IEEE Transactions on Medical Imaging

GraphRegNet: Deep Graph Regularization Networks on Sparse Keypoints for Dense Registration of 3D Lung CTs

Journal:	*IEEE Transactions on Medical Imaging*
Manuscript ID	TMI-2020-2318
Manuscript Type:	Regular Paper
Date Submitted by the Author:	27-Dec-2020
Complete List of Authors:	Hansen, Lasse; Universität zu Lübeck, Institute of Medical Informatics Heinrich, Mattias Paul; University of Luebeck, Institute of Medical Informatics
Keywords:	X-ray imaging and computed tomography < Imaging modalities, Lung < Object of interest, Machine learning < General methodology, Neural network < General methodology, Registration < General methodology

IEEE TRANSACTIONS ON MEDICAL IMAGING, VOL. XX, NO. XX, XXXX 2020 1

GraphRegNet: Deep Graph Regularization Networks on Sparse Keypoints for Dense Registration of 3D Lung CTs

Lasse Hansen, and Mattias P. Heinrich

***Abstract*— In the last two years learning-based methods have started to show encouraging results in different supervised and unsupervised medical image registration tasks. Deep neural networks enable (near) real time applications through fast inference times and have tremendous potential for increased registration accuracies by task-specific learning. However, estimation of large 3D deformations, for example present in inhale to exhale lung CT or interpatient abdominal MRI registration, is still a major challenge for the widely adopted U-Net-like network architectures. Even when using multi-level strategies, current state-of-the-art DL registration results do not yet reach the high accuracy of conventional frameworks. To overcome the problem of large deformations for deep learning approaches, in this work, we present GraphRegNet, a sparse keypoint-based geometric network for dense deformable medical image registration. Similar to the successful 2D optical flow estimation of FlowNet or PWC-Net we leverage discrete dense displacement maps to facilitate the registration process. In order to cope with enormously increasing memory requirements when working with displacement maps in 3D medical volumes and to obtain a well-regularized and accurate deformation field we 1) formulate the registration task as the prediction of displacement vectors on a sparse irregular grid of distinctive keypoints and 2) introduce our efficient GraphRegNet for displacement regularization, a combination of convolutional and graph neural network layers in a unified architecture. In our experiments on exhale to inhale lung CT registration we demonstrate substantial improvements (TRE below 1.4 mm) over other deep learning methods. Our code is publicly available at `https://github.com/multimodallearning/graphregnet`.**

***Index Terms*— Deformable Registration, Graph Learning, Thoracic CT**

I. Introduction

THE automated analysis of multiple thoracic CT scans plays an important role for diagnosis and treatment planning of pulmonary diseases including, lung cancer [1], COPD [2], emphysema or pneumonia [3]. Comparing normal dose inspiration CT with (ultra-)low dose expiration scans helps to reveal subtle local differences in air flow, important for COPD and asthma diagnosis and treatment, that are otherwise only measurable globally or with highly complex functional imaging modalities (xenon CT or helium MRI [4]). Accurate deformable intra-patient registration between different respiratory levels is vital for localised ventilation measurements [5]. In order to improve accuracy and robustness as well as wide adoption of thoracic image registration in clinical practice, deep machine learning could play an important role. Recently, multi-resolution pyramid networks [6] showed great success at a multi-task registration challenge [7] (e.g. ranking 1st for large deformation estimation in abdominal CT), however, as other state-of-the-art DL-based registration algorithms it fails to produce acceptable registration accuracy for large motion estimation for the task of inspiration to exhale CT and only provides reasonable robustness for shallow breathing. As discussed below, graphical optimisation models have helped to overcome these challenging for conventional registration approaches and graphical deep learning methods are destined to become the key element in further improving the applicability of learning based 3D medical image registration within the thorax.

This work was partially supported by the German Federal Ministry for Economic Affairs and Energy as part of the AI space for intelligence healthcare systems (KI SIGS) consortium. Grant number: 01MK20012B. L. Hansen and M. P. Heinrich are with the Institute of Medical Informatics, Universität zu Lübeck, 23562 Lübeck, Germany. (e-mail: {hansen, heinrich}@imi.uni-luebeck.de).

A. Related Work

Discrete graphical optimisation models or Markov Random Fields (MRF), that include graph cuts [8], message passing [9], [10] and mean-field inference, are able to solve complex global regularisation tasks given a suitable and densely sampled unary cost term. Yet, for 3D medical image registration the degrees of freedom are enormous with thousands of deformation control points (nodes in graph) and up to tens of thousands of potential displacements (labels in MRF solution space, cf. [11]). Hence, besides requiring additional fine-tuning stages for subvoxel accuracy [12] MRF-based registration tends to be slow and memory extensive. This may also prevent the use of complex graphical models for 3D registration in end-to-end trainable geometric learning models. So far graph models have been mainly limited to post-hoc regularisation of segmentation predictions (CRF as RNN) [13], [14], which used approximate mean-field inference of conditional random fields (CRF). In [15] an architecture with unary and pairwise convolutional neural networks (CNNs) was proposed for two-view stereo estimation. Our own prior work [16], [17], addressed DL-based 3D registration with a discretised displacement space and differentiable CRF regularisation, but was limited to coarsely and robustly aligning abdominal organs across patients, which is very different to the detail required for lung

vessel alignment. In the context of regional object detection, an end-to-end trainable parts-based model was designed with CNNs in [18], where the MRF inference based on distance-transform regularisation was unrolled. Subsequent research has demonstrated the ability to integrate probabilistic graphical models into a deep network [19] and to infer relations of temporal time points in video analysis.

Most recent DL approaches that tackle intra-patient CT lung registration rely on multi-resolution, cascaded U-Net architectures [20], [21] that have a large number of trainable parameters, but still fall short of the accuracy of efficient conventional registration techniques [22]. Recurrent networks in combination with parametrized transformation models are investigated in [23]. In computer vision, using a discretised displacement space, the so-called correlation layer, for DL-based optical flow estimation is currently considered state-of-the-art for complex and large motion estimation, cf. [24], [25]. However, as discussed in [16] they are not easy to adapt to 3D motion due to processing all displacements as flattened feature channels, which leads to a huge number of trainable parameters and severe memory limitation in medical scans.

For a comprehensive introduction into current geometric deep learning using graph convolutional networks (GCNs), we refer to [26]. Most relevant in our context are the PointNet [27], which ignores the local connectivity of point clouds for simplicity, the diffusion based graph convolution network that is restricted to isotropic graph filters [28] and newer edge weighted graph networks [29]. Graph attention networks introduced in [30] are another related and promising research direction based on the attention mechanism that was popularised in machine translation and medical image analysis [31]. Both latter methods achieve a more dynamic information propagation that can be considered close to the expressiveness of MRF message passing by learning a function that predicts pairwise edge weights based on previous features of both considered nodes. A number of 3D vision applications have been recently addressed using graph convolutions, in particular semantic point segmentation [32], [33].

B. Contribution

In this work, we substantially extend our short paper submission at MIDL 2020, a proof of concept demonstrating that the estimation of large deformations is possible with high accuracy from compact PCA displacement embeddings [34]. Here, we focus on a novel concept of learning informative spatial relations on a sparse irregular grid of distinctive keypoints for the prediction of dense displacement fields in an unsupervised setting and therefore make three important contributions:

1) Propose a lightweight network architecture particular designed for learning-based medical image registration, called GraphRegNet, composed of convolutional and graph neural network layers that act on the discrete displacement space and spatial dimensions, respectively, to predict regularized displacement vectors on a sparsely sampled irregular grid.
2) In contrast to our preliminary work in [34], learn a low dimensional displacement embedding in an unsupervised fashion, which enables compressed message passing (using GCNs) on the inherent geometric structure of the generated keypoint graph.
3) Formulate an unsupervised *dense* warping loss on the fixed and moving image that enables gradient flow through the predicted displacements (by integral regression) at the *sparse* keypoint locations (”inverse gridsample”).

Experiments on the challenging task of exhale to inhale CT lung registration suggest that our unsupervised learning approach is able to extract similar (and even more accurate) information from the keypoint graph as competing methods that use exact graphical message passing. A series of ablation studies justify our different architectural choices for the GraphRegNet and we advance the state of the art for deep learning registration methods on the two widely used DIR-Lab 4D CT [35] and COPDgene [36] datasets to average TREs of 1.39 mm and 1.34 mm, respectively.

