UNIVERSITÄT ZU LÜBECK
INSTITUTE OF MEDICAL INFORMATICS

From the Institute of Medical Informatics
of the University of Lübeck
Director: Prof. Dr. rer. nat. habil. Heinz Handels

Point Clouds and Keypoint Graphs in 3D Deep Learning for Medical Image Analysis

Dissertation
for
Fulfillment of Requirements for the Doctoral Degree
of the University of Lübeck

from the Department of Computer Sciences and Technical Engineering

Submitted by
Lasse Hansen
from Oldenburg i. H.

Lübeck, 2022

First referee: Prof. Dr. Mattias P. Heinrich
Second referee: Prof. Dr. Ben Glocker

Date of oral examination: 30/08/2022

Approved for printing. Lübeck, 01/09/2022

Abstract

For decades, general technological progress has ensured that automated and computer-assisted medical image analysis has and will continue to become an increasingly important aspect of clinical practice. In recent years, machine learning and deep learning methods in particular have succeeded in setting new standards of quality in the field. The focus is thereby largely on processing dense 2D and 3D medical scans, both in clinical applications as well as in research. However, in many cases, sparse data structures such as point clouds or keypoint graphs provide a more natural, direct representation of anatomical structures, e.g. the vascular system of the heart/eye or the airways of the lung. This also offers the potential for reduced memory and more efficient image analysis. Thus, also due to recent methodological advances in the field of geometric deep learning, i.e. adapting successful machine learning strategies to irregular domains, there is an increased interest in the use of graph-based approaches for medical image processing.

This thesis investigates which, how, and to what extent medical image analysis tasks can benefit from point cloud and graph representations in combination with deep learning methods. Various novel methods are developed and presented, covering a wide range of medical applications. Point clouds from multiple time-of-flight cameras are analysed for context monitoring in the operating room. Graph learning for semantic segmentation is investigated for both body part extraction from point cloud models as well as surface prediction on dense medical scans. Another important part of this work is the incorporation of keypoint graphs into deep learning based registration frameworks. Combinations of decoupled feature extraction using convolutional/graph neural networks and efficient message passing algorithms have been developed and a new deep learning architecture for optimising correspondence maps on a sparse keypoint graph is being proposed.

The methodological contributions of this work all demonstrate the feasibility and potential advantages of graph-based approaches in terms of improved accuracy as well as memory and runtime efficiency. In many applications, they can provide clear added value and should therefore be considered more often as a replacement or to complement the prevailing dense, fully convolutional deep learning models. The findings of this work and further research questions that can be derived from the methods presented indicate great potential for future developments of graph learning in medical image analysis.

Zusammenfassung

Der allgemeine technologische Fortschritt sorgt seit Jahrzehnten dafür, dass die automatisierte und computergestützte medizinische Bildanalyse ein wichtiger Aspekt der klinischen Praxis ist und bleiben wird. In den letzten Jahren ist es den Methoden des maschinellen Lernens und insbesondere des Deep Learning gelungen, neue Qualitätsmaßstäbe in diesem Bereich zu setzen. Der Fokus liegt dabei vor allem auf der Verarbeitung von dichten medizinischen 2D- und 3D-Scans, sowohl in klinischen Anwendungen als auch in der Forschung. In vielen Fällen stellen jedoch Datenstrukturen wie Punktwolken oder Keypoint-Graphen eine natürlichere, direktere Darstellung anatomischer Strukturen, z.B. des Gefäßsystems des Herzens/Auges oder der Atemwege der Lunge) dar. Darüber hinaus bieten sie im Allgemeinen das Potenzial für einen reduzierten Speicherbedarf und eine effizientere Verarbeitung medizinische Bilder. Auch aufgrund der jüngsten methodischen Fortschritte auf dem Gebiet des Geometric Deep Learning, insbesondere der erfolgreichen Generalisierung von Machine Learning Methoden auf irregulären Domänen, besteht ein zunehmendes Interesse an der Verwendung von Graph-basierten Ansätzen in der medizinischen Bildverarbeitung.

In dieser Arbeit wird untersucht inwieweit medizinische Bildanalyseaufgaben von Punktwolken- und Graph-Repräsentationen in Kombination mit Deep-Learning-Methoden profitieren können. Es werden verschiedene Methoden entwickelt und vorgestellt, die ein breites Spektrum an medizinischen Anwendungen abdecken. Punktwolken von mehreren Time-of-Flight-Kameras werden für die Kontextüberwachung im Operationssaal analysiert. Graph-Lernen zur semantischen Segmentierung wird sowohl für die Extraktion von Körperregionen aus Punktwolkenmodellen als auch für die Oberflächenprädiktion auf medizinischen Scans untersucht. Ein weiterer wichtiger Teil dieser Arbeit ist die Einbeziehung von Keypoint-Graphen in Deep-Learning-basierte Registrierungsverfahren. Dazu werden Kombinationen aus entkoppelter Merkmalsextraktion unter Verwendung von Convolutional/Graph Neural Networks und effizienten Message Passing Algorithmen entwickelt und eine neue Deep-Learning-Architektur zur Optimierung von Korrespondenzen auf einem Keypoint-Graphen vorgestellt.

Die methodischen Beiträge dieser Arbeit demonstrieren die potentiellen Vorteile von Graph-basierten Ansätzen in Bezug auf verbesserte Genauigkeit sowie Speicher- und Laufzeit-Effizienz. In vielen Anwendungen können sie einen klaren Mehrwert bieten und sollten entsprechend häufiger als Ersatz oder Ergänzung zu den vorherrschenden dichten Deep-Learning-Modellen in Betracht gezogen werden. Die Ergebnisse dieser Arbeit und weitere Forschungsfragen, die sich aus den vorgestellten Methoden ableiten lassen, weisen auf ein großes Potenzial für das Graph-Lernen in der medizinischen Bildverarbeitung hin.

Contents

Chapter 1

Introduction

1.1 Motivation

For decades, general technological progress has ensured that automated and computer-assisted medical image analysis has and will continue to become an increasingly important aspect of clinical practice. In recent years, machine learning and deep learning methods in particular have succeeded in setting new standards in the field, including diagnosis, disease monitoring, preoperative planning, intraprocedural guidance, and treatment [De Fauw et al., 2018; Fauser et al., 2019; Lou et al., 2019; Nguyen et al., 2020; Sardar et al., 2019]. By far the most commonly generated imaging data in medical facilities worldwide are gridded two-dimensional and volumetric scans, e.g. X-ray or ultrasound images and MRI or CT scans, respectively. In Germany alone, about 130 million X-ray and 12 million MRI examinations are performed every year [Nekolla et al., 2017; OECD, 2021]. The past and current focus of research and applied methods on these input modalities and dense processing of the data is therefore rarely questioned. However, the description of medical (image) data as continuous, sparse point clouds and graph structures offers enormous potential in many respects and attracts an increasing amount of attention. Figure 1.1 shows an exemplary overview of different types and use cases of point clouds and graphs in medical applications.

Often, graphs are direct and *natural representations of human anatomical structures*. The vascular system of the heart and the eye or the airways of the lungs are typical examples of anatomies based on graph (or tree) structures, but at least part of this inherent geometric information is lost when stored in a dense, fixed grid (angiograms, fundus images, CT). By extracting and transferring the anatomy into a more appropriate graph representation, the rich structural information can be discovered and exploited by learning-based automatic algorithms.

Point clouds and graphs are an *alternative type of input modality* in the context of medical imaging that enables a number of exciting applications. With the availability of affordable distance measuring devices, such as time-of-flight cameras, it is now feasible to directly capture the geometry (represented as a point cloud) of an observed scene by measuring the distances between the camera and the objects in the field of view. Monitoring context information, such as body poses of surgeons and patients,

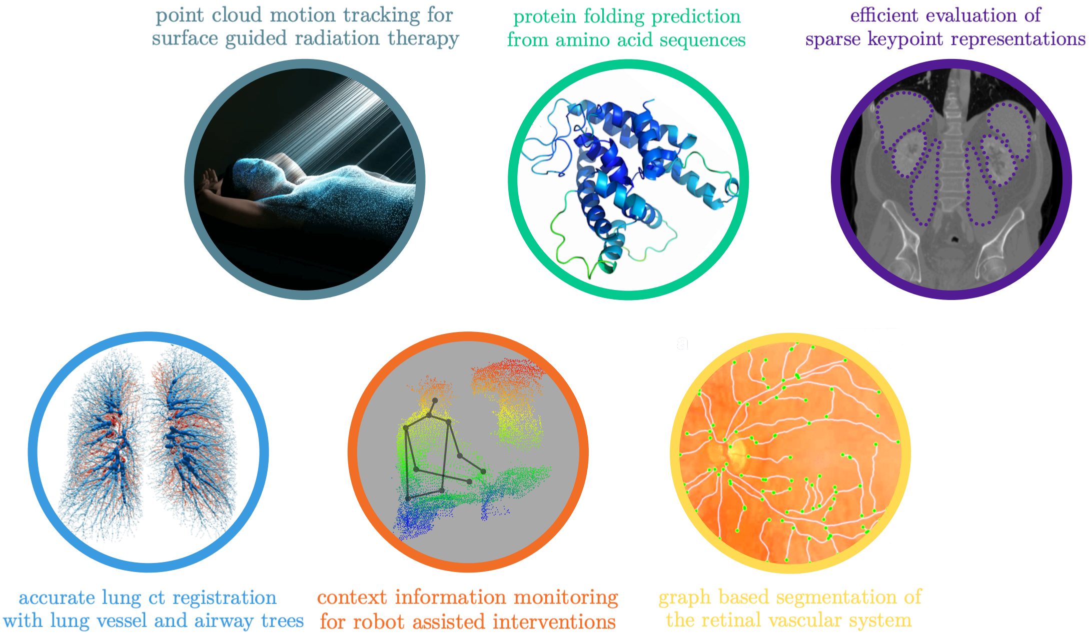

point cloud motion tracking for surface guided radiation therapy
protein folding prediction from amino acid sequences
efficient evaluation of sparse keypoint representations
accurate lung ct registration with lung vessel and airway trees
context information monitoring for robot assisted interventions
graph based segmentation of the retinal vascular system

