Robust medical image registration and motion modeling based on machine learning

Julian Krebs

UNIVERSITÉ CÔTE D'AZUR

ÉCOLE DOCTORALE
SCIENCES ET TECHNOLOGIES DE L'INFORMATION ET DE LA COMMUNICATION

THÈSE DE DOCTORAT

Le recalage robuste d'images médicales et la modélisation du mouvement basée sur l'apprentissage profond

Robust Medical Image Registration and Motion Modeling based on Machine Learning

Julian KREBS

INRIA, Équipe EPIONE

Thèse dirigée par Hervé DELINGETTE et co-dirigée par Nicholas AYACHE

Soutenue le 15 juin 2020

Présentée en vue de l'obtention du grade de DOCTEUR EN AUTOMATIQUE, TRAITEMENT DU SIGNAL ET DES IMAGES de l'UNIVERSITÉ CÔTE D'AZUR.

Devant le jury composé de :

Daniel RUECKERT	Imperial College London	Rapporteur
Ivana IŠGUM	University of Amsterdam	Rapporteur
Tom VERCAUTEREN	King's College London	Examinateur
Nikos PARAGIOS	Centrale Supelec, Paris	Président
Hervé DELINGETTE	Inria Sophia Antipolis	Directeur de thèse
Nicholas AYACHE	Inria Sophia Antipolis	Co-directeur de thèse
Tommaso MANSI	Siemens Healthineers, Princeton	Co-encadrant
Hiroshi ASHIKAGA	Johns Hopkins University, Baltimore	Invité

Inria

SIEMENS Healthineers

Abstract

This thesis presents new computational tools for quantifying deformations and motion of anatomical structures from medical images as required by a large variety of clinical applications. Generic deformable registration tools are presented that enable deformation analysis useful for improving diagnosis, prognosis and therapy guidance. These tools were built by combining state-of-the-art medical image analysis methods with cutting-edge machine learning methods.

First, we focus on difficult inter-subject registration problems. By learning from given deformation examples, we propose a novel agent-based optimization scheme inspired by deep reinforcement learning where a statistical deformation model is explored in a trial-and-error fashion showing improved registration accuracy.

Second, we develop a diffeomorphic deformation model that allows for accurate multi-scale registration and deformation analysis by learning a low-dimensional representation of intra-subject deformations. The unsupervised method uses a latent variable model in form of a conditional variational autoencoder (CVAE) for learning a probabilistic deformation encoding that is useful for the simulation, classification and comparison of deformations.

Third, we propose a probabilistic motion model derived from image sequences of moving organs. This generative model embeds motion in a structured latent space, the motion matrix, which enables the consistent tracking of structures and various analysis tasks. For instance, it leads to the simulation and interpolation of realistic motion patterns allowing for faster data acquisition and data augmentation.

Finally, we demonstrate the importance of the developed tools in a clinical application where the motion model is used for disease prognosis and therapy planning. It is shown that the survival risk for heart failure patients can be predicted from the discriminative motion matrix with a higher accuracy compared to classical image-derived risk factors.

Keywords: medical imaging, image registration, motion modeling, artificial intelligence, machine learning, variational autoencoder, sudden cardiac death.

Résumé

Cette thèse présente de nouveaux outils informatiques pour quantifier les déformations et le mouvement de structures anatomiques à partir d'images médicales dans le cadre d'une grande variété d'applications cliniques. Des outils génériques de recalage déformable sont présentés qui permettent l'analyse de la déformation de tissus anatomiques pour améliorer le diagnostic, le pronostic et la thérapie. Ces outils combinent des méthodes avancées d'analyse d'images médicales avec des méthodes d'apprentissage automatique performantes.

Dans un premier temps, nous nous concentrons sur les problèmes de recalages inter-sujets difficiles. En apprenant à partir d'exemples de déformation donnés, nous proposons un nouveau schéma d'optimisation basé sur un agent inspiré de l'apprentissage par renforcement profond dans lequel un modèle de déformation statistique est exploré de manière itérative montrant une précision améliorée de recalage.

Dans un second temps, nous développons un modèle de déformation difféomorphe qui permet un recalage multi-échelle précis et une analyse de déformation en apprenant une représentation de faible dimension des déformations intra-sujet. La méthode non supervisée utilise un modèle de variable latente sous la forme d'un autoencodeur variationnel conditionnel (CVAE) pour apprendre une représentation probabiliste des déformations qui est utile pour la simulation, la classification et la comparaison des déformations.

Troisièmement, nous proposons un modèle de mouvement probabiliste dérivé de séquences d'images d'organes en mouvement. Ce modèle génératif décrit le mouvement dans un espace latent structuré, la matrice de mouvement, qui permet le suivi cohérent des structures ainsi que l'analyse du mouvement. Ainsi cette approche permet la simulation et l'interpolation de modèles de mouvement réalistes conduisant à une acquisition et une augmentation des données plus rapides.

Enfin, nous démontrons l'intérêt des outils développés dans une application clinique où le modèle de mouvement est utilisé pour le pronostic de maladies et la planification de thérapies. Il est démontré que le risque de survie des patients souffrant d'insuffisance cardiaque peut être prédit à partir de la matrice de mouvement discriminant avec une précision supérieure par rapport aux facteurs de risque classiques dérivés de l'image.

Mots-clés: imagerie médicale, recalage d'images, modélisation du mouvement, intelligence artificielle, apprentissage profond, autoencodeur variationnel, mort cardiaque subite.

Acknowledgement

I would like to thank my thesis advisers Hervé Delingette, Tommaso Mansi and Nicholas Ayache whose guidance was invaluable for the success of this thesis. Thank you for always listening to my questions and constructive discussions when I encountered difficulties. Your passion, motivation and scientific expertise of the field are and will always be very inspiring to me. Tommaso, I am very happy and grateful you selected me for this opportunity. Thank you for your enthusiastic support during and also before my Ph.D. thesis, for always teaching me, motivating me and for pushing me when things seemed hopeless. Hervé, it has been a great pleasure to work with you. Thank you for the countless scientific and non-scientific discussions and for sharing your stimulating scientific ingenuity and deep insights. I would also like to thank Nicholas for your valuable advises, for accepting me in the Asclepios/Epione team with its outstanding and friendly research environment and for sending me to conferences and summer schools around the world.

I am extremely grateful to Prof. Ivana Išgum and Prof. Daniel Rueckert for spending their valuable time on reviewing this manuscript and providing constructive feedback and insights. I am also thankful to Prof. Tom Vercauteren and Prof. Nikos Paragios for accepting to be members of my jury. Thank you, it has been a great honor for me to have such an outstanding jury.

A sincere thank to Dr. Hiroshi Ashikaga and Dr. Katherine Wu, with whom I worked on the clinical project. Hiroshi, I really enjoyed working with you. Thank you for helping me to understand what is important from the clinical point of view.

I thank Dorin Comaniciu for supporting this project and giving me the chance to do four internships in his outstanding team in Princeton. I thank all my colleagues and interns at Siemens Healthineers who made my internships successful, enjoyable and unforgettable. I especially thank Shun Miao, Boris Mailhé, Bin Lou, Li Zhang, Rui Liao, Yue Zhang, Florin Ghesu, Guillaume Chabin, Serkan Çimen, Sasa Grbic, Dong Yang, Ingmar Voigt, Ali Kamen and Sebastien Piat for their help and support.

Moreover, I would like to thank the other members of the Epione team. Thank you to Xavier Pennec, Maxime Sermesant and Marco Lorenzi for stimulating discussions, your ideas and precious insights. I am grateful to Isabelle Strobant for all the help

in organizing my numerous relocations and travels. I am also grateful to the shared time and experiences with colleagues and friends in but also outside of the lab. In particular, I like to thank Roch Mollero for taking care of me when I first arrived in the lab and housing me until I found an apartment. I also thank Raphaël Sivera, Wen Wei, Shuman Jia, Marc-Michel Rohé, Loïc Cadour, Thomas Demarcy, Sophie Giffard-Roisin, Pawel Mlynarski, Nicholas Cedilnik, Manon Muntanter, Yann Thanwerdas, Luigi Antelmi, Clement Abi-Nader, Tania Bacoyannis, Jaume Banus-Cobo, Zihao Wang, Benoit Audelan, Santiago Silva-Rincon, Sara Garbarino, Nicolas Guigui, Gaëtan Desrues, Buntheng Ly, Yingyu Yang and many more.

Finally, I would like to thank my family and friends that have not been mentioned yet. Thank you for your love, support and encouragement: my mother, my father and my sister. You always believed in me and supported me during all the years of education.

Thank you to my grandfather and grandmother, who stood behind me and have taught me the important lessons of life that no book but only life experience can give. You will be always in my heart and memory. This thesis is dedicated to you.

Sofia, thank you for being my biggest supporter, for understanding me and being there despite the sometimes difficult moments and the cumbersome long-distance relationship. Thank you for helping me to grow and making me happy every day.

Thank you!

Financial Support

This work was partially funded by Siemens Healthineers, Digital Technology & Innovation, Princeton, NJ 08540 USA.

Furthermore, this work has been supported by the French government, through the 3IA Côte d'Azur Investments in the Future project managed by the National Research Agency (ANR) with the reference number ANR-19-P3IA-0002 and the grant AAP Santé 06 2017-260 DGA-DSH, and by the Inria Sophia Antipolis - Méditerranée, "NEF" computation cluster.

Contents

1 Introduction

Contents

In this chapter, we introduce the clinical context and objectives of the thesis. First, we discuss current needs in clinical routine that motivate our work. We show the steps from patient to diagnosis, to prognosis and therapy. Then, we explain and give concrete examples to these steps that raise the main objectives of this thesis:

Can we automatically derive relevant information from medical images to learn accurate deformation and motion models that can be helpful for diagnosis, prognosis and therapy planning?

1.1 Clinical Context

The typical clinical workflow consists of four main stages: diagnosis, prognosis, therapy planning and therapy. Starting point is the patient who feels sick and arrives at the hospital to get cured. The first main task of the physician is to collect relevant information such as symptoms, patient health history, vital signs, lab parameters and medical images. The physician analyzes all these information with the help of multi-dimensional analysis tools to form a diagnosis. After a potential disease has been identified, a prognosis is made to evaluate the future impact of the disease such as duration and likely outcomes. In the next step, therapy planning is done by taking into account all potential treatment measures given the previous information. Finally, the therapy which has the best chances of curing the patient or reducing the symptoms is carried out. These four stages will repeat until the endpoint is reached where the patient is healthy again (or has died).

Medical images are increasingly important to help clinicians at all four stages of the clinical workflow for a large variety of applications in healthcare. The importance and additional insights clinicians gain from medical images is also shown in the fact that more than one billion radiological exams are performed worldwide every year [Krupinski, 2010]. In the US, medical imaging accounts for over 40% of all hospital procedures

reported in the discharge report leading to a market volume at the size of $56 billion which is 0.5% of the GDP (in 2004) [Krupinski, 2010].

Today, many different medical image acquisition devices and protocols are used in healthcare. An overview of these systems and the physics behind is given in [Webb, 2003] and more recently in [Maier, 2018].

With the rise of medical imaging systems comes the need for maximizing the insights clinicians can obtain from images. Automatic computational image analysis has a high value for diagnosis, prognosis and therapy as it can extract information in a fast and objective fashion that cannot be measured directly. This not only overcomes the problem of a high inter-rater variability but also allows the processing and comparison of a large amount of data which would be too time-consuming to be done manually. Thus, the last decades have seen large advances and progress in the automatic analysis of medical images using computer vision techniques.

One major need is to deal with multiple images of the same or overlapping body regions. Deformable registration and motion analysis tools aim at finding corresponding locations in images to define the mapping from one to the other image(s). This mapping describes the image deformations and is essential for the comparison, integration or fusion of medical images which can support diagnosis, prognosis and therapy of various diseases. In general, multiple medical images are acquired to get more accurate information for a better understanding and examination of the human body. The images can be taken from various fields of view from the same image modality (mono-modal). For example in x-ray imaging, multiple images help to not oversee structures and abnormalities that are invisible in projections from certain directions. Sometimes, images are acquired from different modalities (multi-modal) to benefit from the advantages of each imaging principles, such as ultrasound and magnetic resonance images (MRI) of the same organ (e.g. prostate [Puech, 2013; Marks, 2013]). Another example is fusing anatomical and functional features provided by Computer Tomography (CT) scans and Positron Emission Tomography (PET) scans respectively.

During a surgery, images need to be registered to other images taken before the surgery from the same or a different modality in order to guide the surgeons. In another example of registration, one would like to compare images of a patient with a reference image with known information (such as structure boundaries, anatomical landmarks or disease) or with images from a population of patients suffering from the same disease for prognostic or therapeutic reasons. This type of registration is known as inter-subject registration. Fusion of mono- or multi-modal and on the other hand of intra- or inter-subject information is required in numerous clinical applications such as in the investigation of organ function and pathologies. The medium of such analysis tasks are typically image deformations in the regions around the organ of interest or pathology. Using two images to extract such

deformations is known as pairwise registration. Having multiple images to register is referred to as motion or group-wise registration.

Sequences of images that are acquired to track structural changes over time are of particular clinical interest to study. In longitudinal studies, for example, registered images acquired over longer time intervals allow to measure disease progression (e.g. neurodegenerative diseases) or tumor growth. Sequences of images are also acquired to analyze the motion of moving organs or to compensate for motion that introduces artifacts such as respiratory motion. One organ of particular interest for studying motion is the heart [Zerhouni, 1988] as cardiovascular diseases are one of the most common disease groups around the world. An impaired heart function such as in heart failure (HF) patients can cause large implications and even lead to the death of a patient. Motion analysis can be very useful in HF as for example certain heart motion features (e.g. ventricular ejection fraction) that are computed manually from images are able to predict outcomes such as sudden cardiac death (SCD) [Adabag, 2012].

In conclusion, all these applications of deformable registration are examples where the integration or fusion of two or more images is an essential task for improving one or many of the four stages of the clinical routine: diagnosis, prognosis, therapy planning and therapy.

1.2 Objectives and Organization of the Thesis

In this thesis, we present tools based on artificial intelligence for the study of deformations between two images and motion from a time series of images. These tools allow for robust image registration but also aim to improve the estimation of motion indices (such as ejection fraction or cardiac strain). This can help to directly guide the diagnosis, prognosis or therapy of diseases, not only but especially for dynamic organs. Given the context above, we first focus on the development of a robust registration tool for difficult registration scenarios by learning from existing examples of deformations.

Then, we study probabilistic deformation and motion models derived from a large database of images which capture population-specific representations of those deformation characteristics. The interest for such learned models is multiple as they allow to quantify, simulate and compare deformation and motion patterns of different patients. For example, this could support diagnosis by detecting similar patients with known diagnosis. Furthermore, predicting the disease progression in a patient could be used for therapy planning such as for survival risk prediction for cardiac diseases. In particular, we investigate the following research questions in the remaining chapters of this thesis:

- In many registration problems such as for inter-subject registration, a large variability in appearance and large deformations increase the difficulty for successful registration. Can we learn a robust registration algorithm from given examples that explores the solution space in small steps by trial-and-error to register images more accurately? (Chapter 3)

- Often, deformable registration is used for subsequent analysis tasks supporting diagnosis and prognosis. Can we learn a deformation model from images that inherently contains knowledge of physiological deformation patterns allowing for analysis tasks such as disease classification or simulation? (Chapter 4)

- Beyond pairwise registration, can we obtain a probabilistic motion model which is useful for consistent tracking of structures and motion simulation? Can we use the model to reconstruct motion from missing data? (Chapter 5)

- Having a compact motion model learned from images without supervision, does it capture discriminative factors that are useful for predicting disease outcomes? For example, can we predict the survival risk of heart failure patients? (Chapter 6)

The thesis is organized in the following way in accordance with the mentioned research questions:

In **Chapter 2**, the technical background of this thesis is discussed. We introduce a state-of-the-art of registration and motion methods including recent deep learning based approaches for deformable registration.

In **Chapter 3**, we investigate how a decision-making agent could help in difficult organ-specific deformable registration problems. An artificial agent is trained to solve an inter-subject registration task by exploring the parametric space of a statistical deformation model built from training data. Since it is difficult to extract trustworthy ground-truth deformation fields, we also present a training scheme with a large number of synthetically deformed image pairs requiring only a small number of real inter-subject deformations. The proposed method has been evaluated on the difficult task of inter-subject prostate MR registration to solve motion compensation or atlas-based segmentation problems in prostate diagnosis. The method showed state-of-the-art registration accuracy in terms of structure overlaps and distance measures. The chapter was presented at MICCAI 2017, Quebec City, Canada [Krebs, 2017].

In **Chapter 4**, we propose to learn a low-dimensional probabilistic deformation model from data which can be used for registration and the analysis of deformations. The latent variable model maps similar deformations close to each other in an encoding space. It

enables to compare deformations, generate normal or pathological deformations for any new image or to transport deformations from one image pair to any other image. Additionally, our framework is diffeomorphic and provides multi-scale velocity field estimations. We have applied our framework on cardiac intra-subject MR registration and demonstrate state-of-the-art registration accuracy, regularity and the model's potentials for disease clustering, deformation simulation and transport. The chapter is published in the journal IEEE TMI [Krebs, 2019b] and is based on the previous conference presentation at Deep Learning in Medical Image Analysis DLMIA (in conjunction with MICCAI 2018, Granada, Spain) [Krebs, 2018].

In **Chapter 5**, we extend our pairwise deformation model to a probabilistic latent motion model learned from a sequence of images for spatio-temporal registration problems. Our model encodes motion in a low-dimensional probabilistic space – the motion matrix – which enables various motion analysis tasks such as simulation and interpolation of realistic motion patterns allowing for faster data acquisition and data augmentation. Furthermore, the motion matrix allows to transport deformations from one subject to another simulating for example a pathological motion in a healthy subject without the need of inter-subject registration. The diffeomorphic motion model was analyzed by using cardiac cine-MRI showing state-of-the-art registration regularity and accuracy. Furthermore, motion simulation and interpolation are demonstrated. The chapter is based on the previous conference presentation at Statistical Atlases and Computational Models of the Heart STACOM (in conjunction with MICCAI 2019, Shenzhen, China) [Krebs, 2020c] and has been submitted to IEEE TMI [Krebs, 2020b].

In **Chapter 6**, we present a learning-based method for personalized risk and survival prediction based on our motion model. We use the 4 chamber-view cine-MRI of a patient cohort suffering from heart failure to build a motion fingerprint, the motion matrix. We demonstrate the discriminative power of this compact representation by predicting risk scores from the fingerprint for disease outcomes. We show that such an image-derived risk score is a more predictive feature for HF endpoints such as hospitalization and sudden cardiac death than any relevant clinical factors. Based on the preliminary material presented in this chapter, a clinical journal submission is in preparation.

In **Chapter 7**, the main contributions of this thesis are summarized. Finally, potential future work and perspectives are discussed.

1.3 Publications and Awards

The described contributions led to the following peer-reviewed publications, patent applications and awards.

Journal Articles

- [Krebs, 2019b] **J. Krebs**, H. Delingette, B. Mailhé, N. Ayache, and T. Mansi. Learning a probabilistic model for diffeomorphic registration. *In IEEE Transactions on Medical Imaging, 38.9, 2019, pp. 2165-2176. (Selected as featured article on the front-page of* `ieee-tmi.org`*)*

- [Krebs, 2020b] **J. Krebs**, H. Delingette, N. Ayache, and T. Mansi. Learning a Generative Motion Model from Image Sequences based on a Latent Motion Matrix. *Submitted to IEEE Transactions on Medical Imaging.*

- Risk Prediction for Heart Failure Outcomes from Cardiac Motion Features. *In preparation for submission to a clinical journal.*

Conference Papers

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

- [Krebs, 2018] **J. Krebs**, T. Mansi, B. Mailhé, N. Ayache, and H. Delingette. Unsupervised Probabilistic Deformation Modeling for Robust Diffeomorphic Registration. *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Springer, 2018, pp. 101–109.*

- [Krebs, 2020c] **J. Krebs**, T. Mansi, N. Ayache, and H. Delingette. Probabilistic Motion Modeling from Medical Image Sequences: Application to Cardiac Cine-MR. *Statistical Atlases and Computational Models of the Heart. Multi-Sequence CMR Segmentation, CRT-EPiggy and LV Full Quantification Challenges. Springer, 2020, pp. 176–185.*

- [Bacoyannis, 2019] T. Bacoyannis, **J. Krebs**, N. Cedilnik, H. Cochet, M. Sermesant. Deep Learning Formulation of ECGI for Data-driven Integration of Spatiotemporal Correlations and Imaging Information. *International Conference on Functional Imaging and Modeling of the Heart. Springer, 2019, pp. 20-28.*

Patent Applications

- [Krebs, 2020a] **J. Krebs** and T. Mansi. Method and System for Deep Motion Model Learning in Medical Images. *U.S. Patent Application No. 16/131,465, 2020*.

- [Krebs, 2019a] **J. Krebs**, H. Delingette, N. Ayache, T. Mansi and S. Miao. Medical Imaging Diffeomorphic Registration based on Machine Learning. *U.S. Patent Application No. 16/233,174, 2019*.

- [Liao, 2017a] R. Liao, S. Miao, P. de Tournemire, **J. Krebs**, L. Zhang, B. Georgescu, S. Grbic, F. C. Ghesu, V. K. Singh, D. Xu, T. Mansi, A. Kamen, and D. Comaniciu. Method and System for Image Registration Using an Intelligent Artificial Agent. *U.S. Patent Application No. 15/587,094, 2017*.

Awards

- The *Best Oral Presentation Award* was awarded at the Statistical Atlases and Computational Models of the Heart STACOM (workshop in conjunction with MICCAI) 2019, Shenzhen, China. For the paper and talk: Probabilistic Motion Modeling from Medical Image Sequences: Application to Cardiac Cine-MR.

- An excellence prize *Prix d'excellence 2019* was awarded from the University Côte d'Azur, Nice, France. For the paper: Probabilistic Motion Modeling from Medical Image Sequences: Application to Cardiac Cine-MR.

2 Deformable Registration in Medical Image Analysis

Contents

2.1 Introduction

As shown before, the integration or fusion of medical images is essential for many diagnostic and interventional tasks. Therefore, research groups have been investigating deformable registration and motion modeling in great detail over the past 30 years. A tremendous number of methods and innovations have been proposed since then. However, the task of non-rigid registration is still mostly considered as an unsolved problem [ElGamal, 2016]. Classifications and reviews of traditional deformable registration algorithms can be found in [Modersitzki, 2004; Sotiras, 2013; Oliveira, 2014; ElGamal, 2016]. Recently, over the past 3-4 years many deep learning-based (DL) approaches have been proposed for image registration. Specific review papers aim to summarize the contributions in this new group of algorithms [Haskins, 2020; Fu, 2019; Boveiri, 2020]. In their recent paper, Boveiri et al. [Boveiri, 2020] counted 80 contributions in DL-based image registration, combining rigid and non-rigid registration. In the remainder of this chapter, we aim to summarize and draw connections between both, traditional and DL-based registration. First, the general methodology for registration and motion modeling algorithms is introduced before we focus on the state-of-the-art of DL-based image registration.

2.2 Registration Algorithms

Registration is referred to as finding the spatial correspondences between two images where one is the moving M and the other the fixed image F. In order to be registered to F, the moving image M is deformed by applying a spatial transformation, the deformation field ϕ: $M \circ \phi$ where $\circ$ denotes the warping functionality. The deformation field is defined by the sum of identity transform and displacement vector field u: $\phi(x) = x + u(x), x \in \Omega$ for every position x in the image domain Ω. The registration process is illustrated in Fig.2.1. Typically, an optimization problem is solved in order to find the optimal deformation field $\phi \in \mathcal{T}$ within a set of possible transformations $\mathcal{T}$ which best aligns M to the fixed image F. Traditionally, one seeks to minimize an objective function of the following form:

$$\arg\min_{\phi \in \mathcal{T}} \mathcal{D}(F, M \circ \phi) + \mathcal{R}(\phi), \tag{2.1}$$

where $\mathcal{D}$ is a dissimilarity (or similarity) metric which measures how well the fixed and the deformed moving image are aligned and $\mathcal{R}$ denotes a regularizer enforcing pre-defined transformation properties such as the desired level of transformation smoothness. Due to the ill-posed nature of the high-dimensional registration problem, the deformation field ϕ needs to be regularized in order to obtain plausible transformations [Sotiras, 2013]. Many different metrics have been proposed for both terms as shown below. Most image registration algorithms consist of three parts: a deformation model determining the set of allowed transformations $\mathcal{T}$, an objective function with suitable dissimilarity $\mathcal{D}$ and regularization $\mathcal{R}$ metrics and an appropriate optimization strategy to find its minimum [Sotiras, 2013]. The choice for these elements is highly dependent on the registration problem to be solved. Some deformation models, dissimilarity and regularization metrics might be better suited for mono-modal than for multi-modal registration. On the other hand, intra-subject registration may require different models than inter-subject problems. Typically, the optimization problem is solved by iterative gradient descent, derivative-free optimizers or by statistical, machine-learning based strategies.

2.2.1 Similarity Metrics

One can distinguish 2 main types of dissimilarity metrics. The first type, geometric methods, are based on the matching of corresponding features such as landmarks placed at anatomical meaningful locations. The difficulty hereby lies in the robust detection of landmarks. One way to automatically obtain landmarks is the SIFT algorithm and its variants [Juan, 2007]. Because of the need for extrapolating the deformation field between sparse landmarks and therefore resulting in a decrease in accuracy, landmark-based similarity metrics have lost popularity [Sotiras, 2013]. However, with the rise of DL-based algorithms, they have gained popularity again due to the fact that in learning-

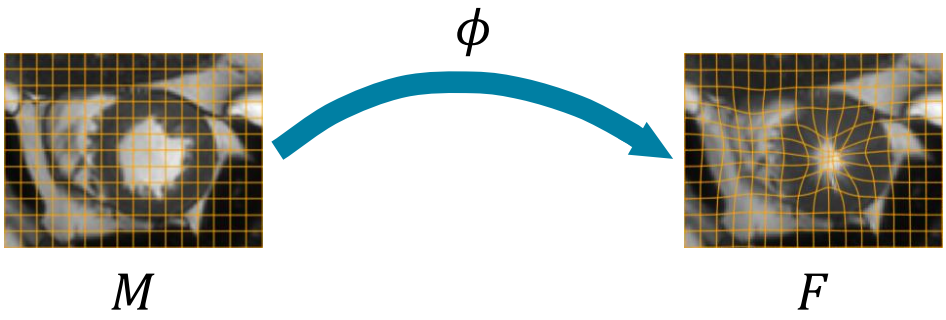


Fig. 2.1: The registration process: the moving image M is matched to the fixed image F by applying the deformation field ϕ. (MR image origin [Bernard, 2018])

based algorithms, landmarks could be used as a support during training while not requiring them at test time (cf. 2.4.3).

The second type of dissimilarity metrics relies on intensity-based quantities such as sum-of-squared or absolute differences (SSD or SAD), cross-correlation (CC) or mutual information (MI). The choice depends on the assumed relation between the signal intensities. In mono-modal registration for example, the noise assumption and the assumed correspondence between intensities dictate the choice. SSD assumes Gaussian noise while CC assumes a linear relation between intensities. In multi-modal registration, these metrics would not be a good choice as the same structures may have very different intensities in images from different modalities. That is why information theoretic approaches have been proposed for multi-modal registration. The most popular metric is MI [Viola, 1997; Maes, 1997] as it assumes a non-parametric statistical relationship between image intensities. However, its generality can turn into drawbacks that have been tried to tackle in numerous works as discussed in Sotiras et al. [Sotiras, 2013]. Besides SSD as one of the most commonly used similarity metrics for mono-modal registration, local cross-correlation (LCC) has been applied successfully due to its implicit estimation of the local affine scaling parameters as a good trade-off between SSD and MI [Lorenzi, 2013]. In this work, LCC is defined as:

$$\mathcal{D}_{\text{LCC}}(F, M \circ \phi) = \int_{\Omega} \frac{\overline{F(M \circ \phi)}}{\sqrt{\overline{F^2}} \cdot \sqrt{\overline{(M \circ \phi)^2}}} \tag{2.2}$$

with the local mean images $\overline{F}$ obtained from a mean filter with kernel size k.

Besides geometric and intensity-based dissimilarity metrics, many approaches build on hybrid models that combine both criteria [Sotiras, 2013]. Most algorithms in the group of weakly supervised DL-based methods fall into this hybrid category and are discussed in 2.4.3.

2.2.2 Regularization and Deformation Models

The choices for a suitable deformation model and regularization metric determine the degrees of freedom (DoF) and complexity of the estimated deformation. The deformation model can limit the set of possible transformations $\mathcal{T}_\theta$ by parameterizing the transformations with parameters θ. These parameterizations can take very different forms and can range from a very small number of parameters (DoF), forming simple or very restricted deformation models to high-dimensional models including thousands or millions of parameters θ. Often no parameterization is used and the space of all dense deformation fields belongs to $\mathcal{T}$. However, the more DoF a deformation model has the more the computational complexity rises and the need for a suitable regularization metric becomes necessary to obtain a well-posed problem [Sotiras, 2013].

Interpolation-based Models

The number of parameters θ can be as small as 6 in the case of 3D rigid registration (3 rotation and 3 translation parameters). Affine registration adds another 3 scaling parameters. But in the case of deformable registration, the dimensionality of θ rises typically to thousands or millions. To keep the number reasonably low and constrain computational complexity, interpolation-based transformation models are commonly used. These models are for example based on radial basis functions [Yang, 2011b; Bookstein, 1991], elastic-body splines [Davis, 1997] or, most commonly, free-from deformations (FFD, [Rueckert, 1999; Schnabel, 2001; Wang, 2007]) where only displacements of sparse control points need to be predicted while the dense deformation field is obtained using interpolation.

Physically-inspired Models

In contrast to interpolation-based methods, many methods are derived from physical models [Sotiras, 2013]. In most of these models, the deformation model allows to estimate the full number of possible parameters determined by all values in the deformation field. However, depending on the underlying physical assumptions on how the image is allowed to deform, the estimated deformations are regularized. Typically, physical models are elastic- [Davatzikos, 1997; Pennec, 2005], fluid- [Christensen, 1996] or diffusion-based [Thirion, 1998; Fischer, 2002; Vercauteren, 2007a]. Diffusion-based models are based on the fact that the Gaussian kernel is the Green's function of the diffusion equation. Under this assumption, non-parametric registration regularization can be efficiently applied using Gaussian filtering of the deformation field [Thirion, 1998].

Statistical Deformation Models

Another category of deformation models consists of statistically-constrained models [Sotiras, 2013]. Statistical deformation models (SDM) have the power to reduce the dimensionality of deformations tremendously allowing for a simpler subsequent deformation analysis. However, a statistical model needs to be trained from an existing database whereby it is limited to the observations in this training set. Before the era of DL-based registration, the size of such databases were typically relatively small due to limited computational powers. A broadly applied statistical dimensionality reduction method is principal component analysis (PCA) which has been used to learn an SDM from FFDs [Rueckert, 2003]. In active shape models, the shape variability is learned from annotated points by using PCA [Cootes, 1995]. PCA has been also used in a generative manner by generating intermediate images through sampling along the PCA axes. By doing so, the registration process can be initialized for instance by projecting the moving image to the closest target image [Tang, 2009]. Similarly, Kim et al. [Kim, 2012] estimated the intermediate target image by using support vector regression.

Diffeomorphims and other Deformation Constraints

In addition, to the presented deformation models and regularization energies, constraints on the transformations have been applied to obtain special properties that are important in medical image analysis problems. Among others, such properties are for example inverse consistency, deformation symmetry, and diffeomorphisms [Sotiras, 2013]. Since deformation are not inverse consistent in general, symmetric algorithms enforce symmetry by optimizing the objective function either 2 times in both directions (by exchanging moving and fixed image) or by constructing symmetry in the objective function, for example by registration to the midpoint between both images (cf. e.g. [Vercauteren, 2008; Lorenzi, 2013]). Diffeomorphisms are topology-preserving and invertible transformations which makes them suitable for many medical registration problems in which foldings are physically implausible [Vercauteren, 2009]. Popular parameterizations of diffeomorphisms include the Large Deformation Diffeomorphic Metric Mapping (LDDMM) [Beg, 2005; Cao, 2005; Zhang, 2015], a symmetric normalization approach [Avants, 2008] or stationary velocity fields (SVF) [Arsigny, 2006; Vercauteren, 2009; Lorenzi, 2013]. SVFs provide an efficient formulation of diffeomorphisms while still maintaining the desirable properties of time-varying LDDMMs. An SVF is not able to capture all possible diffeomorphisms, however, in practice, SVFs are often chosen due to their computational efficiency. SVFs are described as the exponential map of the velocity field v: $\phi = \exp(v)$ which can be efficiently computed by the *scaling and squaring* algorithm [Arsigny, 2006].

2.3 Motion: Adding the Temporal Dimension

Estimating the deformations within a sequence of images is highly related to pairwise registration – the mapping between 2 images. Consistent temporal registration is useful for tracking moving structures or organs, for motion compensation and for detecting pathological motion patterns. Traditionally, one can separate proposed approaches for motion estimation by physically-motivated or interpolation-based and biomechanically- or biophysically-inspired motion models. In principle, the former group extends interpolation-based or physically-motivated methods for pairwise registration by an additional temporal dimension t denoted as 2D+t or 3D+t registration. Most approaches are based on FFDs due to their efficiency where applications range from intra-subject motion estimation [LedesmaCarbayo, 2005; Vandemeulebroucke, 2011; De Craene, 2012], to inter-subject sequence registration [Perperidis, 2005; Peyrat, 2010] and group-wise registration with the purpose of defining a reference frame [Metz, 2011]. Another example for a spatio-temporal physically-motivated model (besides [Peyrat, 2010]) computes cardiac strain from image sequences [Mansi, 2011].

On the other hand, biophysical models are exploiting anatomical and physiological knowledge. Many models apply finite element methods (FEMs) for different organs and applications, for instance for tumor growth modeling, breast imaging or the prostate and its surrounding [Bharatha, 2001]. Also, a biomechanical model was used to generate synthetic training data for learning a statistical model of the prostate [Mohamed, 2002]. Electromechanical models also exist in cardiac imaging where motion analysis can help in diagnosis and therapy planning of many diseases [Sermesant, 2008].

2.4 Deep Learning-based Registration

The main difference between *classical* and learning-based, especially DL-based, registration is the transition from relying only on one pair of images to exploiting a large database of image pairs. Introducing this tremendous amount of data, the optimization strategy is mostly shifted to a training phase in order to retrieve rich implicit prior knowledge that allows to register a new image pair in almost real-time. This speed-up is regarded as one of the major benefits of using DL-based registration.

In general, a neural network is a function approximator which is parameterized by a large number of parameters ω, the network weights [Goodfellow, 2016]. Applied to image registration, the deformation field ϕ can be obtained by a simple evaluation of such a trained function f_ω that takes the image pair (F, M) as input:

$$\phi = f_\omega(F, M). \tag{2.3}$$

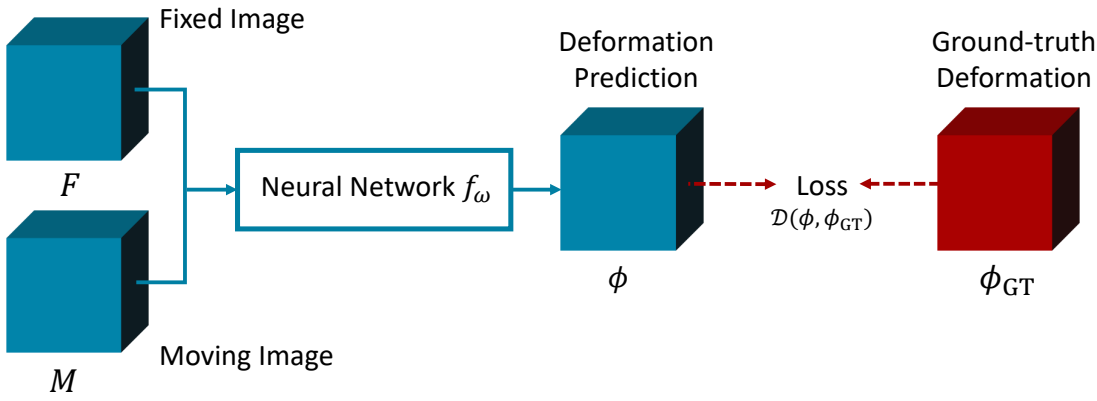


Fig. 2.2: The supervised end-to-end registration model. During training, the known deformation field ϕ_{GT} of an image pair (F, M) is regressed (red arrows).

In order to select the optimal network parameters ω^*, the neural network is trained with respect to an objective function – the loss function. In DL-based registration, one can differentiate 3 classes of approaches [Haskins, 2020] on how to choose the loss function in order to learn the parameterized registration function f_ω: Supervised, unsupervised and weakly supervised approaches. Hereby, supervision refers to the fact that extra information such as ground-truth deformation fields or labels are required during training (but typically not during testing). The different classes of approaches are discussed in the following.

2.4.1 Supervised

Supervised DL-based registration approaches aim at learning a similarity metric between the two images by providing a *ground-truth* deformation field ϕ_{GT}. In this case, the learning objective turns into a regression problem of the following form:

$$\omega^* = \arg\min_{\omega} \mathbb{E}_{p(F,M,\phi_{GT})} \left[\mathcal{D}(f_\omega(F, M), \phi_{GT}) \right], \tag{2.4}$$

where $p(F, M, \phi_{GT})$ is the empirical data distribution of image pairs and *ground-truth* deformation and $\mathcal{D}$ describes a distance metric such as SSD or CC. The idea of regressing deformation fields directly, originates from optical flow estimations in the computer vision community, where large datasets with ground-truth flow fields exist [Dosovitskiy, 2015; Weinzaepfel, 2013]. Supervised approaches can be further differentiated as end-to-end or non-end-to-end depending on whether the learned similarity metric is used in a *classical* registration algorithm or directly applied for registration.

Non-End-to-End

As one of the first DL-based approaches, [Wu, 2015] learned application-specific features that were used in traditional registration methods instead of manually extracted features. In a similar way, learned features were used for estimating the registration error in [Eppenhof, 2018b]. While these approaches are learning features that still need to be matched using a distance metric such as SSD or CC, Simonovsky et al. [Simonovsky, 2016] proposed to learn a similarity metric for inter-subject brain MR T1-T2 registration which showed improved results compared to MI. Wright et al. [Wright, 2018] used recurrent spatial co-transformer networks to iteratively register MR and US volumes showing a better quantified image similarity than self-similarity context descriptors for multi-modal registration.

End-to-End

To overcome the need of slow iterative registration procedures, supervised end-to-end approaches have been proposed that mostly have near real-time performance during testing. According to Eq. 2.4, approaches in this category require registered image pairs during training. In Fig. 2.2, a graphical representation of the typical supervised approach is shown. Due to the difficulty of finding dense ground truth voxel mappings, supervised methods need to rely on deformation predictions either from existing algorithms [Yang, 2017; Rohé, 2017; Cao, 2017], simulations [Sokooti, 2017; Uzunova, 2017; Eppenhof, 2018a] or a combination of both [Mahapatra, 2018; Krebs, 2017]. Instead of predicting the deformation field ϕ, diffeomorphic approaches predict parameterizations based on patches of the initial momentum of LDDMMs [Yang, 2017] or dense SVFs [Rohé, 2017]. In order to reduce the complexity but therefore limiting the use for large deformations, patch-wise approaches have been proposed [Cao, 2017; Sokooti, 2017; Yang, 2017]. In case of simulation-based approaches, Sokooti et al. [Sokooti, 2017] used random transformations based on Gaussian kernels. Random transformations limit the realism and task-specificity of deformations such that, more sophisticated simulations were used by multi-scale, random transformations of aligned image pairs [Eppenhof, 2018a] or applying a statistical deformation model for data augmentation [Uzunova, 2017; Krebs, 2017].

Another way of optimizing Eq. 2.4 is by using deep reinforcement learning (DRL) and implicitly quantifying image similarity through an agent [Haskins, 2020]. Hereby, an agent takes consecutive decisions on actions to apply based on the current state and future reward. This strategy allows to follow a trajectory towards the optimal transformation parameters while allowing to recover from mistakes. Due to limitations on the action space, most approaches have considered rigid registration only [Liao, 2017b; Ma, 2017;

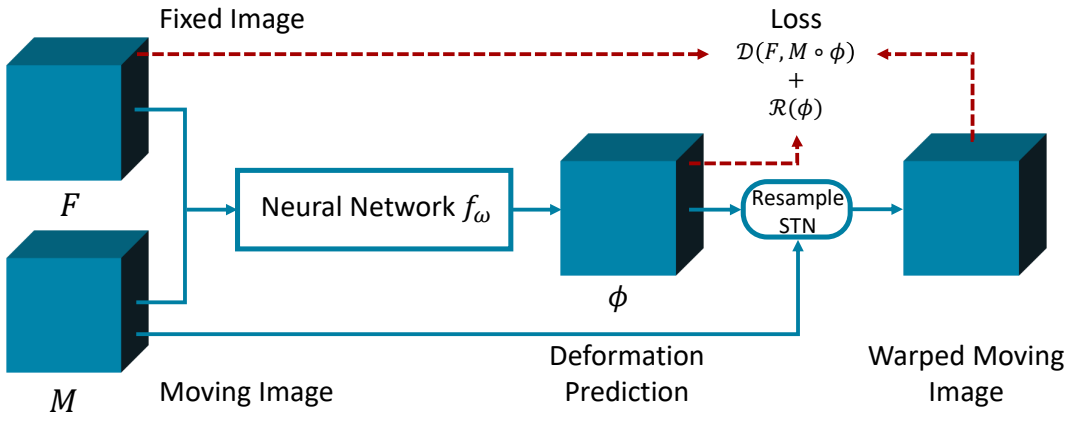


Fig. 2.3: The unsupervised end-to-end registration model. During training, the similarity $\mathcal{D}$ between warped moving and fixed image are optimized together with a deformation regularizer $\mathcal{R}$ (red arrows).