C. Paper Outline

Details on the individual steps of our proposed deep-learning based framework for exhale to inhale CT lung registration are given in Section II. These include first of all the preprocessing of the raw lung CT scans, e.g. to account for different volume sizes, and the generation of a distinctive keypoint graph from the fixed (inhale) image. Next, we outline the computation of a set of discrete dense displacement maps from descriptive image features (here: MIND-SSC features [37]). We then describe the estimation of the final dense deformation field from the individual displacement maps in our methodological contributions, the GraphRegNet and the unsupervised sparse warping loss. In Section III our registration framework is extensively evaluated. In ablation studies different network architecture choices are examined and our method is compared to other recent deep learning registration approaches. A thorough discussion of the main findings of this work and a final conclusion follow in Section V.

II. Methods

Let I_F denote the fixed and I_M the moving image. In this work we focus on voluminous and single channel lung CT images, i.e. $I_F, I_M : \mathbb{R}^3 \rightarrow \mathbb{R}$. The fixed and moving image are defined as the inhale and exhale scan of the paired setting, respectively. The aim of the intrapatient registration process is to estimate a displacement field $D : \mathbb{R}^3 \rightarrow \mathbb{R}^3$ that best aligns the inhale and exhale image. Figure 1 shows an overview of our proposed deep learning registration framework. Individual steps and details of our method are described in the next sections.

A. Preprocessing

In a first step, inhale and exhale images are affinely aligned. For this purpose, lung segmentations are computed on all images using thresholding and morphological filters. Images are cropped to their lung mask bounding boxes (BB) (+ a

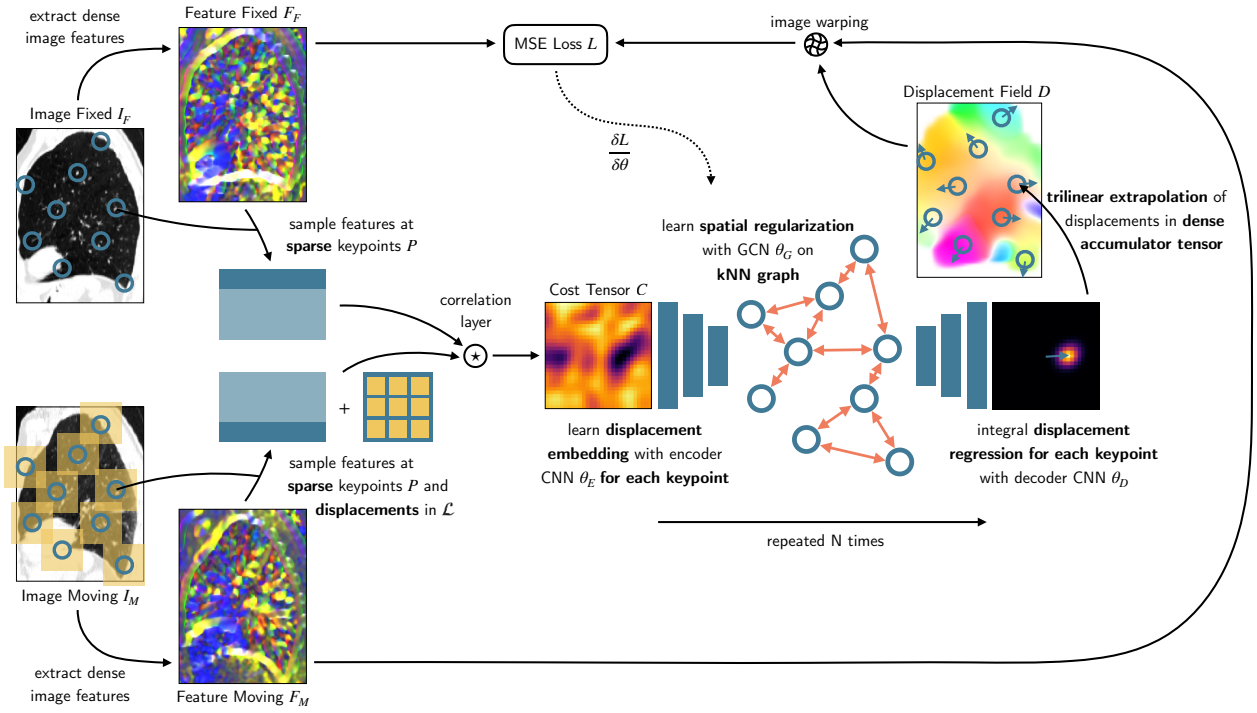

Fig. 1. Overview of our novel deep learning framework for keypoint-based dense deformable image registration. Feature maps $\boldsymbol{F_F}$ and $\boldsymbol{F_M}$ (here: MIND features [38]) are extracted from both, the fixed ($\boldsymbol{I_F}$) and moving ($\boldsymbol{I_M}$) image. Additionally, a set of sparse keypoints $\boldsymbol{P}$ is identified at distinctive locations in the fixed image using the Foerstner operator [39]. Correlating the sampled MIND features at the keypoints in the fixed image and dense displaced locations $\boldsymbol{\mathcal{L}}$ in the moving image, yields a cost tensor $\boldsymbol{C}$ for each keypoint. We then predict displacement vectors with our proposed GraphRegNet $\boldsymbol{\theta}$, that consists of three neural network modules. First, an encoder CNN $\boldsymbol{\theta_E}$ learns a low-dimensional displacement embedding for each cost tensor, then we employ a GCN $\boldsymbol{\theta_G}$ that distributes the learned embeddings across the kNN graph of the keypoints to achieve spatial regularization. Final displacement vectors are obtained via integration over the predefined displacement space $\boldsymbol{\mathcal{L}}$, weighted by probabilities of the predicted softmax map of the decoder CNN $\boldsymbol{\theta_D}$. All displacment vectors of the sparse keypoints are accumulated in the displacement field tensor $\boldsymbol{D}$ using trilinear extrapolation (+ spatial smoothing), which makes this densification operation fully differentiable and enables the use of an MSE loss $\boldsymbol{L}$ on the fixed and warped moving MIND image, guiding the training process in an unsupervised fashion.

fixed margin). For each scan pair an affine transformation is determined that fits the BB of the exhale lung mask to the BB of the inhale lung mask and is applied to the exhale image. This leaves the estimation of nonlinear deformations for the registration framework. Subsequently, all images are resampled to a fixed volume size of D×H×W and grayscale values (in Hounsfield units (HU)) are clipped and normalized to lie in the range of 0 to 1.

B. Distinctive Keypoints

Well distributed, distinctive and informative keypoints are of importance for the proposed graph based registration. Moreover, working on a sparse set of keypoints allows to cope with the large memory requirements of a correlation based approach on 3D medical data. For the keypoint extraction we follow previous works in lung registration [11], [40], employing the Foerstner interest operator [39], which is run-time efficient and led to state-of-the-art results in lung registration for conventional methods [12]. Alternatively, a keypoint graph can be constructed from vessel segmentations (cf. [21]). The keypoints are computed from the spatial gradients ∇I_F of the fixed image, that are smoothed with a Gaussian kernel G_{σ_1}, σ_1 describing the variance. A distinctive score S for each voxel in I_F is given by

$$S(I_F) = \frac{1}{\mathrm{Tr}((G_{\sigma_1} * (\nabla I_F \nabla I_F^T))^{-1})}. \quad (1)$$

High responses in S correspond to distinctive locations. To obtain a well distributed set of keypoints P we apply a max pooling operation with kernel size d and stride 1 to S, which yields S_{max}, and only add keypoints $\mathbf{p} = (p_x, p_y, p_z)$ to P that have equal response in S and S_{max} (non-max suppression). Additionally, we restrict the location of keypoints to the lung region (given by the precomputed inhale masks). Finally, the number of keypoints in P is adapted to a fixed number N_P by farthest point sampling (if $|P| >= N_P$) or insertion of random points already present in P (if $|P| < N_P$). The farthest point sampling algorithm starts with a random point of a point set and iteratively adds points to the sampling that have the farthest distance to all currently sampled points. Thus a well distributed coverage of the original point set is guaranteed.

C. Image Feature Extraction

Since in the context of this work we are strongly focusing on the prediction and regularization of displacement vectors in an unsupervised learning setting, we use the well

established modality independent neighbourhood descriptor (MIND) as image features and leave a deep learning based feature extraction or evaluation of further handcrafted image features (e.g. NGF [41]) as a future task. The MIND descriptor was first proposed in [42] for the task of multi-modal image registration and extended in [37], [38]. It uses the concept of self-similarity by defining six neighbors around the central voxel of interest and compares the patch-wise intensity difference between neighbors of a certain distance, resulting in a 12 channel feature map. We extract MIND descriptors for both, the fixed and moving image, yielding feature representations $F_F, F_M : \mathbb{R}^3 \rightarrow \mathbb{R}^{12}$. In addition to the subsequent feature correlation, the extracted MIND descriptors also serve as signals for the unsupervised warping loss (see Section II-F).

