

# A Deep Learning Framework for Unsupervised Affine and Deformable Image Registration

Bob D. de Vos<sup>1</sup>, Floris F. Berendsen<sup>2</sup>, Max A. Viergever<sup>1</sup>, Hessam Sokooti<sup>2</sup>, Marius Staring<sup>2</sup>, Ivana Išgum<sup>1</sup>

## Abstract

Image registration, the process of aligning two or more images, is the core technique of many (semi-)automatic medical image analysis tasks. Recent studies have shown that deep learning methods, notably convolutional neural networks (ConvNets), can be used for image registration. Thus far training of ConvNets for registration was supervised using predefined example registrations. However, obtaining example registrations is not trivial. To circumvent the need for predefined examples, and thereby to increase convenience of training ConvNets for image registration, we propose the Deep Learning Image Registration (DLIR) framework for *unsupervised* affine and deformable image registration. In the DLIR framework ConvNets are trained for image registration by exploiting image similarity analogous to conventional intensity-based image registration. After a ConvNet has been trained with the DLIR framework, it can be used to register pairs of unseen images in one shot. We propose flexible ConvNets designs for affine image registration and for deformable image registration. By stacking multiple of these ConvNets into a larger architecture, we are able to perform coarse-to-fine image registration. We show for registration of cardiac cine MRI and registration of chest CT that performance of the DLIR framework is comparable to conventional image registration while being several orders of magnitude faster.

*Keywords:* deep learning, unsupervised learning, affine image registration, deformable image registration, cardiac cine MRI, chest CT

## 1. Introduction

Image registration is the process of aligning two or more images. It is a well-established technique in (semi-)automatic medical image analysis that is used to transfer information between images. Commonly used image registration approaches include intensity-based methods, and feature-based methods that use handcrafted image features (Sotiras et al., 2013; Viergever et al., 2016). Since recently, supervised and unsupervised deep learning techniques have been successfully employed for image registration (Cao et al., 2017; de Vos et al., 2017; Eppenhof et al., 2018; Jaderberg et al., 2015; Krebs et al., 2017; Liao et al., 2017; Miao et al., 2016; Sokooti et al., 2017; Wu et al., 2016; Yang et al., 2017).

Deep learning techniques are well suited for image registration, because they automatically learn to aggregate the information of various complexities in images that are relevant for the task at hand. Additionally, the use of deep learning techniques potentially yields high robustness, because local optima may be of lesser concern in deep learning methods, i.e. zero gradients are often (if not always) at saddle points (Dauphin et al., 2014). Moreover,

deep learning methods like convolutional neural networks are highly parallelizable which makes implementation and execution on GPUs straight-forward and fast. As a consequence deep learning enhanced registration methods are exceptionally fast making them interesting for time-critical applications; e.g. for emerging image guided therapies like High Intensity Focused Ultrasound (HIFU), the MRI Linear Accelerator (MR-linac), and MRI-guided proton therapy.

Although not explicitly introduced as a method for image registration, the spatial transformer network (STN) proposed by Jaderberg et al. (2015) was one of the first methods that exploited deep learning for image alignment. The STN is designed as part of a neural network for classification. Its task is to spatially transform input images such that the classification task is simplified. Transformations might be performed using a global transformation model or a thin plate spline model. In the application of an STN, image registration is an implicit result; image alignment is not guaranteed and only performed when beneficial for the classification task at hand. STNs have been shown to aid classification of photographs of traffic signs, house numbers, and handwritten digits, but to the best of our knowledge they have not yet been used to aid classification of medical images.

In other studies deep learning methods were explicitly trained for image registration (Cao et al., 2017; Hu et al.,

<sup>1</sup>Image Sciences Institute, University Medical Center Utrecht and Utrecht University, Utrecht, the Netherlands

<sup>2</sup>Division of Image Processing of the Leiden University Medical Center, Leiden, The Netherlands

2018a,b; Krebs et al., 2017; Liao et al., 2017; Miao et al., 2016; Sokooti et al., 2017; Yang et al., 2017). For example, convolutional neural networks (ConvNets) were trained with reinforcement learning to be agents that predicted small steps of transformations toward optimal alignment. Liao et al. (2017) applied these agents for affine registration of intra-patient cone-beam CT (CBCT) to CT and Krebs et al. (2017) applied agents for deformable image registration of inter-patient prostate MRI. Like intensity-based registration, image registration with agents is iterative. However, ConvNets can also be used to register images in one shot. For example, Miao et al. (2016) used a ConvNet to predict parameters in one shot for rigid registration of 2D CBCT to CT volumes. Similarly, ConvNets have been used to predict parameters of a thin plate spline model. Cao et al. (2017) used thin plate splines for deformable registration of brain MRI scans and Eppenhof et al. (2018) used thin plate splines for deformable registration of chest CT scans. Furthermore, in the work of Sokooti et al. (2017) it has been demonstrated that a ConvNet can be used to predict a dense displacement vector field (DVF) directly, without constraining it to a transformation model. Similarly, Yang et al. (2017) used a ConvNet to predict the momentum for registration with large deformation diffeomorphic metric mapping (Beg et al., 2005). Recently, Hu et al. (2018a) presented a method that employs segmentations to train ConvNets for global and local image registration. In this method a ConvNet takes fixed and moving image pairs as its inputs and it learns to align the segmentations. This was demonstrated on global and deformable registration of ultrasound and MR images using prostate segmentation.

While the aforementioned deep learning-based registration methods show accurate registration performance, the methods are all supervised, i.e. they rely on example registrations for training or require manual segmentations, unlike conventional image registration methods that are typically unsupervised. Training examples for registration have been generated by synthesizing transformation parameters for affine image registration (Miao et al., 2016) and deformable image registration (Eppenhof et al., 2018; Sokooti et al., 2017), or require manual annotations (Hu et al., 2018a,b). However, generating synthetic data may not be trivial as it is problem specific. In contrast to supervised methods, training examples can be obtained by using conventional image registration methods (Cao et al., 2017; Krebs et al., 2017; Liao et al., 2017; Yang et al., 2017). Alternatively, unsupervised deep learning methods could be employed. Wu et al. (2016) exploited unsupervised deep learning by employing a convolutional stacked auto-encoder (CAE) that extracted features from fixed and moving images. It improved registration with Demons (Vercauteren et al., 2009) and HAMMER (Shen and Davatzikos, 2002) on three different brain MRI datasets. However, while the CAE is unsupervised, the extracted features are optimized for image reconstruction and not for image registration. Thus, there is no

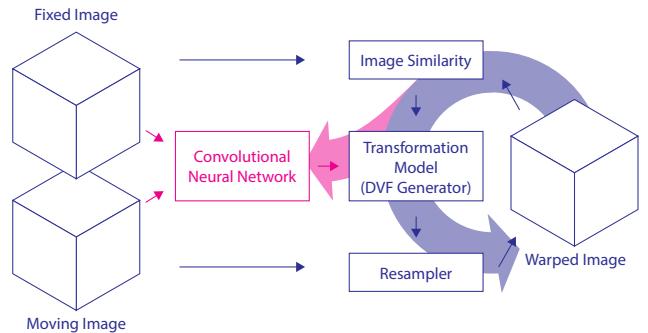


Figure 1: Schematic representation of the deep learning image registration (DLIR) framework. The DLIR training procedure is similar to a conventional iterative image registration framework (blue), but adding a ConvNet in this framework (red) allows unsupervised training for image registration. Unlike in conventional image registration, where image similarity is used to iteratively update the transform parameters directly (large blue arrow), image similarity is used in the DLIR framework to update the weights of the ConvNet by back propagation (large red arrow). Consequently, a trained ConvNet can output a transformation that aligns the input images in one shot.

guarantee that the extracted features are optimal for the specific image registration task.

Unsupervised deep learning has been used to estimate optical flow (Dosovitskiy et al., 2015; Ilg et al., 2017; Yu et al., 2016) or to estimate depth (Garg et al., 2016) in video sequences. Such methods are related to medical image registration, but typically address different problems. They focus on deformations among frames in video sequences. These video sequences are in 2D, contain relatively low levels of noise, have high contrast due to RGB information, and have relatively small deformations between adjacent frames. In contrast, medical images are often 3D, may contain large amounts of noise, may have relatively low contrast and aligning them typically requires larger deformations.

We propose a Deep Learning Image Registration (DLIR) framework: an unsupervised technique to train ConvNets for medical image registration tasks. In the DLIR framework, a ConvNet is trained for image registration by exploiting image similarity between fixed and moving image pairs, thereby circumventing the need for registration examples. The DLIR framework bears similarity with a conventional iterative image registration framework, as shown in Figure 1. However, in contrast to conventional image registration, the transformation parameters are not directly optimized, but indirectly, by optimizing the ConvNet’s parameters. In the DLIR framework the task of a ConvNet is to learn to predict transformation parameters by analyzing fixed and moving image pairs. The predicted transformation parameters are used to make a dense displacement vector field (DVF). The DVF is used to resample the moving image into a warped image that mimics the fixed image. During training, the ConvNet learns the underlying patterns of image registration by optimizing image similarity between the fixed and warped moving image.

ages. Once a ConvNet is trained, it has learned the image registration task and it is able to perform registration on pairs of unseen fixed and moving images in one shot, thus non-iteratively.

The current paper extends our preliminary study of unsupervised deformable image registration (de Vos et al., 2017) in several ways. First, we extend the analysis from 2D to 3D images. Second, we perform B-spline registration with transposed convolutions, which results in high registration speeds, reduces memory footprint, and allows simple implementation of B-spline registration on existing deep learning frameworks. Third, borrowed from conventional image registration where regularization often is an integral part in transformation models (Sotiras et al., 2013), we include a bending energy penalty term that encourages smooth displacements. Fourth, we present ConvNet designs for affine as well as deformable registration. Fifth, we introduce multi-stage ConvNets for registration of coarse-to-fine complexity in multiple-levels and multiple image resolutions by stacking ConvNets for affine and deformable image registration. Such a multi-stage ConvNet can perform registration tasks on fixed and moving pairs of different size, similarly to conventional iterative intensity-based registration strategies. Finally, in addition to evaluation on intra-patient registration of cardiac cine MR images, we conduct experiments on a diverse set of low-dose chest CTs for inter-patient registration, and we evaluate the method on the publicly available DIR-Lab dataset for image registration (Castillo et al., 2010, 2009).

## 2. Method

In image registration the aim is to find a coordinate transformation  $\mathbf{T} : I_F \rightarrow I_M$  that aligns a fixed image  $I_F$  and a moving image  $I_M$ . In conventional image registration similarity between the images is optimized by minimizing a dissimilarity metric  $L$ :

$$\hat{\mu} = \underset{\mu}{\operatorname{argmin}} \{ L(\mathbf{T}_\mu; I_F, I_M) + R(\mathbf{T}_\mu) \},$$

where  $\mathbf{T}_\mu$  is parameterized by transformation parameters  $\mu$  and  $R$  is an optional regularization term to encourage smoothness of the transformation  $\mathbf{T}_\mu$ . Several dissimilarity metrics might be used, e.g. mean squared difference, normalized cross-correlation, and mutual information. Provided the metric is differentiable, optimal transformation parameters can be found by performing (stochastic) gradient descent. In the DLIR framework the ConvNet's task is to predict these transformation parameters using  $I_F$  and  $I_M$  as its inputs:

$$\mu = f_\theta(I_F, I_M),$$

where  $f$  denotes the ConvNet and  $\theta$  the ConvNet's parameters. By minimizing dissimilarity  $L$ , a ConvNet can be trained for image registration as follows:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \{ L(\mathbf{T}_{\mu_\theta}; I_F, I_M) + R(\mathbf{T}_{\mu_\theta}) \}.$$

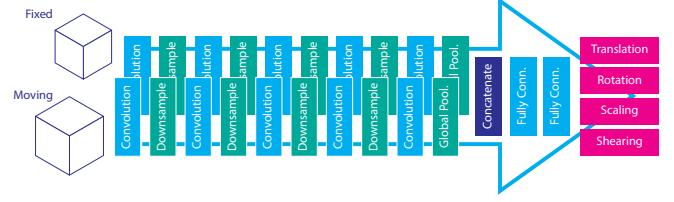


Figure 2: A ConvNet design for affine image registration. The network analyzes pairs of fixed and moving images in separate pipelines. Ending each pipeline with global average pooling enables analysis of input images of different sizes, and allows concatenation with the fully connected layers that have a fixed number of nodes connected to 12 affine transformation parameter outputs.

