CNN-based lung CT registration with multiple anatomical constraints

Alessa Hering [a,b,*], Stephanie Häger [a], Jan Moltz [c], Nikolas Lessmann [b], Stefan Heldmann [a], Bram van Ginneken [b,c]

[a] *Fraunhofer Institute for Digital Medicine MEVIS, Maria-Goeppert-Str 3, Lübeck 23562, Germany*
[b] *Diagnostic Image Analysis Group, Radboud University Medical Center, Nijmegen, The Netherlands*
[c] *Fraunhofer Institute for Digital Medicine MEVIS, Max-von-Laue-Straße 2, Bremen 28359, Germany*

ARTICLE INFO

ABSTRACT

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

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

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

1. Introduction

Image registration is the process of aligning two or more images to achieve point-wise spatial correspondence. This is a fundamental step for many medical image analysis tasks and has been an active field of research for decades (Maintz and Viergever, 1998; Sotiras et al., 2013). Various approaches and tailored solutions have been proposed to a wide range of problems and applications. Typically, image registration is phrased as an optimization problem with respect to a spatial mapping that minimizes a suitable cost function and common approaches estimate solutions by applying iterative optimization schemes. Unfortunately, solving such an optimization problem is computationally demanding and consequently slow.

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

In this paper, we target the challenging task of lung registration. The complexity of this registration task is manifold, as the occurring motion is a superposition of respiratory and cardiac mo-

---

* Corresponding author.
 E-mail address: alessa.hering@mevis.fraunhofer.de (A. Hering).

tion. Moreover, the sliding motion between the lung and rib cage during breathing – more precisely between pleura visceralis and pleura parietalis – is an additional challenge. The scale of the motion within the lungs can often be larger than the structures (vessels and airways) that are used to guide the optimization process. This may cause a registration algorithm to get trapped in a local minimum (Heinrich et al., 2013; Polzin et al., 2013). This makes the problem even more difficult.Therefore, a registration method needs to be able to estimate a displacement field that accounts for substantial breathing motion but also aligns small structures like individual pulmonary blood vessels precisely.

2. Related work

Most deep-learning-based approaches aim to learn a registration function in form of a convolutional neural network to predict spatial deformations warping a moving image to a fixed image. All these works have contributed improving deep-learning-based image registration and have been applied to different registration applications including brain MR (Balakrishnan et al., 2018; 2019; Yang et al., 2017b), cardiac MR (de Vos et al., 2017), cardiac MR-CT (Hering et al., 2019c), prostate MR-US (Hu et al., 2018a), thorax-abdomen CT (Heinrich, 2019), thorax CT (de Vos et al., 2019; Eppenhof and Pluim, 2018; Eppenhof et al., 2019; Hering and Heldmann, 2019; Hering et al., 2019a; Sentker et al., 2018) and CT-CBCT registration (Kuckertz et al., 2020). Existing approaches can be classified as *supervised, unsupervised*, and *weakly-supervised* techniques based on how much supervision is available.

Supervised methods use ground-truth deformation fields for training. The ground truth can be generated in different ways. In Eppenhof et al. (2018) and Sokooti et al. (2017) the network is trained on synthetic random transformations. A drawback is that the randomly generated ground truth is artificial and may not be able to reproduce all possible deformations. Alternatively, conventional registration methods can be used to produce deformations by registering images (Sentker et al., 2018; Yang et al., 2017a) or other image features like landmarks or segmentations (Rohé et al., 2017). Another way to create a ground truth is to combine simulations with existing algorithms (Krebs et al., 2017). Consequently, the performances of all these approaches is upper bounded by the quality of the initial registration algorithm or the realism of the synthetic deformations.

In contrast, *unsupervised methods* – also called *self-supervised methods* – do not require any ground truth. The idea is to use the cost function of conventional image registration (similarity measure and regularization term) as the loss function to train the neural network. An important milestone for the development of these methods was the introduction of the spatial transformer network (Jaderberg et al., 2015) to differentiably warp images. This differentiable warping has actually been part of most conventional registration methods for a long time (e.g. König et al. (2018); Modersitzki (2004, 2009)). The concept of an unsupervised deep-learning-based registration method was first introduced with the DIRNet (de Vos et al., 2017) for 2D image registration using the normalized cross-correlation image similarity measure as loss function. In Li and Fan (2018) the approach has been extended by adding diffusion regularization to the loss function forcing smooth deformations. The method has successfully been demonstrated for registration of 3D brain subvolumes. The idea of unsupervised deep-learning-based image registration has been further evolved in several works (Balakrishnan et al., 2018; de Vos et al., 2019; Ferrante et al., 2018; Hering and Heldmann, 2019; Krebs et al., 2019).

Weakly-supervised methods do not rely on ground-truth deformation fields either but training is still supervised with prior information. In Hu et al. (2018a) and Hu et al. (2018b), a set of anatomical labels is used in the loss function. The labels of the moving image are warped by the deformation field and compared with the fixed labels. All anatomical labels are only required during training. In Balakrishnan et al. (2019) and Hering et al. (2019a,c), the complementary strengths of global semantic information and local distance metrics were combined to improve the registration accuracy.

In conventional registration approaches, multilevel continuation and scale-space techniques have been proven very efficient to avoid local minima during the optimization process of the cost function, to reduce topological changes or foldings, and to speed up runtimes (Bajcsy and Kovačič, 1989; Haber and Modersitzki, 2004; Kabus and Lorenz, 2010; Schnabel et al., 2001) – explaining the popularity of multi-level strategies in conventional registration methods. As a lot of deep-learning-based registration methods are build on top of U-Net (e.g. Balakrishnan et al. (2019); Hering and Heldmann (2019); Hering et al. (2019b); Rohé et al. (2017)), they are also multi-leveled in their nature. The first half of the "U" (the encoder) generates features on different scales starting at the highest resolution and reducing the resolution through pooling operations. In this procedure, however, only feature maps on different levels are calculated but neither are different image resolutions used nor deformation fields computed. Only a few approaches implement a multi-resolution or hierarchical strategy in the sense of multilevel strategies associated with conventional methods. In Hu et al. (2018a), the authors proposed an architecture that is divided into a global and a local network, which are optimized together. In Eppenhof et al. (2019), a multilevel strategy is incorporated into the training of a U-Net, by growing and training progressively level-by-level. In de Vos et al. (2019), a patch-based approach is presented, where multiple CNNs (ConvNets) are combined additively into a larger architecture for performing coarse-to-fine image registration of patches. The results from the patches are then combined into a deformation field warping the whole image. Another patch-based multilevel approach is presented in Fu et al. (2020). The multilevel framework consists of a CoarseNet and a FineNet which are trained jointly. During training, the estimated deformation field of the CoarseNet and the FineNet are not combined but the moving patch is transformed twice. During inference, if the mean absolute differences between the deformed image patch and the fixed image exceeds a predefined threshold, FineNet is applied again. This leads to a variable number of deformation field patches, which are combined additively. Although previous deep-learning-based registration works (e.g. de Vos et al. (2019); Eppenhof et al. (2019); Sentker et al. (2018)) contributes many efforts to improve the registration accuracy for lung registration, there is still a misalignment of smaller structures in the lung, which leads to a high target registration error of landmarks.

Contribution

We previously introduced an end-to-end deep-learning multi-level registration method that can handle large deformations by computing deformation fields on different scales and functionally composing them Hering et al. (2019a). This initial study, despite its limited evaluation, proved that it is a valid strategy to improve the alignment of vessels and airways – though a gap regarding the target registration error of landmarks with the best conventional registration methods remained. Building on this previous work, and addressing its limitations, we were able to further close that gap.

Our key contributions are as follow:

- We present multiple anatomical constraints to incorporate anatomical priors into the registration framework to obtain more realistic results. We integrate the lung lobe mask to consider the global context. Moreover, the keypoint correspon-
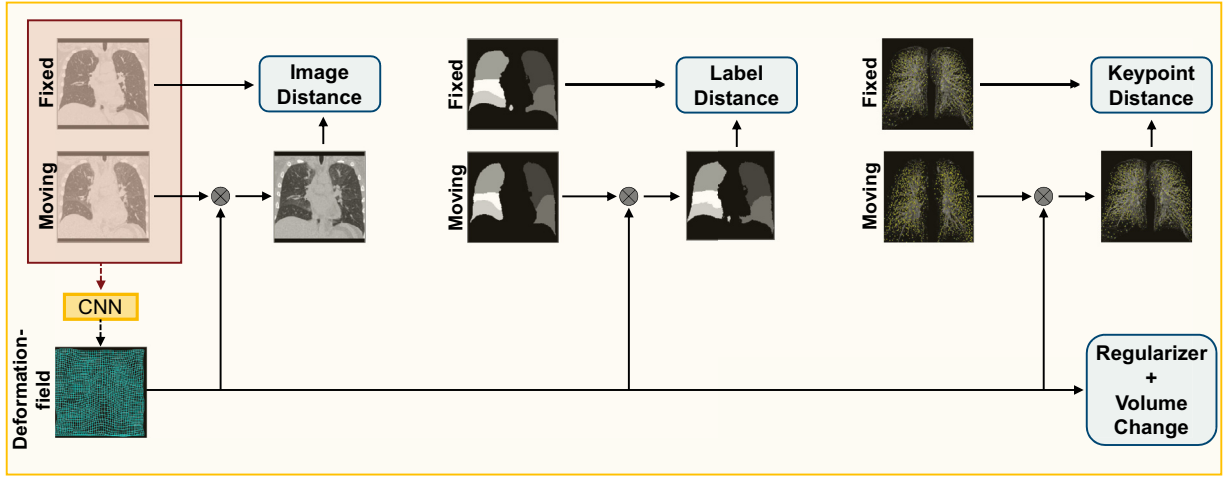
Fig. 1. Schematic representation of the training process. In the loss function, we compare the fixed image, pulmonary lobes mask and keypoints to the deformed moving image, pulmonary lobes mask and keypoints, respectively. To enforce smoothness and to prevent foldings, a regularizer and a volume change penalty are integrated into the loss function. During inference, only the fixed and moving image is required to estimate the deformation field. For a better visualization, we have placed the windowed CT image in the background of the used keypoints. Best viewed in colors.

dences are used to increase the alignment of airways and vessels.

