applied over all three spatial gradients to achieve smooth and plausible deformation fields.

For the proposed two-stream architecture, the benefits of label supervision are examined by further extending the loss function by

$$\mathcal{L}_{\text{LabelSupervision}} = w_{\text{label}}\big(\text{MSE}(Seg_F^{\text{onehot}}, Seg_W^{\text{onehot}})\big) \tag{4.4}$$

computing the MSE between fixed and warped moving one-hot encoded label maps Seg_F^{onehot} and Seg_W^{onehot} (background excluded) weighted inversely proportional to the square root of the class frequency by w_{label}. In Fig. 4.2, the final image registration approach including the resulting loss function

$$\mathcal{L} = \mathcal{L}_{\text{Similarity}} + \lambda_{\text{dr}}\mathcal{L}_{\text{Regularisation}} + \lambda_{\text{ls}}\mathcal{L}_{\text{LabelSupervision}} \tag{4.5}$$

with diffusion regularisation and label supervision weighted by the parameters λ_{dr} and λ_{ls} is illustrated.

4.3 Experiments and Results

The experiments are conducted using the Learn2Reg challenge dataset containing 30 abdominal CT scans with 13 manually labelled abdominal organs [Hering et al., 2022a]. The labelled organs include the spleen ■, the right kidney ■, the left kidney ■, the gall bladder ■, the esophagus ■, the liver ■, the stomach ■, the aorta ■, the inferior vena cava ■, the portal and splenic vein ■, the pancreas ■, the left adrenal gland ■, and the right adrenal gland ■ [Xu et al., 2016].

The data has been linearly pre-registered and is re-sampled to dimensions of $144 \times 112 \times 192$ voxels in order to reduce computational complexity. The dataset is split into 20 training cases and 10 test cases. For evaluation, all possible pairwise combinations of the test cases are considered, which leads to 45 unique pairs. To match the image resolution of $192 \times 160 \times 256$ voxels provided for the Learn2Reg challenge, the predicted deformation fields are upsampled with trilinear interpolation, and evaluation is carried out on the dataset with a resolution of $192 \times 160 \times 256$ voxels.

The networks are trained using Adam and a learning rate of 0.001 (0.0001 for the baseline network of VoxelMorph) for $50,000$ iterations. Diffusion regularisation λ_{dr} is weighted in such a way that the standard deviation of the Jacobian determinant stays below 1.0 on the training set. This reduces the number of singularities and thus leads to a higher plausibility of the predicted deformations. For label supervision, a weighting of $\lambda_{ls} = 2$ is chosen.

In Table 4.1, the average Dice overlap and properties of the Jacobian determinant are reported, as well as the inference time on GPU (Quadro P6000) and the number of trainable parameters. Comparing the registration performance, the first model examined (1-stream (a), unsupervised) is only able to achieve a gain in Dice overlap of $\sim 2.5\,\%$ points compared to the initial overlap of $25.15\,\%$ and is outperformed by the originally proposed VoxelMorph. The original VoxelMorph is therefore trained with the same loss and has a higher number of skip connections and lower number of parameters. Each subsequent modification improves the registration result. The model with an increased number of feature channels (1-stream (b), unsupervised) leads to an improvement of about $2\,\%$ points compared to the first model. Increasing the

Table 4.1: Evaluation results using Dice scores, Jacobian determinant, average inference time on GPU (Quadro P6000), and number of parameters. Dice scores are averaged over all thirteen label classes (background excluded). Initial refers to average values before registration. The quality of the deformation field is evaluated through the Jacobian determinant. Small standard deviations indicate smooth deformation fields and values below 0 indicate singularities, i.e. foldings.

	avg Dice [%]	std $\det(J)$	$\det(J) < 0$ [%]	inf. time/ pair [ms]	param. count
initial	25.14 ± 12.85	-	-	-	-
VoxelMorph unsuperv.	31.70 ± 13.75	0.5853	3.61	117.58	396451
1-stream (a), unsuperv.	27.85 ± 12.56	0.4096	0.89	44.09	746467
1-stream (b), unsuperv.	29.78 ± 12.60	0.4600	1.10	50.17	2985251
1-stream (c), unsuperv.	31.72 ± 13.01	0.4184	0.96	75.09	6814499
2-stream (d), unsuperv.	$\mathbf{35.39 \pm 14.05}$	0.4681	1.38	102.32	7449123
2-stream (d), superv.	$\mathbf{43.85 \pm 11.33}$	0.5012	1.37	102.32	7449123

Table 4.2: Dice overlap in % for the 13 different observed label classes.

	spleen	right kidney	left kidney	gall bladder	esophagus	liver	stomach	aorta	inferior vena cava	portal and splenic vein	pancreas	left adrenal gland	right adrenal gland
initial	42	34	35	2	23	62	24	33	36	5	15	8	9
VoxelMorph, unsuperv.	53	45	45	3	28	72	29	47	44	7	17	12	10
1-stream (c), unsuperv.	54	41	44	4	29	73	31	44	45	8	16	13	10
2-stream (d), unsuperv.	60	49	50	4	28	78	35	50	46	11	19	17	12
2-stream (d), superv.	73	69	71	7	37	83	47	59	52	12	26	20	15

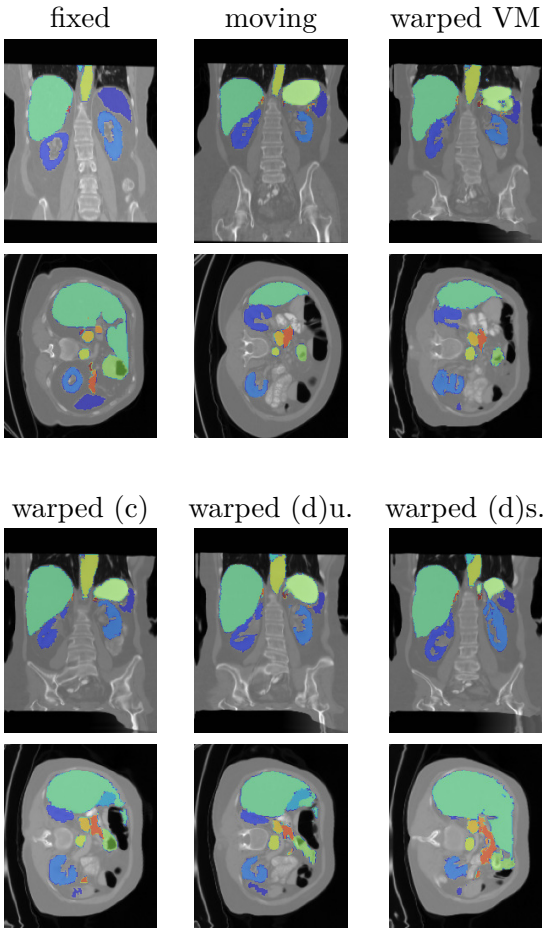


Fig. 4.3: Result visualisation (coronal/axial) of different architectures (VoxelMorph (VM), 1-stream (c), unsupervised 2-stream (d)u., supervised 2-stream (d)s.) for one test pair: fixed and moving image and warped moving images output from the different models.

number of convolution-normalisation-activation blocks per resolution level from one to three (1-stream (c), unsupervised) increases the Dice overlap by another 2 % points. This leads to scores for registration accuracy that are on par with VoxelMorph, while VoxelMorph is outperformed in terms of deformation field smoothness. With the unsupervised 2-stream model (d), an average Dice overlap of 35.39 % is achieved, outperforming VoxelMorph by nearly 4 % points. When training with label supervision, this score could be further improved to 43.85 %. The results clearly show that the proposed two-stream architecture yields a significant gain in registration performance for inter-patient abdominal CT alignment. This result is competitive to many other approaches [Hering et al., 2022a].

Table 4.2 shows Dice scores for the different considered label classes, pointing out that the registration methods show the best improvement of Dice overlap for comparably large and medium-sized organs. For these organs, label supervision during training also shows the highest improvement of registration performance compared to unsupervised training. These findings are also exemplary visually confirmed by Fig. 4.3, which shows the results of different architectures and training settings in coronal and axial view. Corresponding to the increased number of parameters of each architecture modification, the inference time per registered image pair increases. However, the inference time measured for the two-stream architecture is still better compared to VoxelMorph.

4.4 Discussion

The main contribution of this chapter is to outline the impact of basic architectural modifications on the performance of U-Net architectures for pairwise image registration. By performing monomodal inter-patient CT registration as an exemplary use case, it is demonstrated that adjusting simple design options can lead to more precise image registration. Both the increase in the number of feature channels per resolution layer and the increase in the number of convolution operations lead to an improved image registration performance. The highest improvement, however, is achieved by applying the proposed two-stream architecture. It is shown that a separate feature extraction for the fixed and the moving input image within the first network blocks is beneficial compared to direct concatenation of the input data. It is also shown that training the registration model with label supervision leads to better results than an unsupervised training procedure. For both findings, the benefit of initially separated feature extraction modules and the benefit of label supervision can be directly transferred to other deep learning image registration architectures and pairwise image registration tasks. The proposed two-stream approach can also be used for multimodal applications. However, in cases of multimodal alignment, it is important that the weights are not shared between the two encoder modules before concatenation of the extracted features. An according architecture designed for multimodal alignment is presented in Chapter 5.

In addition to the modifications examined, numerous other design options would be conceivable within the scope of U-Net architectures for pairwise image registration. Further options to be investigated could be the depth of the applied U-Net architecture, i.e. the number of resolution layers and the latent size between contracting and expanding network parts. Moreover, the impact of the number of skip connections would be a possibility for further studies. In terms of the two-stream concept, it could be examined which number of convolutional layers for the separated network streams or concatenations at which resolution layer are the most advantageous.

Overall, the results presented in this chapter show an advantage of the proposed two-stream architecture over single-stream VoxelMorph. Nevertheless, the results could

not compete with the results of recent learning and non-learning-based registration approaches used for the same registration task [Hering et al., 2022a]. Especially the approach presented in [Mok et al., 2020b] points out that a U-Net-related architecture can yield significantly increased registration performance for this task when implementing more extensive network design changes like a multi-level framework and, generally, deeper network architectures. In Chapter 6, a method with non-trainable modules for feature alignment and deformation field generation is presented that yields an average Dice score of 68 % for the same evaluation dataset, which is an improvement of 25 % points compared to the two-stream architecture presented in this chapter.

4.5 Conclusion

This chapter compares several architecture modifications for deep learning-based deformable pairwise image registration. Besides the observations that increased numbers of feature channels and numbers of convolution-normalisation-activation sequences lead to improved registration results, it is demonstrated that concatenating the features extracted by separate encoder blocks for the moving and fixed images achieves better results than directly concatenating the input images. This two-stream architecture is able to outperform the simple baseline network for unsupervised pairwise image registration VoxelMorph by almost 4 % points for the Dice overlap while estimating smoother deformation fields. Due to the fact that the experiments are performed on a labelled dataset, it is shown that including label supervision when training the proposed two-stream registration model leads to a further substantial increase of Dice overlap of 8 % points compared to unsupervised training.

Chapter 5

Learning a Metric for Multimodal Registration Based on Cycle Constraints

This chapter is dedicated to a supervision procedure for multimodal medical image registrations that has been introduced in [Siebert et al., 2022b] and [Siebert et al., 2021a]. The approaches introduced in the previous chapters include training procedures that are based on a loss function with a known similarity metric and regularisation term. This form of supervision requires that the individual terms of the loss functions are balanced with the help of one or more weighting parameters. The aim of this chapter is to improve over the use of hand-crafted metric-based losses and introduce a concept for supervision that is applicable to new datasets without domain knowledge. Therefore, the use of synthetic three-way cycles for multimodal image registration is proposed. A differentiable and end-to-end learnable method for estimating large rigid transformations is introduced. The proposed cycle constraints are employed for learning cross-modality features for alignment of intra-patient abdominal CT and MR scans. Similar to the principle of the two-stream feature extraction network part proposed in the previous chapter, the underlying feature extraction CNN in this chapter uses separated feature encoding for each modality within early network blocks. The experiments include a comparison to metric supervision and label supervision, as well as a comparison of inference strategies.

5.1 Introduction

Medical image registration based on deep learning methods has gathered great interest over the last few years. Yet, certain challenges, especially in multimodal registration, need to be addressed for learning-based approaches, as evident from the recent MICCAI challenge *Learn2Reg* [Hering et al., 2022a]. In order to avoid an elaborate comprehensive annotation of all relevant anatomies and label bias, unsupervised metric-based registration networks are common praxis for intra-modal deep learning-based registration [Balakrishnan et al., 2018; Heinrich, 2019].

However, this poses an additional challenge for multimodal registration problems, as currently no universal metric has been developed. Additionally, there has to

be made a trade-off between using local contrast-invariant edge features such as normalised gradient fields (NGF), local correlation coefficient (LCC), and modality independent neighbourhood descriptors (MIND) or more global statistical metrics like mutual information (MI). Metric-based methods also entail the difficulty of tuning hyperparameters that balance similarity costs (ensuring similarity between fixed image and warped moving image) and regularisation (ensuring plausible deformations).

Ground truth deformations for direct supervision are only available if synthetic deformation fields are applied. The now very popular FlowNet [Dosovitskiy et al., 2015] estimates deformation fields between pairs of input images from a synthetically generated dataset that has been obtained by applying affine transformations to images. In the context of medical applications, synthetic deformations have been deployed for monomodal image registration [Eppenhof et al., 2018; Eppenhof et al., 2019; Krebs et al., 2017]. Alternatively, label supervision that primarily maximises the alignment of known structures with expert annotations could be employed [Balakrishnan et al., 2018; Blendowski et al., 2021; Hu et al., 2018b]. This leads to improved registration of anatomies that are well represented. However, it can introduce a bias and deteriorate performance for unseen labels.

On the one hand, the focus of supervised approaches on a limited set of labelled structures may be in particular inadequate for diagnosis of pathologies that are not sufficiently represented in the training data. Using metric supervision, on the other hand, has little potential to improve upon classical algorithms that employ the same metric as similarity term during optimisation. With efficient (parallelised) implementations, adequate runtimes of less than a minute have recently been achieved for classical algorithms.

By learning completely without metric or label supervision, self-supervision could remedy the aforementioned problems and enable the development of completely new registration methods and multimodal feature descriptors without introducing annotation or engineering biases.

5.1.1 Related Work

Self-supervised approaches have been used in medical and non-medical learning-based image processing tasks. Recently, a self-supervised approach for learning pretext-invariant representations for object detection has outperformed supervised pre-training in [Misra et al., 2020]. By minimising a contrastive loss function, the authors learn image representations that are invariant to image patch perturbation, that are similar to the representation of transformed versions of the same image, and that differ from representations of other images. The contrastive loss helps the method to learn transformation invariant image representations by minimising the dissimilarity of representations of an input image and its transformed counterpart. In [Komodakis et al., 2018], semantic features have been learned with self-supervision in order to recognise

the rotation that has been applied to an image, given four possible transformations as multiples of 90°. The learned features are useful for various visual perception tasks. For rigid registration between point clouds, an iterative self-supervised method has been proposed in [Wang et al., 2019b]. Here, partial-to-partial registration problems have been addressed by learning geometric priors directly from data. The method comprises a keypoint detection module, which identifies points that match in the input point clouds based on co-contextual information and aligns common keypoints.

In [Wang et al., 2019a], cycle-consistency in time is used for learning visual correspondences from unlabelled video data for self-supervision. Their idea is to obtain supervision for correspondence by tracking backward and then forward, i.e. along a cycle in time, and use the inconsistency between the start and end points as the loss function. For image-to-image translation, a cycle-consistent adversarial network approach is introduced in [Zhu et al., 2017]. The authors use a cycle consistency loss that induces the assumption that forward and backward translation should be bijective and inverse of each other. Another approach that addresses inconsistency is introduced in [Gass et al., 2015] for medical image registration. It considers information from a complete set of pairwise image registrations, aggregates inconsistency, and minimises the groupwise inconsistency of all pairwise image registrations by means of a regularised least squares algorithm. The idea to measure consistency via registration cycles for monomodal medical image data has been employed in the method of [Christensen et al., 2001], which estimates forward and reverse transformation jointly in a non-deep-learning approach and [Datteri et al., 2012], who use registration circuits to correct registration errors. A monomodal unsupervised medical image registration method to train deep neural networks for deformable registration using CNNs with cycle-consistency is presented in [Kim et al., 2019]. This approach is based on two registration networks that process the two input images as fixed and moving images inversely to each other and gives the deformed volumes to the networks again to re-deform the images to impose cycle-consistency.

Previous deep learning-based registration work has often omitted the step of rigid or affine registration, despite its immense challenges due to often large initial misalignments. Image registration challenges such as [Hering et al., 2022a] provide data that has been pre-aligned with the help of non-deep-learning-based methods, whereas the challenge's image registration tasks are then often addressed with deep learning-based methods. Rigid transformation is in many cases employed as initial step before performing deformable image registration, and only a few works [Vos et al., 2019] investigate deep learning techniques for this step. As evident from the CuRIOUS challenge [Xiao et al., 2019], so far no CNN approach was able to learn a rigid or affine mapping between multimodal scan pairs (MRI and ultrasound of neurosurgery) with an adequate robustness. Besides that, no label bias can occur with rigid alignment since it is a global transformation model. Hence, a learning method for large linear transformations is of great importance.

5.1.2 Contribution

In order to avoid the difficulty of choosing a metric for multimodal image registration, a completely new concept is proposed in this chapter. For learning multimodal features for image registration, the introduced learning method requires neither label supervision nor hand-crafted metrics. It extends upon research that successfully learn monomodal alignment through synthetic deformations, but transforms this concept to multimodal tasks without resorting to complex modality synthesis. In more detail, a cycle-based approach is proposed which includes cycles that for each pair of CT and MRI scans comprise two multimodal transformations to be estimated and one known synthetic monomodal transformation. This chapter is restricted to 3D rigid registration and aims to learn multimodal registration between CT and MRI without metric supervision by minimising the cycle discrepancy. For feature extraction, a CNN is used that initially applies separate encoder blocks for each modality followed by shared weights within the last layers. A correlation layer without trainable weights is used and a differentiable least squares fitting procedure is applied to find an optimal 3D rigid transformation.

5.2 Methods

In this section, a learning concept for multimodal image registration that learns without metric supervision is introduced. The presented method learns with the help of a self-supervised learning procedure using three-way cycles. The basic idea of the learning-based approach is illustrated in Fig. 5.1. It relies on geometric instead of metric supervision and is described in Sec. 5.2.1. In Sec. 5.2.2, an overview of the training pipeline is given. For the presented registration model, the architectural design is described in Sec. 5.2.3 and consists of modules for feature extraction, correlation, and registration. Correlation and transformation computation are explained in detail in Sec. 5.2.4.