D. Feature Correlation

Following previous work on discrete 3D medical registration [10], [11] and the successful methods for 2D optical flow estimation [24], [25] we perform a similarity search for all fixed features $F_F(\mathbf{p})$ sampled at keypoints $\mathbf{p} \in P$ and moving features $F_M(\mathbf{p}+\mathbf{l})$ sampled at potential displacement locations $\mathbf{l} = (l_x, l_y, l_z) \in \mathcal{L} = q \cdot \{-l_{max}, \ldots, -1, 0, 1, \ldots, l_{max}\}^3$. The displacement search region is fully specified by the variables q, the quantisation step size, and $q \cdot l_{max}$, the largest expected displacement. Employing the sum of squared differences (SSD) as similarity metric yields the cost tensor

$$C(\mathbf{p}, \mathbf{l}) = \frac{1}{12} \sum_{i=0}^{11} (F_F^i(\mathbf{p}) - F_M^i(\mathbf{p}+\mathbf{l}))^2 : \mathbb{R}^6 \rightarrow \mathbb{R}, \quad (2)$$

where F_F^i and F_M^i denote the i-th channel of the respective 12 channel feature map. We note that the spatial dimensions [1-3] of the cost tensor are sparse, i.e. defined only for the set of keypoints P, while the displacement dimensions [3-6] are spanned by the dense displacement search space. To account for noisy similarities and ease learning for the following prediction network we smooth the cost tensor C with a Gaussian kernel G_{σ_2} along the displacement dimensions.

E. GraphRegNet

We now aim to predict a sparse displacement field D_S that assigns a displacement vector $\mathbf{d} = (d_x, d_y, d_z)$ to each keypoint $\mathbf{p} \in P$. The searched function $\theta(C) = D_S$ is modeled by our proposed GraphRegNet, a novel deep network architecture, whose parameters are learned in an unsupervised, end-to-end training process. In the following we describe a single layer of the GraphRegNet (for deeper networks, two or more layers can be stacked). The first part of the network architecture is a lightweight encoder CNN θ_E that operates on the displacement dimensions of C (for each keypoint). It consists of three convolutional layers with a stride of 2. Starting with 4 output channels, the number of feature channels is doubled with each layer. Next, the predicted low dimensional displacement embeddings are concatenated with the coordinates of respective keypoints. A GCN θ_G takes the displacement embeddings as input (possibly with shared weights if the embeddings still have dimensions that are non-singleton) and distributes them across the kNN graph of P. We employ three graph convolutions (edge convolutions [29]) in a DenseNet [43] fashion (always concatenating the input tensor of previous and current layers) and keep the number of output feature channels constant. An edge convolution is defined as

$$\mathbf{f}_i' = ReLU(\underset{(i,j)\in E}{\text{avg}} \mathbf{e}_{ij}) \quad (3)$$

following notations in [29] and describes the non-linear transformation of a feature vector $\mathbf{f}_i$ at point $\mathbf{p}_i \in P$. The edge features $\mathbf{e}_{ij} = h_\theta(\mathbf{f}_i, \mathbf{f}_j - \mathbf{f}_i)$ are aggregated per dimension by averaging over all edges $(i,j) \in E$ of the kNN graph. The function h_θ denotes the Euclidean inner product of learnable parameters $\theta = (\theta_1, \ldots, \theta_{|\mathbf{f}_i'|})$ with the concatenated keypoint $(\mathbf{f}_i)$ and local neigbourhood features $(\mathbf{f}_j - \mathbf{f}_i)$. The decoder CNN θ_D predicts the displacement vectors $\mathbf{d}$ using an integral regression approach [44], which combines the advantages of direct (continuous output, end-to-end training) and heatmap (superior performance, constrained output space) regression. Similar to the encoder CNN the decoder operates solely on the displacement dimensions. Two upconvolutions (trilinear upsampling + convolution) and a subsequent single convolutional layer output a single channel feature map $H_\mathbf{p}$ for each keypoint. The final displacement vector $\mathbf{d}$ can now be determined by integration over the displacement search region $\mathcal{L}$ weighted by the (softmax) normalized predictions $\tilde{H}_\mathbf{p}$ as

$$\mathbf{d} = \sum_{\mathbf{l}\in\mathcal{L}} \mathbf{l} \cdot \tilde{H}_\mathbf{p}(\mathbf{l}). \quad (4)$$

Training with direct regression of displacement vectors was also tested but could not converge. To stabilize forward and backward propagation we employ skip connections that concatenate the encoder and decoder feature maps of the same resolution stage and add a further convolution to combine the features (resembles an U-Net architecture but with a GCN in embedding layer, that acts on the spatial dimensions). All convolutions have a kernel size of 3 and are followed by an instance normalization layer [45] and a Leaky ReLU [46]. A detailed overview of the network architecture of the GraphRegNet is given in Section A of the Appendix. To summarize, the GraphRegNet is a novel learnable message passing network architecture that predicts a global optimal (in the sense of the training target) transformation from initial correspondence costs at individual keypoints. Therfore, we employ conventional CNNs (θ_E, θ_D) with trainable filters acting on the dense displacement dimensions of the cost tensor (in an encoder/decoder style) and use a graph neural network (θ_G) to learn to distribute the compressed cost messages across the irregular keypoint graph.

F. Sparse-to-Dense Supervision

All network parts are trained end-to-end in an unsupervised fashion using a mean squared error loss $L = MSE(F_F, D(F_M))$ on the fixed and warped moving MIND features, where D describes a dense displacement field. Additionally, the loss is masked with the precomputed inhale

mask. D is obtained from the predicted sparse displacements D_S. Therefore, all displacements $\mathbf{d} \in D_S$ are accumulated in a dense, low resolution tensor (initialized with zeroes) at respective keypoints $\mathbf{p}$ using trilinear extrapolation. Subsequently, the tensor is smoothed three times using average pooling with a kernel size of 5 and a stride of 1. Trilinear upsampling of the tensor yields the final displacement field D. See Figure 1 for visual comprehension. As all operations (in particular the integral regression and the trilinear extrapolation) are differentiable we can employ the dense warping loss to supervise the graph regularization on the sparse keypoints. An explicit regularization loss as used in many other deep learning frameworks did not show any benefit and is omitted. We attribute this to the trilinear extrapolation and subsequent spatial averaging of the displacement vectors of (only a few thousand) sparse keypoints which imposes an implicit smoothness on the displacement field (similar to spline transformation models with few control points in conventional image registration).

In the inference stage the predicted displacement field D can eventually be used to warp the moving 3D CT image and align the inhale and exhale phase of the observed lung anatomy.

III. Experiments and Results

To assess the accuracy of our method and validate different architectural choices, we conduct several experiments on two challenging inspiration-expiration benchmarks, namely the DIR-Lab 4D CT [35] and COPDgene [36] datasets. Especially the COPDgene dataset is of great importance for the validation of deep learning registration methods to cope with large deformations as it consists of breath-hold CT scans with much larger initial registration errors (in comparison to the 4D CT data acquired from patients with normal resting breathing). Each dataset contains ten scan pairs. As a set of 20 training pairs is considered to be small for a deep learning approach we include 25 additional scans from two public lung registration datasets (Empire10 [47], POPI [48]) and all experiments were carried out using a 5-fold cross validation. Figure 2 shows qualitative results for two scan pairs from the COPDgene dataset.

A. Implementation Details

We implemented our proposed registration framework using the deep learning library PyTorch [49] running on a NVIDIA Titan RTX. During preprocessing all scan pairs are resampled to a fixed volume size of D×H×W $= 192\times160\times192$ and clipped at HU values of -1000 and 1500. All hpyerparameters were tuned on the training cases of the first fold and left unaltered for all other folds and experiments. We extract $N_P = 2048$ keypoints from the fixed image with the described Foerstner method using a max pooling kernel of size $d = 5$ and a Gaussian kernel with variance $\sigma_1 = 1.4$. For the displacement search region $\mathcal{L}$ the quantization step size is set to $q = 2$ and the largest expected displacement to 28 ($l_{max} = 14$). The cost tensor C is smoothed with a Gaussian Kernel with variance $\sigma_2 = 1$. For the regularization network we stack two of the described GraphRegNet layers and the edge convolutions operate on a keypoint graph with $k = 15$ nearest neighbors. An increased accuracy can be reached with a two level approach. For the refinement stage the warped moving image (using the predicted displacement field) defines the new moving image. Additionally, we relax some framework parameters. This includes the number of keypoints $N_p = 3072$, the quantisation step size $q = 1$ and the largest expected displacement $l_{max} = 8$. For the second level a new regularization network is trained from scratch. In each stage we train the GraphRegNet for 150 epochs with a batch size of 1 using the Adam optimizer, which took approximatly 3.5 hours on the GPU. The initial learning rate is set to 0.1. For further details we refer to the publicly available implementation at `https://github.com/multimodallearning/graphregnet`. In our experiments a single GraphRegNet has only ∼33.000 trainable parameters. The total computation time for the final displacement field is less than 2 seconds (including the refinement stage). Most of the time is spent in the SSD computation of the similarities (40%), followed by the extraction of the Foerstner keypoints and the generation of the kNN graph (30%). The extraction of MIND image features and the forward path of the GraphRegNet take up 12% and 17% of the time, respectively. For inference the GPU memory usage is less than 4 GB (less than 11 GB for training when using gradient checkpointing).