- We introduce a novel constraining method to control volume change and therefore avoid foldings inside the deformation field. While the idea of volume change control is not new in conventional registration, we firstly present a suitable version for deep-learning-based image registration.
- We perform comprehensive experiments on three different datasets – the multi-center COPDGene study (Regan et al., 2011) and the DIRLab challenge dataset (Castillo et al., 2009; 2013), and the EMPIRE10 challenge dataset (Murphy et al., 2011) – to assess the accuracy, plausibility, robustness, transferability of our method. We achieve comparable results as state-of-the-art registration approaches.

3. Method

3.1. Variational registration approach

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

$$\mathcal{J}(\mathcal{F}, \mathcal{M}, y) = \mathcal{D}(\mathcal{F}, \mathcal{M}(y)) + \alpha \mathcal{R}(y) \tag{1}$$

with so-called distance measure $\mathcal{D}$ that quantifies the similarity of fixed image $\mathcal{F}$ and deformed moving image $\mathcal{M}(y)$ and so-called regularizer $\mathcal{R}$ that forces smoothness of the deformation typically by penalizing spatial derivatives. Typical examples for the distance measure are the squared L_2 norm of the difference image (SSD), normalized cross correlation (NCC), or mutual information (MI). The cost function can be extended by additional penalty terms to force desired properties or incorporate additional knowledge in form of anatomical constraints (Rühaak et al., 2017). As illustrated in Fig. 1, our method inputs both the fixed and moving image into the network that predicts the dense displacement field. The loss function uses all available information: input images, segmentation masks and keypoints, with additional regularization – in the form of a smoothness prior and a volume consistency constraint – to prevent foldings.

3.2. Loss function

3.2.1. Normalized gradient field distance measure

One of the main challenges of lung registration are the varying intensity changes occurring due to the altered density of lung tissue during breathing. This leads to a violation of the intensity constancy assumption between corresponding points, on which the classic sum of squared differences (SSD) distance measure is built. However, the lung exhibits a rich structure of bronchi, fissures, and especially vessels that can be exploited for the registration, more suited to distance measure that focus on image edges rather than intensities. We follow the approach of Rühaak et al. (2017) and Hering et al. (2019a) using the *normalized gradient fields* (NGF) (Haber and Modersitzki, 2006) distance measure

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

with $\langle f, g \rangle_\epsilon := \sum_{j=1}^3 (f_j g_j + \epsilon^2)$, $\|f\|_\epsilon := \sqrt{\langle f, f \rangle_\epsilon}$. The edge hyper-parameter $\epsilon > 0$ is used to suppress small image noise, without affecting image edges. Therefore, a good strategy is to choose its value relative to the average gradient. In (Haber and Modersitzki, 2006), the following automatic choice is suggested:

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

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

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

3.2.2. Curvature regularizer

Smooth deformation fields are enforced by the second order *curvature regularizer* (Fischer and Modersitzki, 2003b) given by

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

3.2.3. Volume change control

Although the curvature regularization from Equation (2) prefers smooth deformation, foldings may still happen, which is obviously physically impossible. More formally, foldings happen when the
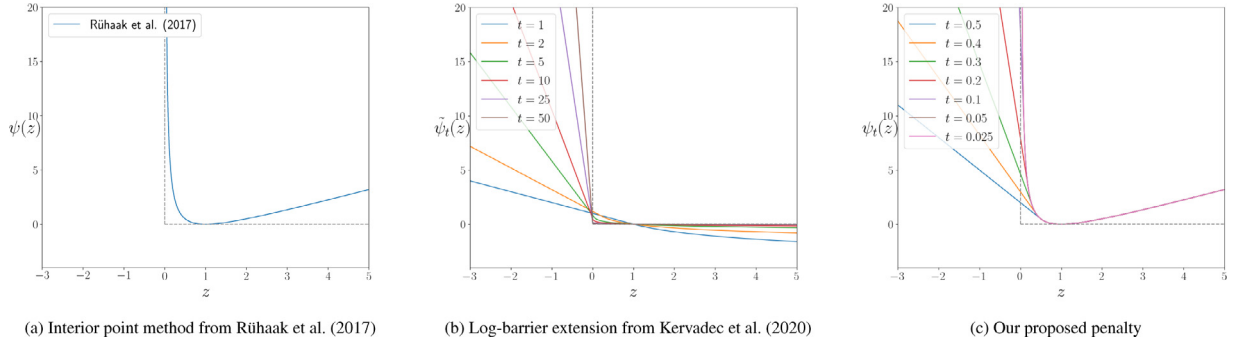
(a) Interior point method from Rühaak et al. (2017) (b) Log-barrier extension from Kervadec et al. (2020) (c) Our proposed penalty

Fig. 2. A graphical illustration of both standard log-barrier (a), the proposed log-barrier extension (b) and examples of penalty functions (c). The solid curves in colors show the approximations for several t values of functions $\tilde{\psi}_t(z)$ and $\psi_t(z)$ respectively.

Jacobian determinant of the deformation field becomes negative. To avoid any foldings, we therefore aim to minimize the distance measure $\mathcal{D}$ and the regularizer $\mathcal{R}$ while keeping the Jacobian determinant positive, for every voxel in Ω. Formally, this can be written as a constrained optimization problem:

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

To achieve this, Rühaak et al. (2017) introduced a *Volume Change Control* (VCC) that could be integrated in their overall objective:

$$\mathcal{V}(y) = \int_\Omega \psi(\det \nabla y(x)) \mathrm{d}x, \tag{3}$$

where

$$\psi(z) = \begin{cases} \dfrac{(z-1)^2}{z} & \text{if } z > 0 \\ +\infty & \text{otherwise.} \end{cases} \tag{4}$$

For the sake of simplicity, the input to ψ in Eq. 3 is substituted with $z = \det \nabla y(x)$ in Eq. 4. Notice that $\psi(z)$ is minimized when $z = 1$ (see Fig. 2 (a)). Therefore, the regularizing effects of the VCC are twofold: i) prevents the formation of foldings, by keeping the determinants positive, ii) limits both shrinkage and expansions by biasing the optimization to keep the same volume.

The method that Rühaak et al. (2017) used falls into the category of *interior-point methods*. Such methods became very popular in constrained optimization (Boyd et al., 2004) as they do not require the expansive primal-dual updates of traditional Lagrangian optimization: the infinity penalty acts as a "barrier", preventing the optimization to go out of bounds.

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

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

Kervadec et al. (2019) (Kervadec, Dolz, Wang, Granger, Ayed, 2020) proposed a parametric log-barrier *extension* (illustrated in Fig. 2 (b)), that does not require an initial feasible solution:

$$\tilde{\psi}_t(z) = \begin{cases} -\dfrac{1}{t}\log(z) & \text{if } z \geq \dfrac{1}{t^2} \\ -tz - \dfrac{1}{t}\log\left(\dfrac{1}{t^2}\right) + \dfrac{1}{t} & \text{otherwise.} \end{cases} \tag{5}$$

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

We propose to keep the property of Eq. (4) to symmetrically penalizes local shrinkage and expansion and make it applicable for neural networks by using the barrier formulation of Eq. (5) for $z < t$ with $t \to 0$ over time (illustrated in Fig. 2 (c)):

$$\psi_t(z) = \begin{cases} \dfrac{(z-1)^2}{z} & \text{if } z \geq t \\ \left(1 - \dfrac{1}{t^2}\right)z + \dfrac{2(1-t)}{t} & \text{otherwise,} \end{cases} \tag{6}$$

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

3.2.4. Mask alignment

Several recent publications (e.g. Balakrishnan et al. (2019); Hering et al. (2019c)) have shown that adding further information in the form of segmentation masks into the loss function can guide the network during the training process. Since the segmentation masks are used in the loss function, they are only required during training and not for registration of unseen images. We integrate segmentation masks by using the SSD loss

$$\mathcal{B}(y) = \frac{1}{2}\int_\Omega \|b_\mathcal{M}(y(x)) - b_\mathcal{F}(x)\|^2 \mathrm{d}x, \tag{7}$$

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

3.2.5. Keypoint loss

For conventional image registration, previous work (e.g. Ehrhardt et al. (2010); Polzin et al. (2013); Rühaak et al. (2017)) has shown that the integration of sparse keypoints during the optimization of the deformation field yields better registration results. In contrast to conventional registration approaches, keypoints can be integrated into the loss function and are therefore, similar

to the segmentation masks for the mask penalty, only needed for training but not during inference. In general, there are several ways to integrate the keypoints into an intensity-based registration approach (e.g. Fischer and Modersitzki (2003a), Rühaak et al. (2017), Papenberg et al. (2009)). We choose to integrate the keypoint information through a least-squares penalty into our model by directly comparing the transformed keypoint of the fixed image with the corresponding moving keypoint:

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

with the moving keypoint $k_{\mathcal{M}}^i$ and the warped fixed keypoint $y\left(k_{\mathcal{F}}^i\right)$ for all $|K|$ keypoints. In general, manually annotated landmarks or automatically generated keypoints can be integrated with this loss function. However, since manual annotation of landmarks is time-consuming, we use the keypoint detection algorithm described in Rühaak et al. (2017) to generate a large number of corresponding keypoints.