Note that the parameters  $\theta$  of the ConvNet are optimized and not the parameters  $\mu$  of the mapping function  $\mathbf{T}$ . The ConvNet predicts the parameters  $\mu$  and is trained by the DLIR framework by exploiting image similarity between pairs of fixed and moving input images, i.e. image dissimilarity is calculated between the fixed and the warped moving images and is used as a loss function for ConvNet training. While the DLIR framework could work with other transformation models, below we work out the architectures for an affine (global) parameterization and a B-spline deformable (local) parameterization of the transformation model. In the following subsections we provide details about the DLIR framework, we describe specific ConvNet designs for affine and for deformable image registration, and finally we describe a ConvNet design consisting of multiple stages to perform multi-resolution and multi-level image registration. We refer the reader to Sotiras et al. (2013) for an extensive review on image registration techniques.

### 2.1. Affine Image Registration

Affine transformation is often the first step in image registration, since it simplifies the optimization of subsequent more complex image registration steps. Considering that the affine transformation model is global, we designed a ConvNet that analyzes a pair of input images globally. Considering that medical images often have different image sizes, the proposed ConvNet analyzes fixed and moving images in separate pipelines. The separate pipelines analyze input images independently and therefore eliminate the need for cropping or padding of input image pairs to the same size. In each pipeline the final feature maps will be of different size. Thus global average pooling (Lin et al., 2014) is applied to output one feature per feature map by taking the average of each feature-map. An additional benefit is that global pooling forces the network to encode orientations and affine transformations globally. Subsequently, the network can be connected to a neural network work that will decode the relative orientations of the fixed and moving images and convert those to 12 affine transformation parameters: three translation, three rotation, three scaling, and three shearing parameters.

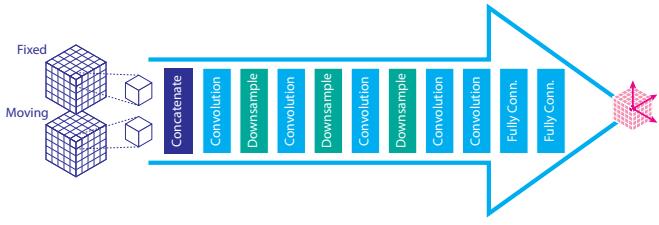


Figure 3: A patch-based ConvNet design for deformable image registration. The ConvNet takes fixed and moving image pairs of equal size as its input—e.g. pre-aligned with affine registration—and outputs a B-spline 3D displacement vector for each patch. The patch and B-spline grid dimensions determine the number of downsampling layers, thus each specific B-spline grid and image resolution requires a dedicated ConvNet design. The fully convolutional patch-based design efficiently generates a B-spline 3D displacement grid of any number of grid points depending on the input images sizes.

Figure 2 illustrates our ConvNet design for affine image registration. The two separate pipelines analyze input pairs of fixed and moving images and each consist of five alternating  $3 \times 3 \times 3$  convolution layers and  $2 \times 2 \times 2$  downsampling layers. The number of these layers may vary, depending on task complexity and input image size. The weights of the layers are shared between the two pipelines to limit the number of total parameters in the network.

## 2.2. Deformable Image Registration

Deformable transformation models can account for local deformations that often occur in medical images. Deformable image registration can be achieved with several transformation models. Here we opt for B-splines (Rueckert et al., 1999) because of their inherent smoothness and local support property: a B-spline control point only affects a specific area in an image, in contrast to e.g. a thin plate spline which has global support. In our ConvNet design we exploit this property by choosing a receptive field that overlaps the support size of the B-spline basis functions, i.e. at least four times the grid spacing for a third order B-spline kernel. The ConvNet takes patches from fixed and moving images and predicts the B-spline control point displacements within that patch. By using a fully convolutional patch-based ConvNet design inspired by Long et al. (2015), input images of arbitrary dimensions can be analyzed efficiently.

The proposed ConvNet design is shown in Figure 3. The ConvNet expects a pair of fixed and moving images of equal size that are concatenated. Depending on the registration problem, moving images might have to be pre-aligned first with e.g. affine registration. After concatenation, alternating layers of  $3 \times 3 \times 3$  convolutions (with 0-padding) and  $2 \times 2 \times 2$  downsampling are applied. The user-chosen B-spline grid spacing determines the amount of required downsampling. A larger grid spacing implies fewer control points, and thus a need for more downsampling layers; by adding more downsampling layers, the receptive field of the ConvNet simultaneously increases. The

two additional  $3 \times 3 \times 3$  convolution layers after the last downsampling layer enlarge the receptive field to the support size of the third order B-spline control points. Thereafter, two  $1 \times 1 \times 1$  convolutional layers are applied, and these are connected to the final convolutional output layer with three  $1 \times 1 \times 1$  kernels that predict the B-spline control points in each of the three directions. The final DVF, used for image resampling, can be generated from the estimated control points by B-spline interpolation.

B-spline interpolation was implemented efficiently by transposed convolutions, also known as fractionally strided convolutions or deconvolutions. Transposed convolutions are the back-bone in ConvNet implementations. They are used to backpropagate loss through the convolutional layers. Due to the  $2 \times 2 \times 2$  downsampling factors resulting in integer grid spacings we can use fixed precomputed B-spline kernels to efficiently upsample B-spline control points to a dense DVF. We use a discrete B-spline kernel as the convolution kernel.

## 2.3. Multi-Stage Image Registration

Conventional image registration is often performed in multiple stages starting with affine registration, followed by coarse-to-fine stages of deformable image registration using B-splines. This hierarchical multi-stage strategy makes conventional iterative image registration less sensitive to local optima and image folding (Schnabel et al., 2001). We adopted this strategy for the DLIR framework by stacking multiple stages of ConvNets, each with its own registration task. For example, a ConvNet for affine registration is followed by multiple ConvNets for coarse-to-fine B-spline registration, each ConvNet with a different B-spline grid spacing and images of different resolution as inputs. When multi-stage registration requires varying input resolutions, we propose average pooling (i.e. windowed averaging), which is a very common building block in deep learning frameworks.

Figure 4 illustrates how such a multi-stage ConvNet can be trained for multi-resolution and multi-level image registration. Training within the DLIR framework is performed sequentially: each stage is trained for its specific registration task, while keeping the weights of ConvNets from preceding stages fixed. After training, the multi-stage ConvNet can be applied for one-shot image registration, similar to a single ConvNet.

## 2.4. Loss Function

The registration ConvNets are trained using mini-batch stochastic gradient descent, hence a differentiable loss is required. Since we perform mono-modal registration experiments, we use normalized cross correlation. Carefully chosen coarse-to-fine levels of multi-stage B-spline registration might prevent image folding and result in smooth deformations (Schnabel et al., 2001). Alternatively, smooth deformations can be encouraged by using a bending energy penalty as proposed by Rueckert et al. (1999). The loss

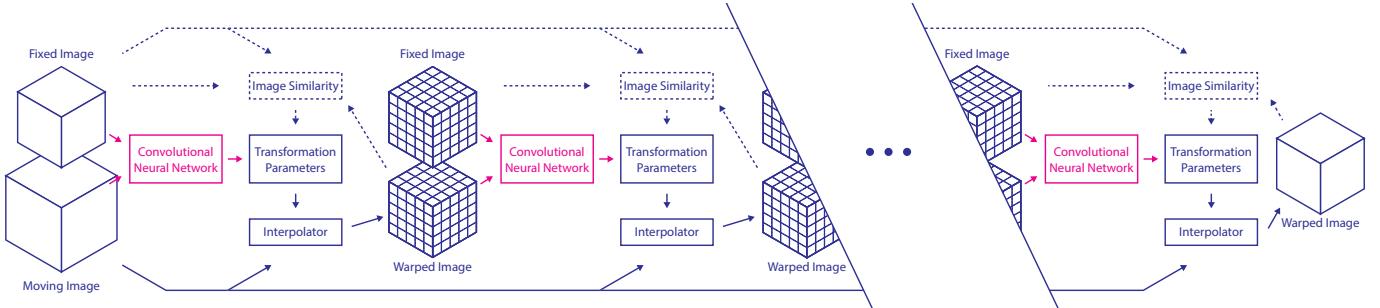


Figure 4: Schematic representation of the DLIR framework applied for hierarchical training of a multi-stage ConvNet for multi-resolution and multi-level image registration. The first stage performs affine registration of an image pair and the subsequent stages perform coarse-to-fine deformable image registration. The ConvNet in each stage is trained for its specific registration task by optimizing image similarity. The weights of the preceding ConvNets are fixed during training. This procedure prevents exploding gradients and conserves memory. Transformation parameters are passed through the network and combined at each stage to create a warped image. The warped image is passed to the subsequent stage and is used as the moving image input.

function we propose combines normalized cross correlation and this penalty:

$$L = L_{NCC} + \alpha P, \quad (1)$$

where  $L_{NCC}$  is the negative normalized cross correlation, and  $P$  the bending energy penalty with  $\alpha = 0$  for affine registration, and  $\alpha$  empirically determined to be 0.05 for all deformable image registration experiments. The bending energy penalty is defined as follows:

$$P = \frac{1}{V} \int_0^X \int_0^Y \int_0^Z \left[ \left( \frac{\partial^2 \mathbf{T}}{\partial x^2} \right)^2 + \left( \frac{\partial^2 \mathbf{T}}{\partial y^2} \right)^2 + \left( \frac{\partial^2 \mathbf{T}}{\partial z^2} \right)^2 + 2 \left( \frac{\partial^2 \mathbf{T}}{\partial xy} \right)^2 + 2 \left( \frac{\partial^2 \mathbf{T}}{\partial xz} \right)^2 + 2 \left( \frac{\partial^2 \mathbf{T}}{\partial yz} \right)^2 \right] dx dy dz,$$

where  $V$  is the volume of the image domain, and  $\mathbf{T}$  the local transformation. Adding this term during registration minimizes the second order derivatives of local transformations of a DVF, thereby resulting in locally affine transformations, thus enforcing global smoothness (Staring et al., 2007):

### 3. Data

Like most deep learning approaches, the DLIR framework requires large sets of training data. Publicly available datasets that are specifically provided to evaluate registration algorithms, contain insufficient training data for our approach. Therefore, we made use of large datasets of cardiac cine MRIs for intra-patient registration experiments, and low-dose chest CTs from the National Lung Screening Trial (NLST) for inter-patient registration experiments. We used manually delineated anatomical structures in these datasets for evaluation of the DLIR framework. Manually obtained delineations in the datasets were only used for final evaluation of registration performance. In addition, we used the publicly available DIR-Lab dataset. The data set is not sufficiently large to demonstrate the

full potential of the proposed method, but it does provide an indication of registration performance and it enables straightforward replication of our work.

#### 3.1. Cardiac Cine MRI

We included publicly available cardiac cine MRI scans from the Sunnybrook Cardiac Data (Radau et al., 2009). The data set contains 45 short-axis cine MRI images distributed over four pathology categories: healthy subjects, patients with hypertrophy, patients with heart failure and infarction, and patients with heart failure without infarction. Each scan contains 20 timepoints (i.e. volumes) encompassing the entire cardiac cycle, which results in  $45 \times 20$  volumes in total. All scans have a slice thickness and spacing of 8 mm and an in-plane resolution of 1.25 mm per voxel. All scans are made with a  $256 \times 256$  matrix and consist of about 10 slices. The data is separated into training, validation, and evaluation sets, each containing 15 scans with equally distributed pathology categories. Provided manual segmentations of left ventricle volumes at end-diastole (ED) and end-systole (ES) were used for evaluation.

#### 3.2. Chest CT

We included 2,060 chest CTs that were randomly selected from a set of scans acquired at baseline in the NLST (The National Lung Screening Trial Research Team, 2011). The dataset is very diverse containing scans of fourteen different CT-scanners from four vendors. All scans were made during inspiratory breath-hold without ECG synchronization and without contrast enhancement. Isotropic in-plane resolution of the  $512 \times 512$  axial slices varied between 0.45 mm to 0.98 mm per voxel. Slice increment ranged from 0.63 mm to 10.0 mm covering the thorax in 26 to 469 axial slices. The scans were divided into 2,000 scans for training and 50 scans for validation during method development. The remaining 10 scans provided 90 image pairs for quantitative evaluation. In each scan the entire visible aorta was delineated, including the ascending aorta,

the aortic arch, and the descending aorta. In addition, ten landmarks were annotated: the carina, the aortic root, the root of the left subclavian artery, the apex of the heart, the tip of the xiphoid, the tops of the left and right lungs, the left and right sterno clavicular joints, and the tip of the spinous process of the T1 vertebra.