5.2.1 Self-supervised learning strategy

The proposed deep learning-based method learns multimodal registration without using metric supervision. Instead, it is based on geometric self-supervision by minimising the cycle discrepancy created through a cycle consisting of two multimodal transformations and one monomodal transformation. The basic cycle idea is illustrated in Fig. 5.1: Initially, a fixed image (*Image 1*) and a moving image (*Image 2*) are considered. The transformation between those images R_{21} is unknown and is to be learned by the method. In each training iteration, the moving image (*Image 2*) is randomly deformed by applying a known random transformation R_{23}. Hereby, a synthetic image (*Image 3*)

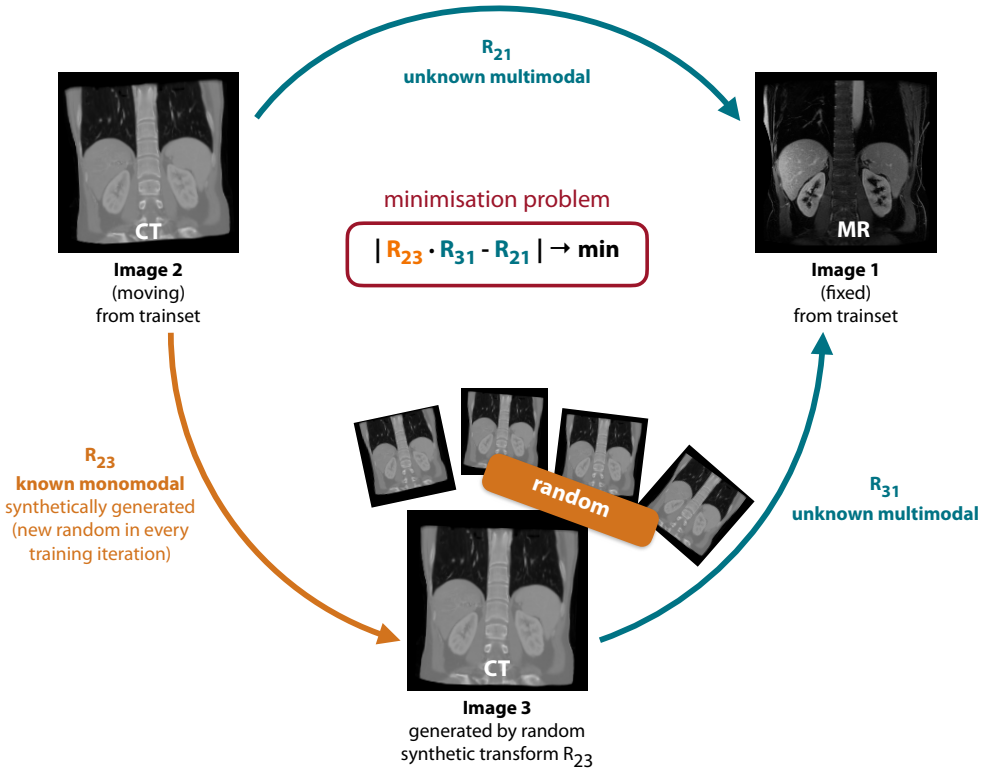


Fig. 5.1: The proposed self-supervised learning concept for multimodal image registration aims to minimise a cycle discrepancy. In every training iteration, another (known) random transformation matrix R_{23} is used to generate a synthetic image. Like this, a cycle consisting of two unknown multimodal transformations with the transformation matrices R_{21} and R_{31} and a known monomodal transformation with the transformation matrix R_{31} is obtained leading to the minimisation problem of $|R_{23} \cdot R_{31} - R_{21}| \rightarrow \min$ that is used for learning.

is obtained. By bringing the individual transformations between the three images into a cycle, the minimisation problem of

$$|R_{23} \cdot R_{31} - R_{21}| \rightarrow \min \tag{5.1}$$

can be derived. Instead of minimising the difference between the transformation combination $R_{23} \cdot R_{31} \cdot R_{12}$ and the identity transformation Id with $|R_{23} \cdot R_{31} \cdot R_{12} - \mathrm{Id}| \rightarrow \min$, the discrepancy is minimised by means of Eq. 5.1 in order to avoid that the method only learns identity warping. For optimisation, the mean squared error loss function is applied to minimise the cycle discrepancy between the two flow fields generated by the transformation matrices R_{21} and $R_{23,31} = R_{23} \cdot R_{31}$.

As the model in this chapter is restricted to rigid registration, the synthetic transformations R_{23} are created by randomly initialising rigid transformation matrices with values that are assumed to be realistic from an anatomical point of view.

The advantages of the proposed learning concept are manifold. First, in contrast to supervising the learning with a hand-crafted similarity metric and regularisation term, the need for balancing a weighting term is not required, and the method is applicable to new datasets without domain knowledge. Second, it enables multimodal learning, which is not feasible using synthetic deformations in conjunction with image-based loss terms as, for example, in [Eppenhof et al., 2018]. Third, it avoids the use of domain discriminators as implemented in the CycleGAN approach [Xu et al., 2020; Zhu et al., 2017], which usually requires a large set of training scans with comparable contrast in each modality and may be sensitive to hyperparameter choices.

On the first sight, it might seem daring to use such a weak guidance. Once suitable features are learned, the loss term enables convergence, since the cycle constraint is fulfilled. Yet, to initiate training towards improved features, the approach primarily relies on the power of randomness by drawing multiple large synthetic deformations and exploratory learning. In addition, the architecture contains a number of stabilising elements: a patch-based correlation layer computation, outlier rejection, and least squares fitting. A detailed description of these methods is provided in Sec. 5.2.3 and Sec. 5.2.4.

5.2.2 Training pipeline

The presented self-supervised learning strategy entails a training procedure, which consists of repeating the steps visualised in Fig. 5.2 within each training iteration: First, a random rigid transformation matrix R_{23} is generated and applied on the moving image in order to obtain the synthetic image. Then, the moving and the fixed image are passed through the feature extraction module that outputs features for each input modality. The extracted features are passed to the correlation layer and the transformation computation module to obtain the transformation matrix R_{21}. The same step is also performed for the fixed and the synthetic image to yield R_{31}. After this, R_{23} and R_{31} are combined to obtain $R_{23,31}$. Finally, the mean squared error of the deformations calculated with help of R_{21} and $R_{23,31}$ is determined as measure for the cycle discrepancy and used for model training. A more detailed description of the individual modules for this training pipeline is given in the subsequent sections Sec. 5.2.3 and Sec. 5.2.4.

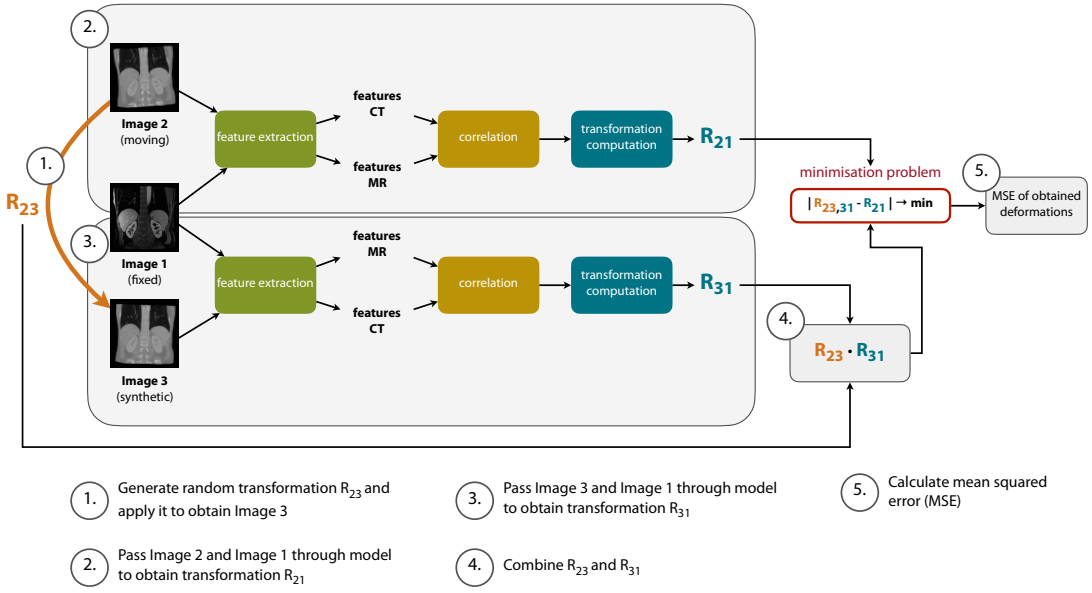


Fig. 5.2: Pipeline to train the registration model: A random transformation matrix R_{23} is generated and used to obtain the synthetic image. The pair of a moving and a fixed image as well as the pair of a synthetic and a fixed image are passed through feature extraction, correlation layer and transformation computation module (see following Sec. 5.2.3 and Sec. 5.2.4) to obtain the transformation matrices R_{21} and R_{31}. Then, R_{23} and R_{31} are combined to obtain $R_{23,31}$. As a final step, the mean squared error (MSE) of the deformations is calculated with the help of R_{21} and $R_{23,31}$.

5.2.3 Architecture

The architecture of the presented registration method consists of three main components for feature extraction, correlation, and computation of the rigid transformation. A schematic overview of these components is given in Fig. 5.3.

For feature extraction, a CNN with initially separate encoder blocks for each modality and shared weights within the last few layers is used. These features are subsequently fed into the correlation layer, which has no trainable weights and whose output is directly converted into displacement probabilities. As the next step, the method employs a robust and differentiable least squares fitting to find an optimal 3D rigid transformation subject to outlier rejection.

For the feature extraction CNN, a Y-shaped architecture [Blendowski et al., 2021] is used (cf. Fig. 5.3), starting with a separate network part for each of the two modalities (ModalityNet). The ModalityNet takes the respective input and passes it through two sequences, which apply the following operations twice: a 3D convolution with a kernel size of three and padding of one, followed by 3D instance normalisation, and leaky

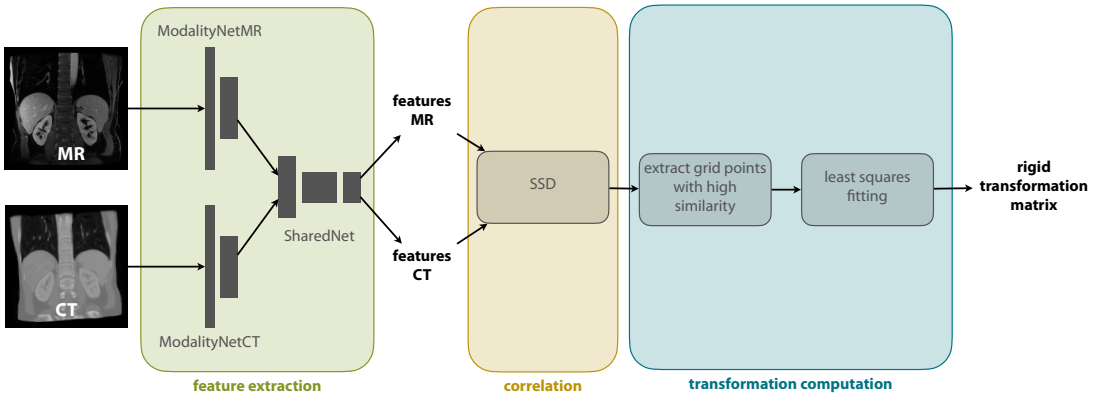


Fig. 5.3: The process of feature extraction, correlation, and computation of the rigid transformation matrix: A CNN is used for feature extraction starting with a separate network part for each modality (ModalitiyNetMR and ModalityNetCT) followed by a module with shared weights (SharedNet). The obtained features are correlated by calculating patch-wise the sum of squared differences (SSD). Subsequently, grid points with high similarity are extracted and used to define point-wise correspondences to calculate the rigid transformation matrix with a least squares fitting.

ReLU activation. The two convolutions of the first sequence are non-strided and output eight feature channels. The first convolution of the second sequence has a stride of two and doubles the number of feature channels to 16. The second convolution of the second sequence is non-strided and keeps the number of 16 feature channels. Whereas the size of the input dimensions is preserved within the first convolution sequence, the strided convolution within the second sequence halves each feature map dimension. The outputs of the ModalityNets are passed into a final module with shared weights (SharedNet), which finalises the feature extraction by applying two sequences of the same structure as used for the separate ModalityNets. Here, the first sequence applies non-strided convolutions that output 16 feature channels, while keeping the spatial dimensions as output by the ModalityNets. The first convolution of the second sequence has a stride of two, again halving the sizes of the spatial dimensions and doubling the number of feature channels to 32. The second convolution of the second sequence is non-strided and keeps the number of 32 feature channels. The output of the SharedNet is given to a $1 \times 1 \times 1$ convolution, providing the final number of 16 feature channels, followed by a sigmoid activation function. As correlation and transformation estimation

techniques without trainable weights are used, the model only comprises 80k trainable parameters within the feature extraction part.

5.2.4 Correlation and transformation computation

As suggested in previous research [Dosovitskiy et al., 2015; Heinrich, 2019], the application of a dense correlation layer that explores a large number of discretised displacements at once is employed to capture large deformations robustly. This way, the learned features are used to define a sum of squared differences cost function akin to metric learning [Simonovsky et al., 2016].

Similar to [Modat et al., 2014] which operates directly on input image pairs and uses normalised cross correlations (NCC), a block-matching technique to find correspondences between the fixed features and a set of transformed moving features is used. The obtained features are correlated by calculating the sum of squared differences (SSD) patch-wise. Points with high similarity are extracted from a coarse grid with a spacing of 12 voxels. The extracted grid points are used to define point-wise correspondences to calculate the rigid transformation matrix with a robust trimmed least squares fitting procedure.

For the correlation layer, a set of $11 \times 11 \times 11$ discrete displacements with a capture range of approximately 40 voxels in the original volumes is chosen. After calculating the SSD cost volume, the obtained SSD costs are sorted. The 50 % of the displacement choices that entail the highest similarity costs are rejected. On the remaining displacement choices, the softmax function is applied to obtain differentiable soft correspondences. While this differentiable approach is used to estimate regularised transformations within a framework that comprises trainable CNN parameters, the learned features could as well be used for other optimisation frameworks [Blendowski et al., 2021].

The displacement candidates output by the softmax function are added to the coarse moving grid points. In a least squares fitting procedure comprising five iterations, the final rigid transformation matrix that serves for transformation of the moving image is determined. The best-fitting rigid transformation can be found by computing the singular value decomposition $S = U\Sigma V^T$ with the matrix $S = X^T Y^T$. Herein, X denotes the centred fixed grid points x_i and Y the centred moving grid points with added displacement candidates y_i. The orthogonal matrices U and V are obtained by the singular value decomposition. This leads to the rotation

$$Q = V \begin{pmatrix} 1 & & & & \\ & 1 & & & \\ & & ... & & \\ & & & 1 & \\ & & & & \det(VU^T) \end{pmatrix} U^T \tag{5.2}$$

and the translation

$$t = \overline{y} - Q\overline{x} \tag{5.3}$$

with $\overline{x}$ being the mean values for fixed grid points and $\overline{y}$ the mean moving grid points with added displacement candidates.

This way, the rigid transformation matrices R_{21} and R_{31} are determined. To combine the synthetic transformation R_{23} and the predicted transformation R_{31}, matrix multiplication is used yielding $R_{23,31}$. With the transformation matrices R_{21} and $R_{23,31}$, the affine grids Φ_{21} and $\Phi_{23,31}$ can be computed, which are then used as input for the MSE-based loss function

$$\mathcal{L}_{\text{CycleDiscrepancy}} = \text{MSE}(\Phi_{21}, \Phi_{23,31}) \tag{5.4}$$

during training. Additionally, the affine grid Φ_{21} computed by R_{21} is used for warping during inference to align the moving image to the fixed image.

This approach has the advantage of being very compact with only 80k parameters, ensuring memory efficiency and fast convergence of training. The multimodal features learned by the proposed model are generally usable for image alignment and can be given to various optimisation methods for image registration once trained with the presented method.

5.3 Experiments and Results

The experiments are performed on 16 paired abdominal CT and MR scans from collections of The Cancer Imaging Archive (TCIA) project [Akin et al., 2016; Clark et al., 2013; Erickson et al., 2016; Linehan et al., 2016]. Labels for four abdominal organs – liver, spleen, left kidney, and right kidney – have been manually created and are used for the evaluation of the proposed methods. The preprocessing comprises reorientation, resampling to an isotropic resolution of 2 mm, and cropping and padding to volume dimensions of $192 \times 160 \times 192$ voxels.

To increase the number of training and testing pairs and model realistic variations in initial misalignment, the scans are augmented with eight random rigid transformations, each that on average reflect the same Dice overlap of approximately 43 % as the raw data. All models are trained for 100 epochs with a mini-batch size of four in less than 45 minutes, all using approximately 8 GByte of GPU memory.

The weights of the CNN used for feature extraction (FeatCNN) are trained using the Adam optimiser with an initial learning rate of 0.001 and a cosine annealing scheduling is employed.

5.3.1 Comparison of training strategies

To train the introduced FeatCNN, four different strategies are compared in a two-fold cross-validation:

1. FeatCNN + Cycle Discrepancy: The proposed self-supervised cycle learning strategy trained for
$$\mathrm{MSE}\Big(\Phi_{21}, \Phi_{23,31}\Big) \rightarrow \min.$$

2. FeatCNN + MI Loss: Learning with metric supervision using Mutual Information (MI) defined as
$$\mathrm{MI}(I_F, I_M, \Phi_{21}) = H(I_F) + H(I_M \circ \Phi_{21}) - H(I_F, I_M \circ \Phi_{21}), \tag{5.5}$$
with H being the entropy and as implemented by [Sandkühler et al., 2018] to yield
$$-\mathrm{MI}\Big(I_F, I_M, \Phi_{21}\Big) \rightarrow \min.$$

3. FeatCNN + NCC^2 Loss: Learning with metric supervision using squared local normalised cross correlations (NCC^2) [Balakrishnan et al., 2019; Modat et al., 2014]. The normalised cross correlation measurement (NCC) is defined as
$$\mathrm{NCC}(I_F, I_M, \Phi_{21}) = \frac{\sum I_F \cdot (I_M \circ \Phi_{21}) - \sum \mathrm{E}(I_F)\mathrm{E}(I_M \circ \Phi_{21})}{|\Omega| \cdot \sum \mathrm{Var}(I_F)\mathrm{Var}(I_M \circ \Phi_{21})}, \tag{5.6}$$
where the sums go over the image domain Ω , E is the expectation value (here: mean value), and Var is the variance [Sandkühler et al., 2018]. The models are trained to aim
$$-\mathrm{NCC}\Big(I_F, I_M, \Phi_{21}\Big)^2 \rightarrow \min.$$

4. FeatCNN + Label Loss: Supervised learning with label supervision using onehot encodings of the provided segmentations S_F and S_M the mean squared error function to learn
$$\mathrm{MSE}\Big(S_F^{\mathrm{onehot}}, S_M^{\mathrm{onehot}} \circ \Phi_{21}\Big) \rightarrow \min.$$

All methods share the same settings for the correlation layer and a trimmed least square transform fitting with five iterations and 50 % outlier rejection. Hyperparameters are determined on a single validation scan for cyclic training and left unaltered for all other experiments.

The same trainable FeatCNN, comprising the layers as described in Sec. 5.2.3, is used to train with the proposed Cycle Discrepancy Loss, MI, NCC^2, and Label Loss. For correlation, corresponding grid points are extracted within a grid with a spacing of 12 voxels. Patches with a radius of 2 voxels are used to calculate the SSD in a

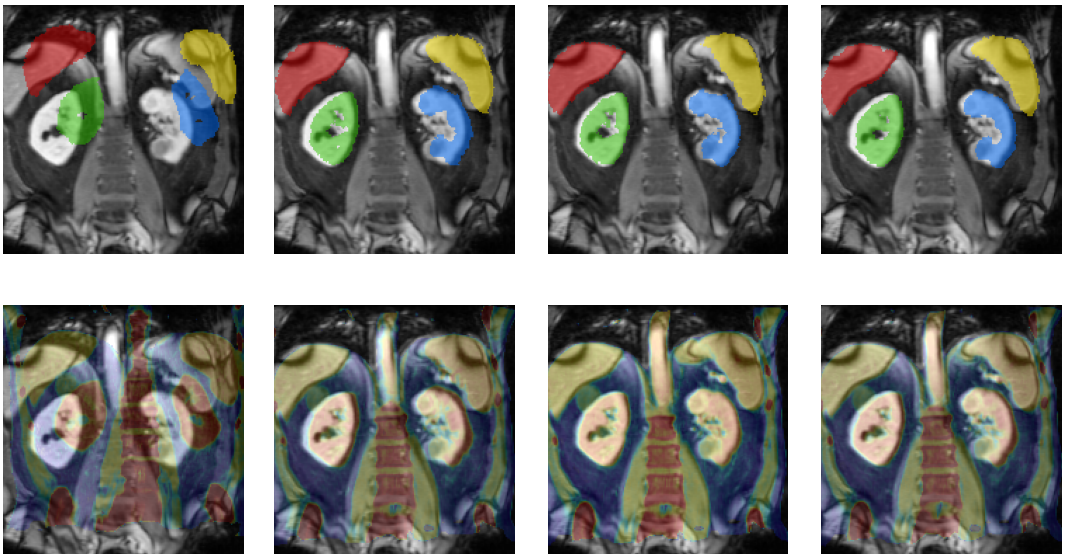
Table 5.1: Results for the cross-validation experiments: Dice scores [%] with standard deviation listed by anatomical structures of the 3D experiments using FeatCNN for feature extraction and MI Loss, NCC^2 Loss, Label Loss, or the Cycle Discrepancy for training.

	liver	spleen	lkidney	rkidney	mean
initial	59.32 ± 14.03	36.90 ± 19.49	36.59 ± 19.53	37.02 ± 22.08	42.46 ± 18.78
FeatCNN + MI Loss	75.07 ± 9.38	63.17 ± 22.13	69.86 ± 26.34	64.46 ± 29.45	68.14 ± 21.92
FeatCNN + NCC^2 Loss	75.08 ± 12.22	61.09 ± 23.69	72.19 ± 27.51	64.04 ± 31.76	68.10 ± 23.80
FeatCNN + Cycle Discrepancy	77.95 ± 8.16	69.89 ± 16.00	70.18 ± 24.34	71.85 ± 34.40	72.30 ± 20.75
FeatCNN + Label Loss	81.24 ± 8.75	73.84 ± 18.32	83.15 ± 26.62	79.97 ± 33.59	79.55 ± 21.82

patch-wise manner. A displacement radius of 4 is used and the set of displacement possibilities is discretised for the correlation layer with a displacement step (respective voxel spacing) of 5. To adjust the smoothness of the soft correspondences, the costs obtained by SSD computation are multiplied by a factor of 150 when given to the softmax function. As the soft correspondences are needed for differentiability only during training, this factor is increased to 750 for inference.