The final loss is given by

$$\mathcal{L}(\mathcal{F}, \mathcal{M}, y) = \mathcal{D}(\mathcal{F}, \mathcal{M}(y)) + \alpha \mathcal{R}(y) + \beta \mathcal{B}(y) + \gamma \mathcal{V}(y)$$
$$+ \delta \mathcal{K}(y). \tag{8}$$

The hyper-parameters α, β, γ and δ have to be chosen manually. However, our experiments showed that a change in the magnitude leads to only slight changes in the results.

3.3. Baseline architecture

Our CNN is based on a U-Net (Ronneberger et al., 2015) which takes the concatenated 3D moving and fixed image as input and predicts a 3D dense displacement field with the same resolution as the input images. The U-net consists of three levels starting with 16 filters in the first layer, which are doubled after each downsampling step. We apply 3D convolutions in both encoder and decoder path with a kernel size of 3 followed by an instance normalization and a ReLU layer. In the encoder path, the feature map downsampling steps use $2 \times 2 \times 2$ average pooling with a stride of 2. In the decoder path, the upsampling steps use transposed convolution with $2 \times 2 \times 2$ filters and half the number of filters than the previous step. The final layer uses a 1x1x1 convolution filter to map each 16-component feature vector to a three-dimensional displacement.

3.4. Multilevel architecture

In conventional image registration, multilevel continuation has been proven very efficient to avoid local minima, to reduce topological changes or foldings, and to speed up runtimes (Bajcsy and Kovačič, 1989; Haber and Modersitzki, 2004; Kabus and Lorenz, 2010; Schnabel et al., 2001). Recent deep-learning-based approaches (de Vos et al., 2019; Fu et al., 2020; Hering et al., 2019a; Jiang et al., 2020; Mok and Chung, 2020) have shown that, besides carrying over these properties, a multilevel scheme helps overcome the limitations of deep-learning-based registration approaches to properly deal with small and local deformations.

As in our previous work (Hering et al., 2019a), we follow the ideas of standard multilevel registration and compute coarse grid solutions that are prolongated and refined on the subsequent finer level. Our multilevel framework is illustrated in Fig. 3 with $L = 3$ levels. The registration starts on the coarsest level L where the deformation $\tilde{y}_L$ is computed from the input images that have been Gaussian-smoothed and downsampled by a factor of 2^{L-1}. On all finer levels $\ell < L$, we incorporate the deformations from all preceding coarse levels as an initial guess by combining them by functional composition and warping the moving image. Subsequently, the fixed and warped moving images are downsampled.

The number of used levels is a hyper parameter which should be chosen depending on the task and the used data. The maximal number of levels that can be used is limited by the GPU memory and the image size. Since the images are downsampled with a factor of two in the multilevel setting and additionally the image features are downsampled three-times in the U-Net, the number may be chosen at most so that the image size is divisible by $2^{3+(L-1)}$. Our experiments (c.f. Section 4.9) have shown that a three-level scheme works best in our application and fits on a 12GB GPU. In our experiments, we use in particular a three-level scheme ($L = 3$, Fig. 1). The three networks are trained progressively. First, the network on the coarsest level is trained for a fixed amount of epochs. Afterwards, the parameters of the middle network are learned while the coarsest network stays fixed and is only used to produce the initial deformation field. The same procedure is repeated on the finest level. The same architecture is used on all levels. The network parameters on the coarsest level are initialized with Xavier uniform (Glorot and Bengio, 2010), whereas all other networks are initialized with the learned parameters of the previous network. Note that the receptive field in voxels is the same for all networks, however, due to the decreased resolution on the coarse levels, the receptive field in mm is much larger.

4. Experiments

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

4.1. Data

We train and validate our method on the data from the COPDGene study (Regan et al., 2011). To prove the robustness and transferability of our method and to compare our method with other registration approaches, we evaluate our registration approach on the publicly available DIRLab dataset (Castillo et al., 2009; 2013) and on the EMPIRE10 challenge as well. On the COPDGene dataset, the evaluation is based on the lobe segmentation masks, and on both of the other datasets, annotated landmarks are available on which we evaluate the target registration error.

4.1.1. COPDGene Dataset

Training, validation, and testing data were acquired from the COPDGene study, a large multi-center clinical trial with over 10,000 subjects with chronic obstructive pulmonary disease (COPD) (Regan et al., 2011). The COPDGene study includes clinical information, blood samples, and chest CT scans. The image dataset was acquired across 21 imaging centers using a variety of scanner makes and models. Each patient had received two breath-hold 3D CT scans, one on full inspiration (200mAs) and one at the end of normal expiration (50mAs). About five years later, follow-up images were acquired from about 6000 subjects. In our study, we use the inspiration and expiration scans of 1000 patients. We split these patients into 750, 50, 200 patients for training, validation, and testing, respectively. The original images have sizes in the range of $512 \times 512 \times \{341, \ldots, 974\}$ voxels. The in-plane resolution of the axial slices varied between 0.5mm to 0.97mm per voxel with a slice thickness of 0.45mm to 0.7mm.

The human lungs are sub dived into five lobes that are separated by visceral pleura called pulmonary fissure. An exemplary inspiration scan and expiration scan of one patient with the lobe segmentation overlay is shown in Fig. 4. For all scans segmentations of the lobes are available, which were computed automatically and manually corrected and verified by trained human analysts.
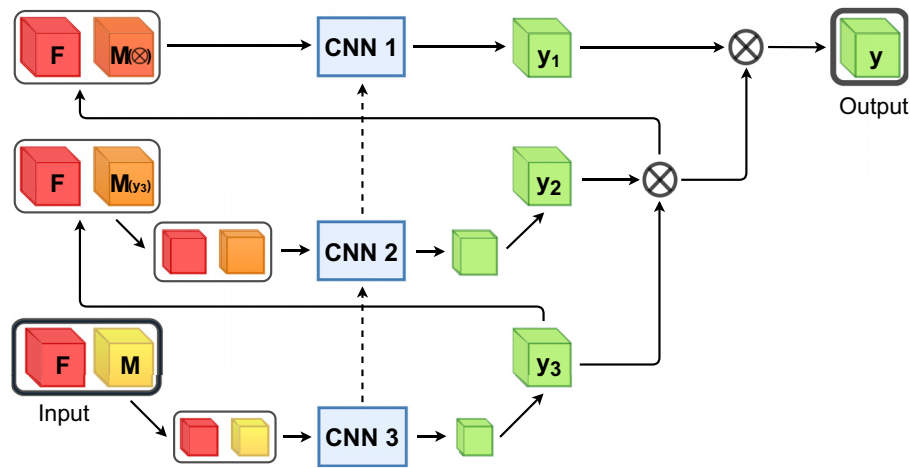
Fig. 3. Overall scheme of the proposed multilevel framework, where $\mathcal{F}$ indicates the fixed image, $\mathcal{M}$ the moving image, y the deformation field and $\mathcal{M}(y)$ the warped image. Each CNNs is trained separately for a fixed amount of epochs and the weights stay fixed afterwards. The deformation fields from all preceding coarse levels are used as an initial guess by combining them by functional composition and warp the moving image on the highest resolution. Subsequently, the warped moving image is downsampled to the current image level. The dotted lines illustrate the initialization of the network weights with the learned parameters of the previous level.
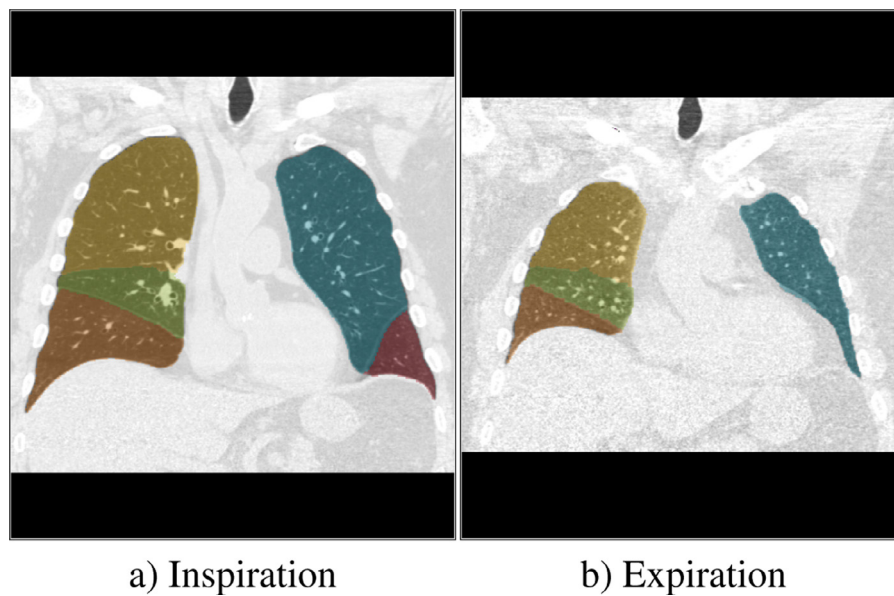

a) Inspiration   b) Expiration

Fig. 4. The Image shows a) an inspiration scan and b) and expiration scan of the lungs subdivided into ■ upper left lobe, ■ lower left lobe, ■ upper right lobe, ■ lower right lobe and ■ middle right lobe.