Fig. 1.1: Medical image analysis methods are in the vast majority developed for processing dense grid-based arrays such as CT and MRI scans. This figure, on the other hand, exemplifies the wide variety of use cases for point clouds and graphs in medical applications. In particular, with the emergence of affordable time-of-flight cameras and methodological advances in machine learning on irregular domains, sparse point clouds and graph structures attract an increasing amount of attention in the field of medical image processing.
Brainlab AG, `brainlab.com/de/journal`, accessed 3 January 2022 (top left)
Deepmind, `deepmind.com/blog`, accessed 3 January 2022 (top center)
Shen et. al. 2021, `github.com/uncbiag/shapmagn`, accessed 3 January 2022 (bottom left)
Ramos et. al. 2018, `pubmed.ncbi.nlm.nih.gov/30458717`, accessed 3 January 2022 (bottom right)

for robot assisted interventions or marker-less motion tracking of a patient's torso for radiotherapy are only two of many clinical use cases [Bauer et al., 2013]. A recent breakthrough in a further medical field, protein folding prediction, is also based in large part on describing the relationships between individual amino acids as a graph, far surpassing previous fixed grid based approaches using distance histograms [Jumper et al., 2021].

Besides the advantage of the inherent available structural information, point clouds and graphs are *efficient data structures* and thus have potential further advantages in terms of speed, memory usage and precision. For grid based representations, a trade-off must be made between the required memory and the distinguishable spatial information. [Liu et al., 2019d] found for a point cloud segmentation benchmark that for a feasible approach ($\leq$ 12 GB of GPU memory) using a voxelised representation with Convolutional Neural Networks (CNNs) up to 50% of original information (distinguishable points) is lost, while processing the full point cloud directly outperforms the method with a 10 times lower memory footprint. When representing medical data as point clouds (e.g. extracted from edge maps or surface segmentations) or evaluating scans on only a subset of informative spatial locations (in arbitrarily shaped regions of interest) the number of required datapoints can be massively reduced from millions (of image voxels in a typical CT/MR scan) to only a few thousand (keypoints).

Recent developments in the field of machine learning enable the facilitated processing of such sparse image representations. Graph Neural Networks (GNNs) [Bronstein et al., 2017] are a class of deep learning methods for irregular domains, such as graphs, pointclouds or meshes, and can be seen as direct counterparts to CNNs for grid based learning. Unlike CNNs, which exploit the structured data input with fixed-size filter kernels, a GNN provides specific aggregation mechanisms to deal with the variable neighbourhood of nodes explicitly defined by the edges of the graph. In addition to conventional methods for processing information on graphs, such as diffusion or message-passing algorithms [Felzenszwalb et al., 2006], GNNs offer a learning-based alternative. A brief introduction to GNNs can be found in Section 2.2.2.

Despite the variety of use cases and potential described previously, the best possible integration of point clouds and graphs into medical (deep learning) algorithms, as well as the advantages for most clinical problems, are in many cases unknown and far less explored than methods that process dense, grid-based medical image data. This thesis therefore investigates *which, how, and to what extent medical image analysis tasks can benefit from point cloud and graph representations in combination with deep learning methods.*

1.2 Objectives

The overall aim of this thesis is the development of methods for medical image analysis tasks under the explicit consideration of sparse representations such as point clouds and graphs. Given the impact that Deep Learning has had on this field in recent years, the implications of learning methods in particular will also be investigated.

Based on the research question formulated at the very end of Section 1.1, the objectives of the work can be summarised in more detail along the following three aspects:

Which... Medical image analysis has become an essential part of clinical practice. The most prominent tasks besides classification include *landmark detection*, *semantic segmentation* and *image registration*, which are important components of a variety of medical imaging pipelines. The goal is to address all three of these areas and assess the feasibility (and ideally the merits) of integrating graph-based solutions. This broader view enables us to identify tasks that particularly benefit or where there is little/no difference in performance compared to conventional grid-based image processing.

New clinical applications through new input data types, such as *point clouds* from time-of-flight cameras, will be considered, but also the advancement of methods applied to common medical images such as CT and MRI scans (e.g. by extracting and processing informative *keypoint graphs*).

How... Sparse (3D) point clouds and keypoint graphs can be processed either as *projections* (cf. X-Ray vs. CT), *dense voxelisations* or *directly.* Each alternative may have its own advantages and disadvantages, which will be investigated. For example, standard 2D and 3D spatial filters (e.g. in CNNs) can be efficiently applied to projections and voxelisations, respectively, but information may be lost and the representation becomes unnecessarily dense. Working directly on the graphs, on the other hand, brings different new challenges. The first problem when extracting keypoint graphs from medical scans is the creation of the graph itself. The optimal sampling of the necessary keypoints is a design choice and depending on the type of application, different methods may be suitable. In this work, three approaches are considered: 1) explicit *prediction of sampling locations*, 2) use of a heuristic *interest point detector*, and 3) *randomised selection* of points within a suitable region of interest (ROI).

The methodological focus of this work is on the automatic analysis of medical image data with the help of graph representations. Various learning and non-learning based methods are used to achieve this goal. Conventional grid-based *2D or 3D feature extractors* (handcrafted, CNNs) can be applied to projection images and dense voxel representation or be evaluated at sparse keypoint locations only later in the process, which also enables a desirable separation of feature extraction and optimisation. When processing information directly on graphs, the use of *Graph Neural Networks* (GNNs) is investigated in particular depth, both for the extraction of informative feature descriptors and for learnable graph-based optimisation. In addition, conventional graph-based methods such as *diffusion* approaches, *Gaussian Mixture Models* (GMMs) or *message passing* algorithms are considered (often together with deep learning based feature extraction).

To what extent... A main goal is to advance medical image analysis methods using graph representations, which can be measured by widely used performance metrics such as the F1 score for semantic segmentation labels or landmark distances for image registration tasks. Since point clouds and graphs are efficient data structures and further offer potential for theoretical improvements in runtime and memory usage (which can be of great benefit in clinical practice), this should also be reflected in the evaluation.

1.3 Organisation and Contributions

Throughout this thesis, a series of methodological approaches for the integration of point clouds and graphs in medical imaging tasks are developed, evaluated and discussed. The organisation can be divided into three parts. First, Chapter 2 introduces general information about the medical image analysis tasks addressed in this thesis and methodological foundations of relevant deep learning and graphical inference methods. The main part comprises the methodological Chapters 3 to 6. Each of these chapters is structured in a consistent manner and begins with a brief introductory summary, placing it in the context of the thesis. Then, new methodological developments on the topic are described, following a self-contained presentation scheme: 1. Introduction (including motivation, relevant related work and scientific contributions), 2. Methods (proposal and explanation of the new approach), 3. Experiments and Results and

