Multi-modal deep learning based 3D image registration for osseous anatomies

Hidde van Ulsen

Technische Universiteit Delft

TU Delft

Multi-modal deep learning based 3D image registration for osseous anatomies

by

Hidde van Ulsen

in partial fulfilment of the requirements to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Tuesday June 14, 2022 at 15:00 PM.

Student number: 4315367

Thesis committee:

Prof. dr. ir. M.J.T. Reinders,	TU Delft, Chair of thesis committee
Dr. ir. M. Staring,	Leiden UMC, TU Delft, supervisor
Dr. M. van Stralen,	MRIguidance b.v., UMC Utrecht, supervisor
Dr. K. Hildebrandt,	TU Delft, committee member

TU Delft

Abstract

Image registration is a fundamental requirement for many medical applications. In recent years, deep learning approaches for registration have shown to be a promising alternative to conventional methods. However, most learning based methods do not consider the different physical properties of various tissues, which can result in unrealistic deformation in anatomical regions where both deformable tissue and rigid bone is present. In this work, we develop and evaluate deep learning methods for intrapatient CT-MR registration while maintaining rigidity of the bones. Unconstrained and locally constrained registration methods are compared in an unsupervised and weakly-supervised setting. The results show that qualitatively and quantitatively accurate registrations can be obtained.

List of Figures

List of Tables

Contents

1

Introduction

Medical imaging techniques are an important tool to obtain insights in the human body. Over the past years, the advancement of machine learning and the continuous increase in use and generation of medical data [46] has inspired numerous developments in image analysis [31].

One such application is the task of synthetic CT generation. For imaging bone in 3D, computed tomography (CT) is the reference modality because of its fast acquisition speed, high resolution and relatively low cost. The main disadvantage of CT is the ionizing radiation dose, which is especially of concern in children [36]. Another disadvantage of CT is the poor soft tissue contrast. Magnetic resonance imaging (MRI) has excellent soft tissue contrast without ionizing radiation deposition, but is limited in imaging cortical bone with discernable contrast due to low proton density [6]. With synthetic MR-derived CT-like imaging, an MR-only workflow can be obtained to simultaneously assess soft tissue as well as bone structure without radiation exposure [14, 19, 54]. This can potentially simplify the workflow and reduce costs for diagnosis and pre-operative planning of musculoskeletal conditions where typically both CT and MR are required. One overarching requirement for many medical applications, and in particular machine learning approaches for synthetic CT due to the dependency on voxel-wise corresponding image pairs [13], is the need for image registration.

Image registration is the task of spatially aligning two or more images. This done by transforming a *moving image* to align with a non-moving *fixed image*. Such a transformation can be of different complexity depending on the intended application; a distinction can be made between *rigid* or *affine* registration, and *deformable* (also known as *non-rigid* or *elastic*) registration. Rigid transformations have up to 6 degrees of freedom, modelling translation and rotation. Affine transformations can have up to 12 degrees of freedom, additionally being able to capture scaling and shearing motion. Deformable transformations use a high degree of freedom to accurately model local deformations. Conventional iterative registration has been studied extensively and is commonly used for medical image analysis [23, 52]. Because registration is an ill-posed problem [12], and can have multiple desired properties and objectives, it remains challenging and an active research area [52]. Recently deep learning has been used to solve registration problems. Deep learning based image registration usually employs convolutional neural networks (CNN) to predict transformation parameters. It has been shown to be able to predict registration faster than conventional approaches, often near real-time [15, 20]. Conventional iterative image registration can take a few minutes for a low degree of freedom and resolution up to multiple hours

for a high degree of freedom/resolution. Deep learning based approaches can also potentially provide more accurate registration. However, most learning based methods that have been proposed in recent years have been focussed on performing registration with the same imaging modality and do not take into account the physical constraints of deformation with respect to different tissue types.

1.1 Research aim

This study investigates multi-modal registration. In particular, we consider registration of MRI and CT for musculoskeletal imaging which is challenged by the composition of rigid bony anatomy and deformable soft tissues leading to complex spatial transformations between images of the same patient in different poses. The aim of this project is to develop and evaluate deep learning methods for multi-modal registration of intrapatient CT-MR images while maintaining rigidity of the bones.

1.2 Outline

The remaining sections in this work are organized as follows. In chapter 2 the necessary background and related work on image registration and deep learning is provided, followed by the proposed method in chapter 3. In chapter 4 the experiments and results will be detailed, followed by the discussion and conclusion in chapter 5.

2

Background

2.1 Medical imaging

Before 1895, physicians did not have any method to investigate and view the internal anatomy of a patient without invasive techniques [4, 39]. With the invention of of Röntgen imaging [40], physicians obtained the possibility to inspect the internals of the body with a minimally invasive technique.

2.1.1 Computed tomography (CT)

In computed tomography imaging, multiple X-ray slices are generated through rotating an X-ray tube and detector around a patient to obtain projections of the attenuated signals from different angles [4]. The attenuation is based on the electron density of the imaged tissue. A 3D image is reconstructed using a reconstruction algorithm such as filtered backprojection. Image intensities are measured in Hounsfield units, which expresses the attenuation coefficients in relation to the known attenuation in water and air.

2.1.2 Magnetic resonance imaging (MRI)

MRI is a non-invasive imaging modality that generates tissue contrast based on magnetic properties of protons (or other nuclei) by using strong magnetic fields. Protons have a magnetic dipole moment, which results in precession around the Larmor frequency under influence of an external static magnetic field [4]. Radio frequency pulses can be utilised to temporarily excite protons, after which the protons realign or relax to an equilibrium. The temporal changes in net magnetism can be measured by receiver coils, which can then be translated into image intensities. As opposed to CT, MRI does not have a calibrated intensity scale. The resulting intensities are dependent on the choice of a specific protocol or sequence and can be based on e.g. variations in longitudinal or transversal relaxation time or proton density.

2.2 Image registration

Image registration is the task of spatially aligning two or more images. Image registration can be used with any set of images that have some degree of structural similarity. Images can be acquired at different points in time, with different subjects and/or with different imaging modalities. We will consider pairwise registration of two images, where in general the goal is to estimate a vector field $\mathbf{u}$ in order to transform a moving image I_M to align with a fixed image I_F using a plausible transformation mapping $\mathbf{T}(\mathbf{x}) = \mathbf{x} + \mathbf{u}$, such that $(I_M \circ \mathbf{T})(\mathbf{x}) = I_M(\mathbf{x} + \mathbf{u})$ and $I_F(\mathbf{x})$ are aligned by having high similarity according to an objective function. Registration is thus typically formulated as an optimisation problem in some form of:

$$\underset{\mathbf{u}}{\operatorname{argmin}}\, S(I_F, I_M, \mathbf{u}) + \gamma R(\mathbf{u}) \tag{2.1}$$

where S is a measure of the (dis)similarity between the images and R is a regularisation term with weight $\gamma \geq 0$ that addresses the ill-posed nature of the problem [12, 48] by enforcing smooth and plausible deformations.

A wide variety of registration methods has been developed [27, 34, 35], where the most suitable method and related parameters is dependent on the specific problem at interest. Main differences in registration methods can be found in three key components; the objective function consisting of the (dis)similarity metric and possibly a regularisation term, the transformation model and in the optimisation algorithm [3, 48].

Similarity measures are based on image intensity or specific features such as landmark points, segmentations or edges of the image. Examples include e.g. sum of squared distances (SSD), mean squared error (MSE), (local) cross-correlation (LCC/CC), mutual information (MI) [33, 53] and normalised gradient fields (NGF) [18].

The transformation model limits the solution space by determining the allowed degrees of freedom (DOF), and can distinguished to be non-parametric or parametric. Non-parametric methods estimate a vector for each pixel or voxel. Parametric methods include the aforementioned rigid (up to 6 DOF) or affine (up to 12 DOF), and non-rigid transformations. For non-rigid registration, instead of directly estimating a vector for each image element, the vector field can also be inferred through interpolation by parametrisation based on control points and basis functions (e.g. B-Splines) which provides some implicit regularisation. Non-parametric methods or a large number of control points result in high dimensional solution spaces which require explicit regularisation to solve [48]. Most regularisation terms are based on the first or second order spatial derivatives of the vector field. Ideally, appropriate regularisation and choice of transformation model results in a diffeomorphic transformation, i.e. a bijective and invertible mapping, but for most methods this is not guaranteed.

Conventional registration methods iteratively optimise equation 2.1 for each image pair, which can result in slow registration for high dimensional solution spaces. Typical optimisation algorithms include first order derivative based methods such as (stochastic) gradient descent or second order methods such as Newton-Raphson or BFGS. Further details can be found in [3, 34, 48].

2.3 Deep learning

Many of the complex tasks in machine learning which have been successfully accomplished to a certain extent in recent years make use of deep neural networks. Inspired by biological neural networks, artificial neural networks consist of many interconnected neurons. In deep learning parameters of neurons are trained on data to fit or "learn" a certain function. Neurons are usually connected in parallel to form layers. Multiple layers are stacked sequentially to form networks, and with enough layers and parameters these networks are then considered to be *deep* neural networks [29, 41].
Computation in neural networks usually consists of two operations, a forward pass and a backward pass (backpropagation). In the forward pass, for a single neuron j first multiple inputs are summed together to obtain a linear combination of n inputs a_i weighted by θ_{ij} and optionally thresholded by bias b_j. Next, a non-linear activation function σ is applied to obtain the output y_j. For a neuron in layer l the output is thus

$$y_j^l = \sigma\left(b_j^l + \sum_{i=0}^{n} \theta_{ij}^l a_i^{l-1}\right) \tag{2.2}$$

The output is then used in the next layers, where the same principle is applied. The weights θ_{ij} and bias b_j are the trainable parameters $\boldsymbol{\theta}$ of the network. Considering an initial input $\mathbf{x}$, the mapping to the output $\mathbf{y}$ of a network can be expressed by $\mathbf{y} = f_{\boldsymbol{\theta}}(\mathbf{x})$.
The error of the predicted output is calculated by a loss function on the training samples. To update the weights in the backward pass, the loss is then propagated trough the network by backpropagation, which essentially calculates the gradient of a loss function with respect to the individual weights using the chain rule.

2.3.1 Convolutional neural nets (CNN)

Most tasks in computer vision are solved by making use of convolutional neural networks (CNN). As opposed to using connections between all neurons, which can quickly become intractable for common input dimensions, in CNN's weights are encoded in learnable convolutional kernels/filters. This provides some form of spatial invariance and reduces the number of trainable parameters in a network. Depending on the dimensionality and structure of the input data, convolutional layers can use one, two or three dimensional kernels. The kernel is moved by taking a step, or stride, on the input grid where the output is calculated by taking the product of the kernel weights and the input at that specific position to output a feature map. In successive layers, the resolution of the intermediate feature maps is usually reduced, increasing the receptive field. This can be achieved by using a kernel with stride > 1 or through non-parametric pooling operations, by taking the maximum or average of a small patch.