For the cycle discrepancy method, the synthetic transformation matrices R_{23} are created by randomly sampling them with values that are assumed to be realistic from an anatomical point of view. Therefore, the maximum rotation is $\pm 23°$ and the maximum translation ± 42 voxels, which equals 84 mm for the performed experiments, in every image dimension.

The results demonstrate a clear advantage of the proposed self-supervised learning procedure with an average Dice of 72.3 % compared to the state-of-the-art MI metric loss with 68.14 % and NCC^2 Loss with 68.1 %, which is suitable for multimodal registration due to its computation involving small local neighbourhoods [Modat et al., 2014]. The result of the model trained with the proposed cycle discrepancy loss comes close to the theoretical upper bound of the model trained with full label supervision with 79.55 %. In Table 5.1, detailed quantitative results are given for all comparison methods and all evaluated anatomical structures. Fig. 5.4 depicts qualitative registration results.



initial	FeatCNN + MI Loss	FeatCNN + Cycle Discrepancy	FeatCNN + Label Loss

Fig. 5.4: Qualitative results of the proposed cycle discrepancy approach and selected comparison methods (coronal slices). The top row shows the fixed MRI and (warped) moving labels for liver ■, spleen ■, left kidney ■, and right kidney ■. The bottom row visualises the (warped) moving CT and a jet colourmap overlay of the fixed MRI scan.

5.3.2 Comparison of inference strategies and increased training dataset

To further enhance the method, experiments are conducted that extend it by a two-level warping approach during inference. More precisely, the input moving and fixed images are presented to the model to warp the moving image. Then, the model is applied to the resulting warped moving image and the fixed image again. For both warping steps, a displacement radius of 7 voxels and a grid spacing of 8 voxels is set. For the first warping step, a displacement discretisation of 4 voxels is used, and then this hyperparameter is refined to 2 voxels for the second warping step.

The dataset is quite small and the proposed method does not require labels. Therefore, when considering an application scenario where a number of MR/CT scan pairs have to be aligned offline, a fine-tuning of the networks on this test data would be feasible. Thus, the aim is to further increase the performance of the method with training on all available paired CT and MR scans without splitting the dataset.

Table 5.2 compares the results of single-level and two-level warping, as well as the cross-validation results and the results when training on the whole available image data.

Table 5.2: Results for the experiments comparing single-level and two-level warping approach, as well as cross-validation and training without withheld data: Dice scores [%] are listed by anatomical structures of the experiments using Cycle Discrepancy for training.

	liver	spleen	lkidney	rkidney	mean
initial	59.32 ± 14.03	36.90 ± 19.49	36.59 ± 19.53	37.02 ± 22.08	42.46 ± 18.78
cross-validation, 1 warp	77.95 ± 8.16	69.89 ± 16.00	70.18 ± 24.34	71.85 ± 34.40	72.30 ± 20.75
cross-validation, 2 warps	80.71 ± 9.33	72.12 ± 17.08	79.33 ± 26.06	74.65 ± 36.91	76.68 ± 22.34
trained without withheld data, 1 warp	81.04 ± 8.22	71.11 ± 18.03	76.27 ± 24.25	76.49 ± 32.64	76.23 ±20.88
trained without withheld data, 2 warps	81.85 ± 0.58	76.77 ± 13.64	79.81 ± 24.52	80.17 ± 34.65	79.65 ± 20.25
NiftyReg *reg_aladin*	83.97 ± 6.19	76.55 ± 12.00	79.83 ± 7.12	79.26 ± 37.55	79.90 ± 15.15

The results achieved by the method are compared with the results achieved using the rigid image registration tool *reg_aladin* of NiftyReg [Modat et al., 2014], applied to the image pairs without the symmetric version and with one registration level.

Introducing a second warping step increases the cross-validation results by more than 4 % points for the proposed method. When training without a withheld test dataset, further improvements by another 3 % points is achieved. These results are on par with the results of the state-of-the-art classic method NiftyReg-*reg_aladin*.

5.4 Discussion

In this chapter, a completely new concept for multimodal feature learning is presented. It is applicable for 3D image registration and operates without label supervision or hand-crafted metrics. The new supervision strategy is based on synthetic random transformations. The method comprises two transformations across modality and one within that together form a triangular cycle. Minimising the two multimodal transformations in such a cycle constraint avoids singular solutions (predicting identity transforms) and enables the learning of large rigid deformations.

Through explorative learning, modality independent feature extractors can be successfully trained that enable highly accurate and fast multimodal medical image alignment by minimising a cycle discrepancy in training. In contrast to other deep learning-based image registration models that often require pre-aligned input data, the presented work is a deep learning model for robustly estimating large misalignments of multimodal scans.

Despite the very promising results, there are a number of potential extensions that could further improve the proposed concepts. The idea of incremental learning and predicting more useful synthetic transformations to improve detail alignment could be considered and has already shown potential in preliminary 2D experiments.

While the gap between training and test accuracy is relatively small due to the robust architectural design, further fine-tuning would be applicable at test time with moderate computational effort since no supervision is required. Combining hand-crafted domain knowledge with self-supervised learning might further boost accuracy. Similarly, domain adaptation through adversarial training could be incorporated to explicitly model the differences of modalities.

5.5 Conclusion

With the introduced method, it is possible to improve over the use of hand-crafted metric-based losses by using synthetic three-way cycles. By minimising the cycle discrepancy, multimodal registration between CT and MRI can be learned without metric supervision. A robust method is created to estimate large rigid transformations. The method is differentiable in end-to-end learning. Intra-patient abdominal CT-MRI registration can be successfully performed while outperforming state-of-the-art metric-supervision. Compared to training with Mutual Information or squared local normalised cross correlations as metric supervision, the average Dice overlap is increased by 6 % when employing the proposed learning strategy. When training without a withheld test dataset and including a second warping step, the achieved results are on par with state-of-the-art method NiftyReg.

Chapter 6

Multitask Medical Image Registration with Little Learning

This chapter presents a method that aims to be applicable to a very wide range of image registration tasks. It combines feature extraction, a coupled convex discrete optimisation procedure, and Adam-based instance optimisation in a self-configuring medical image registration framework. The proposed method is the only image registration approach in this thesis that contains non-differentiable modules and is not end-to-end trainable. It makes use of pretrained semantic feature extraction models for the individual datasets and therefore only requires little learning. The proposed automatic hyperparameter selection scheme examines various hyperparameter combinations and uses a ranking technique to make a choice. Experiments are carried out on all currently available Learn2Reg challenge datasets, and the results are compared to the best challenge submissions. The content of this chapter is related to [Siebert et al., 2022a] and in preparation for journal article submission.

6.1 Introduction

Medical image registration entails a number of requirements and challenges, all of which should ideally be met simultaneously by a versatile image registration method. Methods for medical image registration should be able to align anatomical structures precisely, applying smooth and plausible transformations and operating fast enough to preferably allow real-time application. However, there are particular challenges such as capturing large displacements or movements, as well as handling multimodal or inter-patient data. Especially large three-dimensional image data requires many degrees of freedom for deformable image registration, which can quickly become very computationally intensive. Deep learning techniques have recently facilitated many medical image processing tasks. Registration methods particularly benefit from fast inference times of trained deep learning models that offer the possibility for deformable alignment of large image volumes. Image registration methods usually involve multiple hyperparameters that can control the properties of the generated registration results to meet requirements such as accuracy, smoothness, or time efficiency. The optimal balance of these properties

poses a challenge for current research in the field of learning-based image registration. Another important aspect is the search for a versatile image registration framework that is applicable for various image registration tasks and requires no or only little user intervention.

6.1.1 Related Work

As this work presents an approach for image registration that involves deep learning for feature extraction, as well as correlation, conventional convex optimisation for feature alignment, instance optimisation, and an automatic hyperparameter selection scheme, a brief overview is given on state-of-the-art learning-based image registration methods and furthermore on hyperparameter selection strategies.

A deep learning-based image registration framework called VoxelMorph is presented in [Balakrishnan et al., 2019]. It consists of a U-Net architecture that estimates a deformation field followed by a spatial transformer for image warping and can be trained in unsupervised way or with included label supervision. This approach is often considered as baseline method for deep learning-based deformable pairwise image registration and has inspired other medical image registration methods [Jia et al., 2021; Li et al., 2022; Mok et al., 2020a; Zhao et al., 2019a; Zhu et al., 2022]. Another registration architecture whose basic principles can be found in many other methods is PWC-Net [Sun et al., 2018] that uses feature pyramids, a warping layer, and a cost volume for optical flow estimation with dense correlation for 2D input data. Multi-resolution registration with image pyramids has also been successfully applied in various learning-based registration approaches for 3D image data [Gunnarsson et al., 2020; Hering et al., 2019; Mok et al., 2020b, 2021] and, together with label supervision, yields state-of-the-art registration performance [Mok et al., 2020b]. The principle of using a dense correlation layer for registration learning has been employed in PDD-Net [Heinrich, 2019] for an end-to-end trainable registration approach that requires less trainable weights. In PDD-Net, feature extraction CNNs provide features of both the fixed and the moving image that are passed to the correlation layer and a spatial fitting with differentiable mean-field regularisation. Recently, Transformer-based architectures have shown potential in image registration tasks, as they are able to capture long-range spatial correspondence between image pairs particularly well with the help of self-attention mechanisms [Chen et al., 2022; Mok et al., 2022].

Conventional strategies to determine a learning-based registration method's most suitable hyperparameters involve grid search, random search, or sequential search with training and validating multiple models with multiple sets of hyperparameters. However, such methods often encounter the drawback of being time-consuming. Meta learning aims to learn how to learn across tasks in a systematic, objective, and data-driven way [Vanschoren, 2019]. A method that learns the influence of registration hyperparameters on deformation fields is HyperMorph [Hoopes et al., 2021] which

comprises a meta network, or hypernetwork, that estimates a spectrum of registration models by learning a continuous function of the hyperparameters. While this method strongly increases the number of trainable parameters compared to the method without hypernetworks, a more parameter-efficient method to learn the effect of hyperparameters is proposed in [Mok et al., 2021]. The authors present a framework that learns conditional features which are correlated with the regularisation hyperparameter. Based on the assumptions that deformation field characteristics can be captured and separated by a CNN, conditional image registration modules that contain hidden states are used to shift feature statistics. This enables the control of smoothness regularisation during inference and fast hyperparameter tuning. Other meta learning approaches in the field of medical image registration use meta learning for interactive adaption of network initialisation to improve performance of individual registration tasks formed by data varied by interaction [Baum et al., 2023] or within a gradient-based method to quickly adapt to various image registration tasks on limited datasets from new domains [Park et al., 2022].

6.1.2 Contributions

Despite recent advances, deep learning-based approaches for medical image registration usually involve an elaborate learning procedure. Yet, they often struggle with the versatile usability for a wide range of tasks. This chapter aims for a method that requires little learning, builds upon efficient registration strategies [Heinrich et al., 2014], and is applicable to a wide variety of medical image registration tasks. Therefore, *ConvexAdam* is introduced, which includes a coupled convex optimisation with Adam-based instance optimisation for medical image registration. To apply ConvexAdam for various tasks, an automatic hyperparameter selection scheme is proposed. The Learn2Reg challenge [Hering et al., 2022a] provides the opportunity to evaluate and compare medical image registration algorithms for multiple anatomies and imaging modalities. ConvexAdam is able to achieve the overall 2020 and 2021 Learn2Reg challenge's first place [Hering et al., 2022a; Siebert et al., 2022a]. The main contributions of this chapter can be summarised as follows: Overall, a fast and very versatile method for large-deformation medical image registration that requires little learning is introduced. For the proposed method, feature extraction is decoupled from feature alignment, which allows task-specific adjustments for optimal results. In this chapter, either semantic label features are learned or hand-crafted MIND features are generated. With the help of a correlation layer, a coupled convex optimisation procedure, and instance optimisation, the method is capable of aligning medical image data for various tasks. An automatic hyperparameter selection procedure is proposed, yielding a self-configuring image registration framework. Experiments are conducted on all currently available Learn2Reg challenge datasets, which include the registration of monomodal and multimodal as well as intra-patient and inter-patient image data.

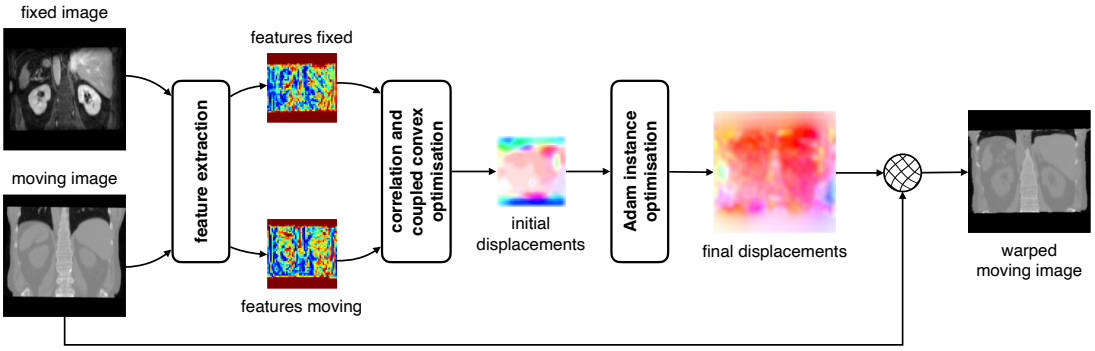


Fig. 6.1: Overview of the presented image registration method: Feature extraction is followed by correlation and coupled convex optimisation. Then, Adam-based instance optimisation is performed.

6.2 Methods

The presented method extracts features from an input image pair with the help of a pretrained segmentation network or by using a hand-crafted descriptor method. The extracted features are passed to a coupled convex optimisation procedure, followed by Adam-based instance optimisation. These steps are included in an automatic hyperparameter selection scheme that examines multiple hyperparameter combination options and leads to a self-configuring and versatile framework for medical image registration.

6.2.1 ConvexAdam: Coupled convex optimisation and instance optimisation with Adam optimisation

The proposed method called ConvexAdam performs image registration in three steps: In step 1, features $feat_M$ and $feat_F$ of an input image pair consisting of a moving image I_M and a fixed image I_F are extracted. In step 2, the extracted features are used to iteratively approximate a displacement field $\hat{\mathbf{u}}$ between the input image pair with a coupled convex discrete optimisation [Heinrich et al., 2014]. In step 3, the preliminary displacement field $\hat{\mathbf{u}}$ is used as the starting point for an Adam-based instance optimisation, which provides the final displacement field $\mathbf{u}$. Fig. 6.1 gives an overview of ConvexAdam. In the following, these three steps are described in more detail.

Step 1: Feature extraction

The presented method initially requires an extraction of the features $feat_F$ and $feat_M$ from the input images I_F and I_M. Although various feature extraction approaches are conceivable, this chapter proposes two different feature extraction strategies that

can precede the further steps of the presented method. One strategy is to compute modality independent neighbourhood descriptors (MIND) [Heinrich et al., 2012]. MIND features ensure versatility with respect to different types of registration tasks as they are hand-crafted and contrast and modality invariant. The other strategy is to use automatic segmentations as provided by the nnU-Net [Isensee et al., 2021] that by its nature can only be applied if labelled image data is available to train the segmentation models. Here, it is avoided to use the expert labels only for the warping loss as done in [Hu et al., 2018a], which may lead to suboptimal results due to limited gradient backflow. Instead, using off-the-shelf segmentation networks for feature extraction has proven to produce best results.

Step 2: Coupled convex discrete optimisation

For coupled convex discrete optimisation, the extracted image features from the input images are fed into a correlation layer. A six-dimensional displacement space volume DSV is extracted, which extends the three spatial image dimensions by three displacement dimensions. The entries of DSV refer to the similarity costs when applying certain displacements to the moving image's voxels. The sum of squared differences (SSD) is used to compute the similarity costs with a box filter. For every voxel, the similarity costs are averaged over a patch to obtain constant motion within local regions. By taking the argmin, an initial displacement field with an initial best displacement for each voxel is provided. For efficient global regularisation, the output of the correlation layer is used to solve two coupled convex optimisation problems: smoothness and similarity optimisation. A combined cost function

$$E_1(\mathbf{v}, \hat{\mathbf{u}}) = DSV(\mathbf{v}) + \frac{1}{2\theta}(\mathbf{v} - \hat{\mathbf{u}})^2 + \alpha|\nabla\hat{\mathbf{u}}|^2 \tag{6.1}$$

is minimised as described in [Heinrich et al., 2014]. To find the searched deformation field $\hat{\mathbf{u}}$, it splits up the optimisation into two convex problems by using an auxiliary deformation field $\mathbf{v}$. The first term of the cost function represents the similarity costs and the last term is for regularisation by penalising steep deformation gradients. For similarity optimisation, a discretised search space is used with quantised displacements $\mathbf{d} = \{0, \pm q, \pm 2q, ..., \pm l_{max}\}$ ranging from 0 to l_{max} with quantisation step q.

For smoothness, the second term introduces a regularisation penalty with a parameter α that controls the diffusivity of the deformation field. In several iterations, smoothness and similarity are optimised in alternation increasing the linking between them with the parameter θ that is decreased in every iteration: A spatially smoothed field with current optimal displacement in terms of minimal SSD costs is generated followed by adding a penalty to the discretised SSD costs based on the discrepancy of this current globally smooth optimum. When $\theta \to 0$, convergence is reached and the searched deformation field $\hat{\mathbf{u}}$ equals the auxiliary deformation field $\mathbf{v}$. In Algorithm 1,

the procedure of the proposed coupled convex discrete optimisation method is outlined.