B. Comparison Methods

In this section we give a brief description of related comparison methods (recent deep learning approaches) for exhale to inhale lung CT registration. Common to all methods is that they only report results on the DIR-Lab 4D CT data but not on the more difficult (in terms of larger deformations/initial errors) COPDgene dataset. Therfore, we also adapt the public implementation[1] of the widely used Voxelmorph registration framework [50] with few extensions and evaluate it on both DIR-Lab datasets.

1) DLIR [51]: The unsupervised Deep Learning Image Registration (DLIR) framework of de Vos et. al. is one of the first deep learning approaches for 3D medical registration. Analogous to conventional image registration an image similarity measure is optimized during the training stage. The authors propose a multi-resolution method by stacking multiple CNNs and providing input images of different resolution to predict the final displacement.

2) Ep18 [52]: Eppenhof et. al. took a supervised approach by applying synthetic transformations to a set of training images and learn to directly predict the known deformation. They could show that with this usage of strong data augmentation and strong supervision a relatively small dataset is sufficient to achieve acceptable registration accuracy.

3) OSL [53]: The work of Fechter et. al. is the most recent proposal for a deep learning framework for medical image registration. It mainly explores the idea of using deep neural networks in a one shot learning (OSL) setting as a drop-in replacement for conventional registration frameworks but

[1] `https://github.com/voxelmorph/voxelmorph`

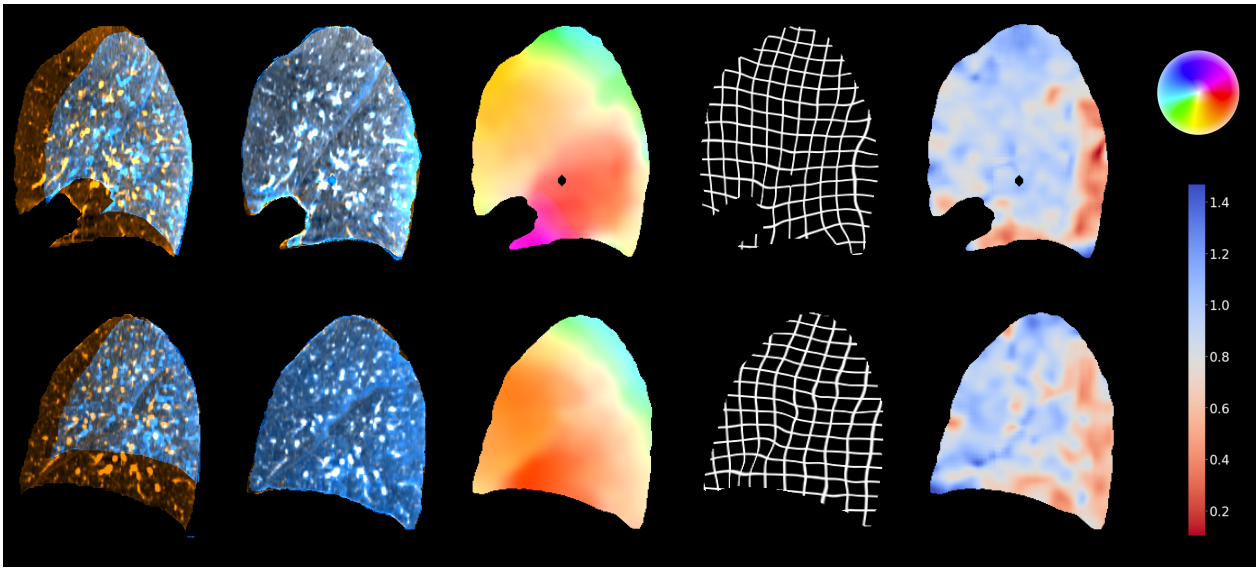

Fig. 2. Qualitative results of our proposed GraphRegNet registration framework. Each row visualizes registration results for one scan pair from the COPDgene dataset in saggital view. The first and second column show color overlays (orange: inhale scan, blue: exhale scan) for the initial and final alignment, respectively. Well aligned structures appear gray or white due to the addition of RGB values. The third and fourth columns are two different representations of the predicted displacement field (after affine prealignment). In the third column 2D displacement vectors of the saggital plane are color coded using the HSV color wheel in the top right. The fourth column shows a deformed regular grid after applying the displacement field. The Jacobian of the transformation within in the lung is visualized in the last column. The color bar in the right indicates the correspondence between color and Jacobian values.

could neither reach the registration accuracy of conventional approaches nor could benefit from the fast runtimes deep learning networks offer in inference.

4) LRN [21]: The LungRegNet of Fu et. al. marks (to our knowledge) the current state of the art for deep learning based methods on the DIR-Lab 4D CT dataset. The authors used a vessel enhancing preprocessing and an additional adversarial network to enforce realistic deformations. A huge drawback of their method is complexity of the architecture that leads to a reported inference time of 1 minute using a powerful NVIDIA Tesla V100 GPU.

5) mlVN [20]: Hering et al. proposed a deep learning based multi-level variational image registration network (mlVIRNET). It uses three resolution stages with progressively trained CNNs (initialized with CNN weights from preceding levels). The training on 500 lung CT scans is supervised with edge based normalized gradient fields (NGF) as image similarity measure, second order curvature regularization and additional manual lobe segmentations.

6) BMRF [54]: The work of Blendowski et. al. is a two-step hybrid approach. First, a deep network is trained to output descriptive binary features in a patch-based landmark retrieval task. The extracted features are then combined with handcrafted MIND-SSC image descriptors and used as input features for the B(inary)MRF-regularised deformable registration framework.

7) VM+ [50]: The Voxelmorph framework of Balakrishnan et. al. is a widely used single-level deep learning method for deformable image registration. As most other approaches it uses a U-Net like CNN to predict a displacement field (from the concatenated fixed and moving image), that warps the moving image and (during training) optimizes an image similarity metric. We extend the implementation (indicated by +) to a multi-level approach (three warps, no end-to-end learning) and employ the same MSE loss on MIND features we use in the rest of our own experiments.

8) LapIRN [6]: LapIRN is a multi-resolution pyramid registration network. In contrast to [20] the CNNs at different levels are trained end-to-end and it maintains full feature maps throughout the coarse-to-fine optimization scheme. The method was the overall winner of the recent multi-task medical image registration challenge "Learn2Reg" [7] and its implementation is publicly accessible[2].

9) FE+ [55]: The FlowNet3D [55] is an end-to-end trainable deep learning network for predicting scene flow between two point clouds. It relies on a flow embedding (FE) layer based on the concatenation of two candidate sets (from connected nodes in the k nearest neighbors graph of the fixed and moving point cloud). Initial experiments lead to unsatisfactory results due to the permutation invariance of the sparse candidates, which is why we extended the FE layer to capture all pairwise combinations of candidates which leads to a higher dimensional intermediate tensor that is fed into 1×1 convolutions and is projected to a meaningful embedding using max-pooling (FE+). For this experiment, Foerstner keypoints are extracted from both, the fixed and the moving image, and MIND feature patches at the sparse keypoints are used as input to the FE+ layer.