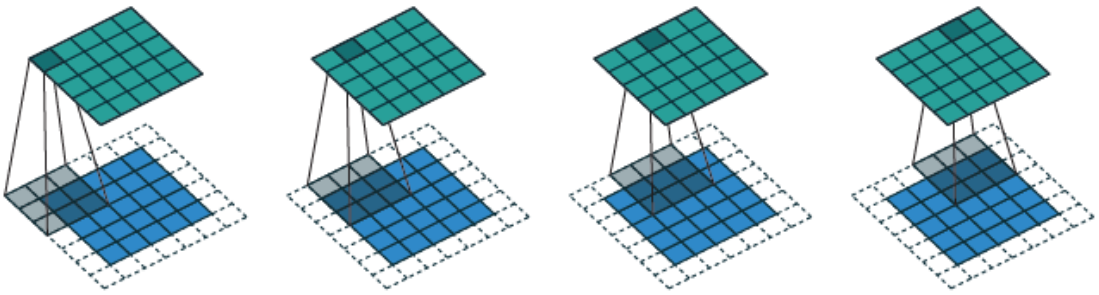


Figure 2.1: Illustration of 2D convolutional operation [9]

2.4 Learning based image registration

Various deep learning based image registration methods have been proposed, each with distinctive properties, advantages and disadvantages, of which a limited overview will be discussed here. Differences in pipelines exist with respect to e.g. learning method ((weakly-) supervised, unsupervised), network architecture and spatial dimensionality (2D, 2D/3D, 3D) [15, 20]. For most methods, the general concept is to learn the parameters $\boldsymbol{\theta}$ of a network to learn a transformation function

$$f_{\boldsymbol{\theta}}(I_M, I_F) = \mathbf{u} \tag{2.3}$$

which can then be used to directly (in a single forward pass) or indirectly estimate the deformation vector field $\mathbf{u}$, instead of iteratively optimising a cost function with respect to (a parametrized version of) $\mathbf{u}$ for each image pair. Training a network is performed by optimising, depending on the type of supervision, similar objective functions as used in conventional image registration.

The specific nuances of different supervision methods in learning based registration needs some elaboration. When considering non-parametric deformable registration, since the output of a network is a vector field $\mathbf{u}$, *supervised* methods in a strict sense use the error between predicted and associated target vector fields as the loss function. These can be acquired by artificially generating representative synthetic sample deformations, i.e. by using a single image to obtain a corresponding artificial image to be registered to. This removes the need for an image similarity metric. However, depending on the assumptions on the degree of realism that is required, this is difficult - especially in the multi-modal case it might be infeasible to generate realistic anatomically correct deformations that are good representations of the complex transformations stemming from acquisition with different scanners, timepoints, field of view and pose.

Alternatively the output vector field of traditional image registration frameworks can be used. In practice, images can also be used instead of vector fields to provide some form of supervision by utilising an intensity based loss between the transformed moving image and a ground-truth registered image. In both approaches, the performance of learning based registration is limited by the quality of the ground-truth vector fields or aligned images. Note that manual expert-knowledge based generation of "ground-truth" deformations can be done for rigid transformations, but in practice is impossible for deformable registration (i.e. estimating vector fields by hand). Also note that by employing vector fields or images from traditional image registration methods as supervision there remains a dependency on image similarity metrics, as these are still used in the complete pipeline.

For mono-modal registration, randomly generated simulated ground truth vector field supervision was used in e.g. Sokooti et al. [47], Eppenhof et al. [11] and Uzunova et al. [51]. In multi-modal registration, Cao et al. [5] used an image based supervised approach for deformable CT-MR registration, by first performing registration using traditional frameworks based on image and label similarity. This dataset was then used to train a network by calculating the intramodal loss over both modalities, i.e. combining warped CT to pre-aligned CT and warped MR to pre-aligned MR losses.

Unsupervised methods only rely on the input moving and fixed image, without the need of a ground-truth, by making use of an image similarity metric as the loss and warping the moving images using interpolation methods introduced in spatial transformer networks [26]. Unsupervised registration has been applied in several works [2, 7, 28, 30]. Due to implicit regularisation of optimisation across a dataset, these methods can potentially avoid local minima which can sometimes be an issue in registration methods [8]. However, especially for use in multi-modal registration, the same challenges with image similarity metrics as in traditional methods remain.

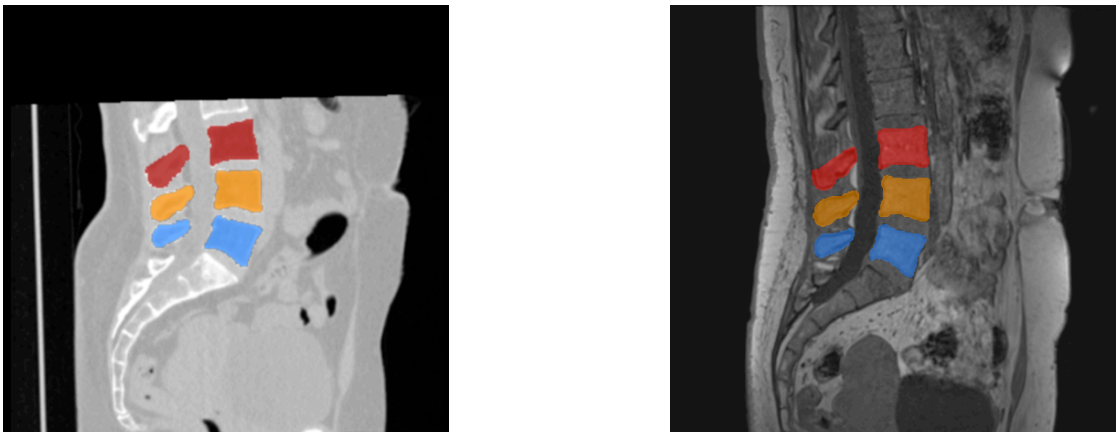
Weak-supervision allows to introduce additional structural information (e.g. landmark points, segmentations) during training, while not relying on these labels during inference. This makes the task more similar to other tasks in computer vision which are usually solved with label supervision (e.g. classification, segmentation), although it should be noted that the goal is not to predict the labels as output. Labels can be acquired (semi)automatically or manually and are independent of modality. Although acquisition of labels can be a difficult and/or labour intensive task, it might be less challenging than generating representative synthetic deformations for full supervision. Labels in the form of landmark points or segmentations are usually only available for a certain region of interest, and not for the full image. Furthermore, for a segmentation based loss, the *change* in overlap drives the optimisation, thus for binary segmentation masks the voxels that contribute most to the loss difference are likely to be concentrated at the edge of a segmentation mask. Hence, the notion of *weak* supervision in this case refers to the *sparsity* of training data on a voxel-wise level, and not necessarily to the *quality* which is determined by the accuracy of the segmentations. In other domains of computer vision weak supervision can have a different meaning. Hu et al. [24] first proposed a method to perform MR-US registration of the prostate using only segmentation labels as supervision. Balakrishnan et al. [2] extended their unsupervised approach to include a segmentation based loss, achieving improved accuracy in mono-modal brain MR registration.

3

Methodology

3.1 Data

The dataset consists of CT and MR images of the same patient. MR images were acquired at 3T with a 3D gradient echo sequence (TE/TR = 2/7 ms) with a resolution of 0.6x0.6x0.8mm. CT images were acquired with a resolution of 0.64x0.64x0.7mm and have a varying field of view. Segmentations of the vertebrae were obtained by using an automatic segmentation network followed by additional manual inspection and correction. An example slice of the dataset including segmentations can be seen in figure 3.1. Images were selected to all have correct segmentations of the L5, L4 and L3 vertebrea within the field of view, resulting in 49 image pairs.



(a) CT **(b)** MR

Figure 3.1: Example sagittal slice of CT and MR images. Segmentations are shown of the L5 (blue), L4 (orange) and L3 (red) vertebrae.

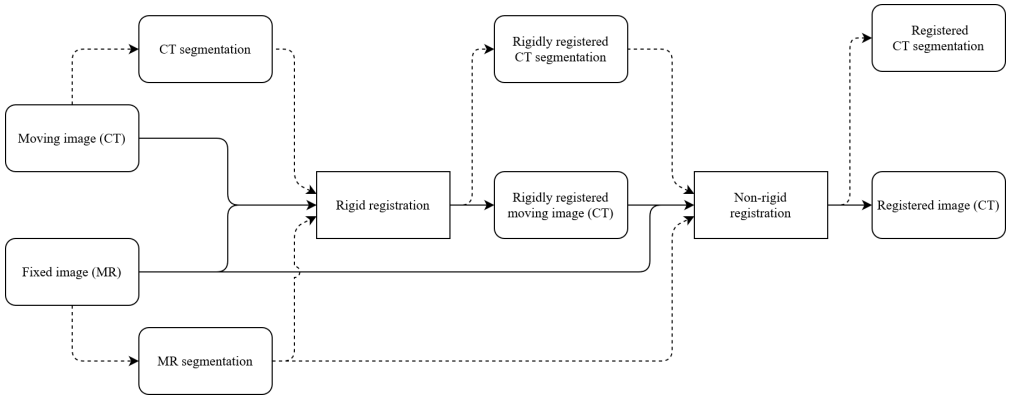


Figure 3.2: Overview of registration pipeline.

3.2 Preprocessing

The images in the dataset have different spatial resolutions, dimensions, intensity scales and can have large initial misalignment, requiring preprocessing for use in neural networks. Both CT and MR are resampled to isotropic 1.0 x 1.0 x 1.0 mm^3 spatial resolution using Lanczos interpolation for images and nearest neighbour interpolation for segmentations. Images and segmentations are cropped to 320 x 320 x 96 voxels. Image intensities are linearly rescaled by mapping the 1st and 99th percentile to [0, 1]. The resulting intensity distribution can be seen in figure 3.3.

3.2.1 Rigid registration

The complete registration pipeline in this work involves two different steps. As a preprocessing step, images and segmentations are first rigidly registered to provide initial alignment. The rigidly aligned images are then used as an input to the network which applies non-rigid registration. An overview can be seen in figure 3.2.