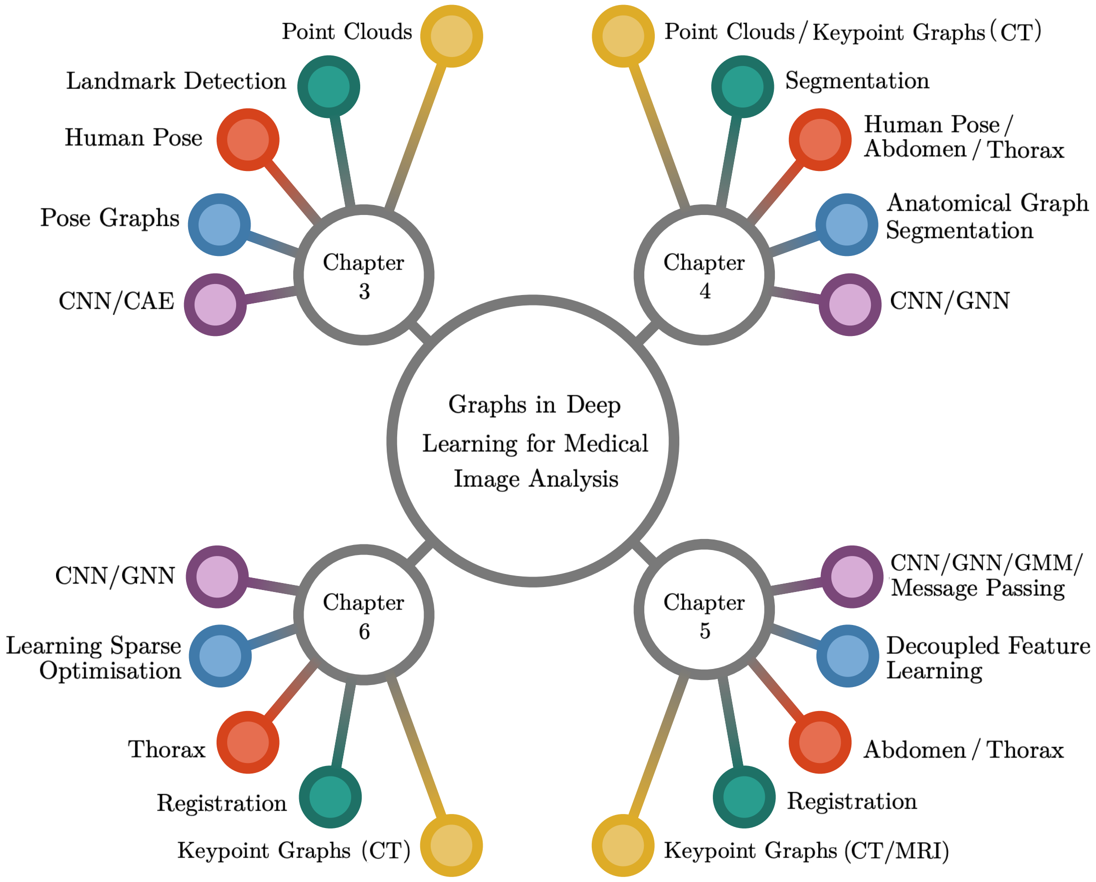

Point Clouds
Landmark Detection
Human Pose
Pose Graphs
CNN/CAE
Chapter 3
Point Clouds / Keypoint Graphs (CT)
Segmentation
Human Pose / Abdomen / Thorax
Anatomical Graph Segmentation
CNN/GNN
Chapter 4
Graphs in Deep Learning for Medical Image Analysis
CNN/GNN
Learning Sparse Optimisation
Thorax
Registration
Keypoint Graphs (CT)
Chapter 6
CNN/GNN/GMM/ Message Passing
Decoupled Feature Learning
Abdomen / Thorax
Registration
Keypoint Graphs (CT/MRI)
Chapter 5

Fig. 1.2: Schematic overview of the methodological chapters. The contents cover a wide range of different input modalities 🟨, medical image analysis tasks 🟩 anatomies 🟥, scientific/clinical topics 🟦 and (learning) methods 🟪.

4. Discussion and Conclusion. Figure 1.1 illustrates the thematic organisation of the individual chapters, while the following chapter outline highlights the main scientific contributions:

- Chapter 3 deals with the detection of anatomical landmarks (of the human upper body pose) in point clouds of an operating scene generated from multiple time-of-flight sensors. The point clouds are converted to grid-based representations (2D projection images and 3D voxel occupancy grids) and processed with conventional CNN architectures. In a multi-view setup the *utilisation of 2D projection images enables the exploitation of widespread pre-trained 2D CNNs*, which boosts the landmark detection accuracy when the results are fused in 3D. For the frequent occurring of implausible poses (e.g. partially swapped left and right body side) a mitigation strategy is developed. A self-supervised Convolutional Autoencoder (CAE) learns an *implicit pose graph embedding that restricts the output space to anatomical plausible predictions.* The developed methods were published in:

  [Hansen et al., 2019b] Hansen, L., Diesel, J., and Heinrich, M. P. "Regularised Landmark Detection with CAEs for Human Pose Estimation in the Operating Room". In: *Bildverarbeitung für die Medizin 2019 –BVM 2019.* 2019, pp. 178–183. *BVM Award for the second best scientific paper.*

  [Hansen et al., 2019e] Hansen, L., Siebert, M., Diesel, J., and Heinrich, M. P. "Fusing Information From Multiple 2D Depth Cameras for 3D Human Pose Estimation in the Operating Room". *International Journal of Computer Assisted Radiology and Surgery* 14 [11], 2019, pp. 1871–1879.
  Invited BVM 2019 special issue paper - IF: 2.924.

- Chapter 4 marks the transition from the grid-based processing to the explicit exploitation of the point cloud and graph structures and properties. In the first part, a point cloud of the human body again forms the input, but in contrast to Chapter 3, a dense pointwise semantic segmentation is to be predicted. For this purpose, a supervised learning method is developed that can *extract semantic features using only diffusion operators and isotropic 1×1 convolutions*, illustrating the feasibility and robustness of learning on graphs. The second part investigates possibilities to transfer graph based learning for semantic segmentation to dense medical images. *CNNs, that extract local features from image based intensity patches, and GNNs based on an anatomical keypoint graph, that enable global communication, are combined* in a supervised learning framework for the task of semantic edge detection in abdominal CT and X-Ray images. In comparison to a fully convolutional dense processing of the scans, the approach shows superior performance both in accuracy and efficiency. The presented methods are based on the following two publications:

[Hansen et al., 2019a] Hansen, L., Diesel, J., and Heinrich, M. P. “Multi-Kernel Diffusion CNNs for Graph-Based Learning on Point Clouds”. In: *Geometry Meets Deep Learning –ECCV 2018 Workshops.* 2019, pp. 456–469.

[Hansen et al., 2019d] Hansen, L. and Heinrich, M. P. “Sparse Structured Prediction for Semantic Edge Detection in Medical Images”. In: *International Conference on Medical Imaging with Deep Learning –MIDL 2019.* 2019, pp. 250–259.
Early accepted (27% of all submissions) - 40% acceptance rate.

- In the previous chapters, feature extraction and task-specific inference are always considered jointly. In contrast, Chapter 5 investigates the decoupling of feature learning and graph-based optimisation for the task of medical image registration. As proposed in the first part of this chapter, sparse keypoint graphs allow for particularly *efficient inference, which can be achieved by a combination of CNN-based supervised feature extraction and iterative graphical message passing algorithms.* Subsequently, methods are presented that use GNNs to extract features on anatomical keypoint graphs, which can then be processed by conventional inference methods (Coherent Point Drift, Loopy Belief Propagation). For the considered task of exhale to inhale lung CT alignment the *sparse keypoint registration methods boosted with learned graph features* surpass the non-learned counterpart as well as dense learning based methods both in accuracy and runtime (even without any image/intensity information). All developed algorithms are described in published works:

  [Hansen et al., 2021c] Hansen, L. and Heinrich, M. P. “Revisiting Iterative Highly Efficient Optimisation Schemes in Medical Image Registration”. In: *Medical Image Computing and Computer Assisted Intervention –MICCAI 2021.* 2021, pp. 203–212.
  33% acceptance rate.

  [Hansen et al., 2019c] Hansen, L., Dittmer, D., and Heinrich, M. P. “Learning Deformable Point Set Registration with Regularized Dynamic Graph CNNs for Large Lung Motion in COPD Patients”. In: *Graph Learning in Medical Imaging –MICCAI 2019 Workshops.* 2019, pp. 53–61.

  [Hansen et al., 2021a] Hansen, L. and Heinrich, M. P. “Deep Learning Based Geometric Registration for Medical Images: How Accurate Can We Get Without Visual Features?” In: *Information Processing in Medical Imaging –IPMI 2021.* 2021, pp. 18–30.
  Oral presentation (9% of all submissions) - 30% acceptance rate.

- While the focus of Chapter 5 is on learned input features and conventional inference, Chapter 6 presents a sparse and learnable optimisation framework for medical image registration. Handcrafted input image features and a *combination of CNNs and GNNs allow for a smooth local and global regularisation of similarity costs on a* *sparse keypoint graph.* The approach sets the current state-of-the-art for learning based methods on breath-hold lung CT and is published in:

[Hansen et al., 2021b] Hansen, L. and Heinrich, M. P. "GraphRegNet: Deep Graph Regularisation Networks on Sparse Keypoints for Dense Registration of 3D Lung CTs". *IEEE Transactions on Medical Imaging* 40 [9], 2021, pp. 2246–2257.
IF: 10.048.

Finally, in Chapter 7, the presented methods are summarised and main findings are discussed with regard to this thesis' research question of *which, how, and to what extent medical image analysis tasks can benefit from point cloud and graph representations in combination with deep learning methods.* To conclude with, ongoing and further promising research directions are highlighted.

Chapter 2

Background

The purpose of this chapter is to provide brief and concise foundations for the algorithms presented in the following methodological chapters. Section 2.1 first summarises the medical image analysis tasks considered in this thesis - landmark detection, semantic segmentation and image registration - and describes datasets and metrics used. Methodological background is then given in Section 2.2, which includes descriptions of Deep Learning using Convolutional Neural Networks, their counterparts for non-regular domains, Graph Convolutinal Networks, as well as non-learning based graphical models.

2.1 Medical Image Analysis Tasks

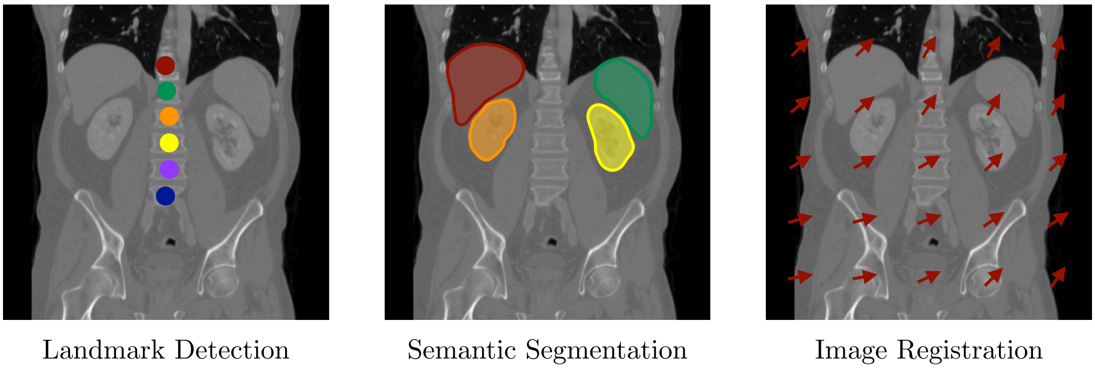


Fig. 2.1: Schematic overview of different exemplary medical image analysis tasks.