10) PDD+ [16]: Own previous work of Heinrich, the PDD net, uses approximate minconvolutions and mean field infer-

[2] https://github.com/cwmok/LapIRN

TABLE I

REGISTRATION RESULTS ON THE DIR-LAB 4D CT [35] AND COPDGENE [36] DATASETS. WE REPORT THE AVERAGE LANDMARK DISTANCE IN MILLIMETERS FOR ALL INDIVIDUAL CASES AS WELL AS THE AVERAGE DISTANCE AND STANDARD DEVIATION OVER ALL CASES OF A DATASET. RESULTS FOR COMPARISON METHODS (WITH EXCEPTION OF VM+, LAPIRN, FE+, PDD+ AND MST) WERE TAKEN FROM LITERATURE. FOR ALL OTHER METHODS/EXPERIMENTS A TEST FOR STATISTICAL SIGNIFICANCE WITH RESPECT TO OUR PROPOSED REGISTRATION FRAMEWORK WAS CONDUCTED USING THE WILCOXON SIGNED-RANK TEST (CALCULATED OVER ALL 3000 AVAILABLE LANDMARK PAIRS OF A DATASET). SIGNIFICANCE LEVELS ARE DEFINED AS $* \; p < 0.05$, $** \; p < 0.01$ AND $*** \; p < 0.001$.

	init.	DLIR [51]	Ep18 [52]	OSL [53]	LRN [21]	mlVN [20]	BMRF [54]	VM+ [50]	LapIRN [6]	FE+ [55]	PDD+ [16]	MST	RW	noreg	coords	sl	unif.	ours
4DCT 01	03.89	1.27	1.45	1.21	0.98	1.33		1.46	1.00	2.20	0.90	**0.82**	1.21	1.40	0.86	0.86	0.89	0.86
4DCT 02	04.34	1.20	1.46	1.13	0.98	1.33		1.51	1.28	3.89	0.91	**0.87**	1.17	1.64	0.98	0.90	0.93	0.90
4DCT 03	06.94	1.48	1.57	1.32	1.14	1.48		2.31	2.18	2.71	1.06	1.09	1.37	1.50	1.11	1.13	**1.05**	1.06
4DCT 04	09.83	2.09	1.95	1.84	**1.39**	1.85		2.72	3.05	2.95	1.66	1.63	1.52	2.05	1.65	1.61	1.51	1.45
4DCT 05	07.48	1.95	2.07	1.80	**1.43**	1.84		2.69	2.36	3.03	1.68	1.58	2.11	2.91	1.73	1.67	1.68	1.60
4DCT 06	10.89	5.16	3.04	2.30	2.26	3.57		3.07	1.78	3.36	1.86	1.71	1.83	2.19	1.60	1.64	**1.59**	**1.59**
4DCT 07	11.03	3.05	3.41	1.91	**1.42**	2.61		3.01	2.24	3.10	1.94	1.73	1.88	2.33	1.67	1.69	1.63	1.74
4DCT 08	14.99	6.48	2.80	3.47	3.13	2.62		6.22	2.24	2.94	1.79	1.55	1.77	2.88	2.28	1.58	**1.43**	1.46
4DCT 09	07.92	2.10	2.18	1.47	**1.27**	2.70		2.94	2.26	2.86	1.94	1.85	2.23	2.23	1.72	1.87	1.72	1.58
4DCT 10	07.30	2.09	1.83	1.79	1.93	2.63		3.00	1.90	2.99	2.03	1.90	1.97	2.43	1.75	1.97	2.26	**1.71**
avg	08.46	2.64	2.17	1.83	1.59	2.19		2.89	2.03	3.00	1.57	1.47	1.70	2.15	1.53	1.49	1.47	**1.39**
std	06.58	4.32	1.89	2.35	1.58	1.62		2.21	1.89	1.70	1.36	1.26	2.38	1.70	1.57	1.30	1.65	1.29
sig. level	***							***	***	***	***	***	***	***	***	***	*	
COPD 01	26.33						1.51	9.95	6.85	4.89	2.57	1.42	3.51	4.32	5.50	1.71	1.80	**1.38**
COPD 02	21.79						2.27	9.96	6.90	7.30	4.01	3.42	5.26	7.27	9.12	2.75	**2.09**	**2.09**
COPD 03	12.64						1.39	4.41	1.51	2.89	1.46	1.32	1.57	1.42	1.40	1.42	**1.18**	1.22
COPD 04	29.58						1.86	7.08	6.38	5.46	2.19	**1.48**	2.51	7.30	4.46	2.06	1.60	1.58
COPD 05	30.08						1.46	9.19	6.81	5.19	2.22	1.44	3.33	4.77	3.44	1.81	1.49	**1.37**
COPD 06	28.46						1.40	8.12	4.19	5.53	1.89	1.47	2.57	3.58	2.96	1.43	1.31	**1.10**
COPD 07	21.60						1.46	7.10	2.73	4.40	1.62	1.37	2.14	2.68	2.99	1.64	1.23	**1.19**
COPD 08	26.46						1.53	7.92	4.32	3.94	1.72	1.33	1.64	4.21	2.22	1.54	1.44	**1.19**
COPD 09	14.86						1.34	6.93	3.60	3.57	1.51	1.22	2.79	3.02	1.68	1.45	1.13	**0.99**
COPD 10	21.81						1.71	9.16	6.59	4.44	2.43	1.55	2.62	7.93	6.95	1.79	1.82	**1.38**
avg	23.36						1.59	7.98	4.99	4.76	2.16	1.60	2.79	4.65	4.07	1.76	1.50	**1.34**
std	11.86						0.27	3.75	3.94	4.06	2.63	2.04	4.51	5.89	5.57	1.57	1.75	1.44
sig. level	***							***	***	***	***	***	***	***	***	***	***	

ence to regularize the displacement similarities on a regular grid. The network consists of only six trainable weighting and offset parameters. For a fair comparison to our proposed method we extended (indicated by the + sign) the original implementation[3] to operate on the irregular keypoint graph, use the same dense warping loss and employ a second (refinement) stage.

11) MST: A non deep learning method that is a close GPU reimplemenation of the CorrField method [11] with a focus on fast computation time. While the computation of the final displacement field took more than a minute using the C++ implementation[4], the GPU version needs less than 5 seconds. For the sake of fast inference, some features (e.g. inverse consistency) are missing from the original method. A minimum spanning tree (MST) is generated from the set of sparse keypoints, which enables exact message passing using belief propagation on the graph to regularize the displacement costs.

C. Ablations Studies

In a second batch of experiments we aim to evaluate the general design choices of our proposed GraphRegNet. All methods in this group use the described keypoint based registration framework and only differ in the way they predict the displacements $\mathbf{d} \in D_S$ from the cost tensor C. Again, we give a short overview over the different ablation studies.

1) RW: In this ablation experiment we replace the GraphRegNet with a non deep learning approach, a random walk (RW) [56] based on the graph Laplacian that distributes the cost tensor over the kNN graph. This method serves as weak baseline for following experiments as a deep learning model should easily learn a better (or at least comparable) solution provided that there are a sufficient number of trainable parameters and the network design itself allows it.

2) noreg: A baseline to show the importance of explicit spatial regularization. In this experiment the GCN θ_G is removed from the architecture. Additionally, we do not add keypoint coordinates to the displacement embeddings. Thus, the network is only able to learn on the displacement dimensions and smooth across the displacements of the cost tensor.

3) coords: In this configuration we set $k = 1$, which means that there are no connections between keypoints on the kNN graph. General spatial information can only be exploited from the concatenated keypoint coordinates (coords) using multi-layer perceptrons (MLPs). Within this setting the GCN θ_G is similar to the PointNet [27] architecture.

4) sl: A single-level (sl) baseline, that does not use a second (refinement) stage.

5) unif.: Instead of extracting distinctive keypoints using the Foerstner operator, we sample keypoints on a uniform (unif.) grid (within the lung mask). The number of keypoints N_P is kept the same.

[3] https://github.com/multimodallearning/pdd_net
[4] http://www.mpheinrich.de/software.html

D. Target Registration Error

Our main evaluation metric is the target registration error (TRE) between (medical) expert-annotated landmarks. Both DIR-Lab datasets provide 300 manually annotated landmark correspondences for each scan pair. Table I provides an overview over quantitative results for all comparison methods and conducted experiments. With an average landmark distance of 8.46 mm and 23.36 mm initial registration errors vary greatly between the 4D CT and COPDgene dataset, respectively. However, most comparison methods only report landmark distances for the 4D CT dataset, for which the various multi-level and multi-resolution approaches based on U-Net like encoder decoder architectures reach accuracies from 2.89 mm (VM+) to 2.03 mm (LapIRN). The currently published state of the art for deep learning methods is at 1.59 mm (LRN). The proposed GraphRegNet within our keypoint based registration framework improves the TRE by ∼13% to 1.39 mm. For the COPDgene dataset we can report the TREs of *Bl19* and our own experiments for *VM+*, *LapIRN*, *FE+*, *PDD+* and *MST* that are at 1.59 mm, 7.89 mm, 4.99 mm, 4.76 mm, 2.16 mm and 1.60 mm, respectively. Here, the GraphRegNet reduces the average landmark distance to 1.34 mm. Figure 3 shows a detailed comparison of all methods of our keypoint-based experiments on the COPDgene dataset. In ablation studies we observe a reduction in TREs of ∼67% when exploiting neighborhood information on the kNN graph using edge convolutions (coords –> ours), ∼24% after a refined alignment with a two level approach (sl –> ours) and ∼11% when sampling keypoints at distinctive instead of uniform locations (unif. –> ours). For experiments where we have registration errors for all landmark pairs available we perform the Wilcoxon signed-rank test and can confirm statistical significance (at least $p < 0.05$) for all methods and ablation studies with respect to our proposed approach.