### 3.3. DIR-Lab 4D Chest CT

We included publicly available 4D chest CT from DIR-Lab (Castillo et al., 2010, 2009). The dataset consists of ten 4D chest CTs that encompass a full breathing cycle in ten timepoints. Isotropic in-plane resolution of  $512 \times 512$  axial slices ranged from 0.97 mm to 1.98 mm per voxel, with a slice thickness and increment of 2.5 mm. Because the dataset is of limited size we did not separate it into separate training, validation, and test sets. Instead, we performed leave-one-out cross-validation during evaluation. Each scan contains 300 manually identified anatomical landmarks annotated in two timepoints, namely at maximum inspiration and maximum expiration. The landmarks serve as a reference to evaluate deformable image registration algorithms.

## 4. Evaluation

The DLIR framework was evaluated with intra-patient as well as inter-patient registration experiments. As image folding is anatomically implausible, especially in intra-patient image registration, after registration, we evaluated the topology of obtained DVFs quantitatively. For this we determined the Jacobian determinant—also known as *the Jacobian*—for every point  $p(i, j, k)$  in the DVF:

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

A Jacobian of 1 indicates that no volume change has occurred. A Jacobian of  $> 1$  indicates expansion, a Jacobian between 0 - 1 indicates shrinkage, and a Jacobian of  $\leq 0$  indicates a singularity: i.e. a place where folding has occurred. By indicating the fraction of foldings per image and by determining the standard deviation of the Jacobian, we can quantify the quality of the DVF.

Additionally, registration performance was evaluated using manually delineated anatomical structures and manually indicated landmarks. By propagating the delineations using obtained DVFs, registration performance can be assessed by measuring label overlap with the Dice coefficient:

$$Dice = \frac{2|P \cap R|}{|P| + |R|},$$

given a propagated segmentation ( $P$ ) and a reference segmentation ( $R$ ).

The surface distance

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

where  $x$  is a point of the propagated surface and  $y$  on the a reference surface ( $R_S$ ), was used to calculate the average symmetric surface distance (ASD)

$$ASD = \frac{1}{|P_S| + |R_S|} \left( \sum_{x \in P_S} d(x, R_S) + \sum_{y \in R_S} d(y, P_S) \right),$$

where  $x$  and  $y$  are points on the propagated surface  $P_S$  and reference surface  $R_S$ . And we calculated the symmetric Hausdorff distance:

$$HD = \max \{d_H(P_S, R_S), d_H(R_S, P_S)\},$$

where

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

For landmarks the registration error was determined as the average 3D Euclidean distance between transformed and reference points.

## 5. Implementation

### 5.1. DLIR Framework

All ConvNets were trained with the DLIR framework using the loss function provided in Section 2.4. The ConvNets were initialized with Glorot's uniform distribution (Glorot and Bengio, 2010) and optimized with Adam (Kingma and Ba, 2015).

Rectified linear units were used for activation in all ConvNets, except in the output layers. The output of the deformable ConvNets were unconstrained to enable prediction of negative B-spline displacement vectors. The outputs of affine ConvNets were constrained as follows: rotation parameters and shearing parameters were constrained between  $-\pi$  and  $+\pi$ , the scaling parameters were constrained between 0.5 and 1.5, and translation parameters were unconstrained.

During training, moving images were warped using linear resampling, during evaluation segmentations were warped using nearest neighbor resampling. All experiments were performed in Python using Pytorch (Paszke et al., 2017) on an NVIDIA Titan-X GPU, an Intel Xeon E5-1620 3.60 GHz CPU with 4 cores (8 threads), and 32 GB of internal memory.

### 5.2. Conventional Image Registration

Registration performance of the DLIR framework was compared with conventional iterative intensity-based image registration using SimpleElastix (Marstal et al., 2016). SimpleElastix enables integration of Elastix (Klein et al., 2010) in a variety of programming languages.

For optimal comparison, settings for conventional registration and DLIR experiments were chosen similar. Thus,

Table 1: Design of deformable image registration (DIR) stages used in single stage and multi-stage intra-patient registration of cardiac cine MRI. For multi-stage registration experiments DIR-1 and DIR-2 were sequentially applied; for single stage experiments one stage equal to DIR-2 was applied. Image resolution, grid spacing, and average number of grid points are given in  $x \times y \times z$  order.

	Single Stage	Multi-Stage	
	DIR	DIR-1	DIR-2
Image resolution (mm)	$1.25 \times 1.25 \times 8$	$2.50 \times 2.50 \times 16$	$1.25 \times 1.25 \times 8$
Grid spacing (mm)	$10 \times 10 \times 8$	$20 \times 20 \times 16$	$10 \times 10 \times 8$
Avg. grid points	$64 \times 64 \times 10$	$32 \times 32 \times 5$	$64 \times 64 \times 10$
Mini-batch size (pairs)	8	16	8

similar grid settings and NCC were used. Adaptive stochastic gradient descent was used for iterative optimization. Registration stages were optimized in 500 iterations, sampling 2,000 random points per iteration. In contrast to multi-stage DLIR experiments, a Gaussian smoothing image pyramid was used in favor of windowed averaging.

## 6. Intra-Patient Registration of Cardiac Cine MRI

Intra-patient registration experiments were conducted using cardiac cine MRIs. The task was to register volumes (i.e. 3D images) within the 4D scans. Experiments were performed with 3-fold cross-validation. In each fold 30 images were used for training and 15 for evaluation. Given that each scan has 20 timepoints, 11,400 different permutations of image pairs were available per fold for training. Performance was evaluated using registration between images at ED and ES by label propagation of manual left ventricle lumen segmentations. In total 90 different registration results were available for evaluation.

### 6.1. ConvNet Design and Training

To evaluate the impact of multi-stage image registration, ConvNets were trained for single stage and multi-stage deformable image registration. Initial global affine registration was not necessary, because cardiac cine MRI images only show local deformations between timepoints. Additionally, experiments were performed to study effect of the bending penalty.

Deformable registration ConvNets were designed as proposed in Section 2. Downsampling was performed using average pooling. To retain information of the through-plane axis, downsampling was applied in the short-axis plane only. Experimental settings are further detailed in Table 1.

All ConvNets were trained with mini-batches consisting of random permutations of two timepoints taken from the same image. Prior to analysis, image intensities were linearly scaled from 0 to 1 based on the minimum intensity and 99<sup>th</sup> percentile of the maximum intensity. During training fixed and moving image pairs were correspondingly augmented by random in-plane rotations of 90, 180, and 270 degrees and random in-plane cropping of at maximum  $\pm 16$  voxels. Registration stages were trained in 10,000 iterations. Each fold was trained in approximately

5 hours for single stage registration and 8 hours for multi-stage registration. Figure 5 shows the development of training and validation NCC between image pairs during training of one of the folds. Overfitting did not occur in the experiments, instead the training error was higher than the validation error due to the random croppings applied on the training set only.

## 6.2. Results

Figure 6 shows single stage image registration results of registering images at ES to ED. The obtained Jacobians show that the bending penalty mitigates image folding of the DLIR framework. Furthermore, quantitative analysis, as shown in Figure 7, reveals that the DLIR framework is not affected by image folding as much as conventional image registration. Nevertheless, even though nearly absent in the DLIR framework, image folding is further reduced by adding a bending penalty. On the other hand, multi-stage registration seems to have no effect on image folding in the DLIR framework, while having a large effect on folding outliers in conventional image registration. However, the label propagation results, shown in Figure 8, show that the DLIR framework also benefits from multi-stage image registration. It improved label overlap as indicated by the increased Dice and decreased ASD. The Hausdorff distance appears to be similar across experiments.

Figure 9 provides additional insight into registration performance of DLIR vs. conventional image registration. The spread shows that there is no correlation between frameworks with respect to registration results of image pairs; some image pairs were well aligned with DLIR and poorly with the conventional approach, and vice versa.

Table 2 provides an overview of all results. Statistical analyses with the Wilcoxon signed rank test indicated that the multi-stage DLIR with bending penalty had significantly less folding and lower standard deviation of the Jacobians ( $p \ll 0.0001$ ) compared to other methods. Dice and ASD were as high as conventional image registration and significantly better compared to single stage experiments. Interestingly, the multi-stage DLIR is approximately 350 times faster than the single stage conventional image registration experiments, and takes only 39 ms for multi-stage image registration, including intermediate and final image resampling.

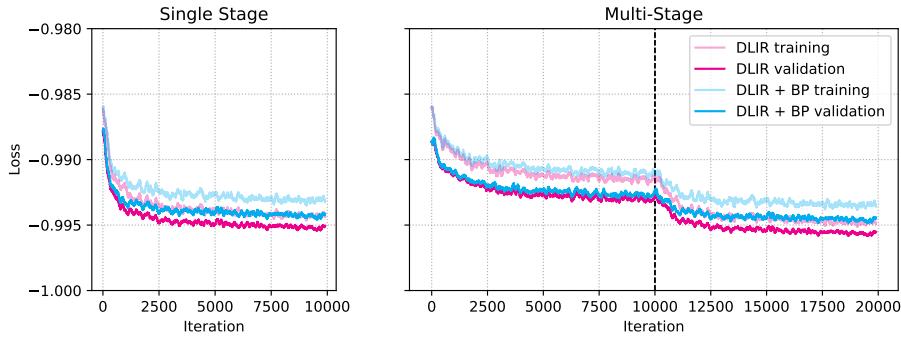


Figure 5: Learning curves showing the negative NCC during training of single stage and multi-stage ConvNets with or without bending penalty (BP) for intra-patient registration of cardiac cine MRI. Learning curves are taken from the one of the folds used in 3-fold cross validation. Augmentations were only applied to training data resulting in a relatively higher training NCC loss.