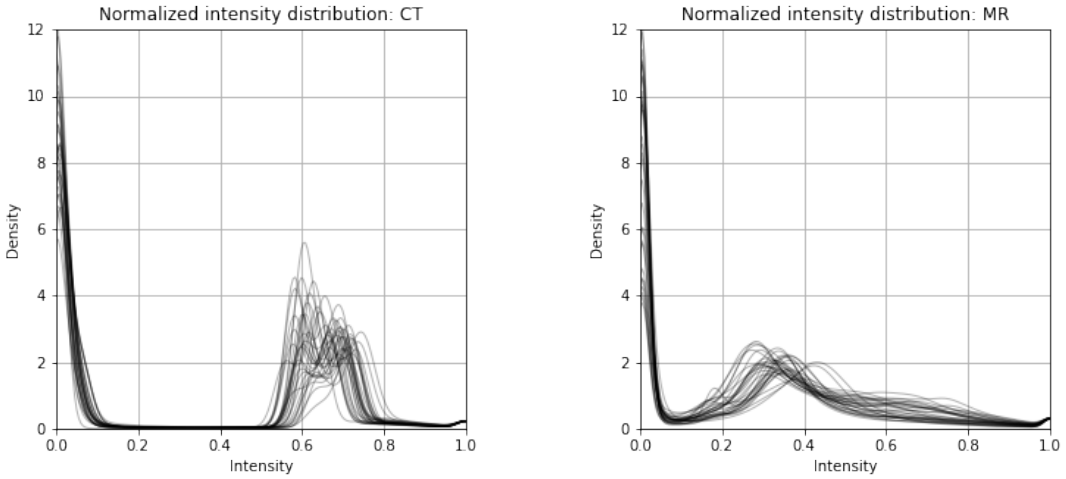
A rigid transform can be described by

$$\mathbf{T}(x, y, z) = \mathbf{R} \begin{bmatrix} x \\ y \\ z \end{bmatrix} + \begin{bmatrix} t_x \\ t_y \\ t_z \end{bmatrix} \tag{3.1}$$

where the matrix $\mathbf{R}$ is composed of rotations

$$\mathbf{R}_x(\phi_1) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\phi_1 & -\sin\phi_1 \\ 0 & \sin\phi_1 & \cos\phi_1 \end{bmatrix}, \ \mathbf{R}_y(\phi_2) = \begin{bmatrix} \cos\phi_2 & 0 & \sin\phi_2 \\ 0 & 1 & 0 \\ -\sin\phi_2 & 0 & \cos\phi_2 \end{bmatrix}, \ \mathbf{R}_z(\phi_3) = \begin{bmatrix} \cos\phi_3 & -\sin\phi_3 & 0 \\ \sin\phi_3 & \cos\phi_3 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{3.2}$$

Rigid registration is performed using Elastix[27] by optimising the error between the L5 segmentation labels.



(a) CT **(b)** MR

Figure 3.3: Intensity distribution (KDE) of normalized CT and MR images where each line represents a different image.

3.3 Non-rigid registration

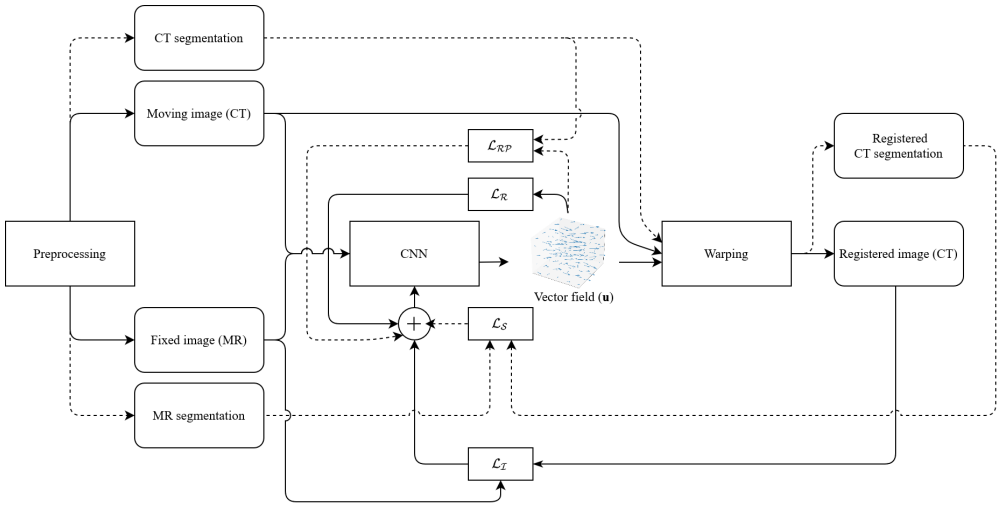


Figure 3.4: Overview of learning based non-rigid image registration method.

An overview of the proposed registration method, which builds upon [16], can be seen in figure 3.4. The moving and fixed images I_M and I_F are used as an input to a convolutional neural network which outputs the vector field. The vector field is used to warp moving image I_M and (optionally) corresponding segmentation S_M. Since image intensities are defined on a discrete grid, interpolation is required to warp the images and segmentations to obtain the new intensity values. Trilinear interpolation is used during training and inference, although to obtain output images with the best quality, different interpolation methods could be used during inference such as tricubic or B-Spline interpolation while keeping the computational efficiency of trilinear interpolation during training. After obtaining the registered image and segmentation, the loss function is calculated and used to update the weights of the network during training. The loss function is based on similarity and regularisation terms, consisting of image similarity $\mathcal{L}_I$, label similarity $\mathcal{L}_S$, global regularisation $\mathcal{L}_R$ and local regularisation $\mathcal{L}_{RP}$:

$$\mathcal{L}(I_F, I_M, S_F, S_M, \mathbf{u}) = \underbrace{-\lambda_1 \mathcal{L}_I(I_F, I_M \circ \mathbf{T}) + \gamma_1 \mathcal{L}_R(\mathbf{u})}_{\text{image}} \underbrace{-\lambda_2 \mathcal{L}_S(S_F, S_M \circ \mathbf{T}) + \gamma_2 \mathcal{L}_{RP}(S_M, \mathbf{T})}_{\text{segmentation}} \tag{3.3}$$

3.4 Loss functions

3.4.1 Regularisation

In order to enforce smoothness of the transformation, a first order regularisation loss is used $\mathcal{L}_R = ||\nabla \mathbf{u}||^2$.

3.4.2 Label similarity

To include information on the of similarity segmentation labels, the Dice loss is used

$$\mathcal{L}_S = \frac{2|S'_M \cap S_F|}{|S'_M| + |S_F| + \epsilon} \tag{3.4}$$

where S'_M is the warped trilinear interpolated moving segmentation label and ϵ is a small term for numerical stability (10^{-7}).

3.4.3 Image similarity

For quantifying image similarity, mutual information is used as the loss function [16, 17]. Mutual information is a measure of the mutual dependence between two or more (stochastic) variables, which are the intensity distributions in the respective images in the context of image registration [37]. The use of intensity distributions instead of using a direct (linear) relation between intensities makes for a relatively general applicable similarity metric and thus suitable for multimodal registration, where as aforementioned there often is no evident linear relationship between intensities. Mutual information can be defined as

$$MI(I_F, I_M) = H(I_F) + H(I_M) - H(I_F, I_M) \tag{3.5}$$

where $H(I)$ is the Shannon entropy [45] for an image with intensity distribution $p(i)$ defined as

$$H(I) = -\sum_{i \in I} p(i) \log p(i) \tag{3.6}$$

and $H(I_F, I_M)$ is the joint entropy of the images with joint intensity distribution $p(i_f, i_m)$

$$H(I_F, I_M) = -\sum_{i_f \in I_F} \sum_{i_m \in I_M} p(i_f, i_m) \log p(i_f, i_m) \tag{3.7}$$

which results in

$$MI(I_F, I_M) = \sum_{i_f \in I_F} \sum_{i_m \in I_M} p(i_f, i_m) \log \left[\frac{p(i_f, i_m)}{p(i_f)p(i_m)} \right] \tag{3.8}$$

The estimate of the intensity distribution in images is usually achieved by calculating the (joint) histogram, i.e. discretizing into a number of bins. To improve optimisation characteristics histograms are further approximated as a continuous function by using kernel density estimation (also known as Parzen-Rosenblatt windowing). For samples $x_1, ..., x_n$ with distribution $p(x)$ the estimated probability then becomes

$$\hat{p}(x) = \frac{1}{n} \sum_{i=1}^{n} K(x - x_i) \tag{3.9}$$

where the kernel K can be any non-negative function such that $\int K(x)dx = 1$, commonly chosen as the Gaussian kernel defined as

$$K(x - x_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-x_i)^2}{2\sigma^2}} \tag{3.10}$$

where σ denotes the bandwidth.

3.4.4 Local regularisation: rigidity penalty

In order to improve local rigidity of transformations, we implement a loss function to penalize local non-rigid transformations based on methods in conventional image registration [32, 43, 44, 49]. The idea is to add an additional regularisation term, which does not act on the global image domain but only locally, to impose constraints in areas with tissue that is known to be rigid and therefore should not be able to deform non-rigidly. This requires prior knowledge of tissue properties. For conventional image registration, this prior knowledge would have to be available for each image, while using a learning based approach allows to implicitly incorporate this prior knowledge into a trained network thus not requiring explicit knowledge of tissue properties during inference.
We follow the approach in [49] to measure departure from multiple conditions that must hold for a vector field to represent a rigid transformation. For a transformation to be rigid, it must hold that the transformation is affine, the Jacobian matrix is orthonormal and orientation is preserved [49]. The segmentation label corresponding to the moving image is used to locally apply constraints on the spatial derivatives of the vector field. Due to the nature of implementation with tensor operations in deep learning frameworks requiring compatible dimensions, gradients are calculated for each location in the vector field, while only the values within the segmentation label are of interest and used in the subsequent calculations.

Linearity Recall that a rigid transformation can be written as an affine function T(x) = Rx + t, i.e. a composition of a linear function and a translation, which implies second order spatial derivatives should be zero. Non-linearity can be measured with the bending energy [44].

$$p_{BE} = \frac{1}{S_M} \iiint_{S_M} \left(\frac{\partial^2 \mathbf{T}}{\partial x^2}\right)^2 + \left(\frac{\partial^2 \mathbf{T}}{\partial y^2}\right)^2 + \left(\frac{\partial^2 \mathbf{T}}{\partial z^2}\right)^2 + 2\left(\frac{\partial^2 \mathbf{T}}{\partial xy}\right)^2 + 2\left(\frac{\partial^2 \mathbf{T}}{\partial xz}\right)^2 + 2\left(\frac{\partial^2 \mathbf{T}}{\partial yz}\right)^2 dx\, dy\, dz \tag{3.11}$$