The image analysis tasks addressed in this work include landmark detection, semantic segmentation and image registration (see Figure 2.1). Landmark detection is investigated in the context of human pose estimation, specifically predictions of joints from point clouds, for clinical context monitoring in Chapter 3. For the task of semantic segmentation, two different applications are discussed in Chapter 4, namely body part prediction, again in point cloud data, and edge detection on CT and X-Ray scans. The last two methodological chapters both deal with medical image registration, more specifically with registration of inter- and intra-patient scans on CT and MR and shallow and breath-hold lung CT scans.

2.1.1 Landmark Detection

Landmark detection is the task of localising and identifying specific keypoints in an image. It is closely related to the more general problem of object detection but instead of a bounding box, a single coordinate must be predicted for each landmark. Landmark detection is a fundamental problem in computer vision, with two main applications being face recognition and human pose estimation. Respective surveys give a broad overview of the topics [Chen et al., 2020; Wu et al., 2019]. Medical image analysis has a wide range of applications for landmark detection, for example as initial step in automatic determination of standard (2- and 4-chamber) views in cardiac MRI [Le et al., 2017] or for computation of fetal biometric parameters such as biparietal diameter, femur diaphysis length and head circumference [Salomon et al., 2019]. A comprehensive overview is included in [Litjens et al., 2017]. Landmark detection methods can be broadly divided into atlas-, appearance- and image-based approaches (cf. [Alansary et al., 2019]). Atlas-based methods rely on accurate registration between a target and reference image to transfer landmarks. Appearance and/or shape models leverage spatial priors to infer keypoints, while image-based methods include all approaches that learn descriptive features for the landmarks directly from the raw input. The last category includes, in particular, deep learning methods using convolutional neural networks, which are now predominantly used and are also applied in this work. For deep learning approaches, different variants exist to eventually regress the landmarks from the learned image features (see Figure 2.2). Formally, we can define the task of predicting a set of N distinctive landmarks $L = l_1, \ldots, l_N$ from an input volume V as $L = \Theta(V)$, where Θ is a universal function approximator, parametrised by the weights of a CNN. In this simplest setting, the landmark coordinates are regressed directly. However, it has been shown that structured predictions can lead to more accurate localisations in many applications [Newell et al., 2016]. For this purpose, Gaussian heatmaps $H = H_1, \ldots, H_N$, which represent the landmark positions L, are used as supervision for training the network, i.e. $\tilde{H} = \Theta(V)$, where $\tilde{H}$ are the heatmaps predicted by the network. Then, during inference, a landmark coordinate l_i can be obtained as $l_i = arg\,max\,\tilde{H}_i$. A potential problem for certain applications is that the argmax function is not differentiable and thus no further loss functions or the like can be defined directly on the landmark coordinates. The alternative is to use an argsoftmax function where the heatmap (normalised to sum to one) is integrated over the voxel coordinates c. Thus, a landmark l_i can be derived as

$$l_i = \sum_c c * \tilde{H}_i(c). \tag{2.1}$$

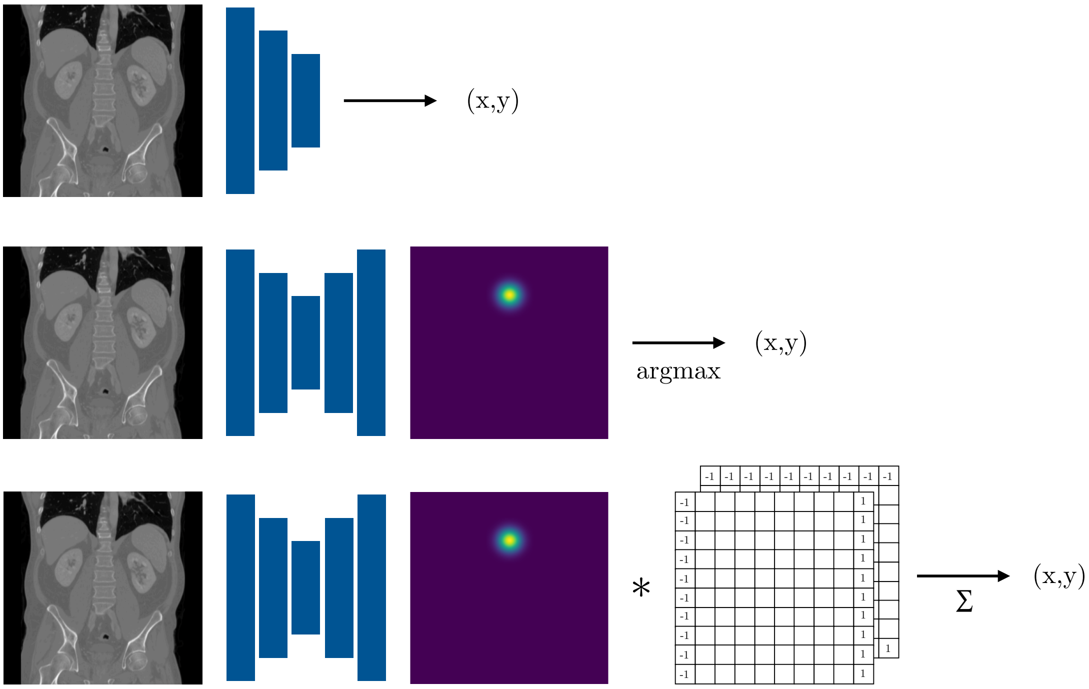


Fig. 2.2: Illustration of commonly used landmark regression variants in CNN-based methods.

Datasets This thesis addresses the task of landmark detection in the context of clinical monitoring of surgeons in an operating room by predicting body joints. The dataset used for training and evaluation is called MVOR (Multi-View Operating Room) [Srivastav et al., 2018] and consists of 732 multi-view frames from three RGB-D cameras providing registered depth and anonymised rgb images synchronised in time. The images were recorded in an operating room at the University Hospital of Strasbourg over a period of four days. For the task of 2D and 3D pose estimation the ground truth annotations consist of 2926 and 1061 upper-body poses, respectively. An upper-body pose is defined by a total of ten joints (upper head, neck, shoulders, elbows, wrists and hips).

Metrics For evaluation in 3D the mean joint position error (MPJP), i.e. the Euclidean distance between the predicted and ground truth landmarks in centimetres, can be used. However, for landmark detection on 2D images (as in Section 3.1) the metric must account for different scales: The percentage of correct keypoints (PCK) with a threshold of α assumes a prediction as correct if it falls within $\alpha \cdot \max(bbox_h, bbox_w)$ pixels of the ground truth annotation, where $bbox_h$ and $bbox_w$ are the height and width of the persons enclosing bounding box, respectively.

2.1.2 Semantic Segmentation

Semantic segmentation describes the pixelwise classification of images and is an important tool for the automatic understanding and further processing of objects in images, which is used in a variety of applications. In the field of computer vision, these are, for example, the segmentation of roads for autonomous driving [Feng et al., 2020], the segmentation of workpieces on production lines for industrial quality inspection [Wu et al., 2020] or the generation of real-time maps from satellite images [Mohanty et al., 2020]. Application areas in medical imaging are similarly numerous and diverse, as discussed in [Litjens et al., 2017]. Prominent examples include quantitative monitoring of tumours in initial and follow-up scans [Myronenko, 2018], segmentation of organs and other anatomical structures for the extraction of clinical parameters (volume, diameter, etc.) [Ghavami et al., 2019] or labelling of target structures for computer-assisted interventions [Wang et al., 2019b]. The simplest methods for semantic segmentation are point-based, which means that only the intensity values of the pixels themselves are considered for classification. Single or multiple thresholds are used to generate the segmentations. To achieve more coherent results, neighbourhoods or regions can be included in the process, as for instance in the region growing algorithm [Adams et al., 1994]. In most cases, this approach requires user interaction, for example to create initial scribbles in the image. The objective of edge-based methods is to delineate the contours of target objects as accurately as possible. This class of methods includes, for example, live wire segmentation [Barrett et al., 1997] or active contours (snakes) [Chan et al., 2001]. All previous methods have mainly processed local image data, whereas model-based approaches can consider global information and a priori knowledge. The most popular of these are statistical shape models [Heimann et al., 2009] and deep learning algorithms, which are now the most widely methods used for semantic segmentation. The U-Net [Ronneberger et al., 2015], a fully convolutional encoder decoder architecture that combines low level and high level information through skip connections, has emerged as a standard tool for this task. With the nnUNet [Isensee et al., 2021], there also exists a sophisticated software framework that makes time-consuming hyperparameter searches obsolete and yields state-of-the-art results for a variety of datasets.

Datasets Three different datasets are used for training and evaluating the semantic segmentation algorithms proposed in Chapter 4. The first dataset, employed for point cloud segmentation in Section 4.1, is the publicly available FAUST dataset [Bogo et al., 2014], which consists of 100 surface meshes of 10 different subjects, each scanned in 10 different poses (training: subjects 1-7 (70 shapes), validation: subject 8, (10 shapes), test: subjects 9-10 (20 shapes)). The 3D meshes have a resolution of 6890 vertices and point-wise correspondences between the shapes have been semi-automatically established for all points. The points were manually labelled on a reference shape and transferred to all other shapes via the known point correspondences. The 15 labels correspond to the head, thorax, abdomen, left/right hand, left/right lower arm, left/right upper arm, left/right foot, left/right lower leg and left/right upper leg. The popular public VISCERAL [Jimenez-del-Toro et al., 2016] and JSRT [Shiraishi et al., 2000] datasets serve for the task of dense segmentation of CT and X-Ray scans in Section 4.2, respectively. The VISCERAL dataset is composed of of CT and MR scans of 30 subjects, with up to 20 expert annotated anatomical structures per scan, such as liver, spleen, and kidneys. The JSRT dataset consists of 247 X-Ray scans, which are extended by [Van Ginneken et al., 2006] with manually delineated contours of the lung, heart and clavicles.

