# Deep-Learning-Based Image Registration And Tumor Follow-Up Analysis

Alessa Hering

The research described in this thesis was carried out at Fraunhofer MEVIS (Lübeck, Germany) and the Diagnostic Image Analysis Group, Radboud University Medical Center (Nijmegen, The Netherlands).

This work was supported by the German Academic Scholarship Foundation.

This book was typeset by the author using LaTeX.

Cover design by Franz Buscha



ISBN: 978-94-6421-853-4

Printed by Ipskamp Printing, Enschede.

# Deep-Learning-based Image Registration and Tumor Follow-Up Analysis

Proefschrift ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus, prof. dr. J. H. J. M. van Krieken,
volgens besluit van het college voor promoties
in het openbaar te verdedigen op

maandag 10 oktober 2022
om 12.30 uur precies

door

Alessa Denise Hering
geboren op 19 december 1990
te Hamburg, Duitsland

| | |
|---|---|
| Promotoren: | Prof. dr. B. van Ginneken |
| | Prof. dr.-ing. H.K. Hahn (Universität Bremen, Duitsland) |
| Copromotoren: | Dr. N. Lessmann |
| | Dr. S. Heldmann (Fraunhofer MEVIS, Duitsland) |
| Manuscriptcommissie: | Prof. dr. A.C.C. Coolen |
| | Dr. H.A. Gietema (Maastricht UMC+) |
| | Prof. dr. J.A. Schnabel (Technische Universität München, Duitsland) |

# Contents

# 1

# CHAPTER 1

# General introduction

## 1.1 Medical image registration

Image registration is the process of aligning two or more images to achieve pointwise spatial correspondence [1]. It is sometimes also called *fusion*, *matching* or *warping*. This is a fundamental step for many tasks in medical image analysis as it links previously unrelated data and enables joint processing of those data. By aligning images from different modalities, complementary information can be fused or propagated from one modality to another. For example, morphological information from a CT image can be fused with functional information from a PET image [2, 3]. Furthermore, image registration can be used to track the progression of the disease overtime, such as it is done with MR images of patients with multiple sclerosis [4], for which an MRI scan of the brain is taken every few months. Another application of image registration is atlas-based segmentation, which aims to transfer label information from one or more atlases to a new image for which no labels are known [5]. Despite ever-improving segmentation methods [6], registration continues to be used for this purpose, especially for the creation of noisy labels for the training process of segmentation networks [7]. Image registration is also an important tool for many further applications and has been an active field of research for decades [8, 9].

This thesis focuses on the registration of exactly two images. The first image is referred to as *fixed image* $\mathcal{F}$. This image remains unmodified during the registration process. The second image is referred to as *moving image* $\mathcal{M}$ and is adapted to match the fixed image by applying a transformation $\phi$. The goal of image registration is to find a *reasonable* transformation $\phi$, such that the transformed moving image $\mathcal{M}(\phi)$ becomes *similar* to the fixed image $\mathcal{F}$. In the past, various approaches and tailored solutions have been proposed to a wide range of problems and applications. The requirements of similarity of the images and reasonability of the deformation field are often explicitly formulated in a cost function using a distance measure $\mathcal{D}$ and a regularizer $\mathcal{R}$ [1, 10–17]. All those approaches require to solve an optimization problem for each image pair, which is a complex and computationally demanding task: this often leads to long processing times. A lot of research has been done to speed-up this process, with more efficient algorithms or better implementation on CPU [18, 19] and GPU [20, 21].

A new approach to solve the optimization problem is to use neural networks. These replace the iterative optimization for each new image pair with a single forward pass through the network.

## 1.2 Deep-learning-based image registration

Artificial intelligence (AI) has received a lot of attention in recent years and has long been present in our everyday lives. Through pioneering achievements in machine learn-

ing, AI is gradually revolutionizing all areas of life; and medicine is no exception [22]. Deep neural networks enable completely autonomous processing of medical image data or serve as support for clinicians. While deep learning has become a methodology of choice in many areas like segmentation or classification, image registration is often still based on conventional methods. One main factor is the lack of ground-truth – needed to *train* the neural networks – which stems from the large variability of plausible deformations.

To better understand this problem, it is helpful to look at different image analysis tasks like segmentation or classification. For classification tasks, a medical expert typically inspects an image or parts of an image and then classifies it with one discrete label. Segmentation tasks require slightly more annotations: one discrete label per voxel is necessary to divide the image into different classes (e.g. organ vs. non-organ). However, during the annotation process, only the border between both classes needs to be drawn, which reduces the annotation effort immensely. In both cases, a medical expert can solve those tasks, albeit with a certain variability between different experts. In contrast, an image registration ground-truth needs to establish a 3D displacement vector for every point of the fixed image, resulting in three continuous labels per voxel. This displacement vector connects the fixed with the moving image and therefore, the solution of the registration problem lies neither in the fixed nor in the moving image, but between them. Due to the absence of dense ground-truth, the registration problem is much less specified than, for example, image classification or segmentation. As a consequence, in mathematical terms, image registration is a so-called ill-posed [23], which roughly means that there is no clear solution.

Nevertheless, several methods – including this thesis – presented in recent years aim to mimic the process of conventional image registration methods by learning a registration function in the form of a convolutional neural network, that predict spatial deformations warping a moving image to a fixed image. Hereby, the methods replace the costly iterative optimization of conventional registration methods for each image pair with one optimization during the training of the convolutional neural network.

As there are no ground-truth deformation fields annotated by a medical expert available to train the registration network, previous works have presented different approaches to mitigate this issue. They can be classified as *supervised* [24–27], *unsupervised* [28–31], and *weakly-supervised* [32–35] registration approaches.

*Supervised* methods use ground-truth deformation fields for training. These fields can either be randomly generated or produced by classic image registration methods. The main limitation of these approaches is that their accuracy is bounded by the performance of existing algorithms or simulations.

In contrast, *unsupervised* methods do not require any ground-truth data. The learning process is driven by image similarity measures or – more generally – by evaluating the cost function of classic variational image registration methods. An important milestone for the development of these methods was the introduction of the

spatial transformer networks [36] for differentiable warping of moving images during training.

*Weakly-supervised* methods also do not rely on ground-truth deformation fields but training is still supervised with prior information. The labels of the moving image are transformed by the deformation field and compared within the loss function with the fixed labels. All anatomical labels are only required during training.

In the last years, deep-Learning-based image registration has been a very active area of research and it has been shown to be equivalent or even superior to conventional approaches in many examples. Nevertheless, deep-learning-based image registration is still a new field and in the clinical settings, conventional image registration is the dominant technology for most applications. In this work, we will explore a primary application from oncology where image registration is an enabling technology: tumor follow-up assessment.

## 1.3 Application: Efficient Tumor Follow-Up Analysis

Cancer is the second leading cause of death [37] and with an estimated number of 17.5 million new cases of cancer diagnosed in 2015 [38], it affects the lives of many people. For this reason, it is highly relevant to investigate approaches that support cancer diagnosis, treatment, and follow-up analysis. Medical images are taken at every stage of the diagnosis and treatment of cancer and have to be read by a radiologist. The radiologist locates, measures, and classifies suspicious lesions. In addition, changes must be assessed in comparison to previous images of the same patient. It must be checked for new lesions and other findings. However, due to the therapy, additional changes like reduction of weight might occur which complicates the comparison. For this purpose, the radiologist typically navigates manually through the slices of three-dimensional images to find corresponding lesions and then measures changes in lesion size and appearance [39]. Image registration can assist the radiologist by automatically providing the locations in one image that corresponds to a specific location in the other image. Just by taking over this tedious task through automation, the radiologist can use their time more efficiently. Furthermore, image registration can be used to highlight changes through the use of difference images or to propagate lesions measurements from previous images to the current image. These results can then be used by other automatic decision support methods to provide more accurate outputs.

## 1.4 Outline of the thesis

The first part of the thesis presents our contributions to the development of deep-learning-based image registration approaches. The first three chapters present the

algorithms we have developed. The fourth chapter compares the approaches of different groups within the Learn2Reg image registration challenge.

**CHAPTER 2** describes a method for a memory-efficient weakly-supervised deep-learning model for multi-modal image registration. The method combines three 2D networks into a 2.5D registration network.

**CHAPTER 3** presents a multilevel approach for deep learning-based image registration.

**CHAPTER 4** describes a method that incorporates multiple anatomical constraints as anatomical priors into the registration network applied on CT lung registration.

**CHAPTER 5** presents the results of the Learn2Reg challenge and compares several conventional and deep-learning-based registration methods.

The second part of the thesis presents steps towards efficient tumor follow-up analysis:

**CHAPTER 6** describes a pipeline that automates the segmentation and measurement of matching lesions, given a point annotation in the baseline lesion. The pipeline is based on a registration approach to locate corresponding image regions and a convolutional neural network to segment the lesion in the follow-up image.

**CHAPTER 7** presents the reader study, which investigates whether the assessment time for follow-up lesion segmentations is reduced by AI-assisted workflow while maintaining the same quality of segmentations.

Finally, **CHAPTER 8** summarizes the presented methods and findings and discusses the results as well as future directions.

2

# CHAPTER 2

# 2.5D Convolutional Transformer Networks for Multi-Modal Registration

**BASED ON:** A. Hering, S. Kuckertz, S. Heldmann, and M. P. Heinrich. "Memory-efficient 2.5 D convolutional transformer networks for multi-modal deformable registration with weak label supervision applied to whole-heart CT and MRI scans," *International journal of computer assisted radiology and surgery*, vol. 14 (2019), pp. 1901–1912.

## Abstract

**PURPOSE** Despite its potential for improvements through supervision, deep-learning-based registration approaches are difficult to train for large deformations in 3D scans due to excessive memory requirements.

**METHODS** We propose a new 2.5D convolutional transformer architecture that enables us to learn a memory-efficient weakly-supervised deep-learning model for multi-modal image registration. Furthermore, we firstly integrate a volume change control term into the loss function of a deep-learning-based registration method to penalize occurring foldings inside the deformation field.

**RESULTS** Our approach succeeds at learning large deformations across multi-modal images. We evaluate our approach on 100 pair-wise registrations of CT and MRI whole heart scans and demonstrate considerably higher Dice Scores (of 0.74 ) compared to a state-of-the-art unsupervised discrete registration framework (deeds with Dice of 0.71 ).

**CONCLUSION** Our proposed memory-efficient registration method performs better than state-of-the-art conventional registration methods. By using a volume change control term in the loss function, the number of occurring foldings can be considerably reduced on new registration cases.

## 2.1 Introduction

Image registration aims to align two or more images to achieve point-wise spatial correspondence. This is a fundamental step for many medical image analysis tasks and has been a very active field of research for decades [40]. Typically, image registration is phrased as an optimization problem w.r.t. a spatial mapping that minimizes a suitable cost function by applying iterative optimization schemes. However, this iterative optimization is time-consuming. Due to substantially increased computational power and availability of image data over the last years, learning-based image registration methods have emerged as an alternative to energy-optimization approaches [41, 42].

We present a non-iterative weakly-supervised deep-learning-based method for multi-modal deformable image registration. The nonlinear alignment of CT and MR is a particularly demanding type of registration. The complexity of this task is two-folded: First, a nonlinear deformation field has to be established between a pair of images to correct changes due to time, deformation and motion. Second, due to the different scanner modalities, the same anatomical structure has a different appearance in the images. We learn a registration function in the form of a convolutional neural network (CNN) to predict spatial deformations that warp a moving image to a fixed image. During training, the weights of the network are optimized with a loss function that reflects an established distance measure and regularizer of conventional image registration [43]. Additionally, we incorporate modality independent prior information in form of segmentation masks of anatomical structures into our loss function. This weakly-supervised approach successfully combines the strengths of prior information (segmentation labels) with an energy-based distance measure. We evaluate our approach on the difficult task of inter-patient CT-MR whole heart registration.

Together with [33] our previous work [44] was the first deep-learning-based image registration method that is not optimized only using label-based information (cf. [32]), but additionally includes an image intensity based distance measure. The presented method is the first to tackle multi-modal alignment with this dual objective. Due to different intensity representations of the same anatomical structures, classical distance measures like Sum-of-Squared-Differences cannot be used to define the similarity of images with different modalities. We therefore make use of an edge-based distance measure, the Normalized Gradient Field [45]. Moreover, we develop a network architecture which is adapted to the task of multi-modal registration by using two separate modality dependent convolution layers at the beginning of our network to extract low-level image features.

Even though computational power has increased dramatically over the past years, volumetric image registration is still a computationally demanding and memory consuming task. Therefore, we propose a memory-efficient 2.5D registration method which trains three independent 2D networks from orthogonal planes and combines the slice-wise results into one 3D deformation field. This provides the possibility of

either using larger networks with more trainable parameters or a larger batch size. Especially for small sized data sets, our 2.5D method can increase the variability of the mini-batches during training by randomly selecting slices from different patients and therefore helps to generalize better. Due to the fact that our 2.5D method only needs 680MB for a batch size of one, it is possible to train the network on a relatively small GPU.

### 2.1.1 Related Work

**DEEP-LEARNING-BASED IMAGE REGISTRATION** Compared to other fields relatively little research has yet been undertaken in deep-learning-based image registration [22] and most of this research has been published since 2017. These methods mostly aim to learn a function in form of a CNN that predicts a spatial deformation, which warps a so-called moving image to a fixed image. Based on how networks are trained, we categorize these approaches into *supervised* [42], *unsupervised* [41, 46, 47] and *weakly-supervised* [32, 33] methods.

*Supervised* methods use ground-truth deformation fields for training. These deformation fields can either be randomly generated or produced by classic image registration methods. The main limitation of these approaches is that their accuracy is limited by the performance of existing algorithms or simulations.

In contrast, *unsupervised* methods do not require any ground-truth data. The learning process is driven by image similarity measures or more general by evaluating the cost function of classic variational image registration methods. An important milestone for the development of these methods was the introduction of the spatial transformer networks [36] for differentiable warping of moving images during training.

*Weakly-supervised* methods also do not rely on ground-truth deformation fields but training is still supervised with prior information. The labels of the moving image are transformed by the deformation field and compared within the loss function with the fixed labels. All anatomical labels are only required during training.

**2.5D CONVOLUTIONAL NETWORKS** The idea of 2.5D methods is not new and has been used, e.g., in the context of segmentation of 3D organs. Here, two types of solutions for volumetric organ segmentation have been proposed. The first ones aim at training 3D networks directly. However, this is computationally expensive and according to [48] less stable in many cases. Therefore, the second group of methods train 2D networks from three orthogonal planes and combines the segmentation results [48–50]. A simple way to combine the results is applying a 3D isotropic Gaussian filtering to propagate the 2D slice-based probabilities to 3D by taking local 3D neighborhoods into account [49]. In [50], the output of the last layer of three otherwise independent CNNs are concatenated to obtain a joint output, which is fed into a softmax classifier. However, the memory-reduction and the ability to parallelize are not optimized because all networks

have to be trained together. Whereas in [48], three 2D segmentation networks on different viewpoints are optimized individually. Subsequently, they use the validation set to train a so-called Volumetric Fusion Net to combine the results of the 2D networks.

## 2.2 Methods

In [44], we have presented a 2D deep-learning-based registration approach for slice-wise 3D registration. This approach is not sufficient for full 3D registration because the existing deformations are mostly three-dimensional. However, training of a real 3D network needs a lot more memory. To tackle the issue of three-dimensional deformation without expanding the network architecture to 3D, we combine three 2D networks to a *2.5D registration network*. Therefore, we train three independent 2D registration networks on the axial, coronal and sagittal slices of the images. During inference, these networks are applied independently yielding three layered 3D deformation fields with one zero component. The final deformation field is created by averaging the respective non-zero components of the deformation fields. In the following, we describe our 2D deep-learning-based registration, which is a modified version of [44]. Subsequently, we formally describe the composition of the 3D deformation field.

### 2.2.1 Deep Deformable 2D Image Registration

Let $\mathcal{F}, \mathcal{M} : \mathbb{R}^2 \to \mathbb{R}$ denote the fixed image and moving image, respectively, and let $\Omega \subset \mathbb{R}^2$ be a domain modeling the field of view of $\mathcal{F}$. We aim to compute a deformation $\mathrm{y} : \Omega \to \mathbb{R}^2$ that aligns the fixed image $\mathcal{F}$ and the moving image $\mathcal{M}$ on the field of view $\Omega$ such that $\mathcal{F}(\mathrm{x})$ and $\mathcal{M}(\mathrm{y}(\mathrm{x}))$ are similar for $\mathrm{x} \in \Omega$. Inspired by recent unsupervised image registration methods (e.g. [41, 42]), we do not employ iterative optimization as in classic registration, but rather train a CNN that takes images $\mathcal{F}$ and $\mathcal{M}$ as input and yields the deformation y as output (cf. Figure 2.1). Thus, in the context of CNNs, we can consider y as a function of input images $\mathcal{F}, \mathcal{M}$ and trainable CNN model parameters $\theta$ to be learned, i.e. $\mathrm{y}(\mathrm{x}) \equiv \mathrm{y}(\theta; \mathcal{F}, \mathcal{M}, \mathrm{x})$. During training, the CNN parameters $\theta$ are optimized so that the deformation field y minimizes the loss function

$$\mathcal{L}(\mathcal{F}, \mathcal{M}, \mathrm{b}_{\mathcal{F}}, \mathrm{b}_{\mathcal{M}}, \mathrm{y}) = \mathcal{D}(\mathcal{F}, \mathcal{M}(\mathrm{y})) + \alpha\mathcal{R}(\mathrm{y}) + \beta\mathcal{B}(\mathrm{b}_{\mathcal{F}}, \mathrm{b}_{\mathcal{M}}(\mathrm{y})) + \gamma\mathcal{V}(\mathrm{y}) \quad (2.1)$$

with a distance measure $\mathcal{D}$ that quantifies the similarity of fixed image $\mathcal{F}$ and deformed moving image $\mathcal{M}(\mathrm{y})$, a regularizer $\mathcal{R}$ that forces smoothness of the deformation, a second distance measure $\mathcal{B}$ that quantifies the similarity of fixed segmentation $\mathrm{b}_{\mathcal{F}}$ and warped moving segmentation $\mathrm{b}_{\mathcal{M}}(\mathrm{y})$ and a volume change control term $\mathcal{V}$ to penalize foldings. The parameters $\alpha, \beta, \gamma \geq 0$ are weighting factors. Note that the segmentations are only used to evaluate the loss function and not used as network input and are therefore only required during training.

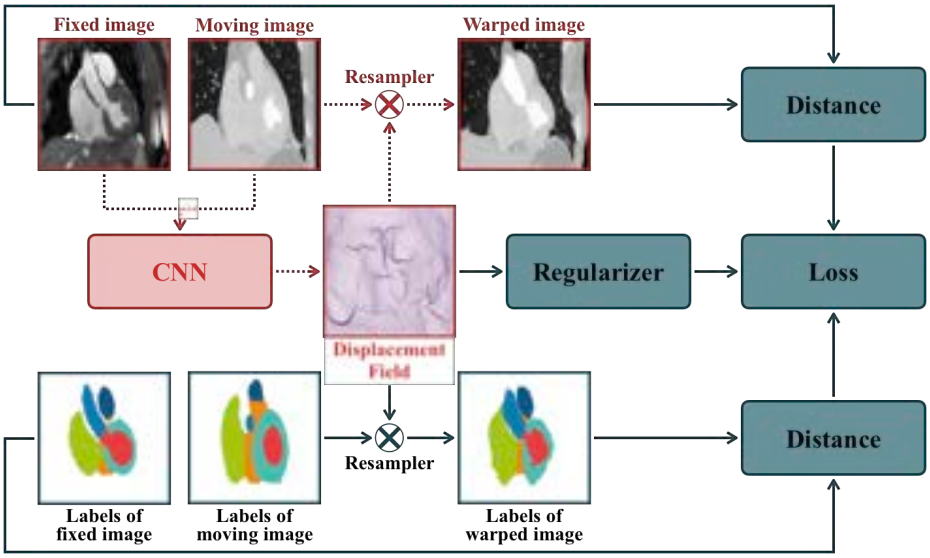



Figure 2.1: Illustration of the training process. For convenience, there is only one output deformation field shown instead of three. While application after the training only flows represented by red-dotted lines and red parts are required.

### 2.2.2 Loss Function

The challenge in multi-modal image registration is that corresponding structures have different appearance. That is, intensities of identical objects are different in images with different modalities. Consequently, classical intensity-based distance measures such as Sum-of-Squared-Differences cannot be used. However, this suggests the use of edge-based distance measures like Normalized Gradient Field (NGF)[45] distance measure

$$\mathcal{D}(\mathcal{F},\mathcal{M}(y)) = \frac{1}{2}\int_{\Omega} 1 - \frac{\langle \nabla\mathcal{M}(y(x)), \nabla\mathcal{F}(x)\rangle^2_{\varepsilon_1,\varepsilon_2}}{\|\nabla\mathcal{M}(y(x))\|^2_{\varepsilon_1}\|\nabla\mathcal{F}(x)\|^2_{\varepsilon_2}}\,dx,$$

with $\langle f,g\rangle_{\varepsilon_1,\varepsilon_2} := \sum_{j=1}^{2} f_j g_j + \varepsilon_1\varepsilon_2$, $\|f\|_{\varepsilon_i} := \sqrt{\langle f,f\rangle_{\varepsilon_i,\varepsilon_i}}$, $i = 1,2$ and so-called modality specific edge parameters $\varepsilon_1, \varepsilon_2 > 0$. Here we follow the work of [45, 51] which has shown that the NGF distance measure is reliable for multi-modal image registration. Furthermore, we use the second order curvature regularizer [52]

$$\mathcal{R}(y) = \frac{1}{2}\int_{\Omega}\sum_{j=1}^{2}\|\Delta y_j\|^2\,dx.$$

to obtain smooth deformation fields. However, it cannot be guaranteed that physically implausible deformations as large volume changes or even foldings happen. Conse-

quently, we extend our loss function with an additional so-called volume change control term $\mathcal{V}$ that measures the change of volume as induced by the transformation y

$$\mathcal{V}(y) = \int_{\Omega} \psi(\det \nabla y(x))dx,$$

with a weighting function

$$\psi(t) = \frac{(t-1)^2}{t} \quad \text{for } t > 0 \text{ and } \psi(t) = \infty \text{ else.}$$

Local volume shrinkage and expansion are symmetrically penalized, due to the symmetry of $\psi$. Additionally, if the Jacobian becomes negative at any point and therefore foldings occur, the volume change control term penalize it by setting the loss value to infinity. Note that since the loss function is evaluated on 2D images, the deformation field is 2D too and therefore only area changes can be measured. The similarity of the segmentation masks is measured using a sum of squared differences of the one-hot-representation of the segmentations

$$\mathcal{B}(y) = \frac{1}{2} \int_{\Omega} \|b_{\mathcal{M}}(y(x)) - b_{\mathcal{F}}(x))\|^2 dx.$$

### 2.2.3 Architecture and Training

In this section, we describe our particular architecture in our experiments, which is illustrated in Figure 2.2. Our network architecture basically follows the structure of a U-Net [53], taking a pair of fixed and moving images as input. In our experiments, the resolution size of the input images is 160 × 160 × 160, but the architecture is not limited or adapted to a particular size. The CNN generates a dense displacement vector field with the same grid resolution as the input images which is used to warp the moving to the fixed image.

In contrast to the standard U-Net architecture, we start with two separate processing streams for the moving and fixed image. In previous work for mono-modal registration, both streams use shared weights (cf. [44]). However, for our purpose of multi-modal registration, we adapt our network architecture by utilizing individual convolution weights for the first layers in order to learn modality specific features. We apply 2D convolution kernels in both the encoder and decoder stage using a kernel size of 3 on all levels. Each convolution is followed by a batch normalization and a ReLU layer. To reduce the spatial resolution of the feature maps in the encoder path, max pooling layers with a stride of 2 are used. Similar to the image pyramid used in conventional image registration, the successive layers of the encoder operate over coarser representations of the input. In the decoder path, we alternate between transposed convolutions, convolutions and concatenating skip connections.

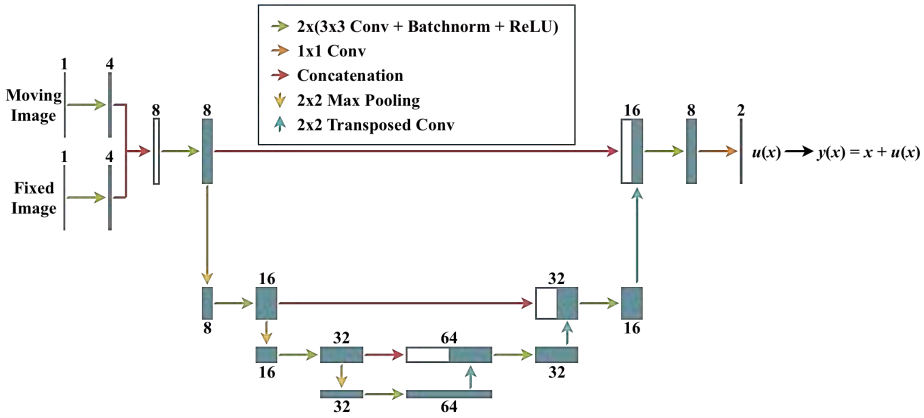



Figure 2.2: Proposed U-Net based architecture of our CNN. Each blue box represents a multi-channel feature map whose width corresponds to the number of channels which is denoted above or below the box.

### 2.2.4 From 2D to 3D Deformations

A main contribution of this work is in estimating 3D deformations from 2D networks. Given 3D images $\mathcal{M}, \mathcal{F} : \mathbb{R}^3 \to \mathbb{R}$, we use above 2D CNN model $y^{2D} : \mathbb{R}^2 \to \mathbb{R}^2$, $x \mapsto y^{2D}(x) \equiv y^{2D}(\theta; \mathcal{F}^{2D}, \mathcal{M}^{2D}, x)$ where $\mathcal{F}^{2D}, \mathcal{M}^{2D} : \mathbb{R}^2 \to \mathbb{R}$ are 2D input images for the computation of a 2D deformation. To this end, we train our model for the registration of 2D axial, coronal and sagittal slices yielding three different sets of parameters $\theta_1, \theta_2, \theta_3$. To be more precise, we define three deformations $y^\ell : \mathbb{R}^2 \to \mathbb{R}$, $\ell = 1, 2, 3$ by fixing either $x_1$, $x_2$ or $x_3$ coordinate of the 3D input images, i.e., we set

$$y^1(x_1, x_2, x_3) := \begin{pmatrix} 0 \\ y_1^{2D}(\theta_1; \mathcal{F}(x_1, \cdot, \cdot), \mathcal{M}(x_1, \cdot, \cdot), x_2, x_3) \\ y_2^{2D}(\theta_1; \mathcal{F}(x_1, \cdot, \cdot), \mathcal{M}(x_1, \cdot, \cdot), x_2, x_3) \end{pmatrix}$$

$$y^2(x_1, x_2, x_3) := \begin{pmatrix} y_1^{2D}(\theta_2; \mathcal{F}(\cdot, x_2, \cdot), \mathcal{M}(\cdot, x_2, \cdot), x_1, x_3) \\ 0 \\ y_2^{2D}(\theta_2; \mathcal{F}(\cdot, x_2, \cdot), \mathcal{M}(\cdot, x_2, \cdot), x_1, x_3) \end{pmatrix}$$

$$y^3(x_1, x_2, x_3) := \begin{pmatrix} y_1^{2D}(\theta_3; \mathcal{F}(\cdot, \cdot, x_3), \mathcal{M}(\cdot, \cdot, x_3), x_1, x_2) \\ y_2^{2D}(\theta_3; \mathcal{F}(\cdot, \cdot, x_3), \mathcal{M}(\cdot, \cdot, x_3), x_1, x_2) \\ 0 \end{pmatrix}.$$

The parameters $\theta_1, \theta_2, \theta_3$ are then computed by training with the 2D slices obtained from the 3D fixed and moving images. Finally, the spatial transformations for axial,

coronal and sagittal registration are averaged into a single 3D vector field

$$y^{2.5D}(x_1, x_2, x_3) := \frac{1}{2}\Big(y^1(x_1, x_2, x_3) + y^2(x_1, x_2, x_3) + y^3(x_1, x_2, x_3)\Big).$$

Note that only two 2D transformations contribute per dimension.

## 2.3 Materials

### 2.3.1 Dataset

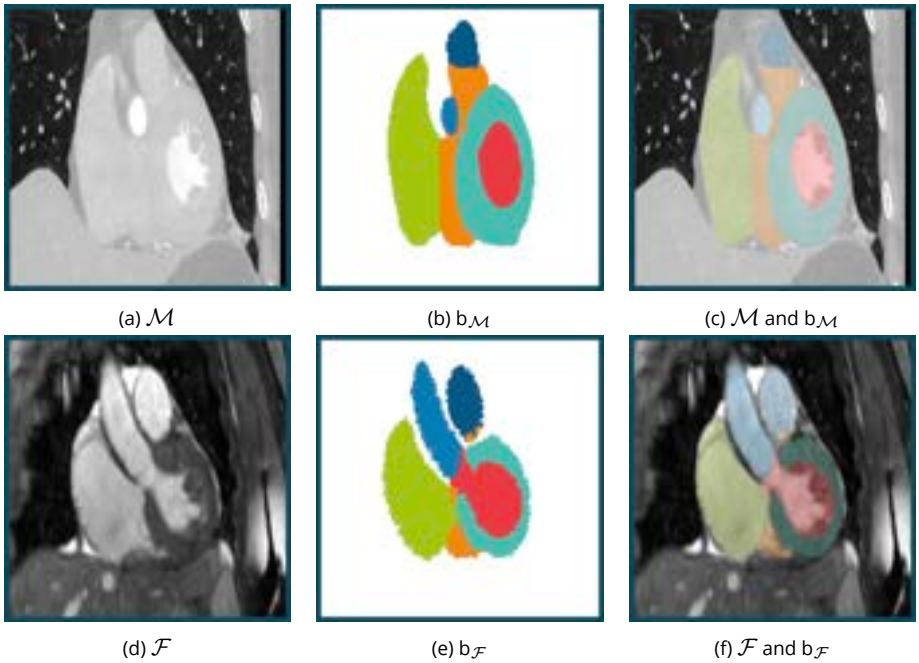
We perform our experiments on the Multi-Modality Whole Heart Segmentation (MM-WHS) dataset [54]. It contains 20 CT and 20 MR whole heart images. The CT data were acquired at Shanghai Shuguang Hospital, China, using routine cardiac CT angiography. The images cover the whole heart from the upper abdominal to the aortic arch with am inplane resolution of $0.78 \times 0.78$ and a average slice thickness of 1.6 mm. The MR images were acquired at St Thomas hospital and Royal Brompton Hospital, London, UK with about 2.0 mm acquisition resolution at each direction and reconstructed (resampled) into about 1.0 mm. The dataset includes segmentations for the following seven structures:

- left ventricle (LV)
- right ventricle (RV)
- left atrium (LA)
- right atrium (RA)
- myocardium (Myo)
- ascending aorta (AA)
- pulmonary artery (PA)

One exemplary scan pair and its segmentation masks are shown in Figure 2.3. Since the CT and MR images are not associated, each CT image could be registered to each MR image. However, evaluation is performed as a k-fold cross-validation with k = 4, leaving out five CT and MR images for testing. This results to 225 and 25 registration pairs during training and testing, respectively.

### 2.3.2 Preprocessing

In this work, we focus on nonlinear deformations, for that reason we perform a linear pre-alignment of fixed and moving image as preprocessing. We choose the CT images as moving image and the MR images as fixed image. The centers of gravity of each label are used as landmarks to solve an linear equation system to obtain a affine transformation matrix. We subsequently warp and resample the MR image on the



(a) $\mathcal{M}$ (b) $b_{\mathcal{M}}$ (c) $\mathcal{M}$ and $b_{\mathcal{M}}$

(d) $\mathcal{F}$ (e) $b_{\mathcal{F}}$ (f) $\mathcal{F}$ and $b_{\mathcal{F}}$

Figure 2.3: Exemplary CT (upper row) and MR (lower row) scans with segmentation masks: left ventricle (LV), right ventricle (RV), left atrium (LA), right atrium (RA), myocardium (Myo), ascending aorta (AA), pulmonary artery (PA).

field of view and resolution of the CT image which yields a pre-registered MR image. Additionally, we resample the images on a resolution of 160 × 160 × 160 and normalize the image intensities to a range of $[0, 1]$.

## 2.4 Experiments and Results

We evaluate our method by computing the volume overlap of the segmentation masks using the Dice Score and the Average Surface Distance (ASD). If a deformation field represents accurate correspondences, the segmentation of the fixed image $b_{\mathcal{F}}$ and the warped segmentation of the moving image $b_{\mathcal{M}}(y)$ should overlap well. Furthermore, the registration should not generate deformations with foldings. Therefore we evaluated the Jacobian Determinant of the 3D deformation fields as it is a local measure for volume change and in particular for (local) change of topology. If $\det \nabla y > 1$ a volume expansion occurrs, if $\det \nabla y < 1$ the volume decreases and for $\det \nabla y \leq 0$ there is a folding.

### 2.4.1 Architecture Adaption for multi-modal Registration

In this experiment, we investigate the need for adaption of the architecture used in [44] for multi-modal image registration. For mono-modal image registration, it is helpful to use the same shared weights for both processing streams to compute the same first feature maps out of the input image (siamese architecture). However, due to different intensities, this assumption may not be valid for multi-modal image registration. For that reason we train a network with and without shared weights for the first convolution layer. The rest of the architecture remains unchanged. Using shared weights (siamese) yields to a Dice Score of 0.72 and an ASD of 3.16 mm on the first fold of the cross-validation, whereas using individual weights achieves a Dice Score of 0.78 and an ASD of 2.62 mm, which shows the advantage of a non-siamese architecture for multi-modal image registration with CNNs.

### 2.4.2 Comparison of 2D to 2.5D Registration

We use the following experiment to evaluate the performance improvement by using a 2.5D registration network instead of a 2D network used in [44]. For this purpose we generate three independent 3D deformation fields for each image through the slice-wise computation and concatenation of 160 2D deformation fields for each direction (axial, coronal and sagittal). Due to the concatenation, each 3D deformation field has one zero component. For the 2D registration we independently apply these three deformation fields to the segmentation of our moving image, compute the Dice Scores and average the three resulting values for each label. For our 2.5D registration, we combine the three deformation fields by averaging the two non-zero components point-wise for each direction. The resulting 3D deformation field is also applied to the segmentation


Figure 2.4: Example segmentations $b_{\mathcal{M}}$ and $b_{\mathcal{F}}$ of moving and fixed input images and the result of a 2D registration [55] in the direction of slicing and our proposed 2.5D registration.

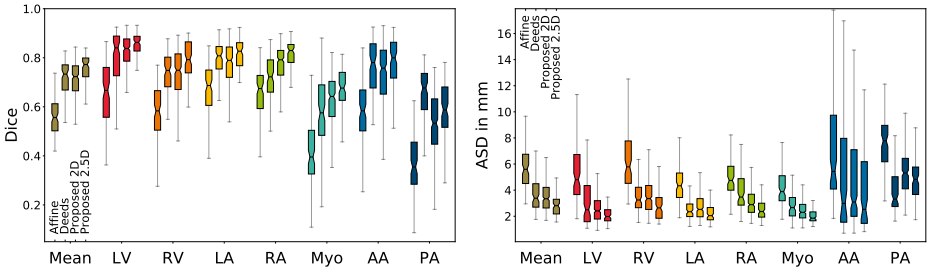



Figure 2.5: Comparison of Dice overlap and Average Surface Distance for all test images and each anatomical label (average of all labels, left ventricle (LV), right ventricle (RV), left atrium (LA), right atrium (RA), myocardium (Myo), ascending aorta (AA), pulmonary artery (PA)). For each one the distributions of Dice coefficients after an affine pre-alignment, a deeds registration, our proposed 2D and after our proposed 2.5D registration are shown.

of our moving image, yielding one 2.5D Dice Score for each label.
Comparing the resulting 2D and 2.5D Dice Scores, we see an increase of 4 percentage points by combining the information of the three deformation fields to one (cf. Table 2.1 and Figure 2.5 for a label-wise comparison). Figure 2.4 shows that the concatenated 2D deformation fields are not able to compensate deformations in the through-plane direction, which yields a situation where labeled structures which are not present in the moving but in the fixed image can not be mapped correctly by these deformations. In contrast to that our 2.5D registration is able to correctly transform these structures, because the combination of the three deformation fields contains information about deformations in all directions. Both the 2D and 2.5D registration can be performed slice- or image-wise, yielding the same registration results. This enables the control between a short runtime and low memory usage (c.f. Table 2.1).

### 2.4.3 Comparison of 2.5D to 3D Registration

In this experiment, we evaluate the performance of our 2.5D registration network compared to a state-of-the-art 3D registration network. For this propose, we expand our proposed U-Net based architecture from 2D to 3D which is comparable to the network used in [33] combined with the 3D version of our loss function. However, some parameters have to be adapted to have the same impact in 3D. The most important parameter is the edge parameter of the NGF image similarity. Since we sum the gradients over all dimensions, this parameter has to be increased to achieve a similar gradient field. Additionally, we slightly increased the impact of the regularization term and the boundary term by increasing $\alpha$ and $\beta$.
Comparing the resulting Dice Scores and Surface Distances, nearly no differences are visible. However, the 3D method shows slightly less foldings (0.1 % to 0.68 % - cf. Table 2.1).

### 2.4.4 Comparison to state-of-the-art Registration

We compare our method to the state-of-the-art 3D unsupervised iterative registration framework deeds [56] with the use of the self-similarity context metric (SSC) as parameterized for CT-MR registration in [57]. It has won the first place in a comprehensive abdominal registration comparison [58] and was ranked second in a recent MR-US registration challenge [59]. By employing a densely sampled discretized search space for displacement vectors, the method can capture large deformations robustly. The default parameters for displacement range and quantization as well as the multi-level grid settings were applied. Figure 2.5 presents the Dice Scores for each label and for the average over all labels as a boxplot and in Table 2.1 the average Dice Scores are shown. Our proposed method achieves slightly better average Dice Scores compared to the deeds method (0.74 to 0.71 ) by considerably shorter execution times (0.19 s to 16 s). To give an indication of the registration results, we show a checkerboard between the fixed image and moving image and between fixed image and the warped image after deeds registration and our proposed registration for two different patients in Figure 2.6. Finally, Figure 2.7 shows the same checkerboard overview with the corresponding segmentation masks.

### 2.4.5 Parameter and Regularity Analysis

To investigate the effect of each term in the loss function, we train our model with different settings of weighting parameters. Therefore, we set some parameters to zero and fix the others to their empirically determined optimal values ($\alpha = 1$, $\beta = 10$ and $\gamma = 0.2$). To reduce training time, we perform this experiment only on one fold of the cross-validation. Table 2.2 illustrates the results of this experiment showing that either the curvature regularizer or the volume change penalty is necessary to obtain


Figure 2.6: Example moving and fixed input images $\mathcal{M}$ and $\mathcal{F}$, checkerboard of fixed and moving images, checkerboard of fixed and warped images after deeds registration and after our proposed 2.5D registration.

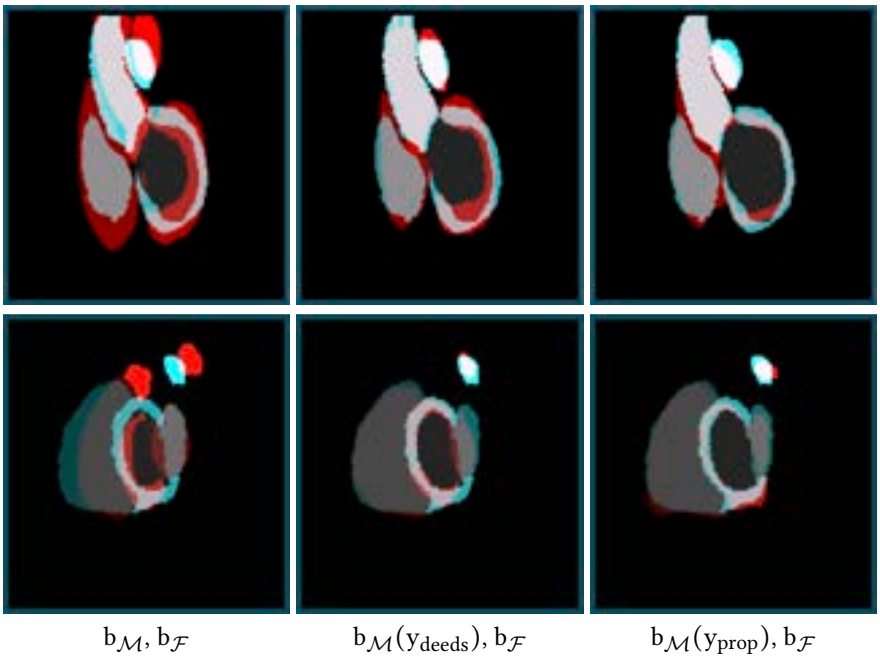



Figure 2.7: Example overlay of fixed and moving segmentations $\mathrm{b}_{\mathcal{F}}$ and $\mathrm{b}_{\mathcal{M}}$, overlay of fixed and warped segmentations after deeds registration and after our proposed 2.5D registration. The color overlays show the fixed image in red and the (registered) moving image in blue. Due to the addition of RGB values, aligned structures appear gray or white.

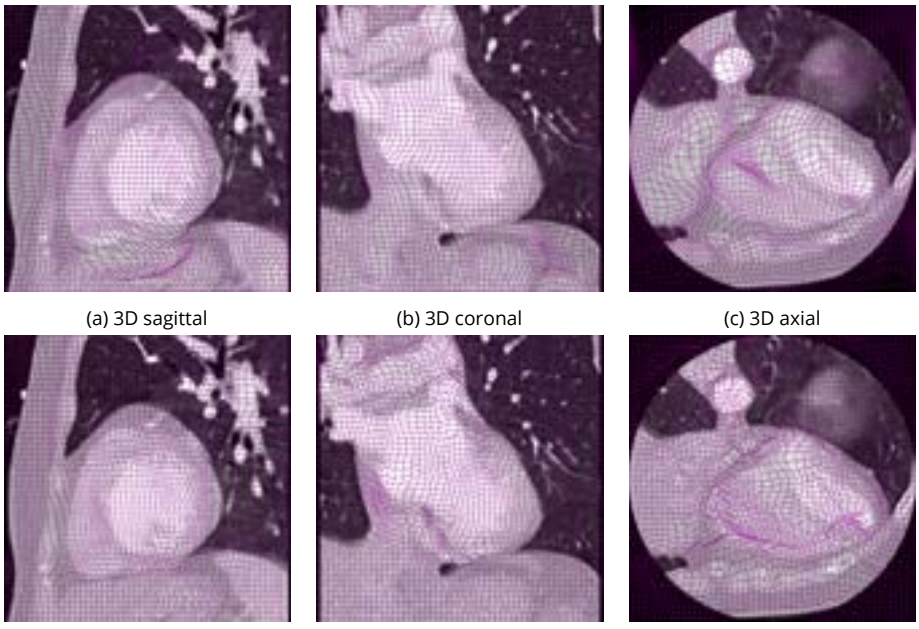
| Method | Dice Score | ASD | Foldings | Memory | Runtime |
|---|---|---|---|---|---|
| Affine pre-alignment | 0.55 | 6.17 mm | 0% | - | - |
| Prop. 2D slice-wise | 0.70 | 3.96 mm | 0.01% | 734 MB | 0.77 s |
| Deeds | 0.71 | 3.82 mm | 0 % | - | 16 s |
| U-Net 3D | 0.74 | 3.46 mm | 0.10% | 1820 MB | 0.18 s |
| Prop. 2.5D image-wise | 0.74 | 3.37 mm | 0.68% | 1840 MB | 0.19 s |
| Prop. 2.5D slice-wise | 0.74 | 3.37 mm | 0.68% | 730 MB | 2.35 s |

Table 2.1: Dice Score, Average Surface Distance (ASD) and foldings of different registration methods are shown evaluated using a cross-validation: after preprocessing (cf. 2.3.2), after our proposed 2D registration (cf. 2.4.2, on GPU), after a deeds registration (cf. 2.4.4, on (multi-core) CPU), after a 3D deep-learning-based registration (cf. 2.4.3) and after our proposed 2.5D registration (cf. 2.4.2, both on GPU). Additionally the runtime and GPU memory usage for registration of one unseen image pair are shown. The 2D and 2.5D CNN registration can be performed slice- or image-wise.

reasonable registration results. Using only the NGF image similarity yields not only a higher number of foldings but also lower Dice Scores due to the point-wise minimization without consideration of the underlying structures. Furthermore, the experiment shows that using global semantic information during training supports the alignment of those structures during the registration of new scan pairs. To give an indication of the regularity of our deformation fields, we show resulting 2.5D and 3D deformation fields in Figure 2.8 as an orthogonal view. Combining the three 2D deformation fields to a 2.5D deformation results in a good but slightly less smooth approximation of the 3D deformation field. Furthermore, we analyze volume changes using the Jacobian Determinant for which no unique optimal distribution exists. However, assuming that the volume of most tissue stays nearly the same, a mean of approximately one is expected. Moreover, we expect some voxels with a increasing ($\det \nabla y > 1$) or decreasing ($\det \nabla y < 1$) volume. For example in the lung due to respiratory motion or in the heart due to heart contraction. Those volume changes should not be too large and especially no foldings should occur. In Figure 2.9, we visualize the distribution of the Jacobian Determinant values for both approaches for the first fold of our cross-validation, showing that our expectations were met in both of them.

## 2.5 Discussion and Conclusion

We have presented a new 2.5D weakly-supervised deep-learning-based method for multi-modal image registration that replaces iterative optimization steps with deep CNN layers. We demonstrated that two independent processing streams for extracting the low-level image features are important to overcome the difficulties of multi-modal images. Later on, a similar network architecture as for mono-modal image registration



(a) 3D sagittal (b) 3D coronal (c) 3D axial

(d) 2.5D sagittal (e) 2.5D coronal (f) 2.5D axial

Figure 2.8: Comparison of 3D (first row) and 2.5D (second row) in-plane deformation fields of an example sagittal (first column), coronal (second column) and axial (third column) slice.