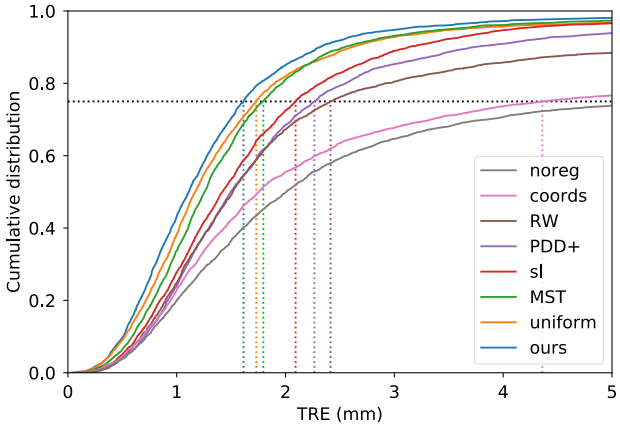

Fig. 3. Cumulative distribution of target registration errors in millimeters for all keypoint based methods on all landmark pairs of the COPDgene [36] dataset. In addition, the dotted lines visualize the **75**th percentiles of the TRE, which are **1.61** mm (ours), **1.73** mm (uniform), **1.80** mm (MST), **2.09** mm (sl), **2.27** mm (PDD+), **2.43** mm (RW), **4.38** mm (coords) and **5.25** mm (noreg).

TABLE II
COMPARISON OF OUR DEEP LEARNING APPROACH TO A SELECTION OF POPULAR CONVENTIONAL REGISTRATION FRAMEWORKS ON THE COPDGENE DATASET. WE REPORT THE MEAN AND STANDARD DEVIATION OF THE TARGET REGISTRATION ERROR IN MILLIMETERS AS WELL AS THE AVERAGE COMPUTATION TIME OF THE ALGORITHMS. FOR A DETAILED RUN-TIME ANALYSIS OF OUR METHOD WE REFER TO SECTION III-A.

	TRE	avg. computation time
DIS-CO [12]	0.82 ± 0.97	5 minutes
ANTs [57]	1.79 ± 2.10	3 hours
Elastix [58]	1.32 ± 1.24	14 minutes
NiftyReg [59]	2.19 ± 2.00	9 minutes
ours	1.34 ± 1.44	2 seconds

E. Jacobian Determinant

For the assessment of realistic and well regularized deformations we evaluate the standard deviation (a value 0 would describe an entirely smooth transformation) and the fraction of negative values (image foldings) of the Jacobian determinant within the lungs. With average fractions of 0.02 % and 0.15 % (maximum: 0.21 % and 0.83 %) the deformation for both datasets, 4D CT and COPDgene, have a small amount of image foldings. The standard deviation of the Jacobian determinant is 0.13 and 0.21, respectively, which compares well with comparison methods *VM+* (0.11 and 0.20), *LapIRN* (0.12 and 0.17), *FE+* (0.15 and 0.32), *PDD+* (0.10 and 0.19) and *MST* (0.10 and 0.18). The smoothness of the transformation and a typical local distribution of the Jacobian determinant can also be visually inspected in Figure 2.

IV. DISCUSSION

Our proposed GraphRegNet shows significantly improved results over a number of comparison and state-of-the-art methods for deep learning based medical registration on the widely used DIR-Lab 4D CT [35] dataset. It surpasses the registration accuracy of the best performing method, the LungRegNet [21] with 1.59 mm, by more than 13% while it needs only a fraction (< 2 seconds) of the reported computation time of approximately 1 minute. The low number of model parameters and fast training times are another advantage over comparison methods. For the more difficult (in terms or larger initial deformations) COPDgene [36] dataset the improvements are even more obvious.

While most comparison methods do not report results on the closely related COPDgene dataset, we conducted own experiments with the state-of-the-art LapIRN registration framework as representative approach for the widely used U-Net like encoder-decoder architectures. Here, our keypoint-based discrete registration could reduce the TRE of 4.99 mm by approximately 70 %. Also, the high TREs of VM+ and LapIRN clearly show the difficulty of U-Net approaches to cope with large deformations in this lung registration task (even when employed in a multi-level and multi-resolution fashion). In contrast, with 1.34 mm the registration accuracy of the GraphRegNet is comparable (and even slightly better) on the COPDgene and 4D CT (1.39 mm) dataset and further highlights the ability of a discrete deep learning registration framework to accurately align images with large

initial deformations. Another related subarea of medical image analysis, anatomical landmark localization, could benefit from this finding as it also often operates on large and discrete search spaces. In comparison to conventional registration approaches, such as the work from Rühaak et. al. [12] (with 0.82 mm the first method that reached the inter-observer variability on the COPDgene dataset), the registration accuracy of the GraphRegNet still lags behind. At the same time our method benefits from the fast inference of deep neural networks and with less than 2 seconds is much faster than [12], which reports a computation time of 5 minutes for a single registration. Well established registration software such as *ANTs* [57], *Elastix* [58] or *NiftyReg* [59] perform worse or on par with our approach (1.79 mm (boosted [60]), 1.32 mm and 2.19 mm (boosted [60]), respectively) while taking from 9 minutes to 3 hours for a single scan pair (values taken from [52], [60]), which shows that the gap in accuracy between conventional and deep learning based registration continues to close. With respect to architectural choices we can conclude that the largest improvement in accuracy stems from our novel deep network architecture that explicitly learns a descriptive embedding on the displacement dimensions (a significant difference compared to works such as FlowNet [24] or PWC-Net [25]) and uses graph convolutions as a data-driven, trainable regularizer (cf. *noreg*/*coords*/*RW*). Further important parts of our framework are the use of a second refinement stage (cf. *sl*) and the sampling of distinctive keypoints (cf. *unif.*). Finally, we would like to highlight the results of our GraphRegNet in comparison to the *MST* method. While *MST* uses exact message passing on the minimum spanning tree of the extracted keypoints, our approach is able to learn this step entirely from data with additional supervision and thus surpasses the exact method. This finding might have identified a general approach for the use of machine learning in the field of discrete optimization.

V. Conclusion and Outlook

In this work, we have presented a novel deep network architecture for learning well-regularized dense displacement fields in a discrete and keypoint-based registration framework. Our GraphRegNet combines CNN and GCN layers in a single network, which allows to learn deep feature embeddings (using a convolutional encoder decoder net that acts on the displacement dimensions) but at the same time distribute information on a sparse and high resolution irregular grid. A novel differentiable sparse-to-dense warping loss allows to supervise the training of our network on a sparse keypoint graph with dense and descriptive image features. In the evaluation on two challenging exhale to inhale lung CT datasets we could advance the state of the art for deep learning methods while also improving the run time for conventional approaches from minutes to seconds. A series of ablation studies demonstrated significant improvements of individual architectural choices and highlights the great potential of discrete and keypoint-based deep learning approaches (in contrast to fully integrated encoder decoder architectures) for 3D medical registration.

While, in this work, we focused mainly on the part of learning accurate displacements from the cost tensor, for future research we especially see potential for further learning and improvements of the input data, i.e. the image features and keypoint graph. Image features could also be learned in an end-to-end training (instead of using fixed MIND features) and a more descriptive keypoint graph (e.g. from vessel trees) would enable a more targeted graphical message passing (compared to the nearest neighbour heuristic on Foerstner keypoints). Finally, we believe that our method generalises well to other registration tasks with large deformations, e.g. the alignment of inter-patient abdominal CT, where keypoints sampled on surfaces of anatomical structures enable the generation of an expressive input graph.