Metrics Commonly used metrics for segmentation evaluation are the Dice similarity coefficient (DSC) and the 95% Hausdorff distance (HD95). The DSC measures the overlap between two binary segmentation labels as $DSC = \frac{2TP}{2TP+FP+FN}$, where TP, FP and FN describe true positive, false positive and false negatives matches. The Hausdorff distance, on the other hand, describes the maximum surface distance between two segmentations. To reduce the effect of outliers, the HD95 metric uses the 95th percentile instead of the maximum distance.

2.1.3 Image Registration

Image registration is the process of spatially transforming an image so that it is as similar as possible to a reference image in terms of a particular image metric. It is an important tool in general image processing, for example when stitching together individual satellite images [Ma et al., 2018] or when creating correspondences in SLAM frameworks [Sarlin et al., 2020], but its main field of application is medical imaging. An overview on the subject of medical image registration can be found, for example, in [Rueckert et al., 2019]. Examples of applications include fusion of scans from different modalities (e.g. CT and MRI) to provide comprehensive visual information during surgery [Heinrich et al., 2013b], progress monitoring of nodules in follow-up lung CT scans after interventions [Zheng et al., 2007] or markerless motion compensation during radiotherapy [Seregni et al., 2017]. Registration of a fixed ($\mathcal{F}$) and a moving ($\mathcal{M}$) image is usually described as an optimisation problem of the form

$$\underset{\varphi}{\operatorname{argmin}}\, \mathcal{D}(\mathcal{F}, \varphi \circ \mathcal{M}) + \lambda \mathcal{R}(\varphi), \tag{2.2}$$

where φ denotes a transformation model (e.g. an affine transformation matrix, a B-spline parametrised model or a dense displacement vector field), $\mathcal{D}$ a distance metric, that measures the similarity of the fixed and warped moving image and $\mathcal{R}$ an additional regularisation term weighted by λ that enforces smoothness of the transformation (typically by constraining spatial derivatives). Continuous optimisation has been one

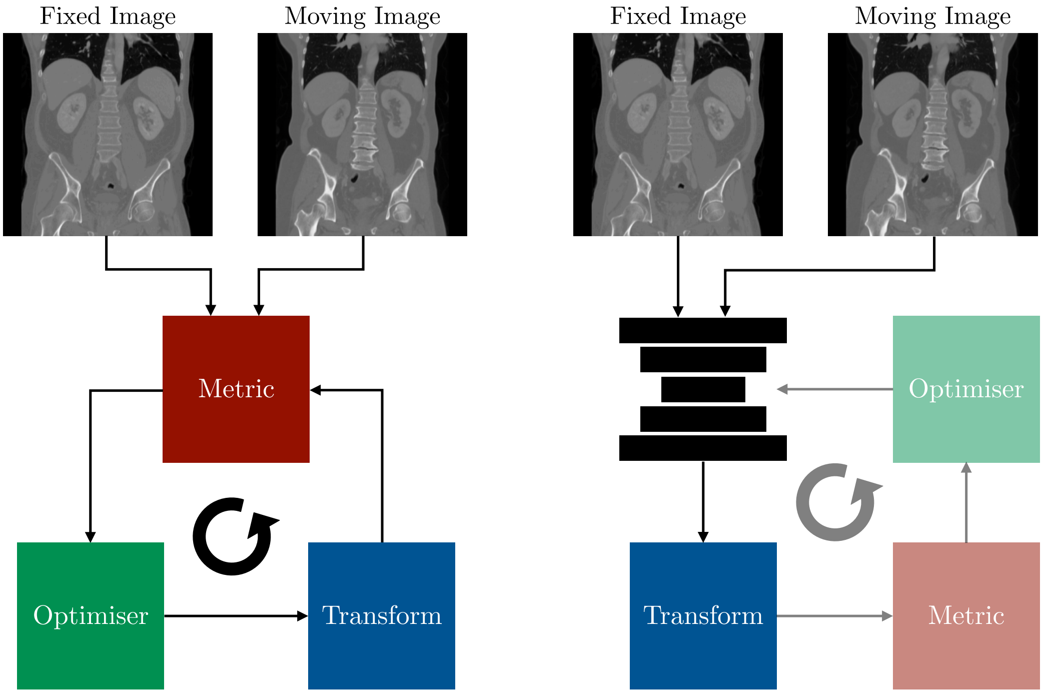


Fig. 2.3: Schematic side-by-side comparison between conventional (iterative) (left) and Deep Learning based image registration (right). Components with reduced opacity indicate that they are only being used during an offline training phase.

of the most widely used registration methods in the last decades. The main idea is that the parameters of a transformation model are iteratively updated using a gradient descent method so that equation 2.2 is minimised. Approaches differ primarily in the choice of the transformation model, similarity metric and optimisation method. Other algorithmic decisions include the use of multi-scale strategies and the type of interpolation method. An example of a popular framework for iterative continuous registration is SimpleElastix [Marstal et al., 2016], which implements a variety of choices for the aspects mentioned. Recently, several deep-learning registration methods have been proposed that drastically reduce runtimes by moving the optimisation to an offline training phase [Balakrishnan et al., 2019; Hu et al., 2018; Vos et al., 2019]. At their core, however, they strongly resemble conventional methods, since the same objective function is minimised (using the same similarity metrics, transformation models, etc.). Figure 2.3 illustrates both approaches. Another class of algorithms to be mentioned, which like continuous optimisation belongs to the conventional (non learning-based) approaches, are discrete methods, based on MRF (see Section 2.2.3 for details) optimisation schemes [Glocker et al., 2011; Heinrich et al., 2013a].

Datasets Two abdominal CT/MRI and two lung CT datasets are used for training end evaluation of registration methods in Chapters 5 and 6. Both abdominal datasets are from the Learn2Reg competition [Hering et al., 2022], the first containing 30 CT scans with 13 manually labelled anatomical structures (spleen, left/right kidney, gallbladder, oesophagus, liver, stomach, aorta, inferior vena cava, portal vein and splenic vein, pancreas and left/right adrenal gland) for interpatient registration, while the second consists of 16 intrapatient CT-MRI scan pairs with liver, spleen and kidney segmentations. All further methods are evaluated on the DIR-Lab 4DCT and COPDgene datasets [Castillo et al., 2009, 2013], each consisting of 10 inhale and exhale scan pairs (shallow breathing for 4DCT and breath-hold scans for COPDgene) with 300 manually annotated landmark correspondences.

Metrics An important accuracy metric for image registration is the target registration error (TRE), that measures Euclidean distances of fixed and corresponding warped moving landmarks. This of course requires manual landmark correspondences that are time-consuming to annotate and therefore only available in a few datasets. An alternative is to evaluate the alignment of segmentation labels after registration, for which the same metrics as for image segmentation (see 2.1.2) can be used. In addition to the registration accuracy, the smoothness of the deformation field, which indicates plausibility of the transformation, is an important evaluation criterion and can be derived from the standard deviation of the Jacobian determinant of the deformation field [Leow et al., 2007].

2.2 Methodological Background

The methods developed in this thesis investigate graph approaches in medical deep learning and build on basic methods from the respective domains. Section 2.2.1 describes Convolutional Neural Networks, as the most prominent deep learning architectures for structured processing, which are extensively adopted as components throughout this work. Graph Neural Networks as counterparts to CNNs for irregular domains are then summarised as they play an equally important role in the development of the methods presented. Finally, graphical models and message-passing algorithms are described, which are mainly used for the decoupled feature learning and optimisation for image registration presented in Chapter 5.

2.2.1 Convolutional Neural Networks

Artificial neural networks represent the current state of the art for many machine learning applications. Especially in the field of computer vision, continuous progress in image understanding has been achieved in recent years with the help of convolutional neural networks (CNNs). Although modern CNNs were described and successfully

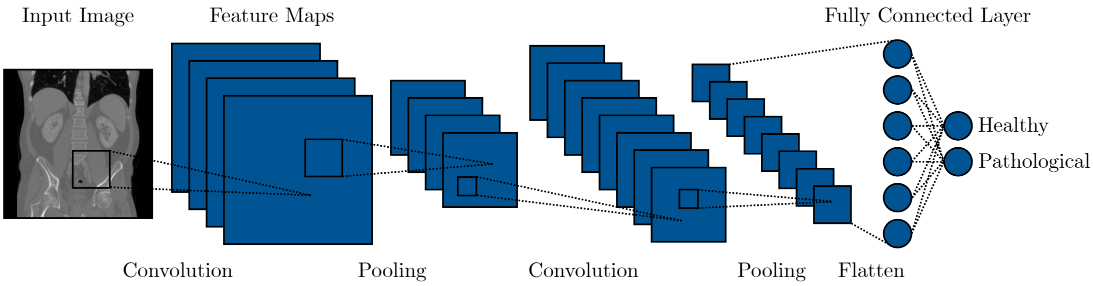


Fig. 2.4: A typical convolutional neural network architecture for classification of pathological indications in medical CT scans.

applied for the recognition of handwritten digits as early as 1989 in [LeCun et al., 1989], the begin of the current technological leap through AI can be attributed in particular to the win of the ImageNet Data Challenge [Deng et al., 2009] by [Krizhevsky et al., 2012] in 2012. The availability of large amounts of annotated image data as well as high performance GPUs for the computationally intensive training process have contributed significantly to this success. For a comprehensive introduction and overview of Deep Learning including CNNs, see [Goodfellow et al., 2016].