Miao, 2018]. However, by using a low-dimensional SDM, we have shown that DRL is useful for difficult inter-subject registration tasks by showing improved registration accuracy compared to state-of-the-art algorithms (cf. Chapter 3, [Krebs, 2017]).

Supervised methods are free from the need of having to define a similarity metric (and most-often regularizer) manually, but the lack of ground-truth deformations either limits the approaches by the performance of existing algorithms or the realism of simulations. Furthermore, retrieving deformations from existing algorithms on a large database is time-consuming and increases the training complexity.

2.4.2 Unsupervised

To overcome the limitations of supervised approaches, end-to-end DL-based approaches that do not require ground-truth deformations have been considered more recently. The introduction of spatial transformer networks (STN, [Jaderberg, 2015]) allowed to integrate transformation models based on B-splines or linear interpolation for dense deformation fields in neural networks directly for an efficient and most importantly differentiable warping of the moving image. With a differentiable warping functionality, loss functions can be applied on the warped moving image, allowing to integrate the classical objective function for registration (Eq. 2.1) in the loss function of neural networks (cf. Fig. 2.3):

$$\omega^* = \arg\min_{\omega} \mathbb{E}_{p(F,M)}\left[\mathcal{D}(F, M \circ f_\omega(F, M)) + \mathcal{R}(f_\omega(F, M))\right], \tag{2.5}$$

where $p(F, M)$ is the empirical data distribution of image pairs. The difference to *classical* registration is that the optimization is done over many training image pairs instead of one test image pair (F, M). Similar learning approaches first appeared in the computer vision

community [Jason, 2016; Liang, 2017] and were recently applied to medical image registration [Vos, 2017; Ghosal, 2017; Yoo, 2017; Balakrishnan, 2018; Krebs, 2018; Dalca, 2018; Mahapatra, 2018; Fan, 2018; Tanner, 2018; Li, 2018; Krebs, 2019b; Vos, 2019; Balakrishnan, 2019; Dalca, 2019a; Sandkühler, 2019]. These approaches cover dense or B-spline [Vos, 2017; Vos, 2019] deformation models, diffeomorphic models [Krebs, 2018; Dalca, 2018; Krebs, 2019b; Dalca, 2019a], single or multi-scale models [Vos, 2019; Krebs, 2019b]. Common similarity and regularization metrics as in classical methods are applied (cf. 2.2.1-2.2.2).

In an iterative fashion using recurrent networks, Sandkuhler et al. [Sandkühler, 2019] obtained a more compact representation and a speedup of 15 compared to B-spline registration for 2D images. Dropping the need for choosing a pre-defined regularizer, Niethammer et al. proposed to learn a spatially adaptive regularizer using multi-Gaussian kernels [Niethammer, 2019].

In a different optimization scheme, adversarial approaches based on generative adversarial networks (GAN, [Goodfellow, 2014]) were used for the difficult case of multi-modal registration [Mahapatra, 2018; Fan, 2018; Tanner, 2018]. Besides, probabilistic approaches were proposed in [Krebs, 2018; Dalca, 2018; Krebs, 2019b; Dalca, 2019a]. In our works [Krebs, 2018; Krebs, 2019b], deformations are encoded in a low-dimensional structured space, similar to an SDM, which allows for a variety of analysis tasks as particularly discussed in the later chapters of this thesis.

2.4.3 Weakly Supervised

All methods that use the unsupervised objective (Eq. 2.5) and additionally make use of some extra information during training such as labels or few ground-truth deformation fields fall in the category of weakly supervised algorithms.

In the latter case, Fan et al. [Fan, 2019] combined supervised and unsupervised objective functions with dynamically changing weights between both, first focusing on learning from supervised deformation fields and later increasing the weight for the objective of Eq. 2.5 for fine-tuning.

On the other side, using the matching of extra labels such as landmarks or segmentation masks, has become popular in very recent approaches due to the fact that such anatomical guidance can improve registration performance in contrast to only intensity-based metrics. Furthermore, an advantage of DL-based approaches is unlike classical methods which are based on geometric similarity metrics, that such extra information are only necessary at the training stage, while test cases do not require labels. Following this principle, Hering et al. [Hering, 2019] introduced a label and similarity metric based loss function

for deformable registration of 2D cine-MR images. Hu et al. [Hu, 2018] proposed to only optimize the matching of labels based on a multi-scale DICE loss and a deformation regularization duplicating the objectives of classical geometric approaches.

More recently, it has been proposed to learn a structure-enhanced representation from segmentations for helping with the registration of hard to register structures [Lee, 2019]. The assumption that segmentation and registration can facilitate each other has let to approaches predicting both by combining the unsupervised registration objective Eq. 2.5 and a segmentation loss [Qin, 2018; Li, 2019]. The latter approach has been successfully applied on cardiac cine-MR sequences and showed solving registration and segmentation in a joint fashion helps to improve both tasks.

3 Robust Registration for Difficult Deformations by Agent-based Action Learning

Contents

In the previous chapter, we showed a state-of-the-art of deformable registration. This chapter focuses on inter-subject registration tasks which are difficult to solve using traditional methods because of the high variability in appearance and large deformations. We try to overcome these difficulties by learning from known deformations and applying a decision-making process in order to optimize the parameters of a learned statistical deformation model. This approach can be classified as a supervised DL-based registration algorithm as it relies on simulated and *ground-truth* deformations. This chapter has been presented at the MICCAI 2017 conference [Krebs, 2017].

3.1 Introduction

Registration of images with focus on the ROI is essential in fusion and atlas-based segmentation (e.g. [Tian, 2015]). Traditional algorithms try to compute the dense mapping between two images by minimizing an objective function with regard to some similarity criterion. However, besides challenges of solving the ill-posed and non-convex problem many approaches have difficulties in handling large deformations or large variability in appearance. Recently, promising results using deep representation learning have been presented for learning similarity metrics [Simonovsky, 2016], predicting the optical flow [Dosovitskiy, 2015] or the large deformation diffeomorphic metric

mapping-momentum [Yang, 2016]. These approaches either only partially remove the above-mentioned limitations as they stick to an energy minimization framework (cf. [Simonovsky, 2016]) or rely on a large number of training samples derived from existing registration results (cf. [Dosovitskiy, 2015; Yang, 2016]).

Inspired by the recent works in reinforcement learning [Mnih, 2015; Ghesu, 2016], we propose a reformulation of the non-rigid registration problem following a similar methodology as in 3-D rigid registration of [Liao, 2017b]: in order to optimize the parameters of a deformation model we apply an artificial agent – solely learned from experience – that does not require explicitly designed similarity measures, regularization and optimization strategy. Trained in a supervised way the agent explores the space of deformations by choosing from a set of actions that update the parameters. By iteratively selecting actions, the agent moves on a trajectory towards the final deformation parameters. To decide which action to take we present a deep dual-stream neural network for implicit image correspondence learning. This work generalizes [Liao, 2017b] to non-rigid registration problems by using a larger number of actions with a low-dimensional parametric deformation model. Since ground-truth (GT) deformation fields are typically not available for deformable registration and training based on landmark-aligned images as in rigid registration (cf. [Liao, 2017b]) is not applicable, we propose a novel GT generator combining synthetically deformed and real image pairs. The GT deformation parameters of the real training pairs were extracted by constraining existing registration algorithms with known correspondences in the ROI in order to get the best possible organ-focused results. Thus, the main contributions of this work are: (1) The creation and use of a low-dimensional parametric statistical deformation model for organ-focused deep learning-based non-rigid registration. (2) A ground truth generator which allows generating millions of synthetically deformed training samples requiring only a few ($<$1000) real deformation estimations. (3) A novel way of fuzzy action control.

3.2 Method

3.2.1 Training Artificial Agents

Image registration consists in finding a spatial transformation $\mathcal{T}_\theta$, parameterized by $\theta \in \mathbb{R}^d$ which best warps the moving image $\mathbf{M}$ as to match the fixed image $\mathbf{F}$. Traditionally, this is done by minimizing an objective function of the form: $\arg\min_\theta \mathcal{F}(\theta, \mathbf{M}, \mathbf{F}) = \mathcal{D}(\mathbf{F}, \mathbf{M} \circ \mathcal{T}_\theta) + \mathcal{R}(\mathcal{T}_\theta)$ with the image similarity metric $\mathcal{D}$ and a regularizer $\mathcal{R}$. In many cases, an iterative scheme is applied where at each iteration t the current parameter value θ_t is updated through gradient descent: $\theta_{t+1} = \theta_t + \lambda \nabla \mathcal{F}(\theta_t, \mathbf{M}_t, \mathbf{F})$ where $\mathbf{M}_t$ is the deformed moving image at time step t: $\mathbf{M} \circ \mathcal{T}_{\theta_t}$.

Inspired by [Liao, 2017b], we propose an alternative approach to optimize θ based on an artificial agent which decides to perform a simple action a_t at each iteration t consisting in applying a fixed increment $\delta\theta_{a_t}$: $\theta_{t+1} = \theta_t + \delta\theta_{a_t}$. If θ is a d-dimensional vector of parameters, we define $2d$ possible actions $a \in \mathcal{A}$ such that $\delta\theta_{2i}[j] = \epsilon_i\delta_i^j$ and $\delta\theta_{2i+1}[j] = -\epsilon_i\delta_i^j$ with $i \in \{0..d-1\}$. In other words the application of an action a_t increases or decreases a specific parameter within θ_t by a fixed amount where δ_i^j is an additional scaling factor per dimension that is set to 1 in our experiments but could be used e.g. to allow larger magnitudes first and smaller in later iterations for fine-tuning the registration.

The difficulty in this approach lies into selecting the action a_t as function of the current state s_t consisting of the fixed and current moving image: $s_t = (\mathbf{F}, \mathbf{M}_t)$. To this end, the framework models a Markov decision process (MDP), where the agent interacts with an environment getting feedbacks for each action. In reinforcement learning (RL) the best action is selected based on the maximization of the quality function $a_t = \arg\max_{a\in\mathcal{A}} Q^\star(s_t, a)$. In the most general setting, this optimal action-value function is computed based on the reward function defined between two states $\mathcal{R}(s_1, a, s_2)$ which serves as the feed-back signal for the agent to quantify the improvement or worsening when applying a certain action. Thus, $Q^\star(s_t, a)$ may take into account the immediate but also future rewards starting from state s_t, as to evaluate the performance of an action a.

Recently, in RL powerful deep neural networks have been presented that approximate the optimal $Q^\star$ [Mnih, 2015]. Ghesu *et al.* [Ghesu, 2016] used deep reinforcement learning (DRL) for landmark detection in 2-D medical images. In the rigid registration approach by Liao *et al.* [Liao, 2017b] the agent's actions are defined as translation and rotation movements of the moving image in order to match the fixed image.

In this work, the quality function $\mathbf{y}_a(s_t) \approx Q^\star(s_t, a)$ is learned in a supervised manner through a deep regression network. More precisely, we adopt a single-stage MDP for which $Q^\star(s_t, a) = \mathcal{R}(s_t, a, s_{t+1})$, implying that only the immediate reward, i.e. the next best action, is accounted for. During training, a batch of random states, pairs of $\mathbf{F}$ and $\mathbf{M}$, is considered with known transformation $\mathcal{T}_{\theta_{GT}}$ (with $\mathbf{F} \approx \mathbf{M} \circ \mathcal{T}_{\theta_{GT}}$). The target quality is defined such that actions that bring the parameters closer to its ground truth value are rewarded:

$$Q^\star(s_t, a) = \mathcal{R}(s_t, a, s_{t+1}) = \|\theta_{GT} - \theta_{s_t}\|_2 - \|\theta_{GT} - \theta^a_{s_{t+1}}\|_2 \ . \tag{3.1}$$

The training loss function consists of the sum of L_2-norms between the explicitly computed Q-values (Eq. 3.1) for all actions $a \in \mathcal{A}$ and the network's quality predictions

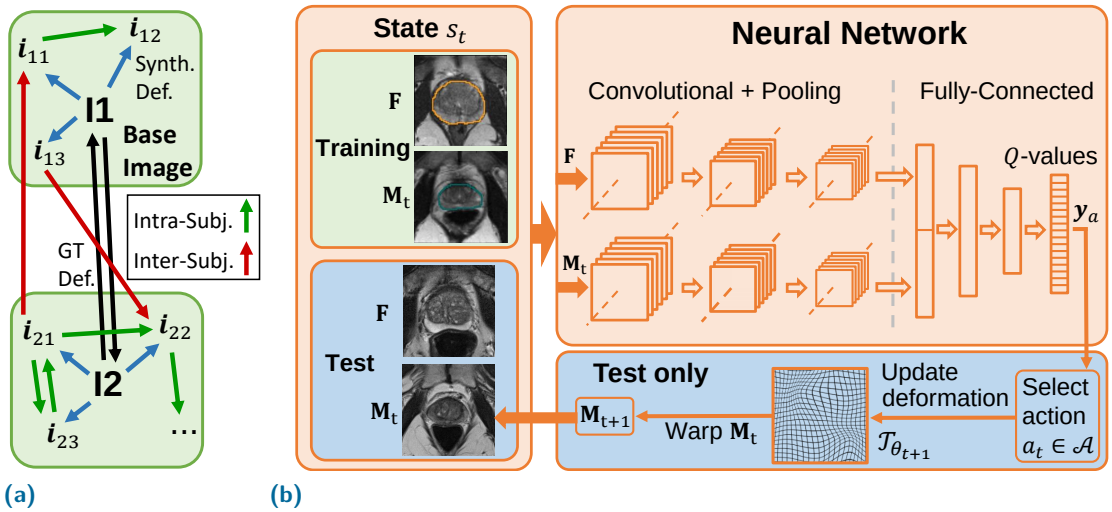


Fig. 3.1: (a) Training Data Generation: Synthetic deformations (blue arrows) and inter-subject GT deformations (black) are used for intra- (green) and inter-subject (red) image pairs for training. (b) Dual-stream network used for Q-value prediction $\mathbf{y}_a$ including complete single-stage Markov Decision Process for testing (blue background).

$\mathbf{y}_a(s_t)$ per action. Having a training batch $\mathcal{B}$ with random states s_b the loss is defined as: $L = \sum_{s_b \in \mathcal{B}} \sum_{a \in \mathcal{A}} \|\mathbf{y}_a(s_b) - Q^\star(s_b, a)\|^2$.

In testing, the agent iteratively selects the best action, updates the parameter θ_t and warps the moving image $\mathbf{M}_t$ as to converge to a final parameter set representing the best mapping from moving to fixed image (see Fig. 3.1b).

3.2.2 Statistical Deformation Model

One challenge of the proposed framework is to find a low dimensional representation of non-rigid transformations to minimize the number of possible actions (equal to $2d$), while keeping enough degrees of freedom to correctly match images. In this work, we base our registration method on statistical deformation models (SDM) defined from Free Form Deformations (FFD). Other parametrizations could work as well. Typically, the dense displacement field is defined as the summation of tensor products of cubic B-splines on a rectangular grid. Rueckert *et al.* [Rueckert, 2003] proposed to further reduce the dimensionality by constructing an SDM through a principal component analysis (PCA) on the B-spline displacements.

We propose to use the modes of the PCA as the parameter vector θ describing the transformation $\mathcal{T}_\theta$ that the agent aims to optimize. The agent's basic increment per action ϵ_i is normalized according to the mean value of each mode estimated in training. To have a stochastic exploration of the parameter space, predicted actions a_t are selected

in a stochastic manner among the 3 best actions with given fixed probabilities (see [Liao, 2017b]).

Fuzzy Action Control

Since parameters θ are the amplitudes of principal components, the deviation of θ_{2m} and θ_{2m+1} from the mean μ_m should stay within k-times the standard deviation σ_m in testing. In order to keep θ inside this reasonable parametric space of the SDM, we propose fuzzy action controlling. Thus, actions that push parameter values of θ outside that space, are stochastically penalized – after being predicted by the network. Inspired by rejection sampling, if an action a moves parameter θ_m to a value f_m, then this move is accepted if a random number generated between $[0, 1]$ is less than the ratio $\mathcal{N}(f_m; \mu_m, \sigma_m)/\mathcal{N}(h; \mu_m, \sigma_m)$ where $h_m = \mu_m + k\sigma_m$, and $\mathcal{N}$ is the Gaussian distribution function. Therefore, if $|f_m - \mu_m| \leq k\sigma_m$, the ratio is greater than 1 and the action is accepted. If $|f_m - \mu_m| > k\sigma_m$ then the action is randomly accepted, but with a decreased likelihood as f_m moves far away from μ_m. This stochastic thresholding is performed for all actions at each iteration and rejection is translated into adding a large negative value to the quality function $\mathbf{y}_a$. The factor k controls the tightness of the parametric space and is empirically chosen as 1.5. By introducing fuzzy action control, the MDP gets more robust since the agent's access to the less known subspace of the SDM is restricted.

3.2.3 Training Data Generation

Since it is difficult to get trustworthy ground-truth (GT) deformation parameters θ_{GT} for training, we propose to generate two different kinds of training pairs: Inter- and intra-subject pairs where in both moving and fixed images are synthetically deformed. The latter pairs serve as a data augmentation method to improve the generalization of the neural network.

In order to produce the ground truth deformations of the available training images, one possibility would be to apply existing registration algorithms with optimally tuned parameters. However, this would imply that the trained artificial agent would only be as good as those already available algorithms. Instead, we make use of manually segmented regions of interest (ROI) available for both pairs of images. By constraining the registration algorithms to enforce the correspondence between the 2 ROIs (for instance by artificially outlining the ROIs in images as brighter voxels or using point correspondences in the ROI), the estimated registration improves significantly around the ROI. From the resulting deformations represented on an FFD grid, the d principal components are extracted. Finally, these modes are used to generate the synthetic training samples by warping the original training images based on randomly drawn

deformation samples according to the SDM. Amplitudes of the modes are bounded to not exceed the variations experienced in the real image pairs, similar to [Rueckert, 2003].

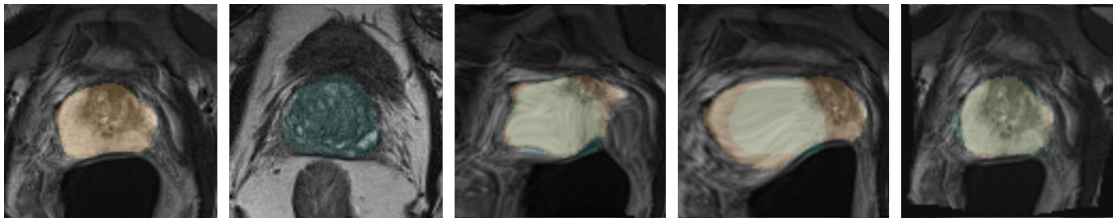
Intra-subject training pairs can be all combinations of synthetically deformed images of the same subject. Since the ground-truth deformation parameters are exactly known, it is guaranteed that the agent learns correct deformations. In the case of inter-patient pairs a synthetic deformed image i_{mb} of one subject I_m is allowed to be paired with any synthetic deformed image i_{nc} of any other subject I_n with b, c denoting random synthetic deformations (see Fig. 3.1a). Thereby, the GT parameters θ_{GT} for image pair (i_{mb}, i_{nc}) are extracted via composition of the different known deformations such that $((i_{mb} \circ \mathcal{T}_\theta^{i_{mb},I_m}) \circ \mathcal{T}_\theta^{I_m,I_n}) \circ \mathcal{T}_\theta^{I_n,i_{nc}}$. Note the first deformation would require the inverse of a known deformation that we approximate by its opposite parameters for reasons of computational efficiency. The additional error due to this approximation, computed on a few pairs, remained below 2% in terms of the DICE score.

Mini-batches are created online – during training – via random image pairing where intra- and inter-subject pairs are selected with the same probabilities. Through online random pairing the experience of new pairs is enforced since the number of possible image combinations can be extremely high (e.g. 10^{12}) depending on the number of synthetic deformations.

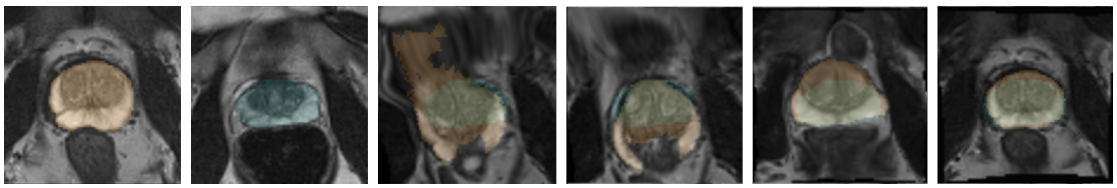
3.3 Experiments

We focused on organ-centered registration of MR prostate images in 2-D and 3-D with the use case of image fusion and atlas-based segmentation [Tian, 2015]. The task is very challenging since texture and anatomical appearance can vary a lot. 25 volumes were selected from the MICCAI challenge *PROMISE12*[1] and 16 from the *Prostate-3T* database[2] including prostate segmentations. Same images and the cases with rectal probes were excluded. Randomly 8 cases were chosen for testing (56 pairs), 33 for training. As preprocessing, translation-based registration for all pairs was carried out in 3-D using the *elastix*-framework[Klein, 2010] with standard parameters followed by cropping and down sampling the images (to 100x100/75x75x20 pixels in 2-D/3-D respectively). For the 2-D experiments, the middle slice of each volume was taken. For the purpose of GT generation mutual information as similarity metric and a bending energy metric was used. The optimization function was further constrained by a Euclidean point correspondence metric. Therefore, equally distributed points were extracted from the given mask surfaces. *elastix* was used to retrieve the solution with the weights 1, 3 and 0.2 for the above-mentioned metrics and a B-spline spacing of 16x16(x8) voxels. As a surrogate measure

[1] https://promise12.grand-challenge.org/
[2] https://wiki.cancerimagingarchive.net/display/Public/Prostate-3T



(a) 2-D: Moving, Fixed, *elastix*-e8 (.84), *elastix*-e16 (.70), ours (.94).



(b) 3-D: Moving, Fixed, *elastix*-e8 (.49), *elastix*-e16 (.59), *LCC-Demons* (.67), ours (.79).

Fig. 3.2: 2-D and 3-D registration results of extreme cases with segmentation masks overlays (fixed: green, moving: orange) and DICE scores in parenthesis.

of registration performance we used the DICE score and Hausdorff distance (HD) on the prostate region. The extracted GT resulted in median DICE coefficients of .96 in 2-D and .88 in 3-D. Given the B-spline displacements, the PCA was trained with $d = 15$ modes in 2-D, $d = 25$ in 3-D (leading to 30 respectively 50 actions with a reconstruction error $< 5\%$ (DICE score) as a compromise to keep the number of modes relatively small.

The network's two independent processing streams contained 3 convolutional (with 32, 64, 64 filters and kernel size 3) and 2 max-pooling layers for feature extraction. The concatenated outputs of the two streams were processed in 3 fully-connected layers (with 128, 128, 64 knots) resulting in an output with size $2d$ (equals the number of actions). Batch normalization and ReLu units were used in all layers. The mini-batch size was 65/30 (2-D/3-D). For updating the network weights, we used the adaptive learning rate gradient-based method *RMSprop*. The learning rate was 0.001 with a decay factor of 0.8 every 10k mini-batch back-propagations. Training took about 12 hours/ 1 day for 2-D and 3-D respectively. All experiments were implemented in *Python* using the deep learning library *Theano* including *Lasagne*[3]. DL tasks ran on GPUs (*NVIDIA GeForce GTX TITAN X*). During testing 200 MDP iterations (incl. resampling of the moving image) took 10 seconds (GPU) in 2-D and 90 seconds in 3-D (GPU). The number of testing steps was set empirically since registration results only change marginally when increasing the number of steps. In empirical 2-D experiments with 1000 steps the agent's convergence was observable (see Fig. 3.5 in the appendix).

For testing, the initial translation registration was done with *elastix* by registering each of the test images to an arbitrarily chosen template from the training base. Table 3.1 shows that our method reaches a median DICE coefficient of .88/.76 in 2-D/3-D and therefore shows similar performance as in [Klein, 2010] with the best reported median DICE of .76

[3] https://lasagne.readthedocs.io/en/latest

Tab. 3.1: Results of prostate MR registration on the 56 testing pairs. 2-D and 3-D results in comparison to *elastix* with B-spline spacing of 8 (e8) or 16 (e16) as proposed in [Klein, 2010] and the *LCC-Demons*[Lorenzi, 2013] algorithm (dem). T are the initial scores after translation registration with *elastix*. 3-D* are results with perfect rigid alignment T*. **nfc** are our results with no fuzzy action control (HD in mm).

	2-D				3-D				
	T	e16	e8	**our**	T	e16	e8	dem	**our**
DICE Mean	.76	.74	.77	**.87**	.62	.63	.64	.67	**.75**
DICE Med.	.78	.79	.81	**.88**	.61	.71	70	.67	**.76**
DICE StD.	.10	.15	.13	**.05**	.11	.22	.20	.11	**.06**
HD Mean	11.6	15.2	14.5	**7.7**	16.1	21.2	25.3	15.9	**11.8**
HD Med.	11.7	13.2	13.0	**7.2**	15.2	18.0	21.7	15.8	**11.2**
HD StD.	4.3	6.8	6.7	**2.5**	3.9	10.7	10.9	3.9	**2.9**
	3-D*								
	T*	e16	e8	dem	**nfc**	**our**			
DICE Mean	.74	.72	.67	.79	**.79**	**.80**			
DICE Med.	.75	.77	.76	.80	**.79**	**.81**			
DICE StD.	.08	.17	.23	.07	**.05**	**.04**			
HD Mean	9.2	13.4	14.5	10.4	**8.9**	**8.0**			
HD Med.	9.0	11.6	13.5	10.8	**8.8**	**7.9**			
HD StD.	2.3	6.8	6.4	2.5	**2.2**	**1.9**			

on a different data set. However, on our challenging test data our method outperformed the *LCC-Demons*[Lorenzi, 2013] algorithm with manually tuned parameters and *elastix*, using similar parameters as proposed for prostate registration [Klein, 2010] using B-spline spacing of 8 and 16 pixels. We found that better rigid registration can significantly improve the algorithm's performance as shown in the experiments with perfect rigid alignment according to the segmentation (3-D*). Extreme results are visually shown in Fig. 3.2. More 2-D and 3-D examples are shown in the appendix, Fig. 3.3-3.4.

Regarding the results of *elastix* and *LCC-Demons*, a rising DICE score was observed while HD increased due to local spikes introduced in the masks (visible in Fig. 3.2b) as we focused on the DICE scores during optimization for fair comparisons. In the 3-D* setting, DICE scores and HDs improved when applying fuzzy action control compared to not applying any constraints (see Table 3.1).

3.4 Conclusion

In this work, we presented a generic learning-based framework using an artificial agent for approaching organ-focused non-rigid registration tasks appearing in image fusion and atlas-based segmentation. The proposed method overcomes limitations of traditional algorithms by learning optimal features for decision-making. Therefore, segmentation or

handcrafted features are not required for the registration during testing. Additionally, we proposed a novel ground-truth generator to learn from synthetically deformed and inter-subject image pairs.

In conclusion, we evaluated our approach on inter-subject registration of prostate MR images showing first promising results in 2-D and 3-D. In future work, the deformation parametrization needs to be further evaluated. Rigid registration as in [Liao, 2017b] could be included in the network or applied as preprocessing to improve results as shown in the experiments. Besides, the extension to multi-modal registration is desirable.

3.5 Appendix

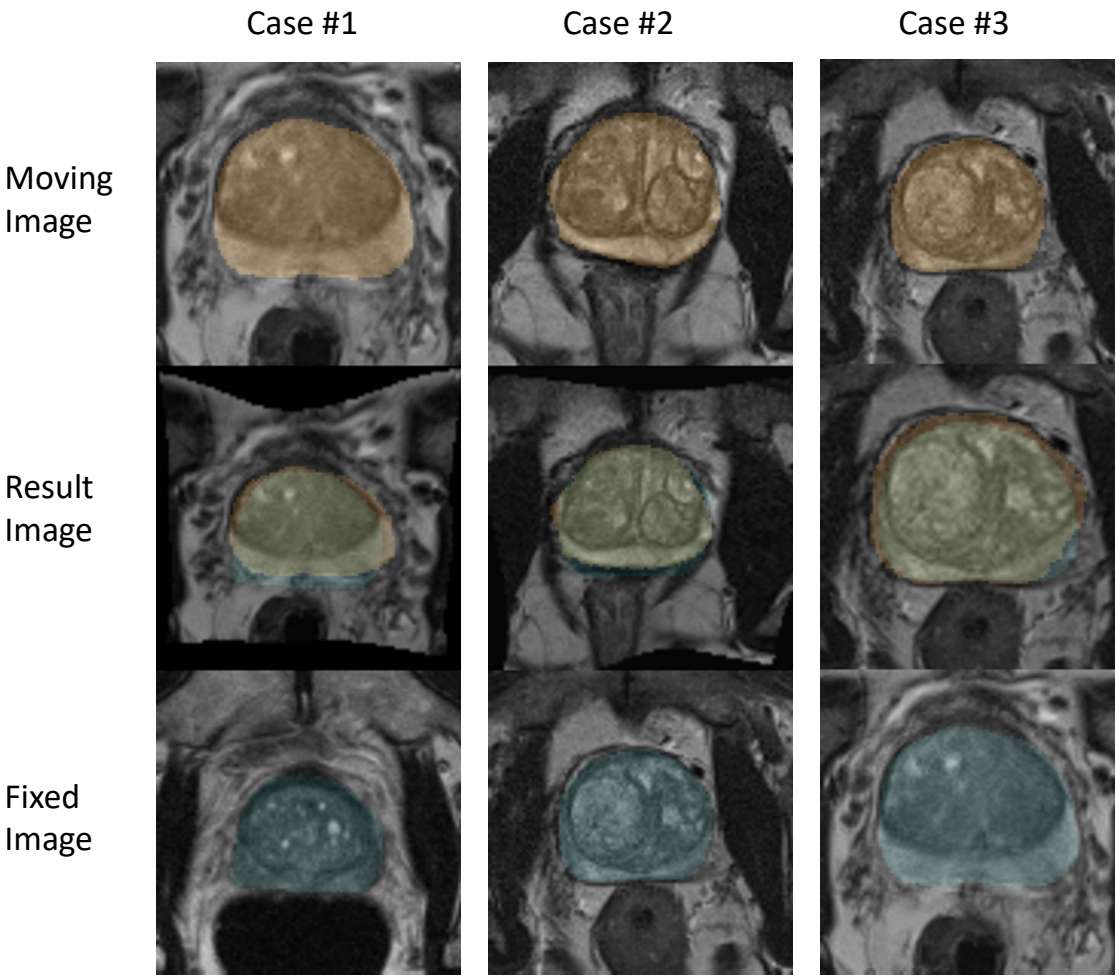


Fig. 3.3: 2-D registration results showing moving and fixed image masks as overlays in orange and green respectively. Final DICE scores for the 3 cases are .90, .93, .92 with initial overlaps of .65, .70, .72.

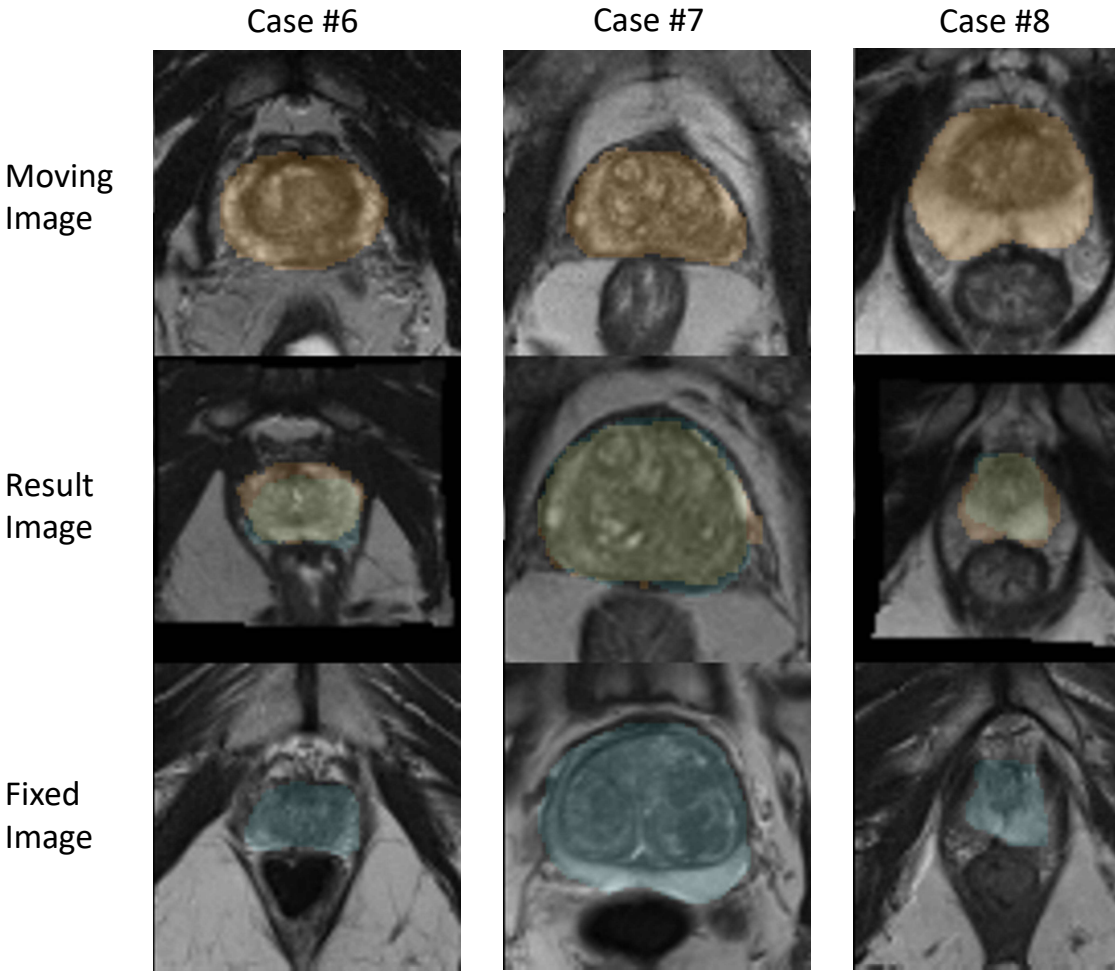


Fig. 3.4: 3-D registration results with final DICE scores for the 3 cases of .83, .85, .83 with initial overlaps of .57, .54, .54.

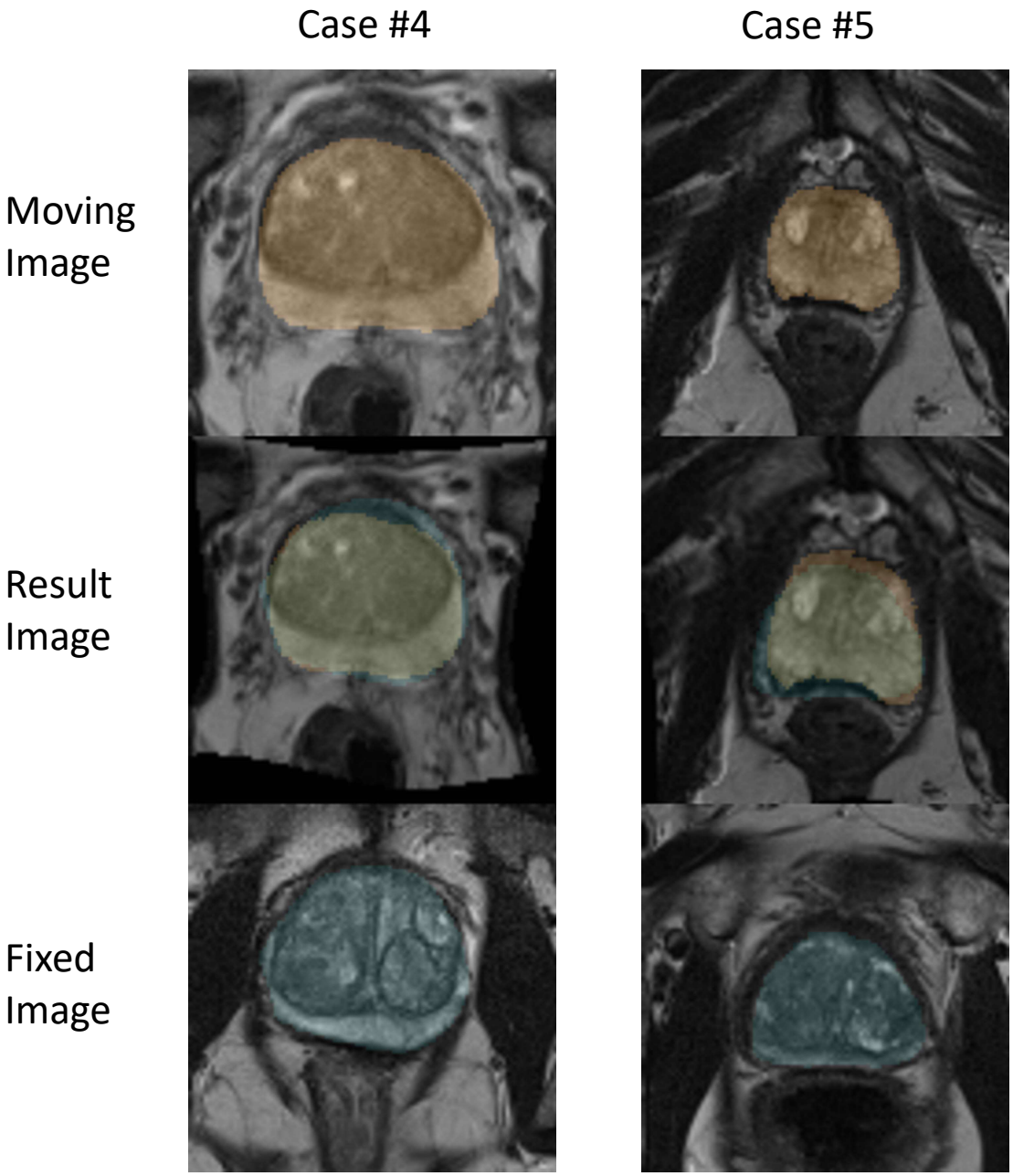


Fig. 3.5: Convergence test, showing 2-D results after 1000 agent steps which is marked by oscillation behavior between neighboring states (final DICE scores .93 and .85).

4 Learning a Probabilistic Model for Diffeomorphic Registration

Contents

In chapter 3, we have shown the successful application of a simple learned statistical deformation model (based on PCA) for difficult registration problems. In this chapter, we focus on learning a more sophisticated statistical deformation model from data that allows deformation analysis tasks such as disease clustering and simulation. We propose to learn a generative deformation model based on a conditional variational autoencoder which can be seen as a non-linear generalization of PCA. This model is trained without requiring ground-truth deformation fields or labels and thus, can be classified as an unsupervised DL-based registration algorithm. The chapter is published in the journal IEEE TMI [Krebs, 2019b] and is based on the previous conference presentation at DLMIA 2018 [Krebs, 2018].

4.1 Introduction

Deformable image registration, the process of finding voxel correspondences in a pair of images, is an essential task in medical image analysis [Sotiras, 2013]. This mapping – the deformation field – can be used for example in pre-op / post-op studies, to find the same structures in images from different modalities or to evaluate the progression of a

disease. The analysis of geometric changes in successive images is important for instance for diagnosing cardiovascular diseases and selecting the most suited therapies. A possible approach is to register sequential images and analyze the extracted deformations for example by parallel transport [Lorenzi, 2014] or by creating an adapted low-dimensional subspace [Rohé, 2018].

We propose a registration algorithm that learns a deformation model directly from training images. Inspired by recent generative latent variable models, our method learns a low-dimensional probabilistic deformation encoding in an unsupervised fashion. This latent variable space encodes similar deformations close to each other and allows the generation of synthetic deformations for a single image and the comparison and transport of deformations from one case to another.

4.1.1 Deformable Image Registration

Traditionally, deformable registration is solved by numerical optimization of a similarity metric which measures the distance between the fixed and the deformed moving image. The moving image is warped given a predefined deformation model in order to get closer to the fixed image. Unfortunately, this results in an ill-posed problem which requires further regularization based on prior assumptions [Sotiras, 2013]. Various regularization energies have been proposed including elastic- [Davatzikos, 1997; Burger, 2013] or diffusion-based methods [Thirion, 1998; Vercauteren, 2007b; Lorenzi, 2013] (cf. [Sotiras, 2013]). Diffeomorphic transforms are folding-free and invertible. The enforcement of these properties in many medical applications has led to the wide use of diffeomorphic registration algorithms. Popular parametrizations of diffeomorphisms include the Large Deformation Diffeomorphic Metric Mapping (LDDMM) [Beg, 2005; Cao, 2005; Zhang, 2015], a symmetric normalization approach [Avants, 2008] or stationary velocity fields (SVF) [Arsigny, 2006; Vercauteren, 2008].

In recent years, learning-based algorithms – notably Deep Learning (DL) – have been proposed to avoid long iterative optimization at test time. In general, one can classify these algorithms as supervised and unsupervised. Due to the difficulty of finding ground truth voxel mappings, supervised methods need to rely on predictions from existing algorithms [Yang, 2017; Rohé, 2017], simulations [Sokooti, 2017; Uzunova, 2017; Eppenhof, 2018a] or a combination of both [Krebs, 2017; Mahapatra, 2018]. The latter can be achieved for example by projecting *B-spline* displacement estimations in the space of a statistical deformation model from which one can extract simulations by sampling of its components [Krebs, 2017]. Diffeomorphic approaches predict patches of the initial momentum of LDDMMs [Yang, 2017] or dense SVFs [Rohé, 2017]. Supervised methods are either limited by the performance of existing algorithms or the realism of simulations.