Appendix

A. Network Architecture

The detailed network architecture of the GraphRegNet with a total number of ∼33.000 trainable parameters is summarized in Figure 4 in combination with Table III. Subindices E1, G1, D1, etc. represent the corresponding layer of the encoder,

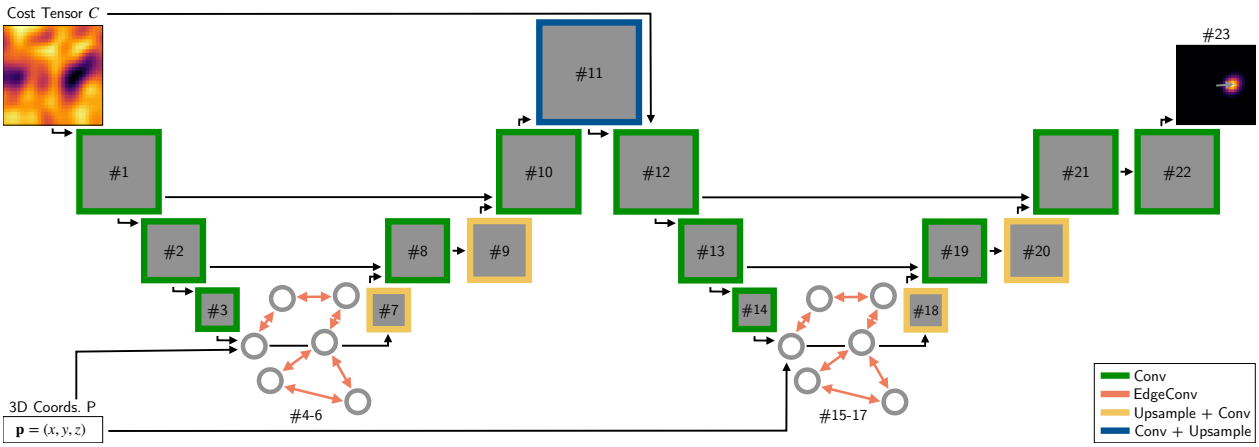

Fig. 4. Block diagram of the GraphRegNet architecture used in our experiments. Detailed information on the individual layers can be found in the corresponding Table III

TABLE III

			Size	Kernel / kNN	Stride	# Ch. (in/out)	Skip	Dim
	3D	Coords.	1			−/3	#4-6, #15-17	
	Cost	Tensor	29^3			−/1	#12	
θ_{E1}	#1	Conv	15^3	3^3	2	1/4	#10	D
	#2	Conv	8^3	3^3	2	4/8	#8	D
	#3	Conv	4^3	3^3	2	8/16	#5,#6	D
θ_{G1}	#4	EdgeConv	4^3	15	1	19/16	#6	S
	#5	EdgeConv	4^3	15	1	35/16		S
	#6	EdgeConv	4^3	15	1	51/16		S
θ_{D1}	#7	Upsample	8^3					D
		Conv	8^3	3^3	1	16/8		D
	#8	Conv	8^3	3^3	1	16/8		D
	#9	Upsample	15^3					D
		Conv	15^3	3^3	1	8/4		D
	#10	Conv	15^3	3^3	1	8/4		D
	#11	Conv	15^3	3^3	1	4/1		D
		Upsample	29^3					D
θ_{E2}	#12	Conv	15^3	3^3	2	2/4	#22	D
	#13	Conv	8^3	3^3	2	4/8	#19	D
	#14	Conv	4^3	3^3	2	8/16	#16,#17	D
θ_{G2}	#15	EdgeConv	4^3	15	1	19/16	#17	S
	#16	EdgeConv	4^3	15	1	35/16		S
	#17	EdgeConv	4^3	15	1	51/16		S
θ_{D2}	#18	Upsample	8^3					D
		Conv	8^3	3^3	1	16/8		D
	#19	Conv	8^3	3^3	1	16/8		D
	#20	Upsample	15^3					D
		Conv	15^3	3^3	1	8/4		D
	#21	Conv	15^3	3^3	1	8/4		D
	#22	Conv	15^3	3^3	1	4/1		D
	#23	IntReg	1			1/3		D

graph network and decoder, respectively. The specific output sizes apply to a displacement space with $l_{max} = 14$. The number of keypoints is omitted for the sake of clarity as it is the same for all layers ($N_P = 2048$). Kernel and kNN sizes correspond to the dimensions of the learnable filters for conventional and to the number of (k) nearest neighbors for edge convolutions, respectively. Skip connections specify layer outputs that are concatenated for further processing. All convolutional layers (except for #23) are followed by instance normalization and a leaky ReLU as non-linear activation function. Upsampling uses trilinear interpolation. Also, we denote if a layer acts on displacement (D) or spatial (S) dimensions. The integral displacement regression layer is abbreviated as *IntReg* and represents the transition from a discrete to a continuous displacement space.

ACKNOWLEDGMENT

The authors thank Tony C. W. Mok and Albert C. S. Chung for the support and provision of the code for the LapIRN registration framework.

REFERENCES

[1] S. Flampouri, S. B. Jiang, G. C. Sharp, J. Wolfgang, A. A. Patel, and N. C. Choi, "Estimation of the delivered patient dose in lung imrt treatment based on deformable registration of 4d-ct data and monte carlo simulations," *Physics in Medicine & Biology*, vol. 51, no. 11, p. 2763, 2006.

[2] C. J. Galbán, M. K. Han, J. L. Boes, K. A. Chughtai, C. R. Meyer, T. D. Johnson, S. Galbán, A. Rehemtulla, E. A. Kazerooni, F. J. Martinez *et al.*, "Computed tomography–based biomarker provides unique signature for diagnosis of copd phenotypes and disease progression," *Nature Medicine*, vol. 18, no. 11, p. 1711, 2012.

[3] F. Pan, T. Ye, P. Sun, S. Gui, B. Liang, L. Li, D. Zheng, J. Wang, R. L. Hesketh, L. Yang *et al.*, "Time course of lung changes on chest ct during recovery from 2019 novel coronavirus (covid-19) pneumonia," *Radiology*, p. 200370, 2020.

[4] E. J. van Beek, J. M. Wild, H.-U. Kauczor, W. Schreiber, J. P. Mugler III, and E. E. de Lange, "Functional mri of the lung using hyperpolarized 3-helium gas," *Journal of Magnetic Resonance Imaging*, vol. 20, no. 4, pp. 540–554, 2004.

[5] J. M. Reinhardt, K. Ding, K. Cao, G. E. Christensen, E. A. Hoffman, and S. V. Bodas, "Registration-based estimates of local lung tissue expansion compared to xenon ct measures of specific ventilation," *Medical Image Analysis*, vol. 12, no. 6, pp. 752–763, 2008.

[6] T. C. Mok and A. C. Chung, "Large deformation diffeomorphic image registration with laplacian pyramid networks," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2020, pp. 211–221.

[7] A. Dalca, Y. Hu, T. Vercauteren, M. Heinrich, L. Hansen, M. Modat, B. de Vos, Y. Xiao, H. Rivaz, M. Chabanas, I. Reinertsen, B. Landman, J. Cardoso, B. van Ginneken, A. Hering, and K. Murphy, "Learn2reg - the challenge," Mar. 2020.

[8] B. Glocker, N. Komodakis, G. Tziritas, N. Navab, and N. Paragios, "Dense image registration through mrfs and efficient linear programming," *Medical Image Analysis*, vol. 12, no. 6, pp. 731–741, 2008.

[9] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient belief propagation for early vision," *International Journal of Computer Vision*, vol. 70, no. 1, pp. 41–54, 2006.

[10] M. P. Heinrich, M. Jenkinson, M. Brady, and J. A. Schnabel, "Mrf-based deformable registration and ventilation estimation of lung ct," *IEEE Transactions on Medical Imaging*, vol. 32, no. 7, pp. 1239–1248, 2013.

[11] M. P. Heinrich, H. Handels, and I. J. Simpson, "Estimating large lung motion in copd patients by symmetric regularised correspondence fields," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015, pp. 338–345.

[12] J. Rühaak, T. Polzin, S. Heldmann, I. J. Simpson, H. Handels, J. Modersitzki, and M. P. Heinrich, "Estimation of large motion in lung ct by integrating regularized keypoint correspondences into dense deformable registration," *IEEE Transactions on Medical Imaging*, vol. 36, no. 8, pp. 1746–1757, 2017.

[13] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. Torr, "Conditional random fields as recurrent neural networks," in *IEEE International Conference on Computer Vision*, 2015, pp. 1529–1537.

[14] K. Kamnitsas, C. Ledig, V. F. Newcombe, J. P. Simpson, A. D. Kane, D. K. Menon, D. Rueckert, and B. Glocker, "Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation," *Medical Image Analysis*, vol. 36, pp. 61–78, 2017.

[15] P. Knobelreiter, C. Reinbacher, A. Shekhovtsov, and T. Pock, "End-to-end training of hybrid cnn-crf models for stereo," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2339–2348.

[16] M. P. Heinrich, "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*, 2019, pp. 50–58.