Algorithm 1: Coupled convex discrete optimisation

$\hat{\mathbf{u}}^0 \leftarrow \mathbf{0}$ ▷ initialisation of the current displacement field to zeros
$DSV \leftarrow \texttt{SSD}(feat_M(\mathbf{d}), feat_F)$ ▷ computation of a correlation volume
$\mathbf{v} \leftarrow \texttt{smooth}(\texttt{argmin}(DSV))$ ▷ auxiliary displacement field
for $i \leq N_{coupled}$ **do**
  $coupled \leftarrow DSV(\mathbf{v}) + \frac{1}{2\theta^i}(\mathbf{v} - \hat{\mathbf{u}}^i)^2$ ▷ regularising coupling term added
  $\hat{\mathbf{u}}^i_{reg} \leftarrow \texttt{argmin}(coupled)$ ▷ argmin operator across all possible displacements
  $\hat{\mathbf{u}}^i_{smooth} \leftarrow \texttt{smooth}(\hat{\mathbf{u}}^i_{reg})$ ▷ spatial smoothing step
  $\hat{\mathbf{u}}^{i+1} \leftarrow \hat{\mathbf{u}}^i_{smooth}$ ▷ update displacement field
end

To enforce inverse consistency, i.e. swapping fixed and moving images while yielding inverted deformation fields, an inverse consistency optimisation is integrated into the method. For this purpose, the forward deformation field $\hat{\mathbf{u}}_{I_M \to I_F}$ and the backward deformation field $\hat{\mathbf{u}}_{I_F \to I_M}$ are calculated independently with the coupled convex discrete optimisation method. The deformation fields are iteratively updated at every position $\mathbf{x}$ while minimising the inverse consistency error $E_{ic} = |\hat{\mathbf{u}}_{I_M \to I_F} - \hat{\mathbf{u}}_{I_F \to I_M}|$ as introduced in [Heinrich et al., 2014]. In detail, the equations

$$\begin{aligned} \hat{\mathbf{u}}^{n+1}_{I_M \to I_F} &= 0.5(\hat{\mathbf{u}}^n_{I_M \to I_F} - \hat{\mathbf{u}}^n_{I_F \to I_M}(\mathbf{x} + \hat{\mathbf{u}}^n_{I_M \to I_F})) \\ \hat{\mathbf{u}}^{n+1}_{I_F \to I_M} &= 0.5(\hat{\mathbf{u}}^n_{I_F \to I_M} - \hat{\mathbf{u}}^n_{I_M \to I_F}(\mathbf{x} + \hat{\mathbf{u}}^n_{I_F \to I_M})) \end{aligned} \tag{6.2}$$

are updated for $n_{\max} = 15$ iterations. This step aims to ensure a one-to-one mapping and that the registration result is independent of the choice for fixed and moving image.

Step 3: Instance optimisation based on Adam optimiser

The resulting displacement field $\hat{\mathbf{u}}$ from step 2 is used as the starting point for an Adam-based instance optimisation. This step is very similar to classic optical flow estimation [Papenberg et al., 2006] and provides the final displacement field $\mathbf{u}$ used for warping of the moving input image I_M. For this purpose, the cost function is linearised to

$$E_2(\mathbf{u}) = |feat_M(\mathbf{u}) - feat_F|^2 + \lambda|\nabla\mathbf{u}|^2 \tag{6.3}$$

and the Adam optimiser [Kingma et al., 2014] is used for gradient descent. This is enabled by a differentiable grid sampling step with trilinear interpolation that computes the cost function with respect to the displacement field $\mathbf{u}$. Smoothness of the displacement field is induced by adding a B-spline deformation model and diffusion regularisation weighted by λ. Algorithm 2 points out the process of Adam-based instance optimisation.

Algorithm 2: Adam-based instance optimisation

for $i \leq N_{Adam}$ **do**
  $\mathbf{u}^i_{smooth} \leftarrow \texttt{BSplineModel}(\mathbf{u}^i)$ ▷ apply B-spline model
  $E^i_{diff} \leftarrow |\nabla \mathbf{u}^i_{smooth}|^2$ ▷ calculate diffusion error
  $feat^i_W \leftarrow feat_M(\mathbf{u}^i_{smooth})$ ▷ warp moving features
  $E^i_{sim} \leftarrow |feat^i_W - feat_F|^2$ ▷ calculate similarity error
  $E^i_2 \leftarrow E^i_{sim} + \lambda E^i_{diff}$ ▷ calculate loss
  $\mathbf{u}^{i+1} \leftarrow \texttt{Adam.step()}$ ▷ update parameters of deformation grid using Adam
end

6.2.2 Automatic hyperparameter selection

The application of ConvexAdam requires the choice of various hyperparameters. To this end, an automatic hyperparameter selection is proposed that uses predefined parameter options. The results of all possible parameter combinations are compared to find the combination that yields the best trade-off in terms of similarity of fixed and warped image, smoothness of the deformation field, and computation speed. Therefore, the presented hyperparameter selection scheme requires that an evaluation procedure is provided comprising at least one similarity metric, a metric that indicates the smoothness of the deformation fields, and a per-case computation time measurement. Fig. 6.2 gives an overview of the self-configuring hyperparameter selection approach. In general, the approach could be divided into three steps: First, a rule-based initialisation of the self-configuring procedure is performed. This is followed by registration with ConvexAdam for all hyperparameter combinations. As a final step, hyperparameter selection is done by evaluation of the results of all hyperparameter combinations and a scoring and ranking procedure. A detailed description of these steps is given in the following.

Rule-based initialisation

Based on several properties of the input data, a rule-based initialisation for the subsequent steps is conducted. The purpose of the rule-based initialisations is to restrict the self-configuring procedure to less possible parameter combinations in order to limit computation complexity. Depending on the modality of the fixed and moving images, the radius r_{MIND} and the dilation d_{MIND} for computation of the MIND features are set. If modalities known for noisy image data (e.g. ultrasound (US)) are involved, it is recommended to use a bigger search region for MIND feature computation and therefore higher values for r_{MIND} and d_{MIND}. The resolution of the input data defines whether hyperparameter options for large or small volumes should be used for coupled convex optimisation. If semantic label features are provided, the following self-configuring hyperparameter selection investigates both, MIND and nnU-Net features. If not, only MIND features are considered as feature extraction option. Based on the metrics that

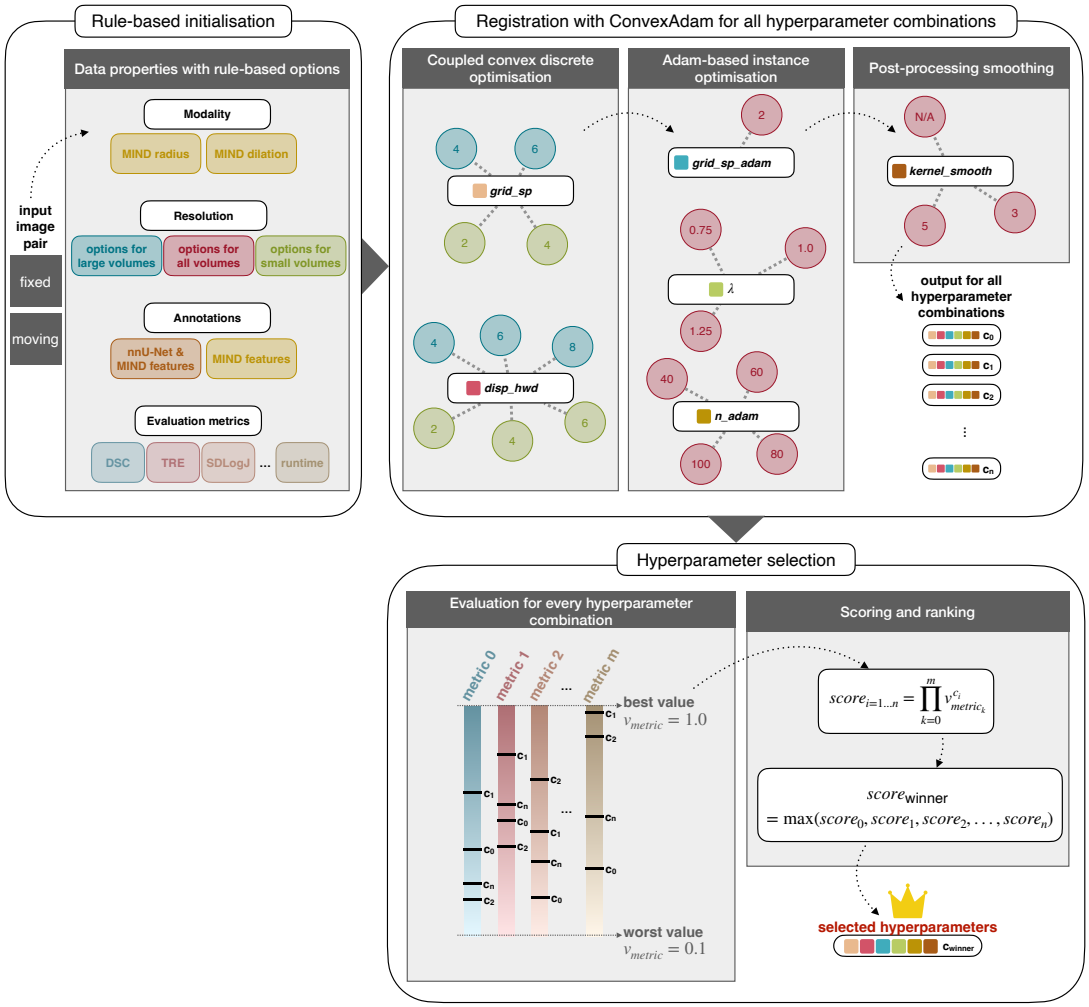


Fig. 6.2: Overview of the proposed self-configuring hyperparameter selection procedure: After a rule-based initialisation based on data properties of the input data, registration with ConvexAdam is performed for all options of hyperparameter combinations (cf. Table 6.1). The final hyperparameter selection is done by evaluating the results of every hyperparameter combination based on the provided evaluation metrics, followed by a scoring and ranking scheme.

are to be used to evaluate a specific registration task, the ranking procedure for final hyperparameter selection will be performed later.

Registration with ConvexAdam for all hyperparameter combinations

Table 6.1 gives an overview of the hyperparameters to be selected for ConvexAdam and the parameter value options proposed for hyperparameter search depending on the

Table 6.1: Hyperparameters to be selected for ConvexAdam and the options that are proposed for the automatic hyperparameter selection.

Coupled convex discrete optimisation	**opt. lrg. vol.**	**opt. sm. vol.**
grid_sp grid spacing used for feature sampling	$\{4, 6\}$	$\{2, 4\}$
disp_hwd displacement range of discretised search space	$\{4, 6, 8\}$	$\{2, 4, 6\}$
Adam-based instance optimisation	**options**	
grid_sp_adam grid spacing used for feature sampling	2	
lambda weight for diffusion regularisation	$\{0.75, 1.0, 1.25\}$	
n_adam number of iterations	$\{40, 60, 80, 100\}$	
Post-processing smoothing	**options**	
kernel_smooth kernel for smoothing with AvgPool	$\{N/A, 3, 5\}$	

resolution of the image volumes to be processed. Generally, the hyperparameters can be divided into three groups: the hyperparameters for coupled convex optimisation, for Adam-based instance optimisation, and for post-processing smoothing. For coupled convex discrete optimisation, the grid spacing used for feature sampling and the displacement range of the discretised search space have to be selected. Adam-based instance optimisation requires that also the grid spacing for feature sampling must be set and, in addition, the weight for diffusion regularisation of the deformation field and the number of iterations for optimisation. Finally, it has to be determined whether post-processing smoothing with B-splines implemented by three consecutive average pooling layers should be used and if so, the kernel size used for the average pooling operation has to be selected.

The proposed hyperparameter options are based on the findings with empirically selected hyperparameters for the Learn2Reg-2021 submission [Siebert et al., 2022a]. Parameter options are distinguished between options for large volumes and options for small volumes by means of a threshold value $th_{\text{vol}} = 1 \times 10^6$ for the total number of voxels contained in one image scan. To keep computation complexity within the correlation step of the method low and due to the fact that large volumes often go hand in hand with high resolution, it is useful to apply larger grid spacing options *grid_sp* and higher displacement range options *disp_hwd* for large volumes and smaller grid spacing options and lower displacement range options for small volumes. In order to preempt the possibility of high-resolution image data depicting fine anatomical structures and low-resolution image data depicting large anatomical structures,

the more extreme combinations $opt_{\text{except\_lrg}} = \{grid\_sp = 2, disp\_hwd = 2\}$ and $opt_{\text{except\_sm}} = \{grid\_sp = 6, disp\_hwd = 8\}$ are added to the options for the proposed automatic hyperparameter selection procedure.

Hyperparameter selection

The ranking method that is used to select the best hyperparameter combination is based on the Learn2Reg challenge's ranking scheme [Hering et al., 2022a]: For each of m metrics applied for a registration task, all n hyperparameter combinations $c_{1,\dots,n}$ are compared against each other and linearly mapped to a numerical score with values between 0.1 and 1.0 with the value $v = 0.1$ indicating the worst and the value $v = 1.0$ indicating the best observed result. Finally, the scores of all m included metrics are multiplied to obtain a score for all n hyperparameter combinations and then select the hyperparameter combination that achieves the highest product

$$score_{i=1,\dots,n} = \prod_{k=0}^{m} v_{metric_k}^{c_i}. \tag{6.4}$$

Unlike the ranking scheme of the Learn2Reg challenge, the results of all investigated cases are employed and, due to strong correlation of all results, not only of statistically significantly different results. The parameter options given in Table 6.1 result in a number of 216 combinations. Together with the 36 additional combinations obtained by including the extreme combinations with $opt_{\text{except\_lrg}}$ or $opt_{\text{except\_sm}}$, this leads to a total number of $n = 252$ combinations considered in the hyperparameter selection process.

6.3 Experiments and Results

Experiments are performed on all seven to date available non-hidden Learn2Reg challenge tasks. The Learn2Reg dataset covers a large variety of medical image registration tasks that include monomodal, multimodal, intra-patient, and inter-patient tasks. It contains scans acquired by ultrasound (US), computed tomography (CT), and magnetic resonance imaging (MRI). Detailed information on the individual datasets and their specific challenges can be found in [Hering et al., 2022a]. A brief summary of the tasks and datasets is described in the following.

CuRIOUS [Xiao et al., n.d., 2017, 2019]: The dataset for this task consists of 22 training and 10 testing datasets of the modalities MR and US with a resolution of $256 \times 256 \times 288$ voxels. The task is to perform multimodal intra-patient registration between MR and US brain scans.

HippocampusMR [Antonelli et al., 2022]: This dataset contains hippocampus MR scans with a size of $64 \times 64 \times 64$ voxels for monomodal inter-patient registration. It

includes label data for two different anatomical structures, 263 training cases, and 131 test cases.

LungCT [Hering et al., 2020]: This task is to perform inspiration-expiration registration on intra-patient lung CT data with a resolution of $192 \times 160 \times 256$ voxels. The dataset comprises 20 training cases and 10 test cases and lung masks are provided as additional data.

AbdomenCTCT [Xu et al., 2016]: The task with this dataset is to perform inter-patient registration of abdominal CT scans with a resolution of $192 \times 160 \times 256$ voxels. This dataset consists of 30 train and 20 test cases with semantic labels for 13 anatomical structures.

AbdomenMRCT [Akin et al., 2016; Clark et al., 2013; Erickson et al., 2016; Linehan et al., 2016]: This dataset includes 8 train and 8 test cases with intra-patient abdominal MR and CT scans for multimodal registration. It has a resolution of $192 \times 160 \times 192$ voxels. For this dataset, four anatomical labels for training and nine anatomical labels for testing, as well as region of interest (ROI) masks are provided.

OASIS [Marcus et al., 2007]: This task deals with the registration of inter-patient T1-weighted brain MRI. It consists of 416 training cases and 39 test cases with a resolution of $160 \times 192 \times 224$ voxels. For this dataset, segmentations of 35 anatomical structures are provided.

NLST [Clark et al., 2013; Team et al., 2011; Team, 2011]: This task is to perform intra-patient registration of follow-up longitudinal lung CT scans. The dataset comprises 220 training cases, 30 test cases, and lung masks and the image data has a resolution of $224 \times 192 \times 224$ voxels.

6.3.1 Automatic hyperparameter selection with validation datasets

The proposed automatic hyperparameter selection approach is applied to the validation cases of all tasks. For the datasets without provided semantic features, MIND features are computed within the feature extraction step. Except for the tasks involving US registration, a radius $r_{\text{MIND}} = 1$ and a dilation $d_{\text{MIND}} = 2$ is chosen. As US images are known to be noisier than other common medical imaging modalities, $r_{\text{MIND}} = 3$ and $d_{\text{MIND}} = 3$ are used for tasks including US data. If masks are provided, they are utilised to mask the image data before MIND computation. For the datasets with available semantic features, nnU-Net models are trained on the training data and the models are used in inference mode within an extended automatic hyperparameter selection procedure. The hyperparameter selection then includes both feature extraction strategies, MIND features and semantic image features, and compares them against each other. If small volumes with $H \times W \times D < th_{\text{vol}}$ are processed, test time augmentation is applied. For large image volumes, test time augmentation is omitted due to strongly increased inference time.

Table 6.2: Comparison of results and hyperparameters on the Learn2Reg validation datasets. Similarity of the aligned images is measured by target registration error (TRE) in mm or Dice score in % of segmentation labels (DSC). Smoothness of the deformation field is indicated by the standard deviation of the logarithmic Jacobian determinant (SDLogJ).

		Validation dataset						
		CuRIOUS	**HippocampusMR**	**LungCT**	**AbdomenCTCT**	**AbdomenMRCT**	**OASIS**	**NLST**
semantic features available		-	✓	-	✓	✓	✓	-
image volume categorisation (lrg./sm.)		lrg.	sm.	lrg.	lrg.	lrg.	lrg.	lrg.
similarity metric for evaluation		TRE	DSC	TRE	DSC	DSC	DSC	TRE
initial	TRE or DSC	7.02 ±6.13	57 ±12	14.64 ±6.08	26 ±7	31 ±16	57 ±5	9.70 ±2.32
MIND features (m)	TRE or DSC	1.53 ±1.00	74 ±6	2.05 ±0.75	41 ±10	82 ±6	73 ±3	1.05 ±0.42
	SDLogJ	0.27 ±0.01	0.06 ±0.01	0.06 ±0.01	0.09 ±0.01	0.11 ±0.01	0.03 ±0.00	0.04 ±0.01
nnUNet features (n)	TRE or DSC	- -	82 ±3	- -	68 ±4	50 ±21	82 ±2	- -
	SDLogJ	- -	0.03 ±0.00	- -	0.05 ±0.01	0.31 ±0.07	0.04 ±0.00	- -
selected feature type		m	n	m	n	m	n	m
selected best hyperparameters	*grid_sp*	6	4	4	6	6	6	6
	disp_hwd	6	2	6	6	6	4	4
	lambda	1.0	1.0	0.75	1.25	0.75	1.25	0.75
	n_adam	80	80	60	60	40	40	80
	kernel_smooth	/	3	3	/	5	3	3
worst result (for selected feature type)	TRE or DSC	4.70 ±5.00	78 ±3	9.94 ±7.17	58 ±7	47 ±37	79 ±2	4.21 ±2.26
	SDLogJ	0.28 ±0.01	0.03 ±0.00	0.10 ±0.03	0.05 ±0.01	0.14 ±0.01	0.04 ±0.00	0.06 ±0.01
	grid_sp	2	6	2	2	2	4	2
	disp_hwd	2	8	2	2	2	8	2
	lambda	0.75	1.0	0.75	1.25	0.75	1.25	0.75
	n_adam	40	40	40	40	40	100	40
	kernel_smooth	/	5	/	5	/	5	/
median result (for selected feature type)	TRE or DSC	2.23 ±2.07	82 ±3	2.19 ±0.74	68 ±4	81 ±10	83 ±2	1.06 ±0.33
	SDLogJ	0.24 ±0.01	0.04 ±0.00	0.05 ±0.10	0.05 ±0.01	0.13 ±0.02	0.05 ±0.00	0.03 ±0.01
	grid_sp	4	4	4	4	4	2	4
	disp_hwd	8	6	6	8	8	2	6
	lambda	1.25	0.75	1.0	1.25	1.0	0.75	0.75
	n_adam	100	60	100	100	100	80	100
	kernel_smooth	5	3	5	/	/	3	5

Table 6.2 presents the results for the Learn2Reg validation datasets and the hyperparameters selected by the presented method. The accuracy of the aligned image pairs is measured by the target registration error (TRE) or the Dice score (DSC) of segmentation labels, if provided. Smoothness of the deformation field is indicated by the standard deviation of the logarithmic Jacobian determinant (SDLogJ). For every task, the results obtained by the best hyperparameter combination when using MIND features and, if available, nnU-Net features are reported. Additionally, the selected feature type and selected hyperparameters are indicated. As comparison to the results for the respective best hyperparameter combination, the table furthermore presents the results for the worst and median performing hyperparameter combinations.