4.1.2. DIRLab Challenge

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

4.1.3. EMPIRE10 Challenge

The EMPIRE10 challenge (Murphy et al., 2011) consists of 30 scan pairs from six different categories: breathhold inspiration scan pairs, breathhold inspiration and expiration scan pairs, 4D data scan pairs, ovine data scan pairs, contrast-noncontrast scan pairs and artificially warped scan pairs. Further information on each category can be found in the challenge paper (Murphy et al., 2011). Each scan pair contains 100 annotated corresponding landmarks.

4.2. Preprocessing

In this work, we focus on non-rigid, non-linear deformations and for that reason we perform a linear prealignment of fixed and moving image as preprocessing. For all methods, the same preprocessing is used. We subsequently warp and resample the moving image on the field of view and resolution of the fixed image, which yields a pre-registered moving image $\hat{\mathcal{M}}$. Lung regions are automatically cropped for each CT and resized to volumes of dimension $192 \times 160 \times 192$ as the network input. However, although the deformation field is computed from low-resolution input, during inference, the output deformation field is up-sampled to the original image resolution using trilinear interpolation and the overall evaluation is performed at full resolution of the original images. We do not perform any further preprocessing like normalization on the images, because the CT images are already in a standardized range (Hounsfield units). On the training data, we use the keypoint detection algorithm described in Rühaak et al. (2017) to automatically compute keypoints inside the lung. These keypoints can be considered noisy labels with residual errors of 1-2mm
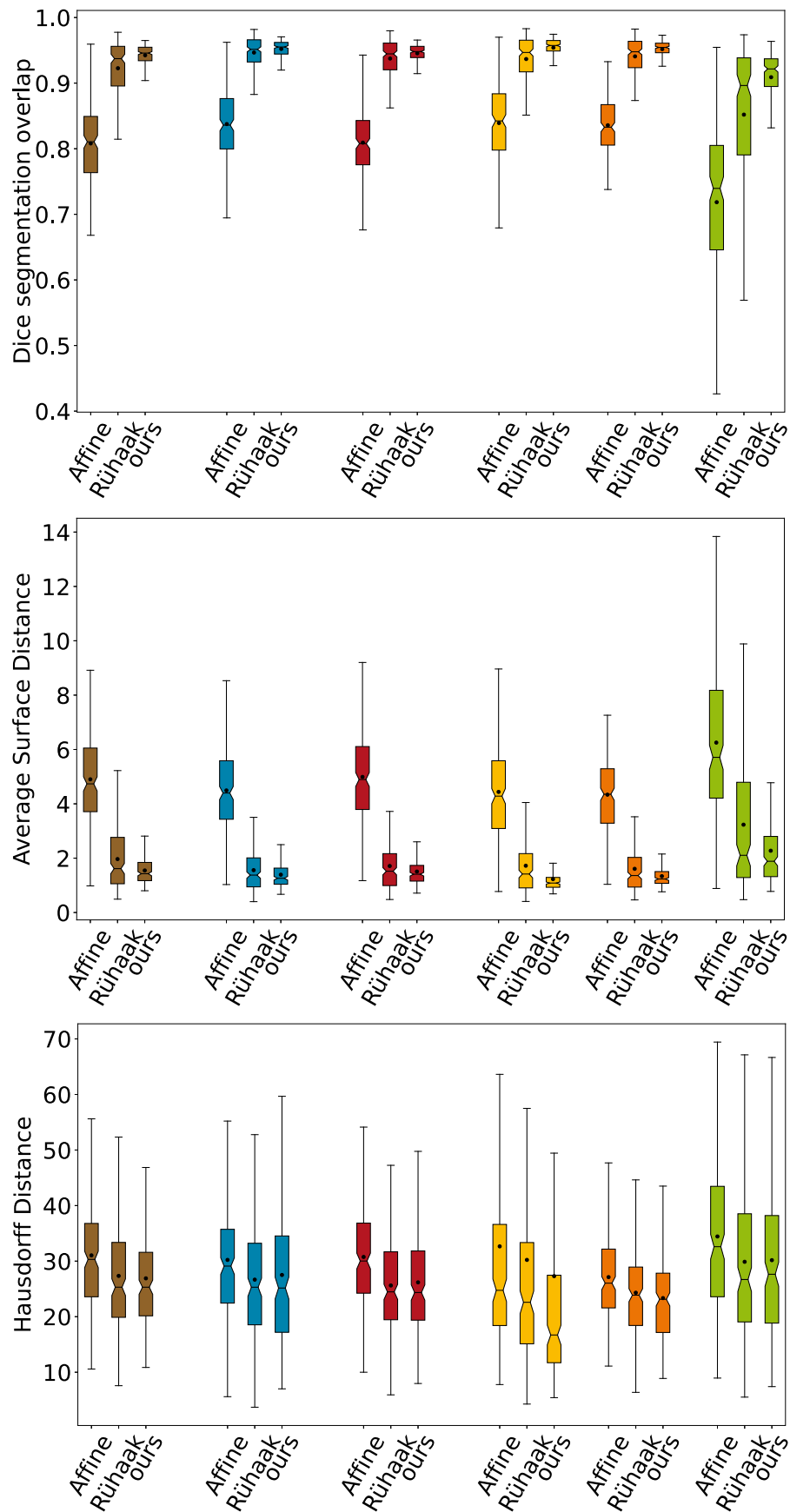
Fig. 5. Comparison of the Dice overlaps, average surface distance and Hausdorff distance for all test images and each anatomical label (■ average of all labels, ■ upper left lobe, ■ lower left lobe, ■ upper right lobe, ■ lower right lobe and ■ middle right lobe). For each one the distributions of the Dice coefficients after affine pre-alignment, after conventional method of Rühaak et al. (2017) and after our proposed registration are shown.

4.3. Implementation details

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

4.4. Accuracy

We evaluate our method by using the propagated lobe segmentation and the fixed lobe segmentation. If a deformation field represents accurate correspondences, the lobe segmentation of the fixed image $b_\mathcal{F}$ and the warped lobe segmentation of the moving image $b_\mathcal{M}(y)$ should overlap well. In contrast to a lung segmentation overlap, the lobe segmentation overlap provides information about inner lung structures. A good alignment of the lobes was shown to be indicative of good alignment of the fissures, which the evaluation of registration quality in Murphy et al. (2011) has shown to be indicative of the overall performance of different registration approaches.

We measure the overlap of the lobes with the Dice coefficient

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

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

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

where d is the surface distance

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

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

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

where

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

We compare our proposed method to the conventional approach of Rühaak et al. (2017) that is currently ranked first in the EMPIRE challenge (Murphy et al., 2011) (https://empire10.grand-challenge.org/Home/). This method performs a discrete keypoint detection and matching which are integrated into a dense continuous optimization framework. Additionally to the keypoint penalty, the method uses an NGF distance measure, curvature regularizer, a volume change penalty, and a mask alignment of the lung segmentations. Note that the lung segmentation has to be available for each pair of images to be registered. This is in contrast to our method, which also uses a boundary loss (Eq. 7), but this requires the masks to be only available during training, not during testing.

4.5. Robustness

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

4.6. Plausibility of the deformation field

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

4.7. Applicability

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

4.8. Transferability and comparison to state-of-the-art

To show the transferability of our method to other datasets and to compare our method to other registration methods, we apply our trained network as-is to the ten images pairs of the DIRLAB 4DCT. To evaluate the registration accuracy, the target registration error of the landmarks was computed.Moreover, we evaluate the impact of the dataset used to train the registration network. Therefore, we train the widely used Voxelmorph (Balakrishnan et al., 2019) framework using the COPDGene data. We adapt the public implementation slightly by choosing a higher regularization weight ($\lambda = 2$) to obtain smooth deformation fields Furthermore, we applied our trained model on the 30 scan pairs of the EMPIRE10 challenge and submitted the displacement fields to the organizers who performed the evaluation which includes a lung boundaries, fissures, landmarks and singularities (foldings).

4.9. Ablation study

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

5. Results

The results of our experiments on the COPDGene dataset are summarized in Table 1. We performed a one-sided Wilcoxon signed-rank test that show that the improvement to the method of Rühaak et al. (2017) is statistically significant for the Dice score, average surface distance (ASD) and Hausdorff distance(HD) and the runtime on the CPU. In the following subsections, we describe the results of each experiment in more detail.

5.1. Accuracy

Our proposed method achieves on average significant better Dice Scores than the conventional registration method (0.95 vs.

Table 1

Registration results of Rühaak et al. (2017) and our method on the COPDGene dataset. We performed a one-sited Wilcoxon signed-rank test to test if improvements to the method of Rühaak et al. (2017) are statistically significance. Significance levels are defined as * $p < 0.05$, ** $p < 0.01$ and *** $p < 0.001$.