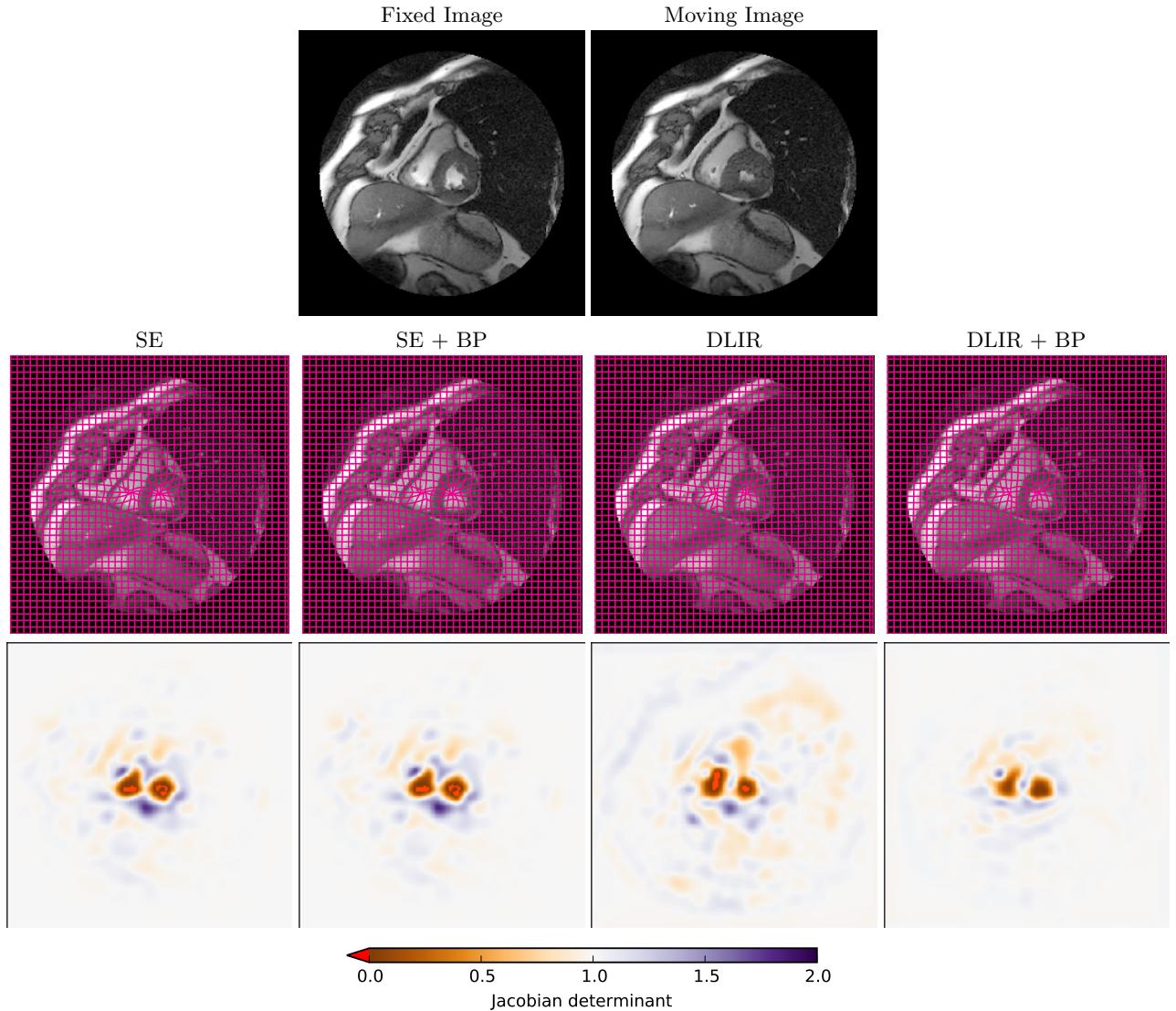


Figure 6: Top row: Cardiac cine MRI of a patient with left ventricular hypertrophy. Center axial slices are taken from the end-diastolic time point (Fixed) and end-systolic (Moving) timepoints. For visualization purposes fixed and moving images are cropped to the heart. Middle row: Registration results with superpositioned deformation grids. Bottom row: Colormap of the Jacobian with singularities (folding) indicated in bright red. From left to right results are shown for SimpleElastix (SE) and DLIR, with and without the bending penalty (BP).

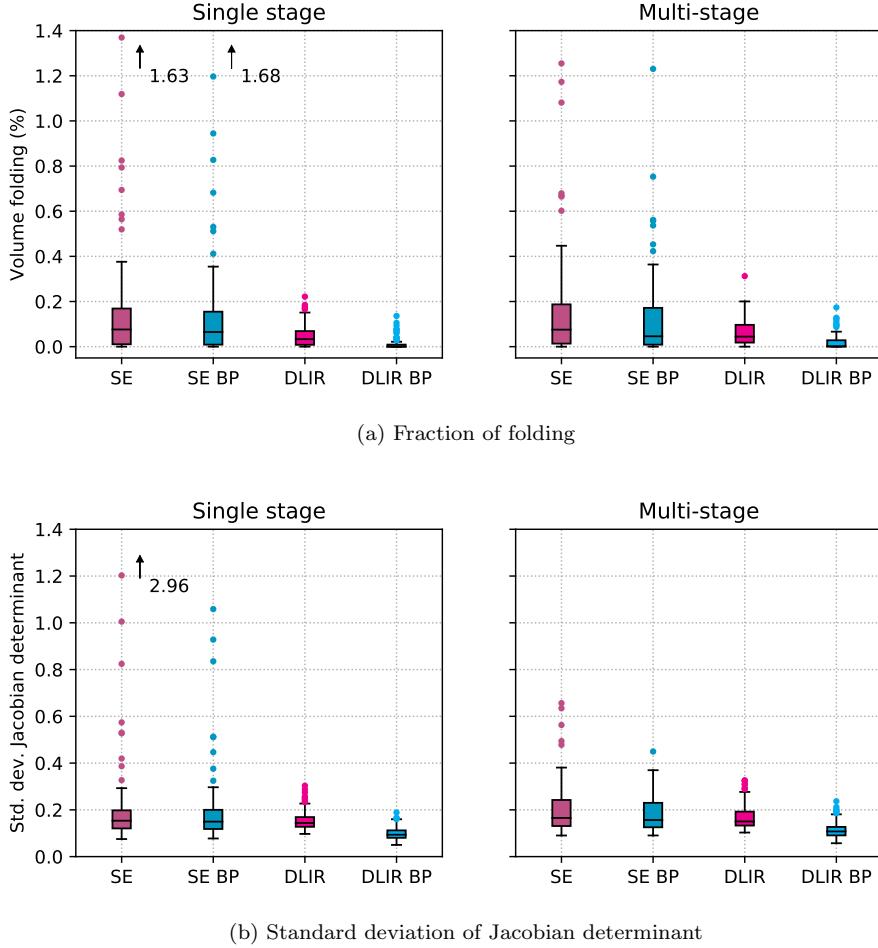
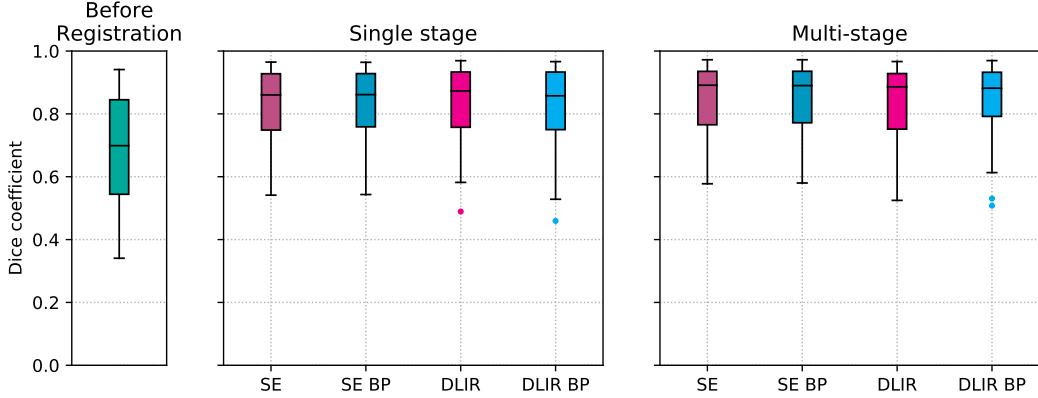


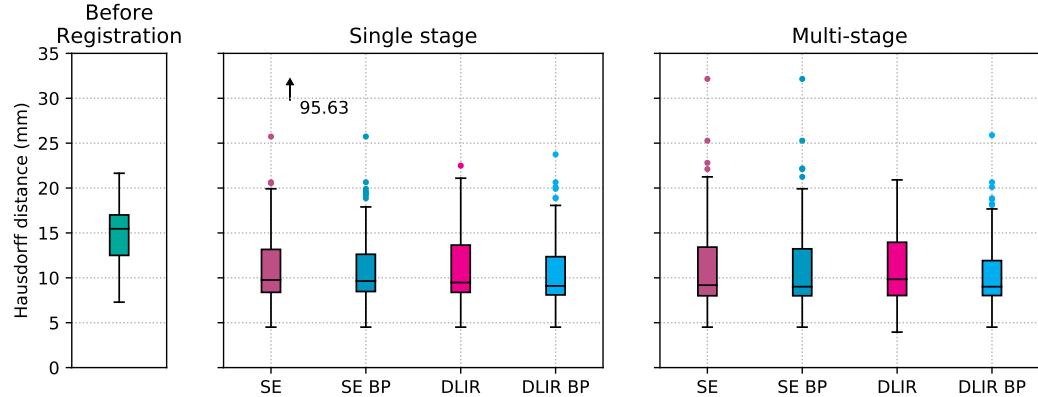
Figure 7: Boxplots showing (a) the volume of singularities and (b) the standard deviation of the Jacobian determinants to evaluate the topology of the DVFs obtained from registration experiments between end-diastole and end-systole cardiac cine MRI. Conventional registration experiments were performed using SimpleElastix (SE) and compared with DLIR registration. Both SE and DLIR experiments were conducted with and without the bending penalty (BP). The necessity of using a mask for conventional registration is illustrated by the results shown in the single stage experiments. For visualization purposes large outliers are indicated with an arrow with their values annotated.

Table 2: Table listing the results of cardiac cine MRI registration experiments. Single stage and multi-stage conventional and DLIR registration are compared with and without bending penalties (BP). Given that the results are not following a normal distribution, median  $\pm$  interquartile ranges are provided. Execution times are provided as mean (standard deviation). Note that the bending penalty is only applied to the DLIR framework during training, thus during application it does not limit execution time.

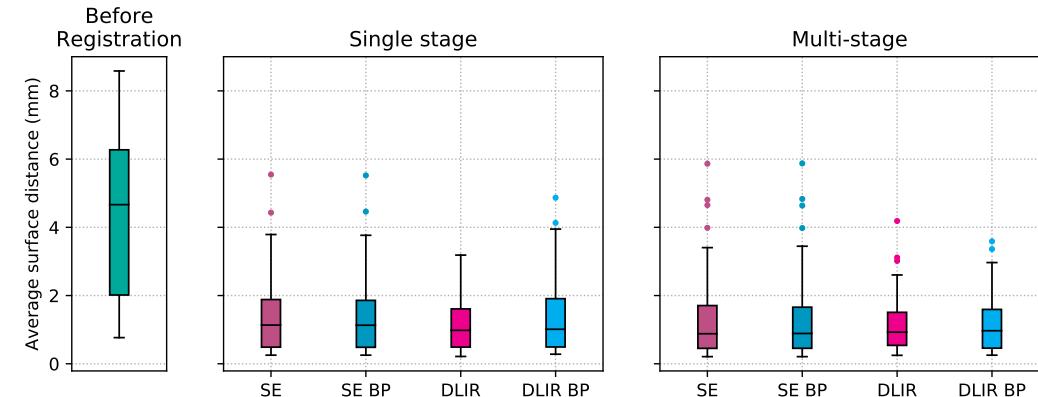
	Dice	HD	ASD	Fraction folding	Std. dev. Jacobian	CPU time (s)	GPU time (s)
Before registration	$0.70 \pm 0.30$	$15.46 \pm 4.50$	$4.66 \pm 4.26$	—	—	—	—
Single stage	SE	$0.86 \pm 0.18$	$9.76 \pm 4.78$	$1.14 \pm 1.40$	$0.08 \pm 0.16$	$0.15 \pm 0.08$	$13.49(3.27)$
	SE + BP	$0.86 \pm 0.17$	$9.64 \pm 4.15$	$1.13 \pm 1.38$	$0.07 \pm 0.15$	$0.15 \pm 0.08$	$14.89(3.07)$
	DLIR	$0.87 \pm 0.18$	$9.47 \pm 5.26$	$0.98 \pm 1.12$	$0.03 \pm 0.06$	$0.14 \pm 0.04$	$1.71(0.45)$
	DLIR + BP	$0.86 \pm 0.18$	$9.10 \pm 4.26$	$1.01 \pm 1.42$	$0.00 \pm 0.01$	$0.09 \pm 0.03$	$0.03 \pm 0.01$
Multi-stage	SE	$0.89 \pm 0.17$	$9.18 \pm 5.42$	$0.88 \pm 1.25$	$0.08 \pm 0.17$	$0.17 \pm 0.11$	$15.51(3.67)$
	SE + BP	$0.89 \pm 0.16$	$9.01 \pm 5.23$	$0.89 \pm 1.21$	$0.05 \pm 0.16$	$0.16 \pm 0.11$	$20.06(3.68)$
	DLIR	$0.89 \pm 0.18$	$9.84 \pm 5.93$	$0.93 \pm 0.97$	$0.05 \pm 0.08$	$0.15 \pm 0.06$	$2.35(0.60)$
	DLIR + BP	$0.88 \pm 0.14$	$9.01 \pm 3.89$	$0.97 \pm 1.14$	$0.002 \pm 0.03$	$0.11 \pm 0.04$	$0.04(0.01)$



(a) Dice coefficient



(b) Hausdorff distance



(c) Average symmetric surface distance

Figure 8: Label propagation results of manual left ventricle lumen annotations of intra-patient cardiac cine MRI registration. Boxplots of (a) Dice, (b) Hausdorff distance, and (c) average surface distance are shown for conventional image registration with SimpleElastix (SE) and the DLIR framework. Single stage and multi-stage registration experiments were performed for conventional registration and DLIR with and without the bending penalty (BP). The large outlier in (b) was indicated with an arrow to improve visualization.