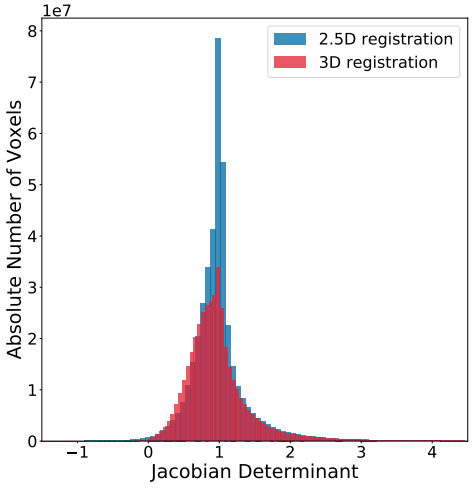



Figure 2.9: Quantitative visualization of the Jacobian Determinant representing the voxel-wise volume change. The histograms for our proposed 2.5D approach and for the 3D registration are shown.

| Method | Dice Score | Surface Distance | Foldings |
|---|---|---|---|
| Affine pre-alignment | 0.57 | 5.54 mm | 0% |
| full loss function | 0.78 | 2.62 mm | 0.52% |
| NGF only ($\alpha = \beta = \gamma = 0$) | 0.45 | 6.10 mm | 47.1% |
| no VCC ($\gamma = 0$) | 0.70 | 3.67 mm | 23.1% |
| no boundary ($\beta = 0$) | 0.63 | 4.47 mm | 0.38% |
| no curv ($\alpha = 0$) | 0.76 | 2.78 mm | 0.53% |

Table 2.2: The quantitative effect of variations of the terms within the loss function $\mathcal{L} = \mathcal{D} + \alpha \cdot \mathcal{R} + \beta \cdot \mathcal{B} + \gamma \cdot \mathcal{V}$ is shown by setting parameters to zero and fixing the others to their empirically determined optimal values ($\alpha = 1$, $\beta = 10$ and $\gamma = 0.2$). Besides the resulting Dice coefficient, the Surface Distance and the percentage of voxels in which foldings ($\det \nabla y \leq 0$) occur is depicted for the first fold of the cross-validation.

can be used. Additionally, we showed that estimating 3D deformations from 2D networks by simply averaging the deformation fields yields sufficient results which can be used for propagating segmentation mask from one modality to another. Although no 3D regularity is required during the training, the 2D regularity conditions (curvature regularizer and volume change control) are sufficient to ensure sufficient smoothness in the combined 3D deformation field. Our 2.5D framework is more memory-efficient than usual 3D methods because deformation fields can be computed slice-wise. This provides the possibility of using larger networks with more learnable parameters and higher batch sizes. Our proposed 2.5D registration method only needs 720MB for a training with a batch size of one. Therefore, the training on a NVIDIA GTX 1080 with 8 GB GPU-memory can process a batch of 200 2D-slices in parallel, while the same network for 3D-processing is limited to a batch size of 1 on this card (using 5160 MB). Especially for small dataset sizes, our 2.5D method can increase the variability of the mini-batches during training by randomly selecting slices from different patients. As demonstrated in previous studies (e.g. [60]) the increased variability of patches from different locations/subjects within one mini-batch greatly improves convergence of deep-learning models. Another advantage of our memory-efficient method is the possibility of training on relatively small GPUs for example directly in the clinic. Moreover, registering a new pair can be performed slice-wise and therefore only requires 730 MB GPU memory, whereas the 3D network needs 1.82 GB. Alternatively to our 2.5D approach, a patch-based approach like [41, 47] could be used to reduce the required memory which has its advantages and disadvantages. In contrast to our approach, it uses a direct 3D input. However, like in [47] only for a limited number of grid points deformation vectors are computed which are combined afterwards by B-spline interpolation. As a consequence, deformations smaller than the grid spacing can not be represented appropriately. Decreasing the grid spacing results in higher memory requirements attenuating the benefit of patch-based registration approaches.

Our approach advances the state-of-the-art in CNN-based deformable registration by firstly integrating a volume change control term into the loss function to explicitly penalize foldings in the deformation fields. We showed that using this additional term in the loss function significantly reduce the percentage of voxels in which foldings occur. Moreover, we combine the complementary strengths of global semantic information (weakly-supervised learning with segmentation labels) and local distance metrics borrowed from conventional medical image registration that supports the alignment of surrounding structures. Despite the increased focus on the alignment of the segmentation masks, it has be shown that the remaining image regions were transformed in a meaningful way. Particularly, the use of a multi-modal distance measurement is important for this aspect. In further experiments without a distance measure, it was observed that the result have deteriorated with regard to the Dice Score and additionally the remaining image regions were visually not well aligned. Additionally, we showed that using only a distance measure leads also to worse results demonstrating that regularization is unavoidable to obtain reasonable registration results. The results of our method demonstrate high Dice Scores (of 0.74), computation times of less than 0.2 second per 3D scan pair and compare favourably to the state-of-the-art unsupervised deeds approach (0.71) [56], which has won the first place in a comprehensive abdominal registration comparison [58]. We also tried to compare our results with the label-driven approach of [32], which is publicly available. Unfortunately, we were only able to achieve a slight improvement of the Dice coefficient compared to the affine pre-alignment.

Our method has two benefits over conventional registration methods. First, in contrast to conventional methods, our deep-learning-based registration method only performs an iterative optimization during the training of the network. After the network parameters have been learned, a registration is performed with a single forward-pass through the combined networks and without further optimization. This results in a very fast registration algorithm with less then 0.2 s for a 3D registration. Second, our method allows integrating label information in form of a penalty term into the loss function, which is only required during the training process and not during inference. In contrast, for conventional registration method such additional information has to be available for each new registration case. However, it increases the registration accuracy considerably.

Another natural idea for future improvement is replacing the simple average of the three 2D deformations into a single 3D vector field by training an additional fusion network. For that purpose, the three 2D networks would be still trained individually and afterwards a fusion network could be trained to combine the 2D deformation fields to a smooth 3D deformation field (for example by using the same loss function as before but in 3D). Hereby, only one network has to be trained at the same time but a smoother

deformation field could be reached in the end.

3

# CHAPTER 3

# mlVIRNET: Multilevel Variational Image Registration Network

**BASED ON:** A. Hering, B. van Ginneken, and S. Heldmann. "mlVIRNET: multilevel variational image registration network," *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2019,* Springer. 2019, pp. 257–265.

## Abstract

We present a novel multilevel approach for deep learning based image registration. Recently published deep learning based registration methods have shown promising results for a wide range of tasks. However, these algorithms are still limited to relatively small deformations. Our method addresses this shortcoming by introducing a multilevel framework, which computes deformation fields on different scales, similar to conventional methods. Thereby, a coarse-level alignment is obtained first, which is subsequently improved on finer levels. We demonstrate our method on the complex task of inhale-to-exhale lung registration. We show that the use of a deep learning multilevel approach leads to significantly better registration results.

## 3.1 Introduction

Image registration is the process of aligning two or more images to achieve point-wise spatial correspondence. This is a fundamental step for many medical image analysis tasks and has been an active field of research for decades. Since recently, deep learning based approaches have been successfully employed for image registration [33, 42, 44, 47, 55, 61]. They have shown promising results in a wide range of application. However, capturing large motion and deformation with deep learning based registration is still an open challenge. In common iterative image registration approaches, this is typically addressed with a multilevel coarse-to-fine registration strategy [11, 14, 62]. Starting on a coarse grid with smoothed and down-sampled versions of the input images a deformation field is computed which is subsequently prolongated on the next finer level as a initial guess. Hereby, a coarse level alignment is obtained first that typically captures the large motion components and which is later improved on finer levels for the alignment of more local details. Most of the recently presented deep learning based approaches also make use of a multilevel strategy as they are based on the U-Net architecture [33, 42, 44, 55]. Thereby, the first half of the "U" is used to generate features on different scales starting at the highest resolution and reducing the resolution through pooling operations. In this procedure, however, only feature maps on different levels are calculated but neither different image resolutions are used nor deformation fields are computed. Only a few approaches implement a multi-resolution or hierarchical strategy in the sense of multilevel strategies associated with conventional methods. In [63] the authors proposed an architecture which is divided into a global and a local network, which are optimized together. In [61] a multilevel strategy is incorporated into the training of a U-Net. Here, a CNN is grown and trained progressively level-by-level. In [47] a patch based approach is presented, where multiple CNNs (ConvNets) are combined additive into a larger architecture for performing coarse-to-fine image registration of patches. The results from the patches are then combined into a deformation field warping the whole image. In this work, we address this challenge and present a multilevel strategy for deep learning based image registration to advance state-of-the-art approaches. The contribution of this paper includes:

- We present deep learning based multilevel registration that is able to compensate and handle large deformations by computing deformation fields on different scales and functionally compose them.
- Our method is a theoretically sound and a direct transference of coarse-to-fine registration from conventional, iterative registration schemes to the deep learning based methods.
- We do not rely on patches. We take the whole image information into account and always consider the full field of view on all levels.

- A robust and fast registration method for the complex task of inhale-to-exhale registration validated on a large dataset of 270 thoracic CT scan pairs of the multi-center COPDGene study and on the publicly available DIR-Lab dataset [64].

## 3.2 Method

Our deep learning based framework for deformable image registration consists of two main building blocks. The first one is the specific design of the convolutional neural network and the loss function. In general, several architectures together with different distance measures, regularizer and penalty terms can be used. However, we focus on a U-Net based architecture, combined with a loss function that has shown good results for the task of pulmonary registration [65]. The second main building block is the embedding into a multilevel approach from coarse to fine. In the following, we give a brief outline of the variational setup, then we describe our particular architecture and loss function and, finally, we present its embedding into a multilevel approach.

### 3.2.1 Variational Registration Approach:

Following [66], let $\mathcal{F}, \mathcal{M} : \mathbb{R}^3 \to \mathbb{R}$ denote the fixed image and moving image, respectively, and let $\Omega \subset \mathbb{R}^3$ be a domain modeling the field of view of $\mathcal{F}$. We aim to compute a deformation $y : \Omega \to \mathbb{R}^3$ that aligns the fixed image $\mathcal{F}$ and the moving image $\mathcal{M}$ on the field of view $\Omega$ such that $\mathcal{F}(x)$ and $\mathcal{M}(y(x))$ are similar for $x \in \Omega$. The deformation is defined as a mimimizer of a suitable cost function that typically takes the form

$$\mathcal{J}(\mathcal{F}, \mathcal{M}, y) = \mathcal{D}(\mathcal{F}, \mathcal{M}(y)) + \alpha \mathcal{R}(y) \tag{3.1}$$

with so-called distance measure $\mathcal{D}$ that quantifies the similarity of fixed image $\mathcal{F}$ and deformed moving image $\mathcal{M}(y)$ and so-called a regularizer $\mathcal{R}$ that forces smoothness of the deformation typically by penalizing of spatial derivatives. Typical examples for the distance measure are, e.g., the squared $L_2$ norm of the difference image (SSD), cross correlation (CC) or mutual information (MI). In our experiments, we follow the approach of [65] using the edge based normalized gradient fields distance measure (NGF) and second order curvature regularization.

### 3.2.2 Deep Learning based Image Registration

In contrast to conventional registration [66], we do not employ iterative optimization during inference of new unseen images but use a convolutional neural network (CNN) that takes images $\mathcal{F}$ and $\mathcal{M}$ as input and yields the deformation $y$ as output. Thus, in the context of CNNs we can consider $y$ as a function of a trainable CNN model parameter vector $\theta \in \mathbb{R}^P$ and input images $\mathcal{F}$, $\mathcal{M}$, i.e. $y(x) \equiv y(\theta; \mathcal{F}, \mathcal{M}, x)$. In an unsupervised learning approach, we set up a loss function $\mathcal{L}$ that depends on $\mathcal{F}$, $\mathcal{M}$

and y, and then $\theta$ is learned by training, i.e., minimizing the expected value of $\mathcal{L}$ among a set of representative input images w.r.t. $\theta$. A natural choice would $\mathcal{L} = \mathcal{J}$. However, in our particular application, we have additional information available during training and we perform a weakly supervised approach. To this end, we define our loss function as suggested in [55]

$$\mathcal{L}(\mathcal{F}, \mathcal{M}, \mathrm{b}_{\mathcal{F}}, \mathrm{b}_{\mathcal{M}}, \mathrm{y}) = \mathcal{J}(\mathcal{F}, \mathcal{M}, \mathrm{y}) + \frac{\beta}{2} \|\mathrm{b}_{\mathcal{F}} - \mathrm{b}_{\mathcal{M}}(\mathrm{y})\|_{\mathrm{L}_2}^2 \tag{3.2}$$

with binary segmentation masks $\mathrm{b}_{\mathcal{F}}$ and $\mathrm{b}_{\mathcal{M}(\mathrm{y})}$ of the fixed and warped moving image, respectively. Note that these segmentations are only used to evaluate the loss function for training and their are not used as network input.

### 3.2.3 Single Level Architecture

Our CNN $\mathrm{y} \equiv \mathrm{y}(\theta, \mathcal{M}, \mathcal{F})$ is based on a U-Net which takes the concatenated 3D moving and fixed image as input and predicts a 3D dense displacement field. The network consists of three resolution levels starting with 16 filters in the first layer, which are doubled after each downsampling step. We apply 3D convolutions in both encoder and decoder stage with a kernel size of 3 followed by a batch normalization and a ReLU layer. For downsampling the feature maps during the encoder path, an $2 \times 2 \times 2$ average pooling operation with a stride of 2 is used. Transposed convolutions upsample and halve the feature maps in the decoder path. At the final layer, a $1 \times 1 \times 1$ convolution is used to map each 16 component feature vector to a three dimensional displacement vector.

### 3.2.4 Multilevel Deep Learning based Registration

Multilevel continuation and scale space techniques have been proven very efficient in conventional variational registration approaches to avoid local minima, to reduce topological changes or foldings and to speed up runtimes [11, 14, 62, 67]. However, beside carrying over these properties, our major motivation here is, to overcome the limitation of deep learning based registration to small and local deformations.

We follow the ideas of standard multilevel registration and compute coarse grid solutions that are prolongated and refined on the subsequent finer level. To this end, first we create image pyramids $\mathcal{F}_\ell, \mathcal{M}_\ell$ for $\ell = 1, \cdots, \mathrm{L}$ with coarsest level L. We start on finest level $\ell = 1$ and subsequently halve image size and resolution from level to level. Registration starts on coarsest level L and we compute deformation $\mathrm{y}_\mathrm{L}$ from images $\mathcal{F}_\mathrm{L}$ and $\mathcal{M}_\mathrm{L}$ as network input. On all finer levels $\ell < \mathrm{L}$, we incorporate the deformations from all preceding coarse levels as initial guess. Therefore, we combine them by functional composition and warp the moving image at current level. Let $\mathrm{X}_\ell$ denote the cell-centered image grid on level $\ell$, we compute the warped moving $\mathcal{M}_\ell(\mathrm{Y}_\ell)$

**Algorithm 1:** Multilevel Deep Learning Registration

**IN** : Fixed image $\mathcal{F}$, moving image $\mathcal{M}$, image grid X

**OUT** : Corse-to-fine deformations $y_L, ..., y_1$, transformed grid $Y = y_1 \circ \cdots \circ y_L(X)$

1 Create image pyramid $\mathcal{F}_\ell, \mathcal{M}_\ell$ for $\ell = 1, 2, ..., L$ with finest level $\ell = 1$ and L coarsest.
2 On coarsest level Compute deformation $y_L = \text{CNN}(\mathcal{F}_L, \mathcal{M}_L)$
3 **for** $\ell = L - 1, L - 2, ..., 1$ **do**
4 Compute transformed grid $Y_\ell = y_{\ell+1} \circ ... \circ y_L(X_\ell)$
5 Compute deformation $y_\ell = \text{CNN}(\mathcal{F}_\ell, \mathcal{M}_\ell(Y_\ell))$
6 **end**

with

$$Y_\ell := y_{\ell+1} \circ y_{\ell+2} \circ \cdots \circ y_L(X_\ell)$$

and use it together with fixed image $\mathcal{F}_\ell$ as network input, yielding the deformation field $y_\ell$ on the current level. The final output deformation y is then given by composition of the whole sequence of coarse-to-fine solutions, i.e., $y = y_1 \circ y_2 \circ \cdots \circ y_L$. To evaluate deformations and images at non-grid grid points, we use trilinear interpolation. Our scheme is summarized in Algorithm 1.

In our experiments we use in particular a three level scheme (L = 3). and we create image pyramid with three reduced resolution images generated from the original 3D images by applying a low-pass filter with a stride of two, four and eight. During training, the three networks are learned progressively. First, the network on the coarsest level is trained for a fixed amount of epochs. Afterwards, the parameters of the middle network are learned while the coarsest network stays fixed and is only used to produce the initial deformation field. The same procedure is repeated on the finest level. The same architecture is used on all levels. The convolution parameters on the coarsest level are initialized with Xavier uniform [68]. Whereas, all other networks are using the learned parameters of the previous network as initialization. Note that the receptive field in voxel is the same for all used networks, however, due to to the decreased resolution on the coarse levels, the receptive field in mm is much higher.

## 3.3 Experiments and Results

We demonstrate our deep learning based registration method by registration of inhale-to-exhale lung CT scans. We use data from 500 patients for training and a disjoint set of 50 patients for validation from the COPDGene study, a large multi-center clinical trial with over 10.000 subjects with chronic obstructive pulmonary disease (COPD) [69]. The dataset was acquired across 21 imaging centers using a variety of scanner makes and models. Each patient had received two breath-hold 3D CT scans, one on full inspiration

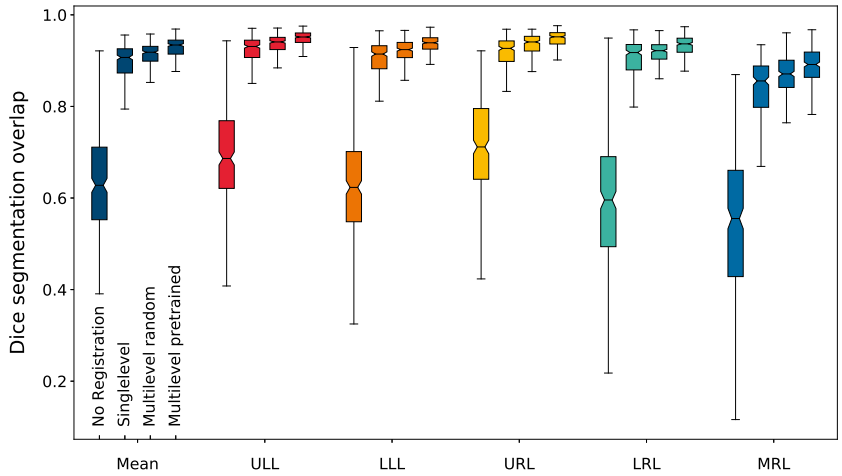



Figure 3.1: Comparison of Dice overlaps for all test images and each anatomical label (average of all labels, upper left lobe (ULL), lower left lobe (LLL), upper right lobe (URL), lower right lobe (LRL), middle right lobe (MRL)). For each one the distributions of Dice coefficients before registration, after single level dl registration, multilevel dl registration without pretrained CNNs and after multilevel registration with pretrained CNNs.

(200 mAs) and one at the end of normal expiration (50 mAs). For all scans segmentations of the lobes are available, which were computed automatically and manually corrected and verified by trained human analysts. The original images have sizes in the range of $512 \times 512 \times \{430, \cdots, 901\}$ voxels. Due to memory and time limitations, we create low-resolution images by resampling to a fixed size of $160 \times 160 \times 160$ voxels. The low-resolution images are then used during training for the computation of the deformation field and for evaluating of the loss function. Note that, our method is generally not limited to any fixed input size. Although we use images with $160^3$ voxels, the computed deformation field are defined on full field-of-view and can be evaluated on grids with arbitrary resolution by using trilinear interpolation. Consequently, we use original full-resolution images for the evaluation of our method.

## Multilevel vs. Single Level

First, we evaluate our multilevel approach on a disjoint subset of 270 patients from the COPDGene study. We compare our proposed method against a single level approach with only one U-Net using the images on finest level as inputs. We train both approaches for 75 epochs with the same hyper-parameters. For the multilevel approach, the epochs are split equally at each level. We also evaluate the effect on how the network parameter are initialized. Therefore, we compare a Xavier initialization for all convolution parameters of all three networks against our proposed progressive

learning strategy. Therefore, only the convolution parameters on the coarsest level are initialized with Xavier initialization and the training of subsequent network is started with the learned parameters form the network of the previous level.

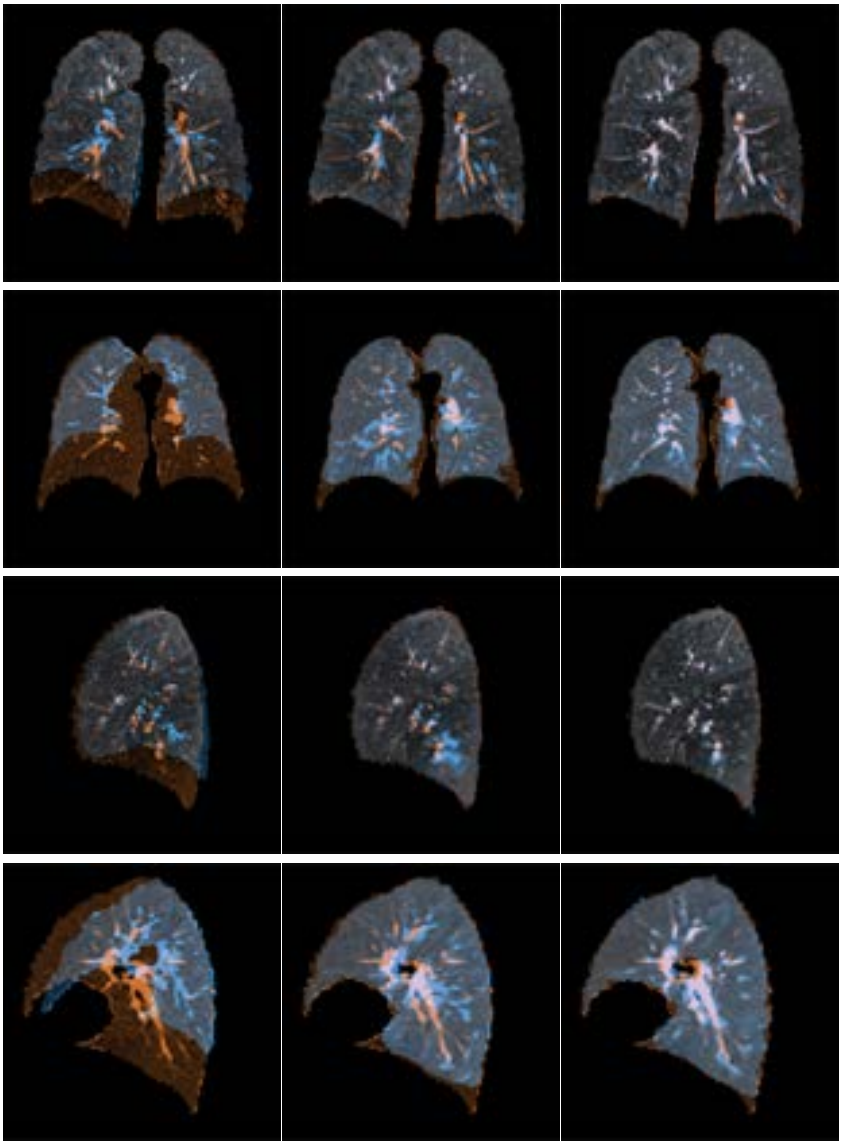
We evaluate our method by measuring the overlap of the lobe masks. The underlying assumption is, that if a deformation yields accurate spatial correspondences, then lobe segmentations of the fixed and the warped lobe segmentation of the moving image should overlap well. Figure 3.1 shows the Dice scores for each label and the average over all labels as a box-plot. Our proposed multilevel approach increase the Dice Score from 63.5 % to 92.1 %. In contrast, the single level method archive a Dice Score of 88.3 %. Furthermore, the multilevel approach produced less foldings (0.3 % to 2.1 %). Figure 3.2 shows representative qualitative results for of two scan pairs before registration and after our single level and multilevel registration. In both cases the respiratory motion was successfully recovered. Although the single level registration produces reasonable Dice scores, it does not well align the inner structures. This is also reflected by the landmark errors in the following section. Comparing the results of the pretrained initialization to the random initialization, an improvement of about 2 % in terms of the Dice Score could be reached.

### 3.3.1 Comparison with state-of-the-art

Additionally, we evaluate our method and compare it to others on the public available DIR-Lab dataset [64]. It is a collection of ten inspiration-expiration cases with 300 expert-annotated landmarks in the lung. The landmarks are used for evaluating our deformable registration method. The mean (and standard deviation) for all ten scans for the deep learning based multi-resolution approaches of Eppenhof [61] and de Vos (DLIR) [47], the single VIRNET and our proposed method are listed in Table 3.1. The overall average landmark error is 2.19 mm with a standard deviation of 1.62 mm. In contrast to the other methods, our mlVIRNET is more robust against outliers and can better handle large initial landmark distances without training on this specific dataset.

## 3.4 Discussion and Conclusion

We presented an end-to-end multilevel framework for deep learning based image registration which is able to compensate and handle large deformations by computing deformation fields on different scales. Our method takes the whole image information into account and predicts a dense 3D deformation field. We validated our framework on the challenging task of large motion inhale-to-exhale registration using large image data of the multi-center COPDGene study. We have shown that our proposed method archives better results than the comparable single level variant. In particular with regard to the alignment of inner lung structures and the presence of foldings. Only less than 0.3 % voxel positions of the images showed a folding. Additionally, we demon-



(a) before (b) single level (c) multilevel

Figure 3.2: Visualization of two inspiration-expiration registration results: The first two rows show coronal views before and after single and multilevel registration, the last two rows sagittal views, respectively. The color overlays show the inhale scan in orange and the exhale in blue; due to addition of RGB values, aligned structures appear gray or white. In both cases the respiratory motion was successfully recovered. However, the single level registration does not well align the inner structures.

| Scan | Initial | Eppenhof [61] | DLIR [47] | single VIRNET | mlVIRNET |
|---|---|---|---|---|---|
| Case 1 | 3.89(2.78) | 2.18(1.05) | 1.27(1.16) | 1.73(0.83) | 1.33(0.73) |
| Case 2 | 4.34(3.90) | 2.06(0.96) | 1.20(1.12) | 2.38(1.11) | 1.33(0.69) |
| Case 3 | 6.94(4.05) | 2.11(1.04) | 1.48(1.26) | 3.01(1.86) | 1.48(0.94) |
| Case 4 | 9.83(4.85) | 3.13(1.60) | 2.09(1.93) | 4.28(2.37) | 1.85(1.37) |
| Case 5 | 7.48(5.50) | 2.92(1.70) | 1.95(2.10) | 3.17(2.2) | 1.84(1.39) |
| Case 6 | 10.89(6.96) | 4.20(2.00) | 5.16(7.09) | 4.85(3.04) | 3.57(2.15) |
| Case 7 | 11.03(7.42) | 4.12(2.97) | 3.05(3.01) | 3.67(1.82) | 2.61(1.63) |
| Case 8 | 14.99(9.00) | 9.43(6.28) | 6.48(5.37) | 5.75(3.93) | 2.62(1.52) |
| Case 9 | 7.92(3.97) | 3.82(1.69) | 2.10(1.66) | 4.90(2.25) | 2.70(1.46) |
| Case 10 | 7.30(6.34) | 2.87(1.96) | 2.09(2.24) | 3.49(2.21) | 2.63(1.93) |
| Total | 8.46(6.58) | 3.68(3.32) | 2.64(4.32) | 3.72(2.45) | 2.19(1.62) |

Table 3.1: Mean (standard deviation) of the registration error in mm determined on DIR-Lab 4D-CT data. From left to right: initial landmark error, the multi-resolution approaches of [61] and [47] and the single level VIRNET and the proposed multilevel VIRNET.

strated that using the network parameter of the previous level as initialization, yields to better registration results. Moreover, we demonstrated the transferability of our approach to new datasets by evaluating our learned method on the publicly available DIR-Lab dataset and showing a lower landmark error than other deep learning based registration methods.

**Acknowledgements:** We gratefully acknowledge the COPDGene Study for providing the data used. COPDGene is funded by Award Number U01 HL089897 and Award Number U01 HL089856 from the National Heart, Lung, and Blood Institute. The COPDGene project is also supported by the COPD Foundation through contributions made to an Industry Advisory Board comprised of AstraZeneca, Boehringer Ingelheim, GlaxoSmithKline, Novartis, Pfizer, Siemens and Sunovion.

# CHAPTER 4

# CNN-based Lung CT Registration with Multiple Anatomical Constraints

**BASED ON:** A. Hering, S. Häger, J. Moltz, N. Lessmann, S. Heldmann, and B. van Ginneken. "CNN-based lung CT registration with multiple anatomical constraints," *Medical Image Analysis* (2021), p. 102139.

## Abstract

Deep-learning-based registration methods emerged as a fast alternative to conventional registration methods. However, these methods often still cannot achieve the same performance as conventional registration methods because they are either limited to small deformation or they fail to handle a superposition of large and small deformations without producing implausible deformation fields with foldings inside.

In this paper, we identify important strategies of conventional registration methods for lung registration and successfully developed the deep-learning counterpart. We employ a Gaussian-pyramid-based multilevel framework that can solve the image registration optimization in a coarse-to-fine fashion. Furthermore, we prevent foldings of the deformation field and restrict the determinant of the Jacobian to physiologically meaningful values by combining a volume change penalty with a curvature regularizer in the loss function. Keypoint correspondences are integrated to focus on the alignment of smaller structures.

We perform an extensive evaluation to assess the accuracy, the robustness, the plausibility of the estimated deformation fields, and the transferability of our registration approach. We show that it achieves state-of-the-art results on the COPDGene dataset compared to conventional registration method with much shorter execution time. In our experiments on the DIR-Lab exhale to inhale lung registration, we demonstrate substantial improvements (TRE below 1.2 mm) over other deep learning methods. Our algorithm is publicly available at https://grand-challenge.org/algorithms/deep-learning-based-ct-lung-registration/.

## 4.1 Introduction

Image registration is the process of aligning two or more images to achieve point-wise spatial correspondence. This is a fundamental step for many medical image analysis tasks and has been an active field of research for decades [8, 9]. Various approaches and tailored solutions have been proposed to a wide range of problems and applications. Typically, image registration is phrased as an optimization problem with respect to a spatial mapping that minimizes a suitable cost function and common approaches estimate solutions by applying iterative optimization schemes. Unfortunately, solving such an optimization problem is computationally demanding and consequently slow.

While deep learning has become the methodology of choice in many areas, relatively few deep-learning-based image registration algorithms have been proposed. One reason for this is the lack of ground truth and the large variability of plausible deformations that can align corresponding anatomies. Therefore, the problem is much less supervised than for example image classification or segmentation. Nevertheless, several methods have been presented in the last years which aim to mimic the process of conventional image registration methods by training a neural network to predict the non-linear deformation function given two new unseen images. As a trained neural networks can process images in real time, this has immense potential for time-sensitive applications such as image guidance in radiotherapy, tracking, or shape analysis through multi-atlas registration.

In this paper, we target the challenging task of lung registration. The complexity of this registration task is manifold, as the occurring motion is a superposition of respiratory and cardiac motion. Moreover, the sliding motion between the lung and rib cage during breathing – more precisely between pleura visceralis and pleura parietalis – is an additional challenge. The scale of the motion within the lungs can often be larger than the structures (vessels and airways) that are used to guide the optimization process. This may cause a registration algorithm to get trapped in a local minimum [70, 71]. This makes the problem even more difficult.Therefore, a registration method needs to be able to estimate a displacement field that accounts for substantial breathing motion but also aligns small structures like individual pulmonary blood vessels precisely.

## 4.2 Related Work

Most deep-learning-based approaches aim to learn a registration function in form of a convolutional neural network to predict spatial deformations warping a moving image to a fixed image. All these works have contributed improving deep-learning-based image registration and have been applied to different registration applications including brain MR [28, 33, 72], cardiac MR [73], cardiac MR-CT [35], prostate MR-US [63], thorax-abdomen CT [74], thorax CT [26, 29, 34, 55, 61, 75] and CT-CBCT registration [76].

Existing approaches can be classified as *supervised*, *unsupervised*, and *weakly-supervised* techniques based on how much supervision is available.

*Supervised methods* use ground-truth deformation fields for training. The ground truth can be generated in different ways. In [24] and [25] the network is trained on synthetic random transformations. A drawback is that the randomly generated ground truth is artificial and may not be able to reproduce all possible deformations. Alternatively, conventional registration methods can be used to produce deformations by registering images [26, 27] or other image features like landmarks or segmentations [42]. Another way to create a ground truth is to combine simulations with existing algorithms [77]. Consequently, the performances of all these approaches is upper bounded by the quality of the initial registration algorithm or the realism of the synthetic deformations.

In contrast, *unsupervised methods* – also called *self-supervised methods* – do not require any ground truth. The idea is to use the cost function of conventional image registration (similarity measure and regularization term) as the loss function to train the neural network. An important milestone for the development of these methods was the introduction of the spatial transformer network [36] to differentiably warp images. This differentiable warping has actually been part of most conventional registration methods for a long time (e.g. [1, 19, 78]). The concept of an unsupervised deep-learning-based registration method was first introduced with the DIRNet [73] for 2D image registration using the normalized cross-correlation image similarity measure as loss function. In [79] the approach has been extended by adding diffusion regularization to the loss function forcing smooth deformations. The method has successfully been demonstrated for registration of 3D brain subvolumes. The idea of unsupervised deep-learning-based image registration has been further evolved in several works [28–31, 55].

*Weakly-supervised methods* do not rely on ground-truth deformation fields either but training is still supervised with prior information. In [63] and [32], a set of anatomical labels is used in the loss function. The labels of the moving image are warped by the deformation field and compared with the fixed labels. All anatomical labels are only required during training. In [33] and [34, 35], the complementary strengths of global semantic information and local distance metrics were combined to improve the registration accuracy.

In conventional registration approaches, multilevel continuation and scale-space techniques have been proven very efficient to avoid local minima during the optimization process of the cost function, to reduce topological changes or foldings, and to speed up runtimes [11, 14, 62, 67] – explaining the popularity of multi-level strategies in conventional registration methods. As a lot of deep-learning-based registration methods are build on top of U-Net (e.g. [33, 42, 44, 55]), they are also multi-leveled in their nature. The first half of the "U" (the encoder) generates features on different scales

starting at the highest resolution and reducing the resolution through pooling operations. In this procedure, however, only feature maps on different levels are calculated but neither are different image resolutions used nor deformation fields computed. Only a few approaches implement a multi-resolution or hierarchical strategy in the sense of multilevel strategies associated with conventional methods. In [63], the authors proposed an architecture that is divided into a global and a local network, which are optimized together. In [61], a multilevel strategy is incorporated into the training of a U-Net, by growing and training progressively level-by-level. In [29], a patch-based approach is presented, where multiple CNNs (ConvNets) are combined additively into a larger architecture for performing coarse-to-fine image registration of patches. The results from the patches are then combined into a deformation field warping the whole image. Another patch-based multilevel approach is presented in [80]. The multilevel framework consists of a CoarseNet and a FineNet which are trained jointly. During training, the estimated deformation field of the CoarseNet and the FineNet are not combined but the moving patch is transformed twice. During inference, if the mean absolute differences between the deformed image patch and the fixed image exceeds a predefined threshold, FineNet is applied again. This leads to a variable number of deformation field patches, which are combined additively. Although previous deep-learning-based registration works (e.g. [26, 29, 61]) contributes many efforts to improve the registration accuracy for lung registration, there is still a misalignment of smaller structures in the lung, which leads to a high target registration error of landmarks.

## Contribution

We previously introduced an end-to-end deep-learning multilevel registration method that can handle large deformations by computing deformation fields on different scales and functionally composing them [34]. This initial study, despite its limited evaluation, proved that it is a valid strategy to improve the alignment of vessels and airways – though a gap regarding the target registration error of landmarks with the best conventional registration methods remained. Building on this previous work, and addressing its limitations, we were able to further close that gap.

Our key contributions are as follow:

- We present multiple anatomical constraints to incorporate anatomical priors into the registration framework to obtain more realistic results. We integrate the lung lobe mask to consider the global context. Moreover, the keypoint correspondences are used to increase the alignment of airways and vessels.
- We introduce a novel constraining method to control volume change and therefore avoid foldings inside the deformation field. While the idea of volume change control is not new in conventional registration, we firstly present a suitable version for deep-learning-based image registration.

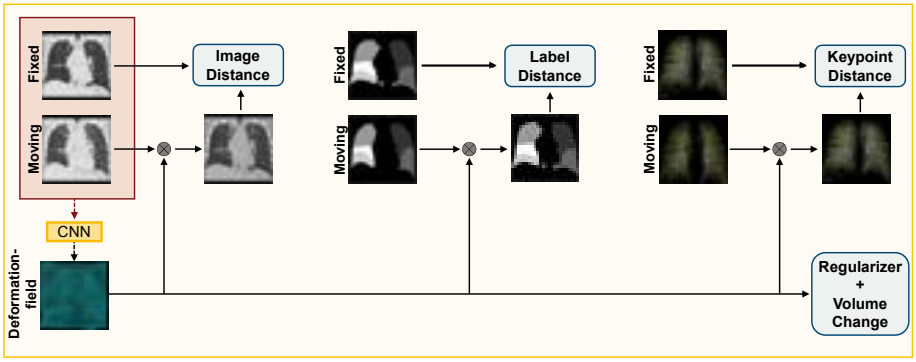



Figure 4.1: Schematic representation of the training process. In the loss function, we compare the fixed image, pulmonary lobes mask and keypoints to the deformed moving image, pulmonary lobes mask and keypoints, respectively. To enforce smoothness and to prevent foldings, a regularizer and a volume change penalty are integrated into the loss function. During inference, only the fixed and moving image is required to estimate the deformation field. For a better visualization, we have placed the windowed CT image in the background of the used keypoints. Best viewed in colors.

- We perform comprehensive experiments on three different datasets – the multi-center COPDGene study [69] and the DIR-Lab challenge dataset [64, 81], and the EMPIRE10 challenge dataset [82] – to assess the accuracy, plausibility, robustness, transferability of our method. We achieve comparable results as state-of-the-art registration approaches.

## 4.3 Method

### 4.3.1 Variational Registration Approach

Let $\mathcal{F}, \mathcal{M} : \mathbb{R}^3 \to \mathbb{R}$ denote the fixed image and moving image, respectively, and let $\Omega \subset \mathbb{R}^3$ be a domain modeling the field of view of $\mathcal{F}$. Registration methods aim to compute a deformation $y : \Omega \to \mathbb{R}^3$ that aligns the fixed image $\mathcal{F}$ and the moving image $\mathcal{M}$ on the field of view $\Omega$ such that $\mathcal{F}(x)$ and $\mathcal{M}(y(x))$ are similar for $x \in \Omega$. The deformation is defined as a minimizer of a suitable cost function that typically takes the form

$$\mathcal{J}(\mathcal{F}, \mathcal{M}, y) = \mathcal{D}(\mathcal{F}, \mathcal{M}(y)) + \alpha \mathcal{R}(y) \tag{4.1}$$

with so-called distance measure $\mathcal{D}$ that quantifies the similarity of fixed image $\mathcal{F}$ and deformed moving image $\mathcal{M}(y)$ and so-called regularizer $\mathcal{R}$ that forces smoothness of the deformation typically by penalizing spatial derivatives. Typical examples for the distance measure are the squared $L_2$ norm of the difference image (SSD), normalized cross correlation (NCC), or mutual information (MI). The cost function can be ex-

tended by additional penalty terms to force desired properties or incorporate additional knowledge in form of anatomical constraints [65]. As illustrated in Figure 4.1, our method inputs both the fixed and moving image into the network that predicts the dense displacement field. The loss function uses all available information: input images, segmentation masks and keypoints, with additional regularization – in the form of a smoothness prior and a volume consistency constraint – to prevent foldings.

### 4.3.2 Loss Function

**NORMALIZED GRADIENT FIELD DISTANCE MEASURE** One of the main challenges of lung registration are the varying intensity changes occurring due to the altered density of lung tissue during breathing. This leads to a violation of the intensity constancy assumption between corresponding points, on which the classic sum of squared differences (SSD) distance measure is built. However, the lung exhibits a rich structure of bronchi, fissures, and especially vessels that can be exploited for the registration, more suited to distance measure that focus on image edges rather than intensities. We follow the approach of [65] and [34] using the *normalized gradient fields* (NGF) [83] distance measure

$$\mathcal{D}(\mathcal{F}, \mathcal{M}(y)) = \int_\Omega 1 - \frac{\langle \nabla \mathcal{M}(y(x)), \nabla \mathcal{F}(x) \rangle_\epsilon^2}{\|\nabla \mathcal{M}(y(x))\|_\epsilon^2 \|\nabla \mathcal{F}(x)\|_\epsilon^2} \, dx,$$

with $\langle f, g \rangle_\epsilon := \sum_{j=1}^3 (f_j g_j + \epsilon^2)$, $\|f\|_\epsilon := \sqrt{\langle f, f \rangle_\epsilon}$. The edge hyper-parameter $\epsilon > 0$ is used to suppress small image noise, without affecting image edges. Therefore, a good strategy is to choose its value relative to the average gradient. In [83], the following automatic choice is suggested:

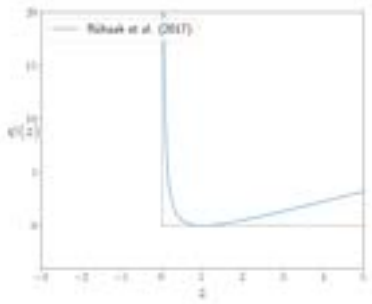
$$\epsilon = \frac{\nu}{V} \int_\Omega \|\nabla I(x)\| dx,$$

where $\nu$ is the estimated noise level in the image and V is the volume of the domain $\Omega$. For CT images, a value in the range of $[0.1, 10]$ is mostly a good choice.

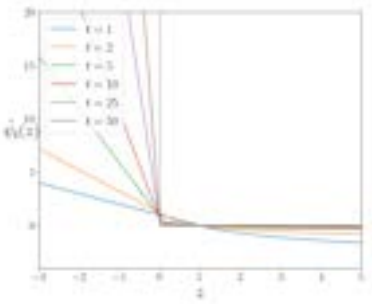
Since we focus on accurate registration inside the lungs and to avoid misalignment artifacts due to sliding motion at the pleura, we restrict $\Omega$ to the support of the lung mask of the fixed image.

**CURVATURE REGULARIZER** Smooth deformation fields are enforced by the second order *curvature regularizer* [84] given by

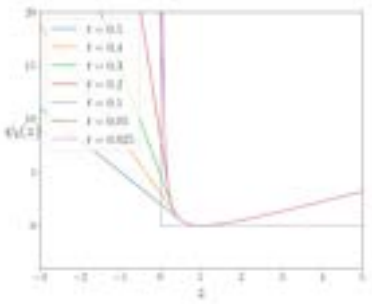
$$\mathcal{R}(y) = \int_\Omega \sum_{j=1}^3 \|\Delta y_j(x)\|^2 \, dx. \tag{4.2}$$



(a) Interior point method from [65]



(b) Log-barrier extension from [85]



(c) Our proposed penalty

Figure 4.2: A graphical illustration of both standard log-barrier (a), the proposed log-barrier extension (b) and examples of penalty functions (c). The solid curves in colors show the approximations for several t values of functions $\tilde{\psi}_t(z)$ and $\psi_t(z)$ respectively.

**VOLUME CHANGE CONTROL** Although the curvature regularization from Equation (4.2) prefers smooth deformation, foldings may still happen, which is obviously physically impossible. More formally, foldings happen when the Jacobian determinant of the deformation field becomes negative. To avoid any foldings, we therefore aim to minimize the distance measure $\mathcal{D}$ and the regularizer $\mathcal{R}$ while keeping the Jacobian determinant positive, for every voxel in $\Omega$. Formally, this can be written as a constrained optimization problem:

$$\begin{aligned} \min_{y} \quad & \mathcal{D}(\mathcal{F}, \mathcal{M}(y)) \; + \; \alpha \mathcal{R}(y) \\ \text{s.t.} \quad & \det \nabla y(x) > 0 \qquad \forall x \in \Omega. \end{aligned}$$

To achieve this, [65] introduced a *Volume Change Control* (VCC) that could be integrated in their overall objective:

$$\mathcal{V}(y) = \int_{\Omega} \psi(\det \nabla y(x)) dx, \tag{4.3}$$

where

$$\psi(z) = \begin{cases} \frac{(z-1)^2}{z} & \text{if } z > 0 \\ +\infty & \text{otherwise.} \end{cases} \tag{4.4}$$

For the sake of simplicity, the input to $\psi$ in equation 4.3 is substituted with $z = \det \nabla y(x)$ in equation 4.4. Notice that $\psi(z)$ is minimized when $z = 1$ (see Figure 4.2 (a)). Therefore, the regularizing effects of the VCC are twofold: i) prevents the formation of foldings, by keeping the determinants positive, ii) limits both shrinkage and expansions by biasing the optimization to keep the same volume.

The method that [65] used falls into the category of *interior-point methods*. Such methods became very popular in constrained optimization [86] as they do not require the expansive primal-dual updates of traditional Lagrangian optimization: the infinity penalty acts as a "barrier", preventing the optimization to go out of bounds.

To be used, interior-points methods require a feasible starting point: all constraints need to be strictly satisfied before starting the optimization procedure. This is usually done in a pre-optimization step (called Phase I) before the actual optimization of Phase II is performed. We can see it as finding a valid initial guess, and then refining it.

In the context of deep neural networks, standard Lagrangian methods are not feasible due to their expensive primal-dual updates, which requires to retrain a neural network (from scratch) at each iteration. Interior-point methods are also not applicable, as solving phase I requires to solve a constrained optimization problem in the first place.

[87] proposed a parametric log-barrier *extension* (illustrated in Figure 4.2 (b)), that does not require an initial feasible solution:

$$\tilde{\psi}_t(z) = \begin{cases} -\frac{1}{t}\log(z) & \text{if } z \geq \frac{1}{t^2} \\ -tz - \frac{1}{t}\log\left(\frac{1}{t^2}\right) + \frac{1}{t} & \text{otherwise.} \end{cases} \tag{4.5}$$

$t$ is a hyper-parameter, controlling the slope of the barrier. By starting with a small initial value, and increasing it as the training progresses, one is able to "raise" the barrier, closing it eventually.

We propose to keep the property of equation (4.4) to symmetrically penalizes local shrinkage and expansion and make it applicable for neural networks by using the barrier formulation of equation (4.5) for $z < t$ with $t \longrightarrow 0$ over time (illustrated in Figure 4.2 (c)):

$$\psi_t(z) = \begin{cases} \frac{(z-1)^2}{z} & \text{if } z \geq t \\ \left(1 - \frac{1}{t^2}\right) z + \frac{2(1-t)}{t} & \text{otherwise,} \end{cases} \tag{4.6}$$

with $t > 0$ which is a hyper-parameter controlling the slope of the linear barrier for $z < t$. This barrier can be raised during the training by decreasing the value of t to penalize foldings more strongly. Note that the linear part for $z < t$ is chosen such that $\psi$ is continuously differentiable provided $t > 0$. In our experiments, we set $t = 0.2$ for the first level of our multilevel architecture and decrease it by the factor of 2 for any further level. For $z \geq t$, we symmetrically penalize local shrinkage and expansion, i.e., $\psi(z) = \psi(1/z)$.