Orthonormality Since a rigid transformation has an orthonormal Jacobian matrix, deviation from orthonormality can be measured with

$$p_{ON} = \frac{1}{S_M} \iiint_{S_M} \left\| \mathbf{J}\left(\mathbf{T}\right) \mathbf{J}\left(\mathbf{T}\right)^T - \mathbf{I} \right\|_F^2 dx\, dy\, dz \tag{3.12}$$

where $||\cdot||_F$ denotes the Frobenius norm, $\mathbf{I}$ is the identity matrix and $\mathbf{J}(\mathbf{T})$ is defined as

$$\mathbf{J}(\mathbf{T}) = \begin{bmatrix} \frac{\partial \mathbf{t_x}}{\partial x} & \frac{\partial \mathbf{t_x}}{\partial y} & \frac{\partial \mathbf{t_x}}{\partial z} \\ \frac{\partial \mathbf{t_y}}{\partial x} & \frac{\partial \mathbf{t_y}}{\partial y} & \frac{\partial \mathbf{t_y}}{\partial z} \\ \frac{\partial \mathbf{t_z}}{\partial x} & \frac{\partial \mathbf{t_z}}{\partial y} & \frac{\partial \mathbf{t_z}}{\partial z} \end{bmatrix} \tag{3.13}$$

which results in

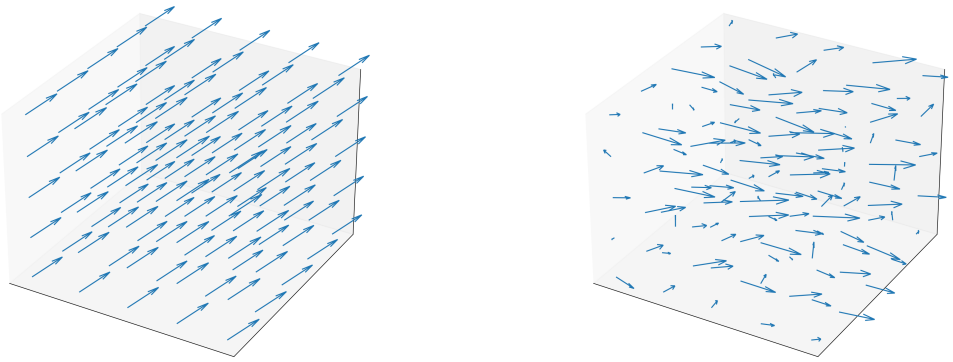
$$\begin{aligned} p_{ON} = \frac{1}{S_M} \iiint_{S_M} & \left(\left(\frac{\partial \mathbf{t_x}}{\partial x}\right)^2 + \left(\frac{\partial \mathbf{t_x}}{\partial y}\right)^2 + \left(\frac{\partial \mathbf{t_x}}{\partial z}\right)^2 - \mathbf{1} \right)^2 + 2\left(\frac{\partial \mathbf{t_x}}{\partial x} \circ \frac{\partial \mathbf{t_y}}{\partial x} + \frac{\partial \mathbf{t_x}}{\partial y} \circ \frac{\partial \mathbf{t_y}}{\partial y} + \frac{\partial \mathbf{t_x}}{\partial z} \circ \frac{\partial \mathbf{t_y}}{\partial z} \right)^2 \\ & + \left(\left(\frac{\partial \mathbf{t_y}}{\partial x}\right)^2 + \left(\frac{\partial \mathbf{t_y}}{\partial y}\right)^2 + \left(\frac{\partial \mathbf{t_y}}{\partial z}\right)^2 - \mathbf{1} \right)^2 + 2\left(\frac{\partial \mathbf{t_x}}{\partial x} \circ \frac{\partial \mathbf{t_z}}{\partial x} + \frac{\partial \mathbf{t_x}}{\partial y} \circ \frac{\partial \mathbf{t_z}}{\partial y} + \frac{\partial \mathbf{t_x}}{\partial z} \circ \frac{\partial \mathbf{t_z}}{\partial z} \right)^2 \\ & + \left(\left(\frac{\partial \mathbf{t_z}}{\partial x}\right)^2 + \left(\frac{\partial \mathbf{t_z}}{\partial y}\right)^2 + \left(\frac{\partial \mathbf{t_z}}{\partial z}\right)^2 - \mathbf{1} \right)^2 + 2\left(\frac{\partial \mathbf{t_y}}{\partial x} \circ \frac{\partial \mathbf{t_z}}{\partial x} + \frac{\partial \mathbf{t_y}}{\partial y} \circ \frac{\partial \mathbf{t_z}}{\partial y} + \frac{\partial \mathbf{t_y}}{\partial z} \circ \frac{\partial \mathbf{t_z}}{\partial z} \right)^2 dx\, dy\, dz \end{aligned} \tag{3.14}$$

Orientation preservation Since a transformation with a orthonormal Jacobian matrix can include reflections/mirroring which is indicated by a negative determinant, orientation can be preserved with

$$p_{JD} = \frac{1}{S_M} \iiint_{S_M} \left(\det \mathbf{J}\left(\mathbf{T}\right) - \mathbf{1}\right)^2 dx\, dy\, dz \tag{3.15}$$

The resulting rigidity penalty loss is then

$$\mathcal{L}_{RP} = p_{BE} + p_{ON} + p_{JD} \tag{3.16}$$



(a) Sample vector field (5x5x5) with equal vectors (translation), $\mathcal{L}_{RP} = 0$

(b) Sample vector field (5x5x5) with random vectors, $\mathcal{L}_{RP} > 0$

Figure 3.5: Illustration of rigid and non-rigid vector fields.

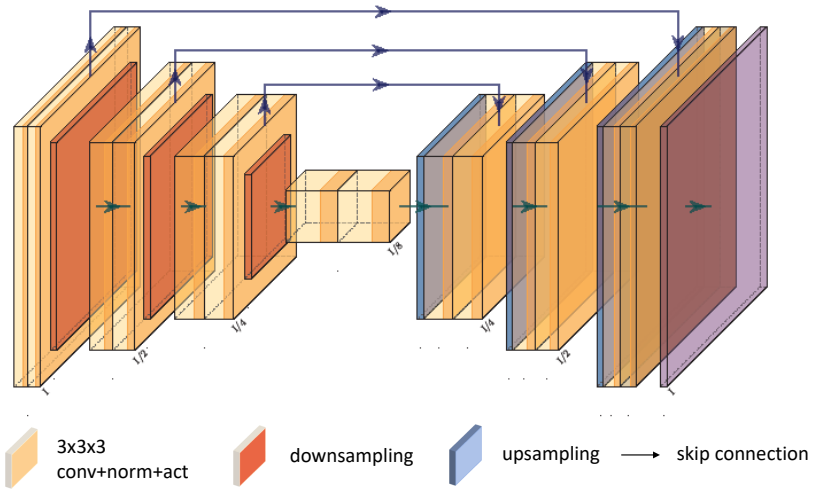


Figure 3.6: Schematic overview of UNet architecture.

3.5 Network architecture

The network architecture which is used in this project is based on a 3D UNet, introduced by Ronneberger et al. [42] for 2D segmentation in small datasets. The UNet architecture follows a fully convolutional encoder-decoder structure, consisting of a contracting and an expanding path where convolutions are respectively applied on progressively reduced and consecutively progressively increased resolutions of the intermediate feature representations.

The input consists of a tensor with the concatenated moving and fixed images. In the contracting path, three repeated blocks consisting of two 3D convolutional layers with stride 1 and a 3x3x3 kernel size are used. Each layer is followed by batch normalisation and ReLU activation. The output is then propagated to the next block by downsampling through a 2x2x2 max pooling operation with stride 2. In the first layers, 4 filters are used, which are doubled after each consecutive block, up to 32 filters in the bottleneck layers.

In the expanding path the same configuration is used in reverse, where 2x2x2 transposed convolutions with stride 2 are used to upsample features while halving the number of filters after each block. Additive skip connections [21] between convolutional layers of the same depth are used to retain some of the detailed feature representations at each resolution level. The last layer is a single 3x3x3 convolutional layer with stride 1 without activation, outputting a 5D tensor which describes the 3D vector field $\mathbf{u}$ at the same resolution as the input for each image pair in a batch. Due to memory constraints, a batch size of 1 was used.

Implementation was done in Tensorflow [1] and Keras using a modified version of the open-source DeepReg framework [16]. Preprocessing, validation and visualisation pipelines were implemented in Python, using Elastix[27], PyTorch/TorchIO [38], OpenCV, and MeVisLab. Training was performed on NVidia RTX3080 and Tesla P100 video cards.

4

Experiments and results

4.1 Experiments

We conduct several experiments to assess the performance of the proposed methods with respect to accuracy and the ability to preserve local rigidity of the transformations. Experiments are done with four different configurations of the loss functions given in section 3.4; we compare unconstrained unsupervised and weakly supervised models and the respective variants with local rigidity constraints.

4.1.1 Experimental set-up

To make an estimate of performance, models were trained using 3-fold cross-validation. The dataset was randomly split in 3 different train and validation sets, and a separate test set which was not utilised for parameter tuning, consisting of 31, 9 and 9 image pairs respectively. Results are reported as an average over folds on the test set. Models were trained using Adam optimisation with a learning rate of $4 \cdot 10^{-4}$, first-order moment estimate decay $\beta_1 = 0.9$ and second-order moment estimate decay $\beta_2 = 0.999$ for 1000 epochs. Mutual information is calculated using histograms discretized into 24 bins.

4.1.2 Evaluation

The results of the registration method are evaluated both qualitatively and quantitatively. In the region of interest accuracy is measured quantitatively based on overlap of and volume change within the segmented structures. For all areas where no structural information is available, results are evaluated by a qualitative investigation of a subset of images and vector fields.

Dice similarity coefficient (DSC)

The Dice coefficient measures spatial overlap between the binary segmentations, defined by

$$DSC = \frac{2|S' \cap S|}{|S'| + |S|} \tag{4.1}$$

where $|\cdot|$ is the number of voxels in the segmentation label. The DSC is a scalar between 0 and 1, where a higher value indicates better overlap. This is equivalent to the notion of the F_1 score

in other domains.

Mean surface distance (MSD)

The mean surface distance is defined as

$$MSD = \frac{1}{|S| + |S'|} \left(\sum_{p \in S} d(p, S') + \sum_{p' \in S'} d(p', S) \right) \tag{4.2}$$

where the error between the surface of label S and S' is given by the absolute surface distance, calculated for each point p on surface S as

$$d(p, S') = \min_{p' \in S'} ||p - p'|| \tag{4.3}$$

Hausdorff distance (HD)

The Hausdorff distance is the largest difference between surface distances, defined as

$$HD = \max \left(d(S, S'), d(S', S) \right) \tag{4.4}$$

where $d(S, S')$ is given by

$$d(S, S') = \max_{p \in S} \min_{p' \in S'} ||p - p'|| \tag{4.5}$$

Since the Hausdorff distance can be sensitive to potential outliers [25, 50] the 99th percentile of the surface distances is used for evaluation.

Jacobian determinant

To investigate topological properties of the transformation, the Jacobian determinant is calculated for each point in the vector field, given in equation 4.6, where partial derivatives are approximated using the discrete central finite difference. As mentioned earlier, the Jacobian determinant measures local volumetric change. A value between 0 and 1 is indicative of compression, a value larger than 1 indicates expansion and a Jacobian ≤ 0 indicates implausible and thus undesired folding in the registered image. For transformation of rigid structures, there should be no volumetric change and as such the Jacobian determinant should ideally be 1.

$$\mathbf{JD}(\mathbf{T}) = \begin{vmatrix} \frac{\partial \mathbf{t_x}}{\partial x} & \frac{\partial \mathbf{t_x}}{\partial y} & \frac{\partial \mathbf{t_x}}{\partial z} \\ \frac{\partial \mathbf{t_y}}{\partial x} & \frac{\partial \mathbf{t_y}}{\partial y} & \frac{\partial \mathbf{t_y}}{\partial z} \\ \frac{\partial \mathbf{t_z}}{\partial x} & \frac{\partial \mathbf{t_z}}{\partial y} & \frac{\partial \mathbf{t_z}}{\partial z} \end{vmatrix} \tag{4.6}$$

4.2 Results

The quantitative results of the experiments are presented in figure 4.2 and table 4.1. To provide insights in the qualitative registration results, figure 4.3-4.5 show exemplary slices of different images and segmentations before and after registration, as well as the colormap and the histogram of the Jacobian determinant, respectively in the full image domain and within the segmented structures. We describe the results in more detail in the next sections.