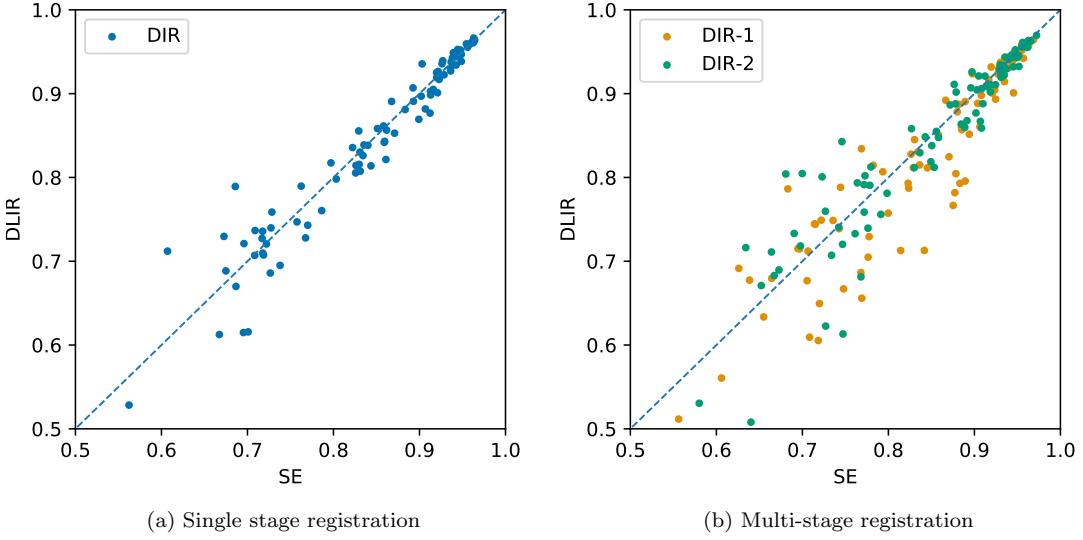


Figure 9: Scatter plots showing a comparison of Dice scores obtained with the DLIR framework and conventional inter-patient cardiac cine MRI registration. The plots show a correlation, but the dispersion of the points indicate that the registration tasks are not equally difficult for the DLIR framework and conventional registration framework.

## 7. Inter-Patient registration of Low-Dose Chest CT

Inter-patient registration was performed with chest CT scans of different subjects from the NLST. In this set large variations in the field of view were present, which were caused by differences in scanning protocol and by the different CT-scanners that were used. Because of these variations, and the variations in anatomy among subjects, affine registration was necessary for initial alignment. Therefore, multi-stage image registration was performed with sequential affine and deformable image registration stages. The test-scans provided 90 registrations for evaluation. Manual delineation of the aorta and 10 landmarks were used to assess registration performance.

### 7.1. ConvNet Design and Training

Inter-patient chest-CT registration requires initial alignment of patient scans. Thus, we implemented a multi-stage ConvNet consisting of an affine registration stage, followed by coarse-to-fine deformable image registration. The full Hounsfield Unit range of CT numbers (-1000 to 3095) was used to rescale input image intensities from 0 to 1. Memory limitations imposed by hardware and software limited the deformable image registration to three stages and a final image resolution of 2 mm. In-plane slice sizes ranged from  $115 \times 115$  to  $250 \times 250$  voxels and the number of slices ranged from 109 to 210. All ConvNets were designed with 32 kernels per convolution layer, but downsampling was performed with strided convolutions with  $2 \times 2 \times 2$  downsampling and  $4 \times 4 \times 4$  kernels instead of the favorable average pooling (de Vos et al., 2017) to further limit memory consumption. The affine registration ConvNet was designed as shown in Figure 2, but

with three downsampling layers in the pipelines. The separate pipelines in the affine registration ConvNet allowed analysis of a fixed and a moving image having different dimensions. The affine ConvNet registers the moving image to the fixed image space. As a result, the fixed and moving pairs can be concatenated and used for subsequent deformable image registration ConvNets, which were designed as specified in Figure 3.

The multi-stage ConvNet was trained in mini-batches consisting of randomly selected image pairs. Given that the training set consisted of 2,000 scans, almost four million possible permutations of image pairs were available for training. Not all permutations were seen during training, but on average each scan was analyzed 674 times. Additionally, random augmentations were performed by randomly cropping 32 mm in any direction. The multi-stage ConvNet was trained in 18 hours using the settings listed in Table 3. The loss curves shown in Figure 10 show no signs of overfitting. The third and fourth stages analyze higher resolution images and output finer B-spline grids. As a consequence the dissimilarity increases and the finer deformations increase the bending penalty, resulting in higher starting losses, compared to previous registration stages.

### 7.2. Results

Ten images with manually segmented aortas resulted in 90 permutations of fixed and moving image pairs that were used for evaluation. Figure 11 shows that the affine stage correctly aligns two images from the evaluation set. The coarse-to-fine deformable stages gradually improves upon this alignment. However, final DVFs obtained in

Table 3: Experimental settings of the DLIR framework for training a multi-stage ConvNet for inter-patient registration of chest CT. The ConvNet consists of an affine image registration (AIR) stage, and three deformable image registration (DIR) stages. Image resolution, grid spacing, and average number of grid points are given in  $x \times y \times z$  order.

Stage	AIR	DIR-1	DIR-2	DIR-3
Input image resolution (mm)	$8 \times 8 \times 8$	$8 \times 8 \times 8$	$4 \times 4 \times 4$	$2 \times 2 \times 2$
Grid spacing (mm)	–	$64 \times 64 \times 64$	$32 \times 32 \times 32$	$16 \times 16 \times 16$
Avg. grid points	–	$5 \times 5 \times 5$	$11 \times 11 \times 10$	$21 \times 21 \times 20$
Mini-batch size (pairs)	16	8	4	2

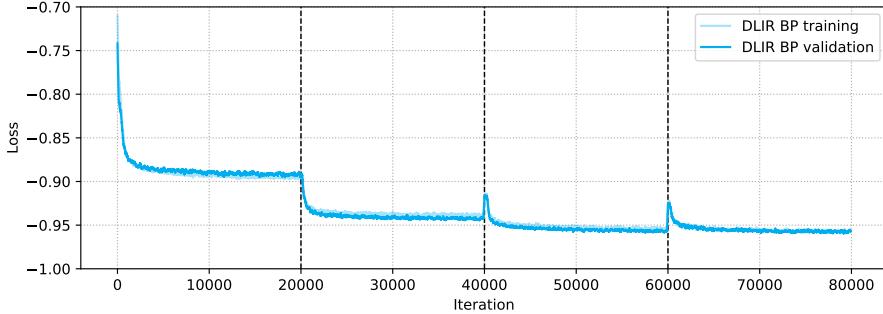


Figure 10: Learning curves during sequential training of the four registration stages of a ConvNet for inter-patient registration of chest CT.

these experiments show some folding, as is visualized in the examples of Figure 12.

Quantitative analysis, shown in Figure 13, of the deformable registration DVF reveals that only the final registration stage is hampered by folding. While in conventional image registration folding gradually increases with each deformable registration stage, the DLIR framework shows zero to limited folding in the first two stages and a large increase in the final stage. A similar pattern is seen in the standard deviations of the Jacobians. The Wilcoxon signed-rank test indicated that for each stage the results were significantly different between conventional image registration and the DLIR framework.

Figure 14 shows that Dice and ASD are similar for affine registration with conventional image registration and DLIR. The first two deformable registration stages have slightly lower Dice and higher ASD. In contrast, HD is lower for DLIR, meaning that segmentations registered with DLIR have less deviations from the reference than segmentations registered with conventional image registration. In the last stage, registration performance is similar for DLIR and conventional registration, but DLIR has less outliers. The Wilcoxon signed-rank test indicated that for the final registration stage, Dice and ASD were not significantly different between conventional registration and DLIR.

Table 4 gives an overview of all registration results and execution times. It shows that registration with the DLIR framework achieves quick registrations. Including image resampling, registration was taken approximately 0.43 s per image pair on a GPU.

Figure 15, shows that a correlation between conventional image registration and DLIR registration with respect to registration quality of image pairs. However, some

registrations are more difficult for conventional image registration, while being correctly performed with DLIR, and vice versa.

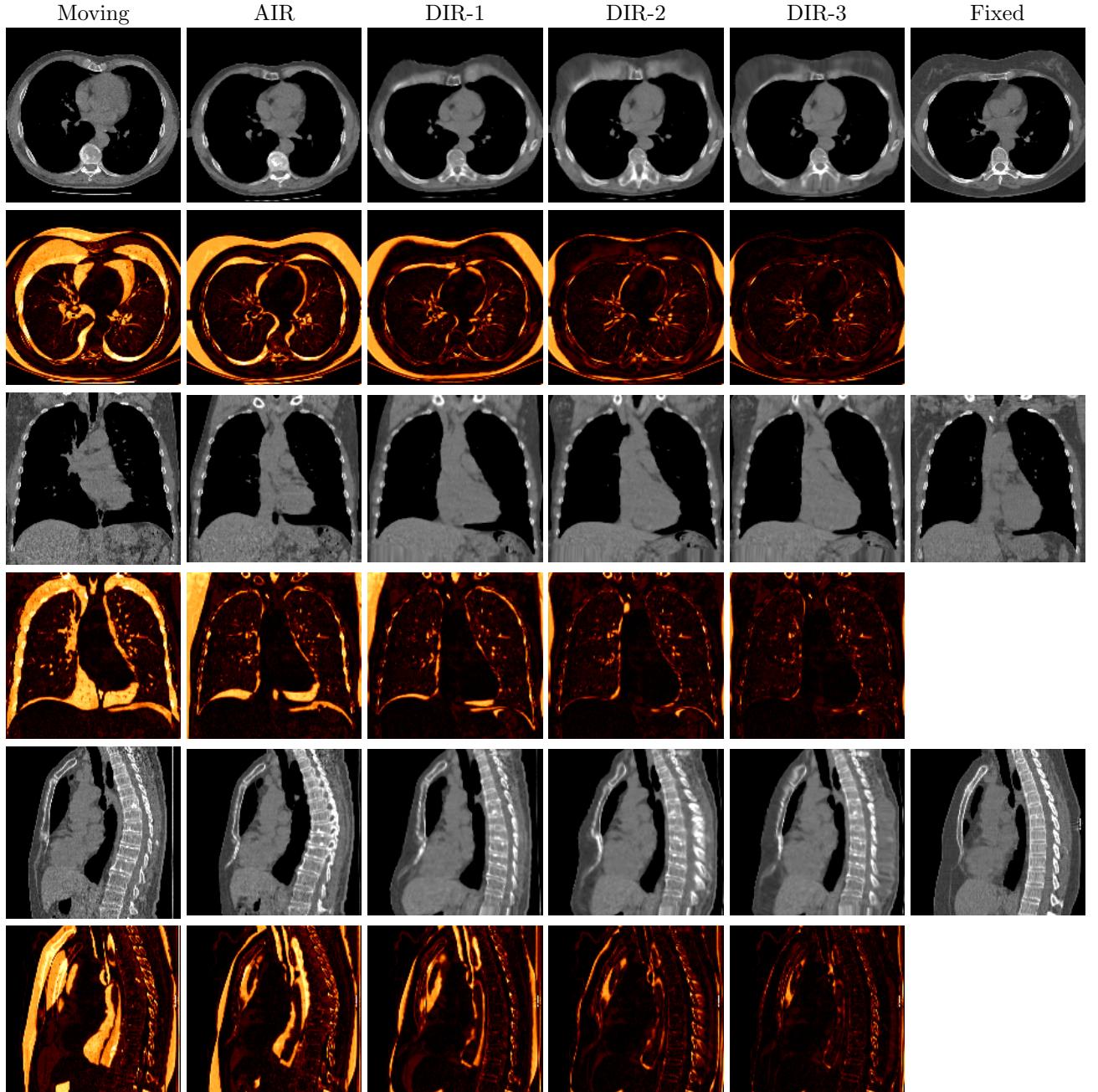


Figure 11: Example results of inter-patient registration of two chest CTs from the NLST test set. The moving image is shown on the left and the target fixed image is shown on the right. Intermediate registration results for each stage are shown in between. The rows show center slices of resp. axial, coronal, and sagittal planes, with in between corresponding heatmaps of the absolute difference with respect to the fixed image. This qualitatively shows increasing alignment at each registration stage. The full scale of Hounsfield units cannot be visualized. Window and level is set to visualize the aorta. As a consequence, complexity of the lungs is not visible in this example.