**MASK ALIGNMENT** Several recent publications (e.g. [33, 35]) have shown that adding further information in the form of segmentation masks into the loss function can guide the network during the training process. Since the segmentation masks are used in the loss function, they are only required during training and not for registration of unseen images. We integrate segmentation masks by using the SSD loss

$$\mathcal{B}(y) = \frac{1}{2} \int_\Omega \left\| b_{\mathcal{M}}(y(x)) - b_{\mathcal{F}}(x) \right\|^2 dx, \tag{4.7}$$

where $b_{\mathcal{F}} : \Omega \longrightarrow [0,1]^k$ and $b_{\mathcal{M}} : \Omega \longrightarrow [0,1]^k$ denote functions of $\mathcal{F}$ and $\mathcal{M}$ that are the one-hot representation of the segmentation mask, with k the number of different labels. For lung registration, we use segmentation of the lungs into the five pulmonary lobes ($k = 5$). During training, we use linear interpolation to warp the one-hot segmentation masks since this results in a smoother loss function at the border of the segmentation. With nearest neighbor interpolation, the loss of each voxel can either be one or zero. Linear interpolation allows for probabilistic loss values between zero and one.

**KEYPOINT LOSS** For conventional image registration, previous work (e.g. [65, 71, 88]) has shown that the integration of sparse keypoints during the optimization of the deformation field yields better registration results. In contrast to conventional registration approaches, keypoints can be integrated into the loss function and are therefore, similar to the segmentation masks for the mask penalty, only needed for training but not during inference. In general, there are several ways to integrate the keypoints into an intensity-based registration approach (e.g. [89], [65], [90]). We choose to integrate the keypoint information through a least-squares penalty into our model by directly comparing the transformed keypoint of the fixed image with the corresponding moving keypoint:

$$\mathcal{K}(y) = \frac{1}{|K|} \sum_{i=1}^{|K|} \left\| k_{\mathcal{M}}^i - y\left(k_{\mathcal{F}}^i\right) \right\|^2$$

with the moving keypoint $k^i_{\mathcal{M}}$ and the warped fixed keypoint $y\left(k^i_{\mathcal{F}}\right)$ for all $|K|$ keypoints. In general, manually annotated landmarks or automatically generated keypoints can be integrated with this loss function. However, since manual annotation of landmarks is time-consuming, we use the keypoint detection algorithm described in [65] to generate a large number of corresponding keypoints.

The final loss is given by

$$\mathcal{L}(\mathcal{F}, \mathcal{M}, y) = \mathcal{D}(\mathcal{F}, \mathcal{M}(y)) + \alpha\mathcal{R}(y) + \beta\mathcal{B}(y) + \gamma\mathcal{V}(y) + \delta\mathcal{K}(y). \tag{4.8}$$

The hyper-parameters $\alpha, \beta, \gamma$ and $\delta$ have to be chosen manually. However, our experiments showed that a change in the magnitude leads to only slight changes in the results.

### 4.3.3 Baseline Architecture

Our CNN is based on a U-Net [53] which takes the concatenated 3D moving and fixed image as input and predicts a 3D dense displacement field with the same resolution as the input images. The U-Net consists of three levels starting with 16 filters in the first layer, which are doubled after each downsampling step. We apply 3D convolutions in both encoder and decoder path with a kernel size of 3 followed by an instance normalization and a ReLU layer. In the encoder path, the feature map downsampling steps use $2 \times 2 \times 2$ average pooling with a stride of 2. In the decoder path, the upsampling steps use transposed convolution with $2 \times 2 \times 2$ filters and half the number of filters than the previous step. The final layer uses a 1x1x1 convolution filter to map each 16-component feature vector to a three-dimensional displacement.

### 4.3.4 Multilevel Architecture

In conventional image registration, multilevel continuation has been proven very efficient to avoid local minima, to reduce topological changes or foldings, and to speed up runtimes [11, 14, 62, 67]. Recent deep-learning-based approaches [29, 34, 80, 91, 92] have shown that, besides carrying over these properties, a multilevel scheme helps overcome the limitations of deep-learning-based registration approaches to properly deal with small and local deformations.

As in our previous work [34], we follow the ideas of standard multilevel registration and compute coarse grid solutions that are prolongated and refined on the subsequent finer level. Our multilevel framework is illustrated in Figure 4.3 with $L = 3$ levels. The registration starts on the coarsest level L where the deformation $\tilde{y}_L$ is computed from the input images that have been Gaussian-smoothed and downsampled by a factor of $2^{L-1}$. On all finer levels $\ell < L$, we incorporate the deformations from all preceding coarse levels as an initial guess by combining them by functional composition and

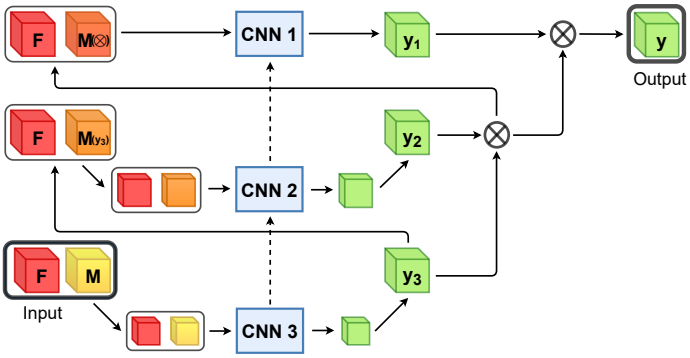



Figure 4.3: Overall scheme of the proposed multilevel framework, where $\mathcal{F}$ indicates the fixed image, $\mathcal{M}$ the moving image, y the deformation field and $\mathcal{M}(y)$ the warped image. Each CNNs is trained separately for a fixed amount of epochs and the weights stay fixed afterwards. The deformation fields from all preceding coarse levels are used as an initial guess by combining them by functional composition and warp the moving image on the highest resolution. Subsequently, the warped moving image is downsampled to the current image level. The dotted lines illustrate the initialization of the network weights with the learned parameters of the previous level.

warping the moving image. Subsequently, the fixed and warped moving images are downsampled.

The number of used levels is a hyper parameter which should be chosen depending on the task and the used data. The maximal number of levels that can be used is limited by the GPU memory and the image size. Since the images are downsampled with a factor of two in the multilevel setting and additionally the image features are downsampled three-times in the U-Net, the number may be chosen at most so that the image size is divisible by $2^{3+(L-1)}$. Our experiments (c.f. section 4.4.9) have shown that a three-level scheme works best in our application and fits on a 12GB GPU. In our experiments, we use in particular a three-level scheme (L = 3, Figure 4.1). The three networks are trained progressively. First, the network on the coarsest level is trained for a fixed amount of epochs. Afterwards, the parameters of the middle network are learned while the coarsest network stays fixed and is only used to produce the initial deformation field. The same procedure is repeated on the finest level. The same architecture is used on all levels. The network parameters on the coarsest level are initialized with Xavier uniform [68], whereas all other networks are initialized with the learned parameters of the previous network. Note that the receptive field in voxels is the same for all networks, however, due to the decreased resolution on the coarse levels, the receptive field in mm is much larger.

## 4.4 Experiments

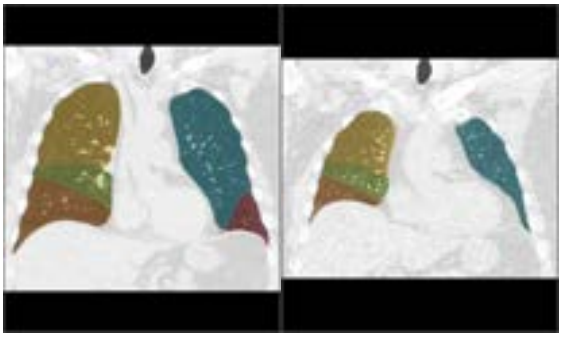
We perform several experiments to assess the accuracy, plausibility, robustness, transferability, and speed of our weakly-supervised deep-learning-based registration approach.

### 4.4.1 Data

We train and validate our method on the data from the COPDGene study [69]. To prove the robustness and transferability of our method and to compare our method with other registration approaches, we evaluate our registration approach on the publicly available DIR-Lab dataset [64, 81] and on the EMPIRE10 challenge as well. On the COPDGene dataset, the evaluation is based on the lobe segmentation masks, and on both of the other datasets, annotated landmarks are available on which we evaluate the target registration error.

**COPDGENE DATASET** Training, validation, and testing data were acquired from the COPDGene study, a large multi-center clinical trial with over 10,000 subjects with chronic obstructive pulmonary disease (COPD) [69]. The COPDGene study includes clinical information, blood samples, and chest CT scans. The image dataset was acquired across 21 imaging centers using a variety of scanner makes and models. Each patient had received two breath-hold 3D CT scans, one on full inspiration (200mAs) and one at the end of normal expiration (50mAs). About five years later, follow-up images were acquired from about 6000 subjects. In our study, we use the inspiration and expiration scans of 1000 patients. We split these patients into 750, 50, 200 patients for training, validation, and testing, respectively. The original images have sizes in the range of $512 \times 512 \times \{341, \cdots, 974\}$ voxels. The in-plane resolution of the axial slices varied between 0.5mm to 0.97mm per voxel with a slice thickness of 0.45mm to 0.7mm. The human lungs are sub dived into five lobes that are separated by visceral pleura called pulmonary fissure. An exemplary inspiration scan and expiration scan of one patient with the lobe segmentation overlay is shown in Figure 4.4. For all scans segmentations of the lobes are available, which were computed automatically and manually corrected and verified by trained human analysts.

**DIR-LAB CHALLENGE** This dataset consists of ten thoracic 4D CT images acquired as part of the radiotherapy planning process for the treatment of thoracic malignancies. In our study we are only using the inspiration and expiration phase of the 4D image, i.e., two of the ten images per 4D scan. The in-plane resolution of the $512 \times 512$ axial slices varied between 0.97mm to 1.16mm per voxel with a slice thickness of 2.5mm. Each scan pair contains 300 manually annotated corresponding landmarks in the lung



a) Inspiration b) Expiration

Figure 4.4: The Image shows a) an inspiration scan and b) and expiration scan of the lungs subdivided into ■ upper left lobe, ■ lower left lobe, ■ upper right lobe, ■ lower right lobe and ■ middle right lobe.

on which we evaluate the target registration error.

**EMPIRE10 CHALLENGE** The EMPIRE10 challenge [82] consists of 30 scan pairs from six different categories: breathhold inspiration scan pairs, breathhold inspiration and expiration scan pairs, 4D data scan pairs, ovine data scan pairs, contrast-noncontrast scan pairs and artificially warped scan pairs. Further information on each category can be found in the challenge paper [82]. Each scan pair contains 100 annotated corresponding landmarks.

### 4.4.2 Preprocessing

In this work, we focus on non-rigid, non-linear deformations and for that reason we perform a linear prealignment of fixed and moving image as preprocessing. For all methods, the same preprocessing is used. We subsequently warp and resample the moving image on the field of view and resolution of the fixed image, which yields a pre-registered moving image $\hat{\mathcal{M}}$. Lung regions are automatically cropped for each CT and resized to volumes of dimension $192 \times 160 \times 192$ as the network input. However, although the deformation field is computed from low-resolution input, during inference, the output deformation field is up-sampled to the original image resolution using trilinear interpolation and the overall evaluation is performed at full resolution of the original images. We do not perform any further preprocessing like normalization on the images, because the CT images are already in a standardized range (Hounsfield units). On the training data, we use the keypoint detection algorithm described in [65] to automatically compute keypoints inside the lung. These keypoints can be considered noisy labels with residual errors of 1-2mm.

### 4.4.3 Implementation Details

We implemente our method in PyTorch. Each network was trained for 25 epochs on an NVIDIA Titan Xp using an ADAM optimizer with a learning rate of $10^{-3}$. The training of all three networks takes about 20 hours. We empirically chose the loss weighting parameters $\alpha = 10$, $\beta = 1$, $\gamma = 0.01$. For the coarsest level, the keypoint weighting parameter $\delta$ was set to zero such that the network can focus on the coarse alignment of larger structures. In the subsequent levels, we chose $\delta = 10^7$. For the edge parameter of the NGF distance measure, we chose $\epsilon = 1$.

### 4.4.4 Accuracy

We evaluate our method by using the propagated lobe segmentation and the fixed lobe segmentation. If a deformation field represents accurate correspondences, the lobe segmentation of the fixed image $b_{\mathcal{F}}$ and the warped lobe segmentation of the moving image $b_{\mathcal{M}}(y)$ should overlap well. In contrast to a lung segmentation overlap, the lobe segmentation overlap provides information about inner lung structures. A good alignment of the lobes was shown to be indicative of good alignment of the fissures, which the evaluation of registration quality in [82] has shown to be indicative of the overall performance of different registration approaches.
We measure the overlap of the lobes with the Dice coefficient

$$\mathrm{DSC} = \frac{|X \cap Y|}{|X| + |Y|}$$

where X is the propagated segmentation $b_{\mathcal{M}}(y)$ and Y is the segmentation of the fixed image $b_{\mathcal{F}}$. Moreover, we evaluate the average symmetric surface distance

$$\mathrm{ASD} = \frac{1}{|X_s| + |Y_s|}\left(\sum_{x \in X_S} d(x, Y_S) + \sum_{y \in Y_S} d(y, X_S)\right),$$

where d is the surface distance

$$d(x, Y_s) = \min_{y \in Y} d(x, y)$$

where x and y are points on the surface of the propagated segmentation surface $X_s$ and the fixed segmentation surface $Y_s$. Additionally, we calculate the symmetric Hausdorff distance

$$\mathrm{HD} = \max\{d_H(X_s, Y_S), d_H(Y_s, X_s)\},$$

where

$$d_H(X_s, Y_s) = \max_{x \in X_s} \min_{y \in Y_s} d(x, y).$$

We compare our proposed method to the conventional approach of [65] that is currently ranked first in the EMPIRE10 challenge [82] (https://empire10.grand-challenge.org/Home/). This method performs a discrete keypoint detection and matching which are integrated into a dense continuous optimization framework. Additionally to the keypoint penalty, the method uses an NGF distance measure, curvature regularizer, a volume change penalty, and a mask alignment of the lung segmentations. Note that the lung segmentation has to be available for each pair of images to be registered. This is in contrast to our method, which also uses a boundary loss (equation 4.7), but this requires the masks to be only available during training, not during testing.

### 4.4.5 Robustness

To analyze the robustness of our method, we evaluate the 30% lowest Dice Scores of all cases. This gives a good overview of the hardest cases and how good our method can register those.

### 4.4.6 Plausibility of the Deformation Field

Besides accurately and robustly transferring anatomical annotations, medical image registration should also provide plausible deformations and therefore should not generate deformations with foldings. Hence, we evaluate the Jacobian determinant as it is a local measure for volume change and in particular for (local) change of topology. If $\det(\nabla y) > 1$ a volume expansion occurred and if $\det(\nabla y) < 1$ the volume decreased and for $\det(\nabla y) \leq 0$ there is a folding.

### 4.4.7 Applicability

In a clinical setting, the registration of two scan pairs has to be available quickly in order not to slow down the routine workflow. In other situations such as screenings, the large number of required registration demands efficient deformable image registration methods. In both cases, the runtime of the algorithms is a crucial factor. For the conventional registration method, we measured the time of the registration without the time needed to load and warp the images. For the network, we measure the time of one forward pass through the cascade of networks. Both measurements were run on the same system with an Intel(R) Core(TM) i7-770K CPU and an Nvidia Titan XP GPU.

### 4.4.8 Transferability and Comparison to state-of-the-art

To show the transferability of our method to other datasets and to compare our method to other registration methods, we apply our trained network as-is to the ten images pairs of the DIR-Lab 4DCT. To evaluate the registration accuracy, the target registration error of the landmarks was computed.Moreover, we evaluate the impact of the dataset used

| | Rühaak [65] | ours |
|---|---|---|
| Dice | 0.92 ± 0.05*** | 0.95 ± 0.03 |
| ASD | 1.97 ± 1.24mm*** | 1.72 ± 0.89mm |
| HD | 27.24 ± 13.70mm* | 26.84 ± 14.27mm |
| Dice30 | 0.86 ± 0.03 *** | 0.93 ± 0.01 |
| Foldings | 0% | 0.06% |
| Runtime CPU | 160s *** | 32s |
| Runtime GPU | - | 0.75s |
| GPU memory | - | 4GB |

Table 4.1: Registration results of [65] and our method on the COPDGene dataset. We performed a one-sited Wilcoxon signed-rank test to test if improvements to the method of [65] are statistically significance. Significance levels are defined as * $p < 0.05$, ** $p< 0.01$ and *** $p< 0.001$.

to train the registration network. Therefore, we train the widely used Voxelmorph [33] framework using the COPDGene data. We adapt the public implementation slightly by choosing a higher regularization weight ($\lambda = 2$) to obtain smooth deformation fields Furthermore, we applied our trained model on the 30 scan pairs of the EMPIRE10 challenge and submitted the displacement fields to the organizers who performed the evaluation which includes a lung boundaries, fissures, landmarks and singularities (foldings).

### 4.4.9 Ablation Study

In an ablation study, we study the impact of the components of the proposed method. We investigate the influence of the number of levels in the multilevel framework on the accuracy and plausibility of the registration results. For all experiments, the overall number of epochs was 75. Furthermore, we explore the additional penalty terms in our loss function by setting the weighting parameters in the loss function to zero and compare it with the proposed loss function.

## 4.5 Results

The results of our experiments on the COPDGene dataset are summarized in Table 4.1. We performed a one-sided Wilcoxon signed-rank test that show that the improvement to the method of [65] is statistically significant for the Dice score, average surface distance (ASD) and Hausdorff distance(HD) and the runtime on the CPU. In the following subsections, we describe the results of each experiment in more detail.

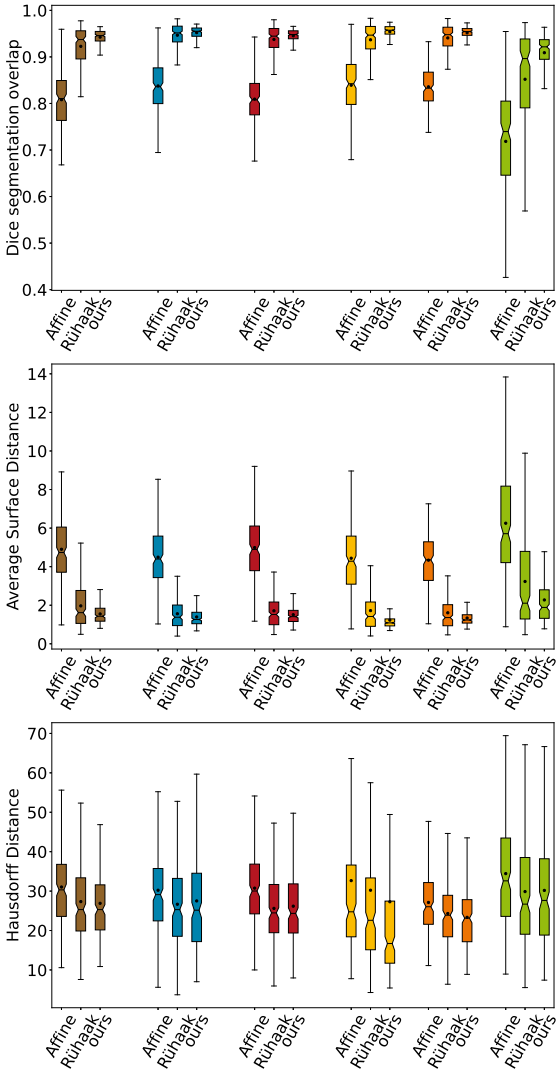



Figure 4.5: Comparison of the Dice overlaps, average surface distance and Hausdorff distance for all test images and each anatomical label (■ average of all labels, ■ upper left lobe, ■ lower left lobe, ■ upper right lobe, ■ lower right lobe and ■ middle right lobe). For each one the distributions of the Dice coefficients after affine pre-alignment, after conventional method of [65] and after our proposed registration are shown.

### 4.5.1 Accuracy

Our proposed method achieves on average significant better Dice Scores than the conventional registration method (0.95 vs. 0.92) with a smaller standard deviation (0.026 vs. 0.046). Also for the symmetric average surface and the Hausdorff distance our method achieves better results (1.72 ± 0.89mm vs. 1.97±1.24 and 26.84 ± 14.27 vs. 27.24 ± 13.70mm, respectively). The distribution of the Dice Scores, of the average surface distance, and of the Hausdorff Distance of both methods are shown in Figures 4.5.

### 4.5.2 Robustness

On the 30% of cases with the lowest Dice Scores, our method achieves an average Dice Score of 0.93 ± 0.01 within a range of $[0.85, 0.94]$. Compared to the method of [65] with an average Dice Score of 0.86 ± 0.03 within a range of $[0.78, 0.90]$, our method propagates the lobes more robustly.

### 4.5.3 Plausibility

For our proposed methods, on average fewer than 0.1% voxel positions of the deformation field showed a negative Jacobian determinant and therefore a folding. The full elimination of foldings as in the conventional registration methods of [65] is not guaranteed. However, the percentage of foldings is within acceptable ranges. Figure 4.6 shows four exemplary Jacobian determinant colormaps overlaid on the fixed image. The volume changes are smooth and mostly around 1 (yellow overlay). Due to large motion in the upper right case, some foldings (dark red overlay) occur on the left inferior border.

### 4.5.4 Applicability

The proposed method needs for registration of an image pair on average 0.75 seconds when executed on a GPU and 32 seconds on the CPU. In contrast, the conventional method takes on average 160 seconds executed on a CPU. Moreover, for the execution, only 4GB of GPU memory are required and therefore our method could also be used on standard computers with less powerful GPUs. The prediction is instantaneous and requires no further manual tuning of parameters. This makes our proposed method very applicable for clinical tasks.

### 4.5.5 Qualitative Results

To illustrate the registration results, we show the difference images $\mathcal{F} - \mathcal{M}(\mathrm{y})$ of four exemplary cases in Figure 4.8. In all cases, the respiratory motion was successfully recovered and most inner structures are well aligned. The first row shows one example

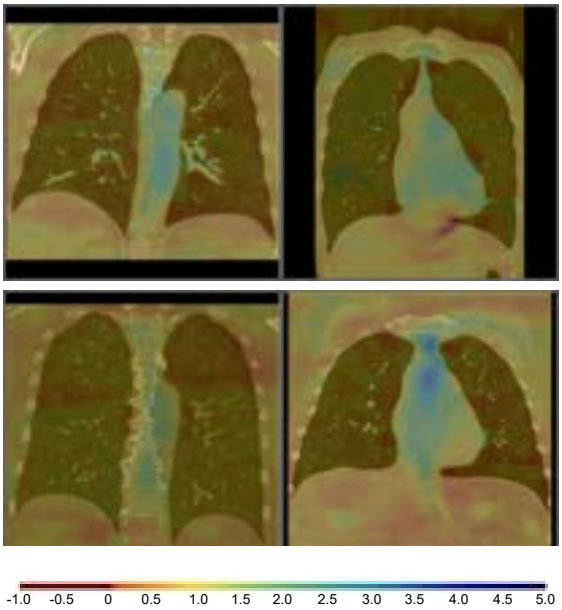



Figure 4.6: Example Jacobian determinant colormaps overlaid on coronal slices of the fixed images. The volume changes are smooth and mostly around 1 (yellow overlay). Due to large motion in the upper right case, some foldings (dark red overlay) occur on the left inferior border.

of a more accurately registered image pairs in terms of the average Dice Score (after affine: 0.85, after: 0.96) and keypoint distance (after affine: 8.51mm, after: 0.99mm). The last row shows the worse case regarding the Dice Score. In this case, the average Dice Score improved from 0.69 to 0.85 and the keypoint distance could be reduced from 13.57mm to 1.9mm, showing also for the cases with large deformations, our registration methods works robustly. Even with masking the distance measure only to the region inside the lung, the surrounding tissue is mostly well aligned. During training, the model learned to align edges and because no lung mask is given during inference, it also aligns edges outside the lung.

### 4.5.6 Ablation Study

We provide an ablation study to further verify the efficiency of proposed components of our method. Results of this ablation experiment on the COPDGene data are presented in Table 4.2. The multi-level experiment shows that increasing the number of level from L = 1 to L = 2 and L = 3 results in a increasing Dice Score from 0.927 to 0.939 to 0.946, a decreasing TRE fron 3.95mm to 2.22mm to 2.00mm, and decreasing number of foldings from 0.1% to 0.09% to 0.06%. The mask alignment loss not only improve the alignment of the pulmonary lobes resulting in a higher Dice Score (0.93 vs 0.946) but also enhance the TRE from 2.16mm to 2.00mm. By integrating our volume change loss, the percentage of foldings can be reduced from 0.3% to 0.06%. Furthermore, it also improve the TRE from 2.16mm to 2.00mm. To further enhance the alignment of smaller structures as vessels and smaller airways, we incorporate keypoint correspondences into the loss function. This decrease the TRE from 4.59mm to 2.00mm. However, the percentage of foldings slightly increase from 0% to 0.06%. Figure 4.7 shows a comparison of the target registration errors of the DIR-Lab 4DCT dataset of all compared settings and after affine registration and the initial errors.

### 4.5.7 Transferability and Comparison to state-of-the-art

In Table 4.4, quantitative results on the DIR-Lab 4DCT dataset of deep-learning-based and conventional registration methods are reported. On average the target registration error (TRE) of our method was 1.14 ± 0.76mm and is therefore better as the currently best deep-learning-based method of [80]. In cases 6, 8, and 10 which have a large initial landmark error, our method achieves much better registration results. Training Voxelmorph on the large COPDGene dataset results in a lower TRE than when trained by leave-one-out on the DIR-Lab dataset (1.71mm vs 3.65mm). The best conventional registration method of [65] has still a lower TRE, however, it needs about 5 minutes to compute the deformation field, whereas our method only needs less than a second. A detailed evaluation of all ten cases for different deep-learning-based registration methods is given in Table 4.5. On the EMPIRE10 challenge data, our method achieves a target registration error of 1.01mm on all cases and a TRE of 0.91mm if ovine data is

| | no mask align. $\beta = 0$ | no VCC $\gamma = 0$ | no keypoint loss $\delta = 0$ | single Level L = 1 | 2-Level L = 2 | **proposed settings** |
|---|---|---|---|---|---|---|
| Dice | 0.93 ± 0.02 *** | 0.95 ± 0.02 | 0.95 ± 0.02 | 0.93 ± 0.02 *** | 0.94 ± 0.02 | 0.95 ± 0.03 |
| TRE KP [mm] | 2.23 ± 1.45 *** | 2.16 ± 1.34 *** | 4.59 ± 2.70 *** | 3.95 ± 1.98 *** | 2.22 ± 1.43 *** | 2.00 ± 1.28 |
| Foldings | 0.04 ± 0.06% | 0.30 ± 0.17% *** | 0.00 ± 0.00% | 0.10 ± 0.14%** | 0.09 ± 0.09% ** | 0.06 ± 0.03% |
| TRE 4DCT [mm] | 1.26 ± 0.82 *** | 1.22 ± 0.84 *** | 1.72 ± 2.31*** | 1.76 ± 1.11*** | 1.26 ± 0.81*** | 1.14 ± 0.76 |

Table 4.2: Results of the ablation study. To demonstrate the impact of the each loss function term, each penalty weight was set to zero once while the remaining parameters were fixed to their empirically determined optimal values. The registration performance is evaluated using the Dice score, the target registration error of the keypoint (TRE KP), and the percentage of foldings on the COPDGene dataset. Moreover the target registration error on the DIR-Lab dataset is compared. We performed a one-sided Wilcoxon signed-rank test to test if improvements to all other settings are statistically significant. We used a Bonferroni correction due to multiples testing. Significance levels are defined as * $p < 0.05$, ** $p < 0.01$ and *** $p < 0.001$.

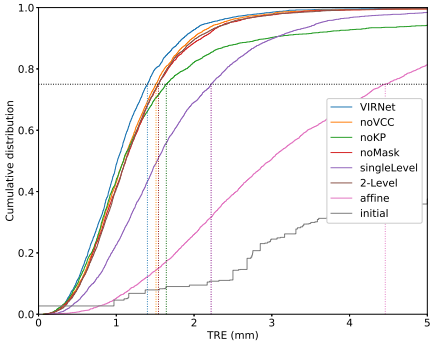



Figure 4.7: Cumulative distribution of target registration error (TRE) in millimeters for all variations of our loss function, after affine registration and initially on all landmark pairs of the DIR-Lab 4DCT dataset. In addition, the dotted lines visualize the 75th percentiles of the TRE, which are 1.40mm (our VIRNet), 1.51mm (noVCC), 1.64mm (noKP), 1.54mm (noMask), 2.22mm (singleLevel), 1.54mm (twoLevel), 4.46mm (affine), 12.55mm (initial)

| | Lung B. | Fissures | Landmarks | Singularities |
|---|---|---|---|---|
| Rühaak | 0.00 | 0.09 | 0.63 | 0.00 |
| ours | 0.07 | 0.09 | 1.01 | 0.02 |

Table 4.3: Results of the EMPIRE10 challenge for the method of [65] and our proposed method. The average score over all 30 cases for the lung boundaries, fissure alignment, landmark error and singularities is shown. Detailed results can be found on the challenge website.

excluded. A summary of the results is shown in Table 4.3 and a more detailed evaluation on the challenge website [1].

## 4.6 Discussion

This paper presents a coarse-to-fine multilevel framework for deep-learning-based image registration that can compensate for and handle large deformations using computing deformation fields on different scales. Our method shares many elements with the conventional registration method of [65]. We have identified key strategies of this method and successfully developed a deep-learning counterpart. The advantage of our deep learning approach is that the expensive annotation and detection of the lobe masks and keypoints is only necessary as training data. This important knowledge is then embedded in our model and therefore the inference is cheap and fast.

[1] `https://empire10.grand-challenge.org/mevis_virnet/`

| | Method | mean TRE (mm) | Foldings | Runtime |
|---|---|---|---|---|
| | initial | 8.46 ± 6.58 | - | - |
| **Conventional** | Schmidt [94] | 1.38 ± 0.87 | - | 83min |
| | Deeds [56] | 1.6 ± 1.7 | 0% | 20min |
| | MRF [70] | 1.43 ± 1.3 | - | 7.97min |
| | Berendsen [95] | 1.36 ± 0.99 | 0% | - |
| | Rühaak [65] | 0.94 ± 1.06 | 0% | 5min |
| **Deep Learning** | Sentker [26] | 2.5 ± 1.16 | - | few seconds |
| | Voxelmorph [33]* | 3.65 ± 2.47 | - | - |
| | Voxelmorph [33]** | 1.71 ± 2.86 | - | - |
| | de Vos [29] | 2.64 ± 4.32 | - | 0.63s |
| | Eppenhof [61] | 2.43 ± 1.81 | 0.42% | 0.56s |
| | mlvirnet [34] | 2.19 ± 1.62 | - | - |
| | Hansen [93] | 1.97 ± 1.42 | - | - |
| | Jiang [91] | 1.66 ± 1.44 | < 0.1% | 1.4s |
| | LungRegNet [80] | 1.59 ± 1.58 | - | 1min |
| | GraphNet [96] | 1.39 ± 1.29 | 0.02% | 2s |
| | **ours** | 1.14 ± 0.76 | < 0.0005% | 0.75s |

Table 4.4: Target registration error values for different conventional and deep learning-based methods on DIR-Lab 4D-CT dataset. All results were extracted from the original papers, besides Voxelmorph* which was reported in [93] and Voxelmorph** which we trained on the COPDGene data. Since the runtime was not measured with the same hardware, it cannot directly be compared. However, it gives an impression of the speed.

| Scan | Initial | Eppenhof | DLIR | mlVIRNet | LungRegNet | GraphRegNet | Voxelmorph | **ours** |
|---|---|---|---|---|---|---|---|---|
| 4DCT 01 | 3.89 ± 2.78 | 2.18 ± 1.05 | 1.27 ± 1.16 | 1.33 ± 0.73 | 0.98 ± 0.54 | 0.86 ± 0.91 | 1.03 ± 1.01 | 0.99 ± 0.47 |
| 4DCT 02 | 4.34 ± 3.90 | 2.06 ± 0.96 | 1.20 ± 1.12 | 1.33 ± 0.69 | 0.98 ± 0.52 | 0.90 ± 0.95 | 1.09 ± 1.87 | 0.98 ± 0.46 |
| 4DCT 03 | 6.94 ± 4.05 | 2.11 ± 1.04 | 1.48 ± 1.26 | 1.48 ± 0.94 | 1.14 ± 0.64 | 1.06 ± 1.10 | 1.40 ± 2.04 | 1.11 ± 0.61 |
| 4DCT 04 | 9.83 ± 4.85 | 3.13 ± 1.60 | 2.09 ± 1.93 | 1.85 ± 1.37 | 1.39 ± 0.99 | 1.45 ± 1.24 | 1.69 ± 2.60 | 1.37 ± 1.03 |
| 4DCT 05 | 7.48 ± 5.50 | 2.92 ± 1.70 | 1.95 ± 2.10 | 1.84 ± 1.39 | 1.43 ± 1.31 | 1.60 ± 1.50 | 1.63 ± 2.44 | 1.32 ± 1.36 |
| 4DCT 06 | 10.89 ± 6.96 | 4.20 ± 2.00 | 5.16 ± 7.09 | 3.57 ± 2.15 | 2.26 ± 2.93* | 1.59 ± 1.06 | 1.60 ± 2.58 | 1.15 ± 1.12 |
| 4DCT 07 | 11.03 ± 7.42 | 4.12 ± 2.97 | 3.05 ± 3.01 | 2.61 ± 1.63 | 1.42 ± 1.16* | 1.74 ± 1.10 | 1.93 ± 2.8 | 1.05 ± 0.81 |
| 4DCT 08 | 14.99 ± 9.00 | 9.43 ± 6.28 | 6.48 ± 5.37 | 2.62 ± 1.52 | 3.13 ± 3.37* | 1.46 ± 1.27 | 3.16 ± 4.69 | 1.22 ± 1.44 |
| 4DCT 09 | 7.92 ± 3.97 | 3.82 ± 1.69 | 2.10 ± 1.66 | 2.70 ± 1.46 | 1.27 ± 0.94 | 1.58 ± 1.07 | 1.95 ± 2.37 | 1.11 ± 0.66 |
| 4DCT 10 | 7.30 ± 6.34 | 2.87 ± 1.96 | 2.09 ± 2.24 | 2.63 ± 1.93 | 1.93 ± 3.06 | 1.71 ± 2.03 | 1.66 ± 2.87 | 1.05 ± 0.72 |
| Mean | 8.46 ± 6.58 | 3.68 ± 3.32 | 2.64 ± 4.32 | 2.19 ± 1.62 | 1.59 ± 1.58 | 1.39 ± 1.29 | 1.71 ± 2.86 | 1.14 ± 0.76 |

Table 4.5: Mean (standard deviation) of the registration error in mm determined on DIR-Lab 4DCT data for several deep-learning-based registration methods: [61], [29], [34], [80], [96] (standard deviations were reported directly by the authors and not included in the their paper) and [33] (trained on the COPDGene data). Asterisk indicates that the FineNet was performed twice for this case and method.

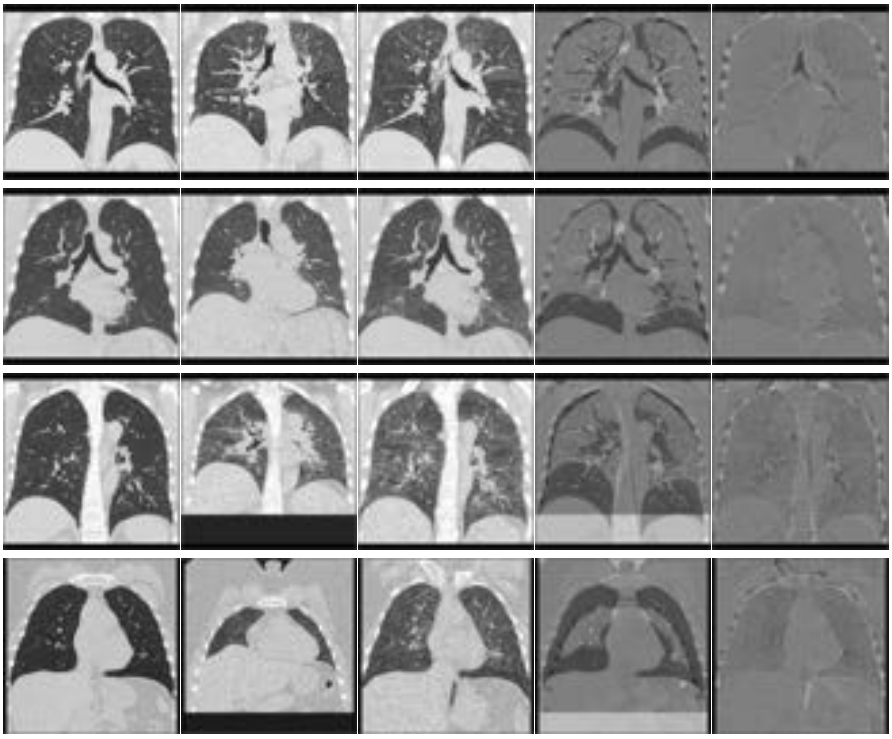



Figure 4.8: Example coronal slices extracted from four exemplary cases. Input images $\mathcal{F}$ and $\mathcal{M}$, the warped moving image $\mathcal{M}(\mathrm{y})$, the difference image $\mathcal{F} - \mathcal{M}$ (fourth column) and the difference image $\mathcal{F} - \mathcal{M}(\mathrm{y})$ after registration with the proposed method (fifth column). In all cases the respiratory motion was successfully recovered and most inner structures are well aligned. Due to altered density of lung tissue during breathing, intensity changes occur and therefore higher values in the difference images are reached without registration errors.

We employ a Gaussian-pyramid-based multilevel framework that can solve the image registration optimization in a coarse-to-fine fashion. To prevent foldings of the deformation field and restrict the determinant of the Jacobian to physiologically meaningful values, we combine the curvature regularizer with a volume change penalty in the loss function. Furthermore, we also integrate weak keypoint correspondences into the loss function to focus more on the alignment of smaller structures. The keypoints are computed automatically and can be considered as noisy labels with residual errors of 1 – 2mm. However, we showed that the use of these noisy labels is nevertheless advantageous and leads to a better alignment of vessels and smaller airways and therefore also results in a better target registration error on the DIR-Lab dataset.

We validated our framework on the challenging task of large motion inspiration-to-expiration registration using image data from the multi-center COPDGene study. To assess the accuracy of our network, we performed an extensive evaluation of 200 pulmonary CT scan pairs from the large-scale COPDGene study and demonstrated that our method can perform accurate registration between two affine pre-aligned images. Especially for the task of lobe propagation, we could show that our method performs better than conventional approaches. It achieves higher Dice scores and lower surface and Hausdorff distances (0.95, 1.72 mm, and 26.8 mm) compared to conventional registration (0.92, 1.97 mm, and 27.2 mm, respectively). This better performance can be explained by the use of the mask-alignment loss. As demonstrated in previous studies (e.g. [33, 35]), the combination of the complementary strength of global semantic information (weakly-supervised learning with segmentation labels) and local distance metrics improves the registration performance during inference. In contrast to conventional registration methods, such additional information only needs to be available in the training dataset.

Furthermore, we have evaluated the proposed method using the DIR-Lab and EMPIRE10 dataset and showed that we achieve excellent TRE of 1.14 mm and 1.01 mm, respectively. Note that our network was not trained on those datasets. This is strong evidence that our network can generalize well. Although previous works (e.g. [26, 29, 61, 91]) contribute much to improving the registration accuracy, there is still a misalignment of smaller structures, which leads to a high TRE. To focus more on the alignment of vessels, [80] introduced a preprocessing step to enhance the vessels in the input images by segmenting vascular structures and increasing the intensity value inside the vessel mask. In their paper, they demonstrated the efficiency of this preprocessing step. Since this step is performed on the input images, it is also required during application. To avoid this problem and thus not increase the execution time, we integrate additional information on smaller structures using the keypoint loss. The advantage of this procedure is that the keypoints, as well as the masks of the boundary loss, are only needed during training. Nevertheless, the best conventional registration methods still achieve lower TRE than our method. One reason for this might be that convectional registration methods mostly work on the original image data. In contrast, for the deep-learning approaches, the input images have to be downsampled due to memory restrictions on the GPU. Especially for smaller structures and small errors (we are speaking about a TRE difference of 0.2-0.4mm), it is easily imaginable that this resolution is not high enough. Moreover, the training data used also influence the performance. Our network was trained on inspiration-expiration scan pairs from humans. In the EMPIRE10 dataset, a variety of lung registration tasks including ovine lung registration has to be performed. Although our method does not register the ovine data perfectly, we achieve a TRE of 1.69 mm on the ovine data which shows that our method is capable of generalizing well. We would assume that with a wider variety in training data, the performance of deep-learning-based registration methods

can further improve. We showed this effect when training the Voxelmorph network. By using the larger COPDGene dataset to train the Voxelmorph network, the target registration error on the DIR-Lab dataset improved from 3.65 mm to 1.71 mm compared to a leave-one-out training on the DIR-Lab dataset. This illustrates the large impact of the training dataset. Since Voxelmorph and our framework are very similar, this experiment also shows that the addition of more loss functions and a multilevel approach is beneficial.

Besides accurately transferring anatomical annotations, medical image registration should also provide plausible transformations and therefore should not generate deformations with foldings. In conventional registration methods, this is achieved by using a regularizer in the cost function. Recently deep-learning-based methods like [61] and [91] also integrated a regularizer into their loss functions to enforce a smooth deformation field resulting in an acceptable amount of foldings (0.42% and 0.1% of foldings). In our work, we additionally use a volume change control which penalizes occurring foldings more severely than the regularizer does, resulting in on average fewer than 0.1% and 0.0005% voxel positions in the deformation field with folding on the COPDGene and DIR-Lab dataset, respectively. Without the volume change control penalty, the deformation fields produced by our method show on average 0.30% of voxel positions with foldings, which is comparable to the values of other deep learning registration methods. This shows that the addition of the volume change control mitigated the occurrence of foldings. The higher number of foldings on the COPDGene dataset can be explained by the noise difference between the expiration and inspiration scan due to different doses during acquisition (see Fig 4.8 for some example images). The full elimination of foldings as in some conventional registration methods is not guaranteed. Another way to reduce the number of foldings was presented in the works of [97], [31], and [98] who are using the scaling and squaring algorithm [99] to integrate the predicted stationary velocity field. With a sufficient number of integration steps, these methods should theoretically guarantee diffeomorphic transformation. However, in the presented works they reported ”nearly no non-negative Jacobian voxels” [97] and 0.023% to 0.151% of voxels with a negative Jacobian determinant [98]. As discussed in [98], this has two major factors. First, the velocity field could be not sufficiently smooth. This can be solved by increasing the regularization weight. However, this often yields a drop in the registration accuracy. Secondly, the number of chosen integration steps was too small. Increasing this can reduce the number of foldings which occur but increase the computational cost as well. In summary, the scaling and squaring approach and the volume-change-control penalty presented achieve similar results preventing foldings. Besides, our approach regulates volume changes.

In our experiments, we focused on the complex task of CT lung registration, as the registration results can be evaluated more accurately than only with an overlap of a

larger structure. However, our method could also be trained for a different task or on a different modality. Except for keypoint detection, no component is lung-specific and the keypoint loss can be used with landmarks in different organs.
In future studies, we will investigate the impact of instance optimization to fine-tune the deformation field for those image pairs for which the registration result is not yet satisfactory.

## 4.7 Conclusion

This paper presents a deep-learning-based registration approach for deformable image registration, targeting in particular the challenging task of lung registration. We introduce a keypoint matching term and a volume change penalty to increase the alignment of smaller structures and to prevent foldings and restrict the deformation field to physiologically meaningful values. Our multi-level registration framework equipped with these components achieves state-of-the-art registration accuracy on the COPDGene and DIR-Lab datasets with a very short execution time.

## Acknowledgements

The authors are deeply grateful to Keelin Murphy, Edward Castillo and Richard Castillo for providing evaluation benchmarks.
This work was supported by the German Academic Scholarship Foundation. We gratefully acknowledge the COPDGene Study for providing the data used. COPDGene is funded by Award Number U01 HL089897 and Award Number U01 HL089856 from the National Heart, Lung, and Blood Institute. COPDGene is also supported by the COPD Foundation through contributions made to an Industry Advisory Board comprised of AstraZeneca, Boehringer-Ingelheim, Genentech, GlaxoSmithKline, Novartis, Pfizer, Siemens, and Sunovion. We gratefully acknowledge the support of the NVIDIA Corporation with their GPU donations for this research.

5

# CHAPTER 5

# Learn2Reg: comprehensive multi-task medical image registration challenge

**BASED ON:** A. Hering, L. Hansen, T. C. Mok, A. Chung, H. Siebert, S. Häger, A. Lange, S. Kuckertz, et al. "Learn2Reg: comprehensive multi-task medical image registration challenge, dataset and evaluation in the era of deep learning," *submitted* (2021).

## Abstract

Image registration is a fundamental medical image analysis task, and a wide variety of approaches have been proposed. However, only a few studies have comprehensively compared medical image registration approaches on a wide range of clinically relevant tasks, in part because of the lack of availability of such diverse data. This limits the development of registration methods, the adoption of research advances into practice, and a fair benchmark across competing approaches. The Learn2Reg challenge addresses these limitations by providing a multi-task medical image registration benchmark for comprehensive characterisation of deformable registration algorithms. A continuous evaluation will be possible at `https://learn2reg.grand-challenge.org`. Learn2Reg covers a wide range of anatomies (brain, abdomen, and thorax), modalities (ultrasound, CT, MR), availability of annotations, as well as intra- and inter-patient registration evaluation. We established an easily accessible framework for training and validation of 3D registration methods, which enabled the compilation of results of over 65 individual method submissions from more than 20 unique teams. We used a complementary set of metrics, including robustness, accuracy, plausibility, and runtime, enabling unique insight into the current state-of-the-art of medical image registration. This paper describes datasets, tasks, evaluation methods and results of the challenge, and the results of further analysis of transferability to new datasets, the importance of label supervision, and resulting bias.