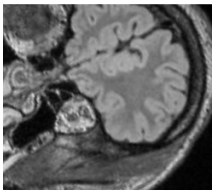
Based on the results on the validation data, the following observations are made: In all tasks, ConvexAdam combined with both MIND features and nnU-Net features is able to give results that are improvements over no registration. This even is the case with the hyperparameter configuration that leads to the worst results in the present study. For three of the four datasets with available semantic label features, superior results are observed if nnU-Net features are used instead of MIND features. If the image volumes are categorised as large, one of the combination of $\{grid\_sp = 6, disp\_hwd = 4\}$, $\{grid\_sp = 4, disp\_hwd = 6\}$, or $\{grid\_sp = 6, disp\_hwd = 6\}$ is selected for every task. Regarding the other hyperparameters, all provided options for *lambda* and *kernel_smooth* have been selected one or several times. For the number of iteration for Adam-based instance optimisation, the options $n\_adam = \{40, 60, 80\}$ have been selected.

6.3.2 Evaluation on test datasets

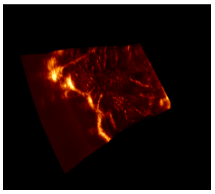
The hyperparameters determined by the automatic hyperparameter selection strategy via the validation datasets are used to apply ConvexAdam on the Learn2Reg test datasets of all tasks. Table 6.3 to 6.9 report the obtained quantitative results averaged over all test dataset cases and show qualitative results for one example image pair respectively. The results for all evaluation metrics are presented that have been considered for the Learn2Reg challenge submission evaluation and, for each task, the proposed method is compared to the four highest-ranked teams. Specifically, the method is compared to *corrField* [Hansen et al., 2021; Heinrich et al., 2015], *PDD-Net* [Heinrich et al., 2020; Heinrich, 2019], *NiftyReg* [Modat et al., 2014], *LapIRN* [Mok et al., 2020b, 2021], *MEVIS* [Häger et al., 2020; Hering et al., 2021], *Estienne* [Estienne et al., 2020a,b], *IWM* [Hering et al., 2022a], *PIMed* [Hering et al., 2022a], *Driver* [Lv et al., 2022], *xi* [Jia et al., 2022b], and *TS_UKE* [Ehrhardt et al., 2015; Werner et al., 2014].

For the datasets with provided keypoints (kp) or landmarks (lm), the target registration error (TRE) indicates the Euclidean distance between corresponding keypoints or landmarks in millimetres. To assess robustness, the 30th percentile in TRE is

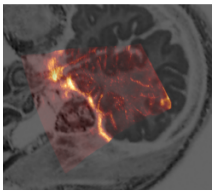
Table 6.3: Results on test dataset for CuRIOUS. Figures: Example images for fixed MR and moving US image and corresponding overlay of images before and after registration with ConvexAdam. Table: Quantitative results averaged over the entire test dataset for ConvexAdam and comparison methods.



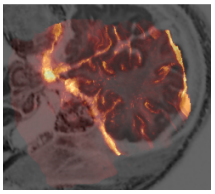
Fixed (MR)



Moving (US)



Initial



Warped

CuRIOUS				
	TRE	TRE30	SDlogJ	time
initial	6.38	12.00	–	–
ConvexAdam	2.78	4.85	0.27	1.33
corrField	2.84	5.29	0.00	2.70
PDD-Net	3.08	6.28	0.00	8.21
NiftyReg	4.09	7.85	0.00	23.1
LapIRN	5.67	11.1	0.00	34.8

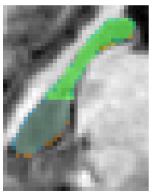
Table 6.4: Results on test dataset for HippocampusMR. Figures: Example images for fixed and moving image and corresponding overlay of fixed image with semantic label data before and after registration with ConvexAdam. Table: Quantitative results averaged over the entire test dataset for ConvexAdam and comparison methods.

HippocampusMR					
	DSC	DSC30	HD95	SDlogJ	time
initial	55	36	3.91	–	–
ConvexAdam	83	82	1.62	0.03	0.48
LapIRN	88	86	1.30	0.05	1.03
MEVIS	85	84	1.55	0.05	0.59
IWM	79	76	2.20	0.08	0.80
Estienne	85	84	1.51	0.09	1.46

Fixed (MR)

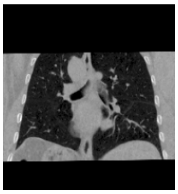
Moving (MR)

Initial



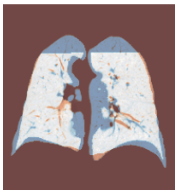
Warped

Table 6.5: Results on test dataset for LungCT. Figures: Example images for fixed inspiration and moving expiration image and corresponding masked overlay of images before and after registration with ConvexAdam. Table: Quantitative results averaged over the entire test dataset for ConvexAdam and comparison methods.



Fixed (CT)

Moving (CT)



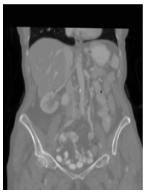
Initial

Warped

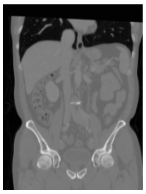
LungCT				
	TRE	TRE30	SDlogJ	time
initial	10.24	16.80	–	–
ConvexAdam	1.84	2.75	0.06	1.82
corrField	1.75	2.48	0.05	2.91
MEVIS	1.68	2.70	0.06	1.82
LapIRN	1.98	2.95	0.06	10.3
PDD-Net	2.46	3.81	0.04	4.22

Table 6.6: Results on test dataset for AbdomenCTCT. Figures: Example images for fixed and moving image and corresponding overlay of fixed image with semantic label data before and after registration with ConvexAdam. Table: Quantitative results averaged over the entire test dataset for ConvexAdam and comparison methods.

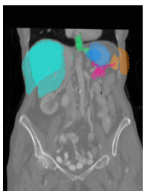
AbdomenCTCT					
	DSC	DSC30	HD95	SDlogJ	time
initial	28	24	21.78	–	–
convexAdam	72	55	8.33	0.05	2.75
LapIRN	67	47	12.51	0.12	3.80
Estienne	69	51	11.77	0.18	8.23
MEVIS	51	24	18.21	0.14	3.49
corrField	49	24	17.22.	0.28	5.40



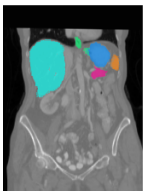
Fixed (CT)



Moving (CT)

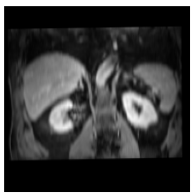


Initial

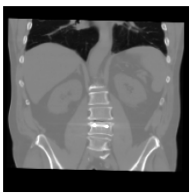


Warped

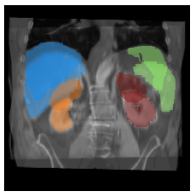
Table 6.7: Results on test dataset for AbdomenMRCT. Figures: Example images for fixed MR and moving CT image and corresponding overlay of images with semantic label data before and after registration with ConvexAdam. Table: Quantitative results averaged over the entire test dataset for ConvexAdam and comparison methods. DSC9 is a special metric that evaluates on nine additional anatomical labels to assess label bias.



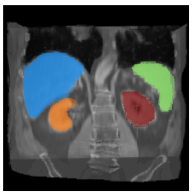
Fixed (MR)



Moving (CT)



Initial

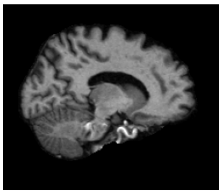


Warped

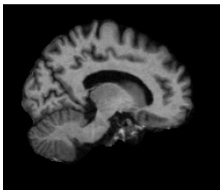
AbdomenMRCT					
	DSC	DSC9	HD95	SDlogJ	time
initial	37	23	42.07	–	–
ConvexAdam	84	76	18.85	0.14	1.30
corrField	84	76	15.67	0.10	2.13
LapIRN	84	72	15.11	0.12	1.50
PIMed	86	71	14.25	0.07	59.22
MEVIS	78	68	20.46	0.15	14.73

Table 6.8: Results on test dataset for OASIS. Figures: Example images for fixed and moving image and corresponding overlay of semantic label data before and after registration with ConvexAdam. Table: Quantitative results averaged over the entire test dataset for ConvexAdam and comparison methods.

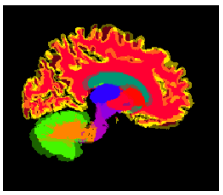
OASIS					
	DSC	DSC30	HD95	SDlogJ	time
initial	56	27	3.86	–	–
ConvexAdam	79	62	1.78	0.05	3.10
LapIRN	82	66	1.67	0.07	1.21
IWM	79	61	1.84	0.05	2.55
Driver	80	62	1.77	0.08	2.02
PIMed	78	58	1.86	0.06	3.47



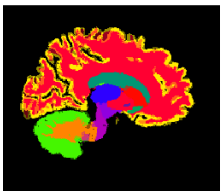
Fixed (MR)



Moving (MR)



Initial

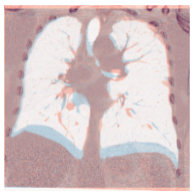


Warped

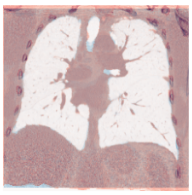
Table 6.9: Results on test dataset for NLST. Figures: Example images for fixed baseline and moving follow-up image and corresponding overlay of images before and after registration with ConvexAdam. Table: Quantitative results averaged over the entire test dataset for ConvexAdam and comparison methods.

Fixed (CT)

Moving (CT)



Initial



Warped

NLST				
	TREkp	TRElm	SDlogJ	time
initial	9.76	10.21	–	–
ConvexAdam	1.05	2.23	0.04	5.83
xi	0.79	1.70	0.05	8.9
TS_UKE	0.84	1.76	0.04	24.7
MEVIS	1.07	2.04	0.02	13.8
LapIRN	0.81	1.67	0.04	–

considered (TRE30). If label data is available, the Dice score (DSC) measures the overlap of segmentations. Therefore, to evaluate robustness, also the 30th percentile in Dice similarity across all considered labels and cases (DSC30) is reported. The 95th percentile of the Hausdorff distance (HD95) gives the 95th percentile of maximum Euclidean distances of labelled surfaces in millimetres. The standard deviation of the logarithmic Jacobian determinant (SDlogJ) is a measurement for the smoothness, and therefore plausibility, of the deformation fields. Runtime is measured on the same hardware with a Xeon Silver 4210R CPU and Quadro RTX 8000 GPU.

The test dataset results are presented in Table 6.3-6.9. Dice scores are given in %, TRE values and Hausdorff distances are given in millimetres, and runtime is given in seconds.

The results for **CuRIOUS** (Table 6.3) show the potential of the presented method to align multimodal brain scans. Whereas it outperforms the comparison methods in terms of TRE and runtime, the deformation field produced by ConvexAdam for this task is less smooth, as indicated by the comparatively high value of the standard deviation of the logarithmic Jacobian determinant.

On the other hand, aligning the dataset of **HippocampusMR** (Table 6.4) with ConvexAdam leads to smoother deformations than all reported comparison methods. The proposed method performs faster image registration, but the Dice scores are slightly lower than most of the comparison methods.

For the task of **LungCT** (Table 6.5), the results of ConvexAdam show a registration performance that is less precise with respect to the TRE compared to *corrField* and *MEVIS*, but better than the rest of the comparison methods. The results for the SDlogJ

are comparable for all reported methods and the runtime of ConvexAdam equals the runtime of *MEVIS* which is faster than for the other reported comparison methods.

The inter-patient and monomodal abdominal registration task **AbdomenCTCT** (Table 6.6) is solved by ConvexAdam best compared to all comparison methods with regard to all evaluated metrics.

For the other abdominal task **AbdomenMRCT** (Table 6.7) of intra-patient multimodal abdominal registration, the presented method performs similar to *corrField* and *LapIRN* in terms of Dice scores. While the value for SDlogJ and Hausdorff distance is higher than for most of the reported comparison methods, ConvexAdam is able to achieve the best runtime.

The results for **OASIS** (Table 6.8) show that ConvexAdam performs less well compared to *LapIRN*, *IWM*, and *Driver* regarding the Dice score, the Hausdorff distance, and the runtime.

The lung registration task **NLST** (Table 6.9) introduced in Learn2Reg-2022 is evaluated with TREs for keypoints (TREkp) and landmarks (TRElm). The results of ConvexAdam for both of the TRE values, although representing a clear improvement over the initial values, could not be ranked within the three top-performing methods. ConvexAdam demonstrates its strength in terms of runtime.

Table 6.10 shows the rank scores of comparison methods that submitted to at least four tasks of Learn2Reg-2020&2021. The reported scores are obtained by running the identical ranking procedure as introduced in [Hering et al., 2022a]. In the last column of Table 6.10, the overall rank score, calculated by averaging over the scores of the in-

Table 6.10: Rank scores of methods that submitted to at least four Learn2Reg-2020&2021 tasks.

	CuRIOUS	HippocampusMR	LungCT	AbdomenCTCT	AbdomenMRCT	OASIS	overall
ConvexAdam	0.70	0.79	0.80	0.95	0.71	0.81	0.79
LapIRN	0.49	0.93	0.75	0.83	0.78	0.92	0.78
corrField	0.86	0.34	0.87	0.49	0.82	0.45	0.64
MEVIS	0.43	0.76	0.79	0.58	0.69	0.52	0.63
NiftyReg	0.57	0.37	0.51	0.40	0.56	0.36	0.46
PIMed	–	–	0.55	0.61	0.75	0.73	0.66
PDD-Net	0.79	0.58	0.62	0.33	–	–	0.58

dividual tasks, is given. The results of self-configuring ConvexAdam are on par with *LapIRN* and outperform the other comparison methods with respect to the overall rank score.

6.4 Discussion

The proposed method shows high performance for all investigated registration tasks. It is able to to align image pairs from challenging datasets, e.g. including image pairs with large initial displacements, multimodal datasets, data without annotated training cases, and image data that requires the alignment of very small anatomical structures.

An automatic hyperparameter selection strategy is proposed to prevent elaborate manual hyperparameter search. The automatic hyperparameter selection strategy leads to top-ranked results across all tasks. Compared to other image registration methods [Hoopes et al., 2021; Mok et al., 2021], the hyperparameters required for the presented method are not selected in a learning-based way. The proposed hyperparameter selection strategy can be considered as a mixture of rule-based and grid search-based procedures and involves the validation of several configurations to achieve the reported results. The options for the individual hyperparameters could be easily adjusted, e.g. if a more refined search is preferred or if knowledge about the deformations to be expected is provided. The method uses a conventional non-differentiable convex optimisation for image feature alignment that converges fast and does not require a large training dataset like needed for deep learning-based alignment. Therefore, it is possible to apply the proposed hyperparameter selection strategy on a small set of validation data without causing extensive (training) times or parameter intensive models and transfer the determined hyperparameters for registration of unseen test data. However, as can be observed for AbdomenMRCT, too small a validation dataset can lead to suboptimal results on the test dataset due to poor generalisation.

Here, it is proposed to use nnU-Net generated segmentations or MIND features as input for the correlation layer that precedes the iterative coupled convex optimisation. In general, any meaningful feature extraction approaches would be conceivable. Nonetheless, the introduced approaches have clear advantages: Hand-crafted MIND features could be computed for any medical image data, independent of whether annotations or further additional data are provided. MIND features do not require separate training of a feature extraction or segmentation network, do not focus on certain anatomical foreground structures, allow multimodal input data, and lead to an image registration framework that belongs to the category of unsupervised image registration methods. On the other hand, learned semantic image features such as nnU-Net features entail a supervised training process and thereby the precondition that annotations are available. Though once trained models exist, this type of image features have the potential to

yield a registration pipeline that often performs even better than with extraction of MIND features.

6.5 Conclusion

The overall results demonstrate that the presented method is capable to overcome challenges in the field of medical image registration, such as the alignment of multimodal, inter-patient, or strongly deformed image pairs. A limitation of the self-configuring implementation of ConvexAdam is that the automatic hyperparameter selection process can be exhaustive. Depending on the number of hyperparameter selection possibilities, the size of the validation dataset, whether different types of feature extraction methods are compared, and the available hardware, this procedure can be very time-consuming. However, a particular benefit of the method is the fast inference time that ranges from 0.84 s to 5.83 s per image pair. Most importantly, the method has the advantage of being extremely versatile. Regardless of the evaluated Learn2Reg challenge task, the method is capable of achieving high-ranked registration results.

Chapter 7

Discussion

In this thesis, four approaches for medical image registration are proposed. Altogether, the experiments that were performed throughout this thesis involve a wide range of registration tasks, including the alignment of different anatomical structures in mono- and multimodal and inter- and intra-patient image data. With respect to the technical components of the proposed registration models, they mainly consist of differentiable and trainable modules, but also methods that incorporate non-trainable or non-differentiable modules are presented. All methods propose some form of decoupling within their architecture. For the trainable modules, different forms of supervision are investigated during the course of this thesis. The main contributions are summarised in Sec. 7.1. In Sec. 7.2, the findings with regard to the objectives defined in Chapter 1.1 are presented. Sec. 7.3 gives a brief overview of ongoing research in the field of deep learning-based image registration and in Sec. 7.4, a final conclusion is given.

7.1 Main Contributions

The method presented in the first methodological chapter (Chapter 3) of this thesis introduces the concept of *Deforming Autoencoders* for deformable medical image registration. Unlike the methods presented in the remaining chapters, this is a method primarily designed for groupwise image registration instead of pairwise image registration. The approach consists of an autoencoder network architecture that is end-to-end trained in an unsupervised way. A particular architectural characteristic of this autoencoder is that it involves a disentanglement of shape and appearance learning within the decoding part. In more detail, an encoder provides a latent vector that is split up in two parts and is then given to two separate decoders. One encoder, the appearance decoder, estimates a template image and the other decoder, the shape decoder, estimates a deformation field. By keeping the number of latent dimensions usable for the appearance decoder small, a universal template should ideally be predicted by the appearance decoder, while enforcing that the shape representation is modelled completely by the shape decoder. Image generation is achieved with the help of a differentiable spatial transformation step, which warps the estimated template with the predicted deformation field. The loss function used for training involves an inverse consistency constraint for robust and

plausible symmetric deformation prediction. In this thesis, the approach is applied for registration of inter-patient brain MR scans. The experiments demonstrate that the concept of deforming autoencoders is applicable for medical image registration, but the reconstruction error also suggests that the decoupling with two separate decoder parts as implemented for this thesis is not yet capable of completely enforcing disentanglement of shape and appearance estimation. The results show the benefit of the introduced inverse consistency constraint for smoother and more precise registration performance, although no diffeomorphism is induced and the generated deformation fields are still not perfectly smooth. For pairwise 3D registration, the mean Dice score is increased by 11 % (~ 6 % points) and the average number of foldings within the predicted deformation fields is decreased by 77 %.

The second methodological chapter (Chapter 4) is dedicated to *architecture design choices* for image registration networks. Various basic architecture design options for U-Net-based pairwise deformable registration learning are considered. Apart from comparing the effect of an increased number of feature channels and convolutions, the concept of not directly concatenating the fixed and the moving image for feature extraction is investigated. For this purpose, a two-stream architecture is proposed that includes partially decoupled feature extraction network blocks to address the problem that basic single-stream U-Net registration architectures struggle with the alignment of inter-patient image data with large deformations. The experiments evaluate the different architecture modifications using inter-patient abdominal CT scans with annotations for several anatomical structures. Furthermore, the results for unsupervised and label supervised training are compared. The main findings that can be derived from the experiments are that using initially separated encoder blocks for the moving and fixed images leads to improved registration results compared to single-stream architectures and that including label supervision in the training procedure is beneficial, given that an annotated training dataset is available. However, to achieve state-of-the-art results for the reported registration task, more elaborate U-Net architectures with, for example, more depth and a multi-level or multi-step registration strategy, would be necessary. Parallel to the work presented in [Siebert et al., 2021b], a two-stream pyramid architecture for multi-level registration learning is introduced in [Kang et al., 2022]. In general, coarse-to-fine approaches based on multi-level pyramids like, for example, PWC-Net [Sun et al., 2018], mlVIRNET [Hering et al., 2019], and LapIRN [Mok et al., 2020b], have proven their benefit for deep learning-based pairwise image registration in the meantime.