[17] M. P. Heinrich and L. Hansen, "Highly accurate and memory efficient unsupervised learning-based discrete ct registration using 2.5d displacement search," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, in press.

[18] R. Girshick, F. Iandola, T. Darrell, and J. Malik, "Deformable part models are convolutional neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 437–446.

[19] M. J. Johnson, D. K. Duvenaud, A. Wiltschko, R. P. Adams, and S. R. Datta, "Composing graphical models with neural networks for structured representations and fast inference," in *Advances in Neural Information Processing Systems*, 2016, pp. 2946–2954.

[20] A. Hering, B. van Ginneken, and S. Heldmann, "mlvirnet: Multilevel variational image registration network," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2019, pp. 257–265.

[21] Y. Fu, Y. Lei, T. Wang, K. Higgins, J. D. Bradley, W. J. Curran, T. Liu, and X. Yang, "Lungregnet: an unsupervised deformable image registration method for 4d-ct lung," *Medical Physics*, vol. 47, no. 4, pp. 1763–1774, 2020.

[22] L. König, J. Rühaak, A. Derksen, and J. Lellmann, "A matrix-free approach to parallel and memory-efficient deformable image registration," *SIAM Journal on Scientific Computing*, vol. 40, no. 3, pp. B858–B888, 2018.

[23] R. Sandkühler, S. Andermatt, G. Bauman, S. Nyilas, C. Jud, and P. C. Cattin, "Recurrent registration neural networks for deformable image registration," in *Advances in Neural Information Processing Systems*, 2019, pp. 8758–8768.

[24] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "Flownet 2.0: Evolution of optical flow estimation with deep networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2462–2470.

[25] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, "Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8934–8943.

[26] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst, "Geometric deep learning: going beyond euclidean data," *IEEE Signal Processing Magazine*, vol. 34, no. 4, pp. 18–42, 2017.

[27] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 652–660.

[28] D. K. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Bombarell, T. Hirzel, A. Aspuru-Guzik, and R. P. Adams, "Convolutional networks on graphs for learning molecular fingerprints," in *Advances in Neural Information Processing Systems*, 2015, pp. 2224–2232.

[29] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, "Dynamic graph cnn for learning on point clouds," *ACM Transactions On Graphics*, vol. 38, no. 5, pp. 1–12, 2019.

[30] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *International Conference on Learning Representations*, 2018.

[31] J. Schlemper, O. Oktay, M. Schaap, M. Heinrich, B. Kainz, B. Glocker, and D. Rueckert, "Attention gated networks: Learning to leverage salient regions in medical images," *Medical Image Analysis*, vol. 53, pp. 197–207, 2019.

[32] L. Tchapmi, C. Choy, I. Armeni, J. Gwak, and S. Savarese, "Segcloud: Semantic segmentation of 3d point clouds," in *International Conference on 3D Vision*, 2017, pp. 537–547.

[33] Q. Huang, W. Wang, and U. Neumann, "Recurrent slice networks for 3d segmentation of point clouds," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2626–2635.

[34] L. Hansen and M. P. Heinrich, "Tackling the problem of large deformations in deep learning based medical image registration using displacement embeddings," in *Medical Imaging with Deep Learning: MIDL 2020–Short Paper Track*, 2020.

[35] R. Castillo, E. Castillo, R. Guerra, V. E. Johnson, T. McPhail, A. K. Garg, and T. Guerrero, "A framework for evaluation of deformable image registration spatial accuracy using large landmark point sets," *Physics in Medicine & Biology*, vol. 54, no. 7, p. 1849, 2009.

[36] R. Castillo, E. Castillo, D. Fuentes, M. Ahmad, A. M. Wood, M. S. Ludwig, and T. Guerrero, "A reference dataset for deformable image registration spatial accuracy evaluation using the copdgene study archive," *Physics in Medicine & Biology*, vol. 58, no. 9, p. 2861, 2013.

[37] M. P. Heinrich, M. Jenkinson, B. W. Papież, M. Brady, and J. A. Schnabel, "Towards realtime multimodal fusion for image-guided interventions using self-similarities," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2013, p. 187.

[38] M. P. Heinrich, M. Jenkinson, M. Bhushan, T. Matin, F. V. Gleeson, M. Brady, and J. A. Schnabel, "Mind: Modality independent neighbourhood descriptor for multi-modal deformable registration," *Medical Image Analysis*, vol. 16, no. 7, pp. 1423–1435, 2012.

[39] W. Förstner and E. Gülch, "A fast operator for detection and precise location of distinct points, corners and centres of circular features," in *Intercommission Conference on Fast Processing of Photogrammetric Data*, 1987, pp. 281–305.

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

[41] E. Haber and J. Modersitzki, "Intensity gradient based registration and fusion of multi-modal images," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2006, p. 726.

[42] M. P. Heinrich, M. Jenkinson, M. Bhushan, T. Matin, F. V. Gleeson, J. M. Brady, and J. A. Schnabel, "Non-local shape descriptor: A new similarity metric for deformable multi-modal registration," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2011, pp. 541–548.

[43] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4700–4708.

[44] X. Sun, B. Xiao, F. Wei, S. Liang, and Y. Wei, "Integral human pose regression," in *European Conference on Computer Vision*, 2018, pp. 529–545.

[45] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," *arXiv preprint arXiv:1607.08022*, 2016.

[46] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *International Conference on Machine Learning*, 2013.

[47] K. Murphy, B. Van Ginneken, J. M. Reinhardt, S. Kabus, K. Ding, X. Deng, K. Cao, K. Du, G. E. Christensen, V. Garcia *et al.*, "Evaluation of registration methods on thoracic ct: the empire10 challenge," *IEEE Transactions on Medical Imaging*, vol. 30, no. 11, pp. 1901–1920, 2011.

[48] J. Vandemeulebroucke, D. Sarrut, P. Clarysse *et al.*, "The popi-model, a point-validated pixel-based breathing thorax model," in *International Conference on the use of Computers in Radiation Therapy*, vol. 2, 2007, pp. 195–199.

[49] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, 2019, pp. 8024–8035.

[50] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, and A. V. Dalca, "Voxelmorph: a learning framework for deformable medical image registration," *IEEE Transactions on Medical Imaging*, vol. 38, no. 8, pp. 1788–1800, 2019.

[51] B. D. de Vos, F. F. Berendsen, M. A. Viergever, H. Sokooti, M. Staring, and I. Išgum, "A deep learning framework for unsupervised affine and deformable image registration," *Medical Image Analysis*, vol. 52, pp. 128–143, 2019.

[52] K. A. Eppenhof and J. P. Pluim, "Pulmonary ct registration through supervised learning with convolutional neural networks," *IEEE Transactions on Medical Imaging*, vol. 38, no. 5, pp. 1097–1105, 2018.

[53] T. Fechter and D. Baltas, "One shot learning for deformable medical image registration and periodic motion tracking," *IEEE Transactions on Medical Imaging*, 2020.

[54] M. Blendowski and M. P. Heinrich, "Combining mrf-based deformable registration and deep binary 3d-cnn descriptors for large lung motion estimation in copd patients," *International journal of computer assisted radiology and surgery*, vol. 14, no. 1, pp. 43–52, 2019.

[55] X. Liu, C. R. Qi, and L. J. Guibas, "Flownet3d: Learning scene flow in 3d point clouds," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 529–537.

[56] D. J. Aldous, "Lower bounds for covering times for reversible markov chains and random walks on graphs," *Journal of Theoretical Probability*, vol. 2, no. 1, pp. 91–100, 1989.

[57] G. Song, N. Tustison, B. Avants, and J. C. Gee, "Lung ct image registration using diffeomorphic transformation models," *Medical Image Analysis for the Clinic: A Grand Challenge*, pp. 23–32, 2010.

[58] S. Klein, M. Staring, K. Murphy, M. A. Viergever, and J. P. Pluim, "Elastix: a toolbox for intensity-based medical image registration," *IEEE Transactions on Medical Imaging*, vol. 29, no. 1, pp. 196–205, 2009.

[59] M. Modat, G. R. Ridgway, Z. A. Taylor, M. Lehmann, J. Barnes, D. J. Hawkes, N. C. Fox, and S. Ourselin, "Fast free-form deformation using graphics processing units," *Computer Methods and Programs in Biomedicine*, vol. 98, no. 3, pp. 278–284, 2010.

[60] S. E. Muenzing, B. van Ginneken, M. A. Viergever, and J. P. Pluim, "Dirboost–an algorithm for boosting deformable image registration: Application to lung ct intra-subject registration," *Medical Image Analysis*, vol. 18, no. 3, pp. 449–459, 2014.