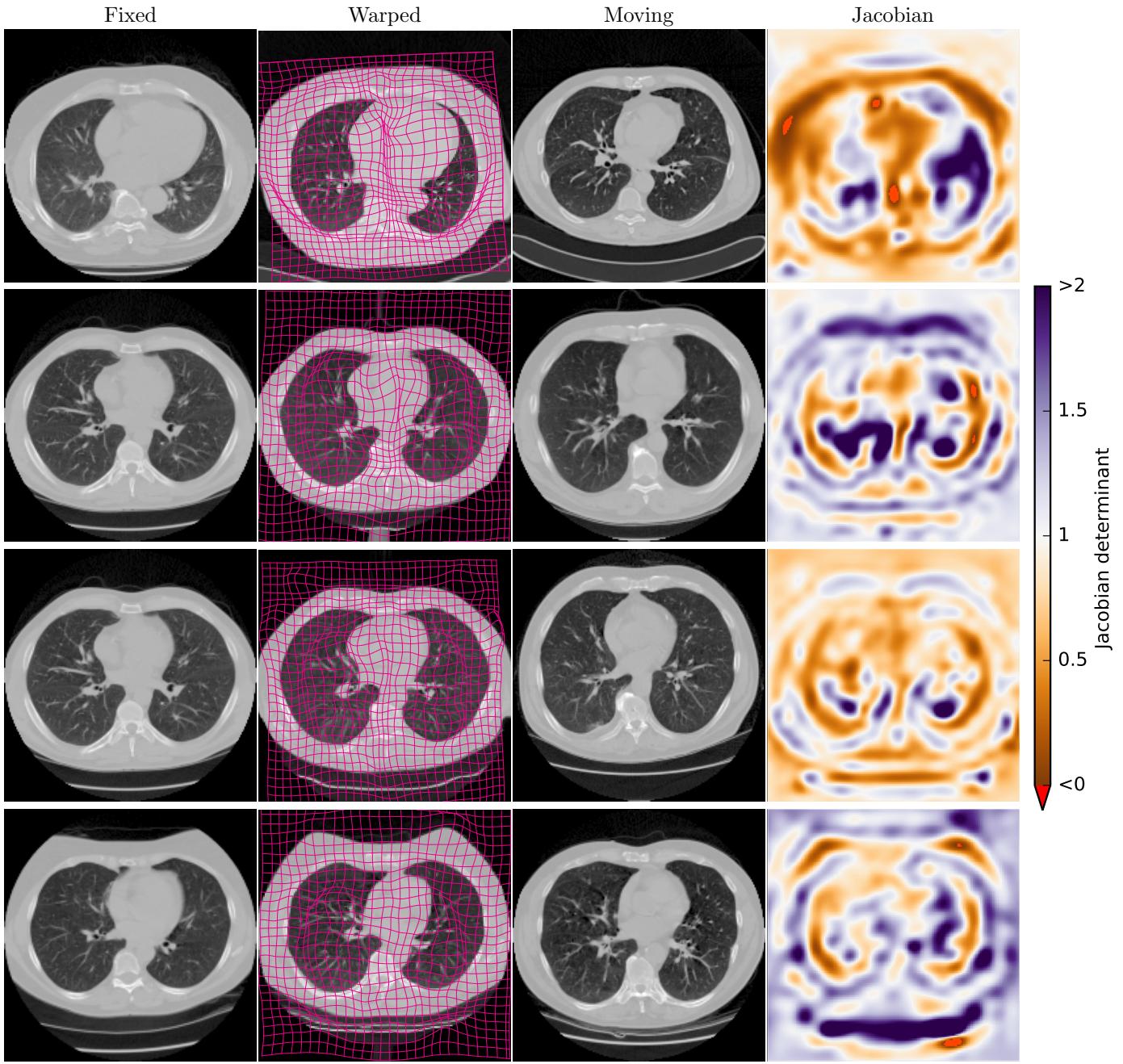


Figure 12: The rows show four inter patient chest-CT registration results. The columns show fixed images, warped images with a deformation grid, moving images, and a colormap of the Jacobian with singularities (folding) indicated in bright red.

Table 4: Results of the inter patient chest-CT registration experiments. DLIR is compared with conventional image registration using SimpleElastix. Results are given for all stages as median  $\pm$  interquartile range. Execution times are presented as mean (standard deviation) in seconds.

	Dice	HD	ASD	Fraction folding (%)	Std. dev. Jacobian	CPU time (s)	GPU time (s)
Before registration	$0.31 \pm 0.21$	$32.62 \pm 12.21$	$9.21 \pm 4.53$	—	—	—	—
SE	AIR	$0.60 \pm 0.19$	$25.81 \pm 15.34$	$4.89 \pm 2.36$	—	—	$3.73(0.26)$
	DIR-1	$0.69 \pm 0.11$	$20.30 \pm 13.26$	$3.39 \pm 1.11$	$0.00 \pm 0.00$	$0.19 \pm 0.11$	$11.67(1.07)$
	DIR-2	$0.75 \pm 0.08$	$21.26 \pm 11.31$	$2.67 \pm 0.87$	$0.00 \pm 0.08$	$0.27 \pm 0.13$	$14.83(3.37)$
	DIR-3	$0.77 \pm 0.08$	$20.83 \pm 11.81$	$2.45 \pm 0.89$	$0.04 \pm 0.19$	$0.30 \pm 0.15$	$20.36(8.41)$
DLIR	AIR	$0.58 \pm 0.16$	$26.79 \pm 13.05$	$5.24 \pm 2.19$	—	—	$1.02(0.29)$
	DIR-1	$0.64 \pm 0.11$	$21.68 \pm 13.09$	$3.86 \pm 1.74$	$0.00 \pm 0.00$	$0.16 \pm 0.09$	$3.85(0.99)$
	DIR-2	$0.70 \pm 0.10$	$19.95 \pm 13.30$	$3.21 \pm 1.15$	$0.00 \pm 0.00$	$0.19 \pm 0.10$	$8.18(2.03)$
	DIR-3	$0.75 \pm 0.08$	$19.34 \pm 13.41$	$2.46 \pm 0.80$	$0.75 \pm 1.08$	$0.45 \pm 0.21$	$15.41(4.38)$

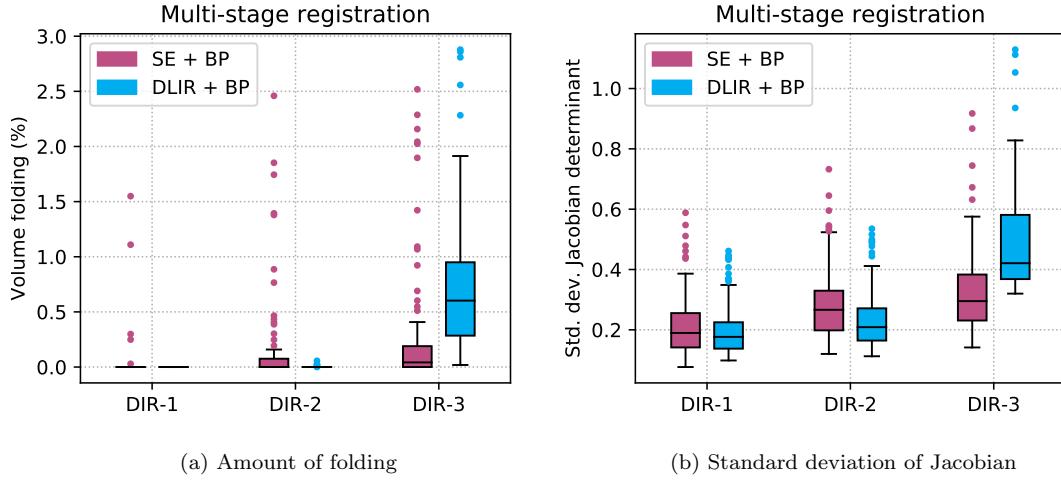


Figure 13: Boxplots showing in (a) the volume fraction of folding and in (b) the standard deviation of the Jacobian determinants of the deformable stages of inter-patient chest CT registration. Conventional registration experiments were performed using SimpleElastix (SE) and compared with DLIR registration both using a bending penalty (BP).

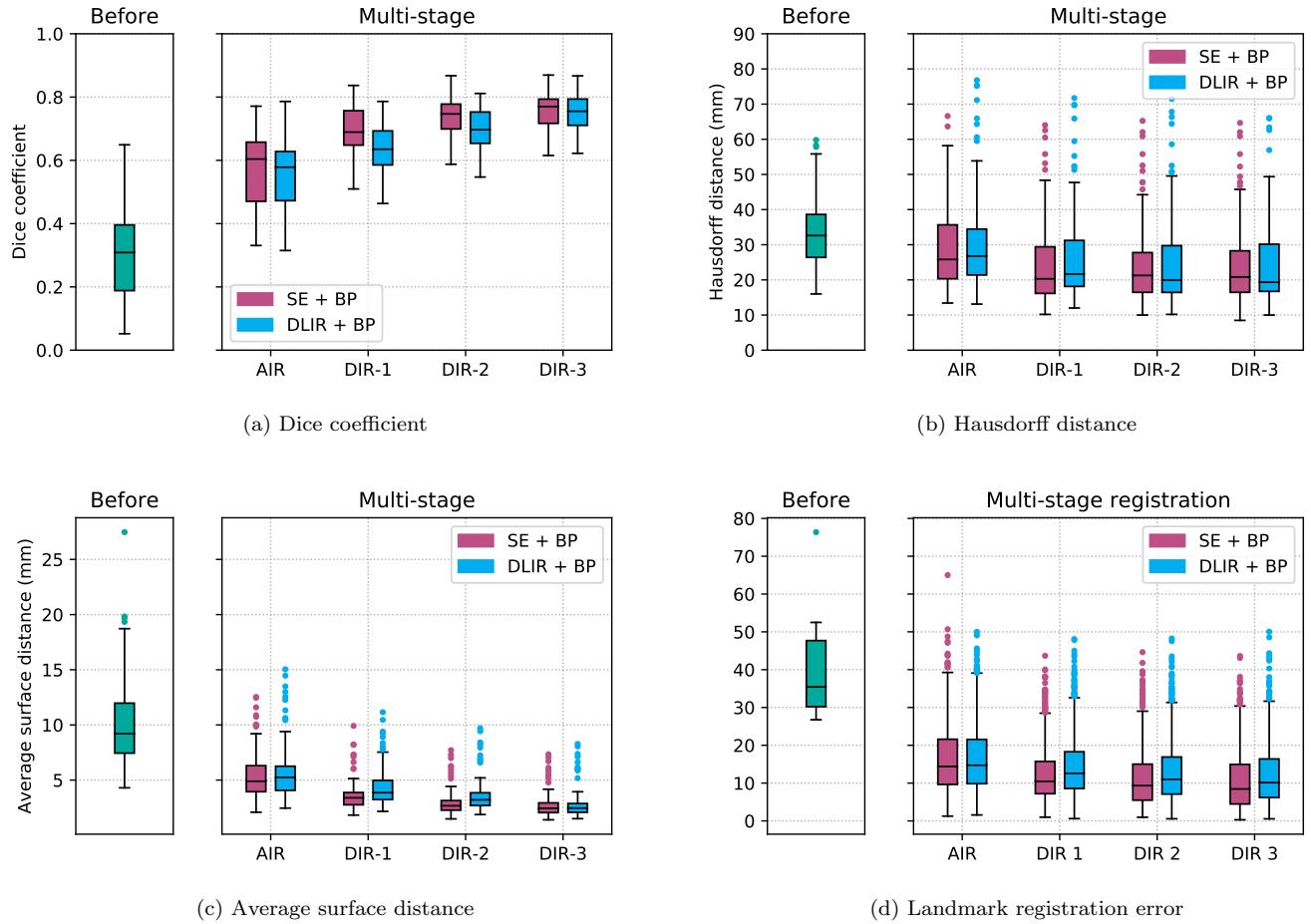


Figure 14: Label propagation results of manual aorta delineations of inter-patient chest CT registration. Boxplots of (a) Dice, (b) Hausdorff distance, (c) average surface distance, and (d) landmark registration error are shown for conventional image registration with SimpleElastix (SE) and the DLIR framework. Left boxplots: results before image registration. Right boxplots: results of multi-stage image registration.