## 5.1 Introduction

Image registration is a fundamental task in medical image analysis and has been an active field of research for decades [8, 9, 100, 101]. Most studies that compared registration methods were focused on specific tasks or algorithmic aspects, and did not comprehensively characterise current approaches. With the recent success of deep learning strategies in image analysis, the degree and dependency of algorithms on (partially) labelled training data is often a crucial aspect in current research. The Learn2Reg challenge aims to gain insight into which methodological components and supervision strategies are best suited for a wide range of clinically useful 3D image registration tasks, and sets a new benchmark to evaluate and distinguish strengths and weaknesses of task-tailored solutions. Learn2Reg covers a wide range of anatomies (brain, abdomen and thorax), modalities (ultrasound, CT, MRI, populations) and auxiliary annotations (e.g. segmentation, keypoints). The challenge also includes both intra- and inter-patient registration tasks. Due to this broad range, it serves as a unique benchmark to evaluate the current state-of-the-art with respect to various qualities of registration algorithms: accuracy, robustness, plausibility and speed. Furthermore, no other medical image registration challenge has thoroughly analysed the benefits and shortcomings of learning- and optimisation-based strategies. To engage a wider participation from new research groups, Learn2Reg removes entry barriers by providing pre-processed and pre-aligned images with additional annotations, as well as evaluation scripts and code for all evaluation metrics.

This overview ranks and scores results from over 65 entries from more than 20 teams throughout 2020 and 2021. We perform additional experiments to analyse the robustness towards cross-dataset transfer, the influence of the bias induced by only labelling certain anatomical regions, and direct comparisons of the supervision level of selected methods.

### 5.1.1 Related Work

In the following a brief overview of important related work on comparing (bio)-medical image registration, and its fundamental methodological choices that differentiate the wide range of metrics, optimisation, and supervision is given. General guidelines for setting up a fair and unbiased challenge have been recently thoroughly discussed in literature [102]. These criteria were adhered to in Learn2Reg and externally reviewed and confirmed by the MICCAI challenge team.

**Challenges** There have previously been four prominent challenges for (bio)-medical image registration. Three challenges were single-task focused challenges: EMPIRE10 (lung CT), CuRIOUS (intra-operative US and MRI), and ANHIR (histology). Each attracted at least 10 participating teams and used various metrics for quantifying the

performance. The EMPIRE10 challenge provided the most comprehensive evaluation including distances of manual landmark pairs, fissure segmentations, and Jacobian determinant values of the deformation field. This challenge also required (original) participants to perform live registrations during the MICCAI workshop in Beijing and therefore employed a time constraint on the computations. The Continuous Registration Challenge co-organised with Workshop of Biomedical Image Registration (WBIR) 2018 aimed at combining multiple tasks from previous benchmarks (lung CT and inter-patient brain MRI). It addressed assessing registration quality as a service but is limited to algorithms that can be incorporated into the SimpleElastix framework and therefore had limited participation.

**Benchmark Papers** Several papers have compared multiple registration algorithms for a given dataset. In contrast to challenges, these benchmark papers did not have (at least originally) an open workshop format that enabled wide-spread participation. Nevertheless, their findings provided meaningful insights. Starting from RIRE [103], which compared rigid-body alignment of head MRI (T1, T2), PET and CT, there have been several brain registration benchmarks - most notably the evaluation of 14 nonlinear iterative registration algorithms [104]. Fewer studies analyzed abdominal registration, and included the evaluation of six affine and non-linear algorithms on inter-patient registration of the ”beyond the cranial vault” dataset [105]. This study revealed large performance gaps and motivated our inclusion of this dataset to study the potential benefit of supervised (learning-based) algorithms. The DIR-Lab datasets [81] have been widely used to benchmark intra-patient CT lung motion estimation and provide a leaderboard for state-of-the-art comparison. All landmarks are publicly available, which makes the dataset prone to overfitting on the test data.

**Survey Papers and Baseline Methods** Surveys on conventional medical image registration [9, 100] have comprehensively reviewed typical categories of approaches including similarity metric, regulariser, and optimiser criteria. Due to the strong increase in the number of deep-learning-based registration paper in the last few years, several new surveys have been published (e.g. [101]) extending the typical categories with deep-learning specific categories like supervision-type and network architecture. Moreover, the training data and thus the registered body region and image modality are more important for deep-learning-based methods and get more into the focus of those survey papers. While few papers have evaluated their proposed registration method on more than two different registration tasks, there is a variety of public methods SyN [106], Elastix [16], NiftyReg [20] and deeds [70], and Voxelmorph [107] that are commonly used as baseline or comparison methods. When comparing only among deep-learning based methods simply re-training specific architectures on new data may be insufficient. Hence the use of a challenge benchmark that incorporates several generally applicable baselines is essential for a fair evaluation.

### 5.1.2 Contributions

Learn2Reg provides both datasets and an easily accessible benchmark for the first comprehensive evaluation of a wide-range of methods for inter- and intra-patient, mono- and multimodal medical registration. We introduce a complementary set of metrics, including robustness, accuracy, plausibility and speed, that follows the principles defined by the BIAS group [102] and could become a de-facto benchmark for new algorithms. Further analysis of label bias (for supervised methods), domain transfer and statistical testing of significant differences across algorithms and types of methods highlight the complementary strength and weaknesses of learning vs. non-learning-based approaches.

## 5.2 Material and Methods

### 5.2.1 Challenge Organisation

The Learn2Reg challenge is organised by Alessa Hering (Fraunhofer MEVIS, Germany and Radboudumc Nijmegen, The Netherlands), Lasse Hansen (Institute of Medical Informatics, Universität zu Lübeck, Germany), Adrian Dalca (Computer Science and Artificial Intelligence Lab, MIT, USA) and Mattias Heinrich (Institute of Medical Informatics, Universität zu Lübeck, Germany) and is associated with the MICCAI 2020 and MICCAI 2021. The Learn2Reg challenge consists of two phases (mainly organised on grand-challenge.org).

- Phase 1 - Validation Phase: The participants downloaded the training and validation scan pairs for each task described in section 5.2.2. The participants trained a registration network or tuned hyperparameters on the training scan pairs in their own facilities. The developed algorithms were used to register the scan pairs of the validation dataset. The resulting displacement fields on the validation dataset were submitted and evaluated using grand-challenge.org. Challenge participants were allowed to create five submissions per day to this phase. The results are continuously published on grand-challenge[1].

- Phase 2 - Test phase: Within one week after the test data release, the participants had to send either the generated displacement fields to the organisers or a Docker container containing the algorithm. A Docker submission was preferred and made more attractive by evaluating the runtime of the algorithm.

To remove entry barriers for new participants with expertise in deep learning but not necessarily registration, the organisers provided pre-preprocessed data (resample, crop, pre-align, etc.). A detailed description of the used preprocessing is given in section 5.2.2. Furthermore, the python evaluation code for voxel displacement fields as well

[1] https://learn2reg.grand-challenge.org/evaluation/challenge/leaderboard/

| | CuRIOUS | | Hippocampus MR | | Abdomen CT-CT | |
|---|---|---|---|---|---|---|
| | Fixed | Moving | Fixed | Moving | Fixed | Moving |
| |  |  |  |  |  |  |
| Modalities | MR T1w & FLAIR/US | | MR T1w/MR T1w | | CT/CT | |
| Intra-/Inter-patient | Intra-patient | | Inter-patient | | Inter-patient | |
| Resolution | 256×256×288 | | 64×64×64 | | 192×160×256 | |
| Voxel size | ~0.5×0.5×0.5mm | | 1×1×1mm | | 2×2×2mm | |
| Cases (Train/Test) | 32 (22/10) | | 394 (263/131) | | 50 (30/20) | |
| Preprocessing | resample | | crop/pad/resample | | canonical affine pre-align crop/pad/resample | |
| Annotations | 9-18 landmarks/case | | 2 anatomical labels | | 13 anatomical labels | |
| Additional data | | | | | | |
| Challenges | ● ● ● | | ● | | ● ● | |
| | Abdomen MR-CT | | OASIS | | Lung CT | |
| | fixed | moving | fixed | moving | fixed | moving |
| |  |  |  |  |  |  |
| Modalities | MR T1w / CT | | MR T1w / MR T1w | | CT / CT | |
| Intra-/Inter-patient | Intra-patient | | Inter-patient | | Intra-patient | |
| Resolution | 192×160×192 | | 160×192×224 | | 192×192×208 | |
| Voxel size | 2×2×2mm | | 1×1×1mm | | 1.75×1.25×1.75mm | |
| Cases (Train/Test) | 16 (8/8) | | 455 (416/39) | | 30 (20/10) | |
| Preprocessing | canonical affine pre-align crop/pad/resample | | | | affine pre-align crop/pad/resample | |
| Annotations | 4-9 anatomical labels | | 35 anatomical labels | | 100 landmarks/case | |
| Additional data | 90 unpaired MR/CT scans 8001 masks | | | | lung masks | |
| Challenges | ● ● ● ● | | ● | | ● ● ● ● | |

Table 5.1: Overview of all six Learn2Reg tasks adressing the imminent challenges of medcial image registration: multi-modal scans ●, few/noisy annotations ●, partial visibility ●, small datasets ●, large deformations ●, small structures ●, unsupervised registration ● and missing correspondences ●.

as a example dockerfile were provided. Members of the organizers' institutes could participate in the challenge but were not eligible for awards. A continuous evaluation for validation and test data will be possible at grand-challenge.org[2].

### 5.2.2 Tasks

Learn2Reg consists of six clinically relevant complementary tasks (datasets). Table 1 summarises the dataset details, and we discuss them in detail below.

[2]https://learn2reg.grand-challenge.org
https://learn2reg-test.grand-challenge.org

**CuRIOUS** EASY-RESECT [108] is a simplified sub-set of the original RESECT dataset [109], previously used in the MICCAI CuRIOUS challenges [110]. The dataset contains 22 training and 10 testing subjects with low-grade brain gliomas, intended to help to develop MR vs. US registration algorithms to correct tissue shift in brain tumor resection. For the Learn2Reg challenge, we included T1w and T2-FLAIR MR scans, and spatially tracked intra-operative ultrasound volumes. All scans were acquired for routine clinical care of brain tumor resection procedures at St Olavs University Hospital (Trondheim, Norway). Matching anatomical landmarks were annotated between T2-FLAIR MR and 3D ultrasound volumes [109] to enable evaluation of the registration accuracy. During pre-processing, for each subject, the T1w scan is rigidly registered to the T2-FLAIR scan, and both scans are resampled to the same coordinate space as the 3D ultrasound volume yielding fixed voxel dimensions for all scans (256×256×288) at an isotropic resolution of approximately 0.5 mm.

**Hippocampus MR** This dataset consists of 394 MR scans of the hippocampus region acquired in 90 healthy adults and 105 adults with a non-affective psychotic disorder taken from the Psychiatric Genotype/Phenotype Project data repository at Vanderbilt University Medical Center (VUMC). The hippocampus head and tail were manually traced in all scans. Previous to the Learn2Reg challenge, the dataset was used as part of the Medical Segmentation Decathlon [111].

**Abdomen CT-CT** This task tackles inter-patient registration of abdominal CT scans, which enables statistical modelling of variations of abdominal organs for abnormality detection, and can provide a canonical atlas space for further investigations. The dataset contains 50 abdominal CT scans (30 for training, 20 for testing) with 13 manually labelled anatomical structures: spleen, right kidney, left kidney, gall bladder, esophagus, liver, stomach, aorta, inferior vena cava, portal and splenic vein, pancreas, left adrenal gland and right adrenal gland [105]. The images were registered affinely in a groupwise manner and resampled to the same voxel resolution and spatial dimensions (192×160×256).

**Abdomen MR-CT** The dataset was compiled from public studies of the cancer imaging archive (TCIA) that contained paired scans of both MRI and CT from the same patients. In particular, 16 MRI and CT scans from the following studies, TCGA-KIRC, TCGA-KIRP, and TCGA-LIHC, are included in Learn2Reg - that cover routine diagnostic scans and follow-up imaging for kidney surgery (donation). The data has been reorientated, resampled to an isotropic resolution of 2 mm, and cropped and padded to achieve voxel dimensions of 192x160x192. We have also manually traced 3D segmentation masks for the liver, spleen, left and right kidney. All scans were pre-aligned using a groupwise affine registration based on the deeds-linear algorithm.

Additional unpaired and segmented training data from two further challenges: BCV-CT (see 5.2.2) and CHAOS-MR were provided for pre-training.

**OASIS** The task employed 416 3D whole-brain MR scans from the Open access series of imaging studies (OASIS) [112], a cross-sectional MRI data study with a wide range of participants from young, middle-aged, nondemented, and demented older adults. We performed standard brain MR pre-processing including skull-stripping, normalisation, pre-alignment, and resampling [113]. Semi-automatic labels with manual corrections of 35 cortical and subcortical brain structures were generated using FreeSurfer [114].

**Lung CT** The aim of the lung CT task was the registration of expiration to inspiration CT scans of the lung. The data consists of 20 training [115] and 10 test scan pairs [116]. The scans were acquired at the Department of Radiology at the Radboud University Medical Center, Nijmegen, The Netherlands. All pairs were affinely pre-registered and resampled to an image size of 192×192×208. Lung segmentation masks and keypoints were provided as additional training information.

### 5.2.3 Challenge Design

To provide a comprehensive evaluation of the registration performance, we consider a number of complementary metrics (see section 5.2.3) that assess the accuracy, robustness, plausibility, and speed of the algorithms. For final task ranks, we further consider the significance of differences in results. The detailed ranking scheme is described in section 5.2.3.

#### METRICS

**DSC** The Dice similarity coefficient (DSC) measures the overlap of two sets of segmentation labels (on the fixed and warped moving scan).

**DSC30** To assess robustness, the DSC30 metric considers the 30th percentile in DSC scores over all cases. For the Abdomen CT-MR task, this robustness metric is replaced with a standard DSC on additional anatomical labels, that were not available during training (DSC9).

**HD95** The Hausdorff distance (HD) indicates the maximum distance in a metric space (here: Euclidean space, distance specified in millimeters (mm)) between two sets of surfaces (segmentation labels on the fixed and warped moving scan). For a robust score, we consider the 95th percentile instead of the maximum distance (HD95).

**TRE** The target registration error (TRE) is defined as the euclidean distance (in millimeters (mm)) between corresponding landmarks in the fixed and warped moving scan.

**TRE30** Similar to the DSC30 score the TRE30 metric collects the 30th percentile of largest landmark distances.

**SDlogJ** The plausibility (smoothness) of a displacement field is captured using the standard deviation of the logarithm of the Jacobian determinant (SDlogJ) of the displacement field ([117, 118]).

**RT** In addition, we are able to measure the test-time registration runtime (RT) on the same hardware (CPU: Xeon Silver 4210R, GPU: Quadro RTX 8000), when methods are submitted as a docker container. Start and stop times are the loading of the first scan and writing of the displacement field to disk, respectively.

**RANKING SCHEME** We ranks methods using statistically significantly different results. For each metric applied in a task, methods are compared against each other (Wilcoxon signed rank test with $p<0.05$), ranked based on the number of "won" comparisons and finally mapped to a numerical metric rank score between 0.1 and 1 (with possible score sharing). A task rank score is then obtained as the geometric mean of individual metric rank scores. All methods for which no metric is available (not submitted to the task, no Docker container submitted) share the lowest possible metric rank score of 0.1.

## 5.3 Challenge Entries

In phase 1, performed using the grand-challenge.org platform, 17 teams submitted displacement fields in 2020 and 22 teams in 2021. In phase 2, two teams submitted displacement fields in 2020 and eight teams submitted their algorithms as docker images. In 2021, three teams submitted displacement fields and 12 teams submitted a docker. Only algorithms that participated in both phases of at least one year were included in this paper. Below is a brief description of each of the 21 algorithms. Table 5.2 provides a summary of important information for each algorithm. For a more detailed description of the algorithms, please refer to the respective articles, in the proceedings of the MICCAI Learn2Reg workshops [119].

**3Idiots** [120] employ a hybrid similarity loss consisting of intensity (SSD), statistical (MI), and label-based (Dice + L1) penalties. A Voxelmorph model with an increased number of feature channels and halved output resolution is trained in a patch-wise manner and applied to the OASIS task.

| | Tasks | | | | | | Type | | Objectives | | | | | | | Reg. | | Optimis. | | | Misc. (used architectures, add. objectives, etc.) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | CuRIOUS | Hippocampus MR | Abdomen CT-CT | Abdomen MR-CT | OASIS | Lung CT | Conventional | Deep Learning | NCC | MIND-SSC | NGF | MI | Dice | Keypoints | Consistency (Inv./Cycl.) | Diffusion | Curvature | Adam | Convex | L-BFGS/Gauss-Newton | |
| 3Idiots | | | | | ● | | | ● | | | | ● | ● | | | ● | | ● | | | Voxelmorph; SSD; |
| Bailiang | | | | | ● | | | ● | ● | | | | ● | | | ● | | ● | | | DeepRegNet; |
| ConvexAdam | ■ | ● | ● | ■ | ● | ■ | ○ | ○ | | ○ | | | ○ | | ○ | ● | | ○ | ○ | | U-Net; Dense corr.; |
| corrField | ■ | ■ | ■ | ■ | ■ | ■ | ● | ○ | | ● | | | | | ● | ● | | | | | Dense corr.; |
| Driver | | | | ● | ● | ● | | ● | ● | | | ● | ○ | ○ | | | | ● | | | PCNet; Cross-entropy loss; |
| Epicure | | | | | | ■ | ● | | ● | | | | | | | | | | | | Bending energy regularisation; |
| Estienne | | ● | ● | | | | | ● | ● | | | | ● | | | ● | | ● | | | U-Net; Multi-Task learning; |
| Gunnarsson | ■ | ● | ● | | | ■ | | ● | ● | | | | ● | | | ● | | ● | | | PWC; |
| Imperial | | | | ● | ● | ■ | | ● | | | ● | | ● | ● | | ● | | ● | | | Structure-guided loss; |
| Joutard | | | ● | | | | | ● | | | | | ● | | | ● | | ● | | | U-Net; EDT similarity; Dense corr.; |
| LapIRN | ■ | ● | ● | ● | ● | ■ | ○ | ○ | ● | ○ | | ○ | ○ | | | ● | | ● | | | U-Net; Conditional NN; |
| LaTIM | | | | ■ | ■ | ■ | ● | | | | | | | | | | | | | | Directional representations; |
| Lifshitz | | | | | | ■ | | ● | ● | | | | | | ● | | | | | | Unrolled $L_1$ regulariser; Dense corr.; |
| lWM | | ● | | | ● | | | ● | | ● | | | ● | | ● | ● | | ● | | | 2-stream NN; |
| MEVIS | ■ | ● | ■ | ■ | ■ | ■ | ○ | ○ | | | ● | | ○ | ○ | ● | | ● | | | ○ | |
| Multi-brain | | | | ■ | ■ | ■ | ● | | | | | | | | | ● | | | | ● | Groupwise registration; Bayesian modelling; |
| NiftyReg | ■ | ■ | ■ | ■ | ■ | ■ | ● | | ○ | ○ | | | | | | | | | | | Bending and Jacobian regularisation; |
| PDD-Net | ■ | ● | ● | | | ■ | | ● | | ● | | | ○ | | | ● | | ● | | | Dense Corr.; |
| PIMed | | | ● | ● | ● | ■ | ○ | ● | | | | | ○ | | | | | ○ | | ○ | U-Net; SSTVD similarity; Dense corr.; |
| Thorley | | | | | ● | | ● | | | | | | | | | | | | | | |
| Winter | | | | ■ | ■ | ■ | ○ | ○ | ● | ● | | ● | | | | ● | | ● | | | Voxelmorph; |

Table 5.2: Methodological overview of all Learn2Reg methods. An entry in the table indicates agreement with the corresponding heading. Unsupervised and supervised challenge entries are marked with a ■ and ● symbol in the *Tasks* subgroup. If a challenge entry uses different approaches for different tasks or mixes them within the method (e.g. Deep Learning + Instance Optimisation) we marked the property with a ○ symbol. For detailed descriptions of the methods see Section 5.3 and the associated references.

**Bailiang** [121] addressed OASIS and is based on the DeepRegNet framework from Project-MONAI. The input of the encoder is the concatenation of fixed and moving images. A dense vector field (DVF) is predicted from summing over different level decoders and integrated using scaling and squaring. The loss function is composed of LNCC, MIND-SCC, Dice, and a diffusion regulariser. `https://github.com/BailiangJ/learn2reg2021_task3`

**ConvexAdam** [122] propose a decoupling of deep learning for semantic feature extraction, using an nnUNet, and a very fast and accurate conventional optimisation. They combine a single-level dense discretised displacement correlation with large capture range and convex global optimisation with a local gradient-based instance refinement using the Adam optimiser. The method is applied to all six tasks and uses diffusion regularisation, an inverse-consistency constraint, and MIND similarity, which is applicable for multi-modal and same-modality intra-patient alignment. The method extends the input features to learned label-supervised representations for inter-patient tasks: Abdomen CT-CT, Hippocampus, and OASIS brain. `https://github.com/multimodallearning/convexAdam`

**corrField** A fast implementation (from [96]) of the corrField method [123] is introduced as a non-learning based unsupervised baseline. The method estimates sparse correspondences on image-based Förstner keypoints with exact message passing on a minimum spanning tree. MIND-SSC features are used for the similarity term. `https://grand-challenge.org/algorithms/corrfield/`

**Driver** [124] use a dual-encoder U-net backbone with separated multi-scale feature extractors that comprises Deformation Field Integration (DFI) and non-rigid feature fusion (NFF) module. It produces multi-scale sub-fields that progressively align fixed and moving features. The DFI module integrates sub-fields through up-sampling, re-weighing, and warping operations. The NFF dynamically fuses features of three pathways based on attention mechanisms. The overall framework comprises a rigid transform network and MI or LNCC similarity, weak label-supervision and regularisation.

**Epicure** [125] addresses the lung CT task using a conventional iterative-based registration approach based on Elastix toolbox optimizing the object function that is composed of the NCC similarity and a bending energy penalty term.

**Estienne** [126, 127] addressed Abdomen CT-CT and Hippocampus with label-supervision. The method combines a diffeomorphic symmetric spatial transformer network with a embedding merging step, that eases the learning by subtracting the embeddings of separately encoded fixed and moving scans and thereby leveraging

the prior knowledge that swapped inputs should yield negated velocity fields. They extend the label-based pre-training by including additional public datasets with at least partial overlap in segmentation classes, using segmentation masks produced by a CNN. https://github.com/TheoEst/abdominal_registration

**Gunnarsson** [128] propose an end-to-end learning-based 3D registration method inspired by the PWC-Net [129]. The method estimates and refines a displacement field from a cost volume at each level of a CNN downsampling pyramid and is supervised by a similarity (NCC) and/or segmentation (Dice) loss, as well as a smoothness penalty. The network is trained and evaluated on scan pairs from three tasks of the 2020 challenge (Lung CT, Abdomen CT-CT and Hippocampus MR) using the same weights for all tasks. https://github.com/ngunnar/learning-a-deformable-registration-pyramid

**Imperial** Imperial uses Image-and-Spatial Transformer Networks (ISTN) as the backbone of their method. In the ISTN, the fixed and moving images are first separately processed by the ITN to generate a segmentation mask and a feature map of the input image. Subsequently, both feature maps are used by the STN to predict the displacement field. The loss function consists of a structural-guided and image similarity and a regularisation loss. https://github.com/biomedia-mira/istn

**Joutard** Joutard addresses the Abdomen CT-CT task with a weakly supervised deep learning approach. A CNN extracts features from the fixed and moving image, which are concatenated with their spatial image coordinates. The feature distributions for each spatial location are then matched between the two images which yield a correspondence matrix from which the average displacement can be derived. The network is supervised by a segmentation (Dice) and a regularisation (L2 norm on gradients) loss.

**LapIRN** [92, 130] propose an image registration method based on Laplacian pyramid registration networks to overcome the large inter-and intra-variations of anatomical structures in the input scans. In 2021, [130] extended their approach by adding a conditional module that enables the input of the regularisation hyperparameter so that the different solutions for different hyperparameter values can be captured by a single convolutional neural network. This fast method won the on-site challenge in both years with robust results across all tasks. https://github.com/cwmok/Conditional_LapIRN

**LaTIM** [131] is an iterative technique exploiting vector-valued directional image representations: smooth edge-based fields oriented towards the main image edges (closely related to vector field convolution for active contour segmentation). The

method is implemented within the Elastix framework and shows improvements compared to directly using intensities.

**Lifshitz** [132] propose a novel solution of learning-based lung CT registration that comprises a 3D extension of ARFlow with multi-resolution warping, dense displacement correlation, and flow estimation. To address edge-preservation of sliding motion an unrolling of the total variation (L1) regularisation loss computation using variable substitution is proposed and shown to stabilise gradients during training.

**lWM** The method of lWM submitted to the Hippocampus MR task uses sequential deformation field composition, while the solution for the OASIS task uses an image pyramid separately applied to both input images and a U-Net with residual blocks. The objective function includes MIND, Dice, inverse consistency and diffusion losses.

**MEVIS** The submission of MEVIS [133] solves all tasks by classical iterative methods and build on cost functions and losses made up from several terms that are selected for the specific task. The methods use a coarse-to-fine multi-level iterative registration scheme where a Gaussian image pyramid is generated for both images to obtain downsampled and smoothed images. Then, a registration is performed on the lowest resolution level and the resulting deformation field serves as the starting point for the following registration on the next highest level. This proceeds to the finest level with quasi-Newton L-BFGS optimization at each level. For the Hippocampus task, a deep learning approach with a weakly supervised trained U-Net was applied using the same cost function as in the iterative approach.

**Multi-brain** [134] use groupwise, fully unsupervised registration techniques based on Bayesian modelling and Gauss-Newton optimisation, which learns priors over image intensities and spatial tissue classes. The method requires no pre-processing of the imaging data and does not utilise label information. The method is applied to Abdomen CT-MR, OASIS, and Lung CT. https://github.com/WTCN-computational-anatomy-group/mb.

**NiftyReg** [20] is applied as conventional baseline for all tasks without label supervision using NCC for CuRIOUS and otherwise MIND as similarity metric. Both bending and Jacobian regularisation penalties are applied and the number of pyramid levels is restricted to yield competitive run times (on multi-core CPU). https://github.com/KCL-BMEIS/niftyreg

**PDD-Net** The PDD-Net (probabilistic dense displacement network) [74, 135] uses a compact deformable convolutional network to extract image features and compute a six-dimensional dissimilarity tensor (three spatial + three displacement dimensions).

A smooth displacement field is obtained from the dissimilarities by interleaved (and twice repeated) steps of mean field inference over spatial dimensions and approximated min-convolutions over displacement dimensions. The method is adapted to four of the six challenge tasks (CuRIOUS, Hippocampus MR, Abdomen CT-CT, and Lung CT). `https://github.com/multimodallearning/pdd_net`

**PIMed** ■ PIMED use a multi-slice segmentation network and train a two-stage registration network for Abdomen CT-MR and Abdomen CT-CT and a residual VoxelMorph model for OASIS. For lung CT they apply a conventional method, with geodesic density regression and adaptation of intensities to lung tissue density [136].

**Winter** ■ Winter address all three tasks from 2021 by employing a traditional method for lung CT and a attention-based DL registration for Abdomen CT-MR and OASIS brain. Improvements are found by a two-step approach that firstly aligns provided ROI masks. The algorithms achieve intermediate ranks for the considered tasks without using any label supervision. The smoothness complexity is large for abdominal registration. `https://github.com/WinterPan2017/ADLReg`

## 5.4 Additional Experiments

**Label Bias** Previous publications on learning-based registration have already discussed the possibility of introducing a bias towards anatomies that are used both for training and evaluation [33]. While this bias is intrinsic to all segmentation approaches, registration is often used as a more generalistic tool in clinical applications that may require accurate alignment of structures that are not defined a priori. To study the effect of adding additional anatomical labels to the evaluation that were not present during method development and training, we extended both abdomen tasks. For the inter-patient CT-CT registration we included the duodenum with the manual annotations provided by [137], for the intra-patient MR-CT task we extended the predominantly large organs by five smaller ones: gallbladder, stomach, aorta, portal vein, pancreas (semi-automatically generated using a specifically trained nnUNet).

**Unsupervised Registration** The top-performing methods are all modular in their use of segmentation labels for supervision. As analysed in the label bias experiment, there is a risk of over-fitting registration performance to the chosen subset of manually annotated anatomies. We, hence, compared the unsupervised counterparts of the following methods: LapIRN and ConvexAdam. ConvexAdam already uses an unsupervised method for all three intra-patient tasks, and LapIRN for CuRIOUS and Lung CT. Therefore the additional comparisons are restricted to the abdomen and brain.

**Transferability** A robust registration method should work well for all scan pairs regardless of acquisition parameters and thus on every comparable dataset. A frequently mentioned limitation of deep-learning-based methods is that they reach higher accuracy on the dataset they are trained on and show a considerable loss of accuracy on other data. As in [138, 139], we evaluate the transferability of the methods submitted to the lung CT-CT task by registering the DIR-Lab 4DCT [81] scan pairs. The DIR-Lab scans are preprocessed in the same way as the scans of the lung CT-CT task. The evaluation is based on the target registration error of the landmarks and the smoothness of the deformation field. Furthermore, this experiment allows comparison to a variety of other lung registration methods, as the DIR-Lab data set is often used as a benchmark (please note that the reduced resolution leads to a general deterioration of TRE of around 0.2-0.3mm).

## 5.5 Results

### 5.5.1 Challenge Outcome

In this section, we will first present and discuss each task separately and subsequently the eight methods that are included in the overall ranking being submitted to at least four of the six tasks. Tables 5.3 to 5.8 give the numerical results and the scores for each algorithm for each task averaged over the number of scan pairs that were registered for that task. The algorithms are listed in order of their final placement per task. Figure 5.1 shows boxplots illustrating the distribution of the accuracy (TRE and Dice) of the different methods for each task. Furthermore, for selected task (Abdomen MR-CT, OASIS, and Lung CT), a bubble chart combines the accuracy, smoothness, and runtime metric.

**CuRIOUS** The registration to be carried out for this task was difficult for several reasons. First of all, it is a multimodal registration between MR and US images and the US images are typically noisier than the MR images. Furthermore, the pre-operative MR scans show a larger region of the brain whereas the intra-operative US volume was obtained to cover the entire tumor region after craniotomy but before dura opening. Due to these difficulties, only four methods were submitted to this task in addition to the three baseline methods. For two of these methods, some cases caused negative outliers and the average TRE was worse than the initial TRE (c.f. Table 5.3). Only the two baseline methods corrField and PDD-Net as well as the ConvexAdam method registered all scan pairs satisfactorily.

**Hippocampus MR** Due to its small volumetric size and reasonably large training dataset with only two anatomical labels, Hippocampus MR appeared to be a good entry-level task for learning-based registration approaches. It was also the only task that

Table 5.3: CuRIOUS

| | TRE↓ | TRE30↓ | SDLogJ↓ | RT↓ | Rank↑ |
|---|---|---|---|---|---|
| Initial | 6.38 | 12.00 | | | |
| corrField | 2.84 | 5.29 | 0.00 | 2.70 | 0.85 |
| PDD-Net | 3.08 | 6.28 | 0.00 | 8.21 | 0.83 |
| ConvexAdam | 3.31 | 5.82 | 0.00 | 1.33 | 0.77 |
| NiftyReg | 4.09 | 7.85 | 0.00 | 23.1 | 0.56 |
| LapIRN | 5.67 | 11.1 | 0.00 | 34.8 | 0.49 |
| MEVIS | 6.55 | 10.4 | 0.00 | 57.8 | 0.42 |
| Gunnarsson | 7.1 | 10.1 | 0.14 | 42.2 | 0.19 |

Table 5.4: Abdomen CT-CT

| | DSC↑ | DSC30↑ | HD95↓ | SDLogJ↓ | RT↓ | Rank↑ |
|---|---|---|---|---|---|---|
| Initial | 0.28 | 0.04 | 21.78 | | | |
| ConvexAdam | 0.69 | 0.45 | 11.03 | 0.06 | 2.75 | 0.94 |
| LapIRN | 0.67 | 0.47 | 12.51 | 0.12 | 3.80 | 0.82 |
| Estienne | 0.69 | 0.51 | 11.77 | 0.18 | 8.23 | 0.67 |
| MEVIS | 0.51 | 0.24 | 18.21 | 0.14 | 3.49 | 0.60 |
| corrField | 0.49 | 0.24 | 17.22 | 0.28 | 5.40 | 0.53 |
| PIMed | 0.49 | 0.23 | 15.75 | 0.05 | | 0.49 |
| PDD-Net | 0.49 | 0.24 | 17.75 | 0.41 | 6.06 | 0.44 |
| Joutard | 0.40 | 0.13 | 17.25 | 0.05 | 3.67 | 0.42 |
| NiftyReg | 0.45 | 0.20 | 20.70 | 0.36 | 17.1 | 0.36 |
| Gunnarsson | 0.43 | 0.17 | 18.55 | 0.13 | 31.5 | 0.33 |

Table 5.5: OASIS

| | DSC↑ | DSC30↑ | HD95↓ | SDLogJ↓ | RT↓ | Rank↑ |
|---|---|---|---|---|---|---|
| Initial | 0.56 | 0.27 | 3.86 | | | |
| LapIRN | 0.82 | 0.66 | 1.67 | 0.07 | 1.21 | 0.92 |
| ConvexAdam | 0.81 | 0.64 | 1.63 | 0.07 | 3.10 | 0.82 |
| lWM | 0.79 | 0.61 | 1.84 | 0.05 | 2.55 | 0.79 |
| Driver | 0.80 | 0.62 | 1.77 | 0.08 | 2.02 | 0.75 |
| PIMed | 0.78 | 0.58 | 1.86 | 0.06 | 3.47 | 0.71 |
| 3Idiots | 0.80 | 0.63 | 1.82 | 0.08 | 1.46 | 0.70 |
| Winter | 0.77 | 0.57 | 2.16 | 0.08 | 2.56 | 0.55 |
| MEVIS | 0.77 | 0.57 | 2.09 | 0.07 | 10.4 | 0.51 |
| Multi-brain | 0.78 | 0.59 | 1.92 | 0.57 | | 0.38 |
| corrField | 0.74 | 0.51 | 2.36 | 0.08 | 5.14 | 0.37 |
| Thorley | 0.77 | 0.60 | 2.21 | 0.31 | | 0.37 |
| NiftyReg | 0.73 | 0.51 | 2.37 | 0.06 | 5.00 | 0.36 |
| Bailiang | 0.67 | 0.42 | 2.74 | 0.04 | 1.38 | 0.33 |
| LaTIM | 0.74 | 0.52 | 2.31 | 0.08 | | 0.32 |
| Imperial | 0.76 | 0.57 | 2.43 | 0.19 | 2610 | 0.29 |

Table 5.6: Hippocampus MR

| | DSC↑ | DSC30↑ | HD95↓ | SDLogJ↓ | RT↓ | Rank↑ |
|---|---|---|---|---|---|---|
| Initial | 0.55 | 0.36 | 3.91 | | | |
| LapIRN | 0.88 | 0.86 | 1.30 | 0.05 | 1.03 | 0.93 |
| MEVIS | 0.85 | 0.84 | 1.55 | 0.05 | 0.59 | 0.78 |
| ConvexAdam | 0.84 | 0.83 | 1.85 | 0.07 | 0.48 | 0.75 |
| lWM | 0.79 | 0.76 | 2.20 | 0.08 | 0.80 | 0.63 |
| Estienne | 0.85 | 0.84 | 1.51 | 0.09 | 1.46 | 0.62 |
| PDD-Net | 0.78 | 0.76 | 2.23 | 0.07 | 0.35 | 0.58 |
| NiftyReg | 0.76 | 0.72 | 2.72 | 0.09 | 4.75 | 0.37 |
| corrField | 0.72 | 0.68 | 2.89 | 0.05 | 1.20 | 0.34 |
| Gunnarsson | 0.74 | 0.67 | 2.82 | 0.16 | 22.0 | 0.25 |

Table 5.7: Abdomen MR-CT

| | DSC↑ | DSC9↑ | HD95↓ | SDLogJ↓ | RT↓ | Rank↑ |
|---|---|---|---|---|---|---|
| Initial | 0.33 | 0.22 | 48.65 | | | |
| ConvexAdam | 0.75 | 0.73 | 24.92 | 0.09 | 1.30 | 0.82 |
| corrField | 0.76 | 0.73 | 23.35 | 0.10 | 2.13 | 0.81 |
| LapIRN | 0.76 | 0.69 | 22.81 | 0.12 | 1.50 | 0.77 |
| PIMed | 0.78 | 0.68 | 21.99 | 0.07 | 59.2 | 0.75 |
| MEVIS | 0.71 | 0.65 | 27.94 | 0.15 | 14.7 | 0.67 |
| Driver | 0.76 | 0.55 | 27.02 | 0.13 | 1.95 | 0.63 |
| NiftyReg | 0.65 | 0.55 | 33.09 | 0.12 | 11.0 | 0.55 |
| LaTIM | 0.54 | 0.49 | 41.17 | 0.13 | | 0.39 |
| Winter | 0.55 | 0.41 | 35.51 | 0.85 | 2.79 | 0.31 |
| Imperial | 0.51 | 0.41 | 48.60 | 0.11 | 278 | 0.30 |
| Multi-brain | 0.54 | 0.44 | 38.21 | 0.48 | | 0.30 |

Table 5.8: Lung CT

| | TRE↓ | TRE30↓ | SDLogJ↓ | RT↓ | Rank↑ |
|---|---|---|---|---|---|
| Initial | 10.24 | 16.80 | | | |
| corrField | 1.75 | 2.48 | 0.05 | 2.91 | 0.87 |
| ConvexAdam | 1.79 | 2.70 | 0.06 | 1.82 | 0.81 |
| MEVIS | 1.68 | 2.37 | 0.08 | 95.4 | 0.78 |
| LapIRN | 1.98 | 2.95 | 0.06 | 10.3 | 0.73 |
| PDD-Net | 2.46 | 3.81 | 0.04 | 4.22 | 0.62 |
| LaTIM | 1.83 | 2.50 | 0.05 | | 0.62 |
| Lifshitz | 2.26 | 3.01 | 0.07 | 2.90 | 0.61 |
| Imperial | 1.81 | 2.54 | 0.11 | 300 | 0.57 |
| PIMed | 2.34 | 3.27 | 0.04 | 623 | 0.55 |
| NiftyReg | 2.70 | 5.28 | 0.10 | 42.2 | 0.51 |
| Driver | 2.66 | 3.50 | 0.10 | 2.66 | 0.44 |
| Winter | 7.41 | 10.11 | 0.09 | 12.0 | 0.40 |
| Epicure | 6.55 | 10.29 | 0.07 | | 0.29 |
| Multi-brain | 6.61 | 8.75 | 0.08 | | 0.27 |
| Gunnarsson | 9.00 | 11.27 | 0.12 | 30.9 | 0.21 |

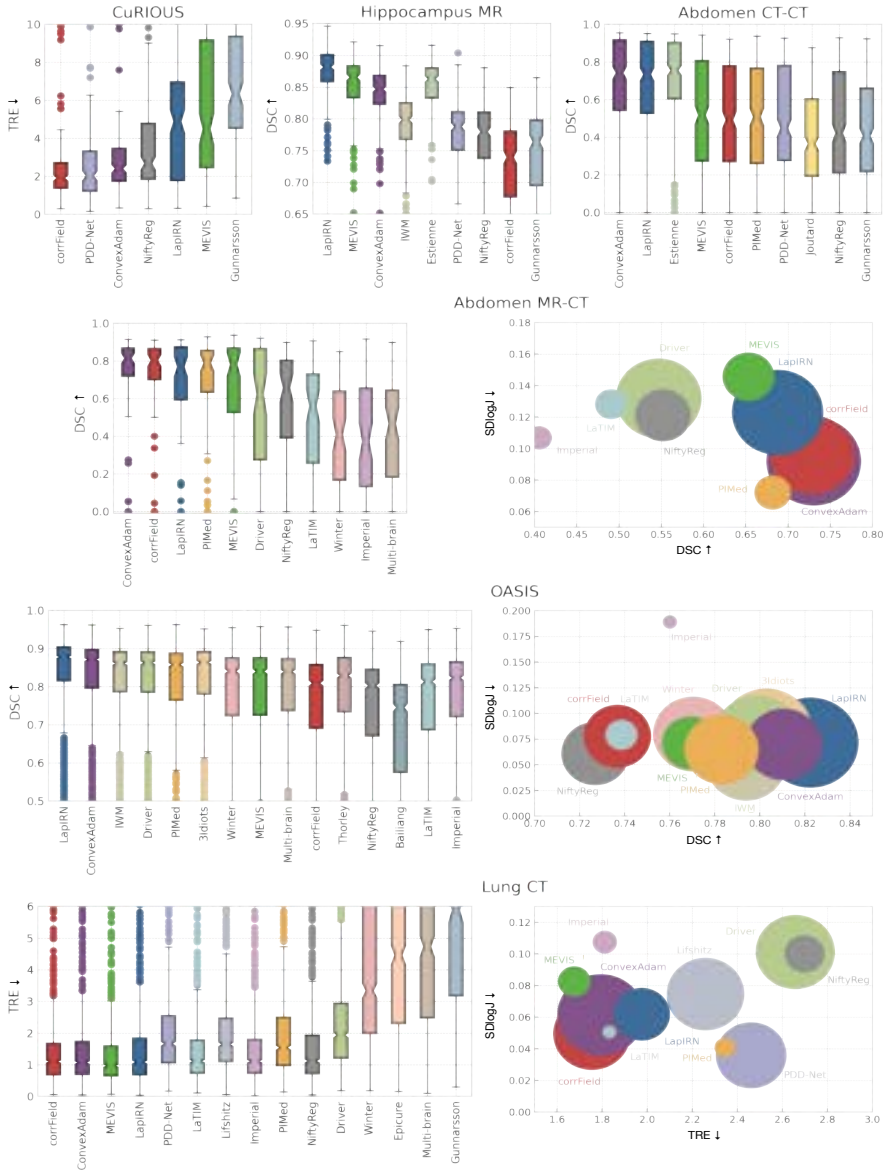



Figure 5.1: Boxplots and (selected) bubble charts visualising the results for the six challenge tasks. While the boxplots show the main accuracy metric (DSC and TRE, respectively), the bubble charts combine the accuracy, smoothness and runtime metric (a larger bubble means a faster runtime). Arrows (↑,↓) indicate the favourable direction of metrics. Comparison methods are color coded: ConvexAdam ■, LapIRN ■, MEVIS ■, corrField ■, NiftyReg ■, PDD-Net ■, PIMed ■, Gunnarsson ■, IWM ■, Estienne ■, Joutard ■, Driver ■, LaTIM ■, Winter ■, Imperial ■, Multi-brain ■, 3Idiots ■, Thorley ■, Bailiang ■, Epicure ■, and Lifshitz ■. Methods are sorted according to final rank scores.

Table 5.9: Overall rank scores of methods submitted to four or more tasks.

| | Cu-RI-OUS | Hip-pocam-pus MR | Ab-domen CT-CT | Ab-domen MR-CT | OA-SIS | Lung CT | Over-all | Intra-Pa-tient | Inter-Pa-tient |
|---|---|---|---|---|---|---|---|---|---|
| ConvexAdam | 0.77 | 0.75 | 0.94 | 0.82 | 0.82 | 0.81 | 0.82 | 0.80 | 0.83 |
| LapIRN | 0.49 | 0.93 | 0.82 | 0.77 | 0.92 | 0.73 | 0.76 | 0.65 | 0.89 |
| MEVIS | 0.42 | 0.78 | 0.60 | 0.67 | 0.51 | 0.78 | 0.61 | 0.61 | 0.62 |
| corrField | 0.85 | 0.34 | 0.53 | 0.81 | 0.37 | 0.87 | 0.59 | 0.84 | 0.41 |
| NiftyReg | 0.56 | 0.37 | 0.36 | 0.55 | 0.36 | 0.51 | 0.44 | 0.54 | 0.36 |
| PIMed | | | 0.49 | 0.75 | 0.71 | 0.55 | 0.35 | 0.39 | 0.33 |
| PDD-Net | 0.83 | 0.58 | 0.44 | | | 0.62 | 0.34 | 0.37 | 0.32 |
| Gunnarsson | 0.19 | 0.25 | 0.33 | | | 0.21 | 0.19 | 0.16 | 0.22 |

enabled sub-second run times. There was a performance gap between most supervised and unsupervised methods (NiftyReg, PDD-Net, and corrField). LapIRN reaches the first rank and MEVIS comes second, which notably used a fully-convolutional solution (without optimisation) only for this task. There is only a very small difference in the results between overall and robustness Dice (of the 30% most difficult instances), which highlights the fact that the problem is well-balanced and it requires fewer specific model adaptations.

**Abdomen CT-CT** Considering the abdomen for inter-patient registration is more challenging than the brain, due to larger anatomical shape differences that require large, complex deformation estimation. Due to the small size of many organs and large initial misalignments, previous work has often reported accuracies of around 40% DSC or less for DL methods [140]. Here, learning-based approaches can leverage anatomical segmentation priors and reach substantial improvement over previous state-of-the-art (as reported in [105]). Estienne, LapIRN, and ConvexAdam achieved 67-69% Dice overlap (across 8 individual labels) nearly 20% points higher Dice scores than all other participants with less than 10 seconds runtimes each and reasonably smooth displacement fields (SDLogJ < 0.2). Unsupervised methods have the advantage of being independent of label bias and reach up to 49% overlap (corrField and PDD-Net).

**Abdomen MR-CT** Multimodal nonlinear registration remains a challenging task in particular for abdominal scans where significant deformations can occur between scans as evident from low initial overlap of DSC-9=22% or DSC-4=33% (for 9 or 4 anatomical labels respectively). However, the provided initialisation masks can already lead to DSC-4>60% overlap using a similarity transform. The multimodal metrics,

MIND [141] and NGF [142] were used by the majority of participants, including the two top-performing methods ConvexAdam and corrField. Those methods also capture larger motion robustly using dense discrete correlation. LapIRN and PIMed (3rd and 4th rank) added a Dice loss, where it became obvious that focusing on the supervision with only 4 organs may lead to over-fitting on those structures and a lower accuracy for further anatomies. The provided additional labeled unpaired datasets (30 BCV CT scans and 30 CHAOS MRI scans [143]) have not yielded the desired advantages due to subtle differences in appearance.

**OASIS** The OASIS inter-subject brain task attracted the most learning-based solutions. The results are summarised in Table 5.5 and visualised in Figure 5.1 showing that most of these methods achieve very similar results in terms of Dice Score for the cases with the highest scores (Dice of 80-90%). The differences are primarily in the more difficult cases and thus in the DSC30 score, where the LapIRN, convexAdam, and the methods of Driver and 3idiots methods perform slightly better than for example PIMed and Winter. The conventional methods of MEVIS and corrField achieve mid-ranked accuracies but have a higher runtime. Figure 5.2 shows an example sagittal slice of the fixed image overlayed with the false-negative segmented voxels (green) and false-positive segmented voxels (yellow) for initial moving segmentation and the propagated segmentations by the methods of Imperial, PIMed, and LapIR. All methods were able to align the small structures of the brain with only very small visible differences.