Furthermore, retrieving deformations from existing algorithms on a large database is time-consuming and increases the training complexity.

Unsupervised approaches to registration aim to optimize an image similarity, often combined with a penalization or smoothing term (regularization). These learning approaches first appeared in the computer vision community [Jason, 2016; Liang, 2017] and were recently applied to medical image registration [Vos, 2017; Balakrishnan, 2018; Fan, 2019; Dalca, 2018; Tanner, 2018]. Unlike traditional methods, learning-based approaches also can include task-specific information such as segmentation labels during training while not requiring those labels at test time. Instead of using an image similarity, Hu et al. [Hu, 2018] proposed to optimize the matching of labels based on a multi-scale DICE loss and a deformation regularization. Fan et al. [Fan, 2019] proposed to jointly optimize a supervised and unsupervised objective by regressing *ground-truth* deformation fields (from an existing algorithm), while simultaneously optimizing an intensity-based similarity criterion. The disadvantage of these *semi-supervised* approaches is that their training complexity is higher since label information needs to be collected, and for example deformations outside the segmented areas are not guaranteed to be captured. Most unsupervised approaches use B-spline grids or dense deformation fields, realized with spatial transformer layers (STN [Jaderberg, 2015]) for an efficient and differentiable linear warping of the moving image. However, it has not been shown yet that these approaches lead to sufficiently regular and plausible deformations.

4.1.2 Deformation Analysis and Transport

Understanding the deformation or motion of an organ goes beyond the registration of successive images. Therefore, it has been proposed to compare and characterize shape and motion patterns by normalizing deformations in a common reference frame [Lorenzi, 2014; Duchateau, 2011] and for example by applying statistical methods to study the variation of cardiac shapes [Bai, 2015]. In the diffeomorphic setting, various dimensionality reduction methods have been proposed. Vaillant et al. [Vaillant, 2004] modeled shape variability by applying PCA in the tangent space to an atlas image. Qiu et al. used a shape prior for surface matching [Qiu, 2012]. While these methods are based on probabilistic inference, dimensionality reduction is done after the estimation of diffeomorphisms. Instead Zhang et al. [Zhang, 2014] introduced a latent variable model for principle geodesic analysis that estimates a template and principle modes of variation while infering the latent dimensionality from the data. Instead of having a general deformation model capable of explaining the deformations of any image pair in the training data distribution, this registration approach still depends on the estimation of a smooth template. Using the SVF parametrization for cardiac motion analysis, Rohé et al. [Rohé, 2018] proposed to build affine subspaces on a manifold of deformations,

the barycentric subspaces, where each point on the manifold represents a 3-D image and the geodesic between two points describes the deformation.

For uncertainty quantification, Wassermann et al. [Wassermann, 2014] used a probabilistic LDDMM approach applying a stochastic differential equation and Wang et al. [Wang, 2018] employed a low-dimensional Fourier representation of the tangent space of diffeomorphisms with a normal assumption. While both approaches contain probabilistic deformation representations, they have not been used for sampling and the representations have not been learned from a large dataset.

In the framework of diffeomorphic registration, *parallel transport* is a promising normalization method for the comparison of deformations. Currently used *parallel transport* approaches are the Schild's [Lorenzi, 2011] or pole ladder [Lorenzi, 2014; Jia, 2018] using the SVF parametrization or approaches based on Jacobi fields using the LDDMM parametrization [Younes, 2007; Louis, 2017]. In general, these approaches aim to convert and apply the temporal deformation of one subject to another subject. However, this *transport* process typically requires multiple registrations, including difficult registrations between different subjects.

4.1.3 Learning-based Generative Latent Variable Models

Alternatively and inspired by recently introduced learning-based generative models, we propose to learn a latent variable model that captures deformation characteristics just by providing a large dataset of training images. In the computer vision community, such generative models as Generative adversarial networks (GAN) [Goodfellow, 2014], stochastic variational autoencoders (VAE) [Kingma, 2013] and adversarial autoencoders (AAE) [Makhzani, 2016] have demonstrated great performance in learning data distributions from large image training sets. The learned models can be used to generate new synthetic images, similar to the ones seen during training. In addition, probabilistic VAEs are latent variable models which are able to learn continuous latent variables with intractable posterior probability distributions (encoder). Efficient Bayesian inference can be used to deduce the posterior distribution by enforcing the latent variables to follow a predefined (simple) distribution. Finally, a decoder aims to reconstruct the data from that representation [Kingma, 2013]. As an extension, conditional variational autoencoders (CVAE) constrain the VAE model on additional information such as labels. This leads to a latent variable space in which similar data points are mapped close to each other. CVAEs are for example used for semi-supervised classification tasks [Kingma, 2014b].

Generative models also showed promising results in medical imaging applications such as in classifying cardiac diseases [Biffi, 2018] or predicting PET-derived myelin content maps from multi-modal MRI [Wei, 2018]. Recently, unsupervised adversarial training

approaches have been proposed for image registration [Mahapatra, 2018; Fan, 2018; Tanner, 2018]. Dalca et al. [Dalca, 2018] developed a framework which enforces a multivariate Gaussian distribution on each component of the velocity field for measuring uncertainty. However, these approaches do not learn global latent variable models which map similar deformations close to each other in a probabilistic subspace of deformations. To the best of the authors' knowledge, generative approaches for registration which allow the sampling of new deformations based on a learned low-dimensional encoding have not been proposed yet.

4.1.4 Probabilistic Registration using a Generative Model

We introduce a generative and probabilistic model for diffeomorphic image registration, inspired by generative latent variable models [Kingma, 2013; Kingma, 2014b]. In contrast to other probabilistic approaches such as [Yang, 2017; Dalca, 2018], we learn a low-dimensional global latent space in an encoder-decoder neural network where the deformation of a new image pair is mapped to and where similar deformations are close to each other. This latent space, learned in an unsupervised fashion, can be used to generate an infinite number of new deformations for any single image from the data distribution and not only for a unique template as in the Bayesian inference procedure for model parameter estimation in [Zhang, 2014]. From this abstract representation of deformations, diffeomorphic deformations are reconstructed by decoding the latent code under the constraint of the moving image. To the best of the author's knowledge, this method describes the first low-dimensional probabilistic latent variable model that can be used for deformation transport from one subject to another. Through applying a latent deformation code of one image pair on a new constraining image, deformation transport (and sampling from the latent space) is useful for instance for simulating cardiac pathologies or synthesizing a large number of pathological and healthy heart deformations for data augmentation purposes.

We use a variational inference method (a CVAE [Kingma, 2014b]) with the objective of *reconstructing* the fixed image by warping the moving image. The decoder of the CVAE is conditioned on the moving image to ease the encoding task: by making appearance information easily accessible in the decoder (in the form of the moving image), the latent space is more likely to encode deformation rather than appearance information. This implicit decoupling of deformation and appearance information allows to transport deformations from one case to another by pairing a latent code with a new conditioning image. The framework provides multi-scale estimations where velocities are extracted at each scale of the decoding network. We use the SVF parametrization and diffeomorphisms are extracted using a vector field exponentiation layer, based on the *scaling and squaring* algorithm proposed in [Arsigny, 2006]. This algorithm has been successfully applied in neural networks in our previous work [Krebs, 2018] and in [Dalca, 2018]. The

framework contains a dense spatial transformer layer (STN) and can be trained end-to-end with a choice of similarity metrics: to avoid asymmetry, we use a symmetric and normalized local cross correlation criterion. In addition, we provide a generic formulation to include regularization terms to control the deformation appearance (if required), such as diffusion regularization in form of Gaussian smoothing [Lorenzi, 2013]. During training, similarity loss terms for each scale and a loss term enforcing a prior assumption on the latent variable distribution are optimized by using the concept of *deep supervision* (cf. [Lee, 2015]). During testing, the low-dimensional latent encoding, multi-scale estimations of velocities, deformation fields and deformed moving image are retrieved in a single forward path of the neural network.

We evaluate our framework on the registration of cardiac MRIs between end-diastole (ED) and end-systole (ES) and provide an intensive analysis on the structure of the latent code and evaluate its application for transporting encoded deformations from one case to another.

This paper extends our preliminary work [Krebs, 2018] by adding:

- Detailed derivations of the probabilistic registration framework including a generic regularization model.
- Deep supervision, multi-scale estimations and a normalized loss function to improve registration performances.
- Analysis of size and structure of the latent variable space.
- Evaluation of the deformation transport by comparing it to a state-of-the-art algorithm [Lorenzi, 2014].

4.2 Methods

In image registration, the goal is to find the spatial transformation $\mathcal{T}_z : \mathbb{R}^3 \rightarrow \mathbb{R}^3$ which is parametrized by a d-dimensional vector $z \in \mathbb{R}^d$. The optimal values of z are the ones which best warp the moving image M in order to match the fixed image F given the transformation $\mathcal{T}_z$. Both images F and M are defined in the spatial domain $\Omega \in \mathbb{R}^3$. Typically, z is optimized by minimizing an objective function of the form: $\arg\min_z \mathcal{F}(z, M, F) = \mathcal{D}\left(F, M \circ \mathcal{T}_z\right) + \mathcal{R}(\mathcal{T}_z)$, where $\mathcal{D}$ is a metric measuring the similarity between fixed F and warped moving image $M \circ \mathcal{T}_z$. $\mathcal{R}$ is a spatial regularizer [Sotiras, 2013]. Recent unsupervised DL-based approaches [Jason, 2016; Vos, 2017; Balakrishnan, 2018] try to learn to maximize such a similarity metric $\mathcal{D}$ using stochastic gradient descent methods and a spatial transformer layer (STN [Jaderberg, 2015]) for warping the moving image M.

In extension, we propose to model registration by learning a probabilistic deformation parametrization vector z from a set of example image pairs (M, F). Thereby, we constrain the low-dimensional z to follow a prior distribution $p(z)$. In other words, our approach contains two key parts: a latent space encoding to model deformations and a decoding function that aims to *reconstruct* the fixed image F from this encoded transformation – by warping the moving image M. In addition, this decoding function is generative as it allows to sample new deformations based on $p(z)$.

4.2.1 Probabilistic model for multi-scale registration

We assume a generative probabilistic distribution for registration $p_{true}(F|M)$, capturing the deformation from M towards F. We aim at learning a parameterized model $p_\theta(F|M)$ with parameters θ which allows us to sample new F's that are similar to samples from the unknown distribution p_{true}. To estimate the posterior p_θ we use a latent variable model parameterized by z. Following the methodology of a VAE [Kingma, 2013], we assume the prior $p(z)$ to be a multivariate unit Gaussian distribution with spherical covariance I:

$$p(z) \sim \mathcal{N}(0, I). \tag{4.1}$$

Using multivariate Gaussians offers a closed-form differentiable solution, however, $p(z)$ could take the form of other distributions. In this work, we parameterize deformation fields ϕ by stationary velocity fields (SVF), denoted by velocities v: $\phi = \exp(v)$ [Arsigny, 2006]. These transformation maps ϕ are given as the sum of identity and displacements $u(x)$ for every position $x \in \Omega$: $\phi(x) = x + u(x)$. In the multi-scale approach, we define velocities v^s at scale $s \in \mathcal{S}$ where $\mathcal{S}$ is the set of different image scales ($s = 1$ describes the original scale for which we omit writing s and $s = 2, 3, ...$ the scale, down-sampled by a factor of 2^{s-1}). For each scale s, a family of *decoding* functions f_v^s is defined, parameterized by a fixed $\theta^s \subset \theta$ and dependent on z and the moving image M^s:

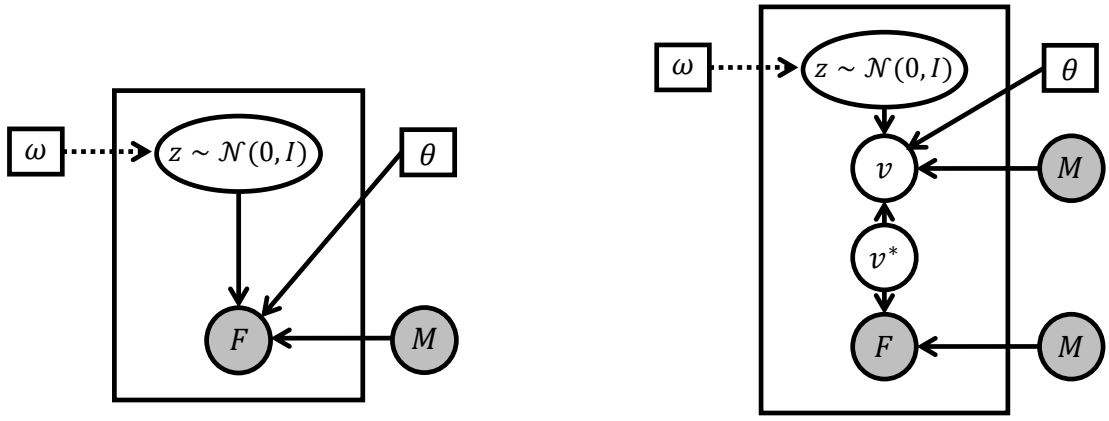
$$v^s = f_v^s(z, M^s; \theta^s). \tag{4.2}$$

In the training, the goal is to optimize θ^s such that all velocities v^s are likely to lead to warped moving images M^{*s} that will be similar to F^s in the training database. M^{*s} is obtained by exponentiation of v^s and warping of the moving image. Using Eq. 4.2, we can define the families of functions f^s:

$$M^{*s} := f^s(z, M^s; \theta^s) = M^s \circ \exp(f_v^s(z, M^s; \theta^s)). \tag{4.3}$$

In order to express the dependency of f^s on z and M^s explicitly, we can define a distribution $p(F^s|z, M^s; \theta^s)$. The product over the different scales gives us the output distribution:

$$p_\theta(F|z, M) = \prod_{s \in \mathcal{S}} p(F^s|z, M^s; \theta^s). \tag{4.4}$$



(a) (b)

Fig. 4.1: (a) Generative process for registration representing the likelihood of the fixed image F given the latent variable vector z and moving M: $p_\theta(F|z, M)$, where ω and θ are fixed parameters. (b) Generative process for regularized image registration where the likelihood depends on the regularized velocities $p_\theta(F|v^*, M)$.

By using the law of total probability, this leads to the following stochastic process for computing $p_\theta(F|M)$ which is also visualized in Fig. 4.1a (cf. [Kingma, 2014b]):

$$p_\theta(F|M) = \int_z p_\theta(F|z, M)p(z)dz. \tag{4.5}$$

The likelihood $p_\theta(F|z, M)$ can be any distribution that is computable and continuous in θ. In VAEs, the choice is often Gaussian, which is equivalent to adopting a sum-of-squared differences (SSD) criterion (cf. [Kingma, 2013]). We propose instead to use a local cross-correlation (LCC) distribution due to its robustness properties and superior results in image registration compared to SSD (cf. [Lorenzi, 2013; Avants, 2011]). Thus, we use the following Boltzmann distribution as likelihood:

$$p^s_\theta(F^s|z, M^s) \sim \exp(-\lambda \mathcal{D}_{LCC}(F^s, M^s, v^s)), \tag{4.6}$$

where $v^s = f^s_v(z, M^s; \theta^s)$ are the velocities and λ is a scalar hyperparameter. The symmetric $\mathcal{D}_{LCC}$ is defined as:

$$\mathcal{D}_{LCC}(F^s, M^s, v^s) = \frac{1}{P}\sum_{x\in\Omega} \frac{\sum_i \left(\left(F^{*s}_{x_i} - \overline{F^{*s}_x}\right)\left(M^{*s}_{x_i} - \overline{M^{*s}_x}\right)\right)^2}{\left(\sum_i \left(F^{*s}_{x_i} - \overline{F^{*s}_x}\right)^2\right)\left(\sum_i \left(M^{*s}_{x_i} - \overline{M^{*s}_x}\right)^2\right) + \tau} - 1, \tag{4.7}$$

with P pixels $x \in \Omega$, the symmetrically warped images $M^{*s} = M^s \circ \exp(v^s/2)$ and $F^{*s} = F^s \circ \exp(-v^s/2)$. The bar $\overline{F_x}$ symbolizes the local mean grey levels of F_x derived by mean filtering with kernel size k at position x. i is iterating through this $k \times k$-window. A small constant τ is added for numerical stability ($\tau = 1e^{-15}$).

Learning the constrained deformation encoding

In order to optimize the parameterized model over θ (Eq. 4.5), two problems must be solved: First, how to define the latent variables z, for example decide what information these variables represent. VAEs assume there is no simple interpretation of the dimensions of z but instead assert that samples of z are drawn from a simple distribution $p(z)$.

Second, the integral over z is intractable since one would need to sample a too large number of z's to get an accurate estimate of $p_\theta(F|M)$. Instead of sampling a large number of z's, the key assumption behind VAEs is to sample only z's that are likely to have produced F and compute $p_\theta(F|M)$ only from those. To this end, one needs to compute the intractable posterior $p(z|F,M)$. Due to this intractability, in VAEs [Kingma, 2013], the posterior is approximated by learning an *encoding* distribution $q_\omega(z|F,M)$, using a neural network with parameters ω (the encoder). This approximated distribution can be related to the true posterior using the Kullback-Leibler divergence (KL) which leads (after rearranging the terms) to the evidence lower bound (ELBO) of the log marginalized likelihood $\log p_\theta(F|M)$ (cf. [Kingma, 2013; Kingma, 2014b]):

$$\log p_\theta(F|M) - \mathrm{KL}\left[q_\omega(z|F,M)\|p(z|F,M)\right] = \mathbb{E}_{z\sim q}\Big[\log p_\theta(F|z,M)\Big] - \mathrm{KL}\left[q_\omega(z|F,M)\|p(z)\right]. \quad (4.8)$$

The KL-divergence on the left hand side gets smaller the better $q_\omega(z|F,M)$ approximates $p(z|F,M)$ and ideally vanishes if q_ω is of enough capacity. Thus, maximizing $\log p_\theta(F|M)$ is equivalent to maximizing the ELBO on the right hand side of the equation consisting of encoder q_ω and decoder p_θ which can be both optimized via stochastic gradient descent.

Optimizing the ELBO

According to the right-hand side of Eq. 4.8, there are two terms to optimize, the KL-divergence of prior $p(z)$ and encoder distribution $q_\omega(z|F,M)$ and the expectation of the reconstruction term $\log p_\theta(F|z,M)$. Since the prior is a multivariate Gaussian, the encoder distribution is defined as $q_\omega(z|F,M) = \mathcal{N}(z|\mu_\omega(F,M), \Sigma_\omega(F,M))$, where μ_ω and Σ_ω are deterministic functions learned in an encoder neural network with parameters ω. The KL-term can be computed in closed form as follows (constraining Σ_ω to be diagonal):

$$\mathrm{KL}[\mathcal{N}(\mu_\omega(F,M), \Sigma_\omega(F,M))\|\mathcal{N}(0,I)] = \frac{1}{2}\Big(\mathrm{tr}(\Sigma_\omega(F,M)) + \|\mu_\omega(F,M)\| - k - \log\det(\Sigma_\omega(F,M))\Big),$$

where k is the dimensionality of the distribution.

The expected log-likelihood $\mathbb{E}_{z\sim q}\left[\log p_\theta(F|z,M)\right]$, the reconstruction term, could be estimated by using many samples of z. To save computations, we treat $p_\theta(F|z,M)$ as $\mathbb{E}_{z\sim q}\left[\log p_\theta(F|z,M)\right]$ by only taking one sample of z. This can be justified as optimization is already done via stochastic gradient descent, where we sample many image pairs (F,M) from the dataset $\mathcal{X}$ and thus witness different values for z. This can be formalized with the expectation over $F, M \sim \mathcal{X}$:

$$\mathbb{E}_{F,M\sim\mathcal{X}}\Big[\mathbb{E}_{z\sim q}\Big[\log p_\theta(F|z,M)\Big] - \mathrm{KL}\left[q_\omega(z|F,M)\|p(z)\right]\Big].$$

To enable back-propagation through the sampling operation $q_\omega(z|F,M)$, the *reparametrization* trick [Kingma, 2013] is used in practice, where $z = \mu_\omega + \epsilon\Sigma_\omega^{1/2}$ (with $\epsilon \sim \mathbf{N}(0,I)$). Thus, for image pairs (F,M) from a training dataset $\mathcal{X}$ the actual objective becomes:

$$\mathbb{E}_{F,M\sim\mathcal{X}}\Big[\mathbb{E}_{\epsilon\sim\mathcal{N}(0,I)}\Big[\log p_\theta(F|z=\mu_\omega(F,M)+\Sigma_\omega^{1/2}(F,M)*\epsilon,M)\Big] - \mathrm{KL}\left[q_\omega(z|F,M)\|p(z)\right]\Big]. \quad (4.9)$$

After insertion of Eq. 4.4, the log of the product over the scales $s \in \mathcal{S}$ results in the sum of the log-likelihood distributions:

$$\mathbb{E}_{F,M\sim\mathcal{X}}\Big[\mathbb{E}_{\epsilon\sim\mathcal{N}(0,I)}\Big[\sum_{s\in\mathcal{S}}\log p_{\theta^s}(F^s|z=\mu_\omega(F,M)+\Sigma_\omega^{1/2}(F,M)*\epsilon,M^s)\Big] - \mathrm{KL}\left[q_\omega(z|F,M)\|p(z)\right]\Big]. \quad (4.10)$$

4.2.2 Introducing regularization on velocities

So far, we have considered that at each scale s, a velocity field v^s is generated by a decoding function $f_v^s(z,M^s;\theta^s)$ through a neural network. To have a better control of its smoothness, we propose to regularize spatially v^s through a Gaussian convolution with standard deviation σ_G:

$$\hat{v}^s = G_{\sigma_G} * v^s \quad (4.11)$$

Gaussian smoothing was applied here, but it could be replaced by any quadratic Tikhonov regularizer or by any functional enforcing prior knowledge about the velocity field.

In the remainder, we show how the regularization of velocities can be inserted into the proposed probabilistic framework. To make the notation less cluttered, we drop the scale

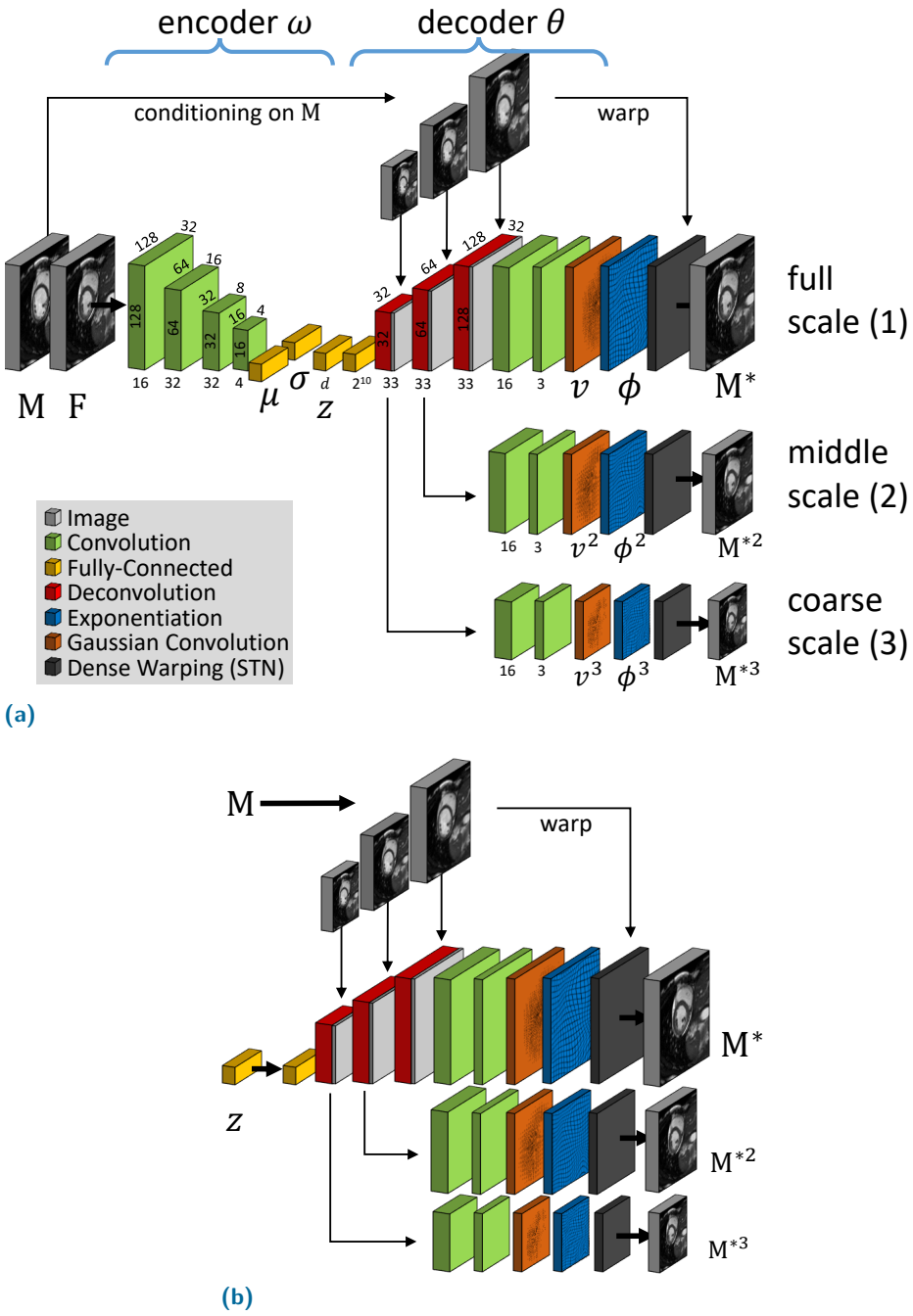


Fig. 4.2: (a) Probabilistic multi-scale registration network based on a CVAE. An encoder maps deformations to latent variables $z \in \mathbb{R}^d$ (with for example $d = 32$) from which a decoder extracts velocities and diffeomorphisms at different scales while being conditioned on the moving image M. (b) After training, the decoder network can be also used to sample and transport deformations: Apply z-code on any new image M.

s superscript in the velocity notations. Until now, the velocities v have been handled as fixed parameters $v = f_v(z, M; \theta)$. We can equivalently assume that velocities v are random variables with a Dirac posterior probability : $p_\theta(v|z, M) \sim \delta_{f_v(z,M;\theta)}(v)$.

We now introduce the random variable v^* which represents the regularized velocities as shown in Fig. 4.1b. This quantity is linked to the regular velocities v through a Gaussian distribution $p(v|v^*) = G(v^*, 1)$ such that v is close to v^* in terms of L_2 norm. Furthermore, we define a diffusion-like regularization prior on v^* [Nielsen, 1997]:

$$\log p(v^*) \propto \int_\Omega \sum_{i=1}^{\infty} \frac{\sigma_G^{2i}}{2^i i!} \left[\frac{\partial^i}{\partial \Omega^i} v^* \right]^2 d\Omega,$$

taking into account the Taylor expansion of the Fourier transform of the Gaussian. The maximum *a posteriori* of the regularized velocities $\hat{v}$ is then obtained through Bayes law:

$$\hat{v} = \arg\max_{v^*} \log p(v^*|v) = \arg\max_{v^*} \log p(v|v^*) + \log p(v^*)$$

which in this case is equivalent to solving the Heat equation [Nielsen, 1997] and leads to a Gaussian convolution: $\hat{v} = G_{\sigma_G} * v$.

Finally, we conveniently assume that the posterior probability of v^* is infinitely peaked around its mode, *i.e.* $p(v^*|v) \sim \delta_{\hat{v}}(v^*)$ (assumption sometimes made for the Expectation-Maximization algorithm [Kurihara, 2009]). In the decoding process, we can now marginalize out the velocity variables v and v^* by integrating over both such that only $\hat{v}$ remains:

$$\begin{aligned} p_\theta(F|M) &= \int_z \int_v \int_{v^*} p(F|v^*, M)\ p(v^*|v)\ p_\theta(v|z, M)\ p(z)\ dv\ dv^*\ dz \\ &= \int_z p(F|\hat{v}, M)\ p(z)\ dz. \end{aligned} \tag{4.12}$$

Thus, the proposed graphical model leads to a decoder working with the regularized velocity field $\hat{v}$ instead of the v generated by the neural network. When combining regularized velocities $\hat{v}^s$ at all scales, we get:

$$p_\theta(F|M) = \int_z \prod_{s \in \mathcal{S}} p(F^s|\hat{v}^s, M^s)\ p(z)\ dz. \tag{4.13}$$

This can be optimized as before and leads to Gaussian convolutions at each scale if considering diffusion-like regularization. Thus, the multi-scale loss function per training image pair (F,M) for one sample ϵ is defined as (cf. Eq. 4.10):

$$\arg\min_{\omega,\theta} \frac{1}{2}\Big(\text{tr}(\Sigma_\omega) + \mu_\omega^\top \mu_\omega - k - \log \det(\Sigma_\omega)\Big) - \lambda \sum_{s \in \mathcal{S}} \mathcal{D}_{LCC}(F^s, M^s, \hat{v}^s), \tag{4.14}$$

where $\hat{v}^s$ depends on v^s and therefore on θ (cf. Eq. 4.2 and 4.11).

4.2.3 Network architecture

The encoder-decoder neural network takes the moving and the fixed image as input and outputs the latent code z, velocities v, the deformation field ϕ and the warped moving image M^*. The last three are returned at the different scales s. The encoder consists of strided convolutions while the bottleneck layers (μ, σ, z) are fully-connected. The deconvolution layers in the decoder were conditioned by concatenating each layer's output with sub-sampled versions of $\mathbf{M}$. Making appearance information of the moving image easily accessible for the decoder, allows the network to focus on deformation information – the differences between moving and fixed image – that need to pass through the latent bottleneck. While it is not guaranteed that the latent representation contains any appearance information, it comes at a cost to use the *small* bottleneck for appearance information. At each decoding scale, a convolutional layer reduces the number of filter maps to three. Then, a Gaussian smoothing layer (cf. Eq. 4.11) with variance σ_G^2 is applied on these filter maps. The resulting velocities v^s (a SVF) are exponentiated by the *scaling and squaring* layer [Krebs, 2018] in order to retrieve the diffeomorphism ϕ^s which is used by a dense STN to retrieve the warped image M^{*s}. The latent code z is computed according to the reparametrization trick. During training, the network parameters are updated through back-propagation of the gradients with respect to the objective Eq. 4.10, defined at each multi-scale output. Finally during testing, registration is done in a single forward path where z is set to μ since we want to execute registration deterministically. One can also think of drawing several z using σ and use the different outputs for uncertainty estimation as in [Dalca, 2018] which we do not further pursue in this work. The network architecture can be seen in Fig. 4.2a. Besides registration, the trained probabilistic framework can be also used for the sampling of deformations as shown in Fig. 4.2b.

4.3 Experiments

We evaluate our framework on cardiac intra-subject registration. End-diastole (ED) frames are registered to end-systole (ES) frames from cine-MRI of healthy and pathological subjects. These images show large deformations. Additionally, we evaluate the learned encoding of deformations by visualizing the latent space and transporting encoded deformations from one patient to another. All experiments are in 3-D.

Data

We used the 334 ED-ES frame pairs of short-axis cine-MRI sequences. 184 cases were acquired from different hospitals and 150 cases were used from the Automatic Cardiac

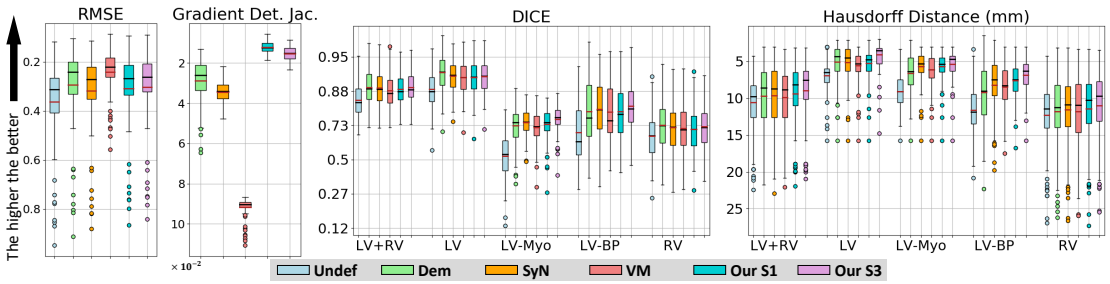


Fig. 4.3: Boxplots of registration results comparing the undeformed (Undef) case to the different algorithms: lcc-demons (Dem), SyN, voxelmorph (VM) and our single scale (S1) respectively multi-scale (S3) using RMSE, gradient of the determinant of the Jacobian, DICE scores (logit-transform) and Hausdorff distances (HD in mm). Mean values are denoted by red bars. Higher values are better.

Diagnosis Challenge (ACDC) at STACOM 2017 [Bernard, 2018], mixing congenital heart diseases with images from adults. We used 234 cases for training and for testing 100 cases from ACDC, that contain segmentation and disease class information. The testing set contained 20 cases of each of the following cardiac diseases: dilated cardiomyopathy (DCM), hypertrophic cardiomyopathy (HCM), previous myocardical infarction (MINF), abnormal right ventricle (RV) and healthy (Normal). All images were resampled with a spacing of $1.5 \times 1.5 \times 3.15$ mm and cropped to a size of $128 \times 128 \times 32$ voxels, by equally removing voxels from all sides. These dimensions were chosen to save computation time and are not a limitation of the framework.

Implementation details

Our neural network consisted of four encoding convolutional layers with strides (2, 2, 2, 1) and three decoding deconvolutional layers. Each scale contained two convolutional layers and a convolutional Gaussian layer with $\sigma_G = 3mm$ (kernel size 15) in front of an exponentiation and a spatial transformer layer using trilinear interpolation (cf. Fig. 4.2a). The dimensionality of the latent code z was set to $d = 32$ as a compromise of registration quality and generalizability (cf. experiment on latent vector dimensionality). The number of trainable parameters in the network was ∼420k. LeakyReLu activation functions and L2 weight decay of $1 * 10^{-4}$ were applied on all layers except the last convolutional layer in each scale where a *tanh* activation function was used in order to avoid extreme velocity values during training. All scales were trained together, using linearly down-sampled versions of the input images for the coarser scales. In all experiments, the number of iterations in the exponentiation layer was set to $N = 4$ (evaluated on a few training samples according to the formula in [Arsigny, 2006]). During the training, the mean filter size of the LCC criterion was $k = 9$. The loss hyper parameter was empirically chosen as $\lambda = 5000$ such that the similarity loss was optimized while the latent codes roughly had zero means and variances of one. We applied a learning rate of $1.5 * 10^{-4}$

with the Adam optimizer and a batch size of one. For augmentation purposes, training image were randomly shifted, rotated, scaled and mirrored. The framework has been implemented in *Tensorflow* using *Keras*[1]. Training took ∼24 hours and testing a single registration case took 0.32s on a *NVIDIA GTX TITAN X* GPU.

Registration

We compare our approach with the LCC-demons (Dem, [Lorenzi, 2013]) and the ANTs software package using Symmetric Normalization (SyN, [Avants, 2008]) with manually tuned parameters (on a few training images) and the diffeomorphic DL-based method VoxelMorph [Dalca, 2018] (VM) which has been trained using the same augmentation techniques as our algorithm. For the latter, we set $\sigma = 0.05$, $\lambda = 50000$ and applied a reduced learning rate of $5 * 10^{-5}$ for stability reasons while using more training epochs. Higher values for λ led to worse registration accuracy. We also show the improvement of using a multi-scale approach (with 3 scales, S3) compared to a single-scale objective (S1).

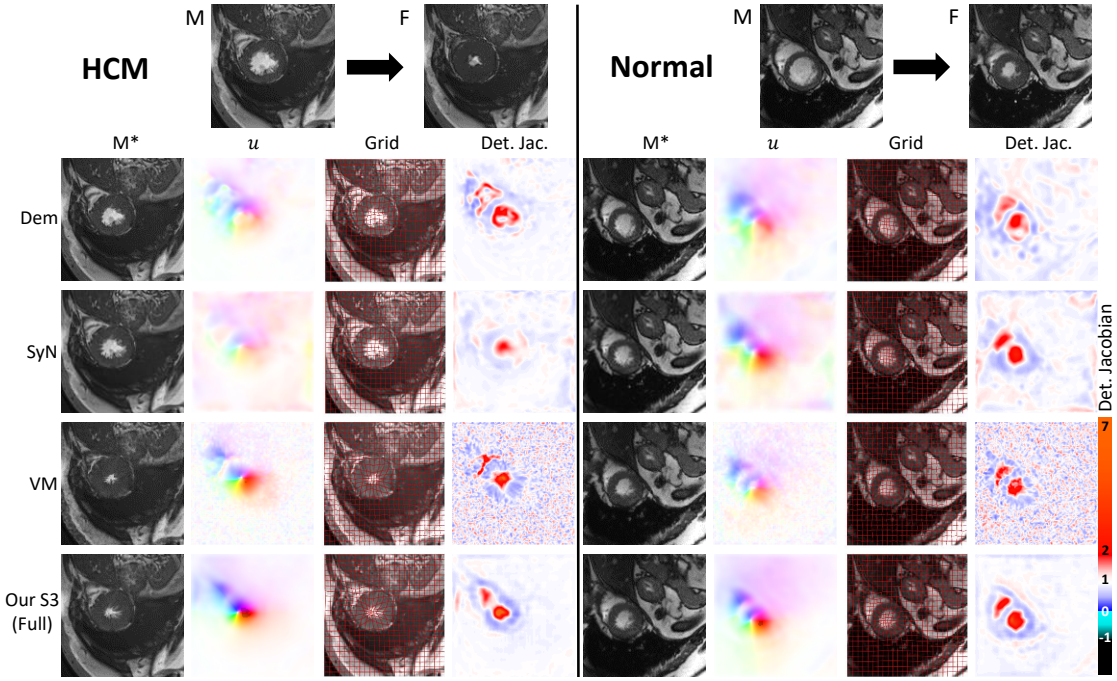
We measure registration performance with the following surrogates: intensity root mean square error (RMSE), DICE score, 95%-tile Hausdorff distance (HD in mm). To quantify deformation regularity, we show the determinant of the Jacobian qualitatively, while we also computed the mean magnitude of the gradients of the determinant of the Jacobian (Grad Det-Jac). We decided to report this second-order description of deformations to better quantify differences in smoothness among the different methods, which are not obvious by taking the mean of the determinant of the Jacobian as bigger and smaller values tend to cancel each other out. DICE and HD scores were evaluated on the following anatomical structures: myocardium (LV-Myo) and epicardium (LV) of the left ventricle, left bloodpool (LV-BP), right ventricle (RV) and LV+RV (Fig. 4.5).

Table 4.1 shows the mean results and standard deviations of all algorithms. In terms of DICE scores, our algorithm using three scales (Our S3) shows the best performances on this dataset while the single-scale version (Our S1) performed similarly compared to the LCC-demons and the SyN algorithm. Hausdorff distances were significantly improved using both of our algorithms. Detailed registration results are shown in Fig. 4.3. Interestingly, we found that the SyN algorithm showed marginally better DICE scores than the LCC-demons which has been also reported on brain data [Lorenzi, 2013].

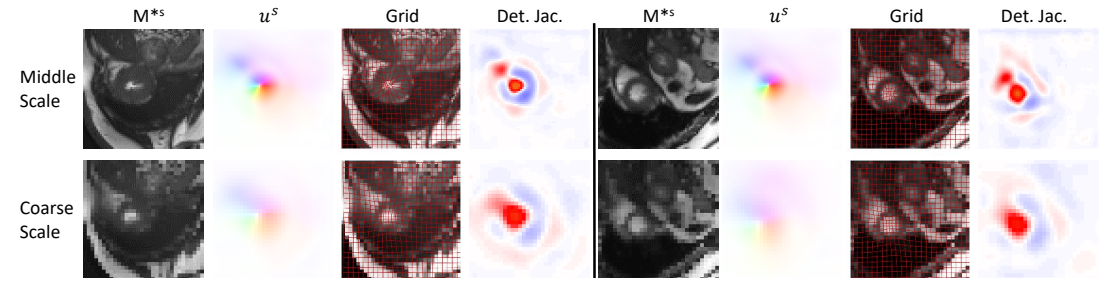
Qualitative registration results of a pathological (HCM) and a healthy case (Normal) are presented in Fig. 4.4a[2]. The warped moving image (with and wihout grid overlay) and

[1] https://keras.io/

[2] Qualitative registration results for all five diseases are also presented in Fig. 10-12 (available in the supplementary files /multimedia tab).



(a)



(b)

Fig. 4.4: (a) Qualitative registration results showing a pathological (hypertrophy) and a normal case. Warped moving image M^*, displacements u, warped moving image with grid overlay and Jacobian determinant are shown for LCC-demons (Dem), SyN, voxelmorph (VM) and our approach using 3 scales (Our S3). (b) Middle and coarse scale predictions of our multi-scale method (Our S3).

the determinant of the Jacobian (Det. Jac.) are shown. Displacements are visualized using the color encoding as typical for the optical flow in computer vision tasks. Middle and coarse scale outputs of our multi-scale method are shown in Fig. 4.4b. We computed the determinant of the Jacobian using *SimpleITK*[3] and found that for all methods no negative values were observed on our test dataset. Compared to the other algorithms, our approach produced smoother and more regular deformations as qualitatively shown by the determinant of the Jacobian in Fig. 4.4a and quantitatively by the significantly smaller mean gradients of the determinant of the Jacobian (Table 4.1)[4]. Despite the fact of being diffeomorphic, the voxelmorph algorithm produced more irregular deformation

[3] `http://www.simpleitk.org/`

[4] Visualization of the gradients of the Jacobian determinant are presented in Fig. 13 (available in the supplementary files /multimedia tab).