4.2.1 Non-rigid registration

Unsupervised

To assess the performance of unsupervised models, a baseline was established by using only images as an input, i.e. with a loss function with image similarity loss $\mathcal{L}_I$ and regularization loss $\mathcal{L}_R$ as

$$\mathcal{L}(I_F, I_M, \mathbf{u}) = -\mathcal{L}_I(I_F, I_M \circ \mathbf{T}) + \gamma_1 \mathcal{L}_R(\mathbf{u}) \tag{4.7}$$

The ratio of the image similarity loss and regularisation provides a balance between smooth deformation fields and the ability to capture local deformations. To investigate the influence of the regularisation weight, models were trained with varying values for γ_1. The percentage of all locations in the vector field with a Jacobian determinant ≤ 0 was calculated, of which the results can be seen in figure 4.1.

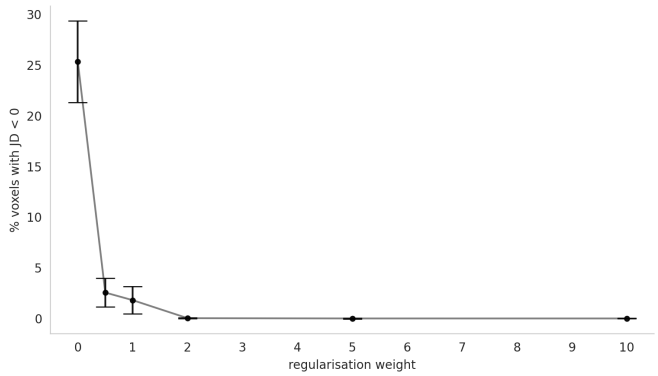


Figure 4.1: Percentage of voxels that show folding for different regularisation weights

As can be seen in figure 4.1, regularisation has a large influence on the plausibility of the registrations as measured by the percentage of locations in the vector field that show folding. While low values for γ_1 result in folding with completely unrealistic results, high values resulted in plausible images with reduced registration accuracy. Consequently $\gamma_1 = 2$ was chosen, with on average 0.02% folding. Visual inspection shows that for all images the alignment of soft tissue improves compared to the initial rigid registration. Figure 4.2 shows that the alignment improves for L3, but a small degradation of registration quality can be observed in the L5 vertebra, which was considered to be registered nearly optimal in the initial registration. Furthermore, local rigidity is not preserved. An example can be seen in figure 4.4a, where unrealistic warping of vertebrae can be observed.

Weakly supervised
To investigate whether including information from the segmentation labels in the loss can improve the alignment of the vertebrae, a weakly supervised approach is investigated by adding the Dice loss $\mathcal{L}_S$ and using the loss function

$$\mathcal{L}(I_F, I_M, S_F, S_M, \mathbf{u}) = (\lambda - 1)\mathcal{L}_I(I_F, I_M \circ \mathbf{T}) - \lambda\mathcal{L}_S(S_F, S_M \circ \mathbf{T}) + \gamma_1\mathcal{L}_R(\mathbf{u}) \tag{4.8}$$

with $\lambda = 0.5$. The registration quality improves for all labels compared to the unsupervised models, achieving an improvement in Dice score from 0.924 to 0.939. Interestingly, what can be seen when visually inspecting the vector field in figure 4.3-4.5c is that for all images the Jacobian determinant at the outer edges of vertebrae deviates from 1, which indicates that local volume change is present.

4.2.2 Locally rigid registration

Unsupervised
To obtain the goal of having locally rigid transformations a rigidity penalty is added. We first assess the influence of the rigidity penalty in the case of unsupervised registration, with the following loss function

$$\mathcal{L}(I_F, I_M, S_F, S_M, \mathbf{u}) = -\mathcal{L}_I(I_F, I_M \circ \mathbf{T}) + \gamma_1\mathcal{L}_R(\mathbf{u}) + \gamma_2\mathcal{L}_{RP}(S_M, \mathbf{T}) \tag{4.9}$$

with $\gamma_2 = 0.2$. The results show similar accuracy as measured by the overlap metrics compared to the unconstrained unsupervised method. The standard deviation of the Jacobian determinant within the segmented structures decreased from 0.095 to 0.025, achieving better preservation of rigidity.

Weakly supervised
To investigate the influence of the rigidity penalty in the case of weak supervision, the following loss function is used

$$\mathcal{L}(I_F, I_M, S_F, S_M, \mathbf{u}) = (\lambda - 1)\mathcal{L}_I(I_F, I_M \circ \mathbf{T}) + \gamma_1\mathcal{L}_R(\mathbf{u}) - \lambda\mathcal{L}_S(S_F, S_M \circ \mathbf{T}) + \gamma_2\mathcal{L}_{RP}(S_M, \mathbf{T}) \tag{4.10}$$

with $\gamma_2 = 0.4$. The results show similar accuracy compared to the unconstrained weakly supervised method. The standard deviation of the Jacobian determinant within the segmented structures decreased from 0.151 to 0.020, achieving better preservation of rigidity. The volume change at the edge of the vertebrae that could be seen in unconstrained weakly-supervised case is not present anymore.

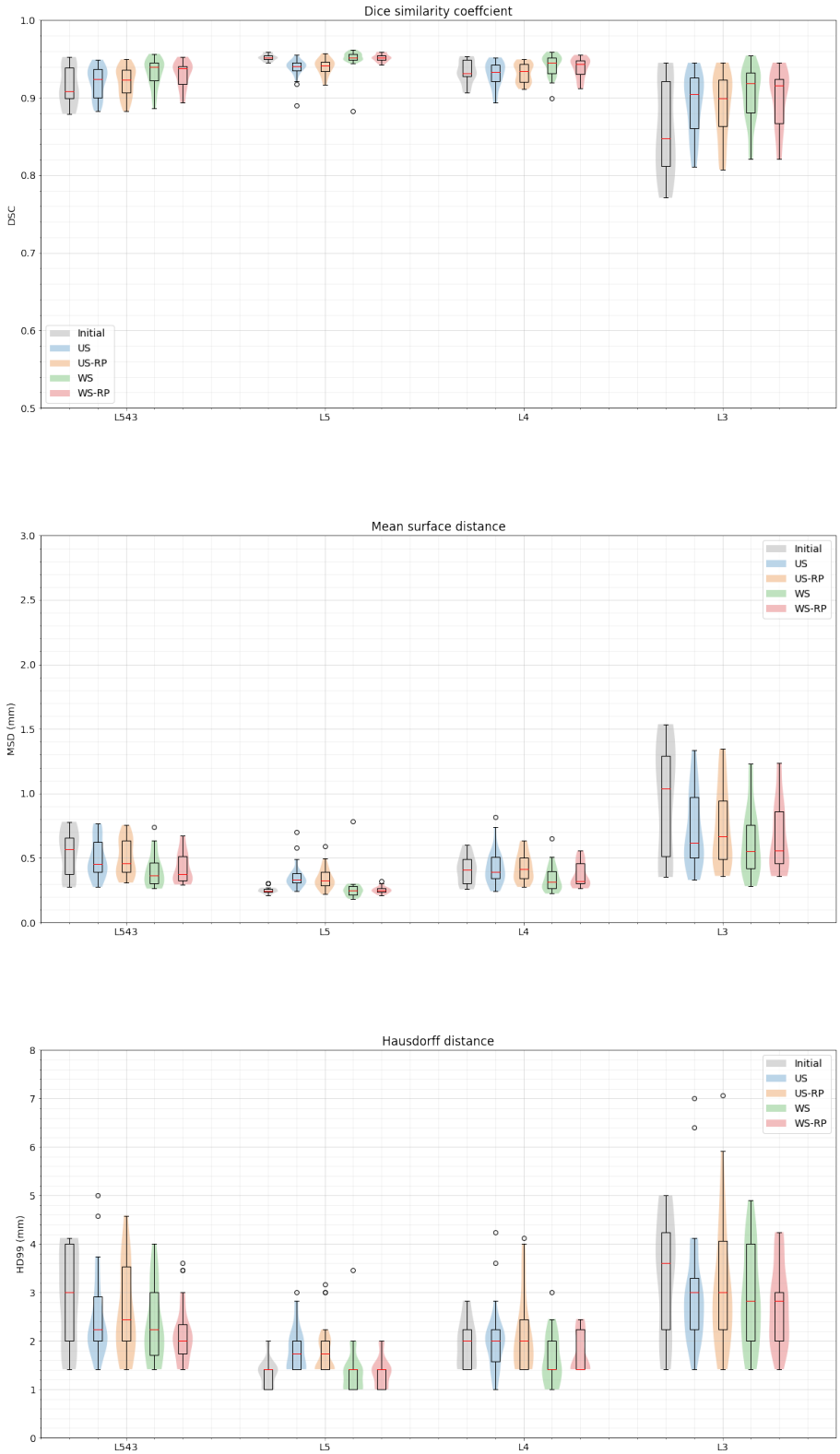
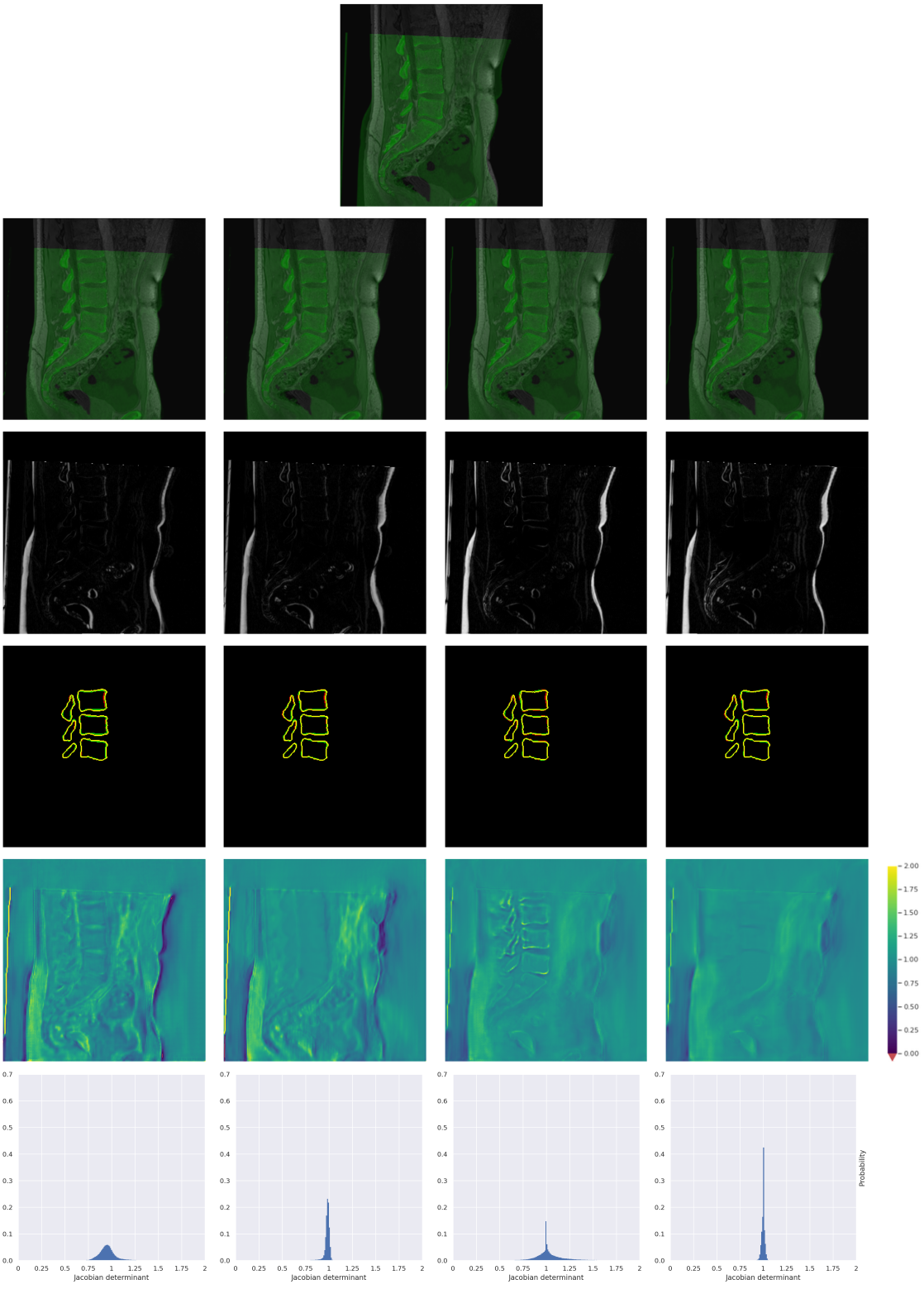