Figure 2.4 illustrates a typical convolutional neural network, exemplified by the classification task of identifying pathological CT scans. It consists of alternating convolutional and pooling layers to extract descriptive features (feature maps) from the input image. This is followed by one or more fully connected layers with the number of output neurons corresponding to the number of possible classes (two in the example: healthy and pathological). Convolutional layers differ from fully connected layers, where each input neuron is connected to each output neuron, by the spatially organised, locally limited filters that are applied to the entire image in a sliding window approach. The advantage of the shared weights is a significant reduction of network parameters and exploitation of the generally assumed translation invariance of image features. After every few convolutional layers, pooling operations are employed, which reduce the resolution of the feature maps and thus not only contribute to further data reduction, but also increase the receptive field and prevent overfitting. The objective of training the CNN is that, given an input image, the output neuron corresponding to the class of the image has the highest activation. This can be achieved in a purely data-driven manner by adjusting all trainable network parameters using a suitable loss function and the backpropagation algorithm. In order to model non-linear decision boundaries and to ease training, activation functions (e.g. Rectified Linear Units (ReLUs)) are usually applied after convolutional layers. Dropout [Srivastava et al., 2014], batch normalisation [Ioffe et al., 2015], residual connections [He et al., 2016] and many other techniques have ensured a continuously improved and more robust training. For pixel-level tasks such as segmentation or registration, special convolutional network architectures have been developed, of which the U-Net [Ronneberger et al., 2015] is the most well-known and commonly used. It consists of an encoder path with convolutional and pooling layers and a symmetric decoder path in which feature maps recover higher resolutions through upsampling (or learnable transposed convolutions). As in the classification task, global features and context are learned in the encoder part and then spatially propagated with each layer in the decoder part. An important part of the architecture are skip connections (implemented as concatenation) from the feature maps of the encoder to the corresponding feature maps of the decoder, which allow important local information to be preserved. Works such as the nnUNet incorporate the UNet in a comprehensive software framework that can be used to configure hyperparameters (number of layers, feature channels) fully automatically and achieve robust state-of-the-art results for medical segmentation tasks. Work such as the nnUNet [Isensee et al., 2021] incorporates the UNet into a comprehensive software framework that, among other things, allows hyperparameters (number of layers, feature channels) to be configured fully automatically and thus robust state-of-the-art results to be achieved for a variety of medical segmentation tasks.

2.2.2 Graph Neural Networks

For the central topic of this thesis, deep learning on graphs for medical image processing, CNNs are only partially applicable, namely whenever dense representations are considered, e.g. in Chapter 3, where point clouds are first voxelised and then processed. For direct learning on point clouds or keypoint graphs, however, it is necessary to employ methods of geometric deep learning, a relatively new research field that aims to generalise neural network models to non-regular domains, comprehensively outlined in [Bronstein et al., 2017]. First approaches were based on the spectrum of the graph Laplacian [Bruna et al., 2014], but an inherent drawback of these methods are that they rely on prior knowledge of the graph structure to define local neighbourhoods. To gain independence from a fixed graph topology, [Kipf et al., 2017; Mikael et al., 2015] proposed to limit the support size of the learned spectral filters. More recent alternative approaches employ spatial instead of spectral filters. These include for example the PointNet [Qi et al., 2017a], which simply ignores local neighbourhoods, or methods that learn local parameterised patches using Gaussian mixture models [Monti et al., 2017] or triangular meshes [Masci et al., 2015]. They can be considered as special cases of EdgeConvs [Wang et al., 2019a], which describe a more general concept of graph convolutions and, apart from initial experiments on learnable diffusion approaches in Chapter 4, are adopted in most of the methods presented in this work. Following the notation of [Wang et al., 2019a], the general concept of graph convolutions is introduced, while specific implementations, including network architectures, are described directly in the respective method chapters. Figure 2.5 illustrates the EdgeConv operation. A Graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is defined as a set of n vertices $\mathcal{V}$ and edges $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$. Each vertex

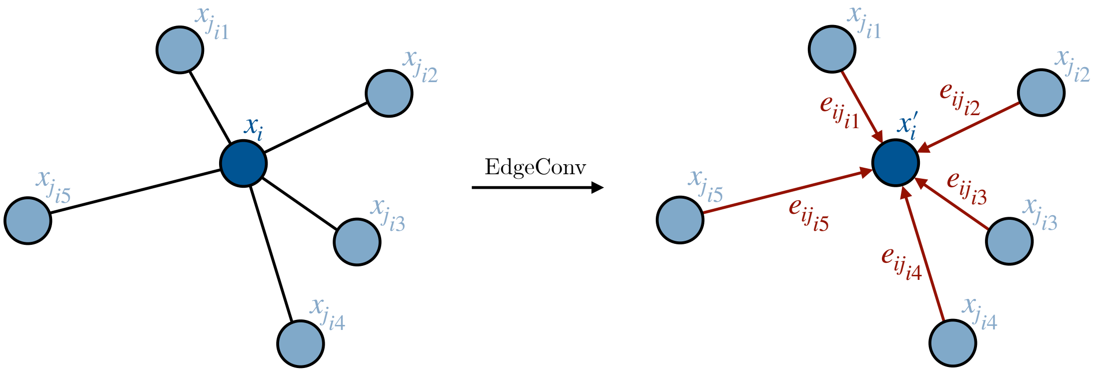


Fig. 2.5: Illustration of an EdgeConv (adapted from [Wang et al., 2019a]).

is represented by an F-dimensional feature vector $x_i \in \mathbb{R}^F$, which in the simplest case, e.g. when extracting keypoints from medical 3D scan, consists of three coordinates, but can be extended depending on the application, e.g. by flattened patches of intensity values. In the methods presented, the graph itself is usually created from the k-nearest neighbors defined on the keypoint coordinates. However, with additional knowledge, it could also be constructed more specifically for a particular application, e.g. if bifurcations in anatomical graphs such as the airway tree are known. The EdgeConv operation for a node x_i consists of two steps: 1) the computation of edge features e_{ij} for all neighboring nodes and 2) a symmetric aggregation function $\square$, i.e. the feature vector is updated as

$$x'_i = \underset{j:(i,j)\in\mathcal{E}}{\square} e_{ij}. \tag{2.3}$$

The aggregation function $\square$ can be, for example, the sum, average or maximum of the features and is applied channel-wise. Edge features are defined as $e_{ij} = h_\Theta(x_i, x_j)$, where $h_\Theta : \mathbb{R}^F \times \mathbb{R}^F \to \mathbb{R}^{F'}$ is a function with learnable parameters Θ, usually implemented as multi-layer perceptrons. While the choice of h_Θ is arbitrary, e.g. the PointNet [Qi et al., 2017a] uses $h_\Theta(x_i, x_j) = h_\Theta(x_i)$, which only considers global features, the function definition employed in this thesis is

$$h_\Theta(x_i, x_j) = h_\Theta(x_i, x_j - x_i), \tag{2.4}$$

which combines both, global features (x_i) and local information ($x_j - x_i$). In particular, edge features e_{ij} can be computed as

$$e_{ij} = \mathrm{ReLU}(\theta \cdot x_i + \phi \cdot (x_j - x_i)) \tag{2.5}$$

with trainable parameters $\Theta = (\theta, \phi)$ and a Rectified Linear Unit (ReLU) as activation function.

2.2.3 Graphical Models and Message Passing

Chapter 5 investigates decoupled feature learning and optimisation for the task of medical image registration. In most cases, the optimisation process is considered as inference on a discrete graphical model. Therefore, a brief summary of the underlying basic principles, namely Markov random fields and belief propagation, is given in this section.

Markov Random Fields Markov random fields are undirected graphs, where each node represents a random variable and the edges model stochastic dependencies [Bishop et al., 2006]. In the presented methods for medical image registration, the graph is usually constructed as k-NN graph from distinctive keypoints extracted from volumetric medical scans. The goal is then to assign discrete labels $l_i \in L$ (quantised 3D displacement vectors for volumetric registration) to each node x_i. A corresponding MRF energy function E_{MRF} can thus be formulated (cf. Equation 2.2) as

$$E_{\mathrm{MRF}}(L) = \sum_i D(l_i) + \sum_i \sum_{j:(i,j)\in\mathcal{E}} R(l_i, l_j), \tag{2.6}$$

where D and R represent unary (data term) and pairwise (regularisation term) potentials, respectively (cf. [Glocker et al., 2008]). In the context of image registration the data term models the correspondence quality, while the regularisation term ensures a smooth deformation field. Specific implementation details are then described in the respective method chapters.

Belief Propagation Belief propagation is a message passing algorithm for inference on graphical models (including MRFs). In this overview as well as the presented methods, the min-sum algorithm is considered. A message m_{ij} from a node x_i to a node x_j is defined as

$$m_{ij}(l_j) = \min_{l_i} \left(D(l_i) + R(l_i, l_j) + \sum_{k:(k,i)\in\mathcal{E}\setminus j} m_{ki}(l_i) \right). \tag{2.7}$$

The marginal distribution can be determined as the sum of all incoming messages plus the respective data term. As evident from the equation, the algorithm is only exact on graphs that have no loops. One possibility is therefore to convert the undirected graphs into trees, e.g. using Prim's algorithm for minimum spanning trees. Alternatively, belief propagation can also be applied iteratively (loopy belief propagation), which leads to an approximate solution whose convergence, however, is not guaranteed. For this, all messages are initialised with ones and the computation of the messages is performed in parallel, as there is no specific order of the nodes. [Felzenszwalb et al., 2006] summarise further ideas for efficient belief propagation in vision tasks.