	Rühaak et al. (2017)	ours
Dice	0.92 ± 0.05***	0.95 ± 0.03
ASD	1.97 ± 1.24mm***	1.72 ± 0.89mm
HD	27.24 ± 13.70mm*	26.84 ± 14.27mm
Dice30	0.86 ± 0.03 ***	0.93 ± 0.01
Foldings	0%	0.06%
Runtime CPU	160s ***	32s
Runtime GPU	-	0.75s
GPU memory	-	4GB

0.92) with a smaller standard deviation (0.026 vs. 0.046). Also for the symmetric average surface and the Hausdorff distance our method achieves better results (1.72 ± 0.89mm vs. 1.97 ± 1.24 and 26.84 ± 14.27 vs. 27.24 ± 13.70mm, respectively). The distribution of the Dice Scores, of the average surface distance, and of the Hausdorff Distance of both methods are shown in Fig. 5.

5.2. Robustness

On the 30% of cases with the lowest Dice Scores, our method achieves an average Dice Score of 0.93 ± 0.01 within a range of $[0.85, 0.94]$. Compared to the method of Rühaak et al. (2017) with an average Dice Score of 0.86 ± 0.03 within a range of $[0.78, 0.90]$, our method propagates the lobes more robustly.

5.3. Plausibility

For our proposed methods, on average fewer than 0.1% voxel positions of the deformation field showed a negative Jacobian determinant and therefore a folding. The full elimination of foldings as in the conventional registration methods of Rühaak et al. (2017) is not guaranteed. However, the percentage of foldings is within acceptable ranges. Fig. 6 shows four exemplary Jacobian determinant colormaps overlaid on the fixed image. The volume changes are smooth and mostly around 1 (yellow overlay). Due to large motion in the upper right case, some foldings (dark red overlay) occur on the left inferior border.

5.4. Applicability

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

5.5. Qualitative results

To illustrate the registration results, we show the difference images $\mathcal{F} - \mathcal{M}(y)$ of four exemplary cases in Fig. 8. In all cases, the respiratory motion was successfully recovered and most inner structures are well aligned. The first row shows one example of a more accurately registered image pairs in terms of the average Dice Score (after affine: 0.85, after: 0.96) and keypoint distance (after affine: 8.51mm, after: 0.99mm). The last row shows the worse case regarding the Dice Score. In this case, the average Dice Score

improved from 0.69 to 0.85 and the keypoint distance could be reduced from 13.57mm to 1.9mm, showing also for the cases with large deformations, our registration methods works robustly. Even with masking the distance measure only to the region inside the lung, the surrounding tissue is mostly well aligned. During training, the model learned to align edges and because no lung mask is given during inference, it also aligns edges outside the lung.

5.6. Ablation study

We provide an ablation study to further verify the efficiency of proposed components of our method. Results of this ablation experiment on the COPDGene data are presented in Table 2. The multi-level experiment shows that increasing the number of level from $L = 1$ to $L = 2$ and $L = 3$ results in a increasing Dice Score from 0.927 to 0.939 to 0.946, a decreasing TRE fron 3.95mm to 2.22mm to 2.00mm, and decreasing number of foldings from 0.1% to 0.09% to 0.06%. The mask alignment loss not only improve the alignment of the pulmonary lobes resulting in a higher Dice Score (0.93 vs 0.946) but also enhance the TRE from 2.16mm to 2.00mm. By integrating our volume change loss, the percentage of foldings can be reduced from 0.3% to 0.06%. Furthermore, it also improve the TRE from 2.16mm to 2.00mm. To further enhance the alignment of smaller structures as vessels and smaller airways, we incorporate keypoint correspondences into the loss function. This decrease the TRE from 4.59mm to 2.00mm. However, the percentage of foldings slightly increase from 0% to 0.06%. Fig. 7 shows a comparison of the target registration errors of the DIRLab 4DCT dataset of all compared settings and after affine registration and the initial errors.

5.7. Transferability and comparison to state-of-the-art

In Table 4, quantitative results on the DIRLAB 4DCT dataset of deep-learning-based and conventional registration methods are reported. On average the target registration error (TRE) of our method was 1.14 ± 0.76mm and is therefore better as the currently best deep-learning-based method of Fu et al. (2020). In cases 6, 8, and 10 which have a large initial landmark error, our method achieves much better registration results. Training Voxelmorph on the large COPDGene dataset results in a lower TRE than when trained by leave-one-out on the DIRLAB dataset (1.71mm vs 3.65mm). The best conventional registration method of Rühaak et al. (2017) has still a lower TRE, however, it needs about 5 minutes to compute the deformation field, whereas our method only needs less than a second. A detailed evaluation of all ten cases for different deep-learning-based registration methods is given in Table 5. On the EMPIRE10 challenge data, our method achieves a target registration error of 1.01mm on all cases and a TRE of 0.91mm if ovine data is excluded. A summary of the results is shown in Table 3 and a more detailed evaluation on the challenge website[1].

6. Discussion

This paper presents a coarse-to-fine multilevel framework for deep-learning-based image registration that can compensate for and handle large deformations using computing deformation fields on different scales. Our method shares many elements with the conventional registration method of Rühaak et al. (2017). We have identified key strategies of this method and successfully developed a deep-learning counterpart. The advantage of our deep learning approach is that the expensive annotation and detection of the

---

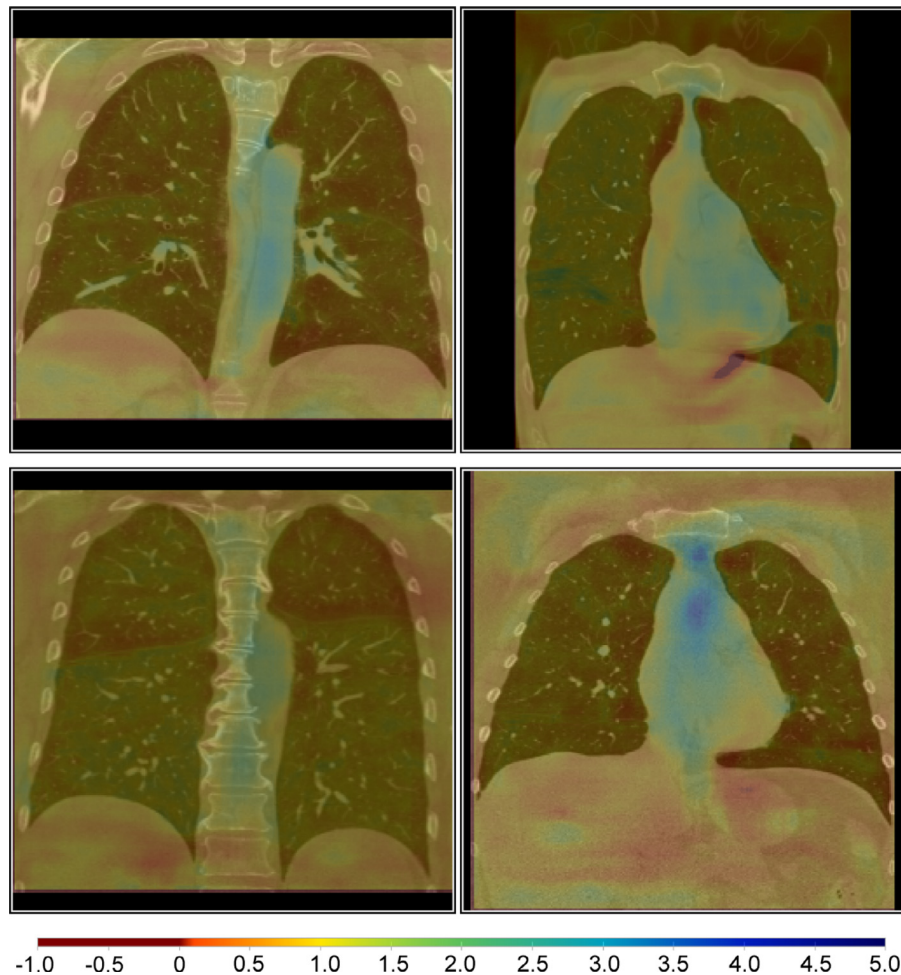[1] https://empire10.grand-challenge.org/mevis_virnet/

Fig. 6. Example Jacobian determinant colormaps overlaid on coronal slices of the fixed images. The volume changes are smooth and mostly around 1 (yellow overlay). Due to large motion in the upper right case, some foldings (dark red overlay) occur on the left inferior border. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2
Results of the ablation study. To demonstrate the impact of the each loss function term, each penalty weight was set to zero once while the remaining parameters were fixed to their empirically determined optimal values. The registration performance is evaluated using the Dice score, the target registration error of the keypoint (TRE KP), and the percentage of foldings on the COPDGene dataset. Moreover the target registration error on the DIRLab dataset is compared. We performed a one-sided Wilcoxon signed-rank test to test if improvements to all other settings are statistically significant. We used a Bonferroni correction due to multiples testing. Significance levels are defined as * $p < 0.05$, ** $p < 0.01$ and *** $p < 0.001$.

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

Table 3
Results of the EMPIRE10 challenge for the method of Rühaak et al. (2017) and our proposed method. The average score over all 30 cases for the lung boundaries, fissure alignment, landmark error and singularities is shown. Detailed results can be found on the challenge website.

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

lobe masks and keypoints is only necessary as training data. This important knowledge is then embedded in our model and therefore the inference is cheap and fast.