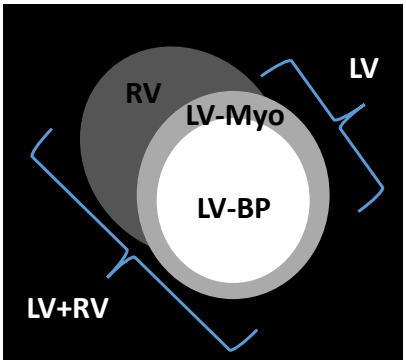


Fig. 4.5: Cardiac structures used only for measuring registration accuracy.

fields compared to all other algorithms. Our single-scale approach resulted in slightly smoother deformations which is probably due to the fact that it performed less accurately in compensating large deformations.

Tab. 4.1: Registration performance with mean and standard deviation scores (in brackets) of RMSE, DICE, Hausdorff Distance (HD in mm) and the mean gradient of the determinant of Jacobians (Grad Det-Jac, $\times 10^{-2}$) comparing the undeformed case (Undef), LCC-demons (Dem), SyN, voxelmorph (VM) and our method.

Method	RMSE	DICE	HD	Grad Det-Jac
Undef	0.37 (0.17)	0.707 (0.145)	10.1 (2.2)	–
Dem	0.29 (0.16)	0.799 (0.096)	8.3 (2.7)	2.9 (1.0)
SyN	0.32 (0.16)	0.801 (0.091)	8.1 (3.6)	3.4 (0.5)
VM	**0.24 (0.08)**	0.790 (0.096)	8.4 (2.6)	9.2 (0.5)
Our S1	0.31 (0.15)	0.797 (0.093)	7.9 (2.6)	**1.2 (0.3)**
Our S3	0.30 (0.14)	**0.812 (0.085)**	**7.3 (2.7)**	1.4 (0.3)

We applied the Wilcoxon signed-rank test with $p < 0.001$ to evaluate statistical significance of the differences in the results of Fig. 4.3. This method is chosen as a paired test without the assumption of normal distributions. For all metrics, the results of our multi-scale algorithm (Our S3) showed significant differences compared to the results of all other methods (including Our S1). With respect to our single-scale algorithm (Our S1), only the differences in DICE scores were not statistically significant in comparison with the LCC-demons (Dem).

Note, that higher DICE and HD scores can be achieved by choosing a higher latent dimensionality (cf. Experiment 4.3), which however comes at the cost of a more complex encoding space, making analysis tasks more difficult. We also tested the first version of voxelmorph [Balakrishnan, 2018] on our dataset. We chose to show the results of the latest version [Dalca, 2018] due to the fact that this version is diffeomorphic and that its DICE and HD results were better (cf. [Krebs, 2018]).

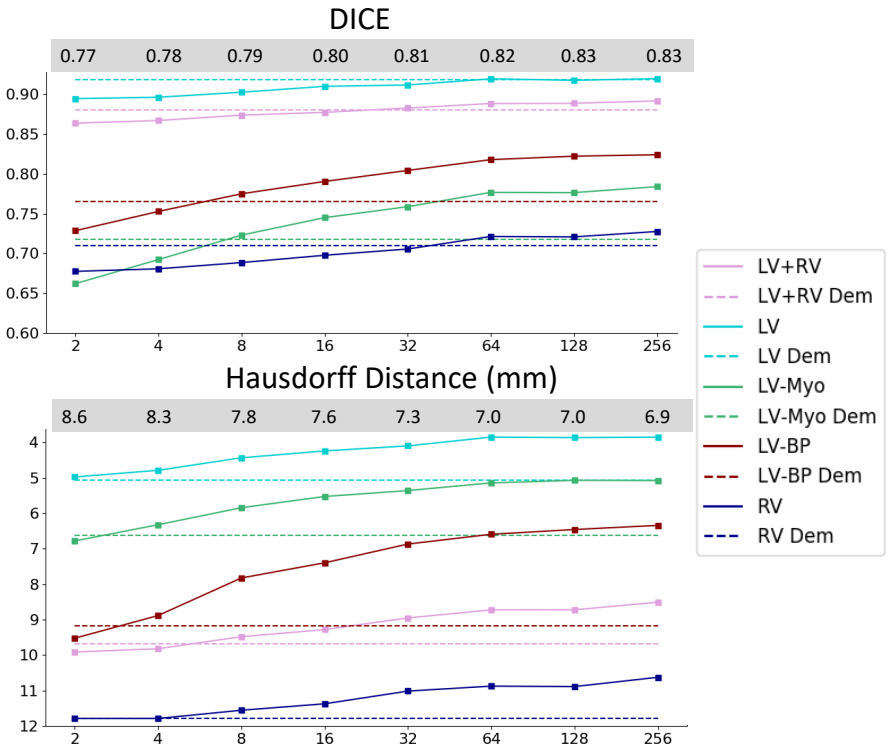


Fig. 4.6: Showing the influence of the latent vector size d on the registration accuracy in terms of DICE and Hausdorff distances in mm of the different anatomical structures with the mean of all structures shown in the grey boxes. The performance of the LCC-demons (Dem) is shown as reference with dashed lines.

Deformation encoding

For evaluating the learned latent space, we investigated (a) the effects of the size of the latent vector on the registration accuracy, (b) the structure of the encoded space by visualizing the distribution of cardiac diseases and showing simulated deformations along the two main axes of variations and (c) we applied our framework on deformation transport and compare its performance with a state-of-the-art algorithm.

Latent Vector Size In Fig. 4.6 we analyzed the influence of the size of the latent code vector with respect to registration accuracy in terms of DICE and HD scores. With a relatively small latent size of $d = 8$, competitive accuracy is achieved. With an increasing dimensionality, performance increases but reaches a plateau eventually. This behavior is expected, since CVAEs tend to ignore components if the dimensionality of the latent space is too high [Kingma, 2014b]. For the cardiac use case, we chose $d = 32$ components as a trade-off between accuracy and latent variable size.

Disease Distribution and Generative Latent Space In this experiment, we used disease information and encoded z-codes of the test images to visualize the learned latent space. Using linear CCA (canonical correlation analysis), we projected the z-codes (32-D) to a

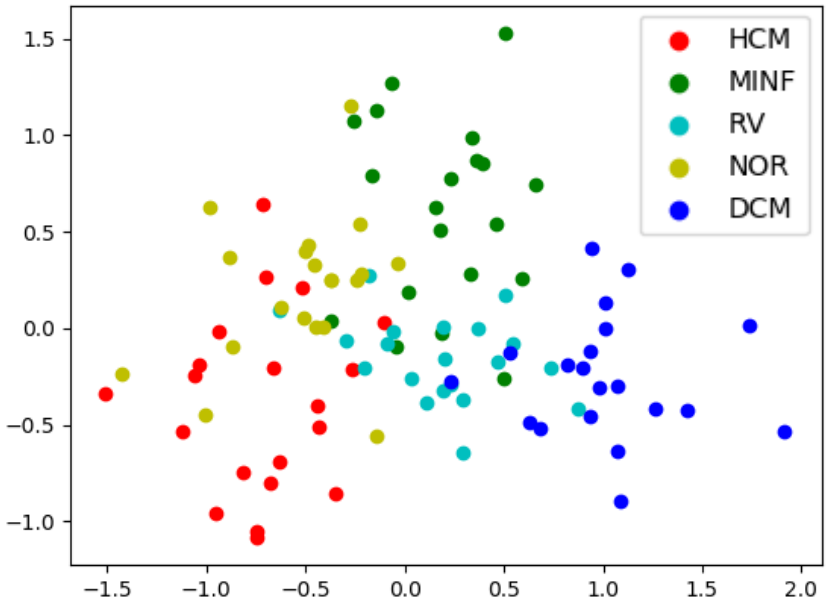


Fig. 4.7: Cardiac disease distribution after projecting the latent variables z of the test images on a 2-D CCA (canonical correlation analysis) space. Using an 8-D CCA and applying SVM with 10-fold cross-validation leads to a classification accuracy of 83%

2-D space by using the two most discriminative CCA components. Fig. 4.7 shows that the 100 test sets are clustered by classes in this space. Taking the 8 most discriminative CCA components into account, the classification accuracy of the five classes is **83%** with 10-fold cross-validation using support vector machines (SVM). In a second experiment, we applied principal component analysis (PCA) on the z-codes of the training dataset. We simulated deformations by sampling equally distributed values in the range of ± 2.5 standard deviations of the two largest principal components and extracting the z-codes through inverse projections. Fig. 4.8 shows reconstructed displacements and deformed images when applying these generated z-codes on a random test image. One can see the different influences of the two eigenvalues. The first eigenvalue (horizontal) focuses on large deformations while the second one focuses on smaller deformations as the right ventricle. The results of these two experiments which are solely based on applying simple linear transformations, suggest that deformations that are mapped close to each other in the deformation latent space have similar characteristics.

Deformation Transport Pathological deformations can be transported to healthy subjects by deforming the healthy ES frames using pathological ED-ES deformations. Our framework allows for deformation transport by first registering the ED-ES frames of a given pathological case (Step 1), which we call *prediction* in this experiment. Secondly, we use the z-codes from these predictions along with the ED frame of a healthy subject to transport the deformations (Step 2, cf. Fig. 4.2b). Note, that this procedure does not require any inter-subject registrations.

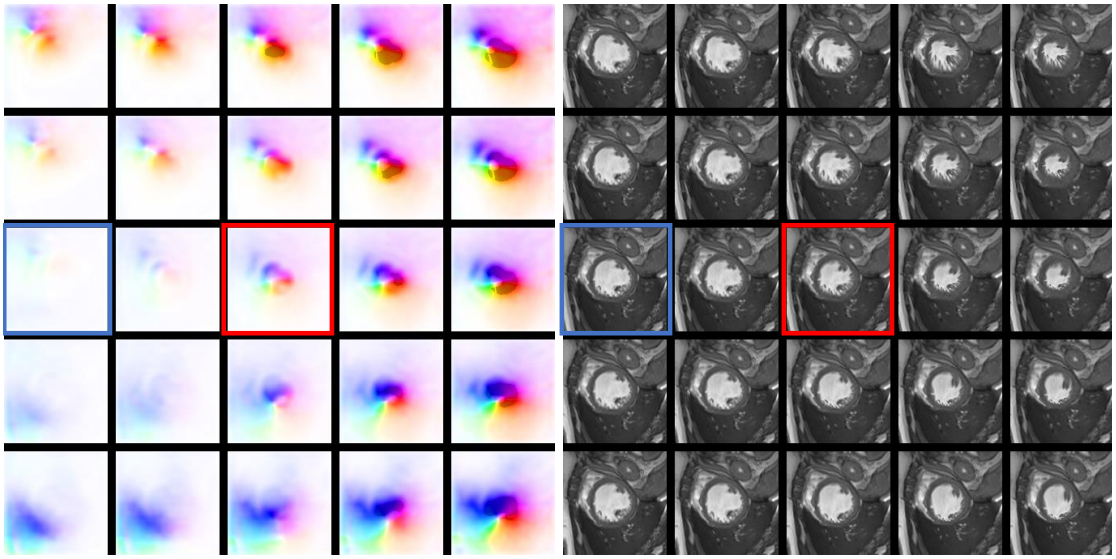


Fig. 4.8: Reconstruction of simulated displacements and an accordingly warped random test image after generating z-codes by equally sampling along the two largest principal components within a range of ± 2.5 sigma around their mean values (red box). The PCA was fitted using all training z-codes. The blue box indicates the image closest to the identity deformation. One can see that the horizontal eigenvalue influences large deformations while the vertical eigenvalue focuses on smaller ones, for example the right ventricle.

We compare our approach with the pole-ladder algorithm (PL [Lorenzi, 2014]). All intra- and inter-subject registrations required by the pole ladder were performed using the LCC-demons [Lorenzi, 2013]. For the inter-subject pairs, we aligned the test data with respect to the center of mass of the provided segmentation and rotated the images manually for rigid alignment. This alignment step was done only for the pole ladder experiment[5].

Qualitative results are shown in Fig. 4.9 where the predicted deformations of one hypertrophy (A, HCM) and one cardiomyopathy (B, DCM) case (step 1) were transported to two healthy (Normal) subjects (step 2, targets C and D). Note that our algorithm automatically determines orientation and location of the heart. In Table 4.2, we evaluated the average ejection fraction (EF) of the ED-ES deformation prediction of the pathologies (step 1) and the average EF after transport to normal subjects (step 2). Hereby, we assume that EFs, as a relative measure, stay similar after successful transport (such that the absolute difference, EF step 1 - EF step 2, is small). The table shows the average of transporting 5 HCM and 5 DCM cases to 20 normal cases (200 transports). For our algorithm, the absolute differences in EFs are much smaller for DCM cases and similarly close in HCM cases in comparison to the pole ladder. All test subjects were not used during training. The EF is computed based on the segmentation masks (warped with the resulting deformation fields). Besides, it can be seen, that predictions done by the

[5]The pipeline for the parallel transport experiment using the pole ladder algorithm is presented in Fig. 14 (available in the supplementary files /multimedia tab).

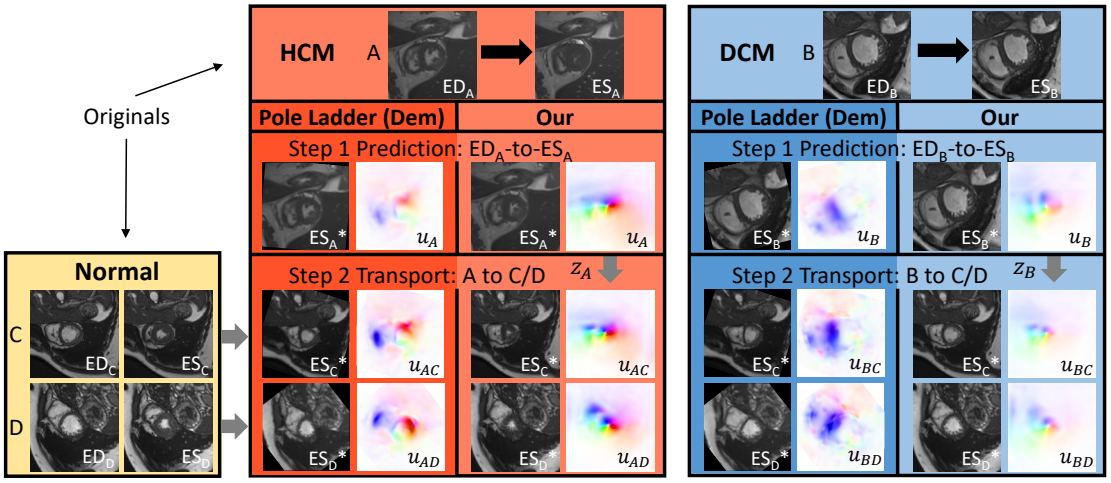


Fig. 4.9: Transport pathological deformation predictions (Step 1, hypertrophy HCM, myopathy DCM) to healthy (Normal) subjects by using the pole ladder (with LCC-demons) and our probabilistic method (Step 2). Note that the pole ladder algorithm requires the registration between pathological and normal subjects while our approach is able to rotate and translate deformations encoded in the latent space z automatically.

Tab. 4.2: Mean Ejection fraction (EF in % with standard deviation in parentheses) of pathological deformation predictions (Step 1) should stay similar to the mean EF after the transport to healthy/normal subjects (Step 2). Our algorithm shows smaller absolute differences compared to the pole ladder (PL).

	Step 1: Prediction		Step 2: Transport		Difference	
Path.	PL (Dem)	Our	PL (Dem)	Our	PL	Our
HCM	29.4 (6)	44.1 (7)	35.5 (8)	38.6 (13)	6.1	**5.5**
DCM	10.8 (2)	12.7 (7)	16.7 (4)	13.9 (7)	5.9	**1.2**

demons are underestimating the EFs for HCM cases which should be >40% according to the ACDC data set specifications.

4.4 Discussion and Conclusions

We presented an unsupervised multi-scale deformable registration approach that learns a low-dimensional probabilistic deformation model. Our method not only allows the accurate registration of two images but also the analysis of deformations. The framework is generative, as it is able to simulate deformations given only one image. Furthermore, it provides a novel way of deformation transport from one subject to another by using its probabilistic encoding. In the latent space, similar deformations are close to each other. The method enables the addition of a regularization term which leads to arbitrarily smooth deformations that are diffeomorphic by using an exponentiation layer for stationary velocity fields. The multi-scale approach, providing velocities, deformation fields and warped images in different scales, leads to improved registration results and a more controlled training procedure compared to a single-scale approach.

We evaluated the approach on end-diastole to end-systole cardiac cine-MRI registration and compared registration performance in terms of RMSE, DICE and Hausdorff distances to two popular algorithms [Lorenzi, 2013; Avants, 2008] and a learning-based method [Dalca, 2018], which are all diffeomorphic. While the performance of our single-scale approach was comparable to the LCC-demons and the SyN algorithm, our multi-scale approach (using 32 latent dimensions) showed statistically significant improvements in terms of registration accuracy. Generally, our approach produced more regular deformation fields, which are significantly smoother than the DL-based algorithm. Using our method with a non-generative U-net style network [Ronneberger, 2015] without a deformation encoding performed similarly compared to the proposed generative model. Adding supervised information such as segmentation masks in the training procedure as in [Hu, 2018; Fan, 2019] led to a marginal increase in terms of registration performance ($\sim$1-2% in DICE scores), so we decided that the performance gain is not large enough in order to justify the higher training complexity. Theoretically, our method allows measurement of registration uncertainty as proposed in [Dalca, 2018] which we did not further investigate in this work.

The analysis of the deformation encoding showed that the latent space projects similar deformations close to each other such that diseases can be clustered. Disease classification could be potentially enforced in a supervised way as in [Biffi, 2018]. Furthermore, our method showed comparable quantitative and qualitative results in transporting deformations with respect to a state-of-the-art algorithm which requires the difficult step of inter-subject registration that our algorithm does not need.

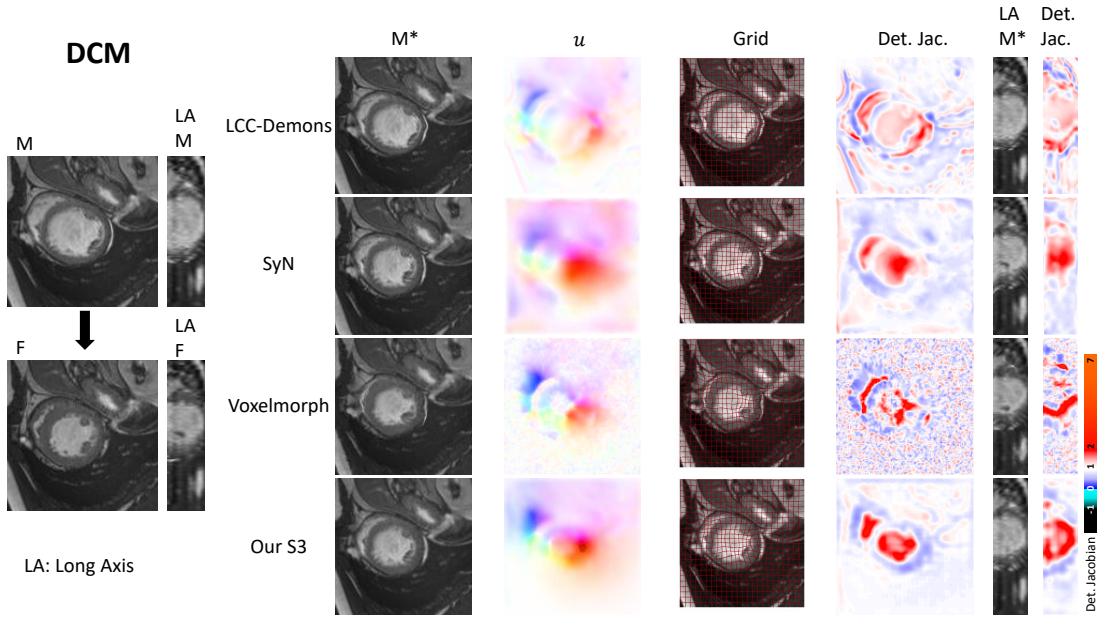
It is arguable if the simple assumption of a multivariate Gaussian is the right choice for the prior of the latent space (Eq. 4.1). Possible other assumptions such as a mixture of Gaussians are subject to future work. The authors think that the promising results of the learned probabilistic deformation model could be also applicable for other tasks such as evaluating disease progression in longitudinal studies or detecting abnormalities in subject-to-template registration. An open question is how the optimal size of the latent vector changes in different applications. In future work, we plan to further explore generative models for learning probabilistic deformation models.

Acknowledgments

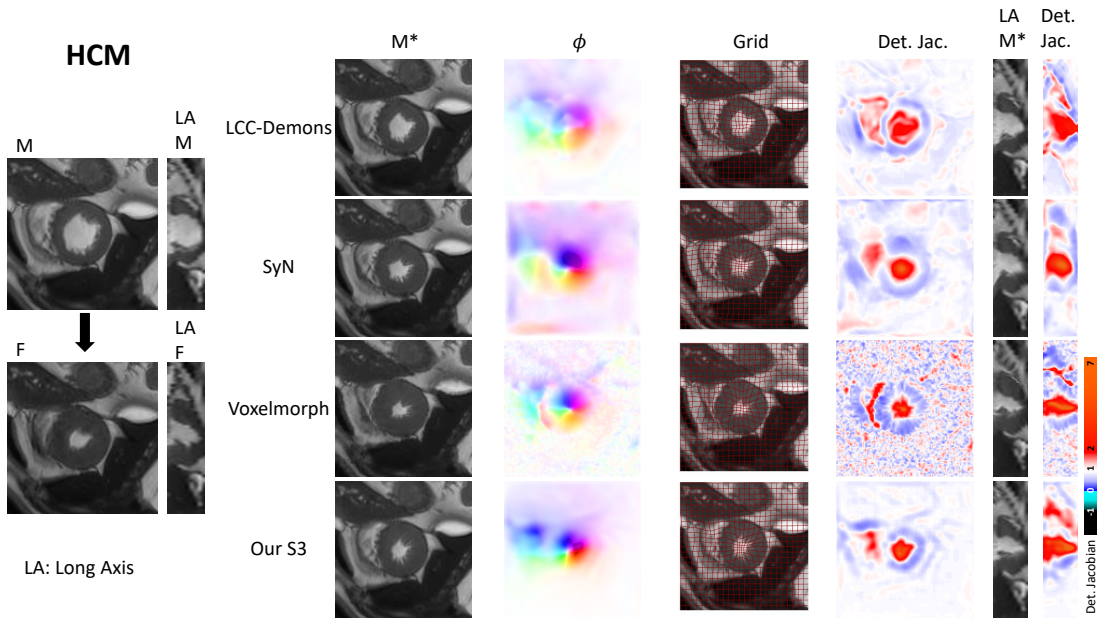
Data used in preparation of this article were obtained from the EU FP7-funded project MD-Paedigree and the ACDC STACOM challenge 2017 [Bernard, 2018]. This work was supported by the grant AAP Santé 06 2017-260 DGA-DSH, and by the Inria Sophia Antipolis - Méditerranée, "NEF" computation cluster. The authors would like to thank

Xavier Pennec for the insightful discussions and Adrian Dalca for the help with the Voxelmorph [Dalca, 2018] experiments.

4.5 Appendix

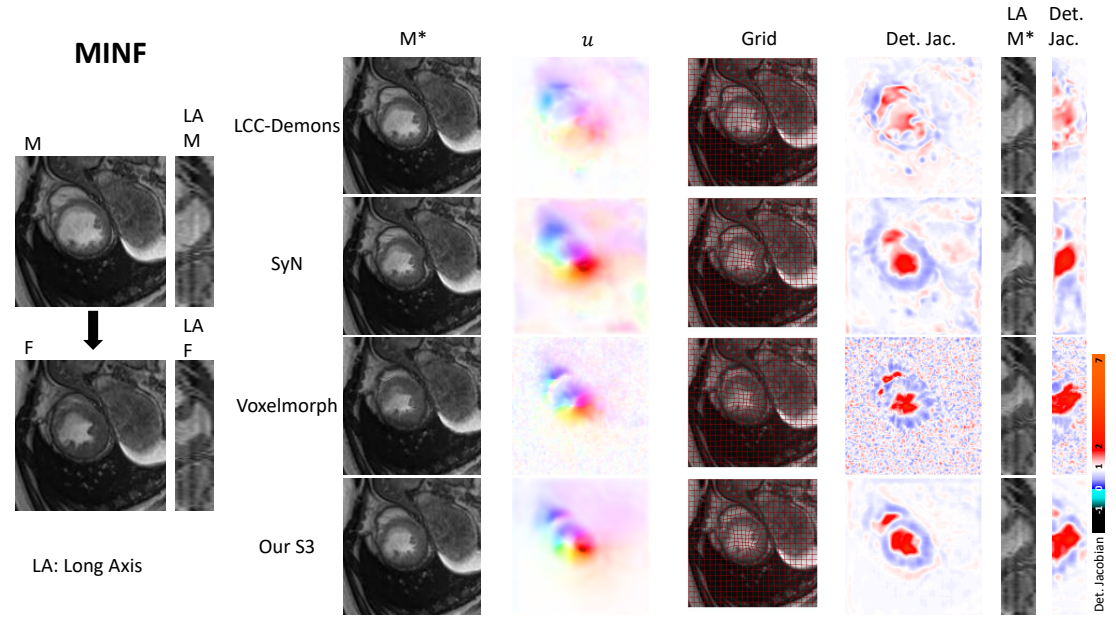


(a)

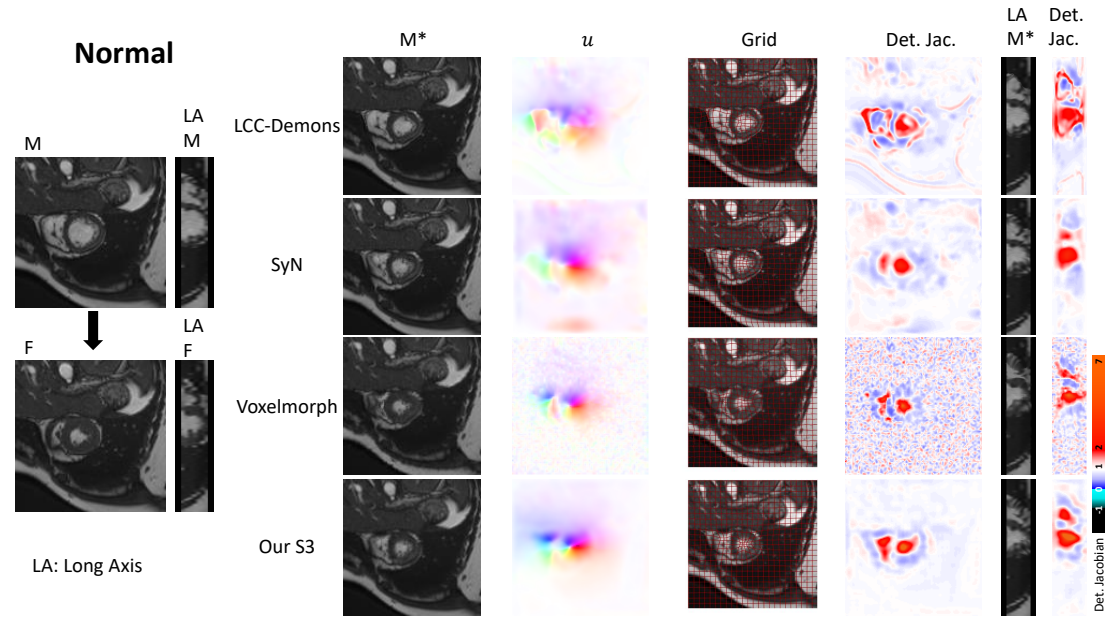


(b)

Fig. 4.10: Qualitative registration results showing a dilated cardiomyopathy (DCM) and a hypertrophic cardiomyopathy (HCM) case. Warped moving image M^*, displacements u, warped moving image with grid overlay and Jacobian determinant are shown for LCC-demons (Dem), SyN, voxelmorph (VM) and our approach using 3 scales (Our S3).



(a)



(b)

Fig. 4.11: Qualitative registration results showing a myocardical infarction (MINF) and healthy (Normal) case. Warped moving image M^*, displacements u, warped moving image with grid overlay and Jacobian determinant are shown for LCC-demons (Dem), SyN, voxelmorph (VM) and our approach using 3 scales (Our S3).

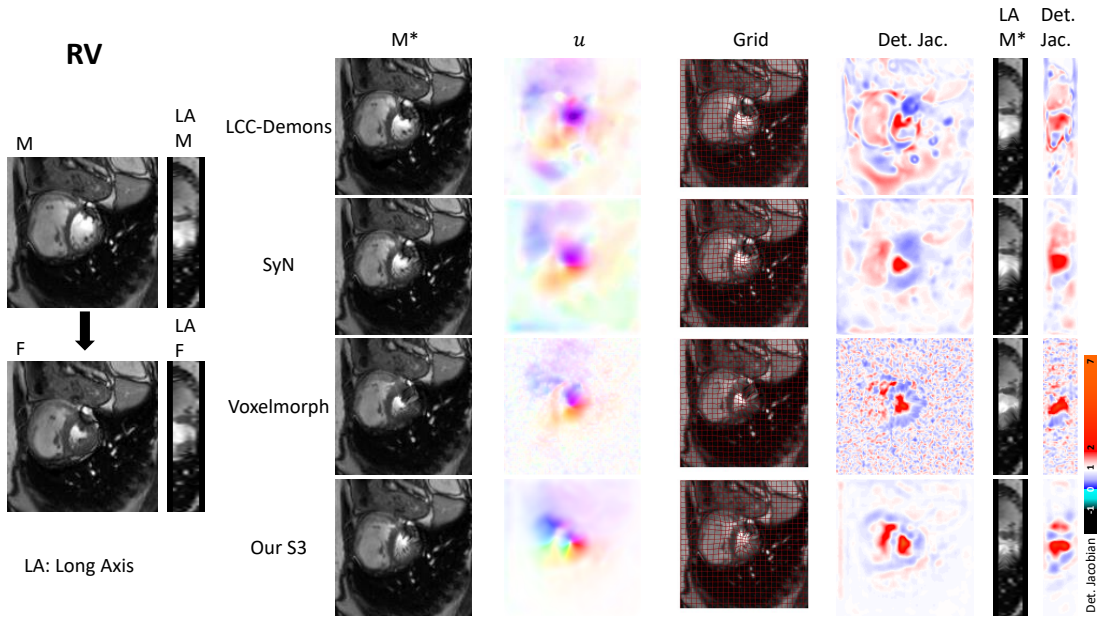


Fig. 4.12: Qualitative registration results showing an abnormal right ventricle case (RV). Warped moving image M^*, displacements u, warped moving image with grid overlay and Jacobian determinant are shown for LCC-demons (Dem), SyN, voxelmorph (VM) and our approach using 3 scales (Our S3).

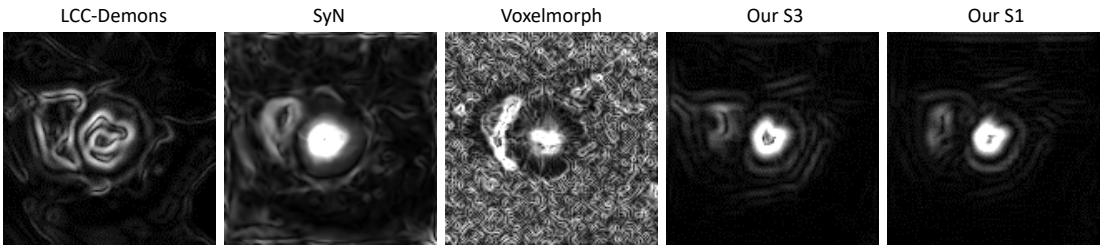
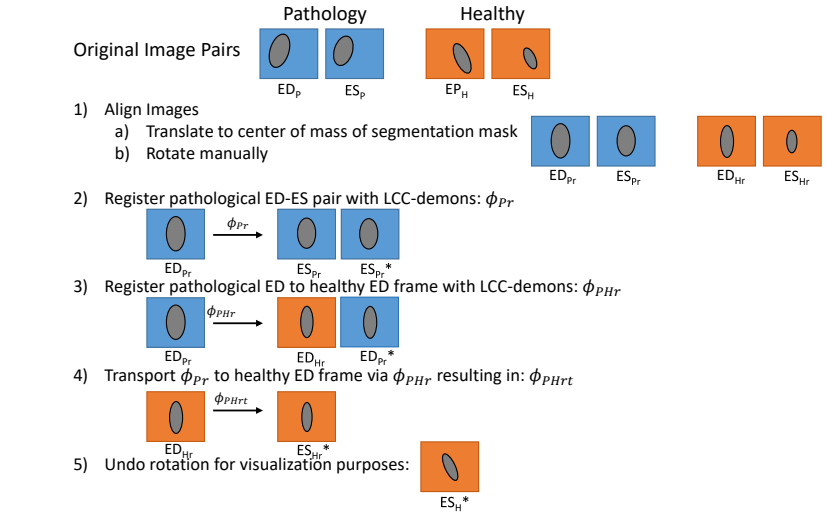
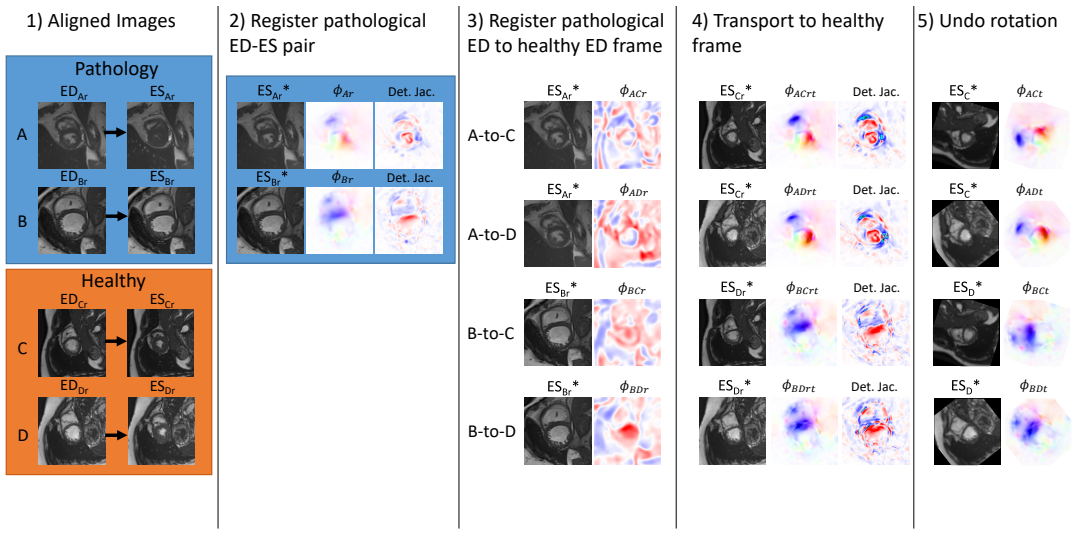


Fig. 4.13: The gradient of the determinant of the Jacobian of a random test case for LCC-demons (Dem), SyN, voxelmorph (VM) and our approach using 1 and respectively 3 scales (Our S1, Our S3). Our single-scale approach shows the most regular deformation.



(a)



(b)

Fig. 4.14: (a) Symbolic pipeline for the parallel transport experiment using the pole ladder approach. (b) Visualization of all pipeline steps for one example.

5 Learning a Generative Motion Model from Image Sequences based on a Latent Motion Matrix

Contents

In chapter 4, we presented a pairwise probabilistic deformation model and showed its applicability to a variety of deformation analysis tasks. In this chapter, we extend the model to a generative motion model that learns population-specific motion patterns from a database of image sequences. Such a motion model enables consistent tracking of structures, the simulation and temporal interpolation of motion. A temporal conditional variational autoencoder is implemented using a novel Gaussian process prior assumption. This chapter is based on the conference presentation at STACOM 2019 [Krebs, 2020b]. However, the presented version includes several methodological advancements and is currently under review as a journal paper.

5.1 Introduction

Motion analysis is an important task in many medical image analysis problems such as organ tracking or longitudinal analysis of various diseases. For moving organs such as the heart, it is not only important to track anatomical structures but also to analyze motion indices that are useful for disease diagnosis or therapy selection [Girija, 2017]. Extracting motion patterns further allows to compensate for motion, handle missing data or do temporal super-resolution and motion simulation.

Motion in medical image sequences is typically analyzed by computing temporally consistent pairwise deformations where each frame in a sequence is registered to a target frame [Girija, 2017]. The resulting series of deformation fields can be utilized to track structures throughout the sequence and to identify abnormal motion patterns, for example by computing clinically relevant variables such as the ejection fraction (EF) of the heart [Rohé, 2018].

5.1.1 State-of-the-art

Registration algorithms typically seek to find the deformation field between two images by solving an optimization problem consisting of a similarity metric and a regularizer. The similarity metric measures the distance between the two images while the regularizer constrains the smoothness of the resulting deformation field. A large variety of registration algorithms using different similarity and regularizing metrics have been proposed [Sotiras, 2013]. One group of registration methods aim to ensure diffeomorphic deformations due to their favorable properties. Diffeomorphisms are topology-preserving and invertible deformations which makes them suitable for many medical registration problems in which foldings are physically implausible [Vercauteren, 2009]. This makes diffeomporphisms also appropriate for tracking anatomical structures in image sequences such as in cardiac imaging [Peyrat, 2010] (assuming structures do not go out of the field of view). Many diffeomorphic registration algorithms have been proposed such as [Beg, 2005; Zhang, 2015; Vercauteren, 2008; Vercauteren, 2009], the SyN algorithm [Avants, 2008] and the LCC-demons [Lorenzi, 2013]. Recently, learning-based algorithms for pairwise diffeomorphic registration have been proposed. These are based on supervised *ground-truth* deformations [Yang, 2017; Rohé, 2017] or on unsupervised learning [Dalca, 2018; Krebs, 2019b]. The latter are trained by minimizing a loss function consisting of an image similarity and a deformation regularizer, similarly to the traditional optimization problem. In these two works, diffeomorphisms are guaranteed by using the stationary velocity field (SVF) parameterization based on the scaling-squaring algorithm [Arsigny, 2006].

For image sequences, one difficulty is to acquire temporally smooth deformations that are fundamental for consistent tracking. That is why registration algorithms with a temporal regularizer have been proposed [LedesmaCarbayo, 2005; Vandemeulebroucke, 2011; De Craene, 2012; Metz, 2011; Qin, 2018; Shi, 2013]. In the computer vision community, temporal video super-resolution and motion compensation are a related research topic [Caballero, 2017; Kappeler, 2016].

However, while these methods are able to capture temporally consistent deformations along a sequence of images, they do not extract intrinsic motion parameters crucial for building a comprehensive motion model that can be used for analysis tasks such as motion simulation, transport or classification as it is for example done in bio-mechanical models such as [Sermesant, 2008]. Yang et al. [Yang, 2011a] generated a motion prior using manifold learning from low-dimensional shapes. Qiu et al. [Qiu, 2011] proposed to build an eigenspace of initial momenta using PCA. In an image-driven fashion, Rohé et al. [Rohé, 2018] introduced a parameterization, the Barycentric Subspaces, for cardiac motion analysis.

5.1.2 Learning a Probabilistic Motion Model

In contrast, we propose a probabilistic motion model that is built in a fully data-driven way from image sequences. Instead of defining a motion parameterization explicitly or learning from pre-processed shapes, our model learns a low-dimensional motion matrix in an unsupervised fashion. The goal is not only to retrieve a compact representation of the motion but to obtain a structured and generative encoding that allows for temporal interpolation (to predict missing frames) and to simulate an indefinite number of new motion patterns. These features could be helpful for data augmentation and to speed-up image acquisition as the model reconstructs a full cyclic motion from missing frames. Besides, the learned probabilistic encoding could be useful for group-wise analysis as it enables to transport motion characteristics to a new subject, simulating for example a pathological motion in a healthy subject.

In this work, we introduce a novel Gaussian Process (GP) prior to extend a conditional variational autoencoder (CVAE [Kingma, 2014b]), a latent variable model, for temporal sequences. A pairwise encoder-decoder neural network applies a temporal convolutional network (TCN) in its latent space in order to learn intrinsic temporal dependencies. Furthermore, we utilize a self-supervised training scheme based on temporal dropout (TD) to enforce temporal consistency and increase generalizability of the motion model. Smooth and diffeomorphic deformations are guaranteed by applying an exponentiation layer [Krebs, 2019b] and spatio-temporal regularization.

The proposed model demonstrates state-of-the-art registration accuracy measured on segmentation overlaps and distances and regularity for diffeomorphic tracking of cardiac cine-MRI. In addition, the potentials of the generated latent motion matrix for motion simulation, interpolation and transport are demonstrated. The main contributions are as follows:

- An unsupervised probabilistic motion model learned from medical image sequences
- A conditional VAE model trained with a novel Gaussian process prior and self-supervised temporal dropout using temporal convolutional networks
- Demonstration of cardiac motion tracking, simulation, transport and temporal super-resolution

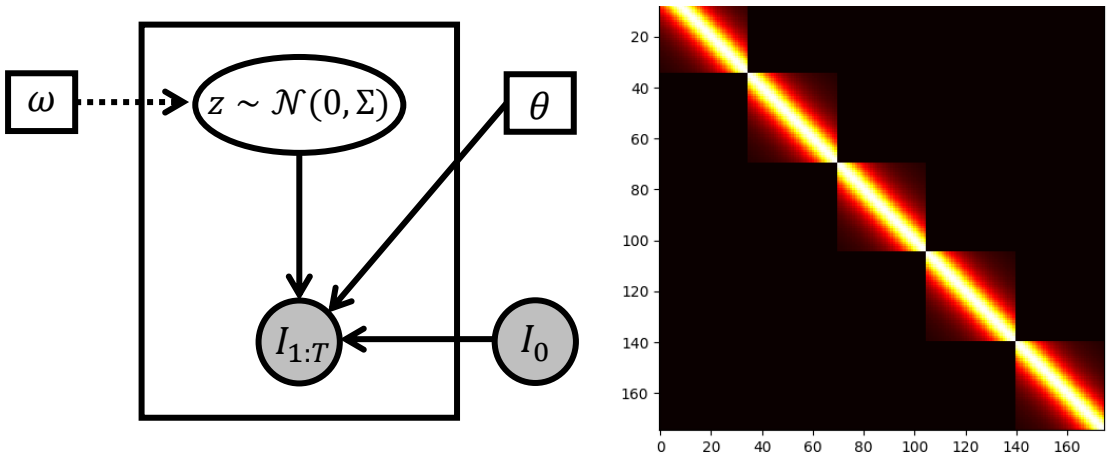
This paper extends our preliminary conference paper [Krebs, 2020c] by replacing the standard unit Gaussian of the CVAE with a novel Gaussian Process Prior. We add detailed derivations of the motion model and show improved tracking accuracy and temporal smoothness. Finally, we show a first generalization of the model to 3-D+t sequences.