**Lung CT** The complexity of this registration task is manifold. First, the fields of view of the fixed and moving scan differ largely since the lungs in the expiration scan are not fully visible. Second, the scale of the motion within the lungs can often be larger than the anatomical structures (vessels and airways) themeselves. Therefore, a registration method needs to estimate large displacements that account for substantial breathing motion and also align small structures like individual pulmonary blood vessels precisely. To measure the accuracy manual landmarks are used that are typically located at the boundary or bifurcation of vessels, airways, and parenchyma. This task was carried out in both years because in 2020 only the MEVIS team, which uses automatically computed keypoints as additional metric, achieved a TRE of less than 2mm (1.72mm), while other teams performed considerably worse (e.g. LapIRN 3.24mm and PDD-Net 2.46mm). In 2021, keypoint correspondences were provided for training and the submissions improved with six teams achieving a TRE of less than 2mm. Especially for the deep-learning-based methods LapIRN and Imperial this is a remarkable result, because they were only trained on a small dataset of 20 scan pairs. In contrast to other lung registration challenges like EMPIRE or DIR-Lab, the TRE is relatively high. This can be explained, in part, by the low resolution of the scans that were chosen in the preprocessing. Figure 5.2 visualises the difference images of an example coronal slices for the methods of Driver, convexAdam, and MEVIS overlayed with the landmarks.

**Overall Ranking** ConvexAdam was among the top 3 on each task (winning Abdomen CT-CT and Abdomen CT-MR) and ranked first overall among the eight methods that were applied to four or more tasks - highlighting the importance of effective optimisation and versatility of using learned semantic or hand-crafted MIND features depending on the application. LapIRN reached the overall second rank and yielded the best result for Hippocampus and OASIS. This demonstrates that a well-designed convolutional feed-forward network (instance optimisation was used only for CuRIOUS and Lung CT) can outperform conventional approaches in particular for inter-patient tasks. MEVIS achieved the third place overall, with top ranks in particular for Lung CT and Hippocampus based on a combination of NGF metric, curvature regularisation, and L-BFGS optimisation with additional learning components only employed for the brain task. CorrField uses no label supervision at all, but relies on highly optimised graph-based registration, and comes fourth overall winning two individual tasks: CuRIOUS and LungCT. It is the best method for intra-patient registration. PIMed's method achieves strong performance on Abdomen MR-CT and OASIS and generalises well to Abdomen CT-CT.

### 5.5.2 Additional Experiments

**Label Bias and Unsupervised Registration** When evaluating the influence of supervision with anatomical labels, we found a clear distinction between intra-patient registration (Abdomen MR-CT) and inter-patient registration (Abdomen CT-CT, Hippocampus and OASIS). The former shows nearly no advantage of including such information and it is therefore possible to avoid a risk of overfitting towards certain anatomies. The latter, however, shows a clear deterioration in accuracy when excluding all or some of the structures from training that were used for evaluation. CorrField, which is unsupervised and achieves the highest scores for intra-patient registration trails nearly all learning-based methods on the remaining inter-patient tasks. LapIRN trained without Dice loss (i.e. without anatomical knowledge) improves upon those results and achieves very strong results for OASIS and Abdomen CT-CT (ranks would be third and fourth respectively). This demonstrates that in particular a large training database and an advanced deep learning architecture (LapIRN uses multi-level Res-Nets, multiple warps and multi-resolution loss functions) can narrow the gap between supervised and unsupervised approaches. We evaluated ConvexAdam (that decouples feature extraction from optimisation) for Abdomen CT-CT in three settings: 1) all 13 labels in training with 8 (7 identical) in test (DSC=69%), 2) 4 labels in training with 8 (thereof 3 identical) in test (DSC=55%) and 3) no labels in training (DSC=45%). This shows that partial supervision clearly leads to improvement of those identical anatomies but can also help to align nearby structures: esophagus which was excluded improved by 16% points (likely through the guidance of liver and aorta) and pancreas overlap was increased by 12% points (possibly by including portal vein and adrenal

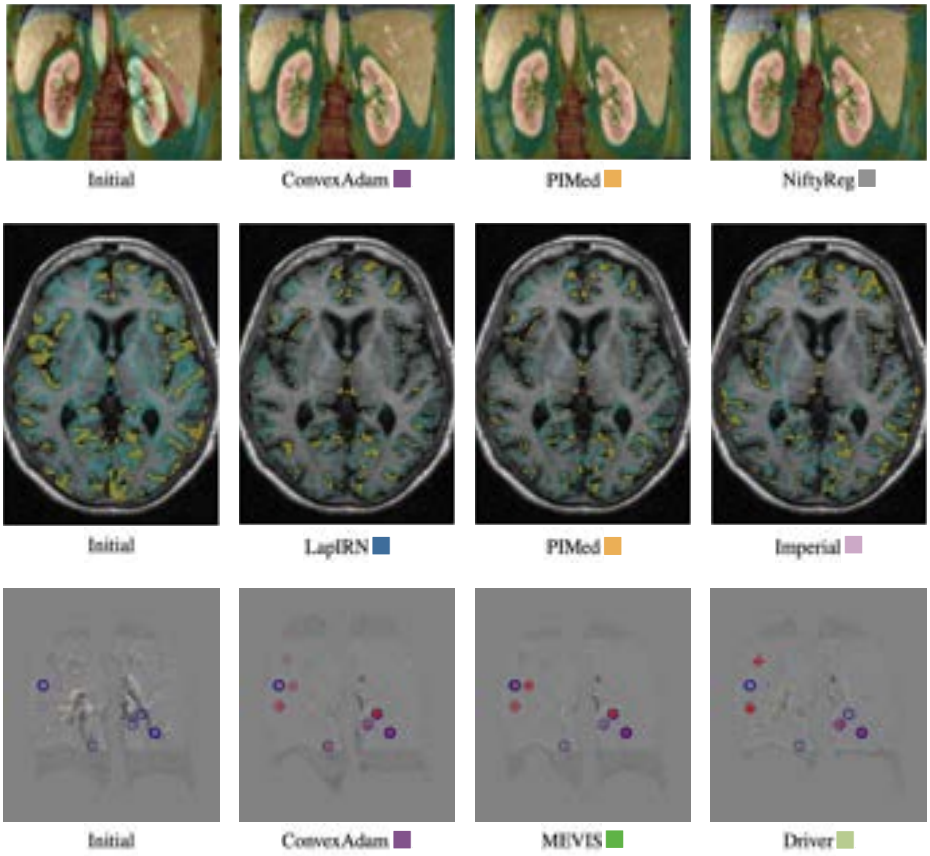



Figure 5.2: Exemplary qualitative results for selected methods and tasks. Top row: Overlay of coronal abdominal MR (gray) and warped CT (color) slices. Middle row: False-negative (green) and false-positive (yellow) voxels of propagated segmentation labels on saggital slices of the OASIS dataset. Bottom row: Coronal slices of difference images between exhale and warped inhale lung CT scans (including exhale (blue circle) and warped inhale (red cross) landmarks).

gland). As mentioned in Sec. 5.5.1 training on 4 and evaluating on 9 abdominal organs for MR-CT fusion results in a moderate performance gap between supervised and unsupervised methods (the latter being about 10% better on this metric).

**Transferability** In this experiment, we were able to show that the three best methods of the lung registration task also perform very well on the DIR-Lab dataset (MEVIS 1.22 mm, convexAdam 1.31 mm, and corrField 1.34 mm) without any further hyperparameter adaptations. Since the inspiration and expiration images of the DIR-Lab dataset are extracted from a 4DCT dataset with shallow breathing, the registration task is easier than the Learn2Reg lung CT task. Therefore, the methods reach a lower TRE on the DIR-Lab dataset compared to the Learn2Reg lung task (improved TRE of 0.46 mm, 0.48 mm, and 0.41 mm for MEVIS, convexAdam, and corrField, respectively). Due to the performed preprocessing and the reduced resolutions, the Learn2Reg methods achieve slightly worse results than state-of-the-art methods evaluated on the DIR-Lab dataset. For example, method of MEVIS included in a registration pipeline registering the original images reaches a TRE of 0.94 mm [65]. LapIRN achieves similar results on both datasets (Learn2Reg lung CT 1.98 mm and DIR-Lab 1.98 mm) showing that the best deep-learning-based methods can also be successfully applied to other datasets without retraining. However, no performance improvement can be observed on the easier dataset mainly due to the limited training data size.

## 5.6 Discussion

In the following, we will discuss specific aspects of the challenge.

**Comparison of Learning- vs Optimisation-based Registration** We argue that Learn2Reg has helped to demystify common beliefs of fundamental differences between learning- and optimisation registration. First and foremost, there is virtually no difference in computational speed. GPU-acceleration brings down computation cost of optimisation-based methods to a few seconds for 3D registration, i.e. the extraction of features using CNNs often outweighs optimisation times. Furthermore, we see a clear trend that learning based on segmentation labels is primarily beneficial for inter-subject registration. For Abdomen CT-CT for instance large improvements of 20%points in Dice overlap compared to previous work [105] could be achieved when incorporating Dice losses. All three highest ranked approaches employ a combination of DL and optimisation: LapIRN primarily uses a deep network, but add instance optimisation for Lung CT, MEVIS mainly use conventional optimisation but a DL network for Hippocampus MR, and ConvexAdam combine discrete optimisation with U-Net-based semantic features for inter-patient tasks. Our current challenge design did not consider any computational constraints (GPU memory, inference time on CPU), which might

limit the practical impact for some applications and should be considered in future studies.

**Algorithmic Design Choices** There are no direct ablation studies possible for the used architectures and loss functions since each method differs in multiple aspects (see Table 5.2), but some general trends are visible nonetheless. Most approaches use a combination of contrast-invariant intensity metrics (LNCC, NGF and MIND) as well as a Dice loss for tasks where anatomical labels are available. To address larger motion (all tasks expect brain) DL registration methods employ multi-scale (and residual) architectures, multiple warps or often dense correlation layers. Two-stream approaches that process both input scans independently are commonplace to deal with multimodality or contrast variations (lung CT).

**Plausibility of Transformations** We analysed the smoothness of transformations with respect to the log-standard deviation of Jacobian determinants for all experiments. While this measure is far from perfect, it enabled a ranking of different solutions to the inherently ambiguous nonlinear registration task that may achieve similar accuracy with large differences in complexity (the common assumption being: the smoother transform is then preferable). As visualised in the bubble-charts in Figure 5.1 there is a tendency that more accurate solutions are also smoother, which indicates that enforcing regularity is an effective means of avoiding overfitting and improves robustness. Some notable exceptions can be found for lung CT, where Imperial appears to suffer from too low regularisation while PDD-Net and PIMed may have reduced accuracy in exchange for overly smooth fields. A potential explanation for the positive correlation of smoothness and accuracy could be the hypothesis that accurate methods are able to establish strong (correct) correspondences at relevant anatomies and extrapolate as smooth as possible in uncertain areas. That means putting emphasis on either surfaces (e.g. based on segmentation estimates) or geometric keypoints (for lung scans) can be beneficial.

**Comparison to Baselines** We evaluated two conventional methods, NiftyReg [20] and corrField [123] (using the GPU implementation of [96]), and two learning-based approaches, PDD-Net [74] and the original version of VoxelMorph [33] as baselines. The latter two were only applied to a subset of tasks. NiftyReg achieves reasonable accuracies across all tasks but falls behind supervised methods on inter-patient tasks. The original VoxelMorph variant reaches an average Dice overlap of 76.88%±2.17 % for OASIS (7th-10th place based on DSC alone) and a TRE of 7.51±3.43 mm for lung CT (13th place). When trained on a large additional lung dataset [138] a TRE of 1.71±2.86 mm was achieved for the additional DIR-Lab lung experiment for which the best performing methods in this challenge achieved 1.3 mm. PDD-Net achieved a second rank for CuRIOUS and fifth place for Lung CT, but lower scores for inter-patient

registration. CorrField achieved the best scores overall for CuRIOUS and LungCT and second place for Abdomen MR-CT (each task without supervision), making it stand out as the best performing intra-patient approach. This demonstrates that conventional methods are still very competitive for datasets without strong label supervision.

**Reducing Entry Barriers** By pre-processing each dataset to the same dimensions and isotropic resolution and providing anatomical annotations for training data wide participation was achieved from research groups across the world. The OASIS inter-subject brain task attracted the most learning-based solutions, which highlights the importance of large, labeled training datasets for deep-learning-based registration and mirrors the focus of recent research. Lung CT intra-patient registration was addressed by the same number but more diverse set of methods, including conventional, fully deep-learning-based, and hybrid approaches. Therefore, we assert that new application areas have been opened for many participants. While adoptions of metrics, fine-tuning, and supervision appeared to be important for methods that were applied across multiple tasks, the consistent performance of the three top-performing groups demonstrates that Learn2Reg enabled effective multi-task solutions. Some aspects of medical image registration, including affine or rigid pre-alignment, dealing with differences in field-of-view of voxel resolutions, and the processing of very high-resolution scans have been omitted due to our challenge design and could be addressed in future.

**Limitations of the Challenge Design** We have identified a number of limitations that should be addressed in future studies. First, for computational reasons the training of algorithms was performed offline by participants. This could introduce a bias when additional data is used by certain teams that is not accessible to others and prevents the use of larger (central) datasets that cannot be made public due to privacy concerns. Enabling docker-based training or fine-tuning of models directly at grand-challenge.org would be desirable. Second, the amount of available annotated training data varied across tasks and made in particular intra-patient tasks harder for learning-based approaches. Decoupling anatomical feature learning from patient-wise optimisation could be a next step, e.g. by providing training data for airway and fissure segmentation for lung CT. Third, the accuracy evaluation is in general limited by inter-observer noise and the difficulty of assessing registration accuracy based on segmentation overlap, which disregards the plausibility of correspondences along the surface or within the structure. Since this problem is inherent to any registration evaluation, we cannot offer any better solution than aiming for further manual annotation efforts. Fourth, the provision of all segmentation classes for training that were used for testing is in our opinion the most problematic limitation of this challenge. This was due to the fact, that for 3 out of 4 tasks with segmentation labels these annotations were already publicly available prior to Learn2Reg and we considered it in-transparent (and biased) to simply not point participants to their availability. While this would not be considered a problem at all

for segmentation tasks (cf. the Medical Decathlon), image registration generally aims to recover deformations for the entire field-of-view. We aimed to mitigate the influence of over-fitting towards labelled anatomies by performing additional experiments for partial supervision.

**Impact and Clinical Adoption** With regard to the five-year-old survey on medical image registration by [100], we can reflect that the shift from surface-based registration to intensity-based approaches has somewhat been reverted with a majority of approaches employing segmentation-based overlap or keypoints as driving force. The establishing of different learning-based strategies, including hybrid approaches that decouple semantic feature extraction from optimisation or combine feed-forward networks with instance optimisation, can be seen as an important new trend. To assess the likelihood of adopting registration in clinical practice, we are encouraged to see that a number of previous obstacles have been successfully addressed by participants of Learn2Reg. First, robustness against variations in scanner protocol and patient characteristics was shown to be very high for top-ranking methods that tackled both multi-centric MRI studies (OASIS) as well as the transferability issue for lung CT. Second, run times have been considerably reduced to a few seconds, which will enable clinicians to interact with those algorithmic solutions by adjusting hyper-parameters, e.g. the strength of regularisation penalties in near realtime (this holds only true for deep-learning-based methods if they are either decoupled or trained with conditioning cf. [130]). Third, it became clear that highly nonrigid transformations are as well solved as rigid alignment, opening up the promise for clinical applications in image-guided surgery, radiotherapy. In fact, it appears as if rigid/pre-alignment remains an active problem in particular for DL solutions.

## 5.7 Conclusion

The Learn2Reg challenge was the first to evaluate a wide-range of methods for various inter- and intra-patient as well as mono- and multimodal medical image registration tasks. The main goal of this challenge was to provide a standardised benchmark on complementary tasks with clinical impact and a platform for comparison of conventional and learning-based medical image registration methods. We established a lower entry barrier for training and validation of 3D registration, which helped us compile results of over 65 individual method submissions from more than 20 unique teams. Although registration is highly dependent on the task, three methods (convexAdam, LapIRN, MEVIS) and a baseline method (corrField) were shown to work robustly on all tasks with only minor adjustments to the hyperparameters. Furthermore, several teams (Estienne, PIMed, Driver, 3idiots, Multi-brain LaTIM, Lifshitz and Imperial) have submitted tailored solutions to individual tasks and achieve very good results with it. For the conventional methods convexAdam, MEVIS, and corrField, it was also shown

that they can be applied directly to new data sets without loss of accuracy. Furthermore, we demystified the common belief that conventional registration methods have to be much slower than deep-learning-based methods. Nevertheless, with LapIRN a deep-learning-based registration method achieves state-of-the-art registration results within seconds. We could not identify any architecture that was advantageous over others. However, it was found that for deep-learning-based methods using a Dice loss for inter-patient registration is particularly useful and instance optimisation helped increasing the accuracy for intra-patient registration. With the Learn2Reg challenge, we have created a dataset for benchmarking future registration papers. Furthermore, the dataset has the potential to allow the development of dataset-independent and self-configuring registration methods.

## Acknowledgment

We thank Yipeng Hu and Tom Vercauteren who co-organised the first Learn2Reg workshop at MICCAI 2019 that was held as a tutorial and initiated the development of the challenge.

# 6

# CHAPTER 6

# Whole-Body Soft-Tissue Lesion Tracking and Segmentation

**BASED ON:** A. Hering, F. Peisen, T. Amaral, S. Gatidis, T. Eigentler, A. Othman, and J. H. Moltz. "Whole-body soft-tissue lesion tracking and segmentation in longitudinal CT imaging studies," *Medical Imaging with Deep Learning*, 2021.

## Abstract

In follow-up CT examinations of cancer patients, therapy success is evaluated by estimating the change in tumor size. This process is time-consuming and error-prone. We present a pipeline that automates the segmentation and measurement of matching lesions, given a point annotation in the baseline lesion. First, a region around the point annotation is extracted, in which a deep-learning-based segmentation of the lesion is performed. Afterward, a registration algorithm finds the corresponding image region in the follow-up scan and the convolutional neural network segments lesions inside this region. In the final step, the corresponding lesion is selected. We evaluate our pipeline on clinical follow-up data comprising 125 soft-tissue lesions from 43 patients with metastatic melanoma. Our pipeline succeeded for 96 % of the baseline and 80 % of the follow-up lesions, showing that we have laid the foundation for an efficient quantitative follow-up assessment in clinical routine.

## 6.1 Introduction

Measurement of metastatic tumors on longitudinal computer tomography (CT) scans is essential to evaluate the efficacy of cancer treatment. The current guideline of metastatic tumor evaluation on CT scans is called response evaluation criteria in solid tumors (RECIST) [144]. Manual measurement of the tumors for the RECIST criteria is often time-consuming and error-prone. However, the diameter-based RECIST criteria also undergo continuous changes. Automated approaches might significantly speed up response evaluation and help to handle the ever-growing mass of image-based staging and follow-up evaluations [145].

Furthermore, radiomics is currently one of the most important topics in radiology. High-throughput extraction of quantitative features resulting in the conversion of medical images into minable data and the subsequent analysis promise new insights into therapy response and hold the potential to revolutionize medical image-based evaluation techniques [146]. Both fields have a huge clinical impact due to rising demand for fast and reliable therapy response evaluations. They, however, share a common bottleneck: automated lesion segmentation. Only if this obstacle is overcome, clinicians will use the mentioned techniques accordingly in a daily manner.

Metastatic malignant melanoma is the perfect entity to implement a pipeline for full-body lesion segmentation. Besides lung and liver, metastatic lesions of melanoma can be found in almost every organ or tissue, such as lymph nodes, adrenal glands, cerebrum, bone, spleen, and soft tissue [147]. Whole-body cross-sectional imaging is part of the standard diagnostic work-up for staging, response assessment, and follow-up in patients with advanced melanoma according to current international guidelines. Malignant melanoma has been increasing fast in the last decades and represents a public health matter in several countries due to its high mortality rates [148].

Among melanoma metastases, soft-tissue lesions provide a particular hurdle. They can arise in a variety of locations (cutaneous, subcutaneous, muscular, retroperitoneal) and shapes (round, multilobular, well defined, invasive), are often primarily small and, if not surrounded by fatty tissue, extremely hard to distinguish. A sufficient segmentation pipeline for soft-tissue metastases in malignant melanoma patients would therefore provide a valuable foundation for further steps towards a full-body lesion segmentation pipeline, that could be transferred to other entities.

To the best of our knowledge, no work has been presented until now that tackles the problem of soft-tissue lesion segmentation in longitudinal CT image series. Lesion segmentation in other anatomical regions, however, has been studied extensively. For example, promising results have been accomplished for liver [149] and kidney lesions [150] in challenges. Currently, the most general and successful available approach is the nnU-Net framework of [6], which has shown impressive results for several organ segmentation tasks such as liver, spleen, kidney, pancreas, heart, or aorta segmentation and also outperforms most methods segmenting different lesion types such as pancreas,

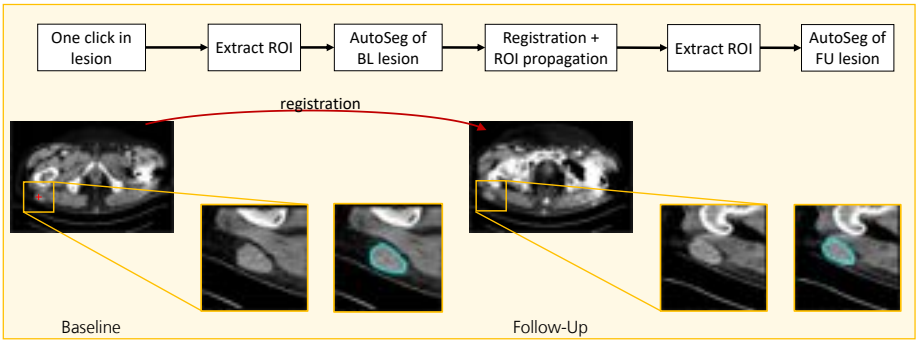



Figure 6.1: Schematic representation of the proposed pipeline for lesion tracking and segmentation.

liver, lung, kidney, or Multiple sclerosis (MS) lesions. nnU-Net [6], initially based on U-Net [53], automatically configures itself, including pre-processing, network architecture, training and post-processing—making it an ideal baseline to build a lesion tracking pipeline.

However, as the lesion segmentation experiments in [6] focus only on segmenting lesions in one organ in one scan, it cannot be used "as is" and requires some modifications. Only few works have been presented on lesion tracking [151] and on lesion tracking and segmentation in longitudinal image scans (e.g. [152–154]. In this work, we tackle the problem of longitudinal tracking and segmentation of soft-tissue lesions in whole-body CT scans.

## 6.2 Method

In our proposed pipeline, soft-tissue lesions are first identified by a radiologist with one click inside the lesion in the baseline CT scan. This step is introduced to avoid annotation of false positive lesions. We then apply our algorithm to automatically segment and measure the diameter in the baseline and follow-up image. This is done by (1) extracting the region of interest (ROI) around the point annotation of the radiologist and applying our CNN to segment the lesion; (2) registering the baseline to the follow-up image; (3) propagating the region of interest to the follow-up image to constrain the search region and applying the CNN on the propagated region of interest in the follow-up image; and (4) selecting the corresponding lesion in the output of the CNN. Figure 6.1 shows an overview of our proposed algorithm. In the following, we describe each step in more detail.

### 6.2.1 Lesion segmentation

To generate the training data, we select for each lesion a bounding box around the point annotation of the radiologist with a size of 100 mm, which is clamped by the image region. Then, we use the nnU-Net framework of [6] to train a 3d full resolution model which consists of a U-Net-like [53] architecture. The main settings are shown in Table 6.1 in the appendix. The trained network is applied to segment the lesion in the baseline and follow-up image on the test dataset.

### 6.2.2 Registration

Propagation of lesion segmentations into follow-up images of the same patient allows for a higher degree of automation because the location and approximate appearance of the lesions are already known. In this scenario, registration algorithms can be employed to find the corresponding image region [153]. For metastatic melanoma, typically full-body or thorax-abdomen CT scans are acquired, which can easily exceed image sizes of $512 \times 512 \times 1000$, which can be a challenge in terms of memory usage and runtime. The registration has to align the global structures but at the same time be locally accurate enough so that the lesion propagation is precise enough. Therefore, we adopted a three-step approach to automatically register the baseline to the follow-up image, which consists of the following steps: (1) a translational alignment; (2) a rigid registration; and (3) a deformable registration. Hereby, the registration pipeline starts with robust methods with fewer degrees of freedom and moves on to more precise, but less robust methods, which require better starting points due to their higher degrees of freedom.

**(1) Translational alignment** The translational prealigment is based on a brute force grid search method named FASTA (Fast Translation Alignment), which evaluates a difference measure (here Sum-of-Squared-Distances (SSD), the squared $\ell_2$ norm of the difference image) on a grid of possible translations. Finer grids allow for more precise translation estimation at the expense of increased computational cost. For faster processing, the moving image is resampled to a maximal image size of $128 \times 128 \times 128$. The fixed image is resampled to the same image resolution as the moving image. For the grid generation, we choose a sampling rate of 3, 3, and 51 in x, y, and z-direction respectively. Since the CT scans are centered around the body center, only the z-translation is used for prealignment.

**(2) Rigid registration** The translational prealignment in z-direction is used as a starting point for a rigid multi-level registration using the SSD distance measure. The method uses a Gauss-Newton optimization scheme to solve the optimization problem.

**(3) Deformable registration** The final step is the matrix-free deformable registration of [19]. The deformation is defined as a minimizer of the cost function

$$\min_{\mathrm{y}} \mathcal{D}^{\mathrm{NGF}}(\mathcal{F}, \mathcal{M}(\mathrm{y})) + \alpha \mathcal{R}^{\mathrm{curv}}(\mathrm{y}), \tag{6.1}$$

with the normalized gradient field distance measure $\mathcal{D}^{\mathrm{NGF}}$ [45] that focuses on the alignment of image gradients of the fixed image $\mathcal{F}$ and the deformed moving image $\mathcal{M}(\mathrm{y})$. The second-order curvature regularizer $\mathcal{R}^{\mathrm{curv}}$ [84] enforces smooth deformation by penalizing spatial derivatives. The parameter $\alpha$ is a weighting factor. The method uses the limited-memory Broyden-Fletcher-Goldfarb-Shannon (L-BFGS) optimization scheme to solve the optimization problem and is embedded in a multi-level scheme.

### 6.2.3 Lesion tracking

We use the registration to propagate the baseline contour to the follow-up scan. While this propagated contour may not be accurate enough due to size changes under therapy, it provides a good initial correspondence. To compensate for registration errors, we enlarge the search region by 50 mm in every direction to ensure that the corresponding lesion is inside this selected region and to include enough information for the CNN.

### 6.2.4 Lesion selection

The CNN is not constrained to segment only one lesion inside the selected region in the follow-up scan. Therefore, we select the lesion whose center is closest to the center of the propagated lesion. To avoid annotation of wrong lesions close by in the case of vanishing lesions under therapy, we only accept a lesion annotated by the network if the Euclidean distance of its center is smaller than 25 mm to the propagated lesion center.

## 6.3 Experiments and Results

### 6.3.1 Dataset

The dataset consists of 206 baseline and follow-up CT scan pairs of patients with metastatic melanoma (Stage IV, AJCC) treated at the Center for Dermato-Oncology at the University Hospital Tuebingen, Germany. All patients received either mono (Nivolumab or Pembrolizumab) or combined (Nivolumab+Ipilimumab) immunotherapy or targeted therapy (Vemurafenib +Cobimetinib or Dabrafenib+Trametinib) before the follow-up scan. The patients were split into 163 training and validation cases and 43 test cases with overall 2408 and 125 manual annotated soft-tissue lesions in the baseline images. Training was performed exclusively on baseline images, whereas testing was done on both baseline and follow-up scans. Therefore, we selected patients with lower

lesion counts for the test set in order to obtain a diverse set of lesions while keeping the annotation effort feasible. For the test cases, 25 of the 125 lesions are gone in the follow-up image.

### 6.3.2 Baseline segmentation

To show the advantage of training the network only on a small region of interest around the lesions, we compare our approach to a network trained on the whole images. However, for the evaluation, we use the closest lesion to the point annotation for both approaches, and therefore, false-positive annotations are not taken into account.

Since the network is not forced to segment anything in the region of interest, we evaluate the percentage of correctly annotated lesions. A lesion counts as correctly annotated if there is an overlap with the segmentation mask. To evaluate the performance of our segmentation network, we use Dice coefficient, average surface distance (ASD), and Hausdorff distance (HD) if the network segmented the correct lesion. Moreover, we evaluate the Surface Dice [155] with a threshold of 1 mm, which is a good approximation for the correction effort given an imperfect segmentation mask of a relatively small structure.

When the nnU-Net is trained only on the small region of interest around the point annotation, the network segments the correct lesion in 96 %, whereas with training on the whole image, only 37.6 % of the lesions are annotated. On the correctly annotated lesions, the network trained on the ROI achieves on average a better Dice Score (0.79 vs. 0.60 ), Surface Dice (0.88 vs. 0.68 ), and average surface distance (1.40 mm vs. 1.77 mm) but a slightly worse Hausdorff distance (5.09 mm vs. 4.59 mm). Note that the number of included lesions for the calculation differs depending on the training mode. Taking all lesions into account the advantage increases to 0.76 vs. 0.23 for the Dice Score and 0.85 vs. 0.26 for the Surface Dice. Figure 6.2 summarizes the quantitative results and Figure. 6.3 shows several visual examples of the results produced by our network.

### 6.3.3 Registration accuracy

We measure the registration accuracy using the center point matching accuracy as in [151], which represents the percentage of correctly matched lesions. A match counts as correct when the Euclidean distance between the center of the propagated baseline lesion and the center of the manually annotated follow-up lesion is smaller than a threshold. Since in this application whole-body CT scans are registered and large volume changes of the lesion happen due to therapy, we set the threshold to 25 mm. For this evaluation, only the lesions which are visible in the follow-up image are taken into account and therefore the number of lesions reduces to 100.

In 95 of the 100 cases, the Euclidean distance was less than the threshold with a mean Euclidean distance of 7.66 mm. The average absolute offset between the center

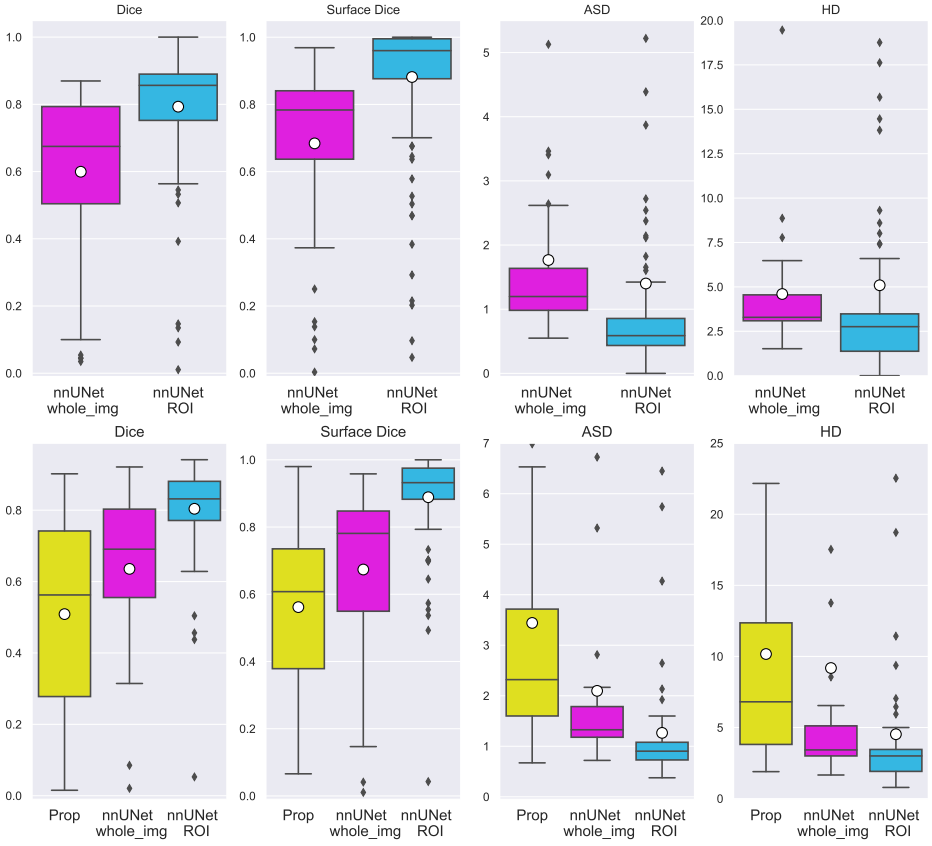



Figure 6.2: Comparison of the evaluation metrics for all baseline lesions (upper row) and follow-up lesions (lower row) in which the correct lesion was annotated with the underlying training mode. Therefore, the number of included lesions for the calculations varies depending on the training mode. For the follow-up lesions, the lesion results by the registration propagated are shown in yellow. The boxplots show the median line and the mean as a white circle.

of the propagated baseline lesion and the center of the manually annotated follow-up lesion is 3.79 mm, 3.16 mm and 4.49 mm in x-, y- and z-direction, respectively. A histogram of the offset is shown in Figure 6.9 in the appendix.

### 6.3.4 Follow-up segmentation

We evaluate the follow-up segmentation in the same way as the baseline segmentation. However, the successful segmentation of the follow-up lesion depends not only on the segmentation accuracy itself but the whole pipeline. For the cases in which the lesion was not propagated accurately enough, segmentation by the nnU-Net was not possible. To evaluate the whole pipeline, those lesions are counted as not correctly annotated lesions. Furthermore, in the 25 cases in which the lesion was fully regressive in the follow-up image, we expect the nnU-Net not to annotate anything.

In 80 % of the lesions, our pipeline successfully annotates the lesion in the follow-up scan with an average Dice Score of 0.80 and an average Surface Dice of 0.89 The lesion propagated by the registration has an overlap to the manual annotation in 77.5 % with an average Dice score of 0.51 and a Surface Dice of 0.56 All quantitative results are summarized in Figure 6.2. All failure cases are visualized in the appendix. In 17 of the 25 cases in which the lesion has disappeared in the follow-up image, the nnU-Net correctly not segment anything.

## 6.4 Discussion and Conclusion

This paper presents a pipeline that automates the segmentation of matching lesions in follow-up CT examinations of cancer patients, given a one-click point annotation in the baseline lesion. We have validated our pipeline on the challenging task of whole-body soft-tissue lesion tracking and segmentation. Our pipeline succeeded for 96 % of the baseline lesions and for 80 % of the follow-up lesions with an average Dice Score of 0.79 and 0.80 , respectively. Furthermore, our pipeline achieves an average Surface dice of 0.88 , which shows that the required correction effort is very low.

All failure cases in the follow-up image are visualized in Figure 6.6 and 6.7 in the appendix showing that the pipeline fails due to different reasons. For some cases, the registration was not accurate enough and therefore a wrong or no lesion was selected even though the correct one was segmented. Other lesions are hard to distinguish from surrounding tissue or they have an untypical shape that might cause problems. In some cases, the lesion split into two smaller lesions in the follow-up scan after the patient received therapy and the nnU-Net segmented both, but just one lesion was selected. In some of these cases, it is also difficult for a radiologist to identify and segment the lesion correctly. Our pipeline has still some limitations which have to be addressed before it could be used in the clinic. Lesions can split or merge over time, however, our pipeline assumes that every lesion in the baseline has zero or

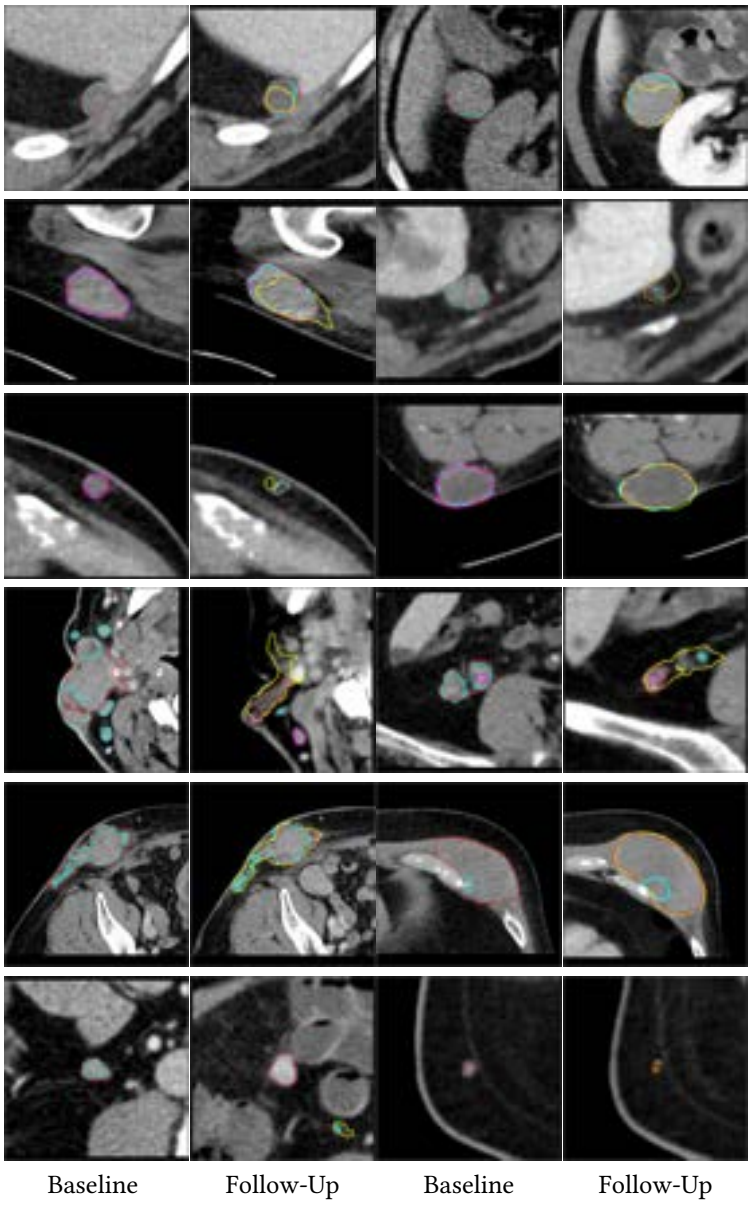



Figure 6.3: Visual examples of results produced by our method. Each example includes the baseline and follow-up lesion and therefore consists of two images (left baseline, right follow-up). On the baseline image, the manual annotation (red curve ■) and the nnU-Net annotation trained on the ROI (blue curve ■) and trained on the whole image (pink curve ■) are shown. On the follow-up image, the manual annotation (red curve ■), the propagated lesion (yellow curve ■) and the results of the presented pipeline (blue curve ■) are visualized.

one corresponding lesion in the follow-up image. This does not always have to be true. Moreover, lesions that are very close to each other could be wrongly assigned in the follow-up scan. These problems will be solved in future work by integrating consistency rules. Besides, our pipeline is not yet capable of detecting new lesions in the follow-up scan. Furthermore, the current pipeline does not take the appearance of the baseline lesion into account. There are different approaches to integrate this information into the model. The transformed baseline image and the corresponding lesion mask could be used as an additional input for the follow-up model. However, this would mean that two models have to be trained; one for segmenting the baseline image and one for the follow-up images. To train the follow-up network, a sufficient number of lesion annotations has to be available. Unfortunately, we only have the annotations that we used for the evaluations and therefore this approach is not suitable. Another approach is a joint-segmentation-registration algorithm as in [156]. We will explore this approach in future work.

We have trained and evaluated our method on soft-tissue lesions, which are particularly challenging due to their diverse appearance and location. Our promising results suggest that we will be able to extend our approach to other lesion types as well. Additionally, for use in clinical routine, it is sufficient to extract the largest diameter from the segmentation, so that detailed corrections will not be necessary. With our work, we have laid the foundation for an efficient automated follow-up assessment according to the RECIST standard and implementation of automated segmentation for Radiomics analysis in clinical routine.

## Acknowledgments

We thank Fabian Isensee, Paul Jäger, Simon Kohl, Jens Petersen, and Klaus Maier-Hein for providing the nnU-Net framework. The research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 428216905 / SPP 2177.

# Appendix

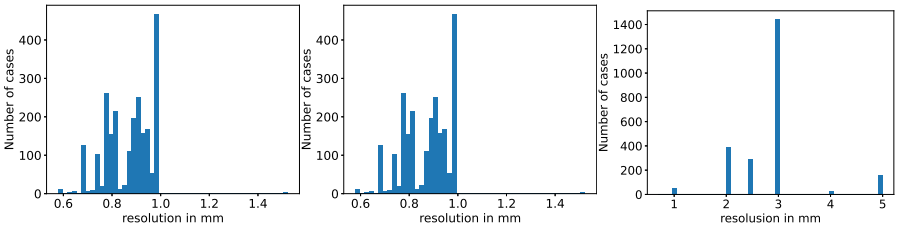



Figure 6.4: Histogram of the image resolution in x-, y- and z-direction.

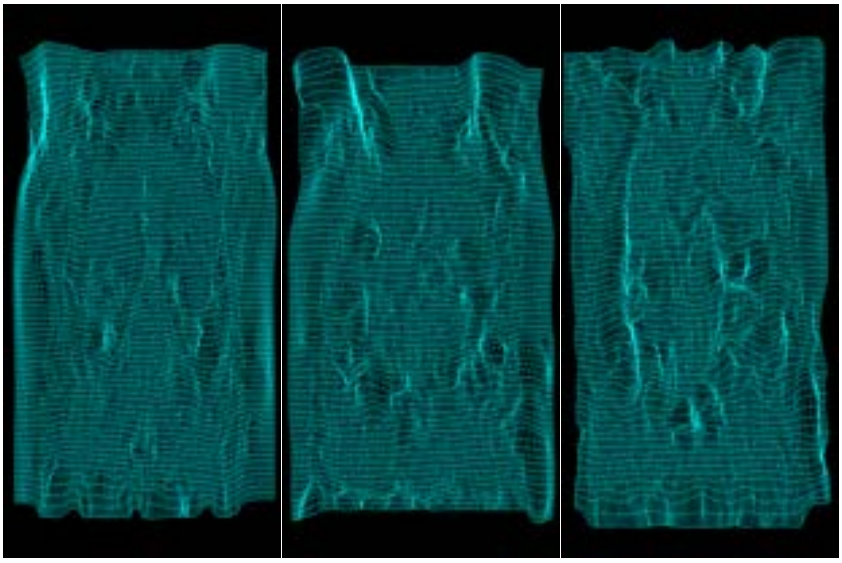


Figure 6.5: Example coronal slices extracted from three deformation fields to give an impression of the smoothness.

Table 6.1: Main settings chosen by the nnUNet framework to train the segmentation network

| Name | Description | Parameter |
|---|---|---|
| net_pool_per_axis | number of pooling operations in z,x,y direction | 3,5,5 |
| base_num_features | number of features after first conv | 32 |
| conv_per_stage | | 2 |
| optimizer | | SGD |
| learning rate | | ≈ 0.00235 |
| max_num_epochs | maximal number of epochs | 1000 |
| num_batches_per_epoch | number of batches used in every epoch | 250 |
| batch_size | number of images per batch | 5 |
| patch_size | z,y,z direction | 56× 128× 128 |
| normalization_schemes | see [6] for details on CT scheme | (0,'CT') |

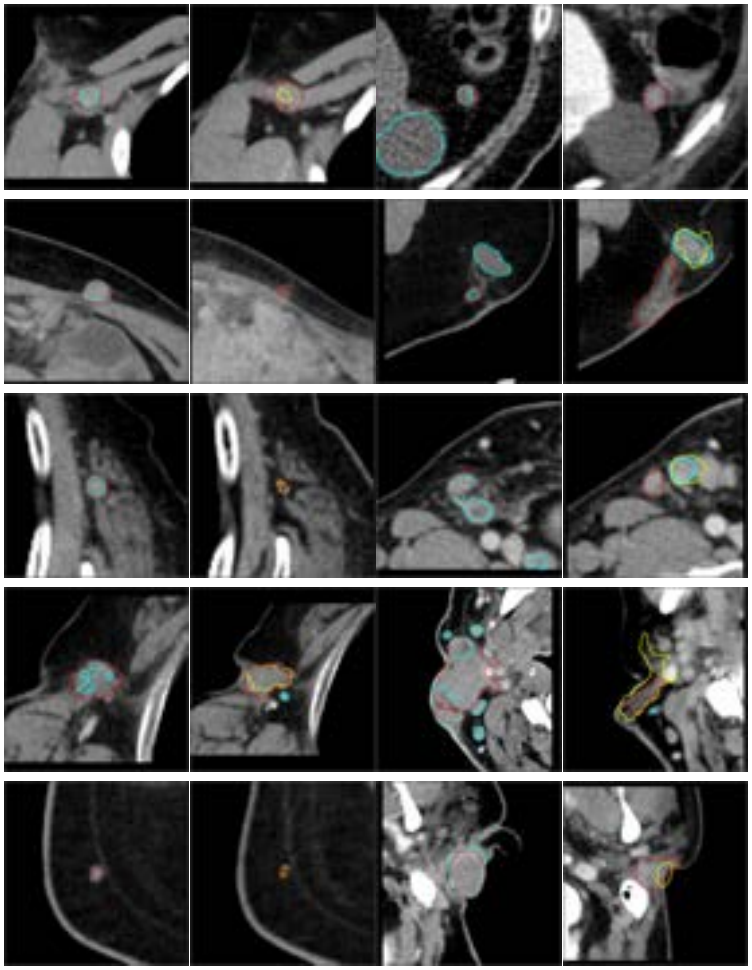



Figure 6.6: Cases in which our pipeline fails to segment the lesion in the follow-up image. Each example includes the baseline and follow-up lesion and therefore consists of two images (left baseline, right follow-up). On the baseline image, the manual annotation (red curve ■) and the nnU-Net annotation trained on the ROI (blue curve ■) are shown. On the follow-up image, the manual annotation (red curve ■), the propagated lesion (yellow curve ■) and the results of the presented pipeline (blue curve ■) are visualized. For these cases, we do not apply the lesion selection and therefore some lesions seem to be correctly segmented, however, they are not selected using our criteria. There are different reasons for these failures. In some cases, the registration was not accurate enough and therefore a wrong or no lesion was segmented. Some lesions are hard to distinguish from surrounding tissue (e.g. last column), but also an untypical shape can be a problem.