We employ a Gaussian-pyramid-based multilevel framework that can solve the image registration optimization in a coarse-to-fine fashion. To prevent foldings of the deformation field and restrict the determinant of the Jacobian to physiologically meaningful values, we combine the curvature regularizer with a volume change penalty in the loss function. Furthermore, we also integrate weak keypoint correspondences into the loss function to focus more on the alignment of smaller structures. The keypoints are computed automatically and can be considered as noisy labels with residual errors of $1 - 2mm$. However, we showed that the use of these noisy labels is nevertheless advantageous and leads to a better alignment of vessels and smaller airways and therefore also results in a better target registration error on the DIRLab dataset.
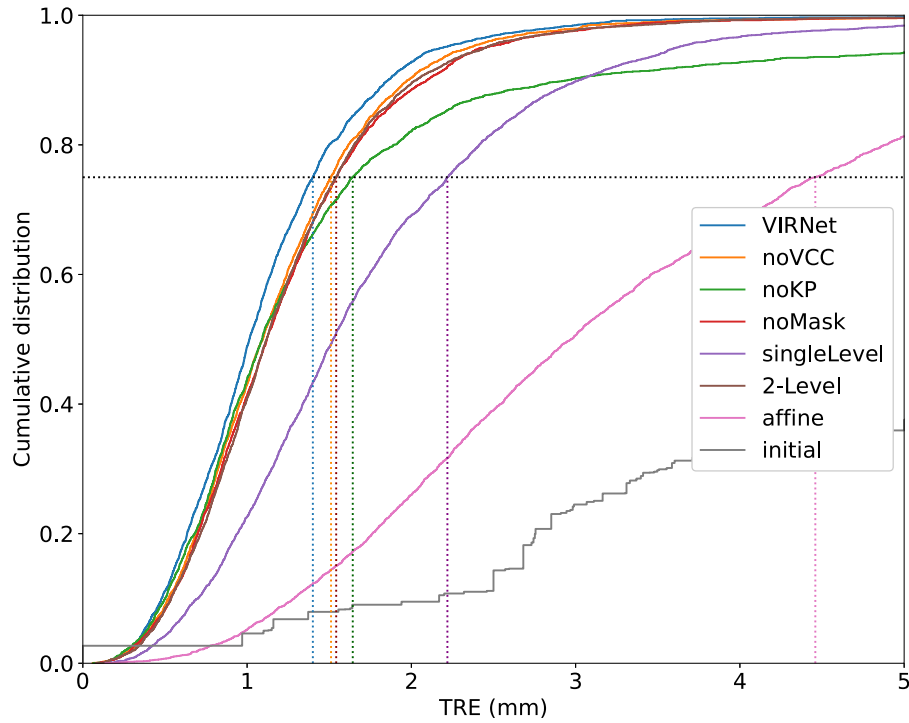
Fig. 7. Cumulative distribution of target registration error (TRE) in millimeters for all variations of our loss function, after affine registration and initially on all landmark pairs of the DIRLAB 4DCT dataset. In addition, the dotted lines visualize the 75th percentiles of the TRE, which are 1.40mm (our VIRNet), 1.51mm (noVCC), 1.64mm (noKP), 1.54mm (noMask), 2.22mm (singleLevel), 1.54mm (twoLevel), 4.46mm (affine), 12.55mm (initial).

Table 4
Target registration error values for different conventional and deep learning-based methods on DIRLAB 4D-CT dataset. All results were extracted from the original papers, besides Voxelmorph* which was reported in Hansen and Heinrich (2020) and Voxelmorph** which we trained on the COPDGene data. Since the runtime was not measured with the same hardware, it cannot directly be compared. However, it gives an impression of the speed.

	Method	mean TRE (mm)	Foldings	Runtime
	initial	8.46 ± 6.58	-	-
Conventional	Schmidt-Richberg et al. (2011)	1.38 ± 0.87	-	83min
	Heinrich et al. (2012)	1.6 ± 1.7	0%	20min
	Heinrich et al. (2013)	1.43 ± 1.3	-	7.97min
	Berendsen et al. (2014)	1.36 ± 0.99	0%	-
	Rühaak et al. (2017)	0.94 ± 1.06	0%	5min
	Sentker et al. (2018)	2.5 ± 1.16	-	few seconds
	Balakrishnan et al. (2019)*	3.65 ± 2.47	-	-
	Balakrishnan et al. (2019)**	1.71 ± 2.86	-	-
	de Vos et al. (2019)	2.64 ± 4.32	-	0.63s
Deep Learning	Eppenhof et al. (2019)	2.43 ± 1.81	0.42%	0.56s
	Hering et al. (2019a)	2.19 ± 1.62	-	-
	Hansen and Heinrich (2020)	1.97 ± 1.42	-	-
	Jiang et al. (2020)	1.66 ± 1.44	< 0.1%	1.4s
	Fu et al. (2020)	1.59 ± 1.58	-	1min
	Hansen and Heinrich (2021)	1.39 ± 1.29	0.02%	2s
	ours	1.14 ± 0.76	< 0.0005%	0.75s

We validated our framework on the challenging task of large motion inspiration-to-expiration registration using image data from the multi-center COPDGene study. To assess the accuracy of our network, we performed an extensive evaluation of 200 pulmonary CT scan pairs from the large-scale COPDGene study and demonstrated that our method can perform accurate registration between two affine pre-aligned images. Especially for the task of lobe propagation, we could show that our method performs better than conventional approaches. It achieves higher Dice scores and lower surface and Hausdorff distances (0.95, 1.72 mm, and 26.8 mm) compared to conventional registration (0.92, 1.97 mm, and 27.2 mm, respectively). This better performance can be explained by the use of the mask-alignment loss. As

demonstrated in previous studies (e.g. Balakrishnan et al. (2019); Hering et al. (2019c)), the combination of the complementary strength of global semantic information (weakly-supervised learning with segmentation labels) and local distance metrics improves the registration performance during inference. In contrast to conventional registration methods, such additional information only needs to be available in the training dataset.

Furthermore, we have evaluated the proposed method using the DIRLab and EMPIRE10 dataset and showed that we achieve excellent TRE of 1.14 mm and 1.01 mm, respectively. Note that our network was not trained on those datasets. This is strong evidence that our network can generalize well. Although previous works (e.g. de Vos et al. (2019); Eppenhof et al. (2019); Jiang et al. (2020);

Table 5
Mean (standard deviation) of the registration error in mm determined on DIR-Lab 4D-CT data for several deep-learning-based registration methods: Eppenhof et al. (2019), de Vos et al. (2019), Hering et al. (2019a), Fu et al. (2020), Hansen and Heinrich (2021) (standard deviations were reported directly by the authors and not included in the their paper) and Balakrishnan et al. (2019) (trained on the COPDGene data). Asterisk indicates that the FineNet was performed twice for this case and method.

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

Sentker et al. (2018)) contribute much to improving the registration accuracy, there is still a misalignment of smaller structures, which leads to a high TRE. To focus more on the alignment of vessels, Fu et al. (2020) introduced a preprocessing step to enhance the vessels in the input images by segmenting vascular structures and increasing the intensity value inside the vessel mask. In their paper, they demonstrated the efficiency of this preprocessing step. Since this step is performed on the input images, it is also required during application. To avoid this problem and thus not increase the execution time, we integrate additional information on smaller structures using the keypoint loss. The advantage of this procedure is that the keypoints, as well as the masks of the boundary loss, are only needed during training. Nevertheless, the best conventional registration methods still achieve lower TRE than our method. One reason for this might be that convectional registration methods mostly work on the original image data. In contrast, for the deep-learning approaches, the input images have to be downsampled due to memory restrictions on the GPU. Especially for smaller structures and small errors (we are speaking about a TRE difference of 0.2-0.4mm), it is easily imaginable that this resolution is not high enough. Moreover, the training data used also influence the performance. Our network was trained on inspiration-expiration scan pairs from humans. In the EMPIRE10 dataset, a variety of lung registration tasks including ovine lung registration has to be performed. Although our method does not register the ovine data perfectly, we achieve a TRE of 1.69 mm on the ovine data which shows that our method is capable of generalizing well. We would assume that with a wider variety in training data, the performance of deep-learning-based registration methods can further improve. We showed this effect when training the Voxelmorph network. By using the larger COPDGene dataset to train the Voxelmorph network, the target registration error on the DIRLab dataset improved from 3.65 mm to 1.71 mm compared to a leave-one-out training on the DIRLab dataset. This illustrates the large impact of the training dataset. Since Voxelmorph and our framework are very similar, this experiment also shows that the addition of more loss functions and a multilevel approach is beneficial.