5.2 Methods

Typically, the motion of an image sequence $I_{0:T}$ with T frames is described by deformation fields between one reference image, for example I_0, and all other images in the sequence. In order to extract consistent sequential deformations ϕ_t with $t \in [1, T]$, we propose a temporal latent variable model that encodes the motion in a low-dimensional probabilistic space, the motion matrix $z \in \mathbb{R}^{D \times \bar{T}}$ with $\bar{T} = T - 1$. Here, we define the reference image I_0 as moving image, while the other frames are fixed images I_t. Each image pair (I_0, I_t) is encoded by D latent variables, the z_t-code, which are the columns of z. Each z_t parameterizes the deformation field ϕ_t while being conditioned on the moving image I_0. The rows z_d with length $\bar{T}$ of the motion matrix z represent the encoded deformation sequence per latent dimension $d \in D$.

Our motion model is learned from data by imposing a Normal prior distribution $p(z)$ on the latent variables z that follows a Gaussian Process (GP) prior in the temporal dimension for each z_d. In addition, we assume independence between the latent variables z_d as in standard VAEs [Kingma, 2013]. Note, when z is written as part of a distribution like $p(z)$, z is used as a vector of size $D\bar{T}$ rather than a matrix for simpler notation.

During training, we follow the learning paradigms of conditional variational autoencoders (CVAE [Kingma, 2014b; Kingma, 2014a]) with the exception of replacing the multivariate unit Gaussian prior with the proposed GP-prior. The approximated posterior is the output of a temporal convolutional neural network (TCN [Bai, 2018]) allowing for temporal



(a) (b)

Fig. 5.1: (a) Generative process for the motion model representing the likelihood of fixed images $I_{1:T}$ given the latent variables z and moving image I_0: $p_\theta(I_{1:T}|z, I_0)$, where ω and θ are fixed parameters and arrows denote dependencies between random variables. (b) Visualization of the covariance matrix Σ of the Gaussian prior $p(z)$ with 5 latent dimensions, a sequence time length of 35 and a length scale of the Cauchy kernel of 7.

regularization. To further facilitate temporal dependencies and handle missing data, temporal dropout (TD) is applied during the training procedure. In the following, the different parts of the method are explained. First, the probabilistic motion model using a GP-prior is defined. Then, posterior and data likelihood distributions are modeled using a encoder-decoder neural network. Lastly, the concept of temporal dropout is introduced.

5.2.1 Generative Motion Model using a Gaussian Process Prior

The proposed motion model consists of an encoder $q_\omega(z|I_{0:T})$ and a decoder $p_\theta(I_{1:T}|z, I_0)$ which are parameterized by ω and θ respectively. The encoder first independently maps each image pair (I_0, I_t) to a latent representation γ_t which is then temporally regularized by mixing all time steps to retrieve the motion matrix z. The decoder p_θ projects the z_t-codes to the deformations ϕ_t while being conditioned on the moving image I_0. The output of the decoder are he reference image I_0 warped with the ϕ_t deformation fields. The encoder approximates the posterior distribution and the decoder the data likelihood of the latent variable model. Using a prior distribution $p(z)$ over latent variables z, we define the following generative process:

$$p_\theta(I_{1:T}|I_0) = \int_z p_\theta(I_{1:T}|z, I_0)p(z)\, dz, \tag{5.1}$$

which is visualized in Fig. 5.1a. In this work, encoder q_ω and decoder p_θ are approximated using neural networks where ω and θ represent the encoder and decoder networks' weights which are optimized using amortized Variational Inference [Kingma, 2013]. The

data likelihood $p_\theta(I_{1:T}|z, I_0)$ can be seen as the fidelity of the reconstruction of the fixed images $I_{1:T}$ by warping the moving image I_0 with appropriate deformations $\phi_{1:T}$. An overview of the motion model can be seen in Fig. 5.2.

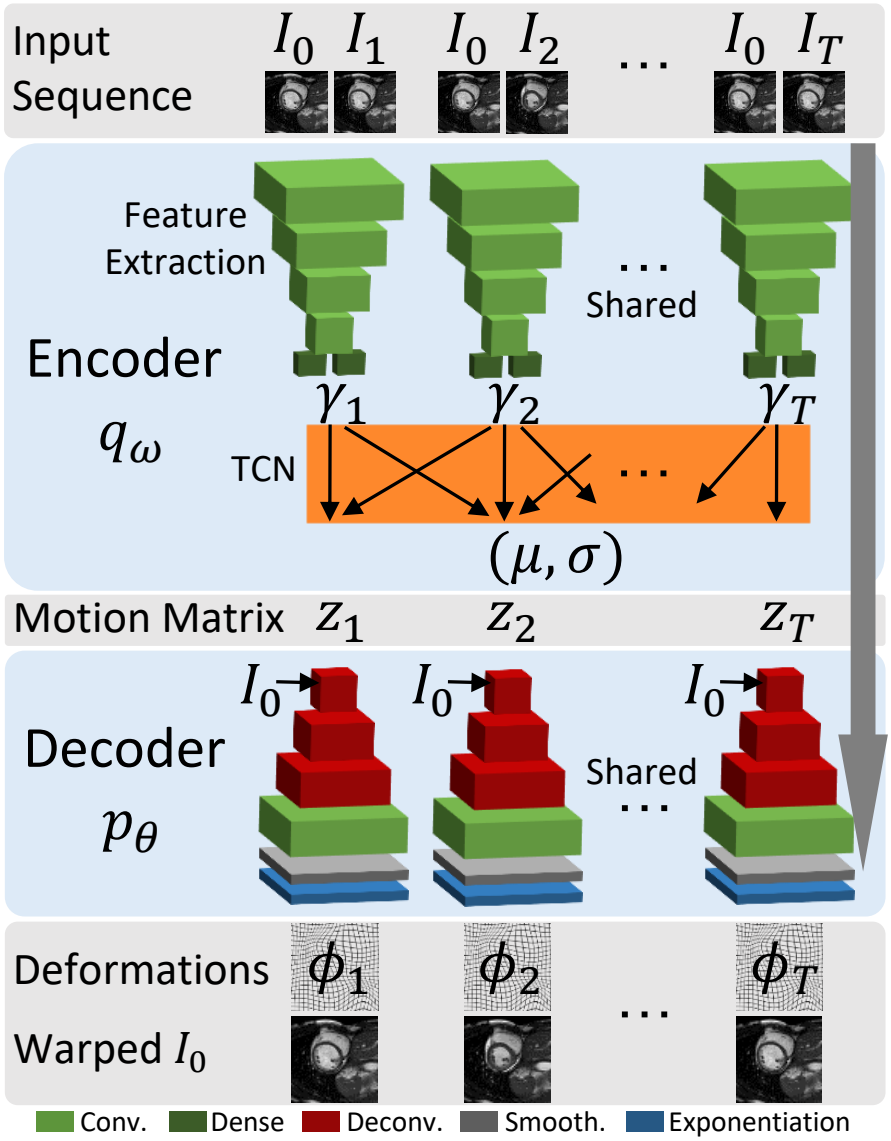


Fig. 5.2: Overview of the motion model including encoder and decoder neural networks. From sequential image pairs, temporally independent feature vectors γ_t are extracted which are fed to a temporally convolutional network (TCN) to obtain the probabilistic motion matrix z. This compact representation is decoded to a sequence of diffeomorphic deformation fields ϕ_t.

Gaussian Process Prior

The prior follows a zero-centered multivariate Gaussian distribution: $p(z) \sim \mathcal{N}(0|\Sigma)$ where the covariance matrix Σ is a diagonal block matrix of dimensions $D\bar{T} \times D\bar{T}$:

$$\Sigma = \text{Diag}_{d=1}^{D}(K_l). \tag{5.2}$$

Each diagonal element of Σ represents the temporal covariance matrix $K_l \in \mathbb{R}^{\bar{T}\times\bar{T}}$ of a Gaussian time-continuous stochastic process whose kernels can be chosen by the user. A typical choice in Gaussian Processes is the squared exponential kernel $K_l^{\text{RBF}}(\tau, \tau') = \sigma_K^2 \exp\left(-|\tau - \tau'|^2/2l^2\right)$ with length scale l and variance σ_K^2. However, due to the fact that we want to model data that varies at multiple time scales, we consider the Cauchy kernel [Rasmussen, 2003; Fortuin, 2019]:

$$K_l^{\text{Cauchy}}(\tau, \tau') = \sigma_K^2 \left(1 - \frac{(\tau - \tau')^2}{l^2}\right)^{-1}, \tag{5.3}$$

with pre-defined σ_K. This covariance matrix Σ allows temporally correlated latent variables while still assuming highest possible independence between the D latent dimensions. In other words, we extended the standard VAE latent space which only consists of the independence assumption between latent variables with a regularized temporal dimension. Latent variables are related over time according to the chosen kernel function K_l while being independent of each other. An example of a covariance matrix can be seen in Fig. 5.1b.

Posterior and Likelihood Distributions

Similar to standard VAEs, the posterior q_ω follows a multivariate Gaussian distribution $q_\omega(z|I_{0:T}) \sim \mathcal{N}(\mu|\Sigma^*(\sigma))$ with data-driven predictions of mean vector $\mu \in \mathbb{R}^{D\bar{T}}$ and variance vector $\sigma \in \mathbb{R}^D$. The full covariance matrix $\Sigma^*(\sigma)$ is defined as a block diagonal matrix of the following form:

$$\Sigma^*(\sigma) = \text{vec}\Big(\big(\sigma \mathbf{1}^\top\big)^\top\Big)\Sigma = \begin{bmatrix} \sigma_1 K_l & 0 & \cdots & 0 \\ 0 & \sigma_2 K_l & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_D K_l \end{bmatrix}, \tag{5.4}$$

where $\mathbf{1}$ defines a vector of ones of size $\bar{T}$ and vec$(\cdot)$ describes the vectorization function. Mean and variance vectors (μ, σ) are the output of the encoder neural network. The kernel K_l is kept the same as in the prior distribution and does not contain predicted parameters to guarantee a user-chosen temporal regularity.

Also, the likelihood p_θ is assumed to follow a multivariate Gaussian distribution $p_\theta(I_{1:T}|z, I_0) \sim \mathcal{N}(I_0 \circ \phi_{1:T}(\theta); 0|\sigma_L * I_{D\bar{T}})$ where $I_{D\bar{T}}$ is the identity matrix of size $D\bar{T}$, f_θ is the decoder neural network that outputs the diffeomorphisms $\phi_{1:T}$ and $\circ$ denotes the image warping operation. The variance σ_L is chosen to be a scalar constant, depicting for example the variance of intensity residuals of well registered images.

Learning the Motion Model via Variational Inference

In order to optimize the parameterized motion model over ω and θ, the evidence lower bound (ELBO) of the log-marginalized likelihood $p_\theta(I_{1:T}|I_0)$ that is conditioned on the moving image I_0, must be maximized (see [Kingma, 2013; Kingma, 2014b; Krebs, 2019b] for details):

$$\mathbb{E}_{z\in q_\omega(\cdot|I_{0:T})}\big[\log p_\theta(I_{1:T}|z, I_0)\big] - \mathrm{KL}\big[q_\omega(z|I_{0:T})\|p(z)\big], \tag{5.5}$$

with KL denoting the Kullback-Leibler Divergence (KL). The first term in Eq. 5.5 enforces that the moving image I_0 is well registered to the fixed images $I_{1:T}$ by maximizing the log likelihood. The second term structures the latent motion encoding by enforcing the posterior distribution $q_\omega(z|I_{0:T})$) to be close to the prior distribution $p(z)$. Following the definition of the KL divergence between 2 multivariate Gaussian distributions, we obtain the closed-from solution (see Appendix A):

$$\mathrm{KL}\big[q_\omega(z|I_{0:T})\|p(z)\big] = \frac{1}{2}\sum_{i=1}^{D} \sigma_i^2\bar{T} + \bar{\mu}_i^\top K^{-1}\bar{\mu}_i - \log\left(\sigma_i^2\right) - \bar{T}, \tag{5.6}$$

with $\bar{\mu}_i$ being the i-th segment of length T in μ.

Recall that the log likelihood $p_\theta(I_{1:T}|z, I_0)$ is also Gaussian. Thus, $\log p_\theta(I_{1:T}|z, I_0) = -\frac{1}{2}\sum_{t=1}^{T}\|I_t - I_0 \circ \phi_t\|^2/\sigma_L$ plus a constant which is equivalent to adopting a sum-of-squared differences (SSD) criterion, commonly used as similarity metric in image registration (for example in [Balakrishnan, 2018]).

During training of the model, parameters ω and θ are updated via stochastic gradient descent and back-propagation. In order to back-propagate through the sampling operation, the reparameterization trick is used [Kingma, 2013]. For full-covariance Gaussian distributions, the covariance matrix must be positive-definite as we use the Cholesky decomposition for the reparameterization (cf. [Kingma, 2019]). The details on how to efficiently compute the Cholesky decomposition of the covariance matrix Σ^* in Eq. 5.4 can be found in Appendix B.

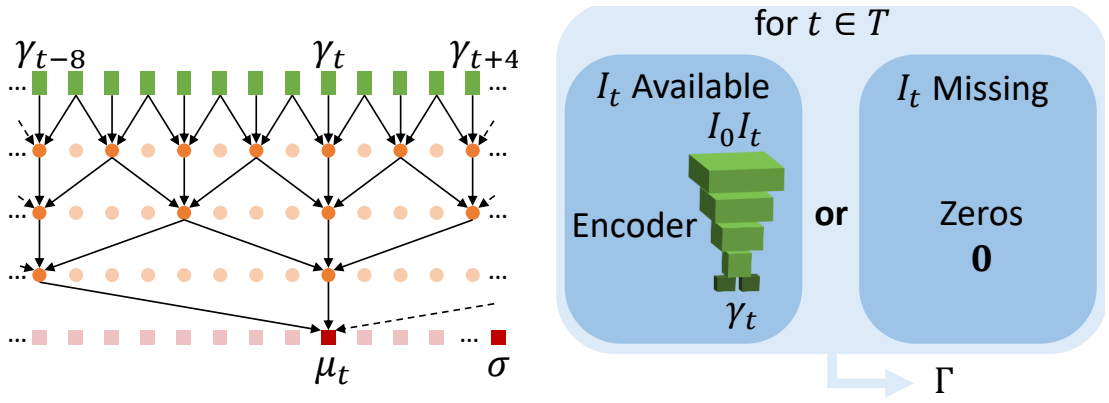


Fig. 5.3: (a) The temporal convolutional network (TCN) allows for temporal regularization of the independently extracted features γ_t per time step t, for retrieving mean vector μ and variance vector σ of the posterior distribution p_θ. (b) Sequences with missing time steps (motion interpolation or simulation) are encoded by a full feature matrix Γ by setting the columns of missing time steps to zero. The TCN handles these missing columns and still predicts a full temporal motion sequence of $\bar{T}$ time steps.

Diffusion-like regularization in spatial and temporal dimensions is applied by Gaussian smoothing kernels. This regularization follows the derivations of [Krebs, 2019b] and is omitted in Fig. 5.1a for reasons of clarity.

Neural Network Architecture

The encoder takes the image pairs (I_0, I_t) as input and outputs the motion matrix z. It consists of a feature extraction part and a temporal regularizer (TCN). The feature extraction part consists of convolutional and fully-connected layers for mean and variance predictions of the posterior [Kingma, 2014b]. These layers are temporally independent and share weights across all image pairs of a sequence. As the output of the feature extraction networks, the extracted features γ_t of size $\mathbb{R}^{2D}$ are merged across different time steps by using a temporal convolutional network (TCN) leading to temporally regularized mean and variance vectors (μ, σ) that define the posterior distribution $q_\omega(z|I_{0:T}) \sim \mathcal{N}(\mu|\Sigma^*(\sigma))$. The size of $2D$ is chosen for γ_t such that each σ value can be influenced by features from the whole sequence. Note, that samples from the posterior distribution are vectors of size $D\bar{T}$ which are reshaped to retrieve the motion matrix z with z_t-columns.

Following the recommended architecture, the TCN consists of multiple 1-D convolutional layers with increasing dilation and skip connections allowing to learn temporal dependencies of the latent variables γ_t that were time-independent before [Bai, 2018]. We use zero-padding and non-causal convolutional layers to also take future time steps into account. The output tensor capturing (μ, σ) is of size $\mathbb{R}^{D\bar{T}+D}$. Our TCN is shown in

Fig. 5.3a. TCNs can handle sequences of varying time lengths and are advantageous compared to recurrent neural networks (RNN) due to a flexible receptive field and more stable gradient computations [Bai, 2018]. Another reason why the authors chose a TCN over RNNs is that RNNs are especially suitable to learn long-distance temporal relationships such as in natural language processing while the focus of this work is on rather short time sequences with higher local dependencies. One could use a cyclic padding instead of zero-padding for cyclic sequences, for example by linking the end of a sequence to its beginning. However, in the case of cardiac cine-MRI, 5-10% of the cardiac cycle are often omitted [Bernard, 2018] such that we chose to not assume cyclic sequences explicitly.

The decoder takes as input samples z_t from the posterior distribution and the moving image I_0 and outputs the diffeomorphisms $\phi_{1:T}$ and the accordingly warped moving image. Deconvolutional and convolutional layers are used in the decoder which are shared across all time steps. It is desired that the latent representation z encodes deformation information on a semantic level, independent of the given subject. That is why the decoder is further conditioned on the moving image I_0 by concatenating down-sampled versions of I_0 with the outputs of the deconvolutional layers at different scales. By providing subject-specific appearance information in form of the moving image, the motion model is driven to encode subject-independent deformation information in the limited dimensionality of z [Krebs, 2019b]. In order to ensure smooth and diffeomorphic deformations, we utilize a Gaussian smoothing layer with standard deviations of σ_G and σ_T in temporal and spatial domains respectively and an exponentiation layer for the stationary velocity field parameterization of diffeomorphisms [Krebs, 2019b]. The linear warping functionality is realized using a spatial transformer network layer [Jaderberg, 2015].

5.2.2 Missing Data and Temporal Dropout

To always predict a full sequence of T deformations, the size of the covariance matrix Σ^* is kept identical across datasets with different time lengths T^*. In case of shorter sequences, the features γ_τ of all available image pairs (I_0, I_τ) with $\tau \in T^*$ are extracted and evenly distributed along T forming the matrix $\Gamma \in \mathbb{R}^{2D \times \bar{T}}$. The remaining missing time steps are filled with a constant (typically zero). On the decoder side, the log-likelihood loss (first part of Eq. 5.5) is evaluated on all available time steps of the original sequence. If a sequence is longer than T, evenly distributed frames would be dropped to reach a lentgh of T. However, this should not happen normally as we assume to put T at least as the maximum experienced length in the data.

In addition, during training, further time steps (i.e. γ_τ) are dropped from Γ using temporal dropout (TD) in order to force the motion model to interpolate motion between

available frames. To encourage the TCN to make use of its temporal connections and search for dependencies across time, our TC drops some of the γ_τ while still trying to recover the deformations ϕ_τ of all available image pairs (I_0, I_τ). More precisely, in TD, instead of extracting features from an image pair (I_0, I_τ), a vector of zeros is chosen as γ_τ while still keeping the loss function on the decoder part for these time steps. A binary Bernoulli random variable r_τ is used to randomly choose at each original time step τ if the zero vector is used instead of the extracted features given (I_0, I_τ). All independent Bernoulli random variables $r \in \mathbb{R}^{T^*}$ have the success probability δ. The latent feature representation γ_t^{TD} using TD can thus be defined as:

$$\gamma_\tau^{TD} = r_\tau * \mathbf{0} + (1 - r_\tau) * \gamma_\tau. \tag{5.7}$$

Note, TD is used only during training as a sort of self-supervision to encourage generalizability and consistent motion simulation and interpolation of missing data. When encountering missing data at test time, one just needs to place the available encoded frame pairs at the desired temporal positions of Γ in order to predict the full motion consisting of T time steps (cf. Fig. 5.3b). A full motion simulation can be generated by setting all elements of Γ to zero. In this case, a sequence of deformations that are plausible with respect to the training data will be predicted given only the original image I_0.

Optional Random Sub-Sequence Training: Since our motion model takes sequences of images as input and outputs a sequence of deformation fields, it comes naturally with high computational costs. This can lead to a model that may not be trainable on standard GPUs. Due to this limitation, we propose to train our model optionally with random sub-sequences. Let $\mathcal{T}$ be the maximum number of frames with which our model can be trained on a given GPU. In each training iteration, a random combination of $\mathcal{T}$ frames is selected from a training subject with T^* frames in case $T^* > \mathcal{T}$. After sorting this combination, the given frame pairs are encoded and placed at their relative temporal position in Γ. In contrast to the TD procedure, only the selected $\mathcal{T}$ time steps are reconstructed in the decoder to limit the requirements of GPU memory. In case of shorter training sequences with $T^* \leq \mathcal{T}$, the full sequence is used. By sampling different sub-sequences in each training epoch, the network will eventually see all parts of a sequence during the training stage.

5.3 Experiments

In this paper, we evaluate the proposed motion model on cardiac cine-MRI. Besides accurate temporal tracking and registration, we show the model's capabilities for motion simulation, interpolation and transport. The improved temporal latent space using the GP prior is demonstrated. Extensive results are presented for 2D+T sequences with more

limited quantitative evaluations on 3D+T sequences due to their heavy computational requirements. In all experiments, the end-diastolic (ED) frame was used as the moving image I_0.

5.3.1 Databases

Two datasets forming 334 cardiac cine-MRI in total were used. First, 184 multi-centric short-axis sequences came from the EU FP7-funded project MD-Paedigree (Grant Agreement 600932), with congenital heart disease and healthy or pathological images from adults. In addition, 150 sequences originated from the Automatic Cardiac Diagnosis Challenge 2017 (ACDC [Bernard, 2018]). The images were acquired in breath hold using 1R-R or 2R-R intervals mixing retrospective or prospective gating. The original sequence lengths varied from 13 to 35 frames. The 100 *training* cases from ACDC that contain ED-ES segmentation information were used for testing while all other sequences were used for training. Slices were resampled with a spacing of 1.5×1.5 mm and cropped to a size of 128×128 pixels. In case of 3D+T sequences, 18 slices were used by adding zero slices at the top and bottom in case of fewer original slices.

5.3.2 Implementation Details

The time-independent neural network parts, the feature extraction part of the encoder and the decoder, followed the architecture proposed in [Krebs, 2019b]. The feature extractor consisted of 4 convolutional layers with (2,2,2,1)-strides and (16,32,32,4)-feature maps and a fully-connected layer of size $2D$, outputting γ_t. The decoder p_θ consisted of a 3 deconvolutional and 1 convolutional layer with (32,32,32,16)-feature maps. The TCN consisted of 4 1-D convolutional layers with (1,2,4,8)-dilations, *same* padding and skip connections (cf. Fig. 5.3a). All (de-)convolutional layers used a kernel size of 3. The last convolutional layer of the decoder was followed by a spatio-temporal Gaussian layer with spatial $\sigma_G = 3$mm and temporal standard deviation $\sigma_T = 1.5$, an exponentiation layer using 6 *scaling-squaring* iterations [Krebs, 2019b] and a linear warping layer [Jaderberg, 2015].

The latent dimensionality was set to $D = 32$ (as in [Krebs, 2019b]). We set the sequence length T to 35, the maximum sequence length found in the training data, resulting in a motion matrix z with $D \cdot \bar{T} = 1088$ elements. All sequences with fewer frames were handled as missing data as described in section 5.2.2. The number of trainable parameters (ω, θ) in the network summed up to ∼210k in 2D+T and ∼456k in 3D+T respectively. L2 weight decay of $1 \cdot 10^{-4}$ and LeakyReLu activation functions were applied on all layers except the last layer of the TCN and the last layer of the decoder. The former used no activation function for the μ-vector but used the exponential of the

σ-vector to guarantee non-negative values close to 1. The last convolutional layer of the decoder p_θ was followed by a *tanh* activation function for stability reasons during training. The Cauchy-kernel parameters were chosen as proposed in [Fortuin, 2019] with $l = 7$ and $\sigma_K = 1.005$. The variance of the data likelihood was set as the variance of intensity residuals of a few well-registered image sequences with $\sigma_L = 0.0045$ in 2D+T and 0.00021 in 3D+T respectively.

For training, we used a first-order gradient-based method for stochastic optimization (Adam [Kingma, 2014a]) with a batch size of one and fixed learning rate of 0.00015. The TD probability δ was 0.5. Random sub-sequence training was only applied for 3D+T with $\mathcal{T} = 18$. Online data augmentation containing randomly shifted, rotated, scaled and mirrored images has been applied. The model was implemented using Keras [Chollet, 2015] and Tensorflow [Abadi, 2016]. The training time was ∼15h in 2D+T and 7 days for 3D+T sequences on a NVIDIA GTX TITAN X GPU.

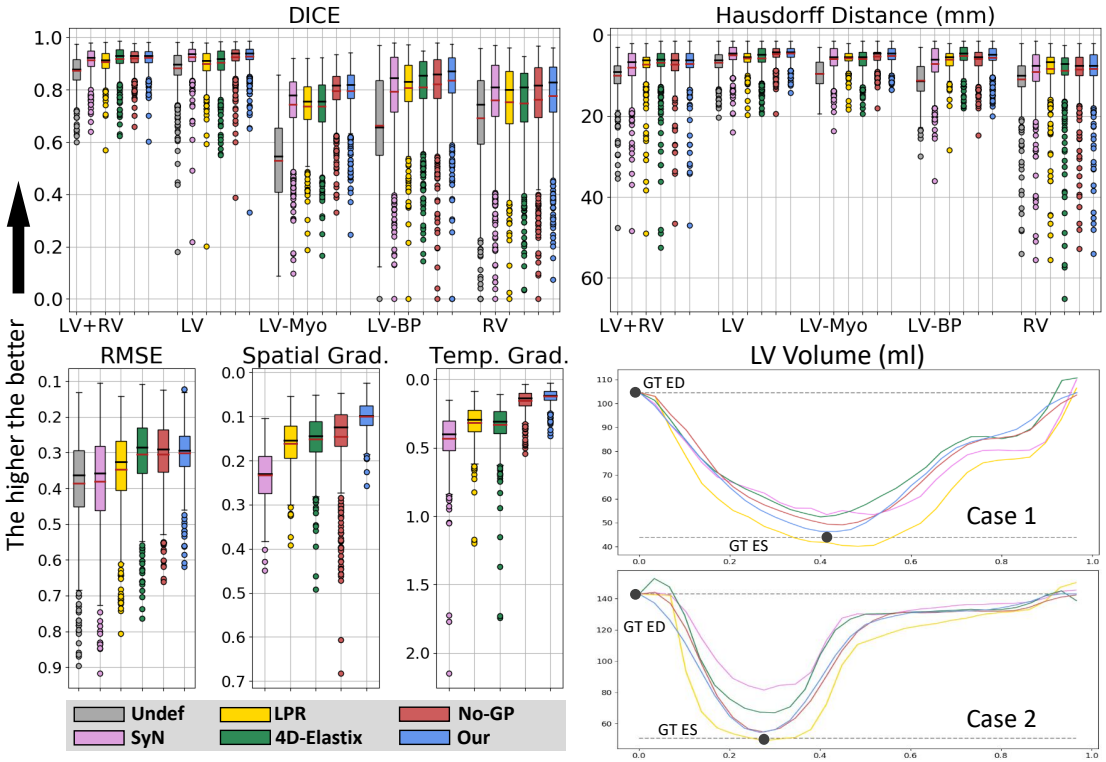


Fig. 5.4: Tracking results showing RMSE, spatial and temporal gradients of the displacement fields, DICE scores and Hausdorff distances for all 2D+T test sequences. The LV volume curves extracted from the warped ED blood pool masks for 2 random test cases in ml, show the temporal smoothness and the distance to the ground-truth ED and ES volumes (marked with black points). The proposed algorithm (Our) shows slightly higher registration accuracy and temporally smoother deformations than the state-of-the-art algorithms: SyN [Avants, 2008], LPR [Krebs, 2019b], 4D-Elastix [Metz, 2011] and the previous version of our method without GP prior (No-GP [Krebs, 2020c]).

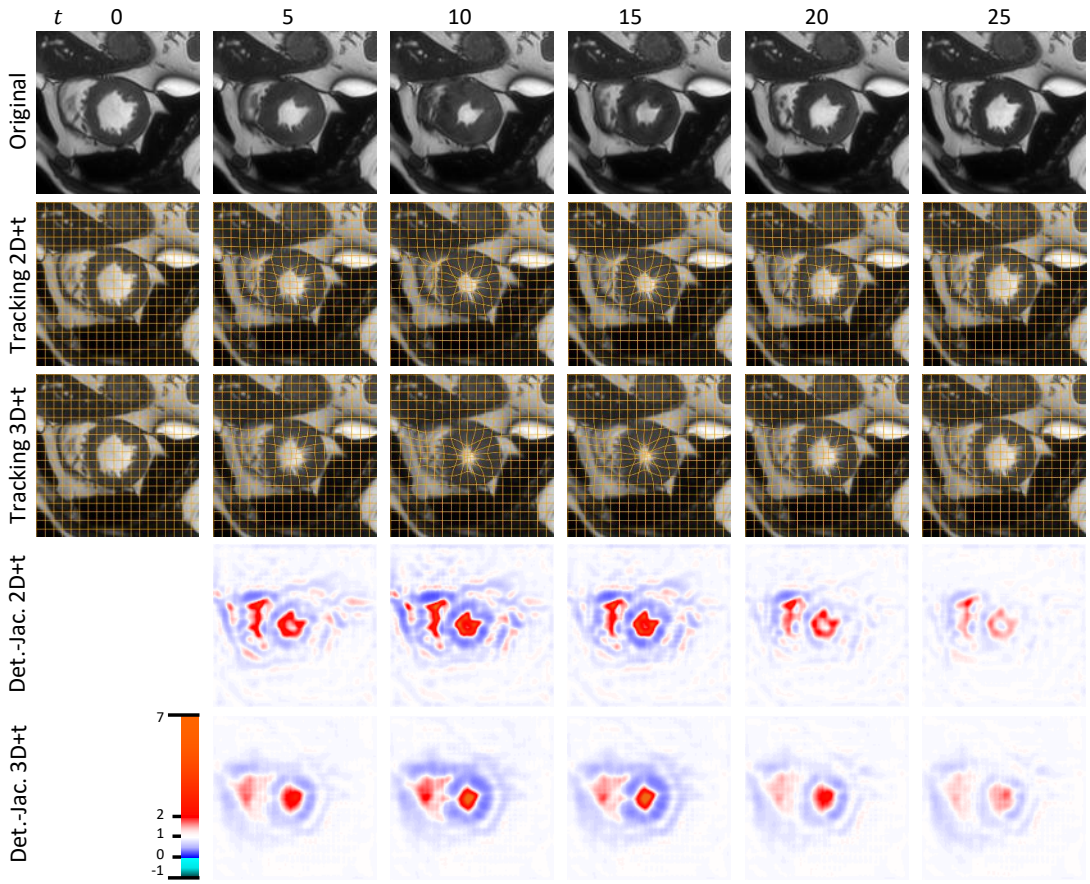


Fig. 5.5: Showing 2D+T and 3D+T tracking results of the warped moving image I_0 with grid overlay and the Jacobian determinant (Det.-Jac.) for a test sequence. In 3D+T, smoother Jacobian determinants were obtained.

5.3.3 Registration and Motion Prediction

We compare our model in terms of registration accuracy and spatio-temporal deformation regularity with 3 state-of-the-art diffeomorphic methods: SyN [Avants, 2008], the learning-based probabilistic pairwise registration (LPR [Krebs, 2019b]) and the temporal B-spline algorithm in elastix (4D-Elastix [Metz, 2011]). We also compare with the previous version of our method with Gaussian Process prior (No-GP [Krebs, 2020c]). SyN and 4D-Elastix have been manually tuned on a few training images following the recommendations in the original papers. The LPR algorithm has been trained on a 2D single scale version using all image pairs of a sequence instead of only the end-diastolic/end-systolic (ED, ES) pairs. We measured registration accuracy using the root mean square error (RMSE) of intensities and segmentation-based DICE scores and 95%-tile Hausdorff distances (HD, in mm) on the five anatomical structures available in ACDC: left ventricle myocardium (LV-Myo), epicardium (LV), left ventricle bloodpool (LV-BP), right ventricle (RV) and LV+RV. In terms of registration regularity, we report spatial (Spatial Grad.) and temporal gradients (Temp. Grad.) of the deformation fields ϕ_t with $t \in [1, T]$.

Tab. 5.1: Registration performance with mean and standard deviation scores of DICE (in %), Hausdorff Distance (HD in mm), spatial and temporal gradients of the deformation fields ($\times 10^{-2}$) comparing the undeformed case (Undef), SyN, learning-based pairwise registration (LPR), 4D-Elastix, our previous version without GP prior (No-GP) and the proposed method for all 2D+T sequences.

Method	DICE	HD	Spat. Grad.	Temp. Grad.
Undef	72.8 ±14	9.70 ±4.20	–	–
SyN	82.7 ±12	7.02 ±4.34	0.23 ±0.06	0.43 ±0.19
LPR	82.1 ±10	6.60 ±3.07	0.16 ±0.06	0.32 ±0.13
4D-Elastix	83.7 ±11	6.27 ±3.91	0.15 ±0.06	0.33 ±0.15
No-GP	84.6 ±10	6.24 ±3.30	0.14 ±0.08	0.15 ±0.08
Our	85.2 ±09	6.11 ±3.28	0.10 ±0.03	0.12 ±0.05

The reported results in Table 5.1 were measured on all 2D test sequences containing at least one mask (resulting in 677 sequences from 100 test subjects). DICE scores and Hausdorff distances are only reported for the frames with available ground-truth segmentation (ES images). Detailed box plots of the results together with LV volume curves are shown in Fig. 5.4. The LV volumes (in ml) were extracted by warping the ED mask according to the extracted deformation fields and computing the blood pool volume for all slices of one subject over time. The results indicate that our model achieves the same (RMSE) or slightly better (DICE and HD) registration accuracy compared to the reference methods while improving spatial and temporal regularity as shown by the deformation field gradients and the volume curves.

Tab. 5.2: 3D+T registration performance with mean and standard deviation scores of RSME, DICE, Hausdorff Distance (HD), spatial and temporal gradients of the deformation fields comparing the undeformed case (Undef), 4D-Elastix and the proposed method.

	RMSE	DICE	HD	Spat. G.	Temp. G.
Undef	0.19	70.1 ±12	7.7 ±2.7	–	–
4D-El.	0.18	79.2 ±10	5.1 ±2.1	0.15 ±0.06	0.62 ±0.32
Our	0.16	79.5 ±09	5.4 ±2.1	0.07 ±0.02	0.09 ±0.03

In Table 5.2, we show the results on the 100 test sequences for our 3D+T model. In comparison to 4D-Elastix, our 3D+T model shows a similar registration accuracy but a significantly improved spatial and temporal regularity. In Fig. 5.5, the warped moving image I_0 and the Jacobian determinant are visualized for one test sequence in 2D+T and 3D+T. One can see, the Jacobian determinants are smoother in 3D+T compared to 2D+T sequences.

The new Gaussian Process prior leads to smoother deformations compared to the previous time-independent prior (No-GP version) while using the same deformation field regularizer. This can be also seen in Fig. 5.6 where the first 5 latent dimensions, the sequences z_d with $d \in [0, 4]$, are visualized for one test case.

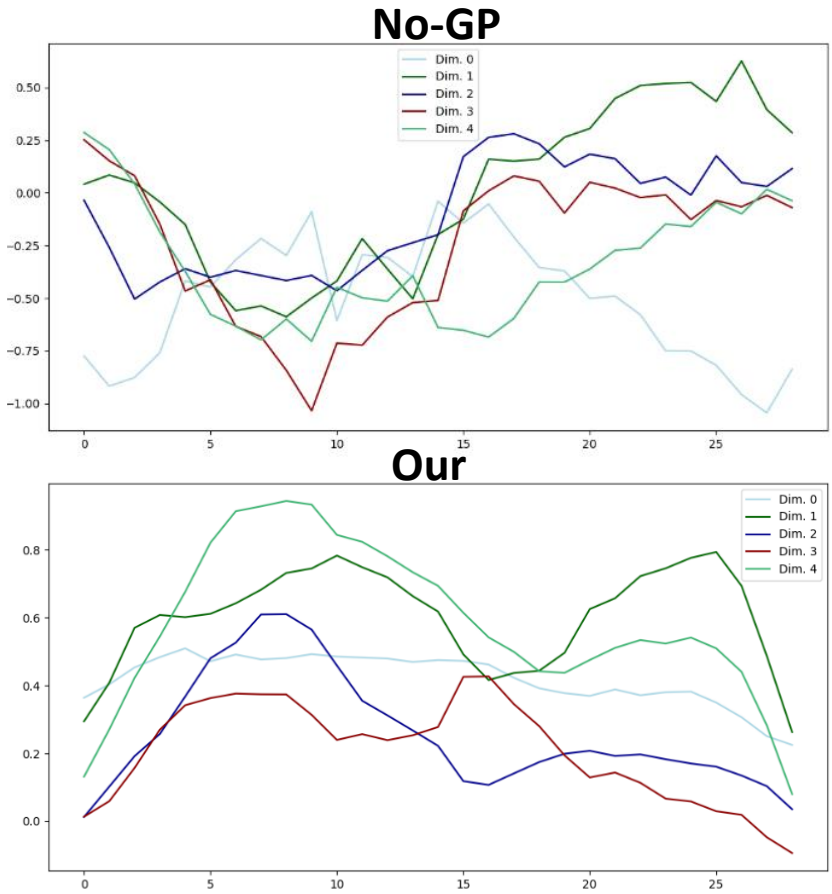


Fig. 5.6: First 5 latent dimensions of the same test sequence shows a temporally smoother motion matrix z for the proposed model trained with the Gaussian process prior compared to the No-GP version.

5.3.4 Motion Simulation, Interpolation and Transport

To evaluate the performance on motion interpolation and simulation, we challenged our model to predict the motion for all time steps from a limited number of input frames. Thus, the goal was to predict motion patterns that are as close as possible to the observed motion of the full sequence (i.e. all registered frames obtained in the all frame model of the previous section 5.3.3). Just as in temporal dropout during training, all the missing frames were represented as zero columns γ_t in the feature matrix Γ as shown in Fig. 5.3b. We compared the motion predictions from various input frame subsets that are provided to the model. First, we provided every 2nd or every 5th frame for motion interpolation. Then, we provided the first 5 frames or only the 10th frame (0th + 10th) to see if the model is able to complete typical cardiac motion patterns. Finally, we tested the full motion simulation by letting the model find a motion sequence given only the moving image I_0 (only 0th) and setting feature matrix Γ to zero everywhere. We compared the simulated motion, with linear and cubic interpolation of the deformation fields (which are taken from the all frame model at the selected time steps). In the top of Fig. 5.7, average LV volume errors (RMSE) with respect to the all frame model were computed

for all 677 test sequences in comparison to linear and cubic interpolation. In the bottom of Fig. 5.7, one can see the results of our model for the different interpolation cases in terms of LV volume curves for two example sequences.