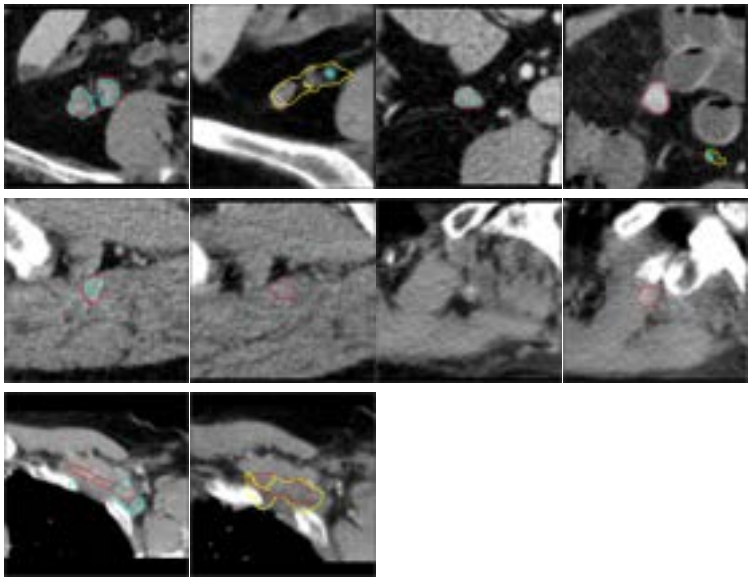



Figure 6.7: Cases in which our pipeline fails to segment the lesion in the follow-up image. Each example includes the baseline and follow-up lesion and therefore consists of two images (left baseline, right follow-up). On the baseline image, the manual annotation (red curve ■) and the nnU-Net annotation trained on the ROI (blue curve ■) are shown. On the follow-up image, the manual annotation (red curve ■), the propagated lesion (yellow curve ■) and the results of the presented pipeline (blue curve ■) are visualized. For these cases, we do not apply the lesion selection and therefore some lesions seem to be correctly segmented, however, they are not selected using our criteria. There are different reasons for these failures. In some cases, the registration was not accurate enough and therefore a wrong or no lesion was segmented. Some lesions are hard to distinguish from surrounding tissue (e.g. last column), but also an untypical shape can be a problem.

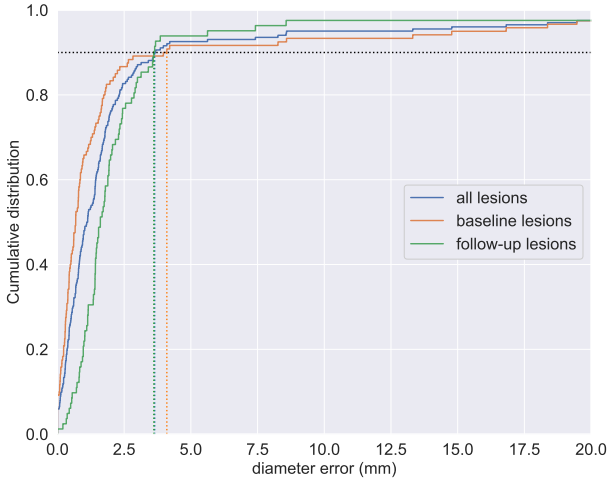



Figure 6.8: Cumulative distribution of diameter error between the manual segmentation and the nnU-Net segmentation. Please note, that in clinical routine the diameter would not be calculated from a segmentation but measured directly which might also introduce some errors. The dotted lines visualize the 90th percentiles of the error, which are 3.6 mm for all lesions, 4.1 mm for the baseline lesions and 3.6 mm for the follow-up lesions.

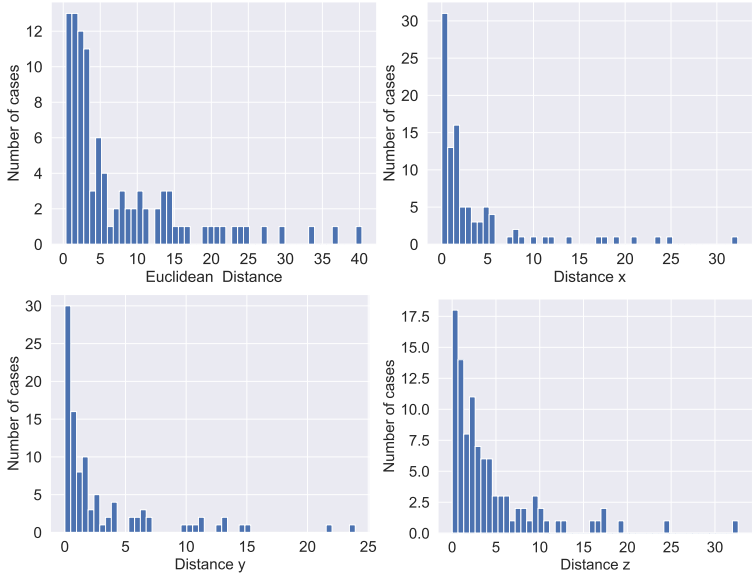



Figure 6.9: Histogram of Euclidean distance and the absolute offset between the center of the propagated lesion and the center of the manually annotated follow-up lesion..

7

# CHAPTER 7

# Workflow-centered evaluation of AI-assisted lesion tracking and segmentation

**BASED ON:** A. Hering, M. Westphal, A. Hänsch, H. Almansour, M. Maurer, T. Kohlbrandt, T. Eigentler, T. Amaral, et al. "Workflow-centered evaluation of an AI-assisted lesion tracking and segmentation software for soft-tissue and lymph node metastases in follow-up CT scans," *in preparation* (2022).

## Abstract

**BACKGROUND** Automated registration and segmentation of lesions in follow up CT scans meets a growing field of application. Limited evidence for AI-assisted lymph node and soft tissue metastases segmentation exists.

**PURPOSE** To evaluate the time efficiency, inter-reader variability and quality of an AI-assisted workflow for registration and segmentation of lymph node and soft tissue metastases in follow-up CTs.

**MATERIALS AND METHODS** 1842 lymph node and 2729 soft tissue metastases from 272 stage IV melanoma patients were retrospectively identified. Manual segmentation served as a reference standard. Results of AI-assisted and manual segmentation by two radiologists were analyzed with focus on time efficiency, inter-reader variability, Dice scores, sensitivity, and positive predictive value (PPV).

**RESULTS** AI-assisted segmentation achieves a significant reduction of user interaction time compared to manual segmentation (1.6 min vs 3.6 min). Results show a high AI-assisted inter-reader agreement (median Dice 1.0 vs 0.82 ). AI-assisted segmentation achieves similar Dice scores (0.83 –0.84 vs 0.82 ), sensitivity (0.93 –0.96 vs 0.92 –0.93 ) and PPV (0.95 vs 0.96 ) compared to manual segmentation.

**CONCLUSION** AI-assisted follow up CT registration and segmentation of lymph node and soft tissue metastases is applicable and significantly reduces the reader's interaction time, as well as inter-reader variability with similar segmentation quality compared to manual segmentation.

**SUMMARY** AI-assisted follow up CT registration and segmentation of lymph node and soft tissue metastases is similarly accurate, less variable and more time efficient compared to manual segmentation.

## 7.1 Introduction

To evaluate the efficacy of cancer treatment, measurement of metastatic tumors on longitudinal computer tomography (CT) scans is essential. Manual measurements for the diameter-based RECIST (Response Evaluation Criteria In Solid Tumors) criteria [144] are often time-consuming and error-prone. However, those criteria and the execution of the measurements undergo continuous changes. Lesion segmentation assistance based on artificial intelligence (AI) might significantly speed up response evaluation and help to handle the ever-growing mass of image-based staging and follow-up evaluations [145].

Additionally, radiomics is a promising topic in radiology. The extraction of multiple quantitative features from medical images obtained from CT, MRI, or PET [157] resulting in the conversion of medical images into minable data and the subsequent analysis promise new insights into therapy response and hold the potential to revolutionize medical image-based evaluation techniques [146].

Both fields have a huge clinical impact, however, share a common bottleneck: an accurate lesion segmentation obtained with minimal manual effort.

U-Nets [53] are one of the current states of the art approaches in deep learning and an established and preferred method for segmentation [6, 158]. Whilst there are many successful applications for organs such as the liver [149], only a few studies investigated the segmentation of lesions, such as lymph nodes [159] and, to our knowledge, no study has evaluated an application for soft tissue metastases, yet. Soft tissue metastases are very common in melanoma patients, however, they provide a particular hurdle for image evaluation, as they can arise in a variety of locations (cutaneous, subcutaneous, muscular, retroperitoneal) and shapes (round, multilobular, well defined, invasive), are often primarily small and, if not surrounded by fatty tissue, extremely hard to distinguish.

The present study evaluates the practical application of a recently introduced U-Net based pipeline [160] for automated registration and segmentation of soft tissue metastases in follow up CTs and extends it to lymph node segmentations. The study's focus was to test the efficacy and applicability of an AI-assisted segmentation pipeline for lymph node and soft tissue metastases in follow up CTs of stage IV melanoma patients. The detection of new metastases was not the scope of the present study. Thus, the three hypotheses of the study were:

1. Assessment time for follow up lesion segmentation is reduced using an AI-assisted workflow.
2. The inter-reader variability of the resulting segmentations is reduced with AI-assistance.
3. The quality of the AI-assisted segmentation is non-inferior to the quality of fully manual segmentation.

| | Total Dataset | Training&Validation | Testing |
|---|---|---|---|
| Age (years,SD) | 63.6 (14.6) | 63.4 (14.9) | 64.7 (13.2) |
| Gender (female) | 44% | 44% | 37% |
| Stage IV (AJCC 8th Edition) | 100% | 100% | 100% |
| Immunotherapy | 72% | 73% | 67% |
| Targeted therapy | 28% | 27% | 33% |
| Number of lesion (total) | 4571 | 4308 | 263 |
| Lymph node | 1842 | 1705 | 137 |
| Soft tissue | 2729 | 2603 | 126 |
| Average number of lesions (per patient) | 7 | 10 | 5 |
| Lymph node | 3 | 4 | 3 |
| Soft tissue | 4 | 6 | 2 |
| Inhouse CT | 81% | 78% | 86% |
| External CT | 19% | 22% | 14% |

Table 7.1: Demographics of the patient collective.

## 7.2 Material and Methods

### 7.2.1 Study Design and Subjects

The Central Malignant Melanoma Registry in Germany (CMMR) was used to retrospectively identify patients diagnosed with stage IV melanoma between 2015-01-01 and 2018-12-31, that were first-line treated at the department of dermatology of the University Hospital Tuebingen, a tertiary referral center for melanoma patients. The institutional ethics board approved the study protocol. Informed consent was waived due to the retrospective study design.

Inclusion criteria were: 1.) patients with stage IV melanoma, 2.) who had undergone baseline (pretreatment) contrast enhanced CT, 3.) who showed lymph node and/or soft tissue metastasis at baseline CT, 4.) with available first follow-up CT after therapy initiation. Inclusion criteria were: 1.) patients with stage IV melanoma, 2.) who had undergone baseline (pretreatment) contrast enhanced CT, 3.) who showed lymph node and/or soft tissue metastasis at baseline CT, 4.) with available first follow up CT after therapy initiation. The study included a training and validation stage for various types of soft tissue and lymph node metastases in baseline and follow-up CTs, as well as an interactive testing stage for follow-up CT segmentation of soft tissue and lymph node metastases in a randomly selected imaging dataset of patients. See Figure 7.1 for a detailed study workflow and Table 7.1 for patient characteristics.

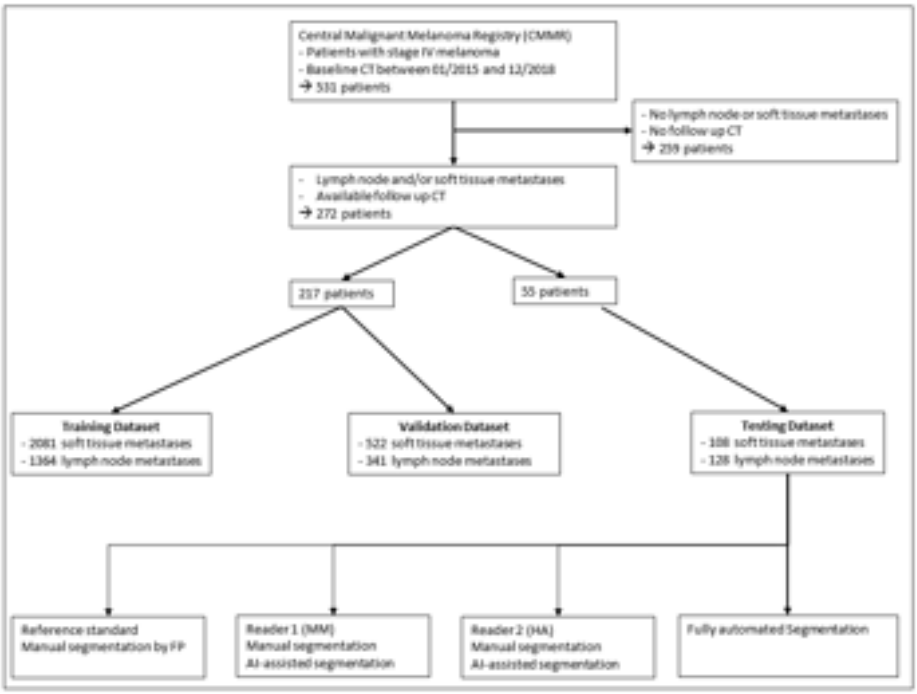



Figure 7.1: Study workflow and selection and split of used dataset.

### 7.2.2 Training and Validation Dataset

Contrast–enhanced baseline and first follow-up CTs were used for training and validation. The training and validation sets included 4308 lesions (2603 soft tissue and 1705 lymph node lesions) split into 3461 (2081 soft tissue and 1364 lymph node) and 866 (522 soft tissue and 341 lymph node) lesions for training and validation, respectively. Patients had various numbers of soft tissue and or lymph node lesions. The datasets included cases from different institutions and were therefore obtained with different CT scanners with various protocols. Typical CT imaging parameters used in our center for staging of melanoma patients are reported in Table 7.7 (appendix). All training and validation segmentations were manually conducted by an experienced resident radiologist (F.P. 4 years) under supervision of A.O and S.G, two senior radiologists with extensive experience in oncologic imaging (A.O. 8 years and S.G. 9 years) using a custom-made image post processing software system (SATORI; Fraunhofer MEVIS, Bremen).

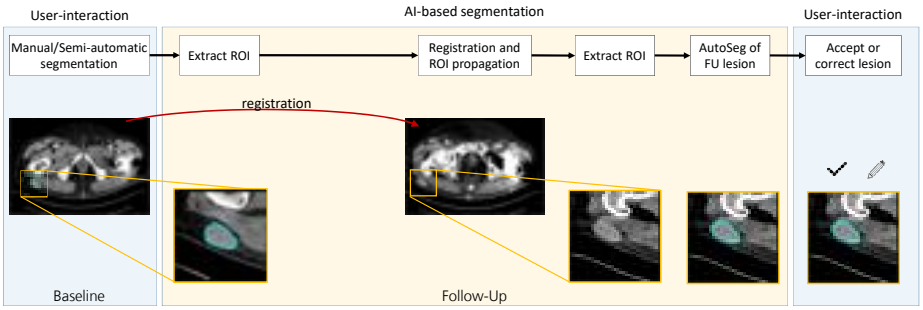



Figure 7.2: Schema of the proposed pipeline for AI-assisted segmentation of lymph node and soft tissue metastases in follow-up CT scans.

### 7.2.3 Testing Dataset

The testing dataset included 55 CTs of patients referred for the first follow-up CT after therapy start that were not included in the training and validation dataset. The patients were randomly selected from the CMMR. For demographic details please refer to Table 7.1. The lesions were stratified by diameter size in the follow-up scan smaller than 10 mm (n = 58), 10 –20 mm (n = 94), and larger than 20 mm in diameter (n = 55), with a mean size of 17.9 mm ±15.2 mm (range: 5.0 – 140.5 mm). 54 lesions showed complete response.

### 7.2.4 Automated Segmentation Pipeline

Baseline segmentation was performed manually by a reference reader (FP). The pipeline then consisted of the following steps: 1.) Extraction of the region of interest (ROI) around the lesion in the baseline scan; 2.) Registration of the baseline to the follow-up image; 3.) Propagation of the ROI to the follow-up image to constrain the search region and apply the trained U-Net to this region; 4.) Selection of the corresponding lesion in the output of the U-Net. See figure 2 for an overview of the proposed pipeline.

The nnU-Net framework [6] was used to train the U-Net [53]. The user could accept or correct the proposed segmentation. If the network did not segment a lesion or segmented a lesion with a diameter smaller than 5 mm, an empty mask was stored that contained the information of the propagated center of gravity of the baseline lesion. Thus, the corresponding region could be displayed by selecting the lesion. Furthermore, the lesion was considered to have disappeared under therapy. Extensive technical details are published in a previous publication [160] and summarized in the appendix.

### 7.2.5 Manual and AI-Assisted Segmentation of the Testing Dataset

For each examination, baseline and follow-up CTs were imported into a custom-made image post processing software system (SATORI; Fraunhofer MEVIS, Bremen) as Digital Imaging and Communications in Medicine (DICOM) files. Radiologists were able to see the baseline CT with the already segmented metastases and the follow-up CT in one shared window (see 7.3). For the manual workflow, masks were created by manually segmenting the lesions on the follow-up CT images using a cursor to contour the lesions with optional interpolation. For AI-assisted segmentations, follow-up CT examinations with lesion masks created by the proposed pipeline were imported into SATORI, and each mask was manually edited by radiologists. They had the choice to (a) accept the automated segmentation as perfect and move on to the next metastasis, (b) accept the automated segmentation as passable and make manual corrections on various slides using a cursor or (c) dismiss the automated segmentation and perform a manual segmentation instead. In the case of metastasis showing complete response in follow-up CTs, the U-Net was supposed to create an empty mask. If a segmentation was falsely computed, the radiologists had the possibility to reject the proposed mask and save an empty mask instead. To assess inter-reader agreement and inter-method variability, the testing set was independently segmented by two radiologists (HA (resident) and MM (specialist), with 2 years and 7 years of experience in oncologic radiology respectively) via the following schema: Firstly, they manually segmented the first 50 % of the testing cohort, followed by the AI-assisted segmentation of the second 50 % of the testing cohort. Two weeks later, they performed AI-assisted segmentation the first 50 % of the testing cohort followed by the manual segmentation of the second 50 % of the cohort to account for recall bias. The patients were sorted by ID and not by number of metastases to account for a random selection. HA and MM were blinded to their previous segmentation results, those of the other reader, and to the reference standard of the follow-up examinations. The reference standard for the testing cohort was manual segmentation of the 126 soft tissue metastases and 137 lymph node metastases by FP under supervision of AO and SG.

## 7.3 Performance Metrics

### 7.3.1 Detection Performance

The study evaluated the annotation of lesions in follow up CTs, given a baseline lesion segmentation. Therefore, lesions could either be present in the follow-up scan or disappear under therapy. The detection performance was evaluated against the reference standard with the following categories:

- true positive (TP): lesions both annotated by the reference reader and the evaluated method.

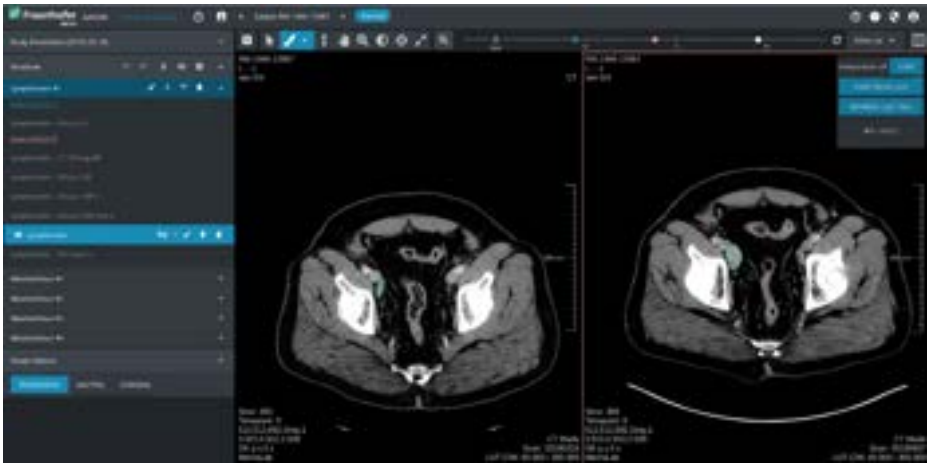


Figure 7.3: Annotation platform SATORI in the assisted session. All lesions annotated on the baseline scan are listed on the left. On the baseline scan (left axial CT reconstruction), only the reference segmentation is shown. For the follow-up study (right axial CT reconstruction), for all series an automatically computed lesion is imported. The user can accept the lesion or manually correct it.

- false negative (FN): lesions annotated by the reference reader but marked as disappeared by the evaluated method.
- false positive (FP): lesion marked as disappeared by the reference reader but annotated by the evaluated method.
- true negative (TN): lesions marked as disappeared both by the reference reader and the evaluated method.

Furthermore, the sensitivity (TP/(TP+FN)) for lesions <10 mm, 10 –20 mm and >20 mm and all lesions as well as the positive predictive value (PPV = TP/(TP+FP) for all lesions were calculated per method.

### 7.3.2 Time Efficiency

User interaction time was recorded for manual segmentations and for manual corrections of automated segmentations per patient.

### 7.3.3 Segmentation Accuracy

Accuracy was evaluated against the reference standard and assessed by using the Dice similarity coefficient for reference lesions that have not disappeared in the follow-up scan. The average symmetric surface distance was evaluated for all reference lesions

detected by both the reader and the reference reader. Measure definitions and formulas are given in the appendix.

### 7.3.4 Segmentation agreement

The inter-reader and inter-method variability of the segmentation accuracy were evaluated using the Dice score.

### 7.3.5 Statistical Analysis

The statistical analysis targeted two (co-primary) endpoints: processing time (seconds) and segmentation accuracy (Dice score). Our initial hypotheses were that the assisted workflow is faster and non-inferior with respect to the Dice score compared to the manual workflow. We considered an average Dice score loss of up to 0.01 as non-inferior. In the secondary analysis, we investigate if the agreement of reader improves with the assisted workflow compared to the manual annotation.

The entire analysis was restricted to two readers. The analysis does not aim to quantify uncertainty with respect to generalizing to new readers but rather to new patients only. For both analyses, a Bayesian hierarchical (mixed-effect) generalized linear model was fit with the statistical software R (version 4.1.2) and the brms package (version 2.16.3). For the processing time analysis, each patient formed an observation. For the Dice analysis, the observation unit was a single lesion. The hierarchical data structure was considered by the statistical model. For the time data, the lognormal distribution was used to model the positive outcome. For the Dice data a zero-inflated Beta regression was performed to adequately deal with the Dice score contained in the unit interval. The brms default (flat or weakly informative) prior distributions were utilized for all analyses. We quantified uncertainty with 95% posterior highest density intervals (HDI). In addition, the posterior probability of each research hypothesis is reported.

## 7.4 Results

### 7.4.1 Detection Performance and Sensitivity

The detection performance and sensitivity for manual, AI-assisted and fully automated segmentation are summarized in Table 7.2 and 7.3. The sensitivity of the fully automated segmentations was lower for small lesions than for larger lesions. For AI-assisted segmentations, there were only negligible differences regarding the lesion size. PPV was highest for manual segmentation (0.96 ) and AI-assisted segmentation (0.95 ), and slightly lower for fully automated segmentation (0.91 ).

| Method | User | TP | FN | FP | TN | PPV |
|---|---|---|---|---|---|---|
| Manual | M1 (MM) | 192 | 17 | 8 | 46 | 0.96 |
| | M2 (HA) | 191 | 18 | 8 | 46 | 0.96 |
| Assisted | A1 (MM) | 198 | 11 | 10 | 44 | 0.95 |
| | A2 (HA) | 194 | 15 | 11 | 44 | 0.95 |
| Automatic | | 180 | 29 | 18 | 36 | 0.91 |

Table 7.2: Detection Performance. Segmentation performed automatically, manual (M) and AI-assisted (A) by reader 1 (MM) and 2 (HA). TP = true positive, FN = false negative, FP = false positive, TN = true negative.

| Method | User | all | < 10 mm | 10 – 20 mm | > 20 mm |
|---|---|---|---|---|---|
| Manual | M1 (MM) | 0.92 [0.87, 0.95] | 0.85 [0.74, 0.92] | 0.94 [0.87, 0.97] | 0.96 [0.88, 0.99] |
| | M2 (HA) | 0.91 [0.87, 0.94] | 0.85 [0.74, 0.92] | 0.91 [0.84, 0.96] | 0.98 [0.90, 1.0] |
| Assisted | A1 (MM) | 0.95 [0.91, 0.97] | 0.90 [0.80, 0.95] | 0.95 [0.88, 0.98] | 1.0 [0.93, 1.0] |
| | A2 (HA) | 0.93 [0.88, 0.96] | 0.90 [0.80, 0.95] | 0.93 [0.85, 0.96] | 0.96 [0.88, 0.99] |
| Automatic | | 0.86 [0.81, 0.90] | 0.78 [0.66, 0.87] | 0.87 [0.79, 0.93] | 0.93 [0.83, 0.97] |

Table 7.3: Sensitivity for segmentation performed automatically, manual (M) and AI-assisted (A) by reader 1 (MM) and 2 (HA). Data in parentheses are 95% confidence intervals. The results are differentiated by the diameter of the reference segmentation smaller than 10 mm, 10-20 mm and larger than 20 mm.

### 7.4.2 Efficiency

Mean interaction time was 3.6 ± 5.0 min. per patient for manual segmentation and 1.6 ± 1.6 min per patient for AI-assisted segmentation (M1: 3.5 ± 4.8 min; M2: 3.7 ± 5.3 min; A1: 1.7 ± 1.8 min; A2: 1.4 ± 1.4 min).

### 7.4.3 Segmentation Accuracy

The results are summarized in Table 7.4 and 7.5 and visualized in figure 7.4. With a median Dice of 0.82, 0.83-0.84, and 0.81 for manual, AI-assisted, and automated segmentation, respectively, all methods achieved comparable results. For small lesions <10 mm, the Dice score was slightly lower than for larger lesions. In figure 7.5, exemplary segmentation results of all readers are shown.

| Method | User | all | < 10 mm | 10 – 20 mm | > 20 mm |
|---|---|---|---|---|---|
| Manual | M1 (MM) | 0.82 [0.69, 0.86] | 0.79 [0.24, 0.82] | 0.81 [0.73, 0.86] | 0.85 [0.79, 0.89] |
| | M2 (HA) | 0.82 [0.63, 0.87] | 0.81 [0.24, 0.86] | 0.80 [0.60, 0.85] | 0.84 [0.76, 0.90] |
| Assisted | A1 (MM) | 0.84 [0.73, 0.89] | 0.80 [0.61, 0.86] | 0.84 [0.72, 0.88] | 0.87 [0.80, 0.91] |
| | A2 (HA) | 0.83 [0.72, 0.88] | 0.81 [0.63, 0.84] | 0.82 [0.71, 0.88] | 0.86 [0.79, 0.90] |
| Automatic | | 0.81 [0.43, 0.88] | 0.78 [0.0, 0.84] | 0.80 [0.60, 0.87] | 0.84 [0.87, 0.90] |

Table 7.4: Segmentation performance. Median Dice and 25%- and 75% quantile in parenthesis. The results are split by the diameter of the reference segmentation (<10mm, 10-20mm, > 20mm).

### 7.4.4 Segmentation Agreement

In Table 7.6, the inter-reader and inter-method agreement is summarized using the Dice score computed between the corresponding segmentation. The median Dice score of the segmentations generated by manual annotation (M1 to M2) was 0.82 In contrast, the assisted segmentation (A1 to A2) achieved a significantly higher median Dice score of 1.0 The inter-method agreement of the automatic segmentation to the respective AI-assisted segmentations (automatic to A1 and automatic to A2) also achieved a median Dice score of 1.0 This means that in more than 50% of the lesions, the reader accepted the segmentation without any further corrections.

### 7.4.5 Statistical Analysis

There is very strong evidence that the AI-assisted workflow is faster compared to the manual workflow. The posterior mean effect (assisted - manual) for the two reader are $-133.4$ (HA; 95 % HDI: $[-400.5, -2.8]$) and $-112.6$ (MM; 95% HDI: $[-337.8, -1.6]$) seconds, respectively. Both HDI lie below zero, supporting our research hypothesis. The posterior probability of superiority is estimated to be at least 0.999 for each reader.

Regarding accuracy, there is strong evidence for a non-inferior Dice Score for the assisted workflow compared to the manual workflow. The posterior mean effect (assisted-manual) for the two reader are 0.008 (HA; 95 % HDI: $[-0.005, 0.024]$) and 0.011 (MM; 95 % HDI: $[0.002, 0.027]$), respectively. Both HDI lie above the non-inferiority margin of $-0.01$, supporting our research hypothesis. The posterior probability of non-inferiority of the assissted workflow is estimated to be at least 99 % (HA: 0.997 ; MM: 0.999 ).

Inter-reader agreement was measured with the Dice score of annotations by the two readers after assisted and manual workflow, respectively. Here, the posterior mean

| Method | User | all | < 10 mm | 10 – 20 mm | > 20 mm |
|---|---|---|---|---|---|
| Manual | M1 (MM) | 0.43 [0.27, 1.] | 0.27 [0.22, 0.53] | 0.45 [0.28, 0.94] | 0.66 [0.37, 1.2] |
| | M2 (HA) | 0.50 [0.25, 1.3] | 0.22 [0.16, 1.1] | 0.50 [0.28, 1.0] | 0.81 [0.41, 1.4] |
| Assisted | A1 (MM) | 0.39 [0.24, 0.86] | 0.28 [0.2, 0.75] | 0.46 [0.24, 0.76] | 0.47 [0.31, 0.9] |
| | A2 (HA) | 0.43 [0.24, 0.88] | 0.28 [0.20, 0.64] | 0.50 [0.23, 0.79] | 0.67 [0.32, 1.2] |
| Automatic | | 0.52 [0.27, 1.3] | 0.33 [0.22, 0.96] | 0.53 [0.24, 1.1] | 0.78 [0.40, 2.6] |

Table 7.5: Segmentation performance. Median Average Surface Distance and 25%- and 75% quantile in parenthesis. The results are split by the diameter of the reference segmentation (<10mm, 10-20mm, > 20mm).

effect (assisted-manual) is 0.050 (95 %-HDI: 0.026 , 0.074 ) in favour of the assisted workflow. The HDI lies above zero, supporting the research hypothesis of an improvement of the inter-reader agreement. Accordingly, the posterior probability of superiority is estimated to be at least 0.999

## 7.5 Discussion

The study's purpose was to evaluate the practical application of an AI-assisted registration and segmentation pipeline for lymph node and soft tissue metastases in follow-up CTs. With the proposed pipeline, mean interaction time for lesion segmentation was significantly reduced from 3.6 min to 1.6 min using AI-assisted segmentation. Vorontsov et al. reported similar effects for the correction of fully automated segmentation of liver lesions in CTs of patients with colorectal cancer liver metastasis using a CNN [161]. Moltz et al. investigated a simpler algorithm for automatic lesion tracking and segmentation in follow-up CTs for lung nodules, liver metastases and lymph nodes and reported a reduction of assessment time through lesion tracking, too [153].

Regarding accuracy, there was strong evidence for a non-inferior Dice score for the assisted workflow compared to the manual workflow. This is in line with a previous publication evaluating automated lesion tracking and segmentation of lung nodules, liver metastases and lymph nodes [153]. Both readers achieved their highest Dice scores in comparison to the reference segmentation using the AI-assisted segmentation pipeline (0.84 and 0.83 , respectively). This effect was present for all three categories of lesion size (see Table 7.4). The achieved Dice scores were comparable to results reported by authors investigating the automated segmentation of liver metastases [162, 163]. The

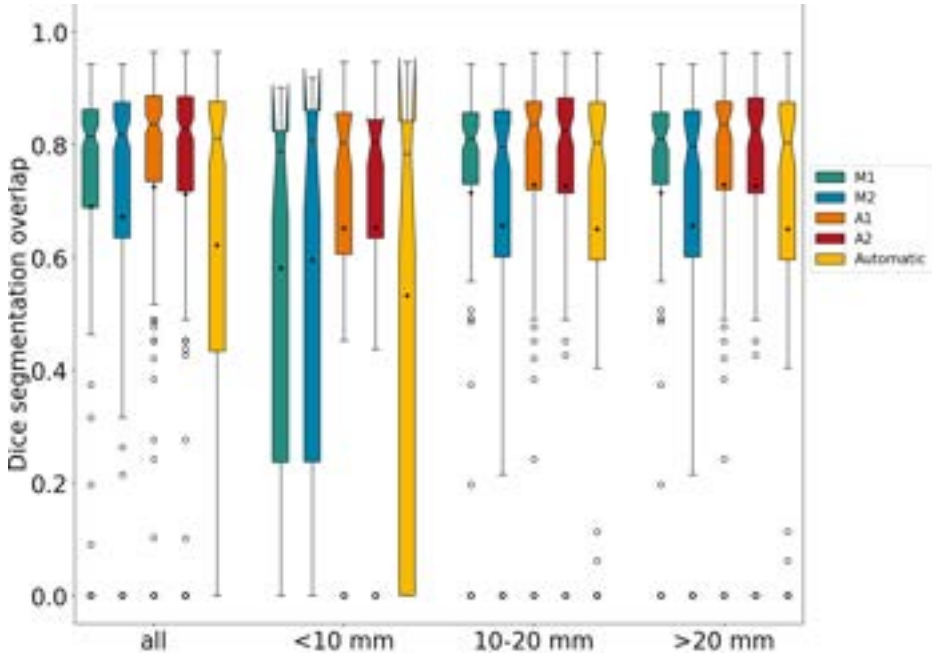



Figure 7.4: Notched boxplots of the Dice score for manual segmentations (M1 and M2), AI-assisted segmentations (A1 and A2) and the fully automated segmentations evaluated against the reference standard split by the diameter of the reference segmentation (<10mm, 10-20mm, >20mm). Mean is symbolized by black dots, median by black horizontal lines.

average surface distance was lowest for AI-assisted segmentations (0.39 mm-0.43 mm). Compared to manual segmentations, the PPV was virtually equal for both readers using AI-assisted segmentation (0.96 vs 0.95 ). However, it is possible that there is a bias of the automated segmentations and thus also the assisted segmentations towards the reference segmentation, since the network for the automatic segmentation was trained on annotations of the reference reader (on a separate training data set).

Furthermore, we found that AI-assisted segmentation reduces inter-reader variability. The inter-reader agreement for the AI-assisted measurements was significantly higher than for manual measurements (A1 vs A2 median Dice 1.0 ; M1 vs M2 median Dice 0.8 (see table 7.6). The reduction of inter-reader variability through AI-assisted segmentation is a well described effect and was reproduced in several publications [153, 164–166]. Over 50 % of the segmentation propositions were accepted with no further corrections.

Fully automated segmentation achieved a similar median Dice score compared to manual segmentation (0.82 vs 0.81 ), but slightly lower compared to AI-assisted segmentation (0.81 vs 0.84 and 0.81 vs 0.83 for reader 1 and 2, respectively). Lower Dice scores were especially present for small lesions <10 mm. Vorontsov et al. reported

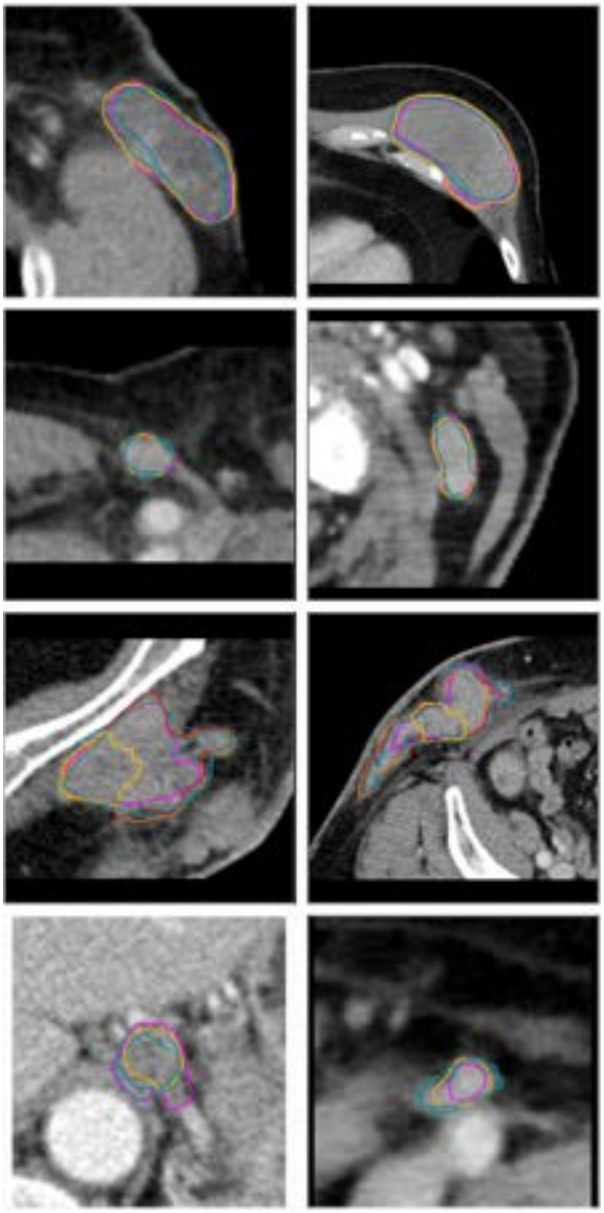


Figure 7.5: Exemplary segmentation results for the reference reader (pink), the manual segmentation M1 (green) and M2 (blue), the assisted segmentation A1 (orange) and A2 (red) as well as the automatic segmentation (yellow). If no assisted segmentation is shown, it is the same as the automatic segmentation. In the examples shown in the first two rows, there is a high similarity of annotations between the readers. The examples in the last two rows show a higher variability.

| **Method** | Reference | M1 | M2 | A1 | A2 | Automatic |
|---|---|---|---|---|---|---|
| Reference | | 0.82 [0.69, 0.86] | 0.82 [0.63, 0.87] | 0.84 [0.73, 0.89] | 0.83 [0.72, 0.88] | 0.81 [0.43, 0.88] |
| M1 | 0.82 [0.69, 0.86] | | 0.82 [0.66, 0.87] | 0.83 [0.71, 0.87] | 0.81 [0.66, 0.87] | 0.78 [0.07, 0.86] |
| M2 | 0.82 [0.63, 0.87] | 0.82 [0.66, 0.87] | | 0.83 [0.64, 0.87] | 0.82 [0.58, 0.87] | 0.79 [0.00, 0.87] |
| A1 | 0.84 [0.73, 0.89] | 0.83 [0.71, 0.87] | 0.83 [0.64, 0.87] | | 1.0 [0.86, 1.0] | 1.0 [0.68, 1.0] |
| A2 | 0.83 [0.72, 0.88] | 0.81 [0.66, 0.87] | 0.82 [0.58, 0.87] | 1.0 [0.86, 1.0] | | 1.0 [0.64, 1.0] |
| Automatic | 0.81 [0.43, 0.88] | 0.78 [0.07, 0.86] | 0.79 [0.00, 0.87] | 1.0 [0.86, 1.0] | 1.0 [0.64, 1.0] | |

Table 7.6: Inter-reader and inter-method agreement displayed by the Dice score computed between the segmentations generated with the respective methods. The median and the 25%- and 75% quantile in parentheses are reported.

similar results [161]. This can be explained by the fact that even a few voxels deviations account for a large percentage in small lesions and user correction attenuates this effect, especially in small lesions.

Our results show that AI-assisted and fully automated segmentation perform very well, whilst only AI-assisted segmentation performs comparable to manual segmentation in all categories. This can be explained through the pipeline that was used. The proposed segmentations for AI-assisted and fully automated segmentation were the same and initially very good, as the numbers for fully automated segmentation show, most likely because a patch-based segmentation pipeline was used. The registration algorithm restricts the search region to an assumed region (patch) in which the segmentation algorithm then tried to identify and outline the given lesion. Such a procedure requires a very precise registration method, otherwise the lesion might not be located in the selected region. The scope of the algorithm was not to identify new lesions. This resulted in an initially very high sensitivity and PPV. Readers could then adjust the already good segmentations and therefore further improve sensitivity, PPV, Dice score and average surface distance.

Summing up, the implemented pipeline significantly reduced the reader's interaction time and inter-reader variability, whilst the quality of the resulting segmentations was not inferior to manual segmentations. This has a huge clinical impact, as several radiological techniques, such as RECIST-measurements and radiomics analysis heavily rely on fast and accurate segmentations [144, 146, 157], and new approaches are in demand to reduce manual effort.

The study has limitations. The study focused on lymph nodes and soft tissue lesions in CT only. However, since the segmentation approach is based on machine learning, other lesion types and imaging modalities could likely be added by providing a sufficient number of reference segmentation in those modalities. The present study used manual segmentations by only one experienced reader as a gold standard. The pipeline focused on the segmentation of lesions in follow up CTs that were already segmented on baseline CTs. New lesions were not detected, and the applied algorithm was not trained and tested for that task. A potential solution might be AI-assisted one click segmentation [167] or a fully automatically detection. An independent evaluation with additional readers is required to support the generalizability of our results to other readers. The analysis is conservative in the sense that further training with the assisted workflow might lead to an additional improvement in one or both outcomes over time.

## 7.6 Conclusion

Our findings support our research hypothesis of an assisted workflow which is superior with respect to processing time and non-inferior with respect to accuracy compared to the manual workflow. An independent evaluation with additional readers is required to support the generalizability of our results to other readers. The analysis is conservative in the sense that further training with the assisted workflow might lead to an additional improvement in one or both outcomes over time.

# Appendix

| | |
|---|---|
| Reference current | 240 mAs |
| Tube voltage | 120 kV |
| Collimation | 128 mm × 0.6 mm |
| Rotation time | 0.5 s |
| Pitch | 0.6 |
| Image reconstruction | Medium smooth kernel |
| Contrast medium phase | Portal venous |

Table 7.7: Inhouse standard CT parameters for melanoma whole body staging

## Detailed description of the segmentation pipeline:

**Registration:** The registration must align the global structures but at the same time be locally accurate enough for a precise lesion propagation. Therefore, we adopted a three-step approach to automatically register the baseline to the follow-up image: 1.) Translational alignment; 2.) Rigid registration; 3.) Deformable registration. Hereby, the registration pipeline starts with robust methods with fewer degrees of freedom and moves on to more precise, but less robust methods, which require better starting points due to their higher degrees of freedom.

**Lesion segmentation:** The registration was used to propagate the baseline mask to the follow-up scan. While this propagated mask may not be accurate enough due to size changes under therapy, it provides a good initial correspondence. To compensate for registration errors, the search region was enlarged by 50 mm in every direction to ensure that the corresponding lesion is inside the selected region and to include enough information for the CNN (nnU-Net framework [6]). The CNN was trained using the lymph node and soft-tissue lesions annotated in the baseline and follow-up scans of the training dataset. The validation data set was only used to monitor the training but not to select hyperparameters.

**Lesion selection:** The CNN was not constrained to segment only one lesion inside the selected region in the follow-up scan. Therefore, the lesion whose center was closest to the center of the propagated lesion was selected. To avoid annotation of wrong close by lesions in the case of complete response, the network accepted only segmentation of lesions if the Euclidean distance of the center was smaller than 25 mm to the propagated lesion center.

## Segmentation performance metrics

Dice Score (DSC) between two segmentation masks X and Y:

$$\mathrm{DSC}(X,Y) = \frac{2|\,\|X \cap Y\|}{\|Y\|\,\|Y\|} \tag{7.1}$$

Average Symmetric Surface Distance (ASD) between two surfaces $X_S$ and and $Y_S$ of segmentation masks X and Y:

$$\mathrm{ASD} = \frac{1}{|X_s| + |Y_s|}\left(\sum_{x \in X_S} d(x, Y_S) + \sum_{y \in Y_S} d(y, X_S)\right),$$

where d is the surface distance

$$d(x, Y_s) = \min_{y \in Y} d(x, y)$$

8

# CHAPTER 8

# Discussion

*This chapter discusses the contributions and advances we have made in the fields of deep-learning-based image registration and efficient tumor follow-up analysis. We also look ahead into the future. The chapter is structured as a series of questions and answers.*

## 8.1 Deep-Learning-Based Image Registration

In this section, we discuss the question of whether deep-learning-based registration methods are the methodology of choice and which registration method is the best. However, this question cannot be answered that easily, which is why we first discuss various aspects that are relevant in answering this question.

### Why is it so difficult to compare different registration methods?

In the last years, a large number of deep-learning-based registration approaches have been presented in which the authors mostly showed that their methods achieve at least state-of-the-art results within shorter execution time than conventional registration methods (e.g. [25, 33, 42]). Consequently, the question arises whether deep-learning-based registration methods are now the methodology of choice. The large number of papers presenting deep-learning-based registration suggests that those methods are the best registration methods and with that the methodology of choice. However, the results and comparisons presented in the literature do not allow such a conclusion for several reasons.

In some papers, the authors did not compare their method to conventional methods but only to previous deep learning-based methods. This is legitimate if the goal of the work is to further develop deep learning-based registration methods. However, this does not allow any statement about the general performance of the developed algorithm. In most papers, a comparison was made to one of the publicly available registration toolboxes or algorithms like FAIR [78], Elastix [16], ANTS [106], ITK [168], NiftyReg [20] or Deeds [56]. Those are mainly not tailored to a specific application meaning that the user has to find good hyper-parameters which is often difficult and the result hardly depends on them. Furthermore, not all conventional methods are publicly available so that it might not be possible to compare the proposed method to the best available method for this task.

Another difficulty regarding comparability lies in the data used for training and testing. For deep-learning-based image registration algorithms, the imaging modality and the body region to be registered are determined with the training data. Thus, it could only be shown in the respective papers that a certain body region can be registered well in the chosen modality, depending on the available training data. No general conclusion can be drawn as to whether these algorithms will also train similarly well on other training data.

This issue also holds for conventional registration methods for which the hyper-parameters are optimized on a training dataset. Furthermore, it is not uncommon that algorithms are evaluated on private test datasets or a new private training-test split of a public dataset so that other researchers cannot compare their work against it. Unfortunately, authors often draw generalized conclusions even though they only used a single dataset in their experiments and compared their results to only a few other methods. These issues also hold true for our own work presented in chapter 2 in which we trained and tested our network on our data split of the Multi-Modality Whole Heart Segmentation (MM-WHS) dataset [54], compared it against the conventional Deeds method [56] and claimed that our method "performs better than state-of-the-art conventional registration methods".