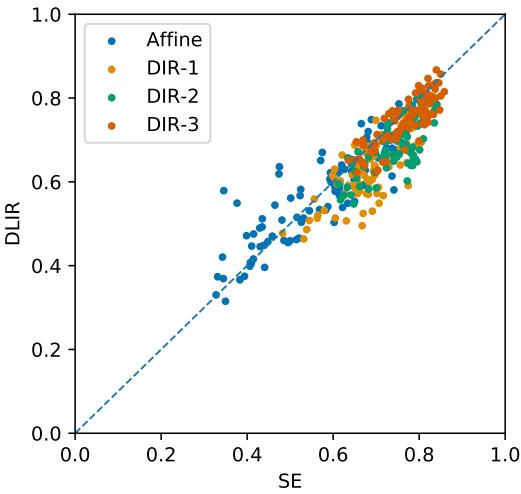


Figure 15: Scatter plot showing a comparison between Dice scores obtained with the DLIR framework and conventional intra-patient registration of chest CT. The plots show a correlation, but the dispersion of the points indicates that the registration tasks are not equally difficult for the DLIR framework and conventional registration framework.

## 8. Intra-Patient Registration of 4D Chest CT

Current registration benchmark datasets unfortunately do not provide sufficient scans to train a ConvNet using the DLIR framework. Nevertheless, to give further insight in the method’s performance and especially to enable reproducing our results, we performed experiments using the publicly available DIR-Lab data. The used dataset consists of ten 4D chest CTs that encompass a full breathing cycle in 10 timepoints. For each scan, 300 manually identified anatomical landmarks in the lungs in two timepoints—at maximum inspiration and maximum expiration—are provided. The landmarks serve as a reference for evaluating deformable image registration algorithms.

### 8.1. ConvNet Design and Training

Because the number of scans is very limited, we performed a leave-one-out cross-validation experiments, where one scan was used for evaluation and the nine remaining scans were used for training. The dataset size was too limited to train a ConvNet for affine registration, thus only ConvNets for deformable image registration were trained. Image intensities were clamped between -1000 and -200 HU and scaled between 0 and 1. This allowed the ConvNet to mainly focus on the anatomy of the lungs. ConvNets were trained for intra-patient registration by taking random timepoints per patient as fixed and moving images. This resulted in only 810 fixed and moving image permutations that were available for training. Ten ConvNets of similar design as used in inter-patient chest CT registration were trained in 15 hours each. Detailed experimental settings are provided in Table 5. The ConvNets were

Table 5: Experimental settings of the DLIR framework for training a multi-stage ConvNet for intra-patient registration of 4D chest CT from DIR-Lab data. The ConvNet consists of four deformable image registration (DIR) stages.

Stage	DIR-1	DIR-2	DIR-3
Input image resolution (mm)	$4 \times 4 \times 5$	$2 \times 2 \times 2.5$	$1 \times 1 \times 2.5$
Grid spacing (mm)	$32 \times 32 \times 40$	$16 \times 16 \times 20$	$8 \times 8 \times 10$
Mini-batch size (pairs)	8	4	2

trained by taking random spatially corresponding image patches of  $128 \times 128 \times 64$  voxels from fixed and moving image pairs to limit memory consumption. Nevertheless, during testing, scans six to ten had to be cropped to the chest to further limit memory consumption.

### 8.2. Results

The results are listed in Table 6, which also shows results of conventional image registration method based on Elastix (Berendsen et al., 2014) and a supervised deep learning based method (Eppenhof et al., 2018). The final average registration error was 2.64 mm with a standard deviation of 4.32. The error is highly influenced by outliers, likely caused by the limited dataset size. Large initial landmark distances were scarcely available for training, which influenced registration performance, as illustrated in Figure 16. By removing 10% of the landmarks with the largest initial registration error more than 17.7 mm—of which 1.47% is coming from scan 8—an adjusted registration error is obtained of 1.63 mm with a standard deviation of 1.67. The average registration time was 0.63 s for multi-stage image registration, including intermediate and final image resampling.

## 9. Discussion

We have presented a new framework for unsupervised training of ConvNets for 3D image registration: the Deep Learning Image Registration (DLIR) framework. The DLIR framework exploits image similarity between fixed and moving image pairs to train a ConvNet for image registration. Labeled training data, in the form of example registrations, are not required. The DLIR framework can train ConvNets for hierarchical multi-resolution and multi-level image registration and it can achieve accurate registration results.

Essentially the DLIR-framework can be viewed as an unsupervised training framework for STNs. The DLIR framework shares many elements with a conventional image registration framework, as is shown in Figure 1. In both frameworks pairs of fixed and moving images are registered by predicting transformation parameters. In both frameworks transformation parameters are inputs for a transformation model that warps the moving image. In both frameworks image similarity between the fixed and the warped moving image is used to improve transformation parameter prediction. However, while a conventional

Table 6: Mean (standard deviation) of the registration error in mm determined on DIR-Lab 4D-CT data. From left to right: initial landmark distances (i.e. prior to registration), results of conventional image registration (Berendsen et al., 2014), results of supervised deep learning method (Eppenhof et al., 2018), and registration results our proposed multi-stage DLIR. Individual registration results are shown for all ten scans. We refer the reader to <https://www.dir-lab.com/Results.html> for a list providing results of other registration methods.

Scan	Initial	Berendsen et al. (2014)	Eppenhof et al. (2018)	DLIR Stage1	Stage2	Stage3
Case 1	3.89(2.78)	1.00(0.52)	1.65(0.89)	2.34(1.76)	1.72(1.37)	1.27(1.16)
Case 2	4.34(3.90)	1.02(0.57)	2.26(1.16)	2.28(1.52)	1.61(1.31)	1.20(1.12)
Case 3	6.94(4.05)	1.14(0.89)	3.15(1.63)	3.89(1.77)	2.32(1.58)	1.48(1.26)
Case 4	9.83(4.85)	1.46(0.96)	4.24(2.69)	3.78(1.95)	2.49(1.90)	2.09(1.93)
Case 5	7.48(5.50)	1.61(1.48)	3.52(2.23)	3.51(2.28)	2.66(2.15)	1.95(2.10)
Case 6	10.89(6.96)	1.42(1.71)	3.19(1.50)	7.58(6.46)	6.04(6.64)	5.16(7.09)
Case 7	11.03(7.42)	1.49(1.06)	4.25(2.08)	5.05(2.36)	3.90(2.46)	3.05(3.01)
Case 8	14.99(9.00)	1.62(1.71)	9.03(5.08)	8.57(3.55)	6.99(4.52)	6.48(5.37)
Case 9	7.92(3.97)	1.30(0.76)	3.85(1.86)	6.12(2.79)	3.51(2.02)	2.10(1.66)
Case 10	7.30(6.34)	1.50(1.31)	5.07(2.31)	3.76(2.36)	2.85(2.11)	2.09(2.24)
Total	8.46(6.58)	1.36(1.01)	4.02(3.08)	5.12(4.64)	3.40(4.17)	2.64(4.32)

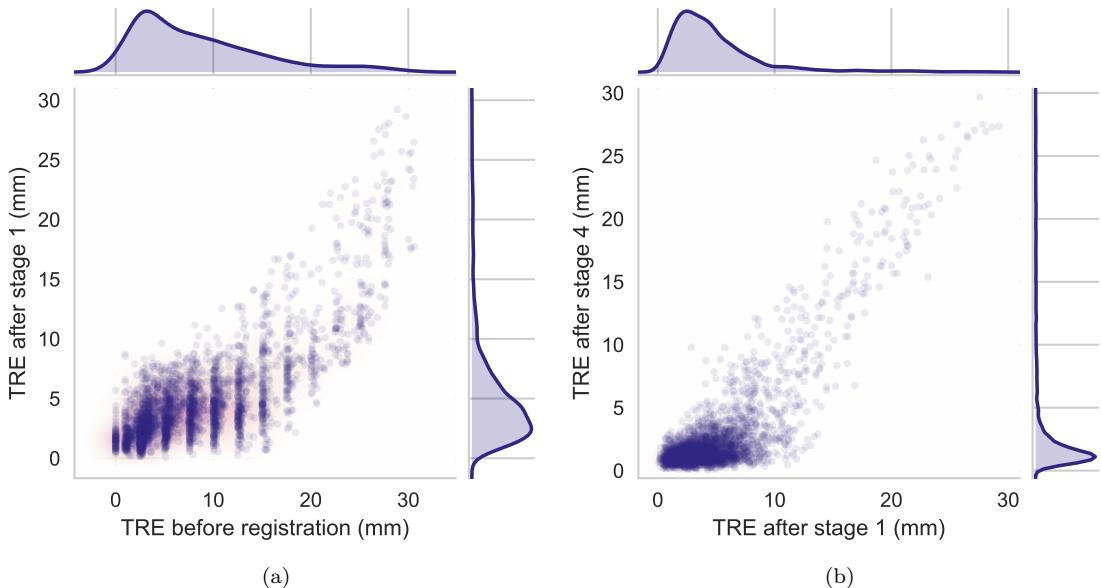


Figure 16: Scatterplots with joint histograms illustrate that large initial deformations are underrepresented in the DIRLab dataset and that the ConvNet is unable to correctly align those with the first registration stage as shown in (a). As a consequence the ConvNet was unable to correct this in later stages as shown in (b). Nevertheless, the majority of landmarks were registered adequately by the ConvNet.

image registration framework is always used during application, the DLIR framework is only used during training of a ConvNet for image registration. After training the ConvNet can be applied for one-shot image registration of unseen images. The DLIR framework allows unsupervised training of ConvNets for affine and deformable image registration. By combining multiple ConvNets, each with its own registration task, a multi-stage ConvNet can be made that is able to perform complex registration tasks like inter-patient image registration.

In this study three multi-stage ConvNets were trained within the DLIR framework for intra-patient registration of cardiac cine MRI, for inter-patient registration of chest CT, and for intra-patient registration of 4D chest CT. In all registration experiments the method showed registration results that are similar to conventional image registration but within exceptionally short execution times, which is especially desirable in time-critical applications.

The DLIR framework matched registration performance of the conventional method in intra-patient cardiac MR registration. Even though evaluation was performed with image pairs having maximum deformation between them, because evaluation pairs were taken from ES and ED time-points; while training was performed with image pairs having limited deformation between them, because training pairs were randomly taken from the full cardiac phase. Results would likely improve by training a ConvNet with a representative data-set of larger deformations, e.g. more image pairs taken from ES and ED timepoints. However, to accomplish this, the number of training scans should be substantially increased. Likewise conclusions can be drawn for 4D chest CT registration experiments with DIR-Lab data. Performance would likely be improved when using a larger data-set with representative training data. Nevertheless, even with this very limited training set size, adequate registration results were obtained within 0.63 s.

In inter-patient registration experiments the DLIR framework had a similar performance as the conventional image registration method in the first stages, most notably in affine registration. However the DLIR framework was slightly outperformed by the conventional method at later stages. This might (partially) be caused suboptimal ConvNet design choices imposed by memory limitations, e.g. the use of strided convolutions for downsampling. Nonetheless, the DLIR framework achieves accurate registration results with limited outliers while performing registrations faster than the conventional iterative method.

Performance of the DLIR framework is highly related to the interplay between the number of training image (or patch) pairs and registration problem complexity. For deformable registration ConvNets training is patch-based, while for affine registration ConvNets training is image based. Hence, deformable registration ConvNets allow extraction of multiple training samples from image pairs, while for affine registration ConvNets each image pair is one training sample. In intra-patient registration of cardiac MRI and inter-patient registration of chest CT the

balance between number of training samples and problem complexity was adequate to match performance of conventional image registration. As expected with the DIR-Lab experiments, conventional image registration outperformed DLIR. Most likely caused by the amount of available representative training data. The employed augmentations were insufficient, and large deformations were not corrected by DLIR. Possibly by adding more training data results will improve.