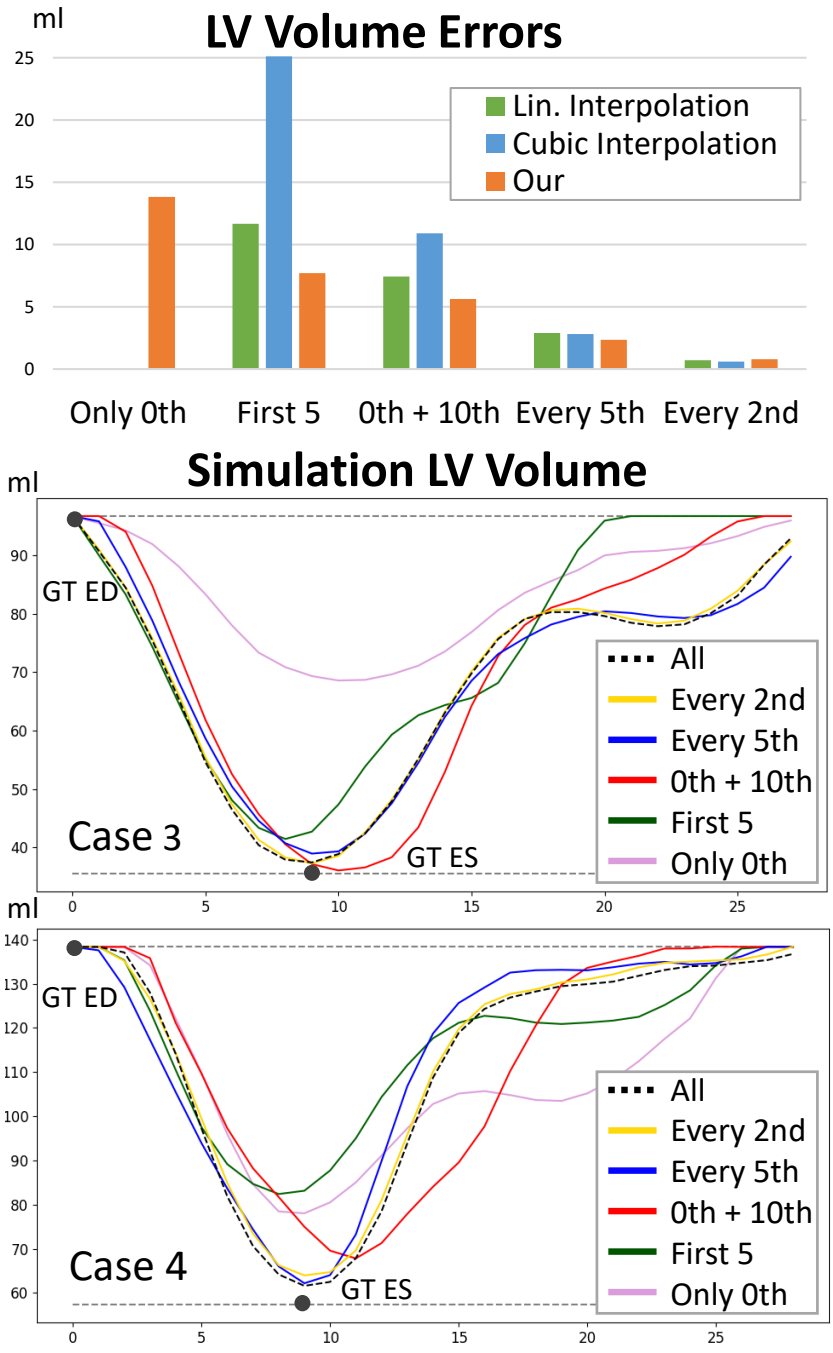


Fig. 5.7: Predicted simulated and interpolated motion from a limited number of frames. Provided frames are decreasing from all frames to only the 0th frame (full motion simulation). The volume errors with respect to the all frame prediction are compared with linear and cubic interpolation of the deformation fields. Two random test subjects are shown in the bottom.

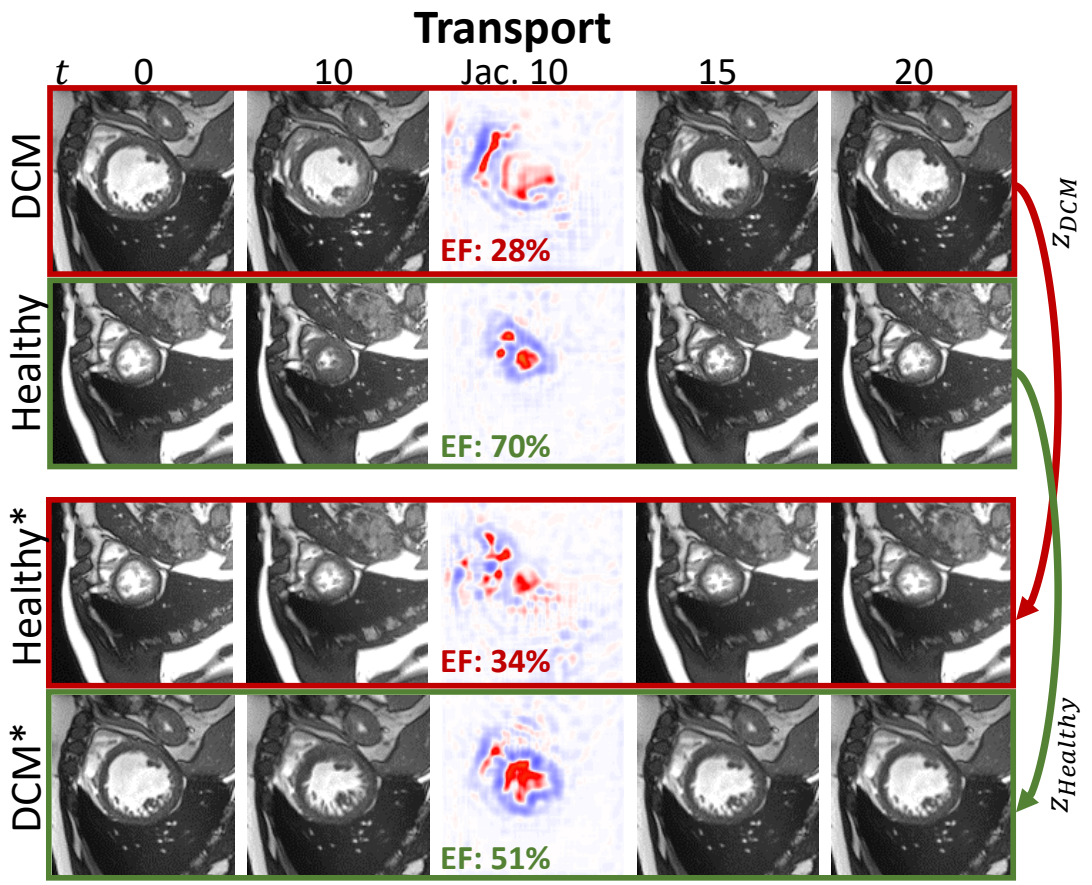


Fig. 5.8: Transporting the motion matrix z from one subject and combining it with the end-diastolic frame of another subject allows for simulating a disease (dilated myopathy, DCM, red motion) in a healthy subject and vice versa (green motion). Ejection fraction (EF) of the simulated cases are more similar to the transported motion.

For the cases of providing every 2nd and every 5th frame, our model interpolated the motion similarly well as linear or cubic interpolation, while providing better results in the cases of providing the 0th+10th and first 5 frames signaling an improved learned cardiac motion model. The full simulation (only 0th) did not result in well fitted volume curves, which is expected as the model has to simulate the full motion sequence from just the ED frame. However, it is observable that the model learned realistic cardiac specific motion patterns as the volume curves for example show the plateau phase before atrial systole which can be also seen in the completed motion for the cases where we provide the first 5 and 0th+10th frames. For the full simulation, our model often slightly under-estimated the motion (cf. case 3 in Fig. 5.7) which can be related to the pathology distribution in the training dataset which contained many cases with reduced cardiac motion.

Furthermore, we demonstrate the model's capacity of motion transport in a qualitative way. Our model allows to transport motion patterns from one subject to another by taking the motion matrix z of one case and applying it on the moving image of another image sequence (ED frame). In this way, for example a pathological motion can be simulated in

a healthy subject or vice versa. In Fig. 5.8, we present 2 subjects from the ACDC dataset, from which one is classified as healthy and the other as a dilated myopathy case (DCM). We extracted the motion matrices for both and applied them on the ED frame of the other case, such that we simulated a DCM typical motion in the healthy case while *curing* the pathological case. This can be seen for example from the LV contraction strengths in the Jacobian determinants or the related ejection fraction (EF). Note, that this form of parallel transport does not require any additional inter-subject registration.

5.4 Discussion and Conclusion

We presented a probabilistic motion model that can be useful for example for spatio-temporal registration, temporal super-resolution, data augmentation, shorter acquisition times and motion analysis. Based on a novel Gaussian Process prior conditional variational autoencoder, the model is learned in an unsupervised fashion from medical image sequences. Intrinsic motion patterns are encoded in a low-dimensional probabilistic space – the motion matrix – which allows for accurate diffeomorphic tracking, temporal interpolation, motion simulation and motion transport.

Our approach has shown state-of-the art registration accuracy and improved deformation regularity temporally and spatially in comparison to 3 state-of-the-art algorithms indicating that the low-dimensional motion encoding helps to regularize the registration problem of image sequences. We have shown that the novel Gaussian Process prior leads to a higher temporal consistency compared to the time-independent prior [Krebs, 2020c] both, in latent and deformation space. A temporally smoother latent space is desirable as it brings more structure and interpretability and is consistent with the temporally smooth motion we experience in deformation space. We have demonstrated motion simulation and interpolation from a very limited number of frames indicating that data acquisition could be speed up as fewer frames are required in order to retrieve an accurate motion. In case of full simulations, our model showed a slightly reduced cardiac motion compared to healthy subjects. The authors believe this is due to a bias introduced from the disease distribution in the training data. To not end up with such a mean motion that merges several pathological motion patterns, one could think of generating disease-specific models. This could be achieved by training different motion models with training sets separated by diseases. As another extension to our previous work, we have shown first results on 3D+T sequences which showed smoother Jacobian determinants than the 2D+T version which can be explained by out-of-plane deformations. However, a limitation is the high computational costs for 3D+T sequences with long training times even for relatively low-dimensional images.

In future work, we aim to reduce this complexity and work on the generalization of the approach to other applications such as respiratory motion estimation. Furthermore, the authors believe the motion matrix as a compact representation of organ motion can be helpful as a quantitative new tool to guide the diagnosis, prognosis or therapy of diseases of dynamic organs.

Acknowledgments

Data used in this article were obtained from the EU FP7-funded project MD-Paedigree and the ACDC STACOM challenge 2017 [Bernard, 2018]. This work has been supported by the French government, through the 3IA Côte d'Azur Investments in the Future project managed by the National Research Agency (ANR) with the reference number ANR-19-P3IA-0002 and the grant AAP Santé 06 2017-260 DGA-DSH, and by the Inria Sophia Antipolis - Méditerranée, "NEF" computation cluster.

5.5 Appendix

5.5.1 KL Divergence using the GP Prior

Given 2 multivariate Gaussian distributions with the same dimensionality, the KL divergence is defined in [Duchi, 2007]. Suppose, we take our prior distribution $p(z)$ with zero-mean $\mathbf{0}$ and covariance Σ of the form of Eq. 5.2 and our posterior distribution q_ω with mean μ and covariance Σ^* with dimensionality $D\bar{T}$:

$$\mathrm{KL}[q_\omega(z|I_{0:T})\|p(z)] = \frac{1}{2}\left(\mathrm{tr}(\Sigma^{-1}\Sigma^*) + \mu^\top\Sigma^{-1}\mu - D\bar{T} + \ln\left(\frac{\det\Sigma}{\det\Sigma^*}\right)\right). \quad (5.8)$$

The determinants of the block diagonal matrices Σ, Σ^* are $\det\Sigma = |K|^D$ and $\det\Sigma^* = |K|^D\prod_{i=1}^D\sigma_i^2$. Thus, the logarithm of the fraction of determinants in Eq. 5.8 becomes:

$$\ln\left(\frac{\det\Sigma}{\det\Sigma^*}\right) = \ln\left(\frac{1}{\prod_{i=1}^D\sigma_i^2}\right) = -\sum_{i=1}^D\ln\sigma_i^2 \quad (5.9)$$

When taking the sum over the D latent dimensions over the remaining terms, Eq. 5.8 simplifies to:

$$\mathrm{KL}[q_\omega(z|I_{0:T})\|p(z)] = \frac{1}{2}\sum_{i=1}^D\sigma_i^2\bar{T} + \bar{\mu}_i^\top K^{-1}\bar{\mu}_i - \bar{T} - \ln(\sigma_i^2) \quad (5.10)$$

with $\bar{\mu}_i$ being the i-th segment of length T in μ. In the case of prior and posterior being identical, thus $\mu = \mathbf{0}$ and $\sigma = \mathbf{1}$ the quantity in Eq. 5.10 becomes 0.

5.5.2 Cholesky Decomposition of Σ^*

The Cholesky decomposition of a symmetric positive-definite matrix X equals the matrix product of a lower-diagonal L and its transposed: $X = LL^\top$. The entries of L can be computed by the Cholesky-Banachiewicz algorithm:

$$L_{j,j} = \sqrt{X_{j,j} - \sum_{k=1}^{j-1} L_{j,k}^2}$$

$$L_{i,j} = \frac{1}{L_{j,j}} \left(X_{i,j} - \sum_{k=1}^{j-1} L_{i,k} Lj, k \right) \qquad \text{for } i > j. \quad (5.11)$$

In case of the block diagonal matrix Σ^* the lower triangular matrix L^* equals a block diagonal matrix with lower triangular matrices that are resulting from the Cholesky decompositions of the diagonal block elements of Σ^*. Thus, in order to compute L^*, the Cholesky decompositions of the $i \in D$ diagonal elements $\sigma_i K$ must be computed. From Eq. 5.11 it follows that $c \cdot X = (\sqrt{c} \cdot L)(\sqrt{c} \cdot L^\top)$. Thus, $\sigma_i K = (\sqrt{\sigma_i} \cdot L_K)(\sqrt{\sigma_i} \cdot L_K^\top)$ and L^* is:

$$L^* = \text{Diag}_{d=1}^{D}(\sqrt{\sigma_d} \cdot L_K). \quad (5.12)$$

Since the kernel matrix K is fixed in our framework, L_K can be pre-computed using Eq. 5.11 and reused keeping the computational efforts minimal even for a large covariance matrix Σ^*.

6 Risk Prediction for Heart Failure Outcomes from Cardiac Motion Features

Contents

This chapter is intended to show one specific clinical example on how the motion model developed in chapter 5 could be used to support prognosis and therapy planning. Using the motion model, the survival risks of heart failure patients can be predicted by obtaining a risk score from the latent motion matrix. Based on this estimated risk, an appropriate therapy can be chosen, for example, whether or not to implant a defibrillator. This demonstrates the discriminative power of the motion model trained on a cohort of heart failure patients. The chapter presents only preliminary results. A clinical journal submission is in preparation.

6.1 Introduction

Sudden cardiac death (SCD) in heart failure (HF) patients is one of the leading causes of natural death. SCD occurs when the electrical system of the heart is malfunctioning causing irregular heartbeats (arrythmias). Emergency treatment includes electric shocks (defibrillation) to restore the normal heart rhythm. For patients with a high risk of SCD, an implantable cardioverter-defibrillator (ICD) can be inserted as a preventive treatment.

An ICD monitors the heart activity and can apply electric shocks in case of extreme arrythmias.

Selecting patients for ICD treatment is a challenging task. It is crucial to predict the risk for SCD to justify potential complications that come along with an ICD treatment such as surgery risks, false shocks and a shorter life expectancy. Accurate SCD risk prediction helps to select only patients for ICD who benefit from it.

Currently, the main quantitative measure used to predict risk for SCD is left ventricular ejection fraction (LVEF), an imaging feature of cardiac structure and function [Myerburg, 2009]. However, among patients receiving a primary prevention ICD based on an LVEF $\leq$35% [Tracy, 2013], the rate of appropriate therapies is very low with 2.6% at 30 months of follow-up [Sabbag, 2015]. In other words, many patients that receive ICD treatment do not require it. In addition, LVEF improvement occurs in up to 25-50% of patients and correlates with diminished SCD risk [Punnoose, 2011]. Thus, LVEF is far from being a comprehensive feature to predict SCD. Recently, other imaging features of cardiac structure and function have been found to be independent predictors of SCD. Such factors are right ventricular (RV) and left atrial (LA) [Rijnierse, 2017] function or the extent of heterogeneous myocardial tissue (gray zone) on late gadolinium enhancement (LGE) cardiac magnetic resonance images [Jablonowski, 2017]. This motivates the assumption that more unidentified SCD predictors are inherently present in cardiac images.

Deep learning is capable of addressing the high-dimensional vector space and extracting unrecognized features from medical images. Lou et al. [Lou, 2019] proposed to extract features from images to predict treatment outcomes in lung cancer patients by incorporating hand-crafted radiomics features in the training. Taking low-dimensional segmentations of the right ventricle as input, Bello et al. [Bello, 2019] predicted the survival risk for patients with pulmonary hypertension. While these approaches rely on hand-crafted features extracted from images, we have shown in our previous work (Chapter 5) that a motion fingerprint containing inherent features of the LV motion can be generated from cine-MRI images using a latent variable model. This population-specific fingerprint can be learned in an unsupervised fashion by training a probabilistic motion model using a conditional variational autoencoder (CVAE) [Krebs, 2020c].

We propose a novel learning-based method for personalized survival risk prediction for SCD that utilizes automatically derived image features from 4 chamber view cine-MRI. Our model generates fingerprints of inherent imaging features of the cardiac motion which are used to predict risk scores for outcomes of HF patients such as hospitalization or SCD. In clinical practice, these risk scores can be used to select high-risk patients for ICD treatment while postponing ICD treatment for low-risk patients. In particular, the novel risk predictor uses an automatically extracted personalized cardiac motion fingerprint in combination with a risk prediction neural network. The risk prediction network is based

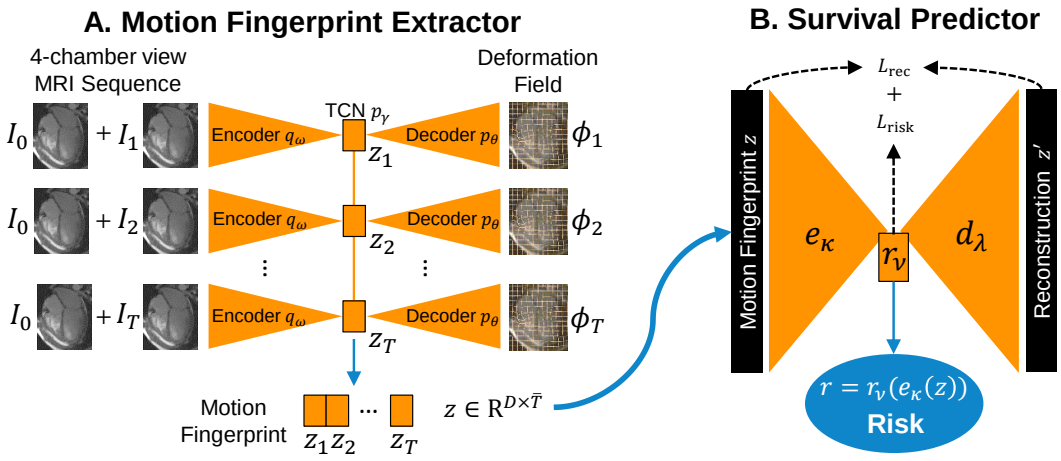


Fig. 6.1: The outcome risk prediction model consisting of learning a motion fingerprint from 4-chamber view cine-MRI (A.) and a survival predictor neural network (B.) which estimates the outcome risk based on the motion fingerprint. The dashed black arrows symbolize training loss computations while the blue arrows symbolize the data flow during testing.

on a non-linear Cox proportional hazard loss to make use of right-censored survival outcome data.

On a non-ischemic cohort of HF patients with clinical criteria for primary prevention ICD, the derived motion risk factor showed the highest statistical significance as an independent predictor for hospitalization among other relevant clinical factors that are associated with HF endpoints.

The main contributions are:

- A novel risk prediction framework for HF patients based on a cardiac motion fingerprint extracted from image sequences in an unsupervised fashion.
- State-of-the-art predictive accuracy for HF hospitalization on a non-ischemic patient cohort.

6.2 Methods

The risk prediction model is composed of two elements: A. a motion fingerprint extractor from image sequences and B. a survival predictor that estimates the risk for a given endpoint (or outcome) from the motion fingerprint. In this work, we apply two independent neural networks for these tasks. First, a probabilistic encoder-decoder neural network [Krebs, 2020c] is trained to learn motion characteristics from image sequences and extract a cardiac motion fingerprint in a fully unsupervised fashion. Second, an autoen-

coder neural network is trained from the motion fingerprint by regressing HF outcomes. To enable the use of censored data, a loss function inspired from the Cox proportional hazards model [Cox, 1972] is utilized. The two steps, A. and B. are schematically shown in Fig. 6.1 and are explained in detail in the following.

6.2.1 Motion Fingerprint Extractor

The motion model used in this chapter has the same inputs and outputs as the one presented in chapter 5. However, it includes some methodological differences as it applies a multivariate unit Gaussian prior, time-independent sampling and explicit time dependence. Detailed derivations of the fingerprint extractor can be found in our conference paper [Krebs, 2020c] and in the appendix 6.5.2.

The motion fingerprint is learned in an encoder-decoder neural network which represents a latent variable model. The input of the network are a sequence of image pairs (I_0, I_t) with $t \in [1, T]$ from image sequences of length T. The output are a sequence of dense deformation fields ϕ_t (between (I_0, I_t)) and a compact deformation representation $z_t \in \mathbb{R}^D$ of dimensionality D per timestep t. The sequence of encoded z_t are combined in the motion matrix $z \in \mathbb{R}^{D \times \bar{T}}$ with $\bar{T} = T - 1$ and D latent dimensions depicting the cardiac motion. In this work, we consider 2-dimensional image sequences of four chamber cine-MRI with a single slice. The model is trained using a conditional variational autoencoder (CVAE, [Kingma, 2014b]). Instead of the left-ventricular motion as in [Krebs, 2020c], we learn a motion fingerprint of the full heart. Furthermore, in contrast to Chapter 5, a temporally independent unit Gaussian prior has been applied.

First, the encoder q_ω with network weights ω maps each of the image pairs (I_0, I_t) independently to a latent space denoted by $\tilde{z}_t \in \mathbb{R}^D$. To this end, the encoder approximates the posterior distribution $q_\omega(\tilde{z}|I_{0:T})$ of the latent variable model. Second, as the key component of temporal modeling, these latent vectors $\tilde{z}_t$ are jointly mapped to the motion matrix or motion fingerprint z by conditioning them on all past and future time steps and on the normalized time $\bar{t}$: $p_\gamma(z|\tilde{z}_{1:T}, \bar{t}_{1:T})$. This regularizing network p_γ with weights γ is realized using a temporal convolutional network (TCN [Bai, 2018]). Finally, the decoder p_θ with trainable network weights θ aims to reconstruct the fixed image I_t by warping the moving image I_0 with the deformation ϕ_t. This deformation ϕ_t is extracted from the temporally regularized z_t-codes. The decoder is further conditioned on the moving image by concatenating the features at each scale with down-sampled versions of I_0. It approximates the data likelihood $p_\theta(I_{1:T}|z, I_0)$.

During training, a lower bound on the data likelihood is maximized with respect to a prior distribution $p(\tilde{z}_t)$ of the latent space $\tilde{z}_t$ (cf. CVAE [Kingma, 2014b]). The prior $p(\tilde{z}_t)$ is assumed to follow a multivariate unit Gaussian distribution with spherical covariance I:

$p(\tilde{z}_t) \sim \mathbb{N}(0, I)$. The loss function of the motion fingerprint extractor results in optimizing the expected log-likelihood p_θ and the Kullback-Leibler (KL) divergence enforcing the posterior distribution q_ω to be close to the prior $p(\tilde{z}_t)$ for all time steps:

$$\mathcal{L}_{\text{Motion}}(\omega, \gamma, \theta) = \sum_{t=1}^{T} -\mathbb{E}_{z_t \sim p_\gamma(\cdot|\tilde{z}_{1:T}, \bar{t}_{1:T})} \Big[\log\, p_\theta(I_t|z_t, I_0)\Big] + \text{KL}\left[q_\omega(\tilde{z}_t|I_0, I_t) \| p(\tilde{z})\right]. \quad (6.1)$$

Unlike the traditional CVAE model, the temporal regularized z_t-code is used in the log-likelihood term p_θ instead of the $\tilde{z}_t$. We model p_θ as a symmetric local cross-correlation Boltzmann distribution with the weighting factor ι. All network weights except the ones in the TCN are shared and thus independent of the time t. Their network architecture consists of convolutional and deconvolutional layers with fully-connected layers for mean and variance predictions in the encoder part [Kingma, 2014b]. We use an exponentiation layer for a stationary velocity field parameterization of diffeomorphisms [Krebs, 2019b], a linear warping layer and diffusion-like regularization with smoothing parameters σ_G in spatial and σ_T in temporal dimension. During training, we apply temporal dropout sampling as described in [Krebs, 2020c] in order to further ensure learning temporal dependencies and increase generalizability.

6.2.2 Survival Predictor

The survival predictor takes the motion fingerprint z, the compact representation of the motion, as input and predicts the survival risk score r which is defined by the logarithm of the hazard ratio in the Cox regression analysis [Cox, 1972]. This ratio contains the hazard $h_z(t)$ of a subject with fingerprint z with respect to the baseline hazard $h_0(t)$:

$$r = \log \frac{h_z(t)}{h_0(t)}, \quad (6.2)$$

where the subject hazard $h_z(t)$ symbolizes the probability of the subject of dying at time t and the baseline hazard describes the survival without an influence of covariates z. The hazard ratio is assumed to be constant over time behind the semi-parametric proportional hazard model of Cox [Cox, 1972]. Thus, the continuous risk score r allows to classify the outcome risk for a new patient at test time.

In contrast to standard Cox regression analysis, we define the risk r as a non-linear combination of input features z: $r = r_\nu(e_\kappa(z))$ where r_ν and e_κ are two neural networks with network weights ν and κ. The full risk model is realized as autoencoder neural networks that reduce the fingerprint's dimensionality $D\bar{T}$ in order to retrieve the risk r. The authors chose an encoder-decoder architecture in contrast to a direct prediction of r

in order to constrain and regularize the risk predictor to avoid over-fitting [Bello, 2019; Lou, 2019].

The encoding and decoding branches of the risk autoencoder are denoted by e_κ and d_λ with network weights κ and λ respectively. A third network with weights ν is applied to obtain the risk score $r_\nu(e_\kappa(z))$ from the latent space of the autoencoder $e_\kappa(z)$. In this work, the three networks consist of fully-connected layers due to the low dimensional fingerprints. In case of larger fingerprints, convolutional and deconvolutional layers in encoder respectively decoder networks could be used. The risk predictor is trained using multi-task learning by aiming to reconstruct the motion fingerprint and to predict the risk r at the same time. Thus, the loss function $\mathcal{L}_{\text{Risk}}(\kappa, \lambda, \nu)$ contains 2 terms, one for the fingerprint reconstruction $\mathcal{L}_{\text{rec}}(\kappa, \lambda)$ and one for risk prediction $\mathcal{L}_{\text{risk}}(\kappa, \nu)$:

$$\mathcal{L}_{\text{Risk}}(\kappa, \lambda, \nu) = \mathcal{L}_{\text{rec}}(\kappa, \lambda) + \alpha \mathcal{L}_{\text{risk}}(\kappa, \nu) \tag{6.3}$$

where α denotes a weighting factor between both terms. For risk prediction, we apply the negative log partial likelihood as survival function over N censored training samples following standard Cox regression analysis [Cox, 1972]:

$$\mathcal{L}_{\text{risk}}(\kappa, \nu) = -\sum_{i=1}^{N} \delta_i \left[r_\nu(e_\kappa(z^i)) - \log \sum_{j=1}^{N} R_{ij} \exp(r_\nu(e_\kappa(z^j))) \right], \tag{6.4}$$

with z^i being the fingerprint of the i-th training subject. The Boolean censoring indicator δ_i equals 1 if the subject experienced SCD (or another endpoint of interest) at the given time τ. A subject is censored $\delta_i = 0$ if the patient was still alive at time τ but removed from the study afterwards. R is the risk matrix where $R_{ij} = 1$ if $\tau_j \geq \tau_i$ and $R_{ij} = 0$ if $\tau_j < \tau_i$, based on N training samples per batch. This represents a non-linear Cox proportional hazard model (cf. to [Lou, 2019]). The fingerprint reconstruction loss term is defined as the mean squared error between fingerprint z and reconstructed fingerprint $z' = d_\lambda(e_\kappa(z))$:

$$\mathcal{L}_{\text{rec}}(\kappa, \lambda) = \frac{1}{N} \sum_{i=1}^{N} \|z^i - d_\lambda(e_\kappa(z^i)\|^2. \tag{6.5}$$

The two modules of fingerprint extractor and survival predictor are trained in 2 steps. First, the motion fingerprint is trained alone and afterwards the survival predictor while fixing the motion fingerprint network. This keeps the motion fingerprint independent of the survival analysis and allows for example the training on additional data for the motion extraction where no survival data is available.

6.3 Experiments

In the experiments, we used a non-ischemic cohort of 167 HF patients with clinical criteria (low LVEF) for primary prevention ICD. These cases were collected from 3 different sites. We used 4 chamber-view cine-MRI which were taken prior to ICD implantation. For the preliminary experiments in this study, we used HF hospitalization as endpoint to evaluate the proposed outcome predictor. HF hospitalizations was defined as the time point when a patient came into hospital with a documented primary diagnosis related to heart failure. In future and once we get clearance for the data, we plan to add results for SCD risk prediction and other HF endpoints. Using HF hospitalization as endpoint, this cohort consisted of 36% of subjects (60 subjects) with event times and right censored data for the remaining ones. Besides the censored data of the endpoint, the following clinical features were used as comparison risk predictors in this study computed with standard clinical tools: graymass (GM), minimum and maximum LA index volume (VminI respectively VmaxI), LA strain rate during LV systole (SRmax), preatrial contraction (SpreA) and atrial contraction (SRA). These features were selected as they are associated for being predictive for HF hospitalization and SCD [Issa, 2017; Rijnierse, 2017; Jablonowski, 2017]. From this cohort, 60 subjects (36%) experienced HF hospitalization.

6.3.1 Implementation Details

The 4 chamber-view cine MRI were resampled to an image size of 128 by 128 pixels with a spacing of 2.2 mm. The implementation of the fingerprint extractor followed the details in [Krebs, 2020c]. We increased the latent dimensionality D to 64 motivated by the fact that 4 chamber view images contain more complex motion details than the LV motion alone. The survival predictor requires motion fingerprints z to have the same size for all patients. In order to retrieve same sized z, we interpolated the cine-MRI in temporal dimension to retrieve a fixed time length T. In this work, we used $T = 25$ as it represents the average sequence length in this cohort. We applied B-spline interpolation for resampling the image sequences that contained less or more than 25 frames. The Gaussian deformation field regularization was applied with $\sigma_G = 3$mm and $\sigma_T = 1.5$. The weighting factor between reconstruction and KL loss terms has been chose empirically as $\iota = 6 \cdot 10^{-4}$.

In total, the neural network of the risk predictor contained 5 fully-connected layers. The encoder e_κ consisted of two consecutive layers with 180 and respectively 10 units whose output created the latent space. On the one side, the decoder d_λ used the latent code and applied two layers to retrieve the reconstructed fingerprint vector z'. On the other side, a single dense layer was used to extract the scalar risk score r from the latent code. The

output layer of the risk network r_ν had a *tanh* activation function while the decoder's output did not apply an activation function. The remaining layers used *relu* activation functions. Furthermore, a dropout factor of 0.3 has been applied on the input layer. Dropout factor and number of units of the fully-connected layers were determined by a hyperparameter search using evolutionary optimization. The model has been trained using the Adam optimizer [Kingma, 2014a] with a learning rate of 0.0001 and batch size of 16. The framework has been implemented using Keras [Chollet, 2015] and Tensorflow [Abadi, 2016].

Tab. 6.1: Predictors of HF hospitalization using univariate and multivariate (for Clinical and Fingerprint+Clinical) Cox proportional hazard models. The results are obtained via 6-fold stratified cross-validation. HR, p-value (reject the null hypothesis that the HR equals one) and average concordance index (C) are reported including a 95% confidence interval (CI) in brackets. The motion fingerprint shows the highest prediction accuracy, independently and together with multiple clinical variables.

Feature	HR	C	p-value
GM	0.92 (0.64-1.32)	0.52 (0.38-0.55)	0.66
VmaxI	1.85 (1.31-2.60)	0.63 (0.55-0.69)	<0.005
VminI	2.03 (1.44-2.87)	0.66 (0.59-0.72)	<0.005
SRmax	2.30 (1.63-3.26)	0.65 (0.59-0.71)	<0.005
SRA	2.02 (1.43-2.85)	0.62 (0.55-0.68)	<0.005
SpreA	1.98 (1.41-2.79)	0.63 (0.56-0.69)	<0.005
Fingerprint	**2.93 (2.05-4.18)**	**0.69 (0.60-0.72)**	<0.005
Best Clinical Params.	2.39 (1.69-3.39)	0.67 (0.61-0.74)	<0.005
Fingerp. + Best Clinical Params.	**3.02 (2.11-4.32)**	**0.70 (0.63-0.75)**	<0.005

6.3.2 Results

We evaluated our risk prediction model in comparison to the other clinical factors by fitting linear univariate and multivariate proportional hazard Cox models [Cox, 1972]. We used 6 fold stratified cross-validation, first for training the fingerprint extractor and second for the Cox models. The fingerprint extractor has been trained first, in a risk independent fashion. The extracted motion for 2 example cases, 1 with HF hospitalization event and one without, can be seen in the appendix 6.5.1. In Fig. 6.3 of the appendix, we further compared the motion model with the 4D Elastix algorithm [Metz, 2011] in terms of matching of intensities and deformation regularity.

For risk prediction, we report the mean concordance index (C) [Harrell, 1982] over the 6 folds and compute hazard ratios (HR) including confidence intervals (CI) and statistical p-value by splitting all test results by their medium risk value, dividing the cohort in a low and high risk group. For the Cox analysis and HR computation, the python package *lifelines* [DavidsonPilon, 2020] has been used.

In Table 6.1, the results for the Cox analysis are shown using the different clinical features and the fingerprint risk score independently. The last two rows in table 6.1 show multivariate Cox analysis results for the joint predictive power of the best-performing combination of clinical features and the combination of these clinical features and the fingerprint risk. We tested all combinations of the 6 clinical features and show only the best combination here denoted by Best Clinical Params. In terms of testing results on C and HR scores, this best combination was found to contain VminI, VmaxI and SpreA features. In case of multivariate models, the linear Cox model showed signs of over-fitting by resulting in much better training but worse testing scores.

With an HR of 2.93 (CI 2.05-4.18) and an C-index of 0.69 (CI 0.60-0.72), the novel fingerprint risk score extracted from the motion model shows the highest prediction accuracy as independent predictor of HF hospitalization. In combination with the best clinical features, the multivariate Cox analysis showed an improved cross-validated C-index of 0.70 (CI 0.63-0.75) and HR of 3.02 (CI 2.11-4.32).

In Fig. 6.2, we further show the Kaplan-Meier plots of 4 independent features and the 2 feature combinations. Kaplan-Meier estimates can be used to measure the fraction of subjects living for a certain amount of time [Goel, 2010]. One can see the capability of different models to recognize low and high risk patients by analyzing the distance between high and low risk survival curves in the Kaplan-Meier plot. We split our cohort into low and high risk groups according to the median risk prognosticated by the Cox model (for test cases). It is shown that for example the gray mass (GM) is not a good predictor since low and high risk survival curves are highly overlapping. The best differentiation between both groups (characterized by a large gap between the survival lines) can be seen for the fingerprint risk score and the combination of fingerprint and clinical features. These results indicate that the proposed risk predictor based on an extracted motion fingerprint can more accurately predict HR hospitalization than other commonly used clinical features.

6.4 Discussion and Conclusions

In this work, we have proposed a novel image-driven risk predictor for personalized survival analysis by using a learned motion fingerprint – a low-dimensional encoding of the motion from a sequence of images. The proposed method showed promising first results in terms of predicting the risks for hospitalization of HF patients. These findings could be the first step to lead to a better patient selection for ICD treatment.

Besides HF hospitalization, we plan to add other endpoints such as SCD to this study. Furthermore, the authors think, the performance could be further improved by adding

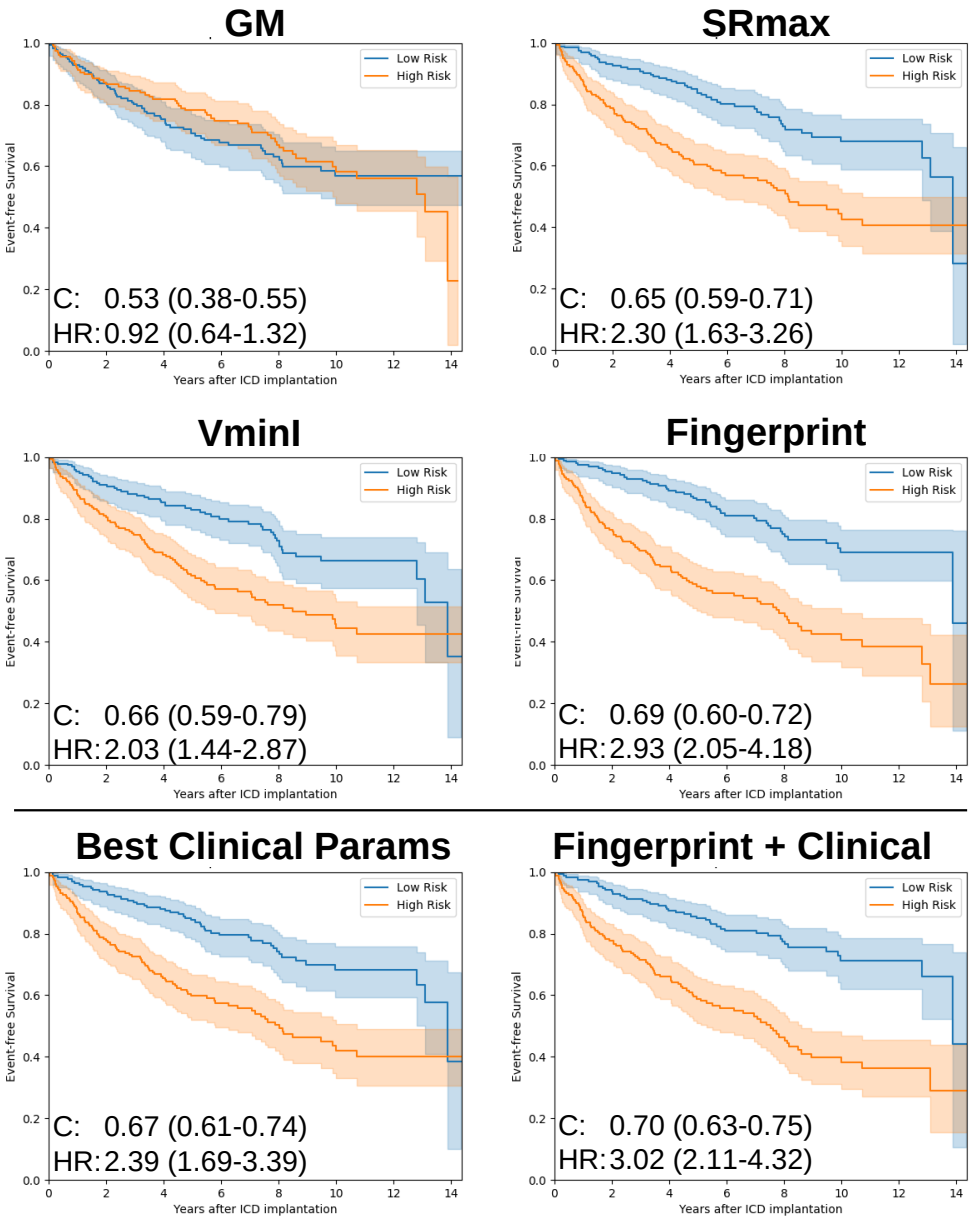


Fig. 6.2: Kaplan-Meier plots showing the average survival risk and its confidence interval for low and high risk patients depending on different predictors: gray mass (GM), SRmax, VminI, motion fingerprint and multivariate risks using clinical and the combination fingerprint and clinical features. The motion fingerprint helps to differentiate between low and high risk patients.

more features to the risk prediction network. One possible way could be by complementing the motion fingerprint with a cardiac structure fingerprint, extracted in a similar fashion as the motion fingerprint from for example late gadolinum enhancement images. The authors think that if multivariate models are used (combination of multiple clinical and motion features) the experienced over-fitting might be resolved by using a bigger dataset for fitting the linear Cox proportional hazard model or by using sparse estimation

of Cox proportional hazards models that select the most relevant features in the survival prediction such as in [Evers, 2008; Su, 2016].

The two modules of fingerprint extractor and survival predictor can be also trained in an end-to-end fashion where all loss terms in Eq. 6.1 and Eq. 6.3 are combined in a single weighted loss function as for multitask training. In this way, the motion fingerprint is fine-tuned for personalized outcome risk prediction. However, the weighting between the different loss terms is more difficult and additional data for training the fingerprint extractor is not easily usable.

In future work, the model's lack of interpretability could be explored. As the model already contains two clearly separated modules of motion fingerprint and risk predictor it would be interesting to see which features are especially used and relevant for predicting the HF risks. Another possible future direction is to investigate the neural network features in depth from a clinical research perspective to potentially find unknown motion features that can be associated with HF or SCD.