To overcome the problem of private data, we used the publicly available DIR-Lab [81] for the evaluation of our method in chapter 3 and 4. This benchmark dataset has enabled us to compare to a variety of other registration methods without conducting the experiments ourselves. Nevertheless, the COPDGene dataset [69] that we used for training and partially for evaluation, is only available to researchers after a formal request to a review board. However, we have shown in our work in chapter 4 how strong the effect of the training data can be. We reduced the target registration error on the DIR-Lab dataset of the Voxelmorph [33] method by half, only by training it on the COPDGene dataset instead of in a leave-one-out experiment as in [93]. This single example illustrates nicely that we should question results from the scientific literature more critically.

Due to the lack of large data sets for training and well-annotated datasets for validation for medical image registration approaches, only insufficient comparisons can be made that do not allow a generally valid statement on the best registration method based on the results in the papers presenting deep-learning-based and conventional image registration approaches.

In the past, several evaluation projects [103, 169, 170], benchmark datasets [64, 81, 105] and challenges [82, 110, 171, 172] have been presented to simplify the comparison of registration methods. Those challenges are mainly single-task focused and provide only the validation dataset. However, in order to fairly compare deep learning-based methods, a consistent training dataset is also necessary. Furthermore, it is desirable to evaluate existing algorithms for several tasks and to be able to make a generally valid statement about the performance.

For that reason, we introduced the Learn2Reg challenge [173] presented in chapter 5 to at least partially overcome the issues. The Learn2Reg challenge provides a multi-task medical image registration benchmark for the comprehensive characterization of deformable registration algorithms. We aimed to find an approach that works well on all (or at least on multiple tasks) that ideally self-configures itself comparable to the nnU-Net framework [6] for segmentation tasks. Nevertheless, a task-specific evaluation seems still be interesting to obtain the *best* registration method for that specific task.

But despite common training and validation datasets, the evaluation of registration methods is not straightforward. Due to the lack of ground-truth deformation fields discussed in chapter 1, the output of different registration approaches cannot directly be evaluated. Depending on the paper, different auxiliary metrics are used to evaluate the registration performance which further reduces comparability. This naturally leads to the question of how best to evaluate a registration method.

## How to best evaluate a registration method?

Evaluating the performance of image registration algorithms is a difficult task, because there is rarely a point-wise correspondence from one image to another available. Therefore, several auxiliary metrics have been introduced to evaluate the registration performance.

The evaluation of the registration accuracy is often based on automatic segmentation produced from image registration [170] or on the target registration error of landmarks annotated by an medical expert [174]. The underlying assumption is that when a registration method enables an accurate propagation of the segmentation masks or the landmarks, the method aligns important structures well and therefore produces meaningful deformation fields. However, when we only focus on the accuracy metric alone, we might get a wrong impression. For example, a registration method could perfectly align the landmarks but produce irregular and implausible deformation in between. Therefore, it is a common practice to also evaluate the plausibility of the deformation field. Since the registration methods are applied on medical images, the transformation should not yield to flipping or disappearing of tissue (apart from a few exceptions such as resections). With the Jacobian Determinant of the deformation field, local volume changes can be measured. A negative Jacobian Determinant of the deformation field means a local folding. The number or percentage of voxels with foldings in the computed deformation field is often reported as a measurement of the plausibility of the presented results [33, 82]. Although this metric gives a good impression of the quality of the deformation field, the deformation field can still be non-smooth despite the absence of foldings. Therefore, several papers additionally report the standard deviation of the Jacobian Determinant [29, 92, 96].

For clinical applications, the robustness of a registration method is also important, because the methods need to work well for nearly all patients in different hospitals and different countries acquired with different scanners. There is, again, no established measure on how to evaluate robustness and different papers suggest different measures. In chapter 4, we interpret the term robustness in the sense that we evaluated how our method performed on datasets on which the network was not trained. Therefore, we applied our registration network on the publicly available DIR-Lab dataset [81] and the EMPIRE10 challenge data [82] and showed that our network registers those images

well. Especially the sheep CT images, which are included in the EMPIRE10 set, are interesting to look at because sheep and human anatomy differ clearly.

Another important metric is the runtime of the algorithms. In many applications, registration would not have to be performed in real time, but the images should be registered without a long waiting time. In addition to these most commonly used metrics, there are a variety of other metrics like *intensity variance*, *inverse consistency* or *transitivity* [170] that can be used. In the Learn2Reg challenge, we used a complementary set including robustness, accuracy, plausibility, and speed, that follows the principles defined by the BIAS group [102] to ensure an evaluation as fair as possible.

With all these auxiliary metrics, the question arises of how to weigh them. For example, by directly applying a zero displacement field, meaning that we do not change anything at all, the smoothness metric and the runtime get good scores while the accuracy is not increased. In most papers, all the scores are reported side-by-side, however, in the Learn2Reg challenge, a weighting scheme was required to determine a winner. To take into account random noise effects, we only ranked a method higher if the results were statistically significantly better. Although this weighting and the evaluation, in general, is still far from perfect, it is a first step towards making the evaluation fair and transparent. For this purpose, all evaluation criteria were announced in advance. While the proposed evaluation scheme helped to improve the comparability of registration algorithms, it is still not perfect. As already described in chapter 5, the accuracy evaluation is in general limited by inter-observer noise and the difficulty of assessing registration accuracy based on segmentation overlap, which disregards the plausibility of correspondences along the surface or within the structure. Furthermore, the evaluation measures were kept as similar as possible across the tasks to avoid complicating the evaluation. However, outside of a challenge, one would have to take a closer look at what the actual purpose of the registration is and adjust the weighting of the measurements accordingly and select the method that fits best for this purpose. For example, for atlas-based segmentation or the generation of (noisy) labels for the training of a segmentation network, we are mostly interested in propagating the segmentation masks well to the new scan and do not care about foldings in the deformation field. Robustness will also be important, as ideally a reasonably good segmentation is needed even for difficult cases. The runtime is probably not all that important. However, if a lot of images are to be processed or real-time applications are considered, then registration should not take too long either. Even for the registration of the same body region, the requirements can be slightly different. For a lung registration to be able to perform a cursor synchronization, a registration accuracy of approximately 2mm could be sufficient. However, if the deformation field is used to calculate a difference image, the registration accuracy should be much higher.

Ultimately, this means that one must first think about the prerequisites and the

goals of a particular registration application and this, of course, requires certain domain knowledge. Nevertheless, the Learn2Reg challenge provides already a good selection of performance measures that follows the principles defined by the BIAS group [102].

### What is the best registration method (according to Learn2Reg)?

Due to all the issues mentioned above, it is difficult to make a definitive statement about the best registration method in general. For this reason, in this section, we focus the results of the Learn2Reg challenge. Nevertheless, the winners of the single-task challenge should also at least be mentioned. The registration method of [175] won ANHIR and [176] performed best on the CuRIOUS dataset. The DIS-CO approach of [65] takes first place in the EMPIRE10 challenge.

To find the best performing method for several tasks, we evaluated in the Learn2Reg challenge the eight algorithms that submitted a solution to at least four of the six tasks. This showed that three methods (convexAdam [122], LapIRN [130], MEVIS [133]) and a baseline method (corrField [123]) were shown to work robustly on all tasks with only minor adjustments to the hyperparameters. ConvexAdam was among the top 3 on each task and ranked first overall highlighting the importance of effective optimization and versatility of using learned semantic or hand-crafted MIND features depending on the application. LapIRN reached the overall second rank and yielded the best result for Hippocampus and OASIS. This demonstrates that a well-designed convolutional feed-forward network can outperform conventional approaches in particular for inter-patient tasks. MEVIS achieved third place overall, with top ranks in particular for Lung CT and Hippocampus based on a combination of NGF metric, curvature regularization, and L-BFGS optimization.

Furthermore, we evaluated the transferability to new datasets in the challenge by applying the submitted methods from the lung task to the DIR-Lab dataset[81]. We found that the conventional methods are directly applicable to this new dataset without any further hyper-parameter tuning. For the deep-learning-based method LapIRN, slightly worse results were obtained. One reason for this might be the limited amount of training data available for the lung task.

The runtime of the methods is evaluated in nearly all papers and is the main selling point for deep-learning-based image registration methods. They often need less than a second to register even large three-dimensional images whereas conventional registration methods typically require several minutes to compute the deformation field. However, we found that there is virtually no difference in computational speed for the best-performing methods. GPU acceleration brings down the computation cost of optimization-based methods to a few seconds for 3D registration. It was also shown in the work of [18–21] that the runtime of conventional methods can be significantly reduced by an efficient implementation.

In summary, the best deep-learning-based registration methods can now keep up with conventional methods in terms of registration accuracy and the fastest conventional methods are not slower than the deep-learning-based methods. By using additional information such as segmentation masks during training, deep-learning-based method can be tailored to one task without this information being available during inference. This is a great advantage over conventional methods. However, the transferability of the algorithms to new data sets is not yet considered fully satisfactory.

## What's next with Learn2Reg?

The aim of finding a self-configuring registration framework similar to the nnU-Net framework [6] could unfortunately not be fulfilled in Learn2Reg 2020 and 2021. Moreover, Learn2Reg has also only been able to provide a limited selection of tasks with sometimes very limited training data (e.g. for the lung task). Therefore, we will continue with Learn2Reg in 2022 with new tasks to further address those problems and to continuously increase the available benchmark datasets. The 2022 version of Learn2Reg is divided into three tasks.

The first task will again deal with a CT lung registration. This time, however, on a much larger data set and on follow-up scan pairs. Moreover, the task will be divided into two phases. In phase 1, participants train or tune their algorithms locally and submit the algorithms via grand-challenge. The best teams of this phase are invited to participate in phase 2. In phase 2, the participants submit a training docker that will be run by the organizers on a larger dataset that includes additional annotations that are not publicly available. The trained networks will be made available via grand-challenge to facilitate reproducibility and further use of the algorithms in the research community.

The second task replicates the challenge of 2021 and serves as a continuation of the benchmark, whereas the last new task explores the possibility of fully automatic self-configuring methods that learn their hyperparameters based on training and validation data and require no user interaction.

The third task aims to find a self-configuring registration framework. For this reason, no further training data is provided. The necessary structure of the data is given by task 2, so that a framework is expected in a Docker container, which can be used to train registration networks on several data sets.

## What is the future of image registration?

Conventional and deep-learning-based registration each have their advantages and disadvantages, but there is little difference between them in terms of overall performance. Consequently, the logical consequence is to combine both approaches to exploit the respective advantages. For example, a registration network that was trained with additional knowledge like segmentation or keypoints could be used to robustly find

an initial approximation on downsampled images. Subsequently, instance optimization [130] or a conventional registration can further align the higher resolution images to obtain a more accurate registration result.

Furthermore, in the near future, a self-configuring registration framework similar to the nnU-Net framework will be proposed. The first efforts have already been presented in [130] and [113]. With such a framework, developments in deep-learning-based registration will continue to evolve into a data-driven discipline. New methodological developments will have a smaller impact on performance enhancements compared to compiling and curating suitable data sets and an accurate data preparation.

## 8.2 Efficient Tumor Follow-Up Analysis

The second part of the thesis presented steps towards efficient tumor follow-up analysis. In this section, we discuss the current state of development, how the methodologies could be further improved, and most importantly what is still missing to use the proposed pipeline in clinical routine in the future.

### How good is our solution?

In the work presented in chapter 6, we evaluated the quality of baseline segmentation, registration accuracy, and quality of follow-up segmentation. The evaluation in this chapter refers only to soft-tissue lesions. We showed that the registration accurately propagates the center of gravity of the lesions from the baseline to the follow scan with a mean Euclidean distance of 7.66 mm which is comparable to the results of several methods presented in [151] on the DeepLesion dataset [177]. More importantly, it is in almost all cases close enough to find the corresponding lesion in the follow-up scan. In the baseline and follow-up scan our segmentation approach achieves an average Dice Score compared to the manual annotation of approximately 0.80.

While these technical measures give a first impression of the performance, they do not yet tell us whether we can add value to the assessment of cancer patients with the proposed pipeline. As a first step to quantify this, the goal of the reader study presented in chapter 7 was to compare the workflow of reading follow-up examinations with and without AI assistance to evaluate the impact of the proposed AI-assisted workflow. In this reader study, we focused on the segmentation not the diameter which is required for current guidelines. However, the diameter can be directly computed out of the segmentation. The three hypotheses of the study were: 1.) Assessment time for follow-up lesion segmentation is reduced using an AI-assisted workflow 2.) The inter-reader variability of the resulting segmentation is reduced with AI assistance. 3.) The quality of the AI-assisted segmentation is non-inferior to a fully manual segmentation. All three hypotheses could be verified in this study. The mean interaction time for lesion segmentation was significantly reduced from 3.5 min to 1.5 min using

AI-assisted segmentation compared to fully manual segmentation while maintaining the same segmentation quality. Furthermore, a reduction of inter-reader variability was achieved. This is an important result in that the therapy response classification of lesion according to RECIST can substantially vary from one radiologist to another [178]. In our study, in more than 50 % of the lesions, the readers have accepted the segmentation without any further corrections and if the correct lesion was annotated only in 11 % the correction changed the Dice Score more than 0.2.

So far, in the paper in chapter 6 and in the reader study in chapter 7, we have only performed an evaluation on the segmentation masks. However, the current guideline of metastatic tumor evaluation on CT scans RECIST [144] is based on the diameter of lesions. Therefore, the next logical step is to further evaluate the accuracy of the calculated diameters.

### Why did we still use conventional image registration?

For this application, the registration has to align the global structures but at the same time be locally accurate enough so that the lesion propagation is precise enough. Therefore, we adopted in chapter 6 a three-step approach to automatically register the baseline to the follow-up image, which consists of the following steps: (1) a translational alignment; (2) a rigid registration; and (3) a deformable registration. Hereby, the registration pipeline starts with robust methods with fewer degrees of freedom and moves on to more precise, but less robust methods, which require better starting points due to their higher degrees of freedom. Most of the recently presented deep-learning-based registration approaches – including the work presented in chapter 2 to 4 – focus on the deformable registration. Nevertheless, there are a few works that presented deep-learning-based approaches for translational and affine alignment. Those approaches could be combined into a pipeline to fulfill the needs of the application. Another hurdle that would need to be addressed by the deep-learning-based approaches is the large image size. For metastatic melanoma, typically full-body or thorax-abdomen CT scans are acquired, which can easily exceed image sizes of 512 × 512 × 1000, which can be a challenge in terms of memory usage. Simply downsampling the images to an image size that fits on current GPUs has the disadvantage that downsampled images lose too much information to be locally accurate enough for the lesion propagation. Therefore, more complex approaches like a multilevel approach that combines the results of the low-resolution full-image registration with the deformation fields of a batch-based method using the high-resolution images might be a solution.

In this initial work, the goal was to investigate whether the combination of registration and subsequent deep-learning-based segmentation of the lesion in the propagated region of interest is a suitable solution for this application. Furthermore, registration is not a time-critical component in the pipeline, as it is calculated in advance. For this

reason, we chose a slightly slower but more robust conventional registration to start with. However, a deep-learning-based registration approach could be integrated into this approach in the future.

### How can we further improve the results?

In the approach presented in chapters 6 and 7, we use the image registration method only to propagate the region of interest to find the approximate location of the lesion in the follow-up scan. However, through the baseline segmentation, we also know the approximate appearance of the lesion in the follow-up scan [153]. Therefore, it might be helpful to integrate this information into the segmentation approach of the follow-up lesion.

There are different conceivable approaches to achieve this. The transformed baseline image and the corresponding lesion mask could be used as an additional input for the segmentation network of the follow-up images. This provides a first approximation of the follow-up segmentation, so that the network only has to correct it. The assumption here is that this is an easier task than completely segmenting it. Furthermore, this additional information could make it easier to decide which lesion is the correct lesion in the case of multiple lesions.

The same idea is pursued by a joint-segmentation-registration approach as in [126, 156]. To compensate for original registration errors, the input images are additionally re-registered locally in a separate decoder path. By learning segmentation and registration together, the two tasks can benefit from each other, and additionally, several new loss terms can be integrated into the training procedure. For both approaches, it would mean that two models have to be trained; one for segmenting the baseline image and one for the follow-up images. This is in general not a problem as long as enough training data is available.

As for all data-driven algorithms, it is also important for our approach to have an appropriate amount of training data available, which is sufficiently diverse. So far, we trained and evaluated on data from patients with metastatic melanoma (Stage IV, AJCC) treated at the Center for Dermato-Oncology at the University Hospital Tübingen, Germany. The number of data seems to be sufficient, but the diversity of the data is not. In particular, for an appropriate evaluation, it needs data from several locations. But also the training and thus the resulting segmentation network would benefit from a higher diversity. A consequent next step is to evaluate our approach on the DeepLesion [177] dataset to compare it to several other methods like [151, 167].

### When will it be ready for clinical routine?

In principle, our developed software component could be integrated into in its current state by an appropriate company in their reporting software after the necessary ap-

proval has been obtained. However, we still have to ask ourselves whether this system will be used by the radiologist. A new system is usually only introduced if it leads to a direct improvement in patient outcomes or for reasons of cost-effectiveness. Although the ultimate goal is, of course, to improve patient outcomes, this is a long process and not easy to show. Due to the rapidly increasing number of imaging procedures and therefore resulting increasing number of readings that have to be carried out per radiologist and at the same time the lack of available healthcare personnel, maintaining the current reading quality by reducing the workload become more and more important. Therefore, we aim to reduce the reading time of the radiologist per case by providing software-assistance.

The radiologist's main task is to define and measure the target lesions to deduce the progression state. One of the most important aspects from a user's point of view is that all necessary information to perform this task is available in one system and is easy accessible. Often it takes a long time to gather all the necessary information like "What preliminary examinations are there?" or "Which structures have already been measured?". This step is simplified by storing all images and annotated metastases in our software. However, the reporting question or clinical information is not yet integrated. Moreover, the generation of the report from the measured data has also not yet been implemented, although this would also further facilitate the work. These are points of connection to the other systems, which would depend very much on the exact type of use of our software.

There are also other points within the proposed workflow that could be improved to further enhance cost-effectiveness. In the current version, the radiologist measures the lesion in the baseline scan by clicking into it. In the future, this step could also be taken over by the software by automatically detecting them [179–182]. Suggestions for target lesions are then provided and the radiologist only has to select which ones to take over. Taken even further, the fully automated selection, annotation, and measurement of baseline lesions could result in all metastases being considered, rather than a small selection as is currently the case. This would further increase the reproducibility between readings from different radiologists [183]. However, such a feature is not indispensable to use the software and requires even more development effort than automatically tracking lesions into follow-up which is already implemented. With automatic follow-up tracking of lesions, the radiologist only needs to be able to efficiently verify the measurements. To do so, all selected target lesions could be for example displayed in a separated window as 2D slice images with the corresponding diameter. In general, the clear presentation of the automatically determined information is an important point that could be further improved.

Another variant to reduce the reading time is that the radiologist trusts the software to such an extent that in cases where the CNN is sure, he or she checks more quickly or only examines random samples. This would require that in addition to the segmentation

result or the diameter, a confidence score is also predicted. We would assume that it is a matter of habituation to what extent one trusts the software. In the beginning, radiologists will be more skeptical about the software and correct the diameters more often. However, as soon as they notice that these changes lead only to little change in the result or within the inter-reader variability, then acceptance will probably increase. Especially in patients where no therapy change is induced by the measurement, the control could be minimized. The capacities freed up could be used for intensive monitoring of patients who will receive a change in therapy.

The detection of new lesions in the follow-up scan is also not yet possible. However, newly formed lesions have a direct impact on the progression status, which is why they are important to detect. Detecting them fully automatically is a difficult task, that ideally can be solved in the future. As an intermediate step, manual visual detection can be facilitated by a change map [184] – an overlay that highlights where changes have occurred between the baseline and the follow-up image.

In addition to the cancer-related reporting, the radiologist must also look at the scan for other findings. Again, a change map can be helpful to highlight relevant changes that need to be reported. To create a meaningful change map, an accurate registration is required to not highlight registration errors as change. Furthermore, it is necessary to filter for relevant change. There might be real change between the two time points like a different filling state of the digestive tract, however, those changes are not relevant for the report.

For all of the above, it is necessary that they can be calculated either very quickly on the fly or in advance so that the radiologist does not have to wait for the calculation.

In addition to the described extensions, extensive tests must still be carried out. So far, in the papers in chapter 6 and the reader study in chapter 7, we have only evaluated the segmentation masks. However, the current guideline of metastatic tumor evaluation on CT scans RECIST [144] is based on the diameter of lesions. Therefore, the first next step is to further evaluate the accuracy of the calculated diameters. Furthermore, an evaluation regarding the fairness of the algorithms is missing so far; does the algorithm treat patients equally regardless of their sex, ethnicity, etc. Since in the current version of the presented workflow, the radiologist still checks the lesions visually, this is not yet as important as with an algorithm that directly provides treatment predictions. Nevertheless, it is important to be aware of these issues and to investigate them through studies.

# Summary

## Deep-Learning-Based Image Registration

The goal of medical image registration is to align anatomical structures of two or more images by establishing spatial correspondences. This is an important step for many tasks in medical image analysis as it links previously unrelated data and enables joint processing of those data. Various approaches and tailored solutions have been proposed to a wide range of problems and applications. Typically, image registration is phrased as an optimization problem with respect to a spatial mapping that minimizes a suitable cost function and common approaches estimate solutions by applying iterative optimization schemes. Unfortunately, solving such an optimization problem is computationally demanding and consequently slow.

Since the availability of image data and computational power has rapidly grown, learning-based image registration methods have emerged as an alternative to conventional approaches. These methods replace the costly iterative optimization of conventional registration methods for each pair of images with one optimization during training of a convolutional neural network. The first part of this thesis describes fast and accurate registration methods using deep learning.

In **CHAPTER 2** describes a 2.5D convolutional transformer architecture that enables to learn a memory-efficient weakly-supervised deep-learning model for multi-modal image registration. The proposed architecture combines three 2D networks to a *2.5D registration network* which are The three networks are independently trained on axial, coronal, and sagittal slices of the images. During the inference, these networks are applied independently yielding three layered 3D deformation fields with one zero component. The final deformation field is created by averaging the respective non-zero components of the deformation field.

To address the multimodality of the task, the standard UNet architecture was adapted such that it has two separate processing streams for the moving and fixed image. The first layers of these streams use individual convolutional weights in order to learn modality-specific features. The later layers share the weight like in mono-modal image registration. We showed that our method succeeds at learning large deformations across multi-modal images.

**CHAPTER 3** presents a 3D deep-learning-based multilevel registration that is able

to compensate and handle large deformations by computing deformation fields on different scales and functionally composing them. The registration starts on the coarsest level using the downsampled network inputs to compute the deformation field on this level. On all finer levels, the deformation fields from all preceding coarse levels are incorporated as an initial guess. To this end, the deformation fields are functionally composed and used to warp the moving image at the current level.

We validated our framework on the challenging task of large motion inhale-to-exhale registration using large image data of the multi-center COPDGene study. We have shown that our proposed method archives better results than the comparable single-level variant. In particular concerning the alignment of inner lung structures and the presence of foldings. Additionally, we demonstrated that using the network parameter of the previous level as initialization yields better registration results.

**CHAPTER 4** identifies important strategies of conventional registration methods for lung registration and successfully developed the deep-learning counterpart. It builds on the method present in the previous chapter and extends it by adding multiple anatomical constraints to incorporate anatomical priors into the registration framework to obtain more realistic results. The lung lobe masks are integrated to consider the global context. Moreover, the keypoint correspondences are used to increase the alignment of smaller structures like airways and vessels. Furthermore, a constraining method was introduced to control volume change and therefore avoid foldings inside the deformation field. We showed that our registration framework equipped with these components achieves state-of-the-art registration accuracy on the COPDGene and DIRLab datasets with a very short execution time.

**CHAPTER 5** presents the results of the Learn2Reg challenge. The Learn2Reg challenge was the first to evaluate a wide range of methods for various inter-and intra-patient as well as mono- and multimodal medical image registration tasks. The main goal of this challenge was to provide a standardized benchmark on complementary tasks with clinical impact and a platform for comparison of conventional and learning-based medical image registration methods. We established a lower entry barrier for training and validation of 3D registration, which helped us compile results of over 65 individual method submissions from more than 20 unique teams.

## Tumor Follow-Up Analysis

Measurement of metastatic tumors on longitudinal computer tomography (CT) scans is essential to evaluate the efficacy of cancer treatment. Manual measurement of the tumors for the RECIST criteria is often time-consuming and error-prone. However, the diameter-based RECIST criteria also undergo continuous changes. AI-assisted approaches might significantly speed up response evaluation and help to handle the

ever-growing mass of image-based staging and follow-up evaluations.

**CHAPTER 6** presents a pipeline that automates the segmentation and measurement of matching lesions, given a point annotation in the baseline lesion. The point annotation is used to extract a region of interest in which the CNN is carried out to segment the lesion. Then, the baseline image is registered to the follow-up image to propagate the region of interest in the follow-up scan. Subsequently, the CNN is applied to the propagated region of interest in the follow-up image. In a final step, the corresponding lesion is selected. We have trained and evaluated our method on soft-tissue lesions from patients with metastatic melanoma, which are particularly challenging due to their diverse appearance and location. We showed that our method archives promising results and therefore laid the foundation for an efficient quantitative follow-up assessment in clinical routine.

The reader study in **CHAPTER 7** evaluates the performance, inter-reader variability, and efficiency of an AI-assisted workflow for segmentation of lymph node and soft tissue metastases in follow-up CTs by comparing it to a fully manual assessment. This workflow builds on the pipeline presented in the previous chapter. Our findings support our research hypothesis of an assisted workflow which is superior with respect to processing time and non-inferior with respect to accuracy compared to the manual workflow.

# Nederlandse samenvatting

## 8.3 Deep-Learning-Based Image Registration

Het doel van medische beeldregistratie is het spatieel uitlijnen van anatomische structuren van twee of meer beelden door het vaststellen van ruimtelijke correspondenties. Dit is een belangrijke stap voor vele taken in de medische beeldanalyse omdat het voorheen niet gerelateerde gegevens verbindt en een gezamenlijke verwerking van die gegevens mogelijk maakt. Verschillende methoden voor registratie zijn voorgesteld voor een breed scala van problemen en toepassingen. Meestal wordt beeldregistratie geformuleerd als een optimalisatieprobleem waarbij een geschikte kostenfunctie wordt geminimaliseerd. Veelgebruikte benaderingen schatten oplossingen door iteratieve optimalisatieschema's toe te passen. Jammer genoeg is het oplossen van een dergelijk optimalisatieprobleem rekenkundig veeleisend en bijgevolg traag.

Het blijkt ook mogelijk om het proces van registratie te leren van voorbeelden door middel van deep learning. Door de toegenomen computerkracht is dit naar voren gekomen als een alternatief voor conventionele benaderingen. Deep learning registratie vervangt de kostbare iteratieve optimalisatie van conventionele registratiemethoden door één optimalisatie die berekend wordt met een convolutienetwerk (CNN). Het eerste deel van dit proefschrift beschrijft snelle en nauwkeurige registratiemethoden met behulp van deep learning.

In **HOOFDSTUK 2** wordt een 2.5D convolutionele transformator architectuur beschreven die het mogelijk maakt om een geheugen-efficiënt zwak-supervised deep-learning model te leren voor multi-modale beeldregistratie. Een 2D deep-learning gebaseerde beeldregistratie aanpak is in de meeste gevallen niet voldoende voor volledige 3D registratie omdat de vervormingen meestal drie-dimensionaal zijn. Het trainen van een 3D netwerk heeft echter veel meer geheugen nodig. In dit hoofdstuk hebben we het probleem van driedimensionale vervormingen aangepakt zonder de netwerkarchitectuur uit te breiden naar 3D. De voorgestelde architectuur combineert drie 2D netwerken tot een 2.5D registratie netwerk. De drie netwerken worden onafhankelijk getraind op axiale, coronale en sagittale doorsneden van de beelden. In de testfaseworden deze netwerken onafhankelijk toegepast, wat drie gelaagde 3D deformatievelden oplevert met één nulcomponent. Het uiteindelijke deformatieveld wordt gecreëerd door het gemiddelde te nemen van de componenten van het deformatieveld die ongelijk aan nul zijn.

Om met multimodale beelden om te kunnen gaan hebben we de standaard UNet-architectuur aangepast. We introduceren twee afzonderlijke verwerkingsstromen voor het bewegende en het vaste beeld. De eerste lagen van deze stromen gebruiken individuele convolutionele gewichten om modaliteitspecifieke kenmerken te leren. De latere lagen delen de parameters. We hebben aangetoond dat onze methode erin slaagt om grote vervormingen te leren over multimodale beelden.

In dit artikel presenteren we een 3D deep-learning gebaseerde multilevel registratie die in staat is om grote vervormingen te compenseren en te verwerken door vervormingsvelden op verschillende schalen te berekenen en deze functioneel samen te stellen. De registratie begint op het grofste niveau en gebruikt de gedownsamplede netwerkinputs om het vervormingsveld op dit niveau te berekenen. Op alle fijnere niveaus worden de vervormingsvelden van alle voorgaande grovere niveaus opgenomen als een initiële schatting. We hebben ons raamwerk gevalideerd op CT scans van de longen die gemaakt zijn in volledige inspiratie en volledige expiratie. Door de grote vervormingen is dit een uitdagende taak. We gebruiken beelden van de multi-center COPDGene studie. We hebben aangetoond dat onze voorgestelde methode betere resultaten oplevertdan de vergelijkbare single-level variant.

**HOOFDSTUK 4** identificeert belangrijke strategieën van conventionele registratiemethoden voor longregistratie en past deze toe in een deep-learning variant. De methode bouwt voort op het vorige hoofdstuk en breidt deze aanpak uit door meerdere anatomische constraints toe te voegen om anatomische priors op te nemen in het registratie raamwerk. Dit levert meer realistische resultaten op. De maskers van de longkwabben worden geïntegreerd in de kostfunctie om rekening te houden met de globale context. Bovendien worden de corresponderende ankerpunten gebruikt om de uitlijning van kleinere structuren zoals luchtwegen en vaten te verbeteren. Tot slot werd een extra term toegevoegd die grote volumeveranderingen tegengaat en daardoor komen onrealistische vervormingen minder voor.. We hebben laten zien dat onze methodehoge nauwkeurigheid bereikt op de COPDGene en DIRLab datasets en bovendien zeer snel werkt.

**HOOFDSTUK 5** presenteert de resultaten van de Learn2Reg challenge. De Learn2Reg challenge was de eerste die een breed scala aan methoden evalueerde voor verschillende inter- en intra-patiënt en mono- en multimodale medische beeldregistratietaken. Het belangrijkste doel van deze challenge was om een gestandaardiseerde benchmark te bieden voor verschillende klinisch belangrijke taken. Met Learn2Reg kunnen conventionele en deep learning beeldregistratiemethoden goed vergeleken worden. De challenge verlaagt de drempel voor onderzoekers om hun methoden te vergelijken, wat ons heeft geholpen om de resultaten te verzamelen van meer dan 65 inzendingen van meer dan 20 teams.

## 8.4 Tumor follow-up analyse

Metingen van uitgezaaide tumoren op opeenvolgende computertomografie (CT) scans is belangrijk om te kunnen vaststellen of een behandeling van kanker doeltreffend is. Radiologen voeren nu manuele metingen uit van de tumoren volgens de RECIST criteria. Dit is tijdrovend en foutgevoelig. AI-ondersteunde benaderingen zouden de evaluatie aanzienlijk kunnen versnellen en kunnen helpen om de steeds groeiende hoeveelheid scans te verwerken.

**HOOFDSTUK 6** presenteert een pijplijn die de segmentatie en meting van overeenkomende laesies automatiseert, waarbij alleen een punt hoeft te worden aangeklikt in de laesie in de baseline scan. De punt annotatie wordt gebruikt om een regio te extraheren waarin het CNN wordt uitgevoerd om de laesie te segmenteren. Vervolgens wordt het basisbeeld geregistreerd op het vervolgbeeld om het interessegebied naar de vervolgscan te propageren. Daar wordt het CNN toegepast op het gepropageerde gebied . Wij hebben onze methode getraind en geëvalueerd op weke delen laesies van patiënten met metastatisch melanoom. We toonden aan dat onze methode veelbelovende resultaten behaalt en dit legt de basis gelegd voor een efficiënte kwantitatieve follow-up beoordeling in de kliniek.

De reader studie in **HOOFDSTUK 7** evalueert de prestaties, inter-reader variabiliteit, en efficiëntie van een AI-ondersteunde workflow voor segmentatie van lymfeklier en weke delen metastasen in follow-up CTs door deze te vergelijken met een volledig handmatige beoordeling. Deze workflow bouwt voort op de pijplijn die in het vorige hoofdstuk is gepresenteerd. Onze bevindingen ondersteunen onze onderzoekshypothese van een geassisteerde workflow die superieur is met betrekking tot verwerkingstijd en niet-inferieur met betrekking tot nauwkeurigheid in vergelijking met de handmatige workflow. Een onafhankelijke evaluatie met extra lezers is nodig om de generaliseerbaarheid van onze resultaten aan te tonen. De analyse is conservatief in de zin dat verdere training met de geassisteerde workflow na verloop van tijd tot een extra verbetering van een of beide uitkomsten zou kunnen leiden.

# Research Data Managment

Studies described in this thesis use publicly available datasets (1-5) that can be accessed online after registration or application and proprietary datasets obtained via clinical cooperation. The primary data used in **Chapter 2 and 5** and the secondary data used in **Chapter 2-7** is stored on a regularly backed-up Fraunhofer MEVIS (FME) server accessible by all FME staff members. The primary data of **Chapter 3 and 4** is stored on a stored on a regularly backed-up Diagnostic Image Analysis Group (DIAG) server accessible by all DIAG staff members. The primary data of **Chapter 6 and 7** is stored on a stored on a regularly backed-up FME server accessible by all FME staff members who worked on the the *DFG Radiomics Melanom* project. The source code used for the experiments presented in **Chapter 2-7** is stored in a private GitLab or SVN repository accessible by FME staff members.

(1) MM-WHS - Multi-Modality Whole Heart Segmentation Dataset (`http://www.sdspeople.fudan.edu.cn/zhuangxiahai/0/mmwhs/`).

(2) DIRLab Dataset (`https://med.emory.edu/departments/radiation-oncology/research-laboratories/deformable-image-registration/index.html`.

(3) EMPIRE10 Challenge (`https://empire10.grand-challenge.org`).

(4) COPDGene Dataset (`http://www.copdgene.org/`).

(5) Learn2Reg Challenge (`https://learn2reg.grand-challenge.org/`).

# Bibliography


1. J. Modersitzki. “Numerical methods for image registration,” Oxford University Press on Demand, 2004 (cited on pp. 8, 48).
2. H. Zaidi, M.-L. Montandon, and A. Alavi. “The clinical role of fusion imaging using PET, CT, and MR imaging,” *PET clinics*, vol. 3 (2008), pp. 275–291 (cited on p. 8).
3. S. J. Gong, G. J. O’keefe, and A. M. Scott. “Comparison and evaluation of PET/CT image registration,” *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*, IEEE. 2006, pp. 1599–1603 (cited on p. 8).
4. M. Chen. “Deformable image registration in the analysis of multiple sclerosis,” PhD thesis. Johns Hopkins University, 2015 (cited on p. 8).
5. J. E. Iglesias and M. R. Sabuncu. “Multi-atlas segmentation of biomedical images: a survey,” *Medical image analysis*, vol. 24 (2015), pp. 205–219 (cited on p. 8).
6. F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein. “nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation,” *Nature Methods* (2020), pp. 1–9 (cited on pp. 8, 105–107, 115, 123, 126, 137, 143, 147).
7. Z. Ding, X. Han, and M. Niethammer. “Votenet: a deep learning label fusion method for multi-atlas segmentation,” *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer. 2019, pp. 202–210 (cited on p. 8).
8. J. A. Maintz and M. A. Viergever. “A survey of medical image registration,” *Medical image analysis*, vol. 2 (1998), pp. 1–36 (cited on pp. 8, 47, 77).
9. A. Sotiras, C. Davatzikos, and N. Paragios. “Deformable medical image registration: a survey,” *Medical Imaging, IEEE Transactions on*, vol. 32 (2013), pp. 1153–1190 (cited on pp. 8, 47, 77, 78).
10. C. Broit. “Optimal registration of deformed images,” University of Pennsylvania, 1981 (cited on p. 8).
11. R. Bajcsy and S. Kovačič. “Multiresolution elastic matching,” *Computer vision, graphics, and image processing*, vol. 46 (1989), pp. 1–21 (cited on pp. 8, 35, 37, 48, 55).
12. Y. Amit. “A nonlinear variational problem for image matching,” *SIAM Journal on Scientific Computing*, vol. 15 (1994), pp. 207–224 (cited on p. 8).
13. D. Rueckert, L. I. Sonoda, C. Hayes, D. L. Hill, M. O. Leach, and D. J. Hawkes. “Non-rigid registration using free-form deformations: application to breast MR images,” *IEEE transactions on medical imaging*, vol. 18 (1999), pp. 712–721 (cited on p. 8).

14. J. A. Schnabel, D. Rueckert, M. Quist, J. M. Blackall, A. D. Castellano-Smith, T. Hartkens, G. P. Penney, W. A. Hall, H. Liu, C. L. Truwit, et al. "A generic framework for non-rigid registration based on non-uniform multi-level free-form deformations," *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2001*, Springer. 2001, pp. 573–581 (cited on pp. 8, 35, 37, 48, 55).

15. T. Vercauteren, X. Pennec, A. Perchant, and N. Ayache. "Diffeomorphic demons: efficient non-parametric image registration," *NeuroImage*, vol. 45 (2009), pp. S61–S72 (cited on p. 8).

16. S. Klein, M. Staring, K. Murphy, M. A. Viergever, and J. P. Pluim. "Elastix: a toolbox for intensity-based medical image registration," *IEEE transactions on medical imaging*, vol. 29 (2009), pp. 196–205 (cited on pp. 8, 78, 142).

17. J. R. Rühaak. "Matrix-free techniques for efficient image registration and their application to pulmonary image analysis," PhD thesis. Jacobs University Bremen, 2017 (cited on p. 8).

18. T. Rohlfing and C. R. Maurer. "Nonrigid image registration in shared-memory multiprocessor environments with application to brains, breasts, and bees," *IEEE transactions on information technology in biomedicine*, vol. 7 (2003), pp. 16–25 (cited on pp. 8, 146).

19. L. König, J. Rühaak, A. Derksen, and J. Lellmann. "A matrix-free approach to parallel and memory-efficient deformable image registration," *SIAM Journal on Scientific Computing*, vol. 40 (2018), B858–B888 (cited on pp. 8, 48, 108, 146).

20. M. Modat, G. R. Ridgway, Z. A. Taylor, M. Lehmann, J. Barnes, D. J. Hawkes, N. C. Fox, and S. Ourselin. "Fast free-form deformation using graphics processing units," *Computer methods and programs in biomedicine*, vol. 98 (2010), pp. 278–284 (cited on pp. 8, 78, 87, 98, 142, 146).

21. D. Budelmann, L. König, N. Papenberg, and J. Lellmann. "Fully-deformable 3d image registration in two seconds," *Bildverarbeitung für die Medizin 2019*, Springer, 2019, pp. 302–307 (cited on pp. 8, 146).

22. G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez. "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42 (2017), pp. 60–88 (cited on pp. 9, 16).

23. J. Hadamard. "Sur les problèmes aux dérivées partielles et leur signification physique," *Princeton university bulletin* (1902), pp. 49–52 (cited on p. 9).

24. K. A. Eppenhof, M. W. Lafarge, P. Moeskops, M. Veta, and J. P. Pluim. "Deformable image registration using convolutional neural networks," *Medical Imaging 2018: Image Processing*, vol. 10574 International Society for Optics and Photonics. (2018), p. 105740S (cited on pp. 9, 48).

25. H. Sokooti, B. de Vos, F. Berendsen, B. P. Lelieveldt, I. Išgum, and M. Staring. "Nonrigid image registration using multi-scale 3D convolutional neural networks," *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer. 2017, pp. 232–239 (cited on pp. 9, 48, 142).

26. T. Sentker, F. Madesta, and R. Werner. "GDL-FIRE 4D: deep learning-based fast 4D CT image registration," *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2018*, Springer. 2018, pp. 765–773 (cited on pp. 9, 47–49, 68, 71).

27. X. Yang, R. Kwitt, M. Styner, and M. Niethammer. "Fast predictive multimodal image registration," *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, 2017, pp. 858–862 (cited on pp. 9, 48).

28. G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, and A. V. Dalca. "An unsupervised learning model for deformable medical image registration," *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 9252–9260 (cited on pp. 9, 47, 48).

29. B. D. de Vos, F. F. Berendsen, M. A. Viergever, H. Sokooti, M. Staring, and I. Išgum. "A deep learning framework for unsupervised affine and deformable image registration," *Medical image analysis*, vol. 52 (2019), pp. 128–143 (cited on pp. 9, 47–49, 55, 68, 69, 71, 144).

30. E. Ferrante, O. Oktay, B. Glocker, and D. H. Milone. "On the adaptability of unsupervised CNN-based deformable image registration to unseen image domains," *International Workshop on Machine Learning in Medical Imaging*, Springer. 2018, pp. 294–302 (cited on pp. 9, 48).

31. J. Krebs, H. Delingette, B. Mailhé, N. Ayache, and T. Mansi. "Learning a probabilistic model for diffeomorphic registration," *IEEE transactions on medical imaging*, vol. 38 (2019), pp. 2165–2176 (cited on pp. 9, 48, 72).

32. Y. Hu, M. Modat, E. Gibson, W. Li, N. Ghavami, E. Bonmati, G. Wang, S. Bandula, C. M. Moore, M. Emberton, et al. "Weakly-supervised convolutional neural networks for multimodal image registration," *Medical image analysis*, vol. 49 (2018), pp. 1–13 (cited on pp. 9, 15, 16, 30, 48).

33. G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, and A. V. Dalca. "VoxelMorph: a learning framework for deformable medical image registration," *IEEE TMI* (2019) (cited on pp. 9, 15, 16, 25, 35, 47, 48, 54, 61, 68, 69, 71, 88, 98, 142–144).

34. A. Hering, B. van Ginneken, and S. Heldmann. "mlVIRNET: multilevel variational image registration network," *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2019*, Springer. 2019, pp. 257–265 (cited on pp. 9, 47–49, 51, 55, 68, 69).

35. A. Hering, S. Kuckertz, S. Heldmann, and M. P. Heinrich. "Memory-efficient 2.5 D convolutional transformer networks for multi-modal deformable registration with weak label supervision applied to whole-heart CT and MRI scans," *International journal of computer assisted radiology and surgery*, vol. 14 (2019), pp. 1901–1912 (cited on pp. 9, 47, 48, 54, 71).

36. M. Jaderberg, K. Simonyan, A. Zisserman, et al. "Spatial transformer networks," *Advances in neural information processing systems*, 2015, pp. 2017–2025 (cited on pp. 10, 16, 48).

37. H. Wang, M. Naghavi, C. Allen, R. M. Barber, Z. A. Bhutta, A. Carter, D. C. Casey, F. J. Charlson, A. Z. Chen, M. M. Coates, et al. "Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980–2015: a systematic analysis for the global burden of disease study 2015," *The lancet*, vol. 388 (2016), pp. 1459–1544 (cited on p. 10).

38. C. Fitzmaurice, C. Allen, R. M. Barber, L. Barregard, Z. A. Bhutta, H. Brenner, D. J. Dicker, O. Chimed-Orchir, R. Dandona, L. Dandona, et al. "Global, regional, and national cancer incidence, mortality, years of life lost, years lived with disability, and disability-adjusted life-years for 32 cancer groups, 1990 to 2015: a systematic analysis for the global burden of disease study," *JAMA oncology*, vol. 3 (2017), pp. 524–548 (cited on p. 10).

39. L. König. "Matrix-free approaches for deformable image registration with large-scale and real-time applications in medical imaging," PhD thesis. Universität zu Lübeck, 2018 (cited on p. 10).

40. A. Sotiras, C. Davatzikos, and N. Paragios. "Deformable medical image registration: a survey," *IEEE transactions on medical imaging*, vol. 32 (2013), pp. 1153–1190 (cited on p. 15).

41. B. D. de Vos, F. F. Berendsen, M. A. Viergever, M. Staring, and I. Išgum. "End-to-end unsupervised deformable image registration with a convolutional neural network," *Deep learning in medical image analysis and multimodal learning for clinical decision support*, Springer, 2017, pp. 204–212 (cited on pp. 15–17, 29).

42. M.-M. Rohé, M. Datar, T. Heimann, M. Sermesant, and X. Pennec. "SVF-Net: learning deformable image registration using shape matching," *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer. 2017, pp. 266–274 (cited on pp. 15–17, 35, 48, 142).

43. J. Rühaak, S. Heldmann, T. Kipshagen, and B. Fischer. "Highly accurate fast lung CT registration," *Medical Imaging 2013: Image Processing*, vol. 8669 International Society for Optics and Photonics. (2013), 86690Y (cited on p. 15).

44. A. Hering, S. Kuckertz, S. Heldmann, and M. P. Heinrich. "Enhancing label-driven deep deformable image registration with local distance metrics for state-of-the-art cardiac motion tracking," *Bildverarbeitung für die Medizin 2019*, Springer, 2019, pp. 309–314 (cited on pp. 15, 17, 19, 23, 35, 48).

45. E. Haber and J. Modersitzki. "Intensity gradient based registration and fusion of multi-modal images," *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer. 2006, pp. 726–733 (cited on pp. 15, 18, 108).

46. 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 (cited on p. 16).

47. B. D. de Vos, F. F. Berendsen, M. A. Viergever, H. Sokooti, M. Staring, and I. Išgum. "A deep learning framework for unsupervised affine and deformable image registration," *Medical image analysis*, vol. 52 (2019), pp. 128–143 (cited on pp. 16, 29, 35, 40, 42).

48. Y. Xia, L. Xie, F. Liu, Z. Zhu, E. K. Fishman, and A. L. Yuille. "Bridging the gap between 2d and 3d organ segmentation with volumetric fusion net," *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer. 2018, pp. 445–453 (cited on pp. 16, 17).

49. H. R. Roth, L. Lu, A. Farag, H.-C. Shin, J. Liu, E. B. Turkbey, and R. M. Summers. "Deeporgan: multi-level deep convolutional networks for automated pancreas segmentation," *International conference on medical image computing and computer-assisted intervention*, Springer. 2015, pp. 556–564 (cited on p. 16).

50. A. Prasoon, K. Petersen, C. Igel, F. Lauze, E. Dam, and M. Nielsen. "Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network," *International conference on medical image computing and computer-assisted intervention*, Springer. 2013, pp. 246–253 (cited on p. 16).