The third method (Chapter 5) addresses supervision for multimodal medical image registration. A self-supervised learning paradigm for rigid registration learning is introduced, which *minimises a discrepancy within registration cycles*. The employed cycles comprise a multimodal input image pair and an image that is generated by synthetically transforming the moving image. As underlying architecture, a feature extraction CNN with initially decoupled feature encoding layers, a correlation layer,

and a least squares fitting for transformation computation is used. With the introduced training procedure, multimodal image registration can be learned, or more specifically suitable feature extractors for its optimisation, without a predefined similarity metric and without balancing of multiple loss function terms with weighting parameters. While initialising the synthetic transformation is an issue that requires an elaborated strategy in order to improve detail alignment, the experiments indicate that the method can rigidly align abdominal intra-patient MR and CT scans and improve over hand-crafted metric-based losses.

The method presented in the fourth methodological chapter (Chapter 6) is an approach, which combines *coupled convex optimisation and Adam-based instance optimisation in a self-configuring registration framework*. It extracts semantic or hand-crafted features for each input image separately. Semantic feature extraction requires that nnUNet models are trained on an annotated training dataset beforehand. The extracted features are fed into a correlation layer to compute similarity costs within a displacement space volume, and a coupled convex optimisation procedure is used for efficient global regularisation. The resulting displacement field is used as the starting point for an Adam-based instance optimisation, which provides the final displacement field. The method is the only image registration approach of this thesis that contains non-differentiable modules. Overall, it requires little learning as it employs pretrained semantic feature extraction models. It is extremely versatile and can be used for a wide range of registration tasks. As this method entails various hyperparameters, an automatic hyperparameter selection with predefined parameter options and a ranking scheme is proposed, resulting in a self-configuring image registration method. The experiments demonstrate that the method is able to achieve good results for all available Learn2Reg challenge tasks in terms of registration accuracy, deformation smoothness, and runtime. Evaluated on the same image pairs like the two-stream method (Chapter 4), the proposed method achieves a Dice score of 68 % compared to the value of 44 %, which is achieved by the two-stream architecture. Inference times for alignment of one image pair range from 0.48 s (HippocampusMR) to 5.83 s (NLST) and, on average over all examined registration tasks, outperform the inference time of the comparison methods significantly.

7.2 Research Findings

The methodological contributions and findings of the four medical image registration approaches proposed in this thesis can be examined in the context of *supervision*, *decoupling*, and *versatility*. This section aims to outline and discuss the respective findings of the individual methods with respect to the research objectives introduced in Chapter 1.1. For clarity, the objectives are mentioned again at the beginning of each subsection.

7.2.1 Findings with Regard to Supervision Concepts

In this subsection, the contributions and findings regarding supervision concepts presented in this thesis with particular focus on

- how an inverse consistency constraint for unsupervised registration learning affects the robustness and plausibility of deformation fields estimated by a deep learning model,
- how registration performance differs when comparing weakly supervised and unsupervised registration learning, and
- how well a self-supervised learning strategy using cycle constraints is applicable for image registration

are outlined and discussed.

Within this thesis, different forms of supervision for training of deep learning-based image registration methods are investigated. In Chapter 3 and Chapter 4, **unsupervised learning** procedures are considered, which aim to optimise image similarities of warped moving image and fixed image and regularise the estimated deformation fields with plausibility constraints. Whereas for training of the deforming autoencoder method in Chapter 3 the absolute differences (ℓ_1 norm) of input and reconstructed image is used as a similarity metric within the loss function, the considered U-Net-based pairwise registration architectures are trained with distances of MIND features as similarity measure, which is more robust to intensity variations of input images. Nevertheless, as deforming autoencoders directly compare an input image with its reconstruction during training, it is possible to use the ℓ_1 norm within the objective function. As regularisation, the objective function applied for training of the U-Net-based registration architectures use diffusion regularisation to obtain smooth deformation fields. The objective function applied for training of deforming autoencoders includes a term that induces smoothness by means of the total variation norm of the deformation fields and penalises large deformations by considering the mean squared error between the average deformation grid and an identity mapping grid.

An **inverse consistency constraint** within the objective function is introduced for deforming autoencoders to obtain further robustness of the registration results in terms of plausibility of the estimated deformation fields. It addresses the problem that image registration is often considered as an asymmetric problem with a specific registration direction. The concept of the proposed inverse consistency constraint is that warping the reconstructed image with the forward deformation field predicted by the shape decoder should ideally result in the template image predicted by the appearance decoder. At the same time, warping the template image with the reversed deformation field should yield the reconstructed image. The proposed loss term therefore sums up the mean squared errors between the reconstructed image and the inversely transformed reconstructed

image, as well as between the template image and the inversely transformed template image. Different from pairwise image registration methods, which integrate inverse consistency by architectural concepts that include deformation field prediction with swapped moving and fixed images (e.g. CycleMorph [Kim et al., 2019]), the proposed inverse consistency constraint only requires the computation of the reversed flow field directly from the forward flow field. Experiments that compare the registration performance for models trained with and without the proposed inverse consistency constraint point out that integrating the constraint into the objective function leads to more precise (increase of 11 % of Dice overlap for 3D experiments) and plausible (reduction of foldings by 77 % for 3D experiments) image alignment. The introduced loss term could be easily adapted for training of any other deep learning-based image registration model. Especially if fast inference times are desired or if models tend to predict implausible deformation fields with foldings, it can be advantageous to integrate the inverse consistency constraint into the loss function instead of during inference via instance optimisation.

For the two-stream U-Net registration architecture that is proposed in Chapter 4, unsupervised training with MIND similarities and diffusion regularisation is compared to weakly supervised training with additional **segmentation label supervision**. The results indicate the benefit of label supervision for the investigated task of inter-patient abdominal CT scan registration in terms of registration accuracy. The resulting Dice overlap is increased by 24 % when including label supervision. The highest improvements are observed for large and medium-sized organs. However, the potential problem of label bias is not evaluated. But as MIND similarities are also included in the employed objective function, the results should nonetheless be reliable in regions without expert labels. In general, given that annotations are at hand, combining label supervision with another similarity metric within the objective function may be beneficial. The method presented in [Schumacher et al., 2022] extends the proposed unsupervised two-stream model by weak bounding box supervision instead of supervision with densely labelled training data. This approach reduces the amount of required annotation effort compared to dense label supervision, while improving Dice scores by $\sim 10\,\%$ compared to unsupervised learning. In [Ha et al., 2020], the authors propose to learn segmentation and registration jointly within a model that comprises a U-Net part to extract semantic information for the input images and a two-step warping framework for large deformation prediction. Their work points out that employing semantic labels for supervision of alignment and segmentation prediction yields improved results compared to only using label information for alignment supervision. Because employing segmentation networks to provide label features for registration guidance has been shown to produce excellent results, ConvexAdam (Chapter 6) uses segmentations predicted by nnUNets [Isensee et al., 2021] as semantic image features that serve as input for the optimisation procedure for deformation field computation.

Balancing the different loss terms within the training procedure is often difficult. Especially the experiments for the deforming autoencoders point out that finding a good balance to obtain precise and plausible registration results is an elaborate issue. The learning strategy proposed in Chapter 5 for multimodal image registration learning addresses this problem. It circumvents the search for a similarity metric for network training by minimising the discrepancy within synthetic three-way registration cycles, yielding a **self-supervised learning** procedure. For rigid multimodal image alignment, the method is able to yield results superior to hand-crafted metric-based losses. Despite its benefits, the method also has its shortcomings. The main drawback is that it requires a good initialisation of the synthetic transformation, especially if no knowledge about the displacement strength to be expected is given. Preliminary experiments on 2D image data show the potential of an incremental learning strategy for generation of the synthetic transformation to improve detail alignment within a deformable registration setting.

ConvexAdam, the method presented in Chapter 6, is the only other approach presented in this thesis that is applied for multimodal image alignment. Both multimodal methods are evaluated on an abdominal CT to MR registration task. Although a different split of the data is used (cross-validation in Chapter 5 and a predefined test dataset in Chapter 6), it is apparent that ConvexAdam, which performs *deformable* **registration guided by MIND features**, yields more precise multimodal image registration. Compared to the respective score for the unaligned evaluation dataset, the Dice overlap is increased by a factor of 1.8 when using the rigid cycle learning registration approach with two warping steps, while an increase by a factor of 2.3 is achieved when using ConvexAdam in conjunction with MIND features.

Brought into context with the inverse consistency constraint, the approaches of both chapters (Chapter 3 and Chapter 5) make use of **registration cycles for supervision**. Whereas the inverse consistency constraint minimises the discrepancy within two-way cycles consisting of forward and backward transformation, the method presented in Chapter 5 uses cycles with transformations between three images: moving image, fixed image, and a random synthetic image. The extension to three-way cycles enables the latter method to be applied within the proposed self-supervised training procedure, while the inverse consistency constraint, whose purpose is to encourage symmetric registration, needs to be applied together with at least a similarity metric to yield a reasonable supervision concept.

7.2.2 Findings with Regard to Architectural Decoupling Techniques

The contributions and findings in terms of decoupling are outlined and discussed with respect to

- how splitting up the latent space of an autoencoder architecture can be used to decouple shape and appearance representation for groupwise image registration,

- how beneficial the use of separated but shared input network blocks for feature extraction for moving and fixed input images might be for pairwise image registration compared to single-stream architectures, and
- how separated trainable feature extraction for input image pairs can be used in combination with non-trainable modules within image registration frameworks.

All architectures of the medical image registration methods presented in this thesis involve a form of decoupling to attain precise and interpretable results. At this point, reference is made to Fig. 1.1, which depicts an overview of the architectures discussed in this thesis.

Apart from the deforming autoencoders (Chapter 3), all methods make use of decoupling within their feature extraction part. Deforming autoencoders **split up the latent vector and use separate decoders for shape and appearance estimation**. Consequently, the method aims to learn meaningful representations for shape and appearance that can be used for groupwise image registration. Reconstruction is performed by warping the template image generated by the appearance decoder with the deformation field generated by the shape decoder. Although the experiments demonstrate that this form of architecture with disentangled latent space is applicable for image registration, it is also outlined that completely enforcing the shape and appearance informations to solely pass through the designated decoder parts would require further extensions of the method like parametrisation of the latent space or learning conditional templates similar to [Dalca et al., 2019b].

The other methods presented in this thesis, which are all for pairwise medical image registration, use **decoupling for feature extraction of the input images**. In Chapter 4, experiments are conducted that compare single-stream U-Net architectures to a two-stream architecture that initially uses separate network blocks to generate features for the fixed and the moving input images. For the investigated monomodal registration task, a clear benefit of this form of decoupling over single-stream processing could be observed with regard to the registration performance. It can be assumed that this type of decoupling is beneficial for many types of single-stream registration architectures.

In Chapter 5, a similar concept of architectural decoupling is used for multimodal feature learning. The Y-shaped architecture generates features for MR and CT images with separate network blocks for each modality followed by shared feature extracting CNN blocks. The decoupling is of particular importance as multimodal image data is processed and modality specific features could be extracted within the separated network blocks. Unlike the proposed two-stream U-Net architecture, this method does not concatenate the extracted features to pass them to further network layers for deformation estimation. Instead, it gives the extracted features to a correlation layer and uses a least squares fitting procedure without trainable weights for deformation estimation. In contrast to the non-trainable modules applied in Chapter 6, the correlation and

deformation computation modules of this method are differentiable and therefore allow backpropagation to train the employed feature extraction network.

The method presented in Chapter 6 performs feature extraction for each input image separately and therefore decouples this step completely. Depending on the registration task and available segmentation annotation, the method either extracts semantic label features with the help of a pretrained segmentation network or extracts hand-crafted MIND features. The strict separation without any backpropagation of feature extraction from subsequent feature alignment allows task-specific adjustments of the hyperparameters required for coupled convex optimisation and Adam-based instance optimisation to obtain overall best performance of all methods applied to the Learn2Reg challenge [Hering et al., 2022a]. Different to the other image registration methods proposed in this thesis, this method uses non-trainable and non-differentiable modules for feature alignment and deformation field generation.

7.2.3 Findings with Regard to Versatility

This subsection describes the findings related, focusing on the objectives for

- basic concepts for image registration that can easily be transferred to various deep learning-based image registration methods, and
- a pairwise image registration approach that can be applied directly to a wide variety of tasks, regardless of whether mono- or multimodal or whether intra- or inter-patient image data are to be processed.

Since this thesis presents diverse concepts for medical image registration, the question of usability and versatility arises. Each of the methods brings with it its own form of versatility. Versatility can be interpreted, on the one hand, that a *methodological concept is easily transferable to other methods* or, on the other hand, that a *method can be utilised for a variety of application tasks.*

In Chapter 3, the benefits of an **inverse consistency constraint** that penalises discrepancy between identity mapping and forward registration followed by reverse registration is outlined. In this thesis, it is included in the objective function to train deforming autoencoders. However, this constraint can be easily implemented as an additional loss term within the training procedure of various deep learning-based image registration approaches like VoxelMorph [Balakrishnan et al., 2019] that do not address the problem of asymmetric registration otherwise.

Chapter 4 introduces a U-Net-based architecture for pairwise image registration with **initially separate network blocks** for the fixed and the moving images. The idea is to improve registration performance over single-stream networks by enabling the network to initially extract meaningful features of both images separately that are concatenated and further processed within a U-Net architecture for deformation

estimation. In this thesis, the architecture in which this two-stream concept is integrated, is similar to VoxelMorph [Balakrishnan et al., 2019] and it is demonstrated that the two-stream concept improves the registration performance. As many pairwise registration architectures are based on the architecture of single-stream VoxelMorph, it can be assumed that an extension to initially two processing streams would be beneficial for various image registration methods. Considered in terms of versatility, the basic concept of decoupled initial feature extraction layers can be transferred to other pairwise image registration architectures. In the case of monomodal image registration, the trainable weights could be shared between the two processing streams.

The supervision concept of **minimising cycle discrepancies** that is presented in Chapter 5 is especially helpful for multimodal registration learning, as the search for a similarity metric is more challenging compared to monomodal registration tasks. Nevertheless, it would also be possible to apply the proposed supervision procedure to monomodal registration tasks. As indicated by the experiments, the underlying architecture consisting of a feature extracting CNN, a correlation layer, and a least squares fitting procedure could also be used in conjunction with metric supervision.

The method presented in Chapter 6 is extremely versatile in terms of applicability to various image registration tasks. With the proposed **self-configuring pipeline** including feature extraction, coupled convex discrete optimisation, and Adam-based instance optimisation, it is possible to align image pairs from a wide variety of medical image datasets. The method shows convincing results for all currently available Learn2Reg challenge datasets and is the overall best performing challenge submission. Besides the Learn2Reg challenge datasets, the proposed method has also been used with empirical hyperparameter settings for the ISBI-2022 and MICCAI-2022 BraTS-Reg challenge [Baheti et al., 2021], where it was ranked second.

7.3 Ongoing research

Most recent medical image registration approaches adopt novel deep learning techniques such as Transformer-based architectures [Chen et al., 2021, 2022; Mok et al., 2022], diffusion models [Kim et al., 2022], neural ODEs [Wu et al., 2022; Xu et al., 2021], or meta learning [Baum et al., 2023; Hoopes et al., 2021; Kanter et al., 2022; Mok et al., 2021; Park et al., 2022][Zou et al., 2022]. This section aims to summarise the key ideas and benefits of several approaches published within the last two years.

One of the currently most trending novelties for image registration is to apply Transformers within deep learning models. Incorporating vision Transformer in image registration architectures helps to capture long-range spatial correspondences with self-attention techniques. Experiments presented in [Chen et al., 2021] demonstrate that replacing the backbone of VoxelMorph with a Transformer-based architecture improves the results for brain MR image alignment. TransMorph [Chen et al., 2022] employs

the Swin Transformer [Liu et al., 2021] within a registration learning framework. The Swin Transformer is a vision Transformer that builds hierarchical feature maps with a window shifting scheme. In TransMorph, the Swin Transformer is used as an encoder to capture the spatial correspondence between moving and fixed images. A CNN decoder then provides a dense displacement field and, with the help of skip connections between encoder and decoder, the flow of localisation information is maintained. The experiments on brain MR and chest-abdomen-pelvis CT demonstrate the robustness and effectiveness of Transformers for medical image registration.

In the field of meta learning, recent learning-based image registration approaches aim to reduce user interaction. HyperMorph [Hoopes et al., 2021] successfully learns to perform hyperparameter selection by learning the influence of registration hyperparameters on deformation fields. The approach of [Mok et al., 2021] is a conditional registration method that conditions the feature maps of the registration networks on a regularisation hyperparameter. Other methods aim to adapt to dataset variations [Baum et al., 2023] or to registration tasks with datasets from new domains [Park et al., 2022]. In [Kanter et al., 2022], a meta learning-based strategy is introduced that combines affine image registration based on iterative conventional approaches with a robust trainable model for estimating good initialisations.

In [Falta et al., 2022], a deep learning-based approach is presented that emulates conventional iterative gradient-based optimisation for deformable image registration. The proposed recurrent framework employs an iterative dynamic cost sampling step and hidden states to imitate information flow during optimisation. An approach that combines a trainable feature backbone with a correlation layer and a differentiable convex optimisation related to [Heinrich et al., 2014] in an end-to-end trainable setting is proposed in [Heinrich et al., 2023]. Similar to [Siebert et al., 2022c], efficient deep learning-based feature generation is combined with a powerful conventional registration approach. The authors of [Jia et al., 2022a] present a network architecture that contains a convolutional encoder, an embedded discrete Fourier transformation to map the low dimensional representation provided by the encoder into the band-limited Fourier domain, and a parameter-free model-driven decoder to reconstruct the deformation field into the full-resolution spatial domain from its band-limited Fourier domain. This approach is very efficient with regard to inference speed while achieving state-of-the-art registration accuracy.

7.4 Conclusion

The end-to-end trainable deep learning techniques that were developed throughout the last four years of this thesis (Chapters 3-5) show their benefits for the respective reported experiments. They are, however, no longer able to compete with the very latest high performing state-of-the-art deep learning approaches in terms of registration

performance. Yet, each method brings its own findings that can contribute to the future development of deep learning-based medical image registration methods. The transferable findings regarding deep learning-based registration methods are to be considered as principles that can be adopted in more elaborate deep learning architectures. Many state-of-the art deep learning-based medical image registration methods use multi-resolution registration with image pyramids [Gunnarsson et al., 2020; Hering et al., 2019; Mok et al., 2020b, 2021], which is a technique that has not been integrated in the methods presented in this thesis. However, as outlined in Sec. 7.2.3, various concepts investigated in this thesis that are beneficial for the presented more basic deep learning architectures, might be easily adapted to current state-of-the-art methods. The most recent techniques described in Sec. 7.3 all lead to enhanced registration performance for the respective reported experiments, but still, there is to date no universally applicable deep learning-based model that is domain and task invariant and not to some extent data-driven. The approach presented in the last methodological chapter (Chapter 6) is no end-to-end trainable deep learning image registration method, as it only involves deep learning for feature extraction and not for the subsequent alignment. But with this method, an approach is presented that is applicable to various medical image registration tasks while only a small validation dataset is needed for hyperparameter selection. Due to the introduced automatic hyperparameter selection strategy, no user intervention is necessary to obtain state-of-the-art registration performance. While setting out with the hypothesis that modelling features, shape and cycle consistency in complex end-to-end frameworks would be ideal, it is to some extent an unexpected conclusion that a modular, separated training of semantic features with subsequent hyperparameter optimisation in fact performed best.