6.5 Appendix

6.5.1 Motion Fingerprint Extraction

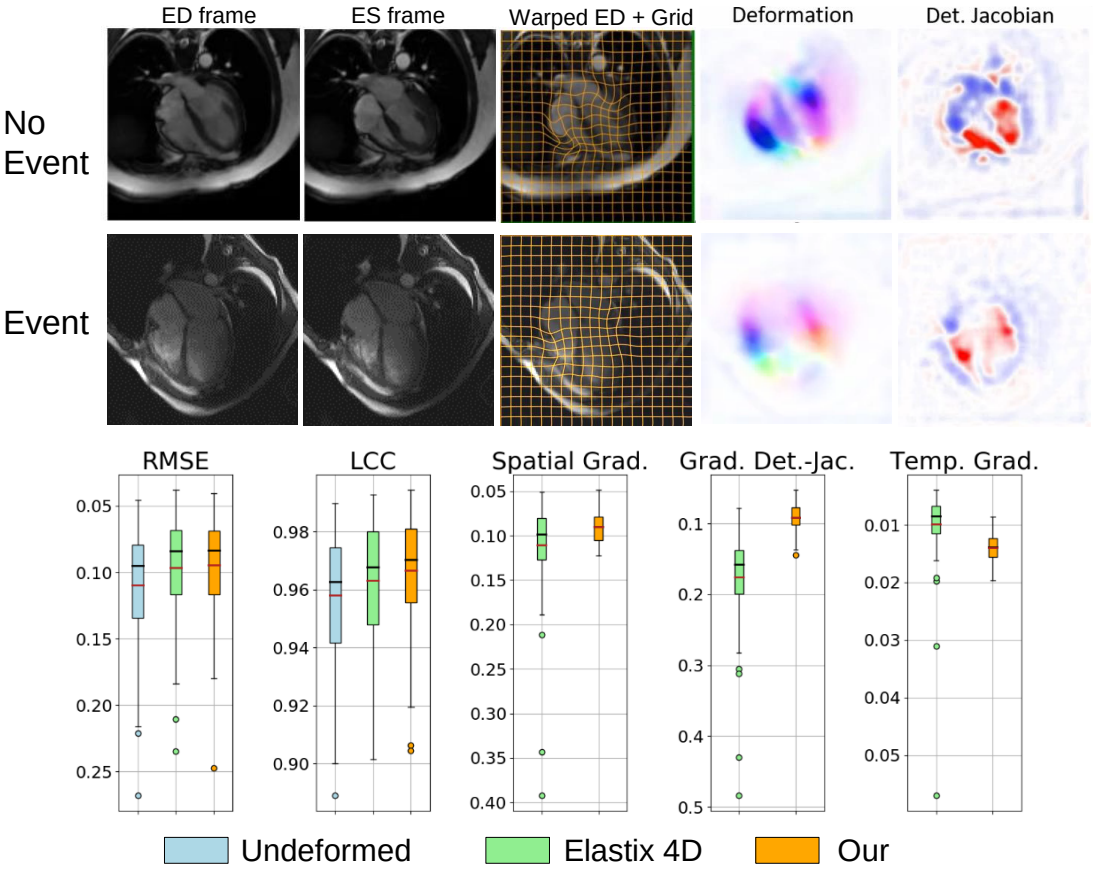


Fig. 6.3: The motion fingerprint extractor is able to learn motion patterns from 4 chamber view cine-MRI. The motion between end-diastolic (ED) and end-systolic (ES) frames are shown for two subjects, one with future HF hospitalization event and one without. The bottom shows boxplots of registration accuracy and deformation regularity in comparison to the 4D elastix algorithm in terms of root mean square (RMSE), local cross-correlation (LCC), gradient of the determinant of Jacobian (Grad. Det. Jac.), spatial and temporal gradients of the deformation field.

6.5.2 Detailed Derivations of the Fingerprint Extractor

Due to the fact that the used motion model in this chapter is different from the one presented in chapter 5 (e.g. not utilizing a Gaussian process prior), we add the full derivations here. The following sections are based on the method section in [Krebs, 2020c].

The motion observed in an image sequence with $T+1$ frames is typically described by deformation fields ϕ_t between a moving image I_0 and the fixed images I_t with $t \in [1, T]$. Inspired by the probabilistic deformation model of [Krebs, 2019b] based

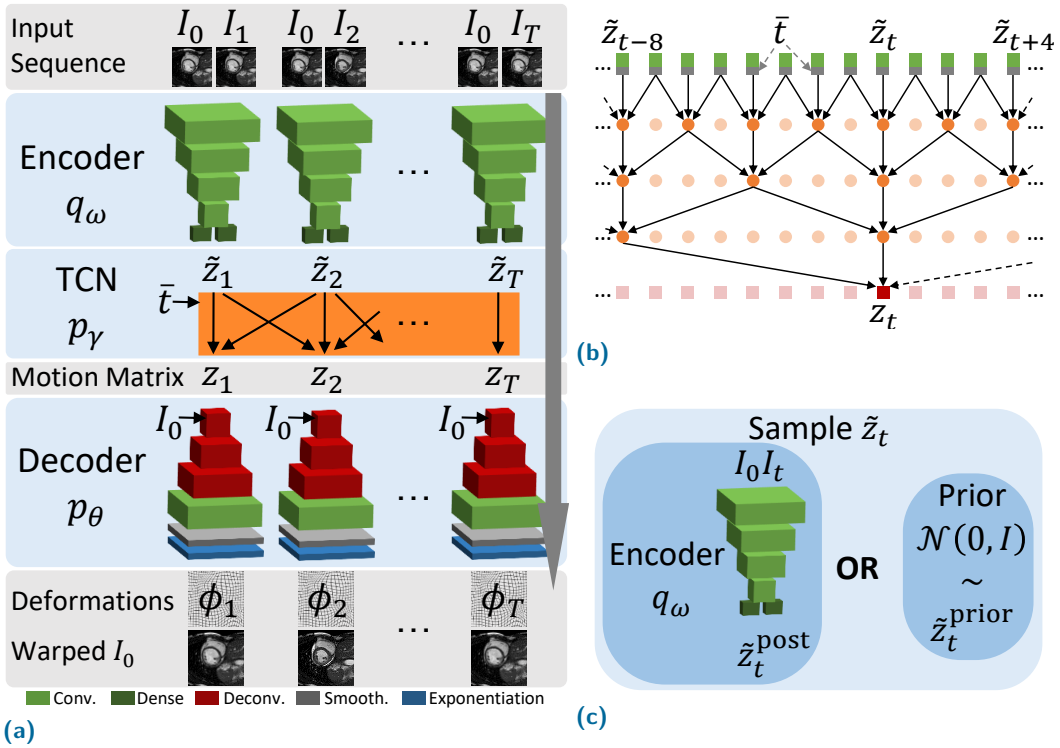


Fig. 6.4: Probabilistic motion model (a): The encoder q_ω projects the image pair (I_0, I_t) to a low-dimensional deformation encoding $\tilde{z}_t$ from which the temporal convolutional network p_γ (b) constructs the motion matrix $z \in \mathbb{R}^{d \times T}$ conditioned on the normalized time $\bar{t}$. The decoder p_θ maps the motion matrix to the deformations ϕ_t. The temporal dropout sampling procedure (c) randomly chooses to sample $\tilde{z}_t$ either from the encoder q_ω or the prior distribution.

on conditional variational autoencoder (CVAE) [Kingma, 2014b], we define a motion model for temporal sequences. The model is conditioned on the moving image and parameterizes the set of diffeomorphisms ϕ_t in a low-dimensional probabilistic space, the motion matrix $z \in \mathbb{R}^{D \times \bar{T}}$, where D is the size of the deformation encoding per image pair adn $\bar{T} = T - 1$. Each column's z_t-code corresponds to the deformation ϕ_t. To take temporal dependencies into account, z_t is conditioned on all past and future time steps. To learn this temporal regularization directly from data, we apply Temporal Convolutional Networks [Bai, 2018] with explicit time dependence and temporal dropout sampling enforcing the network to fill time steps by looking at given past and future deformations. An illustration of the model is shown in Fig. 6.4a.

Probabilistic Motion Model

Our motion model consists of three distributions. First, the encoder $q_\omega(\tilde{z}_t | I_0, I_t)$ maps each of the image pairs (I_0, I_t) independently to a latent space denoted by $\tilde{z}_t \in \mathbb{R}^d$. Second, as the key component of temporal modeling, these latent vectors $\tilde{z}_t$ are jointly mapped to the motion matrix z by conditioning them in all past and future time steps

and on the normalized time $\bar{t}$: $p_\gamma(z|\tilde{z}_{1:T}, \bar{t}_{1:T})$. Finally, the decoder $p_\theta(I_t|z_t, I_0)$ aims to reconstruct the fixed image I_t by warping the moving image I_0 with the deformation ϕ_t. This deformation ϕ_t is extracted from the temporally regularized z_t-codes. The decoder is conditioned on the moving image by concatenating the features at each scale with down-sampled versions of I_0.

The distributions q_ω, p_γ, p_θ are approximated by three neural networks with trainable parameters ω, γ, θ. During training, a lower bound on the data likelihood is maximized with respect to a prior distribution $p(\tilde{z}_t)$ of the latent space $\tilde{z}_t$ (cf. CVAE [Kingma, 2014b]). The prior $p(\tilde{z}_t)$ is assumed to follow a multivariate unit Gaussian distribution with spherical covariance I: $p(\tilde{z}_t) \sim \mathcal{N}(0, I)$. The objective function results in optimizing the expected log-likelihood p_θ and the Kullback-Leibler (KL) divergence enforcing the posterior distribution q_ω to be close to the prior $p(\tilde{z}_t)$ for all time steps:

$$\sum_{t=1}^{T} \mathbb{E}_{z_t \sim p_\gamma(\cdot|\tilde{z}_{1:T}, \bar{t}_{1:T})} \Big[\log\, p_\theta(I_t|z_t, I_0) \Big] - \mathrm{KL}\left[q_\omega(\tilde{z}_t|I_0, I_t) \| p(\tilde{z}) \right]. \tag{6.6}$$

Unlike the traditional CVAE model, the temporal regularized z_t-code is used in the log-likelihood term p_θ instead of the $\tilde{z}_t$. We model p_θ as a symmetric local cross-correlation Boltzmann distribution with the weighting factor λ. Encoder and decoder weights are independent of the time t. Their network architecture consists of convolutional and deconvolutional layers with fully-connected layers for mean and variance predictions in the encoder part [Kingma, 2014b]. We use an exponentiation layer for the stationary velocity field parameterization of diffeomorphisms [Krebs, 2019b], a linear warping layer and diffusion-like regularization with smoothing parameters σ_G in spatial and σ_T in temporal dimension.

Temporal Convolutional Networks with Explicit Time Dependence

Since the parameters of encoder q_ω and decoder p_θ are independent of time, the temporal conditioning p_γ plays an important role in merging information across different time steps. In our work, this regularization is learned by Temporal Convolutional Networks (TCN). Consisting of multiple 1-D convolutional layers with increasing dilation, TCN can handle input sequences of different lengths. TCN have several advantages compared to recurrent neural networks such as a flexible receptive field and more stable gradient computations [Bai, 2018].

The input of the TCN is the sequence of $\tilde{z}$ concatenated with the normalized time $\bar{t} = t/T$. Providing the normalized time explicitly, provides the network with information on where each $\tilde{z}$ is located in the sequence. This supports the learning of a motion model from data representing the same type of motion with varying sequence lengths. The

output of the TCN is the regularized motion matrix z. We use non-causal instead of causal convolutional layers to also take future time steps into account. We follow the standard implementation using zero-padding and skip connections. Each layer contains d filters. A schematic representation of our TCN is shown in Fig. 6.4b. For cyclic sequences, one could use a cyclic padding instead of zero-padding, for example by linking $\tilde{z}_T$ to $\tilde{z}_0$. However, in case of cardiac cine-MRI, one can not assume the end of a sequence coincides with the beginning as 5-10% of the cardiac cycle are often omitted [Bernard, 2018].

Training with Temporal Dropout Sampling

Using Eq. 6.6 for training could lead to learning the identity transform $z \approx \tilde{z}$ in the TCN p_γ such that deformations of the current time step are independent of past and future time steps. To avoid this and enforce the model to search for temporal dependencies during the training, we introduce the concept of temporal dropout sampling (TDS). In TDS, some of the $\tilde{z}_t$ are sampled from the prior distribution $p(\tilde{z})$ instead of only sampling from the posterior distribution $q_\omega(\tilde{z}_t|I_0, I_t)$ as typical for CVAE. At the time steps the prior has been used for sampling, the model has no knowledge of the target image I_t and is forced to use the temporal connections within the TCN in order to minimize the objective.

More precisely, at each time step t, a sample from the prior distribution $\tilde{z}_t^{\text{prior}} \sim p(\tilde{z}_t)$ is selected instead of a posterior sample $\tilde{z}_t^{\text{post}} \sim q_\omega(\tilde{z}_t|I_0, I_t)$ using a binary Bernoulli random variable r_t. All independent Bernoulli random variables $r \in \mathbb{R}^T$ have the success probability δ. The latent vector $\tilde{z}_t$ can be defined as:

$$\tilde{z}_t = r_t * \tilde{z}_t^{\text{prior}} + (1 - r_t) * \tilde{z}_t^{\text{post}}. \tag{6.7}$$

Fig. 6.4c illustrates the TDS procedure. At test time, for each time step independently, one can either draw $\tilde{z}_t$ from the prior or take the encoder's prediction.

Implementation Details and Training of the Fingerprint Extractor

The encoder q_ω consisted of 4 convolutional layers with strides (2, 2, 2, 1) and dense layers of size D for mean and variance estimation of the VAE. The TCN consisted of four 1-D convolutional layers with dilations (1, 2, 4, 8), *same* padding, a kernel size of 3 and skip connections (cf. Fig. 6.4b). The decoder p_θ had 3 deconvolutional and 1 convolutional layer before the exponentiation and warping layers (Fig. 6.4a). The loss weighting factor λ was chosen empirically as $6 \cdot 10^4$. The dropout sampling probability δ was 0.5. We applied a first-order gradient-based method for stochastic optimization

(Adam [Kingma, 2014a]) with a learning rate of 0.00015 and a batch size of one. We performed data augmentation on-the-fly by randomly shifting, rotating, scaling and mirroring images. We implemented the model in Tensorflow [Abadi, 2016] with Keras [Chollet, 2015]. The training time was 15h on a NVIDIA GTX TITAN X GPU.

7 Conclusion

Contents

In this thesis, we presented computational frameworks for the analysis of medical image pairs and image sequences. We built upon state-of-the-art methods for designing accurate and reliable registration and motion analysis tools that can be applied in clinical research by facilitating diagnosis, prognosis and therapy of diseases.

The proposed methods utilize recent machine learning methods showing high computational efficiency. Furthermore, the use of artificial intelligence (AI) in this work demonstrates how powerful compact models can be learned from large datasets of images. The developed tools were designed to find application in difficult inter-subject registration and in intra-subject motion tracking scenarios. For the latter, compact deformation and motion models from sequential images were proposed that enable a variety of analysis tools to quantify and compare deformations. The proposed algorithms were tested on publicly available datasets allowing to benchmark and compare results. While the first 5 chapters are intended for a broader range of applications focusing on the technical contributions of this thesis, in chapter 6, one potential clinical application is shown. It is demonstrated how the proposed motion model can directly support prognosis and therapy planning by predicting the survival risk of patients suffering from heart failure (HF). This could allow for a better patient selection for available therapies.

7.1 Main Contributions

In chapter 3, we proposed **a generic learning-based framework using an artificial agent for difficult inter-subject registration tasks** appearing in organ-focused non-

rigid image fusion and atlas-based segmentation. The proposed method overcomes limitations of traditional algorithms by learning optimal features for registration. Inspired, by deep reinforcement learning the registration problem was reformulated as the iterative optimization of deformation parameters through an artificial agent. Hereby, the agent (a neural network) optimized the parameters of a simple statistical deformation model (SDM) learned from data. In an iterative fashion, the optimal transformation parameters were approached on a trajectory of small deformations. To restrict the agent to a set of reasonable transformations, fuzzy action control has been introduced which sets limits to the parameters of the SDM. During training, a novel ground-truth generator was used. This generator relied on simulated deformations from an SDM and a few *ground-truth* inter-subject deformation fields that were enhanced by segmentations. We showed that the agent-based approach trained with data from the novel ground-truth generator outperformed three state-of-the art registration algorithms in terms of structure overlaps and distances.

We presented an **unsupervised deformable registration approach that learns a low-dimensional probabilistic deformation model** in chapter 4. The deformation model is based on a conditional variational autoencoder (CVAE). It not only allows for accurately registering two images but also for analyzing corresponding deformations efficiently by using a novel generative deformation encoding. In this encoded latent space, similar deformations are close to each other. This enables to cluster and simulate deformations for a given image. Furthermore, it provides a novel way of transporting deformations from one subject to another without requiring inter-subject registration. The model can be seen as a non-linear and richer generalization of a simple statistical deformation model such as PCA. The unsupervised method is based on variational inference. In addition, we introduced a novel exponentiation layer to make DL-based registration algorithms diffeomorphic utilizing the SVF parameterization. An extended version, allows to train the model in a multi-scale fashion which results in higher accuracy. We evaluated the approach on end-diastole to end-systole cardiac cine-MRI registration. In comparison to 3 state-of-the-art algorithms, our multi-scale model showed significantly improved registration accuracy and regularity. The latent encoding showed convincing generative and deformation transport capabilities and showed a 83% classification accuracy for differentiating 5 cardiac diseases.

Beyond pairwise registration, we proposed **a probabilistic motion model** in chapter 5. This model can be useful for spatio-temporal registration, temporal super-resolution, data augmentation, shorter acquisition times and other motion analysis tasks. Intrinsic motion patterns are encoded in a low-dimensional probabilistic space – the latent motion matrix – which allows for accurate tracking of structures, temporal interpolation, motion simulation and motion transport. The diffeomorphic motion model is trained as a temporal latent variable model utilizing a novel Gaussian process prior acting on the latent motion encoding and following the training principles of CVAEs. Applied on

cardiac cine-MRI, our approach has shown state-of-the art registration accuracy and improved temporal and spatial deformation regularity in comparison to 3 state-of-the-art algorithms. These results indicated that the latent motion encoding helps to regularize the registration problem of image sequences. Besides, we demonstrated the model's applicability for motion analysis by simulating realistic motion patterns, by transporting the motion to simulate a pathology in a healthy case and by an improved motion reconstruction from sequences with missing frames.

In chapter 6, we presented how our low-dimensional motion model can be **applied for risk estimation and disease outcome prediction** in heart failure patients. We have proposed a neural network risk predictor based on a non-linear Cox regression loss to estimate different disease endpoints from a motion fingerprint. Hereby, the fingerprint (the motion matrix) was extracted by applying the motion model from the previous chapter to 4 chamber-view cine-MRI. We evaluated the risk predictor on a cohort of heart failure patients with known endpoints such as hospitalization and sudden cardiac death (SCD). We have shown that the risk score predicted from the motion fingerprint is the most predictive independent feature for survival in comparison to other clinical features that have been known to be independently predictive for HF endpoints.

7.2 Perspectives and Future Applications

The proposed methods have proven to be accurate and suitable for the given applications in this thesis. However, one goal was to develop tools that are generalizable and applicable to other data and applications in medical image analysis. Therefore, we believe that the proposed tools could find further application for the study of registration and motion scenarios including different diseases, organs and imaging modalities. Due to the fact that the objective functions for the proposed deformation and motion model can include principally any differentiable similarity and regularization metric (as in traditional registration methods), it makes these models suitable to a large variety of applications including for example multi-modal registration. In addition, future work should focus on the interpretability of the proposed latent variable models.

7.2.1 Motion Model for Cardiac Sequences from other Modalities

In a first step of generalization, one can think of applying the proposed motion model from chapter 5 to cardiac sequences from other modalities such as ultrasound or computer tomography images. Cardiac ultrasound (or echocardiography) is the most widely used and readily available imaging modality to assess cardiac function and structure. We

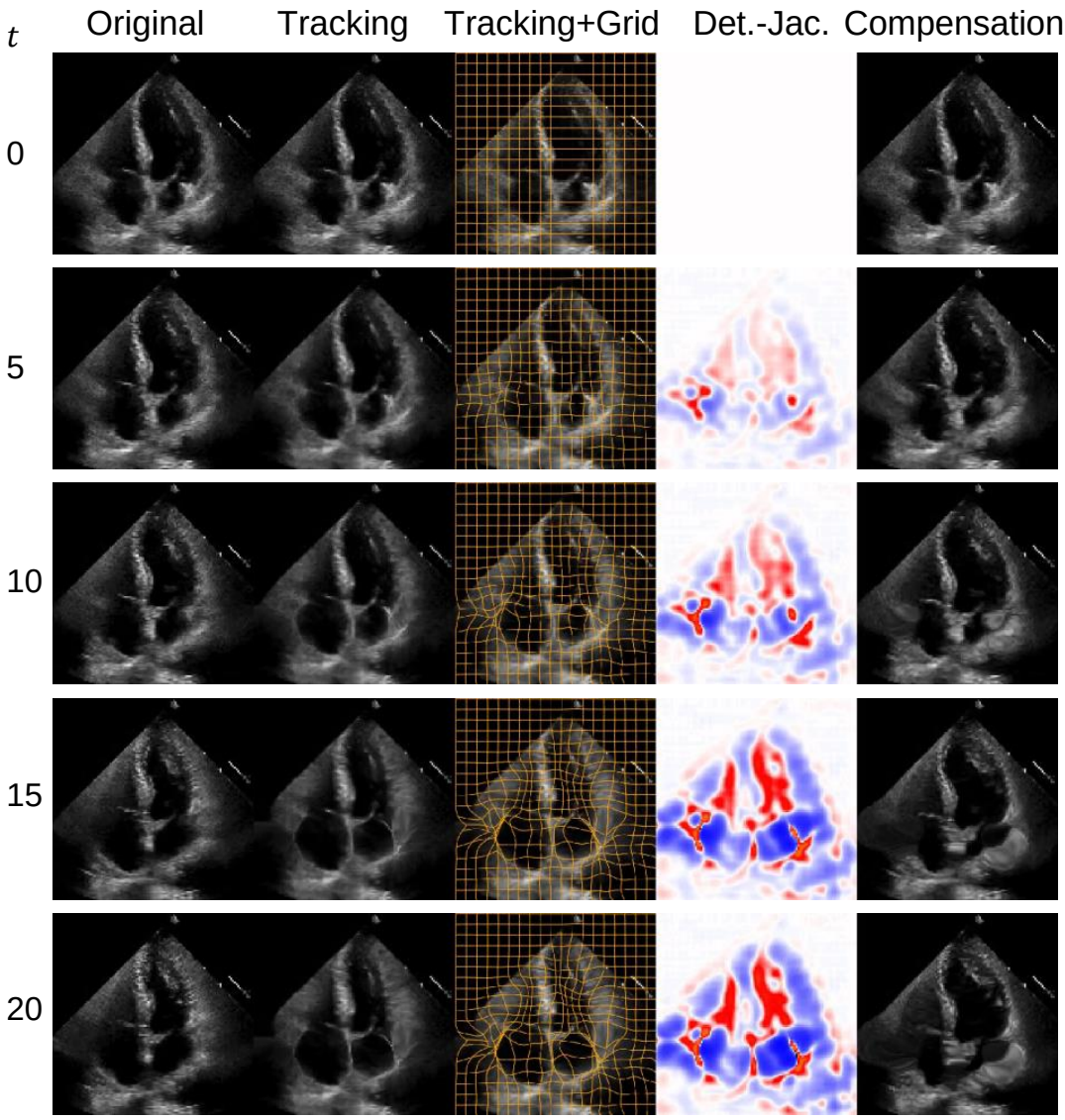


Fig. 7.1: After training the motion model on the Echonet dataset [Ouyang, 2020], an example test sequence with extracted tracking results, Jacobian determinants and motion-compensated images is shown.

already conducted preliminary experiments that show the applicability of our method to a publicly available database of cardiac ultrasound image sequences, the Echonet [Ouyang, 2020]. This dataset contains 10.030 ultrasound videos. We extracted approximately one cardiac sequence given the annotated ED and ES frames. We followed the given division in 7550 training, 1287 validation, 1275 test splits and trained our motion model with the same hyperparameters as described for the cine-MRI. In Fig. 7.1, an example test sequence, the tracking results, Jacobian determinants and motion-compensated images are shown.

Furthermore, simulated motion is shown in case of transporting the motion matrix of a test sequence with high ejection fraction (EF) to one sequence with low ejection fraction

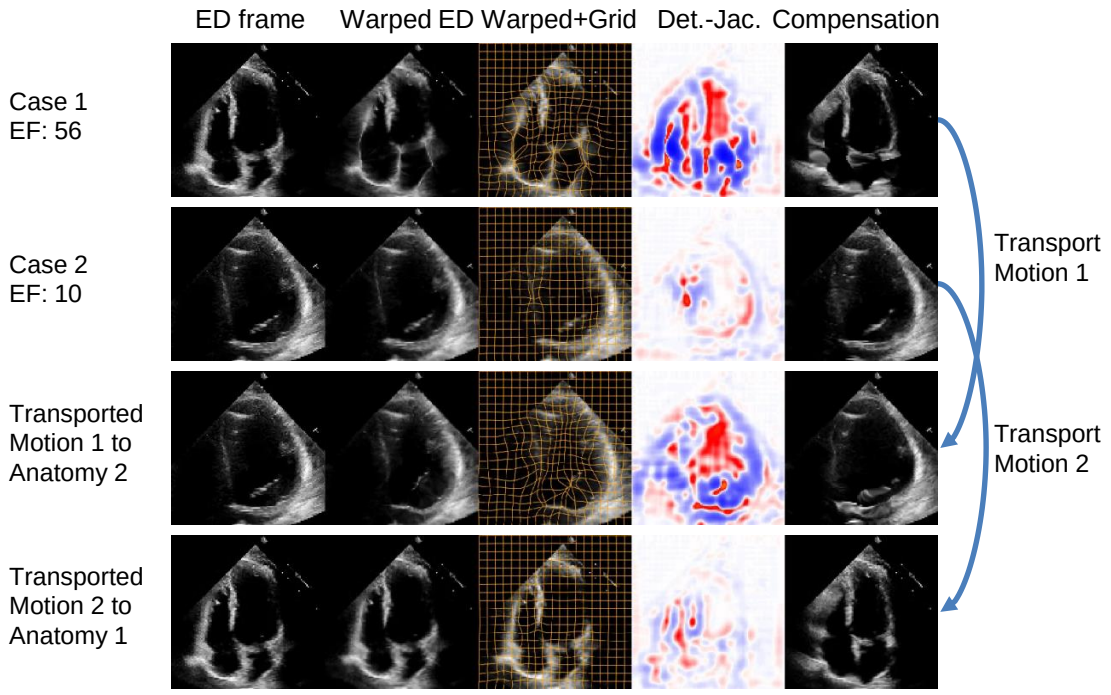


Fig. 7.2: Transported Motion from a case with hihg ejection fraction (EF) to one case with low EF and vice versa.

in Fig. 7.2. These first preliminary results are promising and suggest that the motion model is also applicable to echocardiography sequences of the heart without major adjustments or modifications.

7.2.2 Interpretability and Causability in Deep Latent Variable Models

In PCA, the different variables have a clear structure and the first components are often interpretable: they describe the direction of the largest possible variance of the data and each succeeding component has the highest possible variance while being orthogonal to the preceding components. On the other hand deep latent variable models, such as the ones presented in this thesis, do not offer equivalent means of interpretation for the individual latent variables. By making assumptions of certain prior distributions such as multivariate Gaussians, the latent variables are enforced to be more structured compared to standard autoencoder networks in the sense that similar data points are close to each other in the latent space [Kingma, 2019]. This allows for example to interpolate between data points. However, a deeper interpretability of the latent variables is not available. In future work, it is desirable to investigate the interpretability and causability of latent variable models as this can lead to an improved understanding of the latent encoded motion model and its reasoning [Holzinger, 2019]. The goal is to not only provide a model that can be used for the tasks tackled in this manuscript but to also understand

(in a human explainable way) which features and which characteristics of the images and its deformations are the most important ones for example for predicting disease outcomes. Providing such explainable decisions would make it easier for physicians to trust deep latent variable models and AI-based algorithms in general. A simple way to improve explainability is by looking at the model's feature maps and find distinctive patterns between different pathologies. In addition, it could be helpful to study the model's attention using saliency maps [Simonyan, 2013].

While looking at the network's attention, gives more insights about which parts of the data lead to a certain prediction, a more clinically motivated future direction is to integrate known features into the latent space. In terms of cardiac motion, one could incorporate classical clinical features such as ejection fraction or strain values and thus, enhance interpretability. For risk prediction, the inclusion of clinical features at different stages of a standard autoencoder network has been preliminary investigated by Ji et al. [Jin, 2019].

In another possible approach, one could think of interpreting all or some latent variables as the parameters of a known biomechanical model. In this case, the decoding part to retrieve dense deformation fields could be replaced by the biomechanical model and the motion model would predict optimal parameter values for the biomechanical model solely from a pair or sequence of given images.

7.2.3 Beyond Predicting Heart Failure Disease Outcomes

We have shown the usefulness of the motion model for predicting disease outcomes such as hospitalization and SCD for heart failure patients. However, this is only the first step in the automatic image-driven feature analysis. The proposed risk prediction model and possible variants (e.g. using end-to-end learning) could be useful to identify and reveal unknown clinical features that are significantly predictive for disease outcomes such as the ones mentioned above. Closely related to the interpretability and causability of the model as mentioned above, these features could be extracted by introspection – by analyzing the neural networks' behavior in disease-specific cases. Besides, the diagnosis and prognosis for a single patient, this could impact clinical research directly and lead to a better understanding of heart failure and related risks.

Moreover, multiple other heart diseases that have been associated with an impaired cardiac motion could benefit from the proposed risk model. One example is pulmonary hypertension which is characterized by right ventricular dysfunction [Farber, 2004]. Here, an image-based automatic feature retrieval could also reveal new unknown cardiac motion factors that influence the disease.

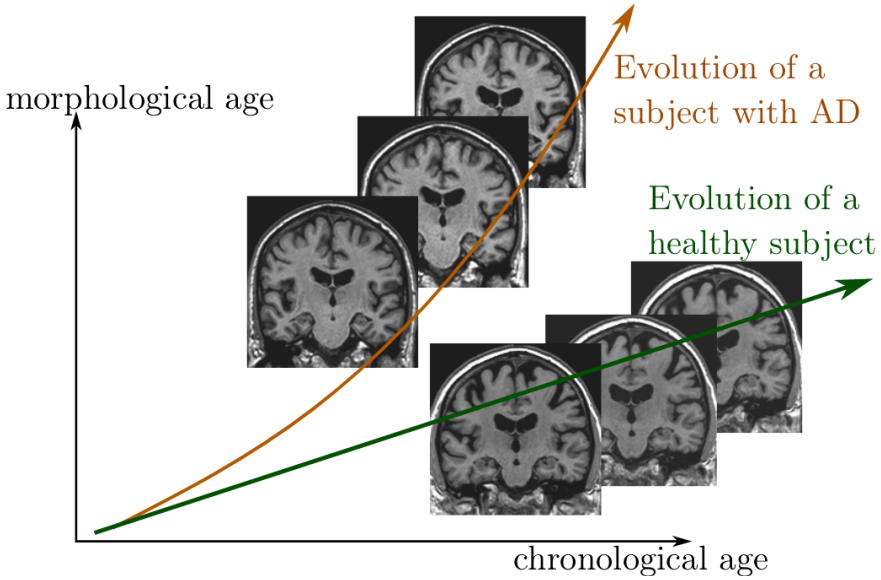


Fig. 7.3: The Alzheimer's disease (AD) is assumed to have a faster morphological degeneration (aging) than healthy people [Sivera, 2019].

7.2.4 Deformation Model for Studying Neurodegenerative Diseases

The thesis focused on the development of a deformation and motion model of the heart from images of the same patient determining intra-subject deformations. While this is helpful for analyzing moving body organs, one future direction could be to learn a deformation model across patients depicting inter-subject deformations. Such a model could be useful for determining disease progression in a patient. An active research topic that comes to mind is the analysis of neurodegenerative diseases. The progress of Alzheimer's disease and aging are known to cause morphological changes in the human brain [Rosen, 2003; Ohnishi, 2001]. These changes can be extracted by image registration of a subject's brain MRI to a template. In combination with a learned template model (e.g. [Dalca, 2019b]), a low-dimensional generative deformation model could provide novel insights in the analysis of brain aging and neurodegenerative diseases. Further on, one could potentially predict the disease progression in a patient by comparing the evolution of healthy and unhealthy brains [Sivera, 2019; Nader, 2020] (cf. Fig. 7.3). Thus, another way of applying the proposed motion model to neurodegenerative diseases is to learn the brain evolution in a patient. This could be done by learning a brain deformation model from longitudinal images where the temporal deformations depict structural brain changes over long time intervals rather than real-time organ motion as from the heart. Using such a low-dimensional temporal deformation model could help in characterizing the personalized disease progression in a patient and guide the therapy. However, modeling the more complex morphological changes in the brain may require

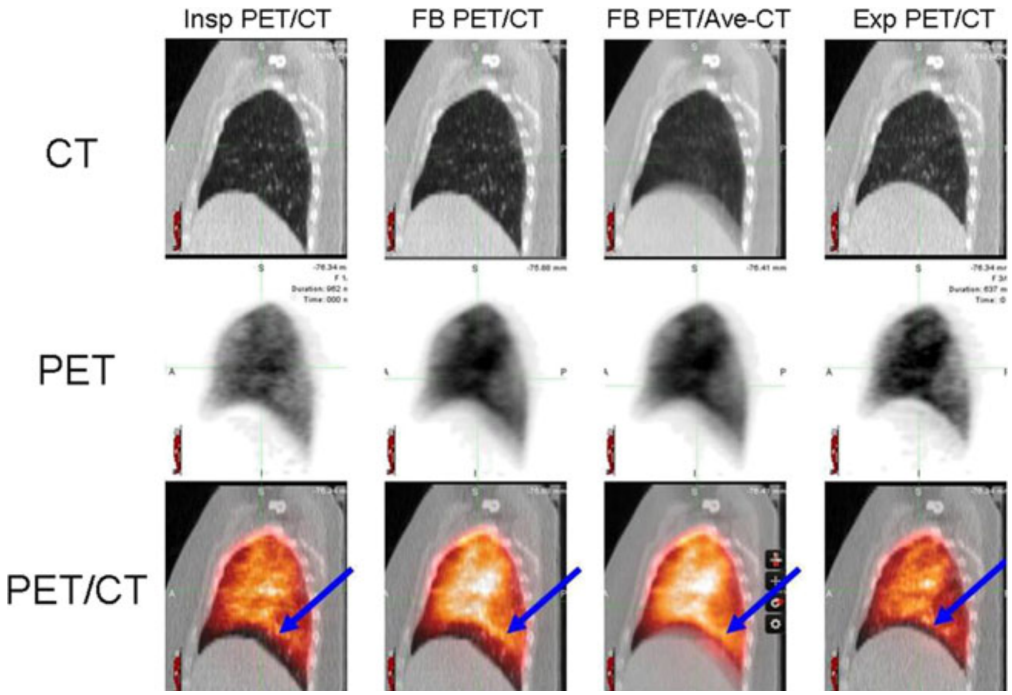


Fig. 7.4: Registration misalignment for CT and PET images induced from respiratory motion. Different breathing states are illustrated, free-breathing (FB), maximal inspiration (Insp) and maximal expiration (Exp) [Callahan, 2014].

an adaptation of the latent *motion* matrix in order to deal with this extra amount of deformation complexity.

7.2.5 Respiratory Motion Model

Another potential area of application is the study of lung and respiratory motion. A crucial need in the analysis of images for lung diseases (such as CT) and for example for tumors in the abdomen is motion compensation [Ozhasoglu, 2002]. In PET imaging for instance, respiratory motion causes artifacts in reconstructed images, which can lead to misinterpretations, imprecise diagnosis or the impairing of fusion with other modalities [Reyes, 2007]. Often PET images need to be registered to CT images in order to map structural to functional images. In Fig. 7.4, one can see the misalignment between CT and PET images induced from respiratory motion as the images were taken at different breathing states [Callahan, 2014]. A comprehensive probabilistic motion model could help in compensating for these motion artifacts within a mono-modal sequence (CT) and allow for reliable multi-modal registration in a second step.

Furthermore, a compact motion model could help in the diagnosis and prognosis of lung diseases. It has been shown that an impaired lung function is associated with increased mortality rates [Beaty, 1985]. The proposed motion model already demonstrated its

capabilities for predicting disease outcomes from cardiac image sequences. Thus, we believe it can be also useful for motion-related lung diseases.

7.2.6 Deformation and Motion Modeling in Personalized Medicine

We conclude with a broader view on the positioning of this work in a long-term outlook. In the past couple of years, machine-learning methods have been successfully applied to a wide variety of applications in medical image analysis [Litjens, 2017]. Typically, these models are gathering experience from large databases in order to solve specific problems. The next logical step for personalized medicine is how to combine this specific knowledge to form something larger, a central system that is able to link information across applications. Already today, a physician has to take the patient's pre-existing conditions, his health history, his age and many other factors into account before reaching conclusions about diagnosis, prognosis and therapy. A system that helps in the analysis of the increasingly growing amount of information can be highly beneficial in the healthcare of tomorrow. Building a supporting system that combines a collection of computational models describing and simulating the human body of a patient has been termed virtual patient or digital twin. While such a model goes far beyond medical image analysis as it involves basically all available data of a patient, medical images, most certainly, will still be of crucial importance.

The models presented in this thesis are already designed to extract relevant information from medical images and create meaningful compact representations or task-specific fingerprints using modern machine learning techniques. Furthermore, we have shown that these personalized fingerprints enable a variety of analysis tasks such as predicting possible outcomes or simulating diseases. Thus, we believe that such personalized fingerprints of a patient can play an important role in the creation of a comprehensive digital twin. In terms of deformation and motion models, one could think of learning an ensemble of organ-specific and/or disease-specific models. This ensemble of fingerprints, could then be one part of the virtual patient helping and supporting the patient's health during all stages of the clinical workflow and whenever necessary.

Bibliography

[Abadi, 2016] Martin Abadi, Ashish Agarwal, Paul Barham, et al. "Tensorflow: Large-scale machine learning on heterogeneous distributed systems". In: *arXiv preprint arXiv:1603.04467* (2016) (cit. on pp. 73, 90, 98).

[Adabag, 2012] Selcuk Adabag, Lindsay G Smith, Inder S Anand, Alan K Berger, and Russell V Luepker. "Sudden cardiac death in heart failure patients with preserved ejection fraction". In: *Journal of cardiac failure* 18.10 (2012), pp. 749–754 (cit. on p. 3).

[Arsigny, 2006] Vincent Arsigny, Olivier Commowick, Xavier Pennec, and Nicholas Ayache. "A log-euclidean framework for statistics on diffeomorphisms". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2006, pp. 924–931 (cit. on pp. 13, 34, 37, 39, 46, 62).

[Avants, 2008] Brian B Avants, Charles L Epstein, Murray Grossman, and James C Gee. "Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain". In: *Medical image analysis* 12.1 (2008), pp. 26–41 (cit. on pp. 13, 34, 47, 54, 62, 73, 74).

[Avants, 2011] Brian B Avants, Nicholas J Tustison, Gang Song, et al. "A reproducible evaluation of ANTs similarity metric performance in brain image registration". In: *Neuroimage* 54.3 (2011), pp. 2033–2044 (cit. on p. 40).

[Bacoyannis, 2019] Tania Bacoyannis, Julian Krebs, Nicolas Cedilnik, Hubert Cochet, and Maxime Sermesant. "Deep Learning Formulation of ECGI for Data-driven Integration of Spatiotemporal Correlations and Imaging Information". In: *International Conference on Functional Imaging and Modeling of the Heart*. Springer. 2019, pp. 20–28 (cit. on p. 6).

[Bai, 2015] Wenjia Bai, Wenzhe Shi, Antonio de Marvao, et al. "A bi-ventricular cardiac atlas built from 1000+ high resolution MR images of healthy subjects and an analysis of shape and motion". In: *Medical image analysis* 26.1 (2015), pp. 133–145 (cit. on p. 35).

[Bai, 2018] Shaojie Bai, J Zico Kolter, and Vladlen Koltun. “An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling”. In: *arXiv preprint arXiv:1803.01271* (2018) (cit. on pp. 64, 69, 70, 86, 95, 96).

[Balakrishnan, 2018] Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. “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 (cit. on pp. 18, 35, 38, 49, 68).

[Balakrishnan, 2019] Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. “VoxelMorph: a learning framework for deformable medical image registration”. In: *IEEE transactions on medical imaging* 38.8 (2019), pp. 1788–1800 (cit. on p. 18).

[Beaty, 1985] TH Beaty, CA Newill, BH Cohen, et al. “Effects of pulmonary function on mortality”. In: *Journal of chronic diseases* 38.8 (1985), pp. 703–710 (cit. on p. 106).

[Beg, 2005] M Faisal Beg, Michael I Miller, Alain Trouvé, and Laurent Younes. “Computing large deformation metric mappings via geodesic flows of diffeomorphisms”. In: *International journal of computer vision* 61.2 (2005), pp. 139–157 (cit. on pp. 13, 34, 62).

[Bello, 2019] Ghalib A Bello, Timothy JW Dawes, Jinming Duan, et al. “Deep-learning cardiac motion analysis for human survival prediction”. In: *Nature machine intelligence* 1.2 (2019), pp. 95–104 (cit. on pp. 84, 88).

[Bernard, 2018] Olivier Bernard, Alain Lalande, Clement Zotti, et al. “Deep Learning Techniques for Automatic MRI Cardiac Multi-structures Segmentation and Diagnosis: Is the Problem Solved?” In: *IEEE Transactions on Medical Imaging* 37.11 (2018), pp. 2514–2525 (cit. on pp. 11, 46, 54, 70, 72, 80, 97).

[Bharatha, 2001] Aditya Bharatha, Masanori Hirose, Nobuhiko Hata, et al. “Evaluation of three-dimensional finite element-based deformable registration of pre- and intraoperative prostate imaging”. In: *Medical physics* 28.12 (2001), pp. 2551–2560 (cit. on p. 14).

[Biffi, 2018] Carlo Biffi, Ozan Oktay, Giacomo Tarroni, et al. “Learning Interpretable Anatomical Features Through Deep Generative Models: Application to Cardiac Remodeling”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2018, pp. 464–471 (cit. on pp. 36, 54).

[Bookstein, 1991] Fred L Bookstein. “Thin-plate splines and the atlas problem for biomedical images”. In: *Biennial International Conference on Information Processing in Medical Imaging*. Springer. 1991, pp. 326–342 (cit. on p. 12).

[Boveiri, 2020] Hamid Reza Boveiri, Raouf Khayami, Reza Javidan, and Ali Reza MehdiZadeh. “Medical Image Registration Using Deep Neural Networks: A Comprehensive Review”. In: *arXiv preprint arXiv:2002.03401* (2020) (cit. on p. 9).