Chapter 3

Pose Graphs for Clinical Context Monitoring

This first methodological chapter addresses the task of human pose estimation from point clouds in the clinical context of an operating room. It is investigated in which way the irregular input data can be processed as dense representations, more precisely projections (depth images) and voxelisations, using deep convolutional neural networks. Another focus of the approach presented here is the regularisation of the implicit output pose graph, which is composed of the individual joint detections. In the first part (Section 3.1) of this chapter it is demonstrated that a state-of-the-art CNN developed for 2D human pose estimation on natural rgb images is also a suitable model for projected point cloud images. It is further shown that frequently occurring predictions of implausible poses (e.g. missing or symmetrically swapped joints) can be reduced by constraining the output pose space by training a CAE with a low dimensional embedding and employing it as post-processing step. The method is published in [Hansen et al., 2019b]. In Section 3.2 the extension of the approach to 3D and a multi-camera setting is described (published in [Hansen et al., 2019e]). For this, 2D joint predictions from each camera view are projected back in 3D space and processed as a fused voxelised 3D point cloud.

3.1 Regularised Landmark Detection with CAEs

3.1.1 Introduction

Human pose estimation is a typical computer vision task that has been studied for decades and is a key component for a variety of higher level applications ranging from motion control in video games or car entertainment systems to video surveillance and behavioural understanding. The conventional estimation process is driven by the underlying image data, capturing local appearance, as well as structured prediction to produce globally plausible poses. Similar to other fields, purely data driven deep learning methods, in particular convolutional neural networks (CNNs) yield impressive results on public datasets and have nowadays replaced hand-crafted features and graphical models. With their success research shifts to more challenging scenarios. Pose estimation in clinical environments offers great opportunities by providing contextual information about the patient or staff for assistance and monitoring systems, but also has to cope with strong occluded and cluttered settings such as the operating room. At the same time clinical pose estimation datasets are much smaller than large-scale annotated datasets like MPII [Andriluka et al., 2014] making it harder to reliably train deep learning models. Explicit regularisation of predicted poses is a possibility to recover plausible poses.

3.1.1.1 Related Work

Early successful methods for human pose estimation relied on hand-crafted image features and sophisticated body part models [Felzenszwalb et al., 2008]. “DeepPose” [Toshev et al., 2014] was the first work that trained a deep neural network to directly regress joint positions. Following the success of CNNs for the image classification task subsequent methods used CNNs in a fully convolutional manner to generate heatmaps of joint locations [Wei et al., 2016]. In general, predicting heatmaps showed to be more robust than directly regressing pixel positions. Recent state-of-the-art pose estimators often adopt variants of stacked hourglass networks (SHGs) [Newell et al., 2016] as basic building blocks. SHGs capture both local and global context within multi-scale CNN architectures. Stacking multiple of these “hourglasses” combined with intermediate supervision further improves the network’s final performance.

The most recent methods described above model the human body only implicitly. In this work we aim at explicitly regularising poses and therefore consider research that uses global priors in neural networks. Convolutional Autoencoders (CAEs) [Masci et al., 2011] were introduced as unsupervised feature extractors. They reduce the input data with spatial filters and pooling operations into a latent space from which the original input must be reconstructed. The learned features from the encoder can then be reused in a supervised classification task. [Oktay et al., 2017] showed that the CAEs latent space could also be used to regularise segmentation of ultrasound images. Therefore during training, the model predictions are forced to follow the distribution of the learnt low dimensional representations of priors. In the field of human pose estimation [Tekin et al., 2016] trained an autoencoder on 3D joint locations to regress the 3D positions from 2D images via its regularised latent space.

3.1.1.2 Contributions

In this work we investigate how SHG, a state-of-the art human pose estimator, performs in a challenging clinical setting. We show that SHG models trained on normalised depth images yield a similar performance on the MVOR dataset [Srivastav et al., 2018] as RGB based methods and can therefore conclude that the use of depth images is a suitable image modality to tackle privacy concerns in visual health care applications. Given the small amount of training data (in contrast to non-clinical settings), we observe frequent anatomically implausible pose predictions. We therefore propose a CAE based post-processing of predicted poses. Our experiments show that this new approach can reliable recover poses from input perturbations such as joints that are missing or symmetrically switched.

3.1.2 Methods

Our architecture for regularised pose estimation consists of two independent components – an SHG that predicts heatmaps from raw input data and a CAE to recover plausible poses from the initial joint locations. See Figure 3.1 for a schematic overview of the two-stage pipeline.

3.1.2.1 Pose Estimation

Pose estimation of input image data, e.g. RGB or depth, is conducted with SHGs from [Newell et al., 2016]. The output of the network is a heatmap with N channels, where N is the number of landmarks to detect. The final predictions are given by the maximum activation of the heatmap. Ground truth targets consist of a 2D Gaussian (kernel) centred on the corresponding joint location. One "hourglass" consists of multiple convolution and pooling layers to process features down to a very low resolution, whereby at each pooling step one further convolution is applied at the current level and added to the corresponding layer on the symmetric bottom-up sequence. Multiple "hourglasses" with intermediate supervision can be stacked to further improve the performance.

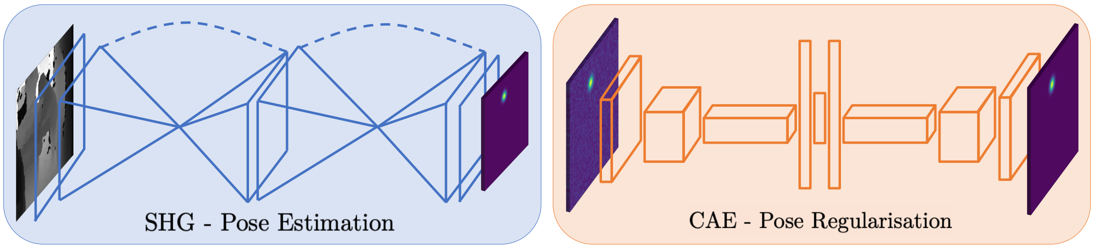


Fig. 3.1: We propose a two-stage architecture for regularised pose estimation from single depth images. A stacked hourglass network infers heatmaps for human joint locations. In a subsequent step a convolutional autoencoder corrects implausible pose predictions. The two networks are trained separately, whereby the pose regulariser is explicitly forced to learn a body model via perturbations of the input data.

3.1.2.2 Pose Regularisation

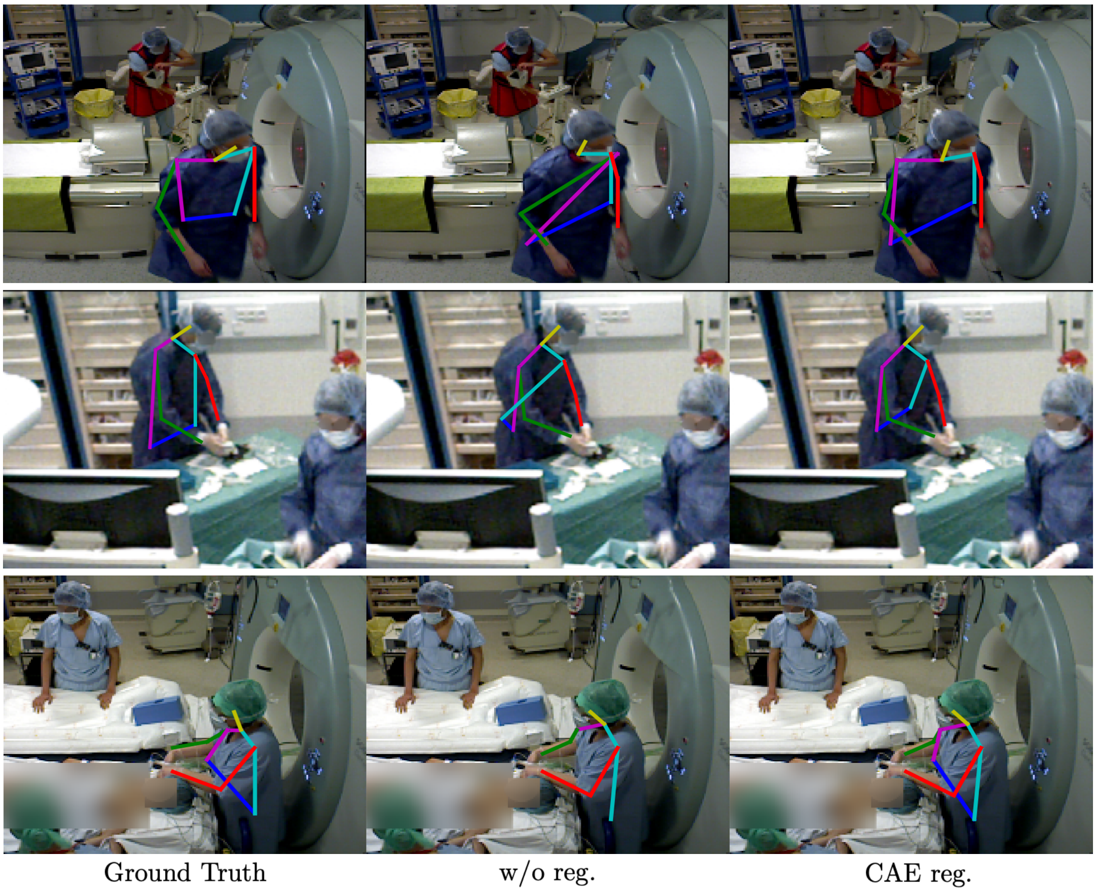
For regularisation of implausible poses we propose a convolutional autoencoder that is trained with different synthetic perturbations on the input data to enforce the explicit learning of a body pose model. The CAE is composed of two parts, the encoder and the decoder. In the encoding step a heatmap with N channels is transformed into a small-low-dimensional feature map (e.g. 4x4 pixels) by a series of convolution and pooling layers. With every convolution the number of features is doubled. After the last pooling layer the spatial relation is broken and the flattened tensor is fed into a two-layer neural network. The number of output features of the last linear layer specifies the dimension of the latent representation and should be small enough to force the CAE to learn a representative body pose embedding. The decoder reverses the operations of the encoding, whereby convolutions and pooling operations are replaced by transposed convolutions and indexed unpooling, respectively. To ensure that the CAE can actually act as a pose regulariser typical pose errors are explicitly incorporated in the training. In this work we propose three different synthetic augmentations: swapping two corresponding joints of the left and right body part, removing a single joint and applying a high random offset to one of the joints. All input channels are randomly scaled and Gaussian noise is added to further reduce the risk of learning an identity mapping.