Figure 4.2: Violinplots of the geometrical overlap metrics of the different registration methods. From top to bottom: Dice similarity coefficient (DSC), mean surface distance (MSD), Hausdorff distance (HD).

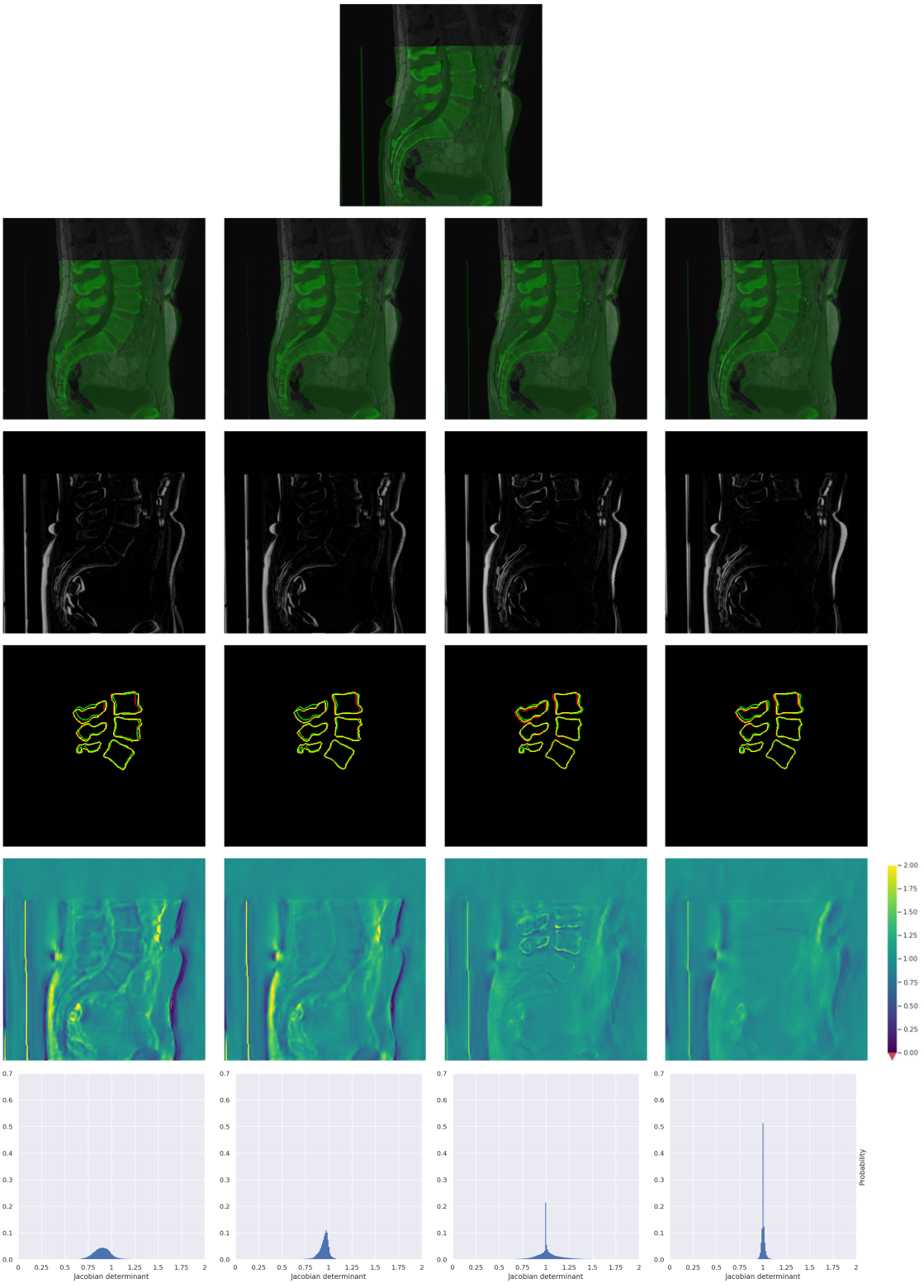
Table 4.1: Comparison of results for initial alignment (Rigid), unsupervised (US), unsupervised with rigidity penalty (US-RP), weakly-supervised (WS), weakly-supervised with rigidity penalty (WS-RP). Results are given in median ± interquartile range of the Dice similarity coefficient (DSC), mean surface distance (MSD), Hausdorff distance (HD99) and the Jacobian determinant (JD) of the local vector field in the vertebrae. Statistically significant differences using a Wilcoxon signed rank test ($p < 0.05$) with the initial alignment (Rigid) are indicated with ‡, whereas significant differences with the unsupervised (US) method are indicated with ⋆.

		Rigid	US	US-RP	WS	WS-RP
DSC	L5	0.952 ± 0.004	0.940 ± 0.010 ‡	0.942 ± 0.012 ‡	0.952 ± 0.008 ⋆	0.951 ± 0.006 ⋆
	L4	0.932 ± 0.021	0.933 ± 0.021	0.935 ± 0.023	0.946 ± 0.021	0.943 ± 0.018
	L3	0.848 ± 0.109	0.905 ± 0.065	0.900 ± 0.059	0.918 ± 0.052 ‡ ⋆	0.917 ± 0.058 ‡ ⋆
	L543	0.908 ± 0.040	0.924 ± 0.037	0.923 ± 0.030	0.939 ± 0.024 ‡	0.937 ± 0.023 ‡
MSD (mm)	L5	0.245 ± 0.023	0.331 ± 0.075 ‡	0.328 ± 0.110 ‡	0.247 ± 0.063 ⋆	0.247 ± 0.032 ⋆
	L4	0.407 ± 0.190	0.394 ± 0.165	0.417 ± 0.161	0.313 ± 0.129	0.355 ± 0.121
	L3	1.037 ± 0.778	0.616 ± 0.467	0.667 ± 0.452	0.552 ± 0.339 ‡ ⋆	0.577 ± 0.345 ‡ ⋆
	L543	0.568 ± 0.282	0.451 ± 0.229	0.503 ± 0.230	0.364 ± 0.158 ‡	0.372 ± 0.185 ‡
HD99 (mm)	L5	1.414 ± 0.414	1.732 ± 0.586	1.732 ± 0.586	1.414 ± 0.414	1.414 ± 0.414
	L4	2.000 ± 0.822	2.000 ± 0.663	2.000 ± 1.035	1.414 ± 0.586	1.414 ± 0.763
	L3	3.606 ± 2.007	3.000 ± 1.067	3.000 ± 1.825	2.828 ± 2.000	2.828 ± 1.000
	L543	3.000 ± 2.000	2.236 ± 0.914	2.449 ± 1.535	2.237 ± 1.293	2.000 ± 0.597
μ JD	L5	1.000 ± 0.000	1.019 ± 0.037	1.000 ± 0.011	1.020 ± 0.013	1.000 ± 0.001
	L4	1.000 ± 0.000	1.013 ± 0.032	0.999 ± 0.014	1.023 ± 0.025	0.998 ± 0.006
	L3	1.000 ± 0.000	1.017 ± 0.035	1.002 ± 0.016	1.017 ± 0.023	0.996 ± 0.009
	L543	1.000 ± 0.000	1.016 ± 0.030	1.001 ± 0.011	1.017 ± 0.022	0.998 ± 0.003
σ JD	L5	0.000 ± 0.000	0.095 ± 0.034	0.026 ± 0.057	0.068 ± 0.038	0.012 ± 0.007
	L4	0.000 ± 0.000	0.095 ± 0.037	0.019 ± 0.020	0.152 ± 0.043	0.018 ± 0.004
	L3	0.000 ± 0.000	0.093 ± 0.031	0.023 ± 0.009	0.189 ± 0.076	0.022 ± 0.007
	L543	0.000 ± 0.000	0.095 ± 0.029	0.025 ± 0.034	0.151 ± 0.054	0.020 ± 0.006