[Burger, 2013] Martin Burger, Jan Modersitzki, and Lars Ruthotto. “A hyperelastic regularization energy for image registration”. In: *SIAM Journal on Scientific Computing* 35.1 (2013), B132–B148 (cit. on p. 34).

[Caballero, 2017] Jose Caballero, Christian Ledig, Andrew Aitken, et al. “Real-time video super-resolution with spatio-temporal networks and motion compensation”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 4778–4787 (cit. on p. 63).

[Callahan, 2014] Jason Callahan, Michael S Hofman, Shankar Siva, et al. “High-resolution imaging of pulmonary ventilation and perfusion with 68 Ga-VQ respiratory gated (4-D) PET/CT”. In: *European journal of nuclear medicine and molecular imaging* 41.2 (2014), pp. 343–349 (cit. on p. 106).

[Cao, 2005] Yan Cao, Michael I Miller, Raimond L Winslow, Laurent Younes, et al. “Large deformation diffeomorphic metric mapping of vector fields”. In: *IEEE transactions on medical imaging* 24.9 (2005), pp. 1216–1230 (cit. on pp. 13, 34).

[Cao, 2017] Xiaohuan Cao, Jianhua Yang, Jun Zhang, et al. “Deformable image registration based on similarity-steered CNN regression”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2017, pp. 300–308 (cit. on p. 16).

[Chollet, 2015] François Chollet et al. *Keras*. `https://keras.io`. 2015 (cit. on pp. 73, 90, 98).

[Christensen, 1996] Gary E Christensen, Richard D Rabbitt, and Michael I Miller. “Deformable templates using large deformation kinematics”. In: *IEEE transactions on image processing* 5.10 (1996), pp. 1435–1447 (cit. on p. 12).

[Cootes, 1995] T.F. Cootes, C.J. Taylor, D.H. Cooper, and J. Graham. “Active Shape Models-Their Training and Application”. In: *Computer Vision and Image Understanding* 61.1 (1995), pp. 38–59 (cit. on p. 13).

[Cox, 1972] David R Cox. “Regression models and life-tables”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 34.2 (1972), pp. 187–202 (cit. on pp. 86–88, 90).

[Dalca, 2018] Adrian V Dalca, Guha Balakrishnan, John Guttag, and Mert R Sabuncu. “Unsupervised learning for fast probabilistic diffeomorphic registration”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2018, pp. 729–738 (cit. on pp. 18, 35, 37, 45, 47, 49, 54, 55, 62).

[Dalca, 2019a] Adrian V Dalca, Guha Balakrishnan, John Guttag, and Mert R Sabuncu. “Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces”. In: *Medical image analysis* 57 (2019), pp. 226–236 (cit. on p. 18).

[Dalca, 2019b] Adrian Dalca, Marianne Rakic, John Guttag, and Mert Sabuncu. "Learning conditional deformable templates with convolutional networks". In: *Advances in neural information processing systems*. 2019, pp. 804–816 (cit. on p. 105).

[Davatzikos, 1997] Christos Davatzikos. "Spatial transformation and registration of brain images using elastically deformable models". In: *Computer Vision and Image Understanding* 66.2 (1997), pp. 207–222 (cit. on pp. 12, 34).

[DavidsonPilon, 2020] Cameron Davidson-Pilon. *CamDavidsonPilon/lifelines: v0.24.0*. Version v0.24.0. Feb. 2020 (cit. on p. 90).

[Davis, 1997] Malcolm H Davis, Alireza Khotanzad, Duane P Flamig, and Steven E Harms. "A physics-based coordinate transformation for 3-D image matching". In: *IEEE transactions on medical imaging* 16.3 (1997), pp. 317–328 (cit. on p. 12).

[De Craene, 2012] Mathieu De Craene, Gemma Piella, Oscar Camara, et al. "Temporal diffeomorphic free-form deformation: Application to motion and strain estimation from 3D echocardiography". In: *Medical image analysis* 16.2 (2012), pp. 427–450 (cit. on pp. 14, 63).

[Dosovitskiy, 2015] A. Dosovitskiy, P. Fischer, E. Ilg, et al. "FlowNet: Learning Optical Flow with Convolutional Networks". In: *IEEE International Conference on Computer Vision (ICCV)*. 2015 (cit. on pp. 15, 21, 22).

[Duchateau, 2011] Nicolas Duchateau, Mathieu De Craene, Gemma Piella, et al. "A spatiotemporal statistical atlas of motion for the quantification of abnormal myocardial tissue velocities". In: *Medical image analysis* 15.3 (2011), pp. 316–328 (cit. on p. 35).

[Duchi, 2007] John Duchi. "Derivations for linear algebra and optimization". In: *Berkeley, California* 3 (2007), pp. 2325–5870 (cit. on p. 80).

[ElGamal, 2016] F. E. A. El-Gamal, M. Elmogy, and A. Atwan. "Current trends in medical image registration and fusion". In: *Egyptian Informatics Journal* 17.1 (2016), pp. 99–124 (cit. on p. 9).

[Eppenhof, 2018a] Koen AJ Eppenhof, Maxime W Lafarge, Pim Moeskops, Mitko Veta, and Josien PW Pluim. "Deformable image registration using convolutional neural networks". In: *Medical Imaging 2018: Image Processing*. Vol. 10574. International Society for Optics and Photonics. 2018, 105740S (cit. on pp. 16, 34).

[Eppenhof, 2018b] Koen AJ Eppenhof and Josien PW Pluim. "Error estimation of deformable image registration of pulmonary CT scans using convolutional neural networks". In: *Journal of Medical Imaging* 5.2 (2018), p. 024003 (cit. on p. 16).

[Evers, 2008] Ludger Evers and Claudia-Martina Messow. "Sparse kernel methods for high-dimensional survival data". In: *Bioinformatics* 24.14 (2008), pp. 1632–1638 (cit. on p. 93).

[Fan, 2018] Jingfan Fan, Xiaohuan Cao, Zhong Xue, Pew-Thian Yap, and Dinggang Shen. "Adversarial Similarity Network for Evaluating Image Alignment in Deep Learning Based Registration". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2018, pp. 739–746 (cit. on pp. 18, 37).

[Fan, 2019] Jingfan Fan, Xiaohuan Cao, Pew-Thian Yap, and Dinggang Shen. "BIRNet: Brain image registration using dual-supervised fully convolutional networks". In: *Medical image analysis* 54 (2019), pp. 193–206 (cit. on pp. 18, 35, 54).

[Farber, 2004] Harrison W Farber and Joseph Loscalzo. "Pulmonary arterial hypertension". In: *New England Journal of Medicine* 351.16 (2004), pp. 1655–1665 (cit. on p. 104).

[Fischer, 2002] Bernd Fischer and Jan Modersitzki. "Fast diffusion registration". In: *Contemporary Mathematics* 313 (2002), pp. 117–128 (cit. on p. 12).

[Fortuin, 2019] Vincent Fortuin, Gunnar Rätsch, and Stephan Mandt. "Multivariate time series imputation with variational autoencoders". In: *arXiv preprint arXiv:1907.04155* (2019) (cit. on pp. 67, 73).

[Fu, 2019] Yabo Fu, Yang Lei, Tonghe Wang, et al. "Deep Learning in Medical Image Registration: A Review". In: *arXiv preprint arXiv:1912.12318* (2019) (cit. on p. 9).

[Ghesu, 2016] F. C. Ghesu, B. Georgescu, T. Mansi, et al. "An artificial agent for anatomical landmark detection in medical images". In: *Medical Image Computing and Computer Assisted Intervention (MICCAI)*. 2016 (cit. on pp. 22, 23).

[Ghosal, 2017] Sayan Ghosal and Nilanjan Ray. "Deep deformable registration: Enhancing accuracy by fully convolutional neural net". In: *Pattern Recognition Letters* 94 (2017), pp. 81–86 (cit. on p. 18).

[Girija, 2017] J Girija, GN Krishna Murthy, and P Chenna Reddy. "4D medical image registration: A survey". In: *2017 International Conference on Intelligent Sustainable Systems (ICISS)*. IEEE. 2017, pp. 539–547 (cit. on p. 62).

[Goel, 2010] Manish Kumar Goel, Pardeep Khanna, and Jugal Kishore. "Understanding survival analysis: Kaplan-Meier estimate". In: *International journal of Ayurveda research* 1.4 (2010), p. 274 (cit. on p. 91).

[Goodfellow, 2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, et al. "Generative adversarial nets". In: *Advances in neural information processing systems*. 2014, pp. 2672–2680 (cit. on pp. 18, 36).

[Goodfellow, 2016] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. `http://www.deeplearningbook.org`. MIT Press, 2016 (cit. on p. 14).

[Harrell, 1982] Frank E Harrell, Robert M Califf, David B Pryor, Kerry L Lee, and Robert A Rosati. “Evaluating the yield of medical tests”. In: *Jama* 247.18 (1982), pp. 2543–2546 (cit. on p. 90).

[Haskins, 2020] Grant Haskins, Uwe Kruger, and Pingkun Yan. “Deep learning in medical image registration: a survey”. In: *Machine Vision and Applications* 31.1 (2020), p. 8 (cit. on pp. 9, 15, 16).

[Hering, 2019] Alessa Hering, Sven Kuckertz, Stefan Heldmann, and Mattias P Heinrich. “Enhancing label-driven deep deformable image registration with local distance metrics for state-of-the-art cardiac motion tracking”. In: *Bildverarbeitung für die Medizin 2019*. Springer, 2019, pp. 309–314 (cit. on p. 18).

[Holzinger, 2019] Andreas Holzinger, Georg Langs, Helmut Denk, Kurt Zatloukal, and Heimo Müller. “Causability and explainability of artificial intelligence in medicine”. In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 9.4 (2019), e1312 (cit. on p. 103).

[Hu, 2018] Yipeng Hu, Marc Modat, Eli Gibson, et al. “Weakly-Supervised Convolutional Neural Networks for Multimodal Image Registration”. In: *Medical Image Analysis* (2018) (cit. on pp. 19, 35, 54).

[Issa, 2017] Omar Issa, Julio G Peguero, Carlos Podesta, et al. “Left atrial size and heart failure hospitalization in patients with diastolic dysfunction and preserved ejection fraction”. In: *Journal of cardiovascular echography* 27.1 (2017), p. 1 (cit. on p. 89).

[Jablonowski, 2017] Robert Jablonowski, Uzma Chaudhry, Jesper Van Der Pals, et al. “Cardiovascular magnetic resonance to predict appropriate implantable cardioverter defibrillator therapy in ischemic and nonischemic cardiomyopathy patients using late gadolinium enhancement border zone: comparison of four analysis methods”. In: *Circulation: Cardiovascular Imaging* 10.9 (2017), e006105 (cit. on pp. 84, 89).

[Jaderberg, 2015] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. “Spatial transformer networks”. In: *Advances in neural information processing systems*. 2015, pp. 2017–2025 (cit. on pp. 17, 35, 38, 70, 72).

[Jason, 2016] J Yu Jason, Adam W Harley, and Konstantinos G Derpanis. “Back to basics: Unsupervised learning of optical flow via brightness constancy and motion smoothness”. In: *European Conference on Computer Vision*. Springer. 2016, pp. 3–10 (cit. on pp. 18, 35, 38).

[Jia, 2018] Shuman Jia, Nicolas Duchateau, Pamela Moceri, Maxime Sermesant, and Xavier Pennec. “Parallel Transport of Surface Deformations from Pole Ladder to Symmetrical Extension”. In: *ShapeMI MICCAI 2018: Workshop on Shape in Medical Imaging*. 2018 (cit. on p. 36).

[Jin, 2019] Shihao Jin, Nicolò Savioli, Antonio de Marvao, et al. “Joint analysis of clinical risk factors and 4D cardiac motion for survival prediction using a hybrid deep learning network”. In: *arXiv preprint arXiv:1910.02951* (2019) (cit. on p. 104).

[Juan, 2007] Luo Juan and Luo Gwon. “A comparison of sift, pca-sift and surf”. In: *International Journal of Signal Processing, Image Processing and Pattern Recognition* 8.3 (2007), pp. 169–176 (cit. on p. 10).

[Kappeler, 2016] Armin Kappeler, Seunghwan Yoo, Qiqin Dai, and Aggelos K Katsaggelos. “Video super-resolution with convolutional neural networks”. In: *IEEE Transactions on Computational Imaging* 2.2 (2016), pp. 109–122 (cit. on p. 63).

[Kim, 2012] Minjeong Kim, Guorong Wu, Pew-Thian Yap, and Dinggang Shen. “A general fast registration framework by learning deformation–appearance correlation”. In: *IEEE Transactions on Image Processing* 21.4 (2012), pp. 1823–1833 (cit. on p. 13).

[Kingma, 2013] Diederik P Kingma and Max Welling. “Auto-encoding variational bayes”. In: *arXiv preprint arXiv:1312.6114* (2013) (cit. on pp. 36, 37, 39–42, 64, 65, 68).

[Kingma, 2014a] Diederik P Kingma and Jimmy Ba. “Adam: A method for stochastic optimization”. In: *arXiv preprint arXiv:1412.6980* (2014) (cit. on pp. 64, 73, 90, 98).

[Kingma, 2014b] Diederik P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. “Semi-supervised learning with deep generative models”. In: *Advances in Neural Information Processing Systems (NIPS)*. 2014, pp. 3581–3589 (cit. on pp. 36, 37, 40, 41, 50, 63, 64, 68, 69, 86, 87, 95, 96).

[Kingma, 2019] Diederik P Kingma, Max Welling, et al. “An introduction to variational autoencoders”. In: *Foundations and Trends® in Machine Learning* 12.4 (2019), pp. 307–392 (cit. on pp. 68, 103).

[Klein, 2010] Stefan Klein, Marius Staring, et al. “Elastix: a toolbox for intensity-based medical image registration”. In: *IEEE transactions on medical imaging* 29.1 (2010), pp. 196–205 (cit. on pp. 26–28).

[Krebs, 2017] Julian Krebs, Tommaso Mansi, Hervé Delingette, et al. “Robust non-rigid registration through agent-based action learning”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2017, pp. 344–352 (cit. on pp. 4, 6, 16, 17, 21, 34).

[Krebs, 2018] Julian Krebs, Tommaso Mansi, Boris Mailhé, Nicholas Ayache, and Hervé Delingette. “Unsupervised Probabilistic Deformation Modeling for Robust Diffeomorphic Registration”. In: *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer, 2018, pp. 101–109 (cit. on pp. 5, 6, 18, 33, 37, 38, 45, 49).

[Krebs, 2019a] Julian Krebs, Hervé Delingette, Nicholas Ayache, Tommaso Mansi, and Shun Miao. *Medical Imaging Diffeomorphic Registration based on Machine Learning*. US Patent App. 16/233,174. July 2019 (cit. on p. 7).

[Krebs, 2019b] Julian Krebs, Hervé Delingette, Boris Mailhé, Nicholas Ayache, and Tommaso Mansi. "Learning a Probabilistic Model for Diffeomorphic Registration". In: *IEEE Transactions on Medical Imaging* 38.9 (Sept. 2019), pp. 2165–2176 (cit. on pp. 5, 6, 18, 33, 62, 63, 68–70, 72–74, 87, 94, 96).

[Krebs, 2020a] Julian Krebs and Tommaso Mansi. *Method and System for Deep Motion Model Learning in Medical Images*. US Patent App. 16/131,465. Mar. 2020 (cit. on p. 7).

[Krebs, 2020b] Julian Krebs, Tommaso Mansi, Nicholas Ayache, and Hervé Delingette. "Learning a Generative Motion Model from Image Sequences based on a Latent Motion Matrix". In: *Submitted to IEEE Transactions on Medical Imaging* (Mar. 2020) (cit. on pp. 5, 6, 61).

[Krebs, 2020c] Julian Krebs, Tommaso Mansi, Nicholas Ayache, and Hervé Delingette. "Probabilistic Motion Modeling from Medical Image Sequences: Application to Cardiac Cine-MRI". In: *Statistical Atlases and Computational Models of the Heart. Multi-Sequence CMR Segmentation, CRT-EPiggy and LV Full Quantification Challenges*. Cham: Springer, Jan. 2020, pp. 176–185 (cit. on pp. 5, 6, 64, 73, 74, 79, 84–87, 89, 94).

[Krupinski, 2010] Elizabeth A Krupinski. "Current perspectives in medical image perception". In: *Attention, Perception, & Psychophysics* 72.5 (2010), pp. 1205–1217 (cit. on pp. 1, 2).

[Kurihara, 2009] Kenichi Kurihara et al. "Bayesian *k*-Means as a "Maximization-Expectation" Algorithm". In: *Neural Computation* 21.4 (2009), pp. 1145–1172 (cit. on p. 44).

[LedesmaCarbayo, 2005] Maria J Ledesma-Carbayo, Jan Kybic, Manuel Desco, et al. "Spatio-temporal nonrigid registration for ultrasound cardiac motion estimation". In: *IEEE transactions on medical imaging* 24.9 (2005), pp. 1113–1126 (cit. on pp. 14, 63).

[Lee, 2015] Chen-Yu Lee, Saining Xie, Patrick Gallagher, Zhengyou Zhang, and Zhuowen Tu. "Deeply-supervised nets". In: *Artificial Intelligence and Statistics*. 2015, pp. 562–570 (cit. on p. 38).

[Lee, 2019] Matthew CH Lee, Ozan Oktay, Andreas Schuh, Michiel Schaap, and Ben Glocker. "Image-and-Spatial Transformer Networks for Structure-Guided Image Registration". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2019, pp. 337–345 (cit. on p. 19).

[Li, 2018] Hongming Li and Yong Fan. "Non-rigid image registration using self-supervised fully convolutional networks without training data". In: *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE. 2018, pp. 1075–1078 (cit. on p. 18).

[Li, 2019] Bo Li, Wiro J Niessen, Stefan Klein, et al. "A hybrid deep learning framework for integrated segmentation and registration: evaluation on longitudinal white matter tract changes". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2019, pp. 645–653 (cit. on p. 19).

[Liang, 2017] Xiaodan Liang, Lisa Lee, Wei Dai, et al. “Dual Motion GAN for Future-Flow Embedded Video Prediction”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 1744–1752 (cit. on pp. 18, 35).

[Liao, 2017a] Rui Liao, Shun Miao, Pierre De Tournemire, et al. *Method and System for Image Registration Using an Intelligent Artificial Agent*. US Patent App. 15/587,094. Nov. 2017 (cit. on p. 7).

[Liao, 2017b] Rui Liao, Shun Miao, Pierre de Tournemire, et al. “An Artificial Agent for Robust Image Registration”. In: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)*. 2017 (cit. on pp. 16, 22, 23, 25, 29).

[Litjens, 2017] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, et al. “A survey on deep learning in medical image analysis”. In: *Medical image analysis* 42 (2017), pp. 60–88 (cit. on p. 107).

[Lorenzi, 2011] Marco Lorenzi, Nicholas Ayache, and Xavier Pennec. “Schild’s ladder for the parallel transport of deformations in time series of images”. In: *Biennial International Conference on Information Processing in Medical Imaging*. Springer. 2011, pp. 463–474 (cit. on p. 36).

[Lorenzi, 2013] Marco Lorenzi, Nicholas Ayache, Giovanni B Frisoni, et al. “LCC-Demons: a robust and accurate symmetric diffeomorphic registration algorithm”. In: *NeuroImage* 81 (2013), pp. 470–483 (cit. on pp. 11, 13, 28, 34, 38, 40, 47, 52, 54, 62).

[Lorenzi, 2014] Marco Lorenzi and Xavier Pennec. “Efficient parallel transport of deformations in time series of images: from Schild’s to pole ladder”. In: *Journal of mathematical imaging and vision* 50.1-2 (2014), pp. 5–17 (cit. on pp. 34–36, 38, 52).

[Lou, 2019] Bin Lou, Semihcan Doken, Tingliang Zhuang, et al. “An image-based deep learning framework for individualising radiotherapy dose: a retrospective analysis of outcome prediction”. In: *The Lancet Digital Health* 1.3 (2019), e136–e147 (cit. on pp. 84, 88).

[Louis, 2017] Maxime Louis, Alexandre Bône, Benjamin Charlier, Stanley Durrleman, Alzheimer’s Disease Neuroimaging Initiative, et al. “Parallel transport in shape analysis: a scalable numerical scheme”. In: *International Conference on Geometric Science of Information*. Springer. 2017, pp. 29–37 (cit. on p. 36).

[Ma, 2017] Kai Ma, Jiangping Wang, Vivek Singh, et al. “Multimodal image registration with deep context reinforcement learning”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2017, pp. 240–248 (cit. on p. 16).

[Maes, 1997] Frederik Maes, Andre Collignon, Dirk Vandermeulen, Guy Marchal, and Paul Suetens. “Multimodality image registration by maximization of mutual information”. In: *IEEE transactions on Medical Imaging* 16.2 (1997), pp. 187–198 (cit. on p. 11).

[Mahapatra, 2018] Dwarikanath Mahapatra, Bhavna Antony, Suman Sedai, and Rahil Garnavi. “Deformable medical image registration using generative adversarial networks”. In: *Biomedical Imaging (ISBI 2018), 2018 IEEE 15th International Symposium on*. IEEE. 2018, pp. 1449–1453 (cit. on pp. 16, 18, 34, 37).

[Maier, 2018] Andreas Maier, Stefan Steidl, Vincent Christlein, and Joachim Hornegger. *Medical Imaging Systems: An Introductory Guide*. Vol. 11111. Springer, 2018 (cit. on p. 2).

[Makhzani, 2016] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, and Ian Goodfellow. “Adversarial Autoencoders”. In: *International Conference on Learning Representations*. 2016 (cit. on p. 36).

[Mansi, 2011] Tommaso Mansi, Xavier Pennec, Maxime Sermesant, Hervé Delingette, and Nicholas Ayache. “iLogDemons: A demons-based registration algorithm for tracking incompressible elastic biological tissues”. In: *International journal of computer vision* 92.1 (2011), pp. 92–111 (cit. on p. 14).

[Marks, 2013] Leonard Marks, Shelena Young, and Shyam Natarajan. “MRI–ultrasound fusion for guidance of targeted prostate biopsy”. In: *Current opinion in urology* 23.1 (2013), p. 43 (cit. on p. 2).

[Metz, 2011] CT Metz, Stefan Klein, Michiel Schaap, Theo van Walsum, and Wiro J Niessen. “Nonrigid registration of dynamic medical imaging data using nD+ t B-splines and a groupwise optimization approach”. In: *Medical image analysis* 15.2 (2011), pp. 238–249 (cit. on pp. 14, 63, 73, 74, 90).

[Miao, 2018] Shun Miao, Sebastien Piat, Peter Fischer, et al. “Dilated fcn for multi-agent 2d/3d medical image registration”. In: *Thirty-Second AAAI Conference on Artificial Intelligence*. 2018 (cit. on p. 17).

[Mnih, 2015] V. Mnih, K. Kavukcuoglu, D. Silver, and others. “Human-level control through deep reinforcement learning”. In: *Nature* 518.7540 (2015), pp. 529–533 (cit. on pp. 22, 23).

[Modersitzki, 2004] J. Modersitzki. *Numerical methods for image registration*. Oxford University Press on Demand, 2004 (cit. on p. 9).

[Mohamed, 2002] Ashraf Mohamed, Christos Davatzikos, and Russell Taylor. “A combined statistical and biomechanical model for estimation of intra-operative prostate deformation”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2002, pp. 452–460 (cit. on p. 14).

[Myerburg, 2009] Robert J Myerburg, Vivek Reddy, and Agustin Castellanos. “Indications for implantable cardioverter-defibrillators based on evidence and judgment”. In: *Journal of the American College of Cardiology* 54.9 (2009), pp. 747–763 (cit. on p. 84).

[Nader, 2020] Clément Abi Nader, Nicholas Ayache, Philippe Robert, Marco Lorenzi, Alzheimer's Disease Neuroimaging Initiative, et al. "Monotonic Gaussian Process for spatio-temporal disease progression modeling in brain imaging data". In: *NeuroImage* 205 (2020), p. 116266 (cit. on p. 105).

[Nielsen, 1997] Mads Nielsen, Luc Florack, and Rachid Deriche. "Regularization, scale-space, and edge detection filters". In: *Journal of Mathematical Imaging and Vision* 7.4 (1997), pp. 291–307 (cit. on p. 44).

[Niethammer, 2019] Marc Niethammer, Roland Kwitt, and Francois-Xavier Vialard. "Metric learning for image registration". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019, pp. 8463–8472 (cit. on p. 18).

[Ohnishi, 2001] Takashi Ohnishi, Hiroshi Matsuda, Takeshi Tabira, Takashi Asada, and Masatake Uno. "Changes in brain morphology in Alzheimer disease and normal aging: is Alzheimer disease an exaggerated aging process?" In: *American Journal of Neuroradiology* 22.9 (2001), pp. 1680–1685 (cit. on p. 105).

[Oliveira, 2014] Francisco PM Oliveira and Joao Manuel RS Tavares. "Medical image registration: a review". In: *Computer methods in biomechanics and biomedical engineering* 17.2 (2014), pp. 73–93 (cit. on p. 9).

[Ouyang, 2020] David Ouyang, Bryan He, Amirata Ghorbani, et al. "Video-based AI for beat-to-beat assessment of cardiac function". In: *Nature* 580.7802 (2020), pp. 252–256 (cit. on p. 102).

[Ozhasoglu, 2002] Cihat Ozhasoglu and Martin J Murphy. "Issues in respiratory motion compensation during external-beam radiotherapy". In: *International Journal of Radiation Oncology* Biology* Physics* 52.5 (2002), pp. 1389–1399 (cit. on p. 106).

[Pennec, 2005] Xavier Pennec, Radu Stefanescu, Vincent Arsigny, Pierre Fillard, and Nicholas Ayache. "Riemannian elasticity: A statistical regularization framework for non-linear registration". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2005, pp. 943–950 (cit. on p. 12).

[Perperidis, 2005] Dimitrios Perperidis, Raad H Mohiaddin, and Daniel Rueckert. "Spatio-temporal free-form registration of cardiac MR image sequences". In: *Medical image analysis* 9.5 (2005), pp. 441–456 (cit. on p. 14).

[Peyrat, 2010] Jean-Marc Peyrat, Hervé Delingette, Maxime Sermesant, Chenyang Xu, and Nicholas Ayache. "Registration of 4D cardiac CT sequences under trajectory constraints with multichannel diffeomorphic demons". In: *IEEE Transactions on Medical Imaging* 29.7 (2010), pp. 1351–1368 (cit. on pp. 14, 62).

[Puech, 2013] Philippe Puech, Olivier Rouvière, Raphaele Renard-Penna, et al. "Prostate cancer diagnosis: multiparametric MR-targeted biopsy with cognitive and transrectal US–MR fusion guidance versus systematic biopsy—prospective multicenter study". In: *Radiology* 268.2 (2013), pp. 461–469 (cit. on p. 2).

[Punnoose, 2011] Lynn R Punnoose, Michael M Givertz, Eldrin F Lewis, et al. “Heart failure with recovered ejection fraction: a distinct clinical entity”. In: *Journal of cardiac failure* 17.7 (2011), pp. 527–532 (cit. on p. 84).

[Qin, 2018] Chen Qin, Wenjia Bai, Jo Schlemper, et al. “Joint learning of motion estimation and segmentation for cardiac MR image sequences”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2018, pp. 472–480 (cit. on pp. 19, 63).

[Qiu, 2011] Anqi Qiu, Laurent Younes, and Michael I Miller. “Principal component based diffeomorphic surface mapping”. In: *IEEE transactions on medical imaging* 31.2 (2011), pp. 302–311 (cit. on p. 63).

[Qiu, 2012] Anqi Qiu, Laurent Younes, and Michael I Miller. “Principal component based diffeomorphic surface mapping”. In: *IEEE transactions on medical imaging* 31.2 (2012), pp. 302–311 (cit. on p. 35).

[Rasmussen, 2003] Carl Edward Rasmussen. “Gaussian processes in machine learning”. In: *Summer School on Machine Learning*. Springer. 2003, pp. 63–71 (cit. on p. 67).

[Reyes, 2007] M Reyes, G Malandain, PM Koulibaly, Miguel Angel González-Ballester, and J Darcourt. “Model-based respiratory motion compensation for emission tomography image reconstruction”. In: *Physics in Medicine & Biology* 52.12 (2007), p. 3579 (cit. on p. 106).

[Rijnierse, 2017] Mischa T Rijnierse, Mehran Kamali Sadeghian, Sophie Schuurmans Stekhoven, et al. “Usefulness of left atrial emptying fraction to predict ventricular arrhythmias in patients with implantable cardioverter defibrillators”. In: *The American journal of cardiology* 120.2 (2017), pp. 243–250 (cit. on pp. 84, 89).

[Rohé, 2017] Marc-Michel Rohé, Manasi Datar, Tobias Heimann, Maxime Sermesant, and Xavier Pennec. “SVF-Net: Learning Deformable Image Registration Using Shape Matching”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2017, pp. 266–274 (cit. on pp. 16, 34, 62).

[Rohé, 2018] Marc-Michel Rohé, Maxime Sermesant, and Xavier Pennec. “Low-dimensional representation of cardiac motion using Barycentric Subspaces: A new group-wise paradigm for estimation, analysis, and reconstruction”. In: *Medical image analysis* 45 (2018), pp. 1–12 (cit. on pp. 34, 35, 62, 63).

[Ronneberger, 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. “U-net: Convolutional networks for biomedical image segmentation”. In: *International Conference on Medical image computing and computer-assisted intervention*. Springer. 2015, pp. 234–241 (cit. on p. 54).

[Rosen, 2003] Allyson C Rosen, Matthew W Prull, John DE Gabrieli, et al. “Differential associations between entorhinal and hippocampal volumes and memory performance in older adults.” In: *Behavioral neuroscience* 117.6 (2003), p. 1150 (cit. on p. 105).

[Rueckert, 1999] Daniel Rueckert, Luke I Sonoda, Carmel Hayes, et al. “Nonrigid registration using free-form deformations: application to breast MR images”. In: *IEEE transactions on medical imaging* 18.8 (1999), pp. 712–721 (cit. on p. 12).

[Rueckert, 2003] Daniel Rueckert, Alejandro F Frangi, and Julia A Schnabel. “Automatic construction of 3-D statistical deformation models of the brain using nonrigid registration”. In: *IEEE transactions on medical imaging* 22.8 (2003), pp. 1014–1025 (cit. on pp. 13, 24, 26).

[Sabbag, 2015] Avi Sabbag, Mahmoud Suleiman, Avishag Laish-Farkash, et al. “Contemporary rates of appropriate shock therapy in patients who receive implantable device therapy in a real-world setting: From the Israeli ICD Registry”. In: *Heart Rhythm* 12.12 (2015), pp. 2426–2433 (cit. on p. 84).

[Sandkühler, 2019] Robin Sandkühler, Simon Andermatt, Grzegorz Bauman, et al. “Recurrent Registration Neural Networks for Deformable Image Registration”. In: *Advances in Neural Information Processing Systems*. 2019, pp. 8755–8765 (cit. on p. 18).

[Schnabel, 2001] Julia A Schnabel, Daniel Rueckert, Marcel Quist, et al. “A generic framework for non-rigid registration based on non-uniform multi-level free-form deformations”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2001, pp. 573–581 (cit. on p. 12).

[Sermesant, 2008] Maxime Sermesant, Jean-Marc Peyrat, Phani Chinchapatnam, et al. “Toward patient-specific myocardial models of the heart”. In: *Heart failure clinics* 4.3 (2008), pp. 289–301 (cit. on pp. 14, 63).

[Shi, 2013] Wenzhe Shi, Martin Jantsch, Paul Aljabar, et al. “Temporal sparse free-form deformations”. In: *Medical image analysis* 17.7 (2013), pp. 779–789 (cit. on p. 63).

[Simonovsky, 2016] Martin Simonovsky, Benjamin Gutiérrez-Becker, Diana Mateus, Nassir Navab, and Nikos Komodakis. “A Deep Metric for Multimodal Registration”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2016, pp. 10–18 (cit. on pp. 16, 21, 22).

[Simonyan, 2013] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. “Deep inside convolutional networks: Visualising image classification models and saliency maps”. In: *arXiv preprint arXiv:1312.6034* (2013) (cit. on p. 104).

[Sivera, 2019] Raphaël Sivera, Hervé Delingette, Marco Lorenzi, et al. “A model of brain morphological changes related to aging and Alzheimer’s disease from cross-sectional assessments”. In: *NeuroImage* 198 (2019), pp. 255–270 (cit. on p. 105).

[Sokooti, 2017] Hessam Sokooti, Bob de Vos, et al. “Nonrigid image registration using multi-scale 3D convolutional neural networks”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2017, pp. 232–239 (cit. on pp. 16, 34).

[Sotiras, 2013] A. Sotiras, C. Davatzikos, and N. Paragios. “Deformable Medical Image Registration: A Survey”. In: *IEEE Transactions on Medical Imaging* 32.7 (2013), pp. 1153–1190 (cit. on pp. 9–13, 33, 34, 38, 62).

[Su, 2016] Xiaogang Su, Chalani S Wijayasinghe, Juanjuan Fan, and Ying Zhang. “Sparse estimation of Cox proportional hazards models via approximated information criteria”. In: *Biometrics* 72.3 (2016), pp. 751–759 (cit. on p. 93).

[Tang, 2009] Songyuan Tang, Yong Fan, Guorong Wu, Minjeong Kim, and Dinggang Shen. “RABBIT: rapid alignment of brains by building intermediate templates”. In: *NeuroImage* 47.4 (2009), pp. 1277–1287 (cit. on p. 13).

[Tanner, 2018] Christine Tanner, Firat Ozdemir, Romy Profanter, et al. “Generative Adversarial Networks for MR-CT Deformable Image Registration”. In: *arXiv preprint arXiv:1807.07349* (2018) (cit. on pp. 18, 35, 37).

[Thirion, 1998] J-P Thirion. “Image matching as a diffusion process: an analogy with Maxwell’s demons”. In: *Medical image analysis* 2.3 (1998), pp. 243–260 (cit. on pp. 12, 34).

[Tian, 2015] Zhiqiang Tian, LiZhi Liu, and Baowei Fei. “A fully automatic multi-atlas based segmentation method for prostate MR images”. In: *SPIE Medical Imaging*. 2015, pp. 941340–941340 (cit. on pp. 21, 26).

[Tracy, 2013] Cynthia M Tracy, Andrew E Epstein, Dawood Darbar, et al. “2012 ACCF/AHA/HRS focused update incorporated into the ACCF/AHA/HRS 2008 guidelines for device-based therapy of cardiac rhythm abnormalities: a report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines and the Heart Rhythm Society”. In: *Journal of the American College of Cardiology* 61.3 (2013), e6–e75 (cit. on p. 84).

[Uzunova, 2017] Hristina Uzunova, Matthias Wilms, Heinz Handels, and Jan Ehrhardt. “Training CNNs for Image Registration from Few Samples with Model-based Data Augmentation”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2017, pp. 223–231 (cit. on pp. 16, 34).

[Vaillant, 2004] Marc Vaillant, Michael I Miller, Laurent Younes, and Alain Trouvé. “Statistics on diffeomorphisms via tangent space representations”. In: *NeuroImage* 23 (2004), S161–S169 (cit. on p. 35).

[Vandemeulebroucke, 2011] Jef Vandemeulebroucke, Simon Rit, Jan Kybic, Patrick Clarysse, and David Sarrut. “Spatiotemporal motion estimation for respiratory-correlated imaging of the lungs”. In: *Medical physics* 38.1 (2011), pp. 166–178 (cit. on pp. 14, 63).

[Vercauteren, 2007a] Tom Vercauteren, Xavier Pennec, Ezio Malis, Aymeric Perchant, and Nicholas Ayache. “Insight into efficient image registration techniques and the demons algorithm”. In: *Biennial International Conference on Information Processing in Medical Imaging*. Springer. 2007, pp. 495–506 (cit. on p. 12).

[Vercauteren, 2007b] Tom Vercauteren, Xavier Pennec, Aymeric Perchant, and Nicholas Ayache. “Non-parametric diffeomorphic image registration with the demons algorithm”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2007, pp. 319–326 (cit. on p. 34).

[Vercauteren, 2008] Tom Vercauteren, Xavier Pennec, Aymeric Perchant, and Nicholas Ayache. “Symmetric log-domain diffeomorphic registration: A demons-based approach”. In: *International conference on medical image computing and computer-assisted intervention*. Springer. 2008, pp. 754–761 (cit. on pp. 13, 34, 62).

[Vercauteren, 2009] Tom Vercauteren, Xavier Pennec, Aymeric Perchant, and Nicholas Ayache. “Diffeomorphic demons: Efficient non-parametric image registration”. In: *NeuroImage* 45.1 (2009), S61–S72 (cit. on pp. 13, 62).

[Viola, 1997] Paul Viola and William M Wells III. “Alignment by maximization of mutual information”. In: *International journal of computer vision* 24.2 (1997), pp. 137–154 (cit. on p. 11).

[Vos, 2017] Bob D de Vos, Floris F Berendsen, Max A Viergever, Marius Staring, and Ivana Išgum. “End-to-end unsupervised deformable image registration with a convolutional neural network”. In: *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer, 2017, pp. 204–212 (cit. on pp. 18, 35, 38).

[Vos, 2019] Bob D de Vos, Floris F Berendsen, Max A Viergever, et al. “A deep learning framework for unsupervised affine and deformable image registration”. In: *Medical image analysis* 52 (2019), pp. 128–143 (cit. on p. 18).

[Wang, 2007] Jianzhe Wang and Tianzi Jiang. “Nonrigid registration of brain MRI using NURBS”. In: *Pattern Recognition Letters* 28.2 (2007), pp. 214–223 (cit. on p. 12).

[Wang, 2018] Jian Wang, William M Wells, Polina Golland, and Miaomiao Zhang. “Efficient Laplace Approximation for Bayesian Registration Uncertainty Quantification”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2018, pp. 880–888 (cit. on p. 36).

[Wassermann, 2014] Demian Wassermann, Matthew Toews, Marc Niethammer, and William Wells. “Probabilistic diffeomorphic registration: Representing uncertainty”. In: *International Workshop on Biomedical Image Registration*. Springer. 2014, pp. 72–82 (cit. on p. 36).

[Webb, 2003] Andrew Webb and George C Kagadis. “Introduction to biomedical imaging”. In: *Medical Physics* 30.8 (2003), pp. 2267–2267 (cit. on p. 2).

[Wei, 2018] Wen Wei, Emilie Poirion, Benedetta Bodini, et al. “Learning Myelin Content in Multiple Sclerosis from Multimodal MRI through Adversarial Training”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2018, pp. 514–522 (cit. on p. 36).

[Weinzaepfel, 2013] P. Weinzaepfel, J. Revaud, Z. Harchaoui, and C. Schmid. “Deepflow: Large displacement optical flow with deep matching”. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2013, pp. 1385–1392 (cit. on p. 15).

[Wright, 2018] Robert Wright, Bishesh Khanal, Alberto Gomez, et al. “LSTM spatial co-transformer networks for registration of 3D fetal US and MR brain images”. In: *Data Driven Treatment Response Assessment and Preterm, Perinatal, and Paediatric Image Analysis*. Springer, 2018, pp. 149–159 (cit. on p. 16).

[Wu, 2015] G. Wu, M. J. Kim, Q. Wang, B. Munsell, and D. Shen. “Scalable High Performance Image Registration Framework by Unsupervised Deep Feature Representations Learning”. In: *IEEE Transactions on Biomedical Engineering* (2015) (cit. on p. 16).

[Yang, 2011a] Lin Yang, Bogdan Georgescu, Yefeng Zheng, et al. “Prediction based collaborative trackers (PCT): A robust and accurate approach toward 3D medical object tracking”. In: *IEEE transactions on medical imaging* 30.11 (2011), pp. 1921–1932 (cit. on p. 63).

[Yang, 2011b] Xuan Yang, Zhong Xue, Xia Liu, and Darong Xiong. “Topology preservation evaluation of compact-support radial basis functions for image registration”. In: *Pattern Recognition Letters* 32.8 (2011), pp. 1162–1177 (cit. on p. 12).

[Yang, 2016] Xiao Yang, Roland Kwitt, and Marc Niethammer. “Fast Predictive Image Registration”. In: *International Workshop on Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*. Springer. 2016, pp. 48–57 (cit. on p. 22).

[Yang, 2017] Xiao Yang, Roland Kwitt, Martin Styner, and Marc Niethammer. “Quicksilver: Fast predictive image registration–A deep learning approach”. In: *NeuroImage* 158 (2017), pp. 378–396 (cit. on pp. 16, 34, 37, 62).

[Yoo, 2017] Inwan Yoo, David GC Hildebrand, Willie F Tobin, Wei-Chung Allen Lee, and Won-Ki Jeong. “ssemnet: Serial-section electron microscopy image registration using a spatial transformer network with learned features”. In: *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer, 2017, pp. 249–257 (cit. on p. 18).

[Younes, 2007] Laurent Younes. “Jacobi fields in groups of diffeomorphisms and applications”. In: *Quarterly of applied mathematics* (2007), pp. 113–134 (cit. on p. 36).

[Zerhouni, 1988] Elias A Zerhouni, David M Parish, Walter J Rogers, Andrew Yang, and Edward P Shapiro. “Human heart: tagging with MR imaging–a method for noninvasive assessment of myocardial motion.” In: *Radiology* 169.1 (1988), pp. 59–63 (cit. on p. 3).

[Zhang, 2014] Miaomiao Zhang and P Thomas Fletcher. “Bayesian principal geodesic analysis in diffeomorphic image registration”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2014, pp. 121–128 (cit. on pp. 35, 37).

[Zhang, 2015] Miaomiao Zhang and P Thomas Fletcher. “Finite-dimensional Lie algebras for fast diffeomorphic image registration”. In: *International Conference on Information Processing in Medical Imaging*. Springer. 2015, pp. 249–260 (cit. on pp. 13, 34, 62).

List of Figures

List of Tables