Besides accurately transferring anatomical annotations, medical image registration should also provide plausible transformations and therefore should not generate deformations with foldings. In conventional registration methods, this is achieved by using a regularizer in the cost function. Recently deep-learning-based methods like Eppenhof et al. (2019) and Jiang et al. (2020) also integrated a regularizer into their loss functions to enforce a smooth deformation field resulting in an acceptable amount of foldings (0.42% and 0.1% of foldings). In our work, we additionally use a volume change control which penalizes occurring foldings more

severely than the regularizer does, resulting in on average fewer than 0.1% and 0.0005% voxel positions in the deformation field with folding on the COPDGene and DIRLab dataset, respectively. Without the volume change control penalty, the deformation fields produced by our method show on average 0.30% of voxel positions with foldings, which is comparable to the values of other deep learning registration methods. This shows that the addition of the volume change control mitigated the occurrence of foldings. The higher number of foldings on the COPDGene dataset can be explained by the noise difference between the expiration and inspiration scan due to different doses during acquisition (see Fig 8 for some example images). The full elimination of foldings as in some conventional registration methods is not guaranteed. Another way to reduce the number of foldings was presented in the works of Dalca et al. (2018), Krebs et al. (2019), and Qiu et al. (2021) who are using the scaling and squaring algorithm (Arsigny et al., 2006) to integrate the predicted stationary velocity field. With a sufficient number of integration steps, these methods should theoretically guarantee diffeomorphic transformation. However, in the presented works they reported "nearly no non-negative Jacobian voxels" (Dalca et al., 2018) and 0.023% to 0.151% of voxels with a negative Jacobian determinant (Qiu et al., 2021). As discussed in (Qiu et al., 2021), this has two major factors. First, the velocity field could be not sufficiently smooth. This can be solved by increasing the regularization weight. However, this often yields a drop in the registration accuracy. Secondly, the number of chosen integration steps was too small. Increasing this can reduce the number of foldings which occur but increase the computational cost as well. In summary, the scaling and squaring approach and the volume-change-control penalty presented achieve similar results preventing foldings. Besides, our approach regulates volume changes.

In our experiments, we focused on the complex task of CT lung registration, as the registration results can be evaluated more accurately than only with an overlap of a larger structure. However, our method could also be trained for a different task or on a different modality. Except for keypoint detection, no component is lung-specific and the keypoint loss can be used with landmarks in different organs.

In future studies, we will investigate the impact of instance optimization to fine-tune the deformation field for those image pairs for which the registration result is not yet satisfactory.

7. Conclusion

This paper presents a deep-learning-based registration approach for deformable image registration, targeting in particular the challenging task of lung registration. We introduce a keypoint matching term and a volume change penalty to increase
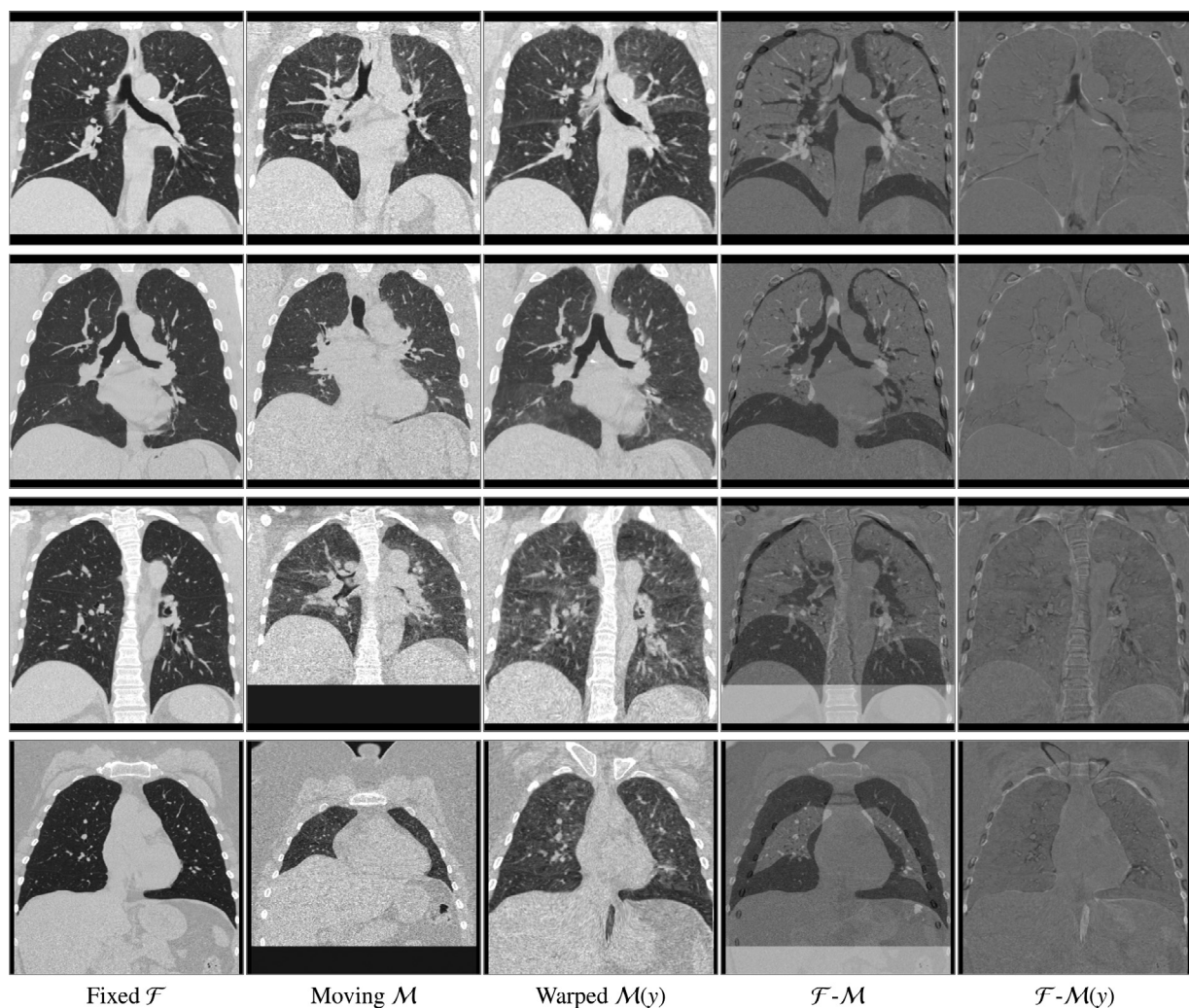
Fixed $\mathcal{F}$ Moving $\mathcal{M}$ Warped $\mathcal{M}(y)$ $\mathcal{F}$-$\mathcal{M}$ $\mathcal{F}$-$\mathcal{M}(y)$

Fig. 8. Example coronal slices extracted from four exemplary cases. Input images $\mathcal{F}$ and $\mathcal{M}$, the warped moving image $\mathcal{M}(y)$, the difference image $\mathcal{F} - \mathcal{M}$ (fourth column) and the difference image $\mathcal{F} - \mathcal{M}(y)$ after registration with the proposed method (fifth column). In all cases the respiratory motion was successfully recovered and most inner structures are well aligned. Due to altered density of lung tissue during breathing, intensity changes occur and therefore higher values in the difference images are reached without registration errors.

the alignment of smaller structures and to prevent foldings and restrict the deformation field to physiologically meaningful values. Our multi-level registration framework equipped with these components achieves state-of-the-art registration accuracy on the COPDGene and DIRLab datasets with a very short execution time.

Declaration of Competing Interest

There's no financial/personal interest or belief that could affect the objectivity of the submitted research results. No conflict of interests exist.

CRediT authorship contribution statement

Alessa Hering: Conceptualization, Methodology, Software, Validation, Investigation, Writing - original draft, Writing - review & editing, Visualization, Funding acquisition. **Stephanie Häger:** Visualization, Writing - review & editing. **Jan Moltz:** Methodology, Writing - review & editing. **Nikolas Lessmann:** Supervision, Writing - review & editing. **Stefan Heldmann:** Conceptualization, Resources, Writing - review & editing, Supervision, Funding acquisition. **Bram van Ginneken:** Conceptualization, Resources, Supervision, Writing - review & editing.

Acknowledgements

References

Arsigny, V., Commowick, O., Pennec, X., Ayache, N., 2006. A log-euclidean framework for statistics on diffeomorphisms. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 924–931.

Bajcsy, R., Kovačič, S., 1989. Multiresolution elastic matching. Comput. Vis. Graphic. Image Process. 46 (1), 1–21.

Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V., 2018. An unsupervised learning model for deformable medical image registration. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9252–9260.

Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V., 2019. Voxelmorph: a learning framework for deformable medical image registration. IEEE Trans. Med. Image..

Berendsen, F.F., Kotte, A.N., Viergever, M.A., Pluim, J.P., 2014. Registration of organs with sliding interfaces and changing topologies. In: Medical Imaging 2014: Image Processing, 9034. International Society for Optics and Photonics, p. 90340E.

Boyd, S., Boyd, S.P., Vandenberghe, L., 2004. Convex optimization. Cambridge university press.

Castillo, R., Castillo, E., Fuentes, D., Ahmad, M., Wood, A.M., Ludwig, M.S., Guerrero, T., 2013. A reference dataset for deformable image registration spatial accuracy evaluation using the copdgene study archive. Phys. Med. Biol. 58 (9), 2861.

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. Phys. Med. Biol. 54 (7), 1849.

Dalca, A.V., Balakrishnan, G., Guttag, J., Sabuncu, M.R., 2018. Unsupervised learning for fast probabilistic diffeomorphic registration. arXiv preprint arXiv:1805.04605.

Ehrhardt, J., Werner, R., Schmidt-Richberg, A., Handels, H., 2010. Automatic landmark detection and non-linear landmark-and surface-based registration of lung ct images. Med. Image Anal. Clinic-A Grand Challenge, MICCAI 2010, 165–174.

Eppenhof, K.A., Lafarge, M.W., Moeskops, P., Veta, M., Pluim, J.P., 2018. Deformable image registration using convolutional neural networks. In: Medical Imaging 2018: Image Processing, 10574. International Society for Optics and Photonics, p. 105740S.

Eppenhof, K.A., Lafarge, M.W., Veta, M., Pluim, J.P., 2019. Progressively trained convolutional neural networks for deformable image registration. IEEE Trans. Med. Imag..