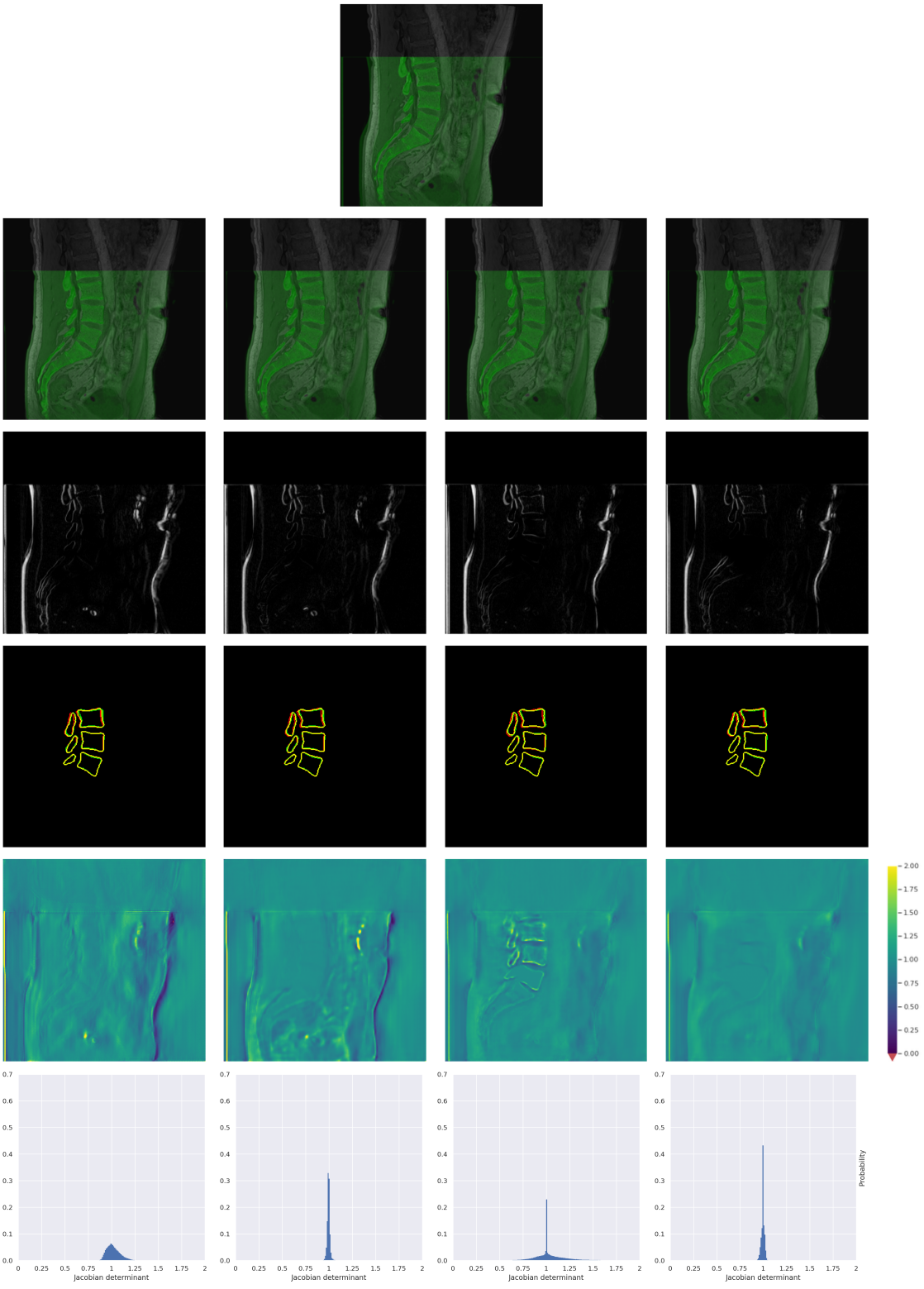
(a) US **(b)** US-RP **(c)** WS **(d)** WS-RP

Figure 4.3: Comparison of different registration methods for a) unsupervised (US), b) unsupervised with rigidity penalty (US-RP), c) weakly-supervised (WS) and d) weakly-supervised with rigidity penalty (WS-RP). From top to bottom: overlay $I_F + I_M$, overlay $I_F + I_M \circ T$, difference $|I_M - I_M \circ T|$, label contours S_F (green) + S_M (red) + $S_M \circ T$ (yellow), colormap of the Jacobian determinant, probability histogram (binwidth = 0.01) of the Jacobian determinant within the segmented structures.



(a) US **(b)** US-RP **(c)** WS **(d)** WS-RP

Figure 4.4: Comparison of different registration methods for a) unsupervised (US), b) unsupervised with rigidity penalty (US-RP), c) weakly-supervised (WS) and d) weakly-supervised with rigidity penalty (WS-RP). From top to bottom: overlay $I_F + I_M$, overlay $I_F + I_M \circ T$, difference $|I_M - I_M \circ T|$, label contours S_F (green) + S_M (red) + $S_M \circ T$ (yellow), colormap of the Jacobian determinant, probability histogram (binwidth = 0.01) of the Jacobian determinant within the segmented structures.



(a) US **(b)** US-RP **(c)** WS **(d)** WS-RP

Figure 4.5: Comparison of different registration methods for a) unsupervised (US), b) unsupervised with rigidity penalty (US-RP), c) weakly-supervised (WS) and d) weakly-supervised with rigidity penalty (WS-RP). From top to bottom: overlay $I_F + I_M$, overlay $I_F + I_M \circ T$, difference $|I_M - I_M \circ T|$, label contours S_F (green) + S_M (red) + $S_M \circ T$ (yellow), colormap of the Jacobian determinant, probability histogram (binwidth = 0.01) of the Jacobian determinant within the segmented structures.

5

Discussion and conclusion

In this study the application of deep learning for multi-modal image registration has been investigated. We present a UNet based approach that employs global and local similarity and regularisation losses to perform registration of MR and CT images while maintaining rigidity of the bones.

The results show that all trained models are able to perform visually accurate registrations. The unsupervised method resulted in an overall improvement although a slight deterioration of registration quality in vertebrae with good initial alignment could be observed. This might be because mutual information is not particularly suited for very precise alignment; rather a global correspondence of intensity distributions is sought after and spatial context is not explicitly taken into account. Adding a segmentation based Dice loss improved upon the accuracy of registration in the region of interest without negatively impacting the overall alignment, in agreement with previous work which has shown the benefit of combining intensity metrics with segmentation based losses [2, 20] where the region of interest was soft deformable tissue. However, our results show that this can introduce local volume change which is undesired for rigid tissue. This is not entirely unexpected, as the model learns to predict transformations guided by change in Dice loss which is the result of change in overlap that mainly occurs at the edge of the segmentation labels. By introducing a loss function that can enforce local rigidity, volume change could be reduced as measured by a reduction of the standard deviation of the Jacobian of the local vector field and therefore the rigid characteristics of bone structures are better preserved.

As aforementioned, deep learning methods for registration to a large extent make use of similar concepts and objectives as conventional methods. The benefit of using a learning based approach is that registration can be performed within seconds and that prior knowledge is embedded in the model through training. As such, the segmentation masks are not required during inference. Note that, as a consequence, a completely locally rigid transformation can not be guaranteed in contrast to conventional methods. Another potential benefit from a learning based approach could be that the models might learn meaningful representations and can learn to recognise which structures should be kept rigid. In the present study, in some cases volume change appeared to be reduced in the models with constraints on rigidity in structures which were not included in the segmentations, such as the L2 vertebra and the sacrum.

We acknowledge several limitations of this study and provide various recommendations for future work. First of all, in general the evaluation of registration methods is difficult due to the lack of a clearly defined ground truth. The common approach is to measure overlap of corresponding segmentations, which naturally relies on the quality of the segmentations to make a proper estimate of accuracy. Evaluation of the experiments could improve by including more segmented regions and by providing quantitative error measurements for the alignment of soft tissue.

Furthermore, while the general findings of the experiments might be expected to hold in different datasets and anatomies, conclusions can only be drawn for the considered dataset and the generalisation ability of the models has to be investigated further. This is not only dependent on the features, data quality and specific normalisation of two images, but also on the respective initial global registration. In the present study, following initial rigid registration, the largest deformation could be found at the outer edge of the body (e.g. 8-14 mm displacements) while within the vertebrae relatively small initial misalignment is present (e.g. 0-4 mm displacements). It would be interesting to investigate the performance with respect to accuracy and the ability to keep structures rigid in anatomies with larger initial misalignment. Moreover, since medical imaging datasets are generally small, the effects of pre-training on large scale diverse datasets with multiple anatomies would be worthwhile to explore.

A main challenge of training networks with high resolution 3D datasets is the high GPU memory requirement, which scales with network size and resolution. In this work, a relatively small architecture ($\sim$160k parameters) is used with an input of almost 20 million voxels. The performance could likely improve with networks with a higher capacity, provided that enough training data is available. To reduce memory requirements, images can be downsampled to a lower resolution. However this leads to a loss of information, and since our interest is in accurate alignment with small deformations in the region of interest this is not a suitable approach for the problem (e.g. consider a misalignment of a few voxels in segmentations which are downsampled using nearest neighbour interpolation). Preliminary experiments on downsampled images - evaluated at full resolution through upsampling and upscaling of the vector fields - seemed to confirm this, achieving visually accurate but quantitatively worse registration accuracy.
Typical medical images might have a resolution in the order of 512^3 voxels, which will not fit in memory of common GPU's. To maintain a full resolution input, a patch based approach could also be utilised. Patch-based approaches are unfortunately not trivial to implement for registration due to the problems associated with boundary conditions, and might be especially difficult for the considered multi-modal setting with both global intensity and local loss terms based on segmentations. With full images, the assumption that there is no deformation at the edge of an image is generally valid, whereas by using patches this would introduce grid-like artefacts. To resolve this, the images would have to be transformed over the boundaries of a patch which evidently results in problems. Further complications could arise when adding additional segmentation based losses where sampling strategies would have to be considered to avoid issues. Future work could try to establish the best methods for highly accurate alignment of high resolution images and investigate memory and parameter efficient architectures.

Obtaining the best performance in registration problems and deep learning problems in general requires tuning of many (hyper)parameters, which in this work was mainly performed manually using a trial-and-error approach. Recent advances in other domains of computer vision such as segmentation or classification have shown promising automatic approaches for parameter optimisation and neural architecture search [10, 22]. It might be interesting to explore whether this can be applied to registration problems as well, although this might be challenging due to the computational demands, the inherent multi-objective nature and difficulties of evaluating registration performance.

5.1 Conclusion

This study presents a learning based method for multi-modal registration while maintaining rigidity of the bones. The results of the experiments suggest that imposing local constraints in selected anatomical regions during training can reduce unrealistic deformation of bones while having similar accuracy compared to unconstrained model variants. Furthermore, the results show that using a similarity loss based on corresponding segmentations can improve accuracy, but this can lead to additional volume change at the edge of segmented regions when no constraints are applied. Based on the experimental results, the best performance could be achieved with the locally constrained weakly-supervised approach, obtaining accurate alignment while keeping bony anatomy nearly rigid. Overall, this study demonstrates the feasibility of deep learning based registration for anatomical regions where both rigid and soft tissue is present, thereby enabling fast and accurate registration without requiring segmentations during inference of the selected anatomical regions that should be kept rigid.

Bibliography

[1] Martin Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. Tensorflow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, pages 265–283, 2016.

[2] Guha Balakrishnan, Amy Zhao, Mert R. Sabuncu, John Guttag, and Adrian V. Dalca. Voxelmorph: A learning framework for deformable medical image registration. *IEEE Transactions on Medical Imaging*, 38(8):1788–1800, Aug 2019.

[3] Lisa Gottesfeld Brown. A survey of image registration techniques. *ACM Comput. Surv.*, 24 (4):325–376, dec 1992.

[4] Jerrold Bushberg. *The essential physics of medical imaging.* Wolters Kluwer Lippincott Williams & Wilkins, Philadelphia, 2021.

[5] Xiaohuan Cao, Jianhuan Yang, Li Wang, Zhong Xue, Qian Wang, and Dinggang Shen. Deep learning based inter-modality image registration supervised by intra-modality similarity. In *Machine Learning in Medical Imaging*, pages 55–63. Springer International Publishing, 2018.