Addition of a bending energy penalty mitigated occurrence of folding. In conventional image registration the penalty is used during application and as consequence it increases execution time. In DLIR the penalty is applied only during training. While it increased memory consumption and therefore in our experiments limited the number of registration stages to three, it had no effect on execution time. Yet, like in conventional image registration, full elimination of folding is not guaranteed. Nevertheless, folding was within acceptable ranges. Additional regularization during training might enforce diffeomorphism (Staring et al., 2007), with no extra cost to execution time.

The last stage of DLIR was subject to increased amounts of folding. Possibly small misregistrations of preceding stages influenced later stages, which ultimately introduced singularities. Fine-tuning the full multi-stage DLIR pipeline end-to-end might reduce this. But, owing to memory limitations, imposed by hardware and software, end-to-end training of the multi-stage ConvNets was impossible. Instead, a hierarchical training approach was used where weights of the ConvNets from preceding stages were fixed. Fixing these weights during training drastically limited memory consumption, which enabled training of large multi-stage ConvNets. Furthermore, end-to-end training of a multi-stage ConvNet could prove to be difficult: exploding gradients hampered end-to-end training in preliminary experiments using highly downsampled data. In future work stringent regularization might allow full end-to-end training of large multi-stage ConvNets, when memory issues have been dealt with.

This work employed coarse-to-fine image registration experiments such that in each registration stage maximum deformations were within the capture range of the B-spline. Given, that ConvNets were designed such that the receptive fields coincided with the B-spline capture range, the receptive fields also captured maximum deformations. In future work it would be interesting to study how DLIR would behave when dealing with deformations that are outside the receptive field and how this would affect registration of areas of uniform intensity.

The DLIR framework is able to recast conventional intensity-based image registration into a learning problem. Thus, the framework can be extended with techniques from conventional image registration and deep learning. Features from conventional image registration, such as different transformation models like thin plate splines or direct DVF estimation can be readily implemented. Ad-

ditionally, different image similarity metrics could be implemented; while in the current paper DLIR was employed on same-modality MR and CT data, the framework readily supports multi-modality image registration, i.e. by replacing the similarity metric for mutual information (Pluim et al., 2003). In our experiments we have used a simple ConvNet design with a limited memory footprint to demonstrate feasibility of the proposed DLIR framework. More complex ConvNet designs could be used, but complex designs are often at the cost of memory. Nevertheless, a large range of designs could be implemented in the proposed framework. In future studies we will investigate impact of other conventional image registration and deep learning techniques on image registration robustness and accuracy.

## 10. Conclusion

We presented the Deep Learning Image Registration framework for unsupervised affine and deformable image registration with convolutional neural networks. We demonstrated that the DLIR framework is able train ConvNets without training examples for accurate affine and deformable image registration within very short execution times.

## Acknowledgment

This work is part of the research programme ImaGene with project number 12726, which is partly financed by the Netherlands Organisation for Scientific Research (NWO).

The authors thank the National Cancer Institute for access to NCI's data collected by the National Lung Screening Trial. The statements contained herein are solely those of the authors and do not represent or imply concurrence or endorsement by NCI.

## References

### References

- Beg, M. F., Miller, M. I., Trouvé, A., Younes, L., Feb 2005. Computing large deformation metric mappings via geodesic flows of diffeomorphisms. *International Journal of Computer Vision* 61 (2), 139–157.
- Berendsen, F. F., Kotte, A. N. T. J., Viergever, M. A., Pluim, J. P. W., 2014. Registration of organs with sliding interfaces and changing topologies. *Proc.SPIE* 9034, 9034 – 7.  
URL <https://doi.org/10.1117/12.2043447>
- Cao, X., Yang, J., Zhang, J., Nie, D., Kim, M., Wang, Q., Shen, D., 2017. Deformable image registration based on similarity-steered cnn regression. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D. L., Duchesne, S. (Eds.), *Medical Image Computing and Computer Assisted Intervention - MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11–13, 2017, Proceedings, Part I*. Springer International Publishing, Cham, pp. 300–308.
- Castillo, E., Castillo, R., Martinez, J., Shenoy, M., Guerrero, T., 2010. Four-dimensional deformable image registration using trajectory modeling. *Physics in Medicine & Biology* 55 (1), 305.
- Castillo, R., Castillo, E., Guerra, R., Johnson, V. E., McPhail, T., Garg, A. K., Guerrero, T., 2009. A framework for evaluation of deformable image registration spatial accuracy using large landmark point sets. *Physics in Medicine & Biology* 54 (7), 1849.
- Dauphin, Y. N., Pascanu, R., Gulcehre, C., Cho, K., Ganguli, S., Bengio, Y., 2014. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., Weinberger, K. Q. (Eds.), *Advances in Neural Information Processing Systems 27*. Curran Associates, Inc., pp. 2933–2941.
- de Vos, B. D., Berendsen, F. F., Viergever, M. A., Staring, M., Işgum, I., 2017. End-to-end unsupervised deformable image registration with a convolutional neural network. In: *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support : Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings*. Springer International Publishing, Cham, pp. 204–212.
- Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., van der Smagt, P., Cremers, D., Brox, T., December 2015. Flownet: Learning optical flow with convolutional networks. In: *The IEEE International Conference on Computer Vision (ICCV)*.
- Eppenhof, K., Lafarge, M., Moeskops, P., Veta, M., Pluim, J., 2018. Deformable image registration using convolutional neural networks. *Proc.SPIE* 10133, 10133 – 6.
- Garg, R., Kumar, B. G. V., Carneiro, G., Reid, I. D., 2016. Unsupervised CNN for single view depth estimation: Geometry to the rescue. In: *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VIII*. pp. 740–756.
- Glorot, X., Bengio, Y., 2010. Understanding the difficulty of training deep feedforward neural networks. In: Teh, Y. W., Titterington, M. (Eds.), *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. Vol. 9 of *Proceedings of Machine Learning Research*. PMLR, Chia Laguna Resort, Sardinia, Italy, pp. 249–256.
- Hu, Y., Modat, M., Gibson, E., Ghavami, N., Bonmati, E., Moore, C. M., Emberton, M., Noble, J. A., Barratt, D. C., Vercauteren, T., April 2018a. Label-driven weakly-supervised learning for multimodal deformable image registration. In: *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. pp. 1070–1074.
- Hu, Y., Modat, M., Gibson, E., Li, W., Ghavami, N., Bonmati, E., Wang, G., Bandula, S., Moore, C. M., Emberton, M., Ourselin, S., Noble, J. A., Barratt, D. C., Vercauteren, T., oct 2018b. Weakly-supervised convolutional neural networks for multimodal image registration. *Medical Image Analysis* 49, 1–13.
- Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., Brox, T., Jul 2017. Flownet 2.0: Evolution of optical flow estimation with deep networks. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.  
URL <http://1mb.informatik.uni-freiburg.de//Publications/2017/IMKDB17>
- Jaderberg, M., Simonyan, K., Zisserman, A., Kavukcuoglu, K., 2015. Spatial transformer networks. In: Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., Garnett, R. (Eds.), *Adv Neural Inf Process Syst 28*. Curran Associates, Inc., pp. 2017–2025.
- Kingma, D., Ba, J., 2015. Adam: A method for stochastic optimization. In: *The International Conference on Learning Representations (ICLR)*.
- Klein, S., Staring, M., Murphy, K., Viergever, M. A., Pluim, J. P. W., 2010. elastix: a toolbox for intensity-based medical image registration. *IEEE Trans Med Imaging* 29 (1), 196–205.
- Krebs, J., Mansi, T., Delingette, H., Zhang, L., Ghesu, F. C., Miao, S., Maier, A. K., Ayache, N., Liao, R., Kamen, A., 2017. Robust non-rigid registration through agent-based action learning. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D. L., Duchesne, S. (Eds.), *Medical Image Computing and Computer Assisted Intervention - MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11–13, 2017, Proceedings, Part I*. Springer International Publishing, Cham, pp. 309–317.

- Conference, Quebec City, QC, Canada, September 11–13, 2017, Proceedings, Part I. Springer International Publishing, Cham, pp. 344–352.
- Liao, R., Miao, S., de Tournemire, P., Grbic, S., Kamen, A., Mansi, T., Comaniciu, D., 2017. An artificial agent for robust image registration. In: Proceedings of the ThirtyFirst AAAI Conference on Artificial Intelligence.
- Lin, M., Chen, Q., Yan, S., 2014. Network in network. In: The International Conference on Learning Representations (ICLR).
- Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
- Marstal, K., Berendsen, F., Staring, M., Klein, S., 2016. SimpleElastix: A user-friendly, multi-lingual library for medical image registration. In: Julia Schnabel and Kensaku Mori (Ed.), International Workshop on Biomedical Image Registration (WBIR). IEEE Conference on Computer Vision and Pattern Recognition Workshops. Las Vegas, Nevada, USA, pp. 574–582.
- Miao, S., Wang, Z. J., Liao, R., 2016. A CNN regression approach for real-time 2D/3D registration. *IEEE Trans Med Imaging* 35 (5), 1352–1363.
- Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A., 2017. Automatic differentiation in pytorch.
- Pluim, J. P. W., Maintz, J. B. A., Viergever, M. A., 2003. Mutual-information-based registration of medical images: a survey. *IEEE Trans Med Imaging* 22 (8), 986–1004.
- Radau, P., Lu, Y., Connolly, K., Paul, G., Dick, A., Wright, G., 2009. Evaluation framework for algorithms segmenting short axis cardiac MRI. *The MIDAS Journal*.
- Rueckert, D., Sonoda, L. I., Hayes, C., Hill, D. L. G., Leach, M. O., Hawkes, D. J., 1999. Nonrigid registration using free-form deformations: application to breast MR images. *IEEE Trans Med Imaging* 18 (8), 712–721.
- Schnabel, J. A., Rueckert, D., Quist, M., Blackall, J. M., Castellano-Smith, A. D., Hartkens, T., Penney, G. P., Hall, W. A., Liu, H., Truwit, C. L., Gerritsen, F. A., Hill, D. L. G., Hawkes, D. J., 2001. A generic framework for non-rigid registration based on non-uniform multi-level free-form deformations. In: Niessen, W. J., Viergever, M. A. (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2001: 4th International Conference Utrecht, The Netherlands, October 14–17, 2001 Proceedings. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 573–581.
- Shen, D., Davatzikos, C., Nov. 2002. HAMMER: hierarchical attribute matching mechanism for elastic registration. *IEEE Trans Med Imaging* 21 (11), 1421–1439.
- Sokooti, H., de Vos, B., Berendsen, F., Lelieveldt, B. P. F., Işgum, I., Staring, M., 2017. Nonrigid image registration using multi-scale 3D convolutional neural networks. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D. L., Duchesne, S. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11–13, 2017, Proceedings, Part I. Springer International Publishing, Cham, pp. 232–239.
- Sotiras, A., Davatzikos, C., Paragios, N., 2013. Deformable medical image registration: A survey. *IEEE Trans Med Imaging* 32 (7), 1153–1190.
- Staring, M., Klein, S., Pluim, J. P. W., 2007. A rigidity penalty term for nonrigid registration. *Medical Physics* 34 (11), 4098–4108.
- The National Lung Screening Trial Research Team, 2011. Reduced lung-cancer mortality with low-dose computed tomographic screening. *New England Journal of Medicine* 365 (5), 395–409.
- Vercauteren, T., Pennec, X., Perchant, A., Ayache, N., Mar. 2009. Diffeomorphic demons: efficient non-parametric image registration. *NeuroImage* 45 (1 Suppl), S61–72.
- Viergever, M. A., Maintz, J. A., Klein, S., Murphy, K., Staring, M., Pluim, J. P., 2016. A survey of medical image registration – under review. *Medical Image Analysis* 33 (Supplement C), 140–144, 20th anniversary of the Medical Image Analysis journal (MedIA).
- Wu, G., Kim, M., Wang, Q., Munsell, B. C., Shen, D., 2016. Scalable high performance image registration framework by unsupervised deep feature representations learning. *IEEE Trans Biomed Eng* 63 (7), 1505–1516.
- Yang, X., Kwitt, R., Styner, M., Niethammer, M., 2017. Quicksilver: Fast predictive image registration – a deep learning approach. *NeuroImage* 158, 378–396.
- Yu, J. J., Harley, A. W., Derpanis, K. G., 2016. Back to basics: Unsupervised learning of optical flow via brightness constancy and motion smoothness. In: Computer Vision - ECCV 2016 Workshops, Part 3.