References

[Akin et al., 2016] Akin, O., Elnajjar, P., Heller, M., Jarosz, R., Erickson, B., Kirk, S., and Filippini, J. "Radiology data from the cancer genome atlas kidney renal clear cell carcinoma [TCGA-KIRC] collection". *The Cancer Imaging Archive*, 2016.

[Allassonnière et al., 2007] Allassonnière, S., Amit, Y., and Trouvé, A. "Towards a coherent statistical framework for dense deformable template estimation". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 69 (1), 2007, pp. 3–29.

[Antonelli et al., 2022] Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Kopp-Schneider, A., Landman, B. A., Litjens, G., Menze, B., Ronneberger, O., Summers, R. M., et al. "The medical segmentation decathlon". *Nature communications* 13 (1), 2022, p. 4128.

[Avants et al., 2008] Avants, B. B., Epstein, C. L., Grossman, M., and Gee, J. C. "Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain". *Medical image analysis* 12 (1), 2008, pp. 26–41.

[Baheti et al., 2021] Baheti, B., Waldmannstetter, D., Chakrabarty, S., Akbari, H., Bilello, M., Wiestler, B., Schwarting, J., Calabrese, E., Rudie, J., Abidi, S., et al. "The brain tumor sequence registration challenge: establishing correspondence between pre-operative and follow-up MRI scans of diffuse glioma patients". *arXiv preprint arXiv:2112.06979*, 2021.

[Balakrishnan et al., 2018] Balakrishnan, G., Zhao, A., Sabuncu, M. R., Guttag, J., and Dalca, A. V. "An unsupervised learning model for deformable medical image registration". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 9252–9260.

[Balakrishnan et al., 2019] Balakrishnan, G., Zhao, A., Sabuncu, M. R., Guttag, J., and Dalca, A. V. "Voxelmorph: a learning framework for deformable medical image registration". *IEEE Trans Med Imag* 38 (8), 2019, pp. 1788–1800.

[Baum et al., 2023] Baum, Z. M. C., Hu, Y., and Barratt, D. C. "Meta-Learning Initializations for Interactive Medical Image Registration". *IEEE Transactions on Medical Imaging* 42 (3), 2023, pp. 823–833.

[Blendowski et al., 2021] Blendowski, M., Hansen, L., and Heinrich, M. P. "Weakly-supervised learning of multi-modal features for regularised iterative descent in 3D image registration". *Medical Image Analysis* 67, 2021, p. 101822.

[Bône et al., 2020] Bône, A., Vernhet, P., Colliot, O., and Durrleman, S. "Learning joint shape and appearance representations with metamorphic auto-encoders". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2020, pp. 202–211.

[Cao et al., 2017] Cao, X., Yang, J., Zhang, J., Nie, D., Kim, M., Wang, Q., and Shen, D. "Deformable image registration based on similarity-steered CNN regression". In: *Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part I 20*. 2017, pp. 300–308.

[Chen et al., 2008] Chen, M., Lu, W., Chen, Q., Ruchala, K. J., and Olivera, G. H. "A simple fixed-point approach to invert a deformation field". *Medical physics* 35 (1), 2008, pp. 81–88.

[Chen et al., 2016] Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., and Abbeel, P. "Infogan: Interpretable representation learning by information maximizing generative adversarial nets". *Advances in neural information processing systems* 29, 2016.

[Che et al., 2019a] Che, T., Zheng, Y., Cong, J., Jiang, Y., Niu, Y., Jiao, W., Zhao, B., and Ding, Y. "Deep Group-Wise Registration for Multi-Spectral Images From Fundus Images". *IEEE Access* 7, 2019, pp. 27650–27661.

[Che et al., 2019b] Che, T., Zheng, Y., Sui, X., Jiang, Y., Cong, J., Jiao, W., and Zhao, B. "DGR-Net: Deep groupwise registration of multispectral images". In: *International Conference on Information Processing in Medical Imaging*. 2019, pp. 706–717.

[Chen et al., 2021] Chen, J., He, Y., Frey, E. C., Li, Y., and Du, Y. "Vit-v-net: Vision transformer for unsupervised volumetric medical image registration". *arXiv preprint arXiv:2104.06468*, 2021.

[Chen et al., 2022] Chen, J., Frey, E. C., He, Y., Segars, W. P., Li, Y., and Du, Y. "Transmorph: Transformer for unsupervised medical image registration". *Medical Image Analysis* 82, 2022, p. 102615.

[Christensen et al., 2001] Christensen, G. E. and Johnson, H. J. "Consistent image registration". *IEEE transactions on medical imaging* 20 (7), 2001, pp. 568–582.

[Christensen et al., 2006] Christensen, G. E., Geng, X., Kuhl, J. G., Bruss, J., Grabowski, T. J., Pirwani, I. A., Vannier, M. W., Allen, J. S., and Damasio, H. "Introduction to the non-rigid image registration evaluation project (NIREP)". In: *Biomedical Image Registration: Third International Workshop,*

WBIR 2006, Utrecht, The Netherlands, July 9-11, 2006. Proceedings 3. 2006, pp. 128–135.

[Clark et al., 2013] Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M., et al. "The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository". *Journal of digital imaging* 26 (6), 2013, pp. 1045–1057.

[Dalca et al., 2019a] Dalca, A. V., Balakrishnan, G., Guttag, J., and Sabuncu, M. R. "Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces". *Medical image analysis* 57, 2019, pp. 226–236.

[Dalca et al., 2019b] Dalca, A. V., Rakic, M., Guttag, J., and Sabuncu, M. R. "Learning Conditional Deformable Templates with Convolutional Networks". *NeurIPS: Neural Information Processing Systems*, 2019.

[Datteri et al., 2012] Datteri, R. D. and Dawant, B. M. "Automatic detection of the magnitude and spatial location of error in non-rigid registration". In: *International Workshop on Biomedical Image Registration*. 2012, pp. 21–30.

[Dosovitskiy et al., 2015] Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., Van Der Smagt, P., Cremers, D., and Brox, T. "Flownet: Learning optical flow with convolutional networks". In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 2758–2766.

[Ehrhardt et al., 2015] Ehrhardt, J., Schmidt-Richberg, A., Werner, R., and Handels, H. "Variational registration: A flexible open-source itk toolbox for nonrigid image registration". In: *Bildverarbeitung für die Medizin 2015: Algorithmen-Systeme-Anwendungen. Proceedings des Workshops vom 15. bis 17. März 2015 in Lübeck*. 2015, pp. 209–214.

[Eppenhof et al., 2018] Eppenhof, K. A., Lafarge, M. W., Moeskops, P., Veta, M., and Pluim, J. P. "Deformable image registration using convolutional neural networks". In: *Medical Imaging 2018: Image Processing*. Vol. 10574. 2018, 105740S.

[Eppenhof et al., 2019] Eppenhof, K. A. J. and Pluim, J. P. W. "Pulmonary CT Registration Through Supervised Learning With Convolutional Neural Networks". *IEEE Trans Med Imag* 38 (5), 2019, pp. 1097–1105.

[Eppenhof et al., 2020] Eppenhof, K. A. J., Lafarge, M. W., Veta, M., and Pluim, J. P. W. "Progressively Trained Convolutional Neural Networks for Deformable Image Registration". *IEEE Trans Med Imag* 39 (5), 2020, pp. 1594–1604.

[Erickson et al., 2016] Erickson, B., Kirk, S., Lee, Y., Bathe, O., Kearns, M., Gerdes, C., and Lemmerman, J. "Radiology Data from The Cancer Genome Atlas Liver Hepatocellular Carcinoma [TCGA-LIHC] collection". *The Cancer Imaging Archive*, 2016.

[Estienne et al., 2020a] Estienne, T., Lerousseau, M., Vakalopoulou, M., Alvarez Andres, E., Battistella, E., Carré, A., Chandra, S., Christodoulidis, S., Sahasrabudhe, M., Sun, R., et al. "Deep learning-based concurrent brain registration and tumor segmentation". *Frontiers in computational neuroscience*, 2020, p. 17.

[Estienne et al., 2020b] Estienne, T., Vakalopoulou, M., Battistella, E., Carré, A., Henry, T., Lerousseau, M., Robert, C., Paragios, N., and Deutsch, E. "Deep learning based registration using spatial gradients and noisy segmentation labels". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2020, pp. 87–93.

[Falta et al., 2022] Falta, F., Hansen, L., and Heinrich, M. P. "Learning Iterative Optimisation for Deformable Image Registration of Lung CT with Recurrent Convolutional Networks". In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VI*. 2022, pp. 301–309.

[Fu et al., 2020] Fu, Y., Lei, Y., Wang, T., Curran, W. J., Liu, T., and Yang, X. "Deep learning in medical image registration: a review". *Physics in Medicine & Biology* 65 (20), 2020, 20TR01.

[Gass et al., 2015] Gass, T., Székely, G., and Goksel, O. "Consistency-based rectification of nonrigid registrations". *Journal of Medical Imaging* 2 (1), 2015, p. 014005.

[Geng et al., 2009] Geng, X., Christensen, G. E., Gu, H., Ross, T. J., and Yang, Y. "Implicit reference-based group-wise image registration and its application to structural and functional MRI". *Neuroimage* 47 (4), 2009, pp. 1341–1351.

[Goodfellow et al., 2016] Goodfellow, I., Bengio, Y., and Courville, A. *Deep learning*. MIT press, 2016.

[Gunnarsson et al., 2020] Gunnarsson, N., Sjölund, J., and Schön, T. B. "Learning a Deformable Registration Pyramid". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2020, pp. 80–86.

[Ha et al., 2020] Ha, I. Y., Wilms, M., and Heinrich, M. "Semantically guided large deformation estimation with deep networks". *Sensors* 20 (5), 2020, p. 1392.

[Haase et al., 2020] Haase, R., Heldmann, S., and Lellmann, J. "Deformable Groupwise Image Registration using Low-Rank and Sparse Decomposition". *arXiv preprint arXiv:2001.03509*, 2020.

[Hagenah et al., 2019] Hagenah, J., Mehdi, M., and Ernst, F. "Generating healthy aortic root geometries from ultrasound images of the individual pathological morphology using deep convolutional autoencoders". In: *2019 Computing in Cardiology (CinC)*. 2019, Page–1.

[Häger et al., 2020] Häger, S., Heldmann, S., Hering, A., Kuckertz, S., and Lange, A. "Variable fraunhofer MEVIS RegLib comprehensively applied to Learn2Reg

challenge". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2020, pp. 74–79.

[Hansen et al., 2021] Hansen, L. and Heinrich, M. P. "GraphRegNet: Deep graph regularisation networks on sparse keypoints for dense registration of 3D lung CTs". *IEEE Transactions on Medical Imaging* 40 (9), 2021, pp. 2246–2257.

[Handels, 2000] Handels, H. *Medizinische Bildverarbeitung*. Springer, 2000.

[He et al., 2012] He, K., Sun, J., and Tang, X. "Guided image filtering". *IEEE transactions on pattern analysis and machine intelligence* 35 (6), 2012, pp. 1397–1409.

[He et al., 2016] He, K., Zhang, X., Ren, S., and Sun, J. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[Heinrich et al., 2012] Heinrich, M. P., Jenkinson, M., Bhushan, M., Matin, T., Gleeson, F. V., Brady, M., and Schnabel, J. A. "MIND: Modality independent neighbourhood descriptor for multi-modal deformable registration". *Medical image analysis* 16 (7), 2012, pp. 1423–1435.

[Heinrich et al., 2013] Heinrich, M. P., Jenkinson, M., Papiez, B. W., Brady, M., and Schnabel, J. A. "Towards realtime multimodal fusion for image-guided interventions using self-similarities". In: *International conference on medical image computing and computer-assisted intervention*. 2013, pp. 187–194.

[Heinrich et al., 2014] Heinrich, M. P., Papież, B. W., Schnabel, J. A., and Handels, H. "Non-parametric discrete registration with convex optimisation". In: *Biomedical Image Registration: 6th International Workshop, WBIR 2014, London, UK, July 7-8, 2014. Proceedings 6*. 2014, pp. 51–61.

[Heinrich et al., 2015] Heinrich, M. P., Handels, H., and Simpson, I. J. "Estimating large lung motion in COPD patients by symmetric regularised correspondence fields". In: *International conference on medical image computing and computer-assisted intervention*. 2015, pp. 338–345.

[Heinrich et al., 2020] Heinrich, M. P. and Hansen, L. "Highly accurate and memory efficient unsupervised learning-based discrete CT registration using 2.5 D displacement search". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2020, pp. 190–200.

[Heinrich et al., 2022] Heinrich, M. P. and Hansen, L. "Voxelmorph++ going beyond the cranial vault with keypoint supervision and multi-channel instance optimisation". In: *Biomedical Image Registration: 10th International Workshop, WBIR 2022, Munich, Germany, July 10–12, 2022, Proceedings*. 2022, pp. 85–95.

[Heinrich et al., 2023] Heinrich, M. P., Siebert, H., Graf, L., Mischkewitz, S., and Hansen, L. "Robust and Realtime Large Deformation Ultrasound Registration

Using End-to-End Differentiable Displacement Optimisation". *Sensors* 23 (6), 2023, p. 2876.

[Heinrich, 2019] Heinrich, M. P. "Closing the gap between deep and conventional image registration using probabilistic dense displacement networks". In: *Proc. MICCAI*. 2019, pp. 50–58.

[Hering et al., 2019] Hering, A., Ginneken, B., and Heldmann, S. "mlvirnet: Multilevel variational image registration network". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2019, pp. 257–265.

[Hering et al., 2020] Hering, A., Murphy, K., and Ginneken, B. *Learn2Reg Challenge: CT Lung Registration - Training Data*. 2020.

[Hering et al., 2021] Hering, A., Lange, A., Heldmann, S., Häger, S., and Kuckertz, S. "Fraunhofer MEVIS Image Registration Solutions for the Learn2Reg 2021 Challenge". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2021, pp. 147–152.

[Hering et al., 2022a] Hering, A., Hansen, L., Mok, T. C. W., Chung, A. C. S., Siebert, H., Häger, S., Lange, A., Kuckertz, S., Heldmann, S., Shao, W., Vesal, S., Rusu, M., Sonn, G., Estienne, T., Vakalopoulou, M., Han, L., Huang, Y., Yap, P.-T., Brudfors, M., Balbastre, Y., Joutard, S., Modat, M., Lifshitz, G., Raviv, D., Lv, J., Li, Q., Jaouen, V., Visvikis, D., Fourcade, C., Rubeaux, M., Pan, W., Xu, Z., Jian, B., De Benetti, F., Wodzinski, M., Gunnarsson, N., Sjölund, J., Grzech, D., Qiu, H., Li, Z., Thorley, A., Duan, J., Großbröhmer, C., Hoopes, A., Reinertsen, I., Xiao, Y., Landman, B., Huo, Y., Murphy, K., Lessmann, N., Van Ginneken, B., Dalca, A. V., and Heinrich, M. P. "Learn2Reg: comprehensive multi-task medical image registration challenge, dataset and evaluation in the era of deep learning". *IEEE Transactions on Medical Imaging*, 2022.

[Hering et al., 2022b] Hering, A., Schnabel, J., Zhang, M., Ferrante, E., Heinrich, M., and Rueckert, D. *Biomedical Image Registration: 10th International Workshop, WBIR 2022, Munich, Germany, July 10–12, 2022, Proceedings*. Vol. 13386. Springer Nature, 2022.

[Higgins et al., 2017] Higgins, I., Matthey, L., Pal, A., Burgess, C. P., Glorot, X., Botvinick, M. M., Mohamed, S., and Lerchner, A. "beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework". In: *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

[Hill et al., 2001] Hill, D. L., Batchelor, P. G., Holden, M., and Hawkes, D. J. "Medical image registration". *Physics in medicine & biology* 46 (3), 2001, R1.

[Hoopes et al., 2021] Hoopes, A., Hoffmann, M., Fischl, B., Guttag, J., and Dalca, A. V. "Hypermorph: Amortized hyperparameter learning for image registration". In:

Information Processing in Medical Imaging: 27th International Conference, IPMI 2021, Virtual Event, June 28–June 30, 2021, Proceedings 27. 2021, pp. 3–17.

[Hu et al., 2018a] Hu, Y., Modat, M., Gibson, E., Ghavami, N., Bonmati, E., Moore, C. M., Emberton, M., Noble, J. A., Barratt, D. C., and Vercauteren, T. "Label-driven weakly-supervised learning for multimodal deformable image registration". In: *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. 2018, pp. 1070–1074.

[Hu et al., 2018b] Hu, Y., Modat, M., Gibson, E., Li, W., Ghavami, N., Bonmati, E., Wang, G., Bandula, S., Moore, C. M., Emberton, M., et al. "Weakly-supervised convolutional neural networks for multimodal image registration". *Medical image analysis* 49, 2018, pp. 1–13.

[Hu et al., 2019] Hu, X., Kang, M., Huang, W., Scott, M. R., Wiest, R., and Reyes, M. "Dual-Stream Pyramid Registration Network". In: *Proc. MICCAI*. 2019, pp. 382–390.

[Ilg et al., 2017] Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., and Brox, T. "Flownet 2.0: Evolution of optical flow estimation with deep networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 2462–2470.

[Isensee et al., 2021] Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J., and Maier-Hein, K. H. "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation". *Nature methods* 18 (2), 2021, pp. 203–211.

[Jaderberg et al., 2015] Jaderberg, M., Simonyan, K., Zisserman, A., et al. "Spatial transformer networks". In: *Advances in neural information processing systems*. 2015, pp. 2017–2025.

[Jia et al., 2021] Jia, X., Thorley, A., Chen, W., Qiu, H., Shen, L., Styles, I. B., Chang, H. J., Leonardis, A., De Marvao, A., O'Regan, D. P., et al. "Learning a model-driven variational network for deformable image registration". *IEEE Transactions on Medical Imaging* 41 (1), 2021, pp. 199–212.

[Jia et al., 2022a] Jia, X., Bartlett, J., Chen, W., Song, S., Zhang, T., Cheng, X., Lu, W., Qiu, Z., and Duan, J. "Fourier-Net: Fast Image Registration with Band-limited Deformation". *arXiv preprint arXiv:2211.16342*, 2022.

[Jia et al., 2022b] Jia, X., Bartlett, J., Zhang, T., Lu, W., Qiu, Z., and Duan, J. "U-Net vs Transformer: Is U-Net Outdated in Medical Image Registration?" *arXiv preprint arXiv:2208.04939*, 2022.

[Kang et al., 2018] Kang, S. K., Seo, S., Shin, S. A., Byun, M. S., Lee, D. Y., Kim, Y. K., Lee, D. S., and Lee, J. S. "Adaptive template generation for amyloid PET using a deep learning approach". *Human brain mapping* 39 (9), 2018, pp. 3769–3778.

[Kang et al., 2022] Kang, M., Hu, X., Huang, W., Scott, M. R., and Reyes, M. "Dual-stream pyramid registration network". *Medical Image Analysis* 78, 2022, p. 102379.

[Kanter et al., 2022] Kanter, F. and Lellmann, J. "A Flexible Meta Learning Model for Image Registration". In: *International Conference on Medical Imaging with Deep Learning*. 2022, pp. 638–652.

[Kim et al., 2019] Kim, B., Kim, J., Lee, J.-G., Kim, D. H., Park, S. H., and Ye, J. C. "Unsupervised deformable image registration using cycle-consistent cnn". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2019, pp. 166–174.

[Kim et al., 2022] Kim, B., Han, I., and Ye, J. C. "DiffuseMorph: Unsupervised Deformable Image Registration Using Diffusion Model". In: *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXI*. 2022, pp. 347–364.

[Kingma et al., 2014] Kingma, D. P. and Ba, J. "Adam: A method for stochastic optimization". *arXiv preprint arXiv:1412.6980*, 2014.

[Klein et al., 2007] Klein, S., Staring, M., and Pluim, J. P. "Evaluation of optimization methods for nonrigid medical image registration using mutual information and B-splines". *IEEE transactions on image processing* 16 (12), 2007, pp. 2879–2890.