[6] Eric Y Chang, Jiang Du, and Christine B Chung. Ute imaging in the musculoskeletal system. *Journal of magnetic resonance imaging*, 41(4):870–883, 2015.

[7] Bob D. de Vos, Floris F. Berendsen, Max A. Viergever, Marius Staring, and Ivana Išgum. End-to-end unsupervised deformable image registration with a convolutional neural network. In M. Jorge Cardoso, Tal Arbel, Gustavo Carneiro, Tanveer Syeda-Mahmood, João Manuel R.S. Tavares, Mehdi Moradi, Andrew Bradley, Hayit Greenspan, João Paulo Papa, Anant Madabhushi, Jacinto C. Nascimento, Jaime S. Cardoso, Vasileios Belagiannis, and Zhi Lu, editors, *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 204–212, Cham, 2017. Springer International Publishing.

[8] Bob D. de Vos, Floris F. Berendsen, Max A. Viergever, Hessam Sokooti, Marius Staring, and Ivana Išgum. A deep learning framework for unsupervised affine and deformable image registration. *Medical Image Analysis*, 52:128–143, Feb 2019.

[9] Vincent Dumoulin and Francesco Visin. A guide to convolution arithmetic for deep learning. March 2016.

[10] Thomas Elsken, Jan Hendrik Metzen, and Frank Hutter. Neural architecture search: A survey. 2018.

[11] Koen AJ Eppenhof, Maxime W Lafarge, Pim Moeskops, Mitko Veta, and Josien PW Pluim. Deformable image registration using convolutional neural networks. In *Medical Imaging 2018: Image Processing*, volume 10574, page 105740S. International Society for Optics and Photonics, 2018.

[12] Bernd Fischer and Jan Modersitzki. Ill-posed medicine—an introduction to image registration. *Inverse Problems*, 24(3):034008, may 2008.

[13] Mateusz C. Florkow, Frank Zijlstra, Linda G. W. Kerkmeijer, Matteo Maspero, Cornelis A. T. van den Berg, Marijn van Stralen, and Peter R. Seevinck. The impact of mri-ct registration errors on deep learning-based synthetic ct generation. In *Medical Imaging: Image Processing*, 2019.

[14] Mateusz C. Florkow, Frank Zijlstra, Koen Willemsen, Matteo Maspero, Cornelis A. T. Berg, Linda G. W. Kerkmeijer, René M. Castelein, Harrie Weinans, Max A. Viergever, Marijn Stralen, and Peter R. Seevinck. Deep learning–based MR-to-CT synthesis: The influence of varying gradient echo–based MR images as input channels. *Magnetic Resonance in Medicine*, 83(4):1429–1441, oct 2019.

[15] Yabo Fu, Yang Lei, Tonghe Wang, Walter J Curran, Tian Liu, and Xiaofeng Yang. Deep learning in medical image registration: a review. *Physics in Medicine & Biology*, 65(20): 20TR01, oct 2020.

[16] Yunguan Fu, Nina Brown, Shaheer Saeed, Adrià Casamitjana, Zachary Baum, Rémi Delaunay, Qianye Yang, Alexander Grimwood, Zhe Min, Stefano Blumberg, and et al. Deepreg: a deep learning toolkit for medical image registration. *Journal of Open Source Software*, 5 (55):2705, Nov 2020.

[17] C. K. Guo. *Multi-modal image registration with unsupervised deep learning.* PhD thesis, MIT, 2019.

[18] Eldad Haber and Jan Modersitzki. Intensity gradient based registration and fusion of multimodal images. In Rasmus Larsen, Mads Nielsen, and Jon Sporring, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2006*, pages 726–733, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

[19] Xiao Han. MR-based synthetic CT generation using a deep convolutional neural network method. *Medical Physics*, 44(4):1408–1419, mar 2017.

[20] Grant Haskins, Uwe Kruger, and Pingkun Yan. Deep learning in medical image registration: a survey. *Machine Vision and Applications*, 31(1-2), jan 2020.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[22] Xin He, Kaiyong Zhao, and Xiaowen Chu. AutoML: A survey of the state-of-the-art. *Knowledge-Based Systems*, 212:106622, jan 2021.

[23] Derek L G Hill, Philipp G Batchelor, Mark Holden, and David J Hawkes. Medical image registration. *Physics in Medicine and Biology*, 46(3):R1–R45, feb 2001.

[24] Yipeng Hu, Marc Modat, Eli Gibson, Nooshin Ghavami, Ester Bonmati, Caroline M. Moore, Mark Emberton, J. Alison Noble, Dean C. Barratt, and Tom Vercauteren. Label-driven weakly-supervised learning for multimodal deformable image registration. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 1070–1074, 2018.

[25] Daniel P. Huttenlocher, Gregory A. Klanderman, and William Rucklidge. Comparing images using the hausdorff distance. *IEEE Trans. Pattern Anal. Mach. Intell.*, 15:850–863, 1993.

[26] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. *CoRR*, abs/1506.02025, 2015.

[27] Stefan Klein, Marius Staring, Keelin Murphy, Max A Viergever, and Josien PW Pluim. Elastix: a toolbox for intensity-based medical image registration. *IEEE transactions on medical imaging*, 29(1):196–205, 2009.

[28] Julian Krebs, Hervé Delingette, Boris Mailhé, Nicholas Ayache, and Tommaso Mansi. Learning a probabilistic model for diffeomorphic registration. *IEEE transactions on medical imaging*, 38(9):2165–2176, 2019.

[29] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553): 436–444, may 2015.

[30] Hongming Li and Yong Fan. Non-rigid image registration using self-supervised fully convolutional networks without training data. *CoRR*, abs/1801.04012, 2018.

[31] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.

[32] Dirk Loeckx, Frederik Maes, Dirk Vandermeulen, and Paul Suetens. Nonrigid image registration using free-form deformations with a local rigidity constraint. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2004*, pages 639–646. Springer Berlin Heidelberg, 2004.

[33] Frederik Maes, Andre Collignon, Dirk Vandermeulen, Guy Marchal, and Paul Suetens. Multimodality image registration by maximization of mutual information. *IEEE transactions on Medical Imaging*, 16(2):187–198, 1997.

[34] J.B.Antoine Maintz and Max A. Viergever. A survey of medical image registration. *Medical Image Analysis*, 2(1):1–36, mar 1998.

[35] Marc Modat, Gerard R. Ridgway, Zeike A. Taylor, Manja Lehmann, Josephine Barnes, David J. Hawkes, Nick C. Fox, and Sébastien Ourselin. Fast free-form deformation using graphics processing units. *Computer Methods and Programs in Biomedicine*, 98(3):278–284, 2010. HP-MICCAI 2008.

[36] Mark S Pearce, Jane A Salotti, Mark P Little, Kieran McHugh, Choonsik Lee, Kwang Pyo Kim, Nicola L Howe, Cecile M Ronckers, Preetha Rajaraman, Alan W Craft, Louise Parker, and Amy Berrington de González. Radiation exposure from CT scans in childhood and subsequent risk of leukaemia and brain tumours: a retrospective cohort study. *The Lancet*, 380(9840):499–505, aug 2012.

[37] J.P.W. Pluim, J.B.A. Maintz, and M.A. Viergever. Mutual-information-based registration of medical images: a survey. *IEEE Transactions on Medical Imaging*, 22(8):986–1004, 2003.

[38] Fernando Pérez-García, Rachel Sparks, and Sébastien Ourselin. Torchio: A python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. *Computer Methods and Programs in Biomedicine*, 208:106236, Sep 2021.

[39] Jerry Prince. *Medical imaging signals and systems*. Pearson Education, Upper Saddle River, NJ, 2015.

[40] WC Röntgen. On a new kind of rays. *Nature*, 53(1369):274–276, jan 1896.

[41] Daniel A. Roberts, Sho Yaida, and Boris Hanin. The principles of deep learning theory. June 2021.

[42] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Lecture Notes in Computer Science*, pages 234–241. Springer International Publishing, 2015.

[43] Dan Ruan, Jeffrey A. Fessler, Michael Roberson, James Balter, and Marc Kessler. Nonrigid registration using regularization that accomodates local tissue rigidity. In Joseph M. Reinhardt and Josien P. W. Pluim, editors, *SPIE Proceedings*. SPIE, mar 2006.

[44] D. Rueckert, L.I. Sonoda, C. Hayes, D.L.G. Hill, M.O. Leach, and D.J. Hawkes. Nonrigid registration using free-form deformations: application to breast mr images. *IEEE Transactions on Medical Imaging*, 18(8):712–721, 1999.

[45] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.

[46] Rebecca Smith-Bindman, Marilyn L. Kwan, Emily C. Marlow, Mary Kay Theis, Wesley Bolch, Stephanie Y. Cheng, Erin J. A. Bowles, James R. Duncan, Robert T. Greenlee, Lawrence H. Kushi, Jason D. Pole, Alanna K. Rahm, Natasha K. Stout, Sheila Weinmann, and Diana L. Miglioretti. Trends in Use of Medical Imaging in US Health Care Systems and in Ontario, Canada, 2000-2016. *JAMA*, 322(9):843–856, 09 2019.

[47] Hessam Sokooti, Bob de Vos, Floris Berendsen, Boudewijn P. F. Lelieveldt, Ivana Išgum, and Marius Staring. Nonrigid image registration using multi-scale 3d convolutional neural networks. In Maxime Descoteaux, Lena Maier-Hein, Alfred Franz, Pierre Jannin, D. Louis Collins, and Simon Duchesne, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2017*, pages 232–239, Cham, 2017. Springer International Publishing.

[48] Aristeidis Sotiras, Christos Davatzikos, and Nikos Paragios. Deformable medical image registration: A survey. *IEEE Transactions on Medical Imaging*, 32(7):1153–1190, 2013.

[49] Marius Staring, Stefan Klein, and Josien PW Pluim. A rigidity penalty term for nonrigid registration. *Medical physics*, 34(11):4098–4108, 2007.

[50] Abdel Aziz Taha and Allan Hanbury. Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool. *BMC Medical Imaging*, 15(1), aug 2015.

[51] Hristina Uzunova, Matthias Wilms, Heinz Handels, and Jan Ehrhardt. Training cnns for image registration from few samples with model-based data augmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 223–231. Springer, 2017.

[52] Max A. Viergever, J.B. Antoine Maintz, Stefan Klein, Keelin Murphy, Marius Staring, and Josien P.W. Pluim. A survey of medical image registration – under review. *Medical Image Analysis*, 33:140–144, oct 2016.

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

[54] Jelmer M. Wolterink, Anna M. Dinkla, Mark H. F. Savenije, Peter R. Seevinck, Cornelis A. T. van den Berg, and Ivana Išgum. Deep mr to ct synthesis using unpaired data. In Sotirios A. Tsaftaris, Ali Gooya, Alejandro F. Frangi, and Jerry L. Prince, editors, *Simulation and Synthesis in Medical Imaging*, pages 14–23, Cham, 2017. Springer International Publishing.