Eppenhof, K.A., Pluim, J.P., 2018. Pulmonary ct registration through supervised learning with convolutional neural networks. IEEE Trans. Med. Imag..

Ferrante, E., Oktay, O., Glocker, B., Milone, D.H., 2018. On the adaptability of unsupervised cnn-based deformable image registration to unseen image domains. In: International Workshop on Machine Learning in Medical Imaging. Springer, pp. 294–302.

Fischer, B., Modersitzki, J., 2003. Combination of automatic non-rigid and landmark-based registration: the best of both worlds. In: Medical Imaging 2003: Image Processing, 5032. International Society for Optics and Photonics, pp. 1037–1048.

Fischer, B., Modersitzki, J., 2003. Curvature based image registration. J. Math. Imag. Vis. 18 (1), 81–85.

Fu, Y., Lei, Y., Wang, T., Higgins, K., Bradley, J., Curran, W., Liu, T., Yang, X., 2020. Lungregnet: an unsupervised deformable image registration method for 4d-ct lung.. Med. Phys. 47 (4), 1763.

Glorot, X., Bengio, Y., 2010. Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, pp. 249–256.

Haber, E., Modersitzki, J., 2004. Cofir: coarse and fine image registration. Real-Time PDE-Constrained Optimization. Citeseer.

Haber, E., Modersitzki, J., 2006. Intensity gradient based registration and fusion of multi-modal images. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2006, 3216, pp. 591–598.

Hansen, L., Heinrich, M.P., 2020. Tackling the problem of large deformations in deep learning based medical image registration using displacement embeddings. arXiv preprint arXiv:2005.13338.

Hansen, L., Heinrich, M.P., 2021. Graphregnet: deep graph regularisation networks on sparse keypoints for dense registration of 3d lung cts. IEEE Trans. Med. Imag..

Heinrich, M.P., 2019. Closing the gap between deep and conventional image registration using probabilistic dense displacement networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 50–58.

Heinrich, M.P., Jenkinson, M., Brady, M., Schnabel, J.A., 2012. Globally optimal deformable registration on a minimum spanning tree using dense displacement sampling. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 115–122.

Heinrich, M.P., Jenkinson, M., Brady, M., Schnabel, J.A., 2013. Mrf-based deformable registration and ventilation estimation of lung ct. IEEE Trans. Med. Imag. 32 (7), 1239–1248.

Hering, A., van Ginneken, B., Heldmann, S., 2019. mlvirnet: multilevel variational image registration network. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 257–265.

Hering, A., Heldmann, S., 2019. Unsupervised learning for large motion thoracic ct follow-up registration. In: Angelini, E.D., Landman, B.A. (Eds.), SPIE Medical Imaging 2019: Image Processing, 10949.

Hering, A., Kuckertz, S., Heldmann, S., Heinrich, M.P., 2019. Enhancing label–driven deep deformable image registration with local distance metrics for state-of-the-art cardiac motion tracking. In: Bildverarbeitung für die Medizin 2019 - Algorithmen - Systeme - Anwendungen. Proceedings des Workshops vom 17. bis 19. März 2019 in Lübeck. Springer, pp. 309–314.

Hering, A., Kuckertz, S., Heldmann, S., Heinrich, M.P., 2019. Memory-efficient 2.5 d convolutional transformer networks for multi-modal deformable registration with weak label supervision applied to whole-heart ct and mri scans. Int. J. Comput. Assist. Radiol. Surg. 14 (11), 1901–1912.

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

Hu, Y., Modat, M., Gibson, E., Li, W., Ghavami, N., Bonmati, E., Wang, G., Bandula, S., Moore, C.M., Emberton, M., et al., 2018. Weakly-supervised convolutional neural networks for multimodal image registration. Med. Image Anal. 49, 1–13.

Jaderberg, M., Simonyan, K., Zisserman, A., et al., 2015. Spatial transformer networks. In: Advances in Neural Information Processing Systems, pp. 2017–2025.

Jiang, Z., Yin, F.-F., Ge, Y., Ren, L., 2020. A multi-scale framework with unsupervised joint training of convolutional neural networks for pulmonary deformable image registration. Phys. Med. Biol. 65 (1), 015011.

Kabus, S., Lorenz, C., 2010. Fast elastic image registration. Med. Image Anal. Clin. 81–89.

Kervadec, H., Dolz, J., Wang, S., Granger, E., Ayed, I.B., 2020. Bounding boxes for weakly supervised segmentation: global constraints get close to full supervision. arXiv preprint arXiv:2004.06816.

Kervadec, H., Dolz, J., Yuan, J., Desrosiers, C., Granger, E., Ayed, I.B., 2019. Constrained deep networks: lagrangian optimization via log-barrier extensions. arXiv:1904.04205.

König, L., Rühaak, J., Derksen, A., Lellmann, J., 2018. A matrix-free approach to parallel and memory-efficient deformable image registration. SIAM J. Sci. Comput. 40 (3), B858–B888.

Krebs, J., Delingette, H., Mailhé, B., Ayache, N., Mansi, T., 2019. Learning a probabilistic model for diffeomorphic registration. IEEE Trans. Med. Imag. 38 (9), 2165–2176.

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: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 344–352.

Kuckertz, S., Papenberg, N., Honegger, J., Morgas, T., Haas, B., Heldmann, S., 2020. Deep learning based ct-cbct image registration for adaptive radio therapy. In: Medical Imaging 2020: Image Processing, 11313. International Society for Optics and Photonics, p. 113130Q.

Li, H., Fan, Y., 2018. Non-rigid image registration using self-supervised fully convolutional networks without training data. arXiv preprint arXiv:1801.04012.

Maintz, J., Viergever, M.A., 1998. A survey of medical image registration. Med. Image. Anal. 2 (1), 1–36.

Modersitzki, J., 2004. Numerical methods for image registration. Oxford University Press on Demand.

Modersitzki, J., 2009. FAIR: Flexible algorithms for image registration, 6. SIAM.

Mok, T.C., Chung, A.C., 2020. Large deformation diffeomorphic image registration with laplacian pyramid networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 211–221.

Murphy, K., Van Ginneken, B., Reinhardt, J.M., Kabus, S., Ding, K., Deng, X., Cao, K., Du, K., Christensen, G.E., Garcia, V., et al., 2011. Evaluation of registration methods on thoracic ct: the empire10 challenge. IEEE Trans. Med. Imag. 30 (11), 1901–1920.

Papenberg, N., Olesch, J., Lange, T., Schlag, P.M., Fischer, B., 2009. Landmark Constrained Non-parametric Image Registration with Isotropic Tolerances. In: Bildverarbeitung für die Medizin 2009. Springer, pp. 122–126.

Polzin, T., Rühaak, J., Werner, R., Strehlow, J., Heldmann, S., Handels, H., Modersitzki, J., 2013. Combining automatic landmark detection and variational methods for lung ct registration. In: Fifth International Workshop on Pulmonary Image Analysis, pp. 85–96.

Qiu, H., Qin, C., Schuh, A., Hammernik, K., Rueckert, D., 2021. Learning diffeomorphic and modality-invariant registration using b-splines.

Regan, E.A., Hokanson, J.E., Murphy, J.R., Make, B., Lynch, D.A., Beaty, T.H., Curran-Everett, D., Silverman, E.K., Crapo, J.D., 2011. Genetic epidemiology of copd (copdgene) study design. COPD: J. Chronic Obstruct. Pulmon. Diseas. 7 (1), 32–43.

Rohé, M.-M., Datar, M., Heimann, T., Sermesant, M., Pennec, X., 2017. Svf-net: Learning deformable image registration using shape matching. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 266–274.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. Springer, pp. 234–241.

Rühaak, J., Polzin, T., Heldmann, S., Simpson, I.J., Handels, H., Modersitzki, J., Heinrich, M.P., 2017. Estimation of large motion in lung ct by integrating regularized keypoint correspondences into dense deformable registration. IEEE Trans. Med. Imag. 36 (8), 1746–1757.

Schmidt-Richberg, A., Werner, R., Ehrhardt, J., Wolf, J.-C., Handels, H., 2011. Landmark-driven parameter optimization for non-linear image registration. In: Medical Imaging 2011: Image Processing, 7962. International Society for Optics and Photonics, p. 79620T.

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., et al., 2001. A generic framework for non-rigid registration based on non-uniform multi-level free–form deformations. In: MICCAI 2001. Springer, pp. 573–581.

Sentker, T., Madesta, F., Werner, R., 2018. Gdl-fire 4d: deep learning-based fast 4d ct image registration. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 765–773.

Sokooti, H., de Vos, B., Berendsen, F., Lelieveldt, B.P., Išgum, I., Staring, M., 2017. Non-rigid image registration using multi-scale 3d convolutional neural networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 232–239.

Sotiras, A., Davatzikos, C., Paragios, N., 2013. Deformable medical image registration: a survey. Med. Imag. IEEE Trans. 32 (7), 1153–1190.

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

de Vos, B.D., Berendsen, F.F., Viergever, M.A., Staring, M., Išgum, 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. Springer, pp. 204–212.

Yang, X., Kwitt, R., Styner, M., Niethammer, M., 2017. Fast predictive multimodal image registration. In: 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017), pp. 858–862.

Yang, X., Kwitt, R., Styner, M., Niethammer, M., 2017. Quicksilver: fast predictive image registration–a deep learning approach. Neuroimage 158, 378–396.