[Klein et al., 2012] Klein, A. and Tourville, J. "101 labeled brain images and a consistent human cortical labeling protocol". *Frontiers in neuroscience* 6, 2012, p. 171.

[Komodakis et al., 2018] Komodakis, N. and Gidaris, S. "Unsupervised representation learning by predicting image rotations". In: *International Conference on Learning Representations (ICLR)*. 2018.

[Krebs et al., 2017] Krebs, J., Mansi, T., Delingette, H., Zhang, L., Ghesu, F. C., Miao, S., Maier, A. K., Ayache, N., Liao, R., and Kamen, A. "Robust non-rigid registration through agent-based action learning". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2017, pp. 344–352.

[Kumar et al., 2017] Kumar, A., Sattigeri, P., and Balakrishnan, A. "Variational inference of disentangled latent concepts from unlabeled observations". *arXiv preprint arXiv:1711.00848*, 2017.

[LeCun et al., 1989] LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. "Backpropagation applied to handwritten zip code recognition". *Neural computation* 1 (4), 1989, pp. 541–551.

[Li et al., 2020] Li, X., Lin, C., Li, R., Wang, C., and Guerin, F. "Latent space factorisation and manipulation via matrix subspace projection". In: *International Conference on Machine Learning*. 2020, pp. 5916–5926.

[Li et al., 2022] Li, Y.-x., Tang, H., Wang, W., Zhang, X.-f., and Qu, H. "Dual attention network for unsupervised medical image registration based on VoxelMorph". *Scientific Reports* 12 (1), 2022, p. 16250.

[Linehan et al., 2016] Linehan, M., Gautam, R., Kirk, S., Lee, Y., Roche, C., Bonaccio, E., and Jarosz, R. "Radiology data from the cancer genome atlas cervical kidney renal papillary cell carcinoma [KIRP] collection". *The Cancer Imaging Archive*, 2016.

[Liu et al., 2021] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. "Swin transformer: Hierarchical vision transformer using shifted windows". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 10012–10022.

[Lv et al., 2022] Lv, J., Wang, Z., Shi, H., Zhang, H., Wang, S., Wang, Y., and Li, Q. "Joint progressive and coarse-to-fine registration of brain MRI via deformation field integration and non-rigid feature fusion". *IEEE Transactions on Medical Imaging* 41 (10), 2022, pp. 2788–2802.

[Maes et al., 1997] Maes, F., Collignon, A., Vandermeulen, D., Marchal, G., and Suetens, P. "Multimodality image registration by maximization of mutual information". *IEEE transactions on Medical Imaging* 16 (2), 1997, pp. 187–198.

[Marcus et al., 2007] Marcus, D. S., Wang, T. H., Parker, J., Csernansky, J. G., Morris, J. C., and Buckner, R. L. "Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults". *Journal of cognitive neuroscience* 19 (9), 2007, pp. 1498–1507.

[Misra et al., 2020] Misra, I. and Maaten, L. v. d. "Self-supervised learning of pretext-invariant representations". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 6707–6717.

[Modat et al., 2014] Modat, M., Cash, D. M., Daga, P., Winston, G. P., Duncan, J. S., and Ourselin, S. "Global image registration using a symmetric block-matching approach". *Journal of Medical Imaging* 1 (2), 2014, p. 024003.

[Mok et al., 2020a] Mok, T. C. and Chung, A. "Fast symmetric diffeomorphic image registration with convolutional neural networks". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 4644–4653.

[Mok et al., 2020b] Mok, T. C. and Chung, A. C. "Large Deformation Diffeomorphic Image Registration with Laplacian Pyramid Networks". In: *Proc. MICCAI*. 2020, pp. 211–221.

[Mok et al., 2021] Mok, T. C. and Chung, A. C. "Conditional deformable image registration with convolutional neural network". In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part IV 24*. 2021, pp. 35–45.

[Mok et al., 2022] Mok, T. C. and Chung, A. “Affine medical image registration with coarse-to-fine vision transformer”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 20835–20844.

[Mouches et al., 2021] Mouches, P., Wilms, M., Rajashekar, D., Langner, S., and Forkert, N. “Unifying brain age prediction and age-conditioned template generation with a deterministic autoencoder”. In: *Medical Imaging with Deep Learning*. 2021, pp. 497–506.

[Papenberg et al., 2006] Papenberg, N., Bruhn, A., Brox, T., Didas, S., and Weickert, J. “Highly accurate optic flow computation with theoretically justified warping”. *International Journal of Computer Vision* 67, 2006, pp. 141–158.

[Park et al., 2022] Park, H., Lee, G. M., Kim, S., Ryu, G. H., Jeong, A., Sagong, M., and Park, S. H. “A meta-learning approach for medical image registration”. In: *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*. 2022, pp. 1–5.

[Rohé et al., 2017] Rohé, M.-M., Datar, M., Heimann, T., Sermesant, M., and Pennec, X. “SVF-Net: learning deformable image registration using shape matching”. In: *Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part I 20*. 2017, pp. 266–274.

[Rohlfing, 2011] Rohlfing, T. “Image similarity and tissue overlaps as surrogates for image registration accuracy: widely used but unreliable”. *IEEE transactions on medical imaging* 31 (2), 2011, pp. 153–163.

[Ronneberger et al., 2015] Ronneberger, O., Fischer, P., and Brox, T. “U-net: Convolutional networks for biomedical image segmentation”. In: *Proc. MICCAI*. 2015, pp. 234–241.

[Rueckert et al., 1999] Rueckert, D., Sonoda, L. I., Hayes, C., Hill, D. L., Leach, M. O., and Hawkes, D. J. “Nonrigid registration using free-form deformations: application to breast MR images”. *IEEE transactions on medical imaging* 18 (8), 1999, pp. 712–721.

[Sandkühler et al., 2018] Sandkühler, R., Jud, C., Andermatt, S., and Cattin, P. C. “AirLab: autograd image registration laboratory”. *arXiv preprint arXiv:1806.09907*, 2018.

[Schumacher et al., 2022] Schumacher, M., Siebert, H., Genz, A., Bade, R., and Heinrich, M. “Learning-based three-dimensional registration with weak bounding box supervision”. *Journal of Medical Imaging* 9 (4), 2022, p. 044001.

[Shu et al., 2018] Shu, Z., Sahasrabudhe, M., Alp Guler, R., Samaras, D., Paragios, N., and Kokkinos, I. “Deforming autoencoders: Unsupervised disentangling of shape and appearance”. In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 650–665.

[Siebert et al., 2020] Siebert, H. and Heinrich, M. P. “Deep Groupwise Registration of MRI Using Deforming Autoencoders”. In: *Bildverarbeitung für die Medizin 2020 –BVM 2020*. Informatik Aktuell. Springer, 2020, pp. 236–241.

[Siebert et al., 2021a] Siebert, H., Hansen, L., and Heinrich, M. P. “Learning a Metric without Supervision: Multimodal Registration using Synthetic Cycle Discrepancy”. In: *International Conference on Medical Imaging with Deep Learning –Extended Abstract Track –MIDL 2021*. 2021.

[Siebert et al., 2021b] Siebert, H., Hansen, L., and Heinrich, M. P. “Architecture Matters: Evaluating Design Choices for Deep Learning Registration Networks”. In: *Bildverarbeitung für die Medizin 2021 –BVM 2021*. Informatik Aktuell. Springer, 2021, pp. 111–116.

[Siebert et al., 2021c] Siebert, H., Rajamani, K. T., and Heinrich, M. P. “Learning inverse consistent 3D groupwise registration with deforming autoencoders”. In: *Medical Imaging 2021: Image Processing*. Vol. 11596. 2021, pp. 89–95.

[Siebert et al., 2022a] Siebert, H., Hansen, L., and Heinrich, M. P. “Fast 3D Registration with Accurate Optimisation and Little Learning for Learn2Reg 2021”. In: *Biomedical Image Registration, Domain Generalisation and Out-of-Distribution Analysis –MICCAI 2021 Challenges*. 2022, pp. 174–179.

[Siebert et al., 2022b] Siebert, H., Hansen, L., and Heinrich, M. P. “Learning a Metric for Multimodal Medical Image Registration without Supervision Based on Cycle Constraints”. *Sensors* 22 (3), 2022.

[Siebert et al., 2022c] Siebert, H. and Heinrich, M. P. “Learn to Fuse Input Features for Large-Deformation Registration with Differentiable Convex-Discrete Optimisation”. In: *Biomedical Image Registration - 10th International Workshop, WBIR 2022, Munich, Germany, July 10-12, 2022, Proceedings*. Vol. 13386. Lecture Notes in Computer Science. Springer, 2022, pp. 119–123.

[Simonovsky et al., 2016] Simonovsky, M., Gutiérrez-Becker, B., Mateus, D., Navab, N., and Komodakis, N. “A deep metric for multimodal registration”. In: *International conference on medical image computing and computer-assisted intervention*. 2016, pp. 10–18.

[Sokooti et al., 2017] Sokooti, H., De Vos, B., Berendsen, F., Lelieveldt, B. P., Išgum, I., and Staring, M. “Nonrigid image registration using multi-scale 3D convolutional neural networks”. In: *Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part I 20*. 2017, pp. 232–239.

[Sotiras et al., 2013] Sotiras, A., Davatzikos, C., and Paragios, N. “Deformable medical image registration: A survey”. *IEEE transactions on medical imaging* 32 (7), 2013, pp. 1153–1190.

[Sudre et al., 2017] Sudre, C. H., Li, W., Vercauteren, T., Ourselin, S., and Jorge Cardoso, M. "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations". In: *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3*. 2017, pp. 240–248.

[Sun et al., 2018] Sun, D., Yang, X., Liu, M.-Y., and Kautz, J. "Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 8934–8943.

[Team et al., 2011] Team, N. L. S. T. R. et al. "The national lung screening trial: overview and study design". *Radiology* 258 (1), 2011, p. 243.

[Team, 2011] Team, N. L. S. T. R. "Reduced lung-cancer mortality with low-dose computed tomographic screening". *New England Journal of Medicine* 365 (5), 2011, pp. 395–409.

[Tewari et al., 2022] Tewari, A., Pan, X., Fried, O., Agrawala, M., Theobalt, C., et al. "Disentangled3D: Learning a 3D Generative Model with Disentangled Geometry and Appearance from Monocular Images". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 1516–1525.

[Uzunova et al., 2017] Uzunova, H., Wilms, M., Handels, H., and Ehrhardt, J. "Training CNNs for image registration from few samples with model-based data augmentation". In: *Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part I 20*. 2017, pp. 223–231.

[Uzunova et al., 2021] Uzunova, H., Handels, H., and Ehrhardt, J. "Guided Filter Regularization for Improved Disentanglement of Shape and Appearance in Diffeomorphic Autoencoders". In: *Medical Imaging with Deep Learning*. 2021, pp. 774–786.

[Vanschoren, 2019] Vanschoren, J. "Meta-learning". *Automated machine learning: methods, systems, challenges*, 2019, pp. 35–61.

[Viola et al., 1995] Viola, P. and Wells, W. M. "Alignment by maximization of mutual information". In: *Proceedings of IEEE International Conference on Computer Vision*. 1995, pp. 16–23.

[Vos et al., 2019] Vos, B. D., Berendsen, F. F., Viergever, M. A., Sokooti, H., Staring, M., and Išgum, I. "A deep learning framework for unsupervised affine and deformable image registration". *Med Image Anal* 52, 2019, pp. 128–143.

[Wang et al., 2019a] Wang, X., Jabri, A., and Efros, A. A. "Learning correspondence from the cycle-consistency of time". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 2566–2576.

[Wang et al., 2019b] Wang, Y. and Solomon, J. M. "PRNet: Self-Supervised Learning for Partial-to-Partial Registration". In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc., 2019.

[Werner et al., 2014] Werner, R., Schmidt-Richberg, A., Handels, H., and Ehrhardt, J. "Estimation of lung motion fields in 4D CT data by variational non-linear intensity-based registration: A comparison and evaluation study". *Physics in Medicine & Biology* 59 (15), 2014, p. 4247.

[Wiputra et al., 2020] Wiputra, H., Chan, W. X., Foo, Y. Y., Ho, S., and Yap, C. H. "Cardiac motion estimation from medical images: a regularisation framework applied on pairwise image registration displacement fields". *Scientific reports* 10 (1), 2020, p. 18510.

[Wu et al., 2022] Wu, Y., Jiahao, T. Z., Wang, J., Yushkevich, P. A., Hsieh, M. A., and Gee, J. C. "Nodeo: A neural ordinary differential equation based optimization framework for deformable image registration". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 20804–20813.

[Xiao et al., n.d.] Xiao, Y., Fortin, M., Unsgård, G., Rivaz, H., and Reinertsen, I. *EASY-RESECT*.

[Xiao et al., 2017] Xiao, Y., Fortin, M., Unsgård, G., Rivaz, H., and Reinertsen, I. "REtroSpective Evaluation of Cerebral Tumors (RESECT): A clinical database of pre-operative MRI and intra-operative ultrasound in low-grade glioma surgeries". *Medical physics* 44 (7), 2017, pp. 3875–3882.

[Xiao et al., 2019] Xiao, Y., Rivaz, H., Chabanas, M., Fortin, M., Machado, I., Ou, Y., Heinrich, M. P., Schnabel, J. A., Zhong, X., Maier, A., et al. "Evaluation of MRI to ultrasound registration methods for brain shift correction: the CuRIOUS2018 challenge". *IEEE transactions on medical imaging* 39 (3), 2019, pp. 777–786.

[Xing et al., 2020] Xing, X., Gao, R., Han, T., Zhu, S.-C., and Wu, Y. N. "Deformable generator networks: unsupervised disentanglement of appearance and geometry". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[Xu et al., 2016] Xu, Z., Lee, C. P., Heinrich, M. P., Modat, M., Rueckert, D., Ourselin, S., Abramson, R. G., and Landman, B. A. "Evaluation of six registration methods for the human abdomen on clinically acquired CT". *IEEE Trans Biomed Eng* 63 (8), 2016, pp. 1563–1572.

[Xu et al., 2020] Xu, Z., Luo, J., Yan, J., Pulya, R., Li, X., Wells, W., and Jagadeesan, J. "Adversarial uni-and multi-modal stream networks for multimodal image registration". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2020, pp. 222–232.

[Xu et al., 2021] Xu, J., Chen, E. Z., Chen, X., Chen, T., and Sun, S. "Multi-scale neural odes for 3d medical image registration". In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part IV 24*. 2021, pp. 213–223.

[Yang et al., 2017] Yang, X., Kwitt, R., Styner, M., and Niethammer, M. "Quicksilver: Fast predictive image registration–a deep learning approach". *NeuroImage* 158, 2017, pp. 378–396.

[Zhao et al., 2019a] Zhao, S., Dong, Y., Chang, E. I., Xu, Y., et al. "Recursive cascaded networks for unsupervised medical image registration". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 10600–10610.

[Zhao et al., 2019b] Zhao, S., Lau, T., Luo, J., Eric, I., Chang, C., and Xu, Y. "Unsupervised 3d end-to-end medical image registration with volume tweening network". *IEEE journal of biomedical and health informatics* 24 (5), 2019, pp. 1394–1404.

[Zhu et al., 2017] Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. "Unpaired image-to-image translation using cycle-consistent adversarial networks". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2223–2232.

[Zhu et al., 2022] Zhu, Y. and Lu, S. "Swin-voxelmorph: A symmetric unsupervised learning model for deformable medical image registration using swin transformer". In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VI*. 2022, pp. 78–87.

[Zou et al., 2022] Zou, J., Gao, B., Song, Y., and Qin, J. "A review of deep learning-based deformable medical image registration". *Frontiers in Oncology* 12, 2022.

List of Publications

The following publication list contains journal articles, conference papers, and short papers published or submitted during the work on this dissertation. An asterisk (*) indicates co-first authorship.

Journal articles as first author

- Siebert, H., Hansen, L., and Heinrich, M. P. "Learning a Metric for Multimodal Medical Image Registration without Supervision Based on Cycle Constraints". *Sensors* 22 (3), 2022.

in preparation:

- Siebert, H., Hansen, L., Großbröhmer, C., and Heinrich, M. P. "ConvexAdam: Coupled Convex Discrete Optimisation with Adam-based Instance Optimisation for Multitask Medical Image Registration".

Conference papers as first author

- Siebert, H., Hansen, L., and Heinrich, M. P. "Fast 3D Registration with Accurate Optimisation and Little Learning for Learn2Reg 2021". In: *Biomedical Image Registration, Domain Generalisation and Out-of-Distribution Analysis –MICCAI 2021 Challenges*. 2022, pp. 174–179.
- Siebert, H., Hansen, L., and Heinrich, M. P. "Architecture Matters: Evaluating Design Choices for Deep Learning Registration Networks". In: *Bildverarbeitung für die Medizin 2021 –BVM 2021*. Informatik Aktuell. Springer, 2021, pp. 111–116.
- Siebert, H., Rajamani, K. T., and Heinrich, M. P. "Learning inverse consistent 3D groupwise registration with deforming autoencoders". In: *Medical Imaging 2021: Image Processing*. Vol. 11596. 2021, pp. 89–95.
- Siebert, H. and Heinrich, M. P. "Deep Groupwise Registration of MRI Using Deforming Autoencoders". In: *Bildverarbeitung für die Medizin 2020 –BVM 2020*. Informatik Aktuell. Springer, 2020, pp. 236–241.

Short papers as first author

- Siebert, H. and Heinrich, M. P. "Learn to Fuse Input Features for Large-Deformation Registration with Differentiable Convex-Discrete Optimisation". In: *Biomedical Image Registration - 10th International Workshop, WBIR 2022, Munich, Germany, July 10-12, 2022, Proceedings*. Vol. 13386. Lecture Notes in Computer Science. Springer, 2022, pp. 119–123.
- Siebert*, H., Hansen*, L., and Heinrich, M. P. "Learning a Metric without Supervision: Multimodal Registration using Synthetic Cycle Discrepancy". In: *International Conference on Medical Imaging with Deep Learning –Extended Abstract Track –MIDL 2021*. 2021.

Journal articles as co-author

- Heinrich, M. P., Siebert, H., Graf, L., Mischkewitz, S., and Hansen, L. "Robust and Realtime Large Deformation Ultrasound Registration Using End-to-End Differentiable Displacement Optimisation". *Sensors* 23 (6), 2023, p. 2876.
- Rajamani, K. T., Rani, P., Siebert, H., ElagiriRamalingam, R., and Heinrich, M. P. "Attention-augmented U-Net (AA-U-Net) for semantic segmentation". *Signal, Image and Video Processing*, 2022, pp. 1–9.
- Schumacher, M., Siebert, H., Genz, A., Bade, R., and Heinrich, M. "Learning-based three-dimensional registration with weak bounding box supervision". *Journal of Medical Imaging* 9 (4), 2022, p. 044001.
- Rajamani, K. T., Siebert, H., and Heinrich, M. P. "Dynamic deformable attention network (DDANet) for COVID-19 lesions semantic segmentation". *J. Biomed. Informatics* 119, 2021, p. 103816.

Conference papers as co-author

- Graf, L. F., Siebert, H., Mischkewitz, S., Keuth, R., and Heinrich, M. P. "Highly accurate deep registration networks for large deformation estimation in compression ultrasound". In: *Medical Imaging 2023: Image Processing*. Vol. 12464. SPIE, 2023, p. 1246426.
- Gupta, P., Siebert, H., Heinrich, M. P., and Rajamani, K. T. "DA-AR-Net: an attentive activation based Deformable auto-encoder for group-wise registration". In: *Medical Imaging 2021: Image Processing*. Vol. 11596. 2021, pp. 181–188.

Short paper as co-author

- Schumacher, M., Siebert, H., Bade, R., Genz, A., and Heinrich, M. P. "Weak Bounding Box Supervision for Image Registration Networks". In: *Biomedical Image Registration - 10th International Workshop, WBIR 2022, Munich, Germany, July 10-12, 2022, Proceedings*. Vol. 13386. Lecture Notes in Computer Science. Springer, 2022, pp. 215–219.