51. J. Rühaak, A. Derksen, S. Heldmann, M. Hallmann, and H. Meine. "Accurate CT-MR image registration for deep brain stimulation: a multi-observer evaluation study," *Medical Imaging 2015: Image Processing*, vol. 9413 International Society for Optics and Photonics. (2015), p. 941337 (cited on p. 18).

52. B. Fischer and J. Modersitzki. "Curvature based image registration," *JMIV*, vol. 18 (2003) (cited on p. 18).

53. O. Ronneberger, P. Fischer, and T. Brox. "U-Net: convolutional networks for biomedical image segmentation," *International Conference on Medical image computing and computer-assisted intervention*, Springer. 2015, pp. 234–241 (cited on pp. 19, 55, 106, 107, 123, 126).

54. X. Zhuang and J. Shen. "Multi-scale patch and multi-modality atlases for whole heart segmentation of MRI," *Medical image analysis*, vol. 31 (2016), pp. 77–87 (cited on pp. 21, 143).

55. A. Hering and S. Heldmann. "Unsupervised learning for large motion thoracic CT follow-up registration," *Medical Imaging 2019: Image Processing*, vol. 10949 International Society for Optics and Photonics. (2019), 109491B (cited on pp. 24, 35, 37, 47, 48).

56. M. P. Heinrich, M. Jenkinson, M. Brady, and J. A. Schnabel. "Globally optimal deformable registration on a minimum spanning tree using dense displacement sampling," *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer. 2012, pp. 115–122 (cited on pp. 25, 30, 68, 142, 143).

57. M. P. Heinrich, O. Maier, and H. Handels. "Multi-modal multi-atlas segmentation using discrete optimisation and self-similarities." *VISCERAL Challenge@ ISBI*, vol. 1390 (2015), p. 27 (cited on p. 25).

58. Z. Xu, C. P. Lee, M. P. Heinrich, M. Modat, D. Rueckert, S. Ourselin, R. G. Abramson, and B. A. Landman. "Evaluation of six registration methods for the human abdomen on clinically acquired CT," *IEEE Transactions on Biomedical Engineering*, vol. 63 (2016), pp. 1563–1572 (cited on pp. 25, 30).

59. M. P. Heinrich. "Intra-operative ultrasound to mri fusion with a public multimodal discrete registration tool," *Simulation, Image Processing, and Ultrasound Systems for Assisted Diagnosis and Navigation*, Springer, 2018, pp. 159–164 (cited on p. 25).

60. M. P. Heinrich, O. Oktay, and N. Bouteldja. "Obelisk-one kernel to solve nearly everything: unified 3D binary convolutions for image analysis," (2018) (cited on p. 29).

61. K. A. Eppenhof, M. W. Lafarge, M. Veta, and J. P. Pluim. "Progressively trained convolutional neural networks for deformable image registration," *IEEE transactions on medical imaging* (2019) (cited on pp. 35, 40, 42, 47, 49, 68, 69, 71, 72).

62. J. Modersitzki and E. Haber. "Cofir: coarse and fine image registration," in edited by L. T. Biegler, O. Ghattas, M. Heinkenschloss, D. Keyes, and B. van Bloemen Waanders. *Computational Science & Engineering: Real-Time PDE-Constrained Optimization* SIAM, 2007. Chap. 14, pp. 277–288 (cited on pp. 35, 37, 48, 55).

63. Y. Hu, M. Modat, E. Gibson, N. Ghavami, E. Bonmati, C. M. Moore, M. Emberton, J. A. Noble, D. C. Barratt, and T. Vercauteren. "Label-driven weakly-supervised learning for multimodal deformable image registration," *Proc. of ISBI 2018*, IEEE. 2018, pp. 1070–1074 (cited on pp. 35, 47–49).

64. R. Castillo, E. Castillo, D. Fuentes, M. Ahmad, A. M. Wood, M. S. Ludwig, and T. Guerrero. "A reference dataset for deformable image registration spatial accuracy evaluation using the copdgene study archive," *Physics in Medicine & Biology*, vol. 58 (2013), p. 2861 (cited on pp. 36, 40, 50, 57, 143).

65. J. Rühaak, T. Polzin, S. Heldmann, I. J. Simpson, H. Handels, J. Modersitzki, and M. P. Heinrich. "Estimation of large motion in lung CT by integrating regularized keypoint correspondences into dense deformable registration," *IEEE TMI*, vol. 36 (2017), pp. 1746–1757 (cited on pp. 36, 51–55, 58, 60–63, 65, 67, 68, 97, 146).

66. J. Modersitzki. "FAIR: flexible algorithms for image registration," SIAM, 2009 (cited on p. 36).

67. S. Kabus and C. Lorenz. "Fast elastic image registration," *Medical Image Analysis for the Clinic: A Grand Challenge* (2010), pp. 81–89 (cited on pp. 37, 48, 55).

68. X. Glorot and Y. Bengio. "Understanding the difficulty of training deep feedforward neural networks," *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 2010, pp. 249–256 (cited on pp. 38, 56).

69. E. A. Regan, J. E. Hokanson, J. R. Murphy, B. Make, D. A. Lynch, T. H. Beaty, D. Curran-Everett, E. K. Silverman, and J. D. Crapo. "Genetic epidemiology of COPD (COPDGene) study design," *COPD: Journal of Chronic Obstructive Pulmonary Disease*, vol. 7 (2011), pp. 32–43 (cited on pp. 38, 50, 57, 143).

70. M. P. Heinrich, M. Jenkinson, M. Brady, and J. A. Schnabel. "MRF-based deformable registration and ventilation estimation of lung CT," *IEEE transactions on medical imaging*, vol. 32 (2013), pp. 1239–1248 (cited on pp. 47, 68, 78).

71. T. Polzin, J. Rühaak, R. Werner, J. Strehlow, S. Heldmann, H. Handels, and J. Modersitzki. "Combining automatic landmark detection and variational methods for lung ct registration," *Fifth International Workshop on Pulmonary Image Analysis*, 2013, pp. 85–96 (cited on pp. 47, 54).

72. X. Yang, R. Kwitt, M. Styner, and M. Niethammer. "Quicksilver: fast predictive image registration–a deep learning approach," *NeuroImage*, vol. 158 (2017), pp. 378–396 (cited on p. 47).

73. B. D. de Vos, F. F. Berendsen, M. A. Viergever, M. Staring, and I. Išgum. "End-to-end unsupervised deformable image registration with a convolutional neural network," *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, Springer, 2017, pp. 204–212 (cited on pp. 47, 48).

74. M. P. Heinrich. "Closing the gap between deep and conventional image registration using probabilistic dense displacement networks," *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer. 2019, pp. 50–58 (cited on pp. 47, 87, 98).

75. K. A. Eppenhof and J. P. Pluim. "Pulmonary ct registration through supervised learning with convolutional neural networks," *IEEE transactions on medical imaging* (2018) (cited on p. 47).

76. S. Kuckertz, N. Papenberg, J. Honegger, T. Morgas, B. Haas, and S. Heldmann. "Deep learning based CT-CBCT image registration for adaptive radio therapy," *Medical Imaging 2020: Image Processing*, vol. 11313 International Society for Optics and Photonics. (2020), 113130Q (cited on p. 47).

77. J. Krebs, T. Mansi, H. Delingette, L. Zhang, F. C. Ghesu, S. Miao, A. K. 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*, Springer. 2017, pp. 344–352 (cited on p. 48).

78. J. Modersitzki. "FAIR: flexible algorithms for image registration," vol. 6 (SIAM, 2009) (cited on pp. 48, 142).

79. H. Li and Y. Fan. "Non-rigid image registration using self-supervised fully convolutional networks without training data," *arXiv preprint arXiv:1801.04012* (2018) (cited on p. 48).

80. Y. Fu, Y. Lei, T. Wang, K. Higgins, J. Bradley, W. Curran, T. Liu, and X. Yang. "Lungregnet: an unsupervised deformable image registration method for 4D-CT lung." *Medical physics*, vol. 47 (2020), p. 1763 (cited on pp. 49, 55, 65, 68, 69, 71).

81. R. Castillo, E. Castillo, R. Guerra, V. E. Johnson, T. McPhail, A. K. Garg, and T. Guerrero. "A framework for evaluation of deformable image registration spatial accuracy using large landmark point sets," *Physics in Medicine & Biology*, vol. 54 (2009), p. 1849 (cited on pp. 50, 57, 78, 89, 143, 144, 146).

82. K. Murphy, B. Van Ginneken, J. M. Reinhardt, S. Kabus, K. Ding, X. Deng, K. Cao, K. Du, G. E. Christensen, V. Garcia, et al. "Evaluation of registration methods on thoracic CT: the EMPIRE10 challenge," *IEEE transactions on medical imaging*, vol. 30 (2011), pp. 1901–1920 (cited on pp. 50, 58–60, 143, 144).

83. E. Haber and J. Modersitzki. "Intensity gradient based registration and fusion of multi-modal images," *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2006*, vol. 3216 (2006), pp. 591–598 (cited on p. 51).

84. B. Fischer and J. Modersitzki. "Curvature based image registration," *Journal of Mathematical Imaging and Vision*, vol. 18 (2003), pp. 81–85 (cited on pp. 51, 108).

85. H. Kervadec, J. Dolz, S. Wang, E. Granger, and I. B. Ayed. "Bounding boxes for weakly supervised segmentation: global constraints get close to full supervision," *arXiv preprint arXiv:2004.06816* (2020) (cited on p. 52).

86. S. Boyd, S. P. Boyd, and L. Vandenberghe. "Convex optimization," Cambridge university press, 2004 (cited on p. 53).

87. H. Kervadec, J. Dolz, J. Yuan, C. Desrosiers, E. Granger, and I. B. Ayed. "Constrained deep networks: lagrangian optimization via log-barrier extensions," *arXiv preprint arXiv:1904.04205* (2019) (cited on p. 53).

88. J. Ehrhardt, R. Werner, A. Schmidt-Richberg, and H. Handels. "Automatic landmark detection and non-linear landmark-and surface-based registration of lung CT images," *Medical Image Analysis for the Clinic-A Grand Challenge, MICCAI*, vol. 2010 (2010), pp. 165–174 (cited on p. 54).

89. B. Fischer and J. Modersitzki. "Combination of automatic non-rigid and landmark-based registration: the best of both worlds," *Medical Imaging 2003: Image Processing*, vol. 5032 International Society for Optics and Photonics. (2003), pp. 1037–1048 (cited on p. 54).

90. N. Papenberg, J. Olesch, T. Lange, P. M. Schlag, and B. Fischer. "Landmark constrained non-parametric image registration with isotropic tolerances," *Bildverarbeitung für die Medizin 2009*, Springer, 2009, pp. 122–126 (cited on p. 54).

91. Z. Jiang, F.-F. Yin, Y. Ge, and L. Ren. "A multi-scale framework with unsupervised joint training of convolutional neural networks for pulmonary deformable image registration," *Physics in Medicine & Biology*, vol. 65 (2020), p. 015011 (cited on pp. 55, 68, 71, 72).

92. T. C. Mok and A. C. Chung. "Large deformation diffeomorphic image registration with laplacian pyramid networks," *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer. 2020, pp. 211–221 (cited on pp. 55, 86, 144).

93. L. Hansen and M. P. Heinrich. "Tackling the problem of large deformations in deep learning based medical image registration using displacement embeddings," *arXiv preprint arXiv:2005.13338* (2020) (cited on pp. 68, 143).

94. A. Schmidt-Richberg, R. Werner, J. Ehrhardt, J.-C. Wolf, and H. Handels. "Landmark-driven parameter optimization for non-linear image registration," *Medical Imaging 2011: Image Processing*, vol. 7962 International Society for Optics and Photonics. (2011), 79620T (cited on p. 68).

95. F. F. Berendsen, A. N. Kotte, M. A. Viergever, and J. P. Pluim. "Registration of organs with sliding interfaces and changing topologies," *Medical Imaging 2014: Image Processing*, vol. 9034 International Society for Optics and Photonics. (2014), 90340E (cited on p. 68).

96. L. Hansen and M. P. Heinrich. "GraphRegNet: deep graph regularisation networks on sparse keypoints for dense registration of 3D lung CTs," *IEEE Transactions on Medical Imaging* (2021) (cited on pp. 68, 69, 85, 98, 144).

97. A. V. Dalca, G. Balakrishnan, J. Guttag, and M. R. Sabuncu. "Unsupervised learning for fast probabilistic diffeomorphic registration," *arXiv preprint arXiv:1805.04605* (2018) (cited on p. 72).

98. H. Qiu, C. Qin, A. Schuh, K. Hammernik, and D. Rueckert. "Learning diffeomorphic and modality-invariant registration using B-splines," (2021) (cited on p. 72).

99. V. Arsigny, O. Commowick, X. Pennec, and N. Ayache. "A log-euclidean framework for statistics on diffeomorphisms," *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer. 2006, pp. 924–931 (cited on p. 72).

100. M. Viergever, J. Maintz, S. Klein, K. Murphy, M. Staring, and J. Pluim. "A survey of medical image registration–under review," *Medical Image Analysis*, vol. 33 (2016), pp. 140–144 (cited on pp. 77, 78, 100).

101. G. Haskins, U. Kruger, and P. Yan. "Deep learning in medical image registration: a survey," *Machine Vision and Applications*, vol. 31 (2020), p. 8 (cited on pp. 77, 78).

102. L. Maier-Hein, A. Reinke, M. Kozubek, A. L. Martel, T. Arbel, M. Eisenmann, A. Hanbury, P. Jannin, H. Müller, S. Onogur, et al. "BIAS: transparent reporting of biomedical image analysis challenges," *Medical image analysis*, vol. 66 (2020), p. 101796 (cited on pp. 77, 79, 145, 146).

103. J. West, J. M. Fitzpatrick, M. Y. Wang, B. M. Dawant, C. R. Maurer Jr, R. M. Kessler, R. J. Maciunas, C. Barillot, D. Lemoine, A. Collignon, et al. "Comparison and evaluation of retrospective intermodality brain image registration techniques," *Journal of computer assisted tomography*, vol. 21 (1997), pp. 554–568 (cited on pp. 78, 143).

104. A. Klein, J. Andersson, B. A. Ardekani, J. Ashburner, B. Avants, M.-C. Chiang, G. E. Christensen, D. L. Collins, J. Gee, P. Hellier, et al. "Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration," *Neuroimage*, vol. 46 (2009), pp. 786–802 (cited on p. 78).

105. Z. Xu, C. P. Lee, M. P. Heinrich, M. Modat, D. Rueckert, S. Ourselin, R. G. Abramson, and B. A. Landman. "Evaluation of six registration methods for the human abdomen on clinically acquired CT," *IEEE Transactions on Biomedical Engineering*, vol. 63 (2016), pp. 1563–1572 (cited on pp. 78, 81, 93, 97, 143).

106. B. B. Avants, C. L. Epstein, M. Grossman, and J. C. Gee. "Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain," *Medical image analysis*, vol. 12 (2008), pp. 26–41 (cited on pp. 78, 142).

107. A. V. Dalca, G. Balakrishnan, J. Guttag, and M. R. Sabuncu. "Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces," *Medical image analysis*, vol. 57 (2019), pp. 226–236 (cited on p. 78).

108. Y. Xiao, M. Fortin, G. Unsgård, H. Rivaz, and I. Reinertsen. "EASY-RESECT," 2020. DOI 10.11582/2020.00025 (cited on p. 81).

109. Y. Xiao, M. Fortin, G. Unsgård, H. Rivaz, and I. Reinertsen. "Retrospective evaluation of cerebral tumors (RESECT): a clinical database of pre-operative MRI and intra-operative ultrasound in low-grade glioma surgeries," *Medical physics*, vol. 44 (2017), pp. 3875–3882 (cited on p. 81).

110. Y. Xiao, H. Rivaz, M. Chabanas, M. Fortin, I. Machado, Y. Ou, M. P. Heinrich, J. A. Schnabel, X. Zhong, A. Maier, et al. "Evaluation of mri to ultrasound registration methods for brain shift correction: the CuRIOUS2018 challenge," *IEEE Transactions on Medical Imaging*, vol. 39 (2019), pp. 777–786 (cited on pp. 81, 143).

111. M. Antonelli, A. Reinke, S. Bakas, K. Farahani, B. A. Landman, G. Litjens, B. Menze, O. Ronneberger, R. M. Summers, B. van Ginneken, et al. "The medical segmentation decathlon," *arXiv preprint arXiv:2106.05735* (2021) (cited on p. 81).

112. D. S. Marcus, T. H. Wang, J. Parker, J. G. Csernansky, J. C. Morris, and R. L. Buckner. "Open access series of imaging studies (OASIS): cross-sectional mri data in young, middle aged, nondemented, and demented older adults," *Journal of cognitive neuroscience*, vol. 19 (2007), pp. 1498–1507 (cited on p. 82).

113. A. Hoopes, M. Hoffmann, B. Fischl, J. Guttag, and A. V. Dalca. "Hypermorph: amortized hyperparameter learning for image registration," *International Conference on Information Processing in Medical Imaging*, Springer. 2021, pp. 3–17 (cited on pp. 82, 148).

114. B. Fischl. "FreeSurfer," *Neuroimage*, vol. 62 (2012), pp. 774–781 (cited on p. 82).

115. A. Hering, K. Murphy, and B. van Ginneken. "Learn2Reg challenge: CT lung registration - training data," 2020. DOI 10.5281/zenodo.3835682 (cited on p. 82).

116. A. Hering, K. Murphy, and B. van Ginneken. "Learn2Reg challenge: CT lung registration - test data," 2020. DOI 10.5281/zenodo.4048761 (cited on p. 82).

117. A. D. Leow, I. Yanovsky, M.-C. Chiang, A. D. Lee, A. D. Klunder, A. Lu, J. T. Becker, S. W. Davis, A. W. Toga, and P. M. Thompson. "Statistical properties of jacobian maps and the realization of unbiased large-deformation nonlinear image registration," *IEEE transactions on medical imaging*, vol. 26 (2007), pp. 822–832 (cited on p. 83).

118. S. Kabus, T. Klinder, K. Murphy, B. van Ginneken, C. Lorenz, and J. P. Pluim. "Evaluation of 4D-CT lung registration," *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer. 2009, pp. 747–754 (cited on p. 83).

119. N. Shusharina, M. P. Heinrich, and R. Huang, eds. "Segmentation, classification, and registration of multi-modality medical imaging data," Springer International Publishing, 2021 (cited on p. 83).

120. L. Han, H. Dou, Y. Huang, and P.-T. Yap. "Deformable registration of brain mr images via a hybrid loss," *arXiv preprint arXiv:2110.15027* (2021) (cited on p. 83).

121. DeepRegNet. "Deepregnet," 2021 (cited on p. 85).

122. H. Siebert, L. Hansen, and M. P. Heinrich. "Fast 3d registration with accurate optimisation and little learning for learn2reg 2021," 2021 (cited on pp. 85, 146).

123. M. P. Heinrich, H. Handels, and I. J. Simpson. "Estimating large lung motion in COPD patients by symmetric regularised correspondence fields," *International conference on medical image computing and computer-assisted intervention*, Springer. 2015, pp. 338–345 (cited on pp. 85, 98, 146).

124. J. Lv, Z. Wang, H. Shi, H. Zhang, S. Wang, Y. Wang, and Q. Li. "Joint progressive and coarse-to-fine registration of brain MRI via deformation field integration and non-rigid feature fusion," *arXiv preprint arXiv:2109.12384* (2021) (cited on p. 85).

125. C. Fourcade, M. Rubeaux, and D. Mateus. "Using Elastix to register inhale/exhale intra-subject thorax CT: a unsupervised baseline to the task 2 of the learn2reg challenge," *International Conference on Medical Image Computing and Computer-Assisted Intervention (Workshops)*, Springer. 2020, pp. 100–105 (cited on p. 85).

126. T. Estienne, M. Lerousseau, M. Vakalopoulou, E. Alvarez Andres, E. Battistella, A. Carré, S. Chandra, S. Christodoulidis, M. Sahasrabudhe, R. Sun, et al. "Deep learning-based concurrent brain registration and tumor segmentation," *Frontiers in computational neuroscience*, vol. 14 (2020), p. 17 (cited on pp. 85, 150).

127. T. Estienne, M. Vakalopoulou, E. Battistella, A. Carré, T. Henry, M. Lerousseau, C. Robert, N. Paragios, and E. Deutsch. "Deep learning based registration using spatial gradients and noisy segmentation labels," *Segmentation, Classification, and Registration of Multi-modality Medical Imaging Data*, vol. 12587 (2020), p. 87 (cited on p. 85).

128. N. Gunnarsson, J. Sjölund, and T. B. Schön. "Learning a deformable registration pyramid," *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer. 2020, pp. 80–86 (cited on p. 86).

129. D. Sun, X. Yang, M.-Y. Liu, and J. Kautz. "Pwc-net: CNNs for optical flow using pyramid, warping, and cost volume," *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8934–8943 (cited on p. 86).

130. T. C. Mok and A. Chung. "Conditional deformable image registration with convolutional neural network," *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer. 2021, pp. 35–45 (cited on pp. 86, 100, 146, 148).

131. V. Jaouen, P.-H. Conze, D. Guillaume, J. Bert, and D. Visvikis. "Regularized directional representations for medical image registration," 2021 (cited on p. 86).

132. G. Lifshitz and D. Raviv. "Cost function unrolling in unsupervised optical flow," 2021 (cited on p. 87).

133. S. Häger, S. Heldmann, A. Hering, S. Kuckertz, and A. Lange. "Variable Fraunhofer MEVIS RegLib comprehensively applied to Learn2Reg challenge," *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer. 2020, pp. 74–79 (cited on pp. 87, 146).

134. M. Brudfors, Y. Balbastre, G. Flandin, P. Nachev, and J. Ashburner. "Flexible bayesian modelling for nonlinear image registration," *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer. 2020, pp. 253–263 (cited on p. 87).

135. M. P. Heinrich and L. Hansen. "Highly accurate and memory efficient unsupervised learning-based discrete CT registration using 2.5 D displacement search," *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer. 2020, pp. 190–200 (cited on p. 87).

136. W. Shao, Y. Pan, O. C. Durumeric, J. M. Reinhardt, J. E. Bayouth, M. Rusu, and G. E. Christensen. "Geodesic density regression for correcting 4DCT pulmonary respiratory motion artifacts," *Medical Image Analysis* (2021), p. 102140 (cited on p. 88).

137. E. Gibson, F. Giganti, Y. Hu, E. Bonmati, S. Bandula, K. Gurusamy, B. Davidson, S. P. Pereira, M. J. Clarkson, and D. C. Barratt. "Automatic multi-organ segmentation on abdominal CT with dense V-networks," *IEEE transactions on medical imaging*, vol. 37 (2018), pp. 1822–1834 (cited on p. 88).

138. A. Hering, S. Häger, J. Moltz, N. Lessmann, S. Heldmann, and B. van Ginneken. "CNN-based lung CT registration with multiple anatomical constraints," *Medical Image Analysis* (2021), p. 102139 (cited on pp. 89, 98).

139. M. Hoffmann, B. Billot, D. N. Greve, J. E. Iglesias, B. Fischl, and A. V. Dalca. "SynthMorph: learning contrast-invariant registration without acquired images," *IEEE Transactions on Medical Imaging* (2021) (cited on p. 89).

140. H. Siebert, L. Hansen, and M. P. Heinrich. "Evaluating design choices for deep learning registration networks," *Bildverarbeitung für die Medizin 2021*, Springer, 2021, pp. 111–116 (cited on p. 93).

141. M. P. Heinrich, M. Jenkinson, M. Bhushan, T. Matin, F. V. Gleeson, M. Brady, and J. A. Schnabel. "MIND: modality independent neighbourhood descriptor for multi-modal deformable registration," *Medical image analysis*, vol. 16 (2012), pp. 1423–1435 (cited on p. 94).

142. E. Haber and J. Modersitzki. "Intensity gradient based registration and fusion of multi-modal images," *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer. 2006, pp. 726–733 (cited on p. 94).

143. A. E. Kavur, N. S. Gezer, M. Barış, S. Aslan, P.-H. Conze, V. Groza, D. D. Pham, S. Chatterjee, P. Ernst, S. Özkan, et al. "CHAOS challenge-combined (CT-MR) healthy abdominal organ segmentation," *Medical Image Analysis*, vol. 69 (2021), p. 101950 (cited on p. 94).

144. E. A. Eisenhauer, P. Therasse, J. Bogaerts, L. H. Schwartz, D. Sargent, R. Ford, J. Dancey, S. Arbuck, et al. "New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1)," *European Journal of Cancer*, vol. 45 (2009), pp. 228–247 (cited on pp. 105, 123, 135, 149, 152).

145. A. W. Moawad, D. Fuentes, A. M. Khalaf, K. J. Blair, J. Szklaruk, A. Qayyum, J. D. Hazle, and K. M. Elsayes. "Feasibility of automated volumetric assessment of large hepatocellular carcinomas' responses to transarterial chemoembolization," *Frontiers in Oncology*, vol. 10 (2020), p. 572 (cited on pp. 105, 123).

146. R. J. Gillies, P. E. Kinahan, and H. Hricak. "Radiomics: images are more than pictures, they are data," *Radiology*, vol. 278 (2016), pp. 563–577 (cited on pp. 105, 123, 135).

147. D. Schadendorf, A. C. van Akkooi, C. Berking, K. G. Griewank, R. Gutzmer, A. Hauschild, A. Stang, A. Roesch, and S. Ugurel. "Melanoma," *The Lancet*, vol. 392 (2018), pp. 971–984 (cited on p. 105).

148. W. H. Ward and J. M. Farma. "Cutaneous melanoma: etiology and therapy [internet]," (2017) (cited on p. 105).

149. P. Bilic, P. F. Christ, E. Vorontsov, G. Chlebus, H. Chen, Q. Dou, C.-W. Fu, X. Han, P.-A. Heng, J. Hesser, et al. "The liver tumor segmentation benchmark (LiTs)," *arXiv preprint arXiv:1901.04056* (2019) (cited on pp. 105, 123).

150. N. Heller, F. Isensee, K. H. Maier-Hein, X. Hou, C. Xie, F. Li, Y. Nan, G. Mu, Z. Lin, M. Han, et al. "The state of the art in kidney and kidney tumor segmentation in contrast-enhanced CT imaging: results of the kits19 challenge," *Medical Image Analysis*, vol. 67 (2021), p. 101821 (cited on p. 105).

151. J. Cai, Y. Tang, K. Yan, A. P. Harrison, J. Xiao, G. Lin, and L. Lu. "Deep lesion tracker: monitoring lesions in 4D longitudinal imaging studies," *arXiv preprint arXiv:2012.04872* (2020) (cited on pp. 106, 109, 148, 150).

152. J. Xu, H. Greenspan, S. Napel, and D. L. Rubin. "Automated temporal tracking and segmentation of lymphoma on serial CT examinations," *Medical physics*, vol. 38 (2011), pp. 5879–5886 (cited on p. 106).

153. J. H. Moltz, M. D'Anastasi, A. Kießling, D. P. Dos Santos, C. Schülke, and H.-O. Peitgen. "Workflow-centred evaluation of an automatic lesion tracking software for chemotherapy monitoring by CT," *European radiology*, vol. 22 (2012), pp. 2759–2767 (cited on pp. 106, 107, 132, 133, 150).

154. L. R. Folio, M. M. Choi, J. M. Solomon, and N. P. Schaub. "Automated registration, segmentation, and measurement of metastatic melanoma tumors in serial CT scans," *Academic radiology*, vol. 20 (2013), pp. 604–613 (cited on p. 106).

155. S. Nikolov, S. Blackwell, R. Mendes, J. De Fauw, C. Meyer, C. Hughes, H. Askham, B. Romera-Paredes, A. Karthikesalingam, C. Chu, et al. "Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy," *arXiv preprint arXiv:1809.04430* (2018) (cited on p. 109).

156. B. Li, W. J. Niessen, S. Klein, M. de Groot, M. A. Ikram, M. W. Vernooij, and E. E. Bron. "A hybrid deep learning framework for integrated segmentation and registration: evaluation on longitudinal white matter tract changes," *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer. 2019, pp. 645–653 (cited on pp. 113, 150).

157. V. Kumar, Y. Gu, S. Basu, A. Berglund, S. A. Eschrich, M. B. Schabath, K. Forster, H. J. Aerts, A. Dekker, D. Fenstermacher, et al. "Radiomics: the process and the challenges," *Magnetic resonance imaging*, vol. 30 (2012), pp. 1234–1248 (cited on pp. 123, 135).

158. P. M. Cheng, E. Montagnon, R. Yamashita, I. Pan, A. Cadrin-Chênevert, F. Perdigón Romero, G. Chartrand, S. Kadoury, and A. Tang. "Deep learning: an update for radiologists," *RadioGraphics*, vol. 41 (2021), pp. 1427–1445 (cited on p. 123).

159. Z. Li and Y. Xia. "Deep reinforcement learning for weakly-supervised lymph node segmentation in CT images," *IEEE Journal of Biomedical and Health Informatics*, vol. 25 (2020), pp. 774–783 (cited on p. 123).

160. A. Hering, F. Peisen, T. Amaral, S. Gatidis, T. Eigentler, A. Othman, and J. H. Moltz. "Whole-body soft-tissue lesion tracking and segmentation in longitudinal CT imaging studies," *Medical Imaging with Deep Learning*, 2021 (cited on pp. 123, 126).

161. E. Vorontsov, M. Cerny, P. Régnier, L. Di Jorio, C. J. Pal, R. Lapointe, F. Vandenbroucke-Menu, S. Turcotte, S. Kadoury, and A. Tang. "Deep learning for automated segmentation of liver lesions at CT in patients with colorectal cancer liver metastases," *Radiology: Artificial Intelligence*, vol. 1 (2019), p. 180014 (cited on pp. 132, 135).

162. D. T. Kushnure and S. N. Talbar. "MS-UNet: a multi-scale U-Net with feature recalibration approach for automatic liver and tumor segmentation in CT images," *Computerized Medical Imaging and Graphics*, vol. 89 (2021), p. 101885 (cited on p. 132).

163. M. Moghbel, S. Mashohor, R. Mahmud, and M. I. B. Saripan. "Automatic liver tumor segmentation on computed tomography for patient treatment planning and monitoring," *EXCLI journal*, vol. 15 (2016), p. 406 (cited on p. 132).

164. G. Chlebus, H. Meine, S. Thoduka, N. Abolmaali, B. Van Ginneken, H. K. Hahn, and A. Schenk. "Reducing inter-observer variability and interaction time of MR liver volumetry by combining automatic cnn-based liver segmentation and manual corrections," *PloS one*, vol. 14 (2019), p. e0217228 (cited on p. 133).

165. M. Kosmin, J. Ledsam, B. Romera-Paredes, R. Mendes, S. Moinuddin, D. de Souza, L. Gunn, C. Kelly, C. Hughes, A. Karthikesalingam, et al. "Rapid advances in auto-segmentation of organs at risk and target volumes in head and neck cancer," *Radiotherapy and Oncology*, vol. 135 (2019), pp. 130–140 (cited on p. 133).

166. L. J. Stapleford, J. D. Lawson, C. Perkins, S. Edelman, L. Davis, M. W. McDonald, A. Waller, E. Schreibmann, and T. Fox. "Evaluation of automatic atlas-based lymph node segmentation for head-and-neck cancer," *International Journal of Radiation Oncology* Biology* Physics*, vol. 77 (2010), pp. 959–966 (cited on p. 133).

167. Y. Tang, K. Yan, J. Xiao, and R. M. Summers. "One click lesion recist measurement and segmentation on CT scans," *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer. 2020, pp. 573–583 (cited on pp. 136, 150).

168. B. B. Avants, N. J. Tustison, M. Stauffer, G. Song, B. Wu, and J. C. Gee. "The insight ToolKit image registration framework," *Frontiers in neuroinformatics*, vol. 8 (2014), p. 44 (cited on p. 142).

169. P. Hellier, C. Barillot, I. Corouge, B. Gibaud, G. Le Goualher, D. L. Collins, A. Evans, G. Malandain, N. Ayache, G. E. Christensen, et al. "Retrospective evaluation of intersubject brain registration," *IEEE transactions on medical imaging*, vol. 22 (2003), pp. 1120–1130 (cited on p. 143).

170. G. E. Christensen, X. Geng, J. G. Kuhl, J. Bruss, T. J. Grabowski, I. A. Pirwani, M. W. Vannier, J. S. Allen, and H. Damasio. "Introduction to the non-rigid image registration evaluation project (NIREP)," *International workshop on biomedical image registration*, Springer. 2006, pp. 128–135 (cited on pp. 143–145).

171. J. Borovec, J. Kybic, I. Arganda-Carreras, D. V. Sorokin, G. Bueno, A. V. Khvostikov, S. Bakas, I. Eric, C. Chang, S. Heldmann, et al. "ANHIR: automatic non-rigid histological image registration challenge," *IEEE transactions on medical imaging*, vol. 39 (2020), pp. 3042–3052 (cited on p. 143).

172. K. Marstal, F. Berendsen, N. Dekker, M. Staring, and S. Klein. "The continuous registration challenge: evaluation-as-a-service for medical image registration algorithms," *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, IEEE. 2019, pp. 1399–1402 (cited on p. 143).

173. L. Hansen, A. Hering, M. P. Heinrich, A. Dalca, et al. "Learn2Reg: 2020 MICCAI registration challenge; 2020," (cited on p. 143).

174. J. M. Fitzpatrick and J. B. West. "The distribution of target registration error in rigid-body point-based registration," *IEEE transactions on medical imaging*, vol. 20 (2001), pp. 917–927 (cited on p. 144).

175. J. Lotz, N. Weiss, and S. Heldmann. "Robust, fast and accurate: a 3-step method for automatic histological image registration," *arXiv preprint arXiv:1903.12063* (2019) (cited on p. 146).

176. W. Wein. "Brain-shift correction with image-based registration and landmark accuracy evaluation," *Simulation, Image Processing, and Ultrasound Systems for Assisted Diagnosis and Navigation*, Springer, 2018, pp. 146–151 (cited on p. 146).

177. K. Yan, X. Wang, L. Lu, and R. M. Summers. "DeepLesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning," *Journal of medical imaging*, vol. 5 (2018), p. 036501 (cited on pp. 148, 150).

178. M. Bellomi, F. De Piano, E. Ancona, A. F. Lodigiani, G. Curigliano, S. Raimondi, and L. Preda. "Evaluation of inter-observer variability according to RECIST 1.1 and its influence on response classification in CT measurement of liver metastases," *European journal of radiology*, vol. 95 (2017), pp. 96–101 (cited on p. 149).

179. Y.-B. Tang, K. Yan, Y.-X. Tang, J. Liu, J. Xiao, and R. M. Summers. "ULDor: a universal lesion detector for CT scans with pseudo masks and hard negative example mining," *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, IEEE. 2019, pp. 833–836 (cited on p. 151).

180. Q. Tao, Z. Ge, J. Cai, J. Yin, and S. See. "Improving deep lesion detection using 3D contextual and spatial attention," *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer. 2019, pp. 185–193 (cited on p. 151).

181. K. Yan, M. Bagheri, and R. M. Summers. "3D context enhanced region-based convolutional neural network for end-to-end lesion detection," *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer. 2018, pp. 511–519 (cited on p. 151).

182. K. Yan, Y. Tang, Y. Peng, V. Sandfort, M. Bagheri, Z. Lu, and R. M. Summers. "MULAN: multitask universal lesion analysis network for joint lesion detection, tagging, and segmentation," *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer. 2019, pp. 194–202 (cited on p. 151).

183. S. Keil, A. Barabasch, T. Dirrichs, P. Bruners, N. L. Hansen, H. B. Bieling, T. H. Brümmendorf, and C. K. Kuhl. "Target lesion selection: an important factor causing variability of response classification in the response evaluation criteria for solid tumors 1.1," *Investigative radiology*, vol. 49 (2014), pp. 509–517 (cited on p. 151).

184. M. Sieren, F. Brenne, A. Hering, H. Kienapfel, N. Gebauer, T. Oechtering, A. Fürschke, F. Wegner, E. Stahlberg, S. Heldmann, et al. "Rapid study assessment in follow-up whole-body computed tomography in patients with multiple myeloma using a dedicated bone subtraction software," *European radiology* (2020), pp. 1–12 (cited on p. 152).

# Publications

## Journal publications

F. Peisen, A. Hänsch, **A. Hering**, A. S. Brendlin, S. Afat, K. Nikolaou, S. Gatidis, T. Eigentler, T. Amaral, J. H. Moltz, et al. "Combination of whole-body baseline ct radiomics and clinical parameters to predict response and survival in a stage-iv melanoma cohort undergoing immunotherapy," *Cancers*, vol. 14 (2022), p. 2992.

**A. Hering**, S. Häger, J. Moltz, N. Lessmann, S. Heldmann, and B. van Ginneken. "CNN-based lung CT registration with multiple anatomical constraints," *Medical Image Analysis* (2021), p. 102139.

M. Sieren, F. Brenne, **A. Hering**, H. Kienapfel, N. Gebauer, T. Oechtering, A. Fürschke, F. Wegner, E. Stahlberg, S. Heldmann, et al. "Rapid study assessment in follow-up whole-body computed tomography in patients with multiple myeloma using a dedicated bone subtraction software," *European radiology* (2020), pp. 1–12.

**A. Hering**, S. Kuckertz, S. Heldmann, and M. P. Heinrich. "Memory-efficient 2.5 D convolutional transformer networks for multi-modal deformable registration with weak label supervision applied to whole-heart CT and MRI scans," *International journal of computer assisted radiology and surgery*, vol. 14 (2019), pp. 1901–1912.

## Conference proceedings

**A. Hering**, F. Peisen, and J. H. Moltz. "Towards more efficient tumor follow-up assessment using ai assistance," *Medical Imaging with Deep Learning*, 2022.

L. Hansen, **A. Hering**, C. Großbröhmer, and M. P. Heinrich. "Continuous benchmarking in medical image registration-review of the current state of the learn2reg challenge," *Medical Imaging with Deep Learning* (2022).

**A. Hering**, A. Lange, S. Heldmann, S. Häger, and S. Kuckertz. "Fraunhofer MEVIS image registration solutions for the Learn2Reg 2021 challenge," *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer. 2021, pp. 147–152.

**A. Hering**, F. Peisen, T. Amaral, S. Gatidis, T. Eigentler, A. Othman, and J. H. Moltz. "Whole-body soft-tissue lesion tracking and segmentation in longitudinal CT imaging studies," *Medical Imaging with Deep Learning*, 2021.

B. Lassen-Schmidt, **A. Hering**, S. Krass, and H. Meine. "Automatic segmentation of the pulmonary lobes with a 3D U-Net and optimized loss function," *Medical Imaging with Deep Learning*, 2020.

S. Häger, S. Heldmann, **A. Hering**, S. Kuckertz, and A. Lange. "Variable Fraunhofer MEVIS RegLib comprehensively applied to Learn2Reg challenge," *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer. 2020, pp. 74–79.

**A. Hering**, B. van Ginneken, and S. Heldmann. "mlVIRNET: multilevel variational image registration network," *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2019*, Springer. 2019, pp. 257–265.

H. Meine and **A. Hering**. "Efficient prealignment of CT scans for registration through a bodypart regressor," *Medical Imaging with Deep Learning*, 2019.

**A. Hering** and S. Heldmann. "Unsupervised learning for large motion thoracic CT follow-up registration," *Medical Imaging 2019: Image Processing*, vol. 10949 International Society for Optics and Photonics. (2019), 109491B.

**A. Hering**, S. Kuckertz, S. Heldmann, and M. P. Heinrich. "Enhancing label-driven deep deformable image registration with local distance metrics for state-of-the-art cardiac motion tracking," *Bildverarbeitung für die Medizin 2019*, Springer, 2019, pp. 309–314.

## Preprints

**A. Hering**, L. Hansen, T. C. Mok, A. Chung, H. Siebert, S. Häger, A. Lange, S. Kuckertz, S. Heldmann, W. Shao, et al. "Learn2reg: comprehensive multi-task medical image registration challenge, dataset and evaluation in the era of deep learning," *arXiv preprint arXiv:2112.04489* (2021).

## Conference abstracts

**A. Hering**, F. Peisen, A. Othman, and J. H. Moltz. "Minimally interactive AI-based segmentation of lymph nodes and soft tissue lesions in follow-up CT," European Congress of Radiology - ECR. 2022.

A. Fürschke, A. Loh, S. Kopelke, N. Gebauer, **A. Hering**, A. Derksen, J. Barkhausen, and A. Frydrychowicz. "Validating a novel, semi-automatic software for quantitation of cross-sectional muscle area," European Congress of Radiology - ECR. 2019.

M. Sieren, F. Brenne, **A. Hering**, T. Oechtering, H. Kienapfel, J. Barkhausen, and A. Frydrychowicz. "Bewertung einer „Smart-Linking "Software zur computergestützten Evaluation von Follow-up Ganzkörper Computertomografien bei Patienten mit ossärem Plamozytomcts," *RöFo-Fortschritte auf dem Gebiet der Röntgenstrahlen und der bildgebenden Verfahren*, vol. 191 Georg Thieme Verlag KG. (2019), WISS–107.

A. Loh, A. Fürschke, A. Frydrychowicz, J. Barkhausen, S. Kopelke, N. Gebauer, **A. Hering**, and A. Derksen. "Validierung einer neuentwickelten Software (SMQ) zur Quantifizierung der thorakalen Muskeldichte im Rahmen der Sarkopenie-Diagnostik," *RöFo-Fortschritte auf dem Gebiet der Röntgenstrahlen und der bildgebenden Verfahren*, vol. 191 Georg Thieme Verlag KG. (2019), WISS–106.

A. Fürschke, A. Loh, S. Kopelke, N. Gebauer, **A. Hering**, A. Derksen, J. Barkhausen, and A. Frydrychowicz. "Validierung einer neuartigen, halbautomatischen Software zur Quantifizierung von Muskelmasse in der Computertomografie," *RöFo-Fortschritte auf dem Gebiet der Röntgenstrahlen und der bildgebenden Verfahren*, vol. 191 Georg Thieme Verlag KG. (2019), WISS–303.

# Acknowledgments

*"The outcome of your Ph.D. trajectory is not only your thesis but also yourself."* I read this somewhere and haven't forgotten it, because it changes the way I look at my time as a Phd candidate. Many people have accompanied and supported me on my way - not only in writing the papers for this thesis but also in growing myself. I am incredibly grateful to all of you.

I would like to thank you, **Bram**, for giving me the opportunity of writing this thesis. I am grateful for your guidance and advice, from which I have learned a lot. I always felt like I could find my own way and do the research I wanted, but I could always count on you if I needed you.

**Nikolas**, you are the person who has had the most influence on my paper writing style and the way I review papers. You have inspired me to present my results in a far more honest and fair way and to value the work of others even more. Thank you, you have helped me to become a better scientist!

The thesis was enabled by and carried out at Fraunhofer MEVIS. Therefore, I would also like to thank my MEVIS supervisors.
**Stefan**, I am incredibly grateful to you for the various and valuable ideas and discussions; for giving me so many opportunities, and that we always found solutions together to achieve everything we wanted.

**Horst**, I would like to thank you for creating such a pleasant working atmosphere at MEVIS, for your enormous trust in every single person, and for always taking time for everyone who needs your advice, despite your immense workload.

In addition to my official supervision team, I was lucky to have you, **Jan**, as an additional mentor. You have accompanied me in my projects since my first day at MEVIS and have not been able to get rid of me. I would like to thank you for your patience, your kindness and the many hours we spent together discussing ideas.

I would also like to thank my thesis committee for taking the time to review my thesis.

A special thanks goes to **Nadine** - even though she, unfortunately, can no longer read this as she passed away far too early. Nadine welcomed me so warmly on my first days at MEVIS and guided me through my first steps towards my PhD by asking me the right questions.

My work would not have been possible without all the previous work that had already been done at MEVIS and on which my work is based. For this, I would like to thank all my colleagues at Fraunhofer MEVIS - especially **Jan R.**, **Lars** and **Alex**.

The best thing about colleagues is when they are not just colleagues. I would like to thank my colleagues from Lübeck and Bremen for all the wonderful moments spent together at lunch, at the coffee machine, at our YES, during a barbecue or a game of Boßeln. For the tough but fair kicker matches even if it meant ending up on the list of shame. I particularly would like to thank **Annkristin** and **Stephanie** for our virtual morning rounds which made the start of the day much more pleasant. Many thanks to **Janine, Alex, Fred, Sven and Sina** for sharing the office and for many entertaining and encouraging conversations.

I am also grateful to my colleagues from Nijmegen and especially the BodyCT group for welcoming me so kindly and for all the meetings also taking place virtually because of me. Due to Covid, I was not able to visit you as often as I had imagined, but I am sure we will make up for it in the future at DIAG days and conferences! I especially would like to thank you **Kiran** for never making me feel like I was bugging you with all my questions and of course for agreeing to be my paranymph and all the interesting conversations.

In addition to the colleagues from the three groups in Lübeck, Bremen and Nijmegen, I have been lucky to have had further people by my side.
**Tanja**, thank you so much for the conferences together - with you it always felt so easy and fun. The discussions with you at other people's posters were so often far more enlightening than with the actual authors.

**Mattias**, I enjoy organizing events with you, whether Learn2Reg, WBIR or MIDL. I have learned a lot from you, but I still don't know how you manage it all with only 24 hours a day. Thank you for everything!

**Jannis**, it's incredibly fun to make great plans with you. I'm pretty sure we'll put them all into action!

**Hoel**, thank you for many shared conference memories, for always having an open ear and for almost always having a solution!

Furthermore, I would like to take this opportunity to thank all the people who have made all the conferences so memorable.

Fortunately, the last few years have not been all work, so I am grateful for the support of the people in my private life. I would like to thank all my friends for always providing a balance to work that was so often desperately needed.

In particular, I would like to thank **Tanja** and **Ramona** for all the shared activities and the evenings with oven cheese and wine.

I would like to thank my family (**Mama, Torben, Alena and Lars**) for always supporting me no matter what.

I am grateful to you, **Christian**, for supporting me so patiently on the last few meters.

Aber am meisten möchte ich dir, **Julius**, danken - dafür, dass du der besten Grund bist nicht zu arbeiten, Spaß zu haben und Unfug zu machen.

# Biography

Alessa Hering was born on December 19, 1990 in Hamburg, Germany. She studied Computational Life Science at the University of Lübeck, Germany. In 2017, she obtained her Master's degree after completing her thesis on regression of shape information in medical images. Afterwards, she started working as a research scientist at Fraunhofer MEVIS. Additionally, she joined the Diagnostic Image Analysis Group at Radboud University Medical Center in Nijmegen as an external PhD candidate to work on deep-learning-based image registration under supervision of Bram van Ginneken. Her research was partially funded by the German Academic Scholarship Foundation. She continues her research as postdoctoral researcher at the Diagnostic Image Analysis Group at Radboud University Medical Center and at Fraunhofer MEVIS.