3.1.3 Experiments and Results

All experiments were conducted on the MVOR dataset [Srivastav et al., 2018]. It consists of 732 multi-view frames from three RGB-D cameras providing registered depth and anonymised rgb images synchronized in time. The images were recorded in an operating room at the University Hospital of Strasbourg over a period of four days. Each day was used as one fold in a 4-fold cross validation. For the task of 2D pose estimation the ground truth annotations consist of 2926 bounding boxes and upper-body poses. As evaluation metric we use the percentage of correct keypoints with a threshold of 0.2 (PCK), whereby a prediction is assumed as correct if it falls within $0.2 \cdot \max(bbox_h, bbox_w)$ pixels of the ground truth annotation.

3.1.3.1 Pose Estimation

We first evaluate the pose estimation network on both image modalities, depth and RGB. The SHG with two stacks is implemented in PyTorch. Training was performed for 50 epochs with the Adam optimizer and an initial learning rate of $2.5e{-}4$. The objective function is the mean squared error between the predicted and ground truth heatmaps. We use random affine transformations to augment the input data. For training with depth images it was crucial to normalise each image independently to correct for different distances of the person to the camera.



	head	shoulder	elbow	wrist	hip	mean
OpenPose$_{\text{rgb}}$	91.0	88.8	74.5	58.1	56.4	73.8
AlphaPose$_{\text{rgb}}$	87.7	88.9	77.8	64.7	61.8	76.2
SHG$_{\text{rgb}}$	97.0	**90.5**	77.4	79.5	**71.7**	83.2
SHG$_{\text{depth}}$	**97.6**	**90.5**	**78.2**	**80.0**	71.2	**83.5**

	w/o reg.	CAE reg.
swap	80.3	**97.1**
remove	90.0	**96.6**
offset	90.7	**96.5**

Fig. 3.2: Qualitative and quantitative results on the MVOR dataset. The example images show ground truth (left), SHG predictions (middle) and CAE regularised joints (right). Body poses were inferred from depth frames and overlayed on rgb images for better visualization. Independent quantitative evaluations (using the PCK metric) were conducted for both the pose estimation and pose regularisation step and are shown in the left and right table, respectively.

Results Fig. 3.2 shows the results of the SHG models. Training with the two image modalities gives a similar PCK of approximately 83.5. For comparison, PCK values of two further pose estimation methods, OpenPose [Cao et al., 2017] and AlphaPose [Fang et al., 2017], are reported from [Srivastav et al., 2018]. We note that both estimators are pretrained on MPII without fine-tuning on MVOR.

3.1.3.2 Pose Regularisation

The general training setting is the same as for our pose estimation network. The CAE encodes body poses in a 16-dimensional latent vector. During training Gaussian noise with a std of 0.05 is added to the input heatmap and one of the three perturbations (swap, remove, offset) is applied with a probability of 0.1 each. We evaluate the capability of the CAE to recover plausible poses for each input perturbation independently by applying the corresponding augmentation strategy on all test images. In a last experiment we regularise the predicted heatmaps from the SHG by feeding them directly into the CAE.

Results The isolated evaluation of our CAE shows improved mean PCK values of 80.3 to 97.1, 90.0 to 96.6 and 90.7 to 96.5 for the swap, remove and offset perturbation, respectively. The subsequent regularisation of the SHG predictions on depth images leads to a mean PCK of 84.4. Qualitative results in Fig. 3.2 visualise the regularisation of implausible and incomplete poses.

3.1.4 Discussion and Conclusion

We successfully validated a state-of-the-art pose estimator, namely SHG, for upper body pose estimation in a new challenging clinical environment. Thereby, we could show that depth frames as input modality reached a better performance than blurred RGB images, implying that depth information can be a natural choice for computer vision algorithms in health care applications that depend on anonymised input data. Our CAE based regularisation can reliable recover plausible poses from a set of input perturbations and as a simple and independent post processing step the CAE leads to a small improvement in mean PCK. To further improve on the results the CAE could be incorporated in an end-to-end training in the pose estimation step, e.g. using the latent space for a regularisation loss as in [Tekin et al., 2016]. Finally, we believe our self-supervised regularisation method has great potential for future use in landmark detection and foresee further research in different domains, e.g. localisation in CT or MRI volumes.

3.2 2D Boosted 3D Pose Graph Estimation

3.2.1 Introduction

Robust human pose estimation is a key component for many of today's and future computer vision tasks. Areas of usage range from surveillance systems to entertainment solutions to autonomous driving. Likewise, precise knowledge of 3D joint positions of clinical personnel is a valuable information for context aware assistance and monitoring systems in clinical settings such as operating rooms. Knowing joint locations of all actors during an operation over time enables many different applications. As an important base feature in action recognition tasks [Yao et al., 2012], the human body pose is essential for the automatic analysis and the improvement of surgical workflows [Padoy et al., 2012]. It may further help to enable a safe human-robot interaction during future collaborative robotic surgeries [Jacob et al., 2013]. Gesture based control systems form a third application area, where human body poses play a major role: [Yusoff et al., 2013] use arm gestures to navigate MRI or CT images during surgery, whereas [Dietz et al., 2016] propose to control the operating light by simple hand gestures.

Real clinical environments present special challenges for the task of human pose estimation. An operating room is a visually complex scene with difficult lighting conditions, frequent clutter and occlusions from different medical devices. Besides the patient, many clinical staff are visible in the scene during an operation, often working simultaneously directly at or near the operating table. Hospital clothing that must be worn during an operation is usually wider and less formfitting than everyday clothes. Moreover, we identified two reasons for possible privacy concerns in using RGB videos in the operating room: 1. While the images used for the computation of body poses may be transient in a final application, in a first step and for clinical studies/trials the video data must be permanently stored and usually manually annotated to train and evaluate today's machine learning methods. 2. Security breaches in hospital networks have become more common [McCoy et al., 2018]. In such worst-case scenarios, attackers may gain access to video material of patients in a most vulnerable situation. This makes it preferable to work with unidentifiable visual data, such as blurred RGB or depth images [Silas et al., 2015]. Finally, clinical datasets for 3D joint localization are much fewer and smaller than for the general setting.

Given the constraints listed above, depth images may serve as input modality for a clinical 3D human pose estimation framework. Depth frames offer informative visual data, while at the same time providing a decent degree of privacy for patients and staff. In contrast to RGB or near infra-red (NIR) images, a clear identification of people is hardly possible [Silas et al., 2015]. Additionally, depth cameras are less sensitive to changing lighting conditions. A network of multiple depth cameras is a straight forward solution to reduce the problem of occlusions by medical devices or other people and can be easily mounted and calibrated in a closed setting such as the operating room. To overcome the limited databases in clinical environments, large scale annotated datasets for joint localization such as MPII [Andriluka et al., 2014] or Human3.6m [Ionescu et al., 2013] can be used to support a learning system wherever possible.

3.2.1.1 Related Work

2D Pose from RGB As a key problem for computer vision tasks, human pose estimation from single RGB images has been studied for more than 15 years. Early methods relied on hand-crafted features and explicit body part models [Andriluka et al., 2009; Felzenszwalb et al., 2008; Mori et al., 2004]. However, as many other topics in computer vision, human pose estimation gained more attention with the breakthrough of deep convolutional neural networks (DCNNs). Consequently, “DeepPose” [Toshev et al., 2014] was the first work that proposed to train a DCNN to directly regress joint positions from an input image. Recent state-of-the-art bottom-up and top-down methods often adopt the approach of [Newell et al., 2016] and generate heatmaps of joint locations in a fully convolutional manner [Cao et al., 2017; Chen et al., 2018b; Newell et al., 2017; Xiao et al., 2018].

3D Pose from RGB Much research has concentrated on inferring 3D prediction from RGB images, which is an ill-posed problem in the presence of perspective distortion. [Pavlakos et al., 2017] adopted stacked hourglass networks from [Newell et al., 2016] for this task by regressing voxel likelihoods for each joint instead of two dimensional heatmaps. Another family of works directly regresses spatial coordinates of the human joints with respect to a known root joint. A simple baseline is proposed from [Xiao et al., 2018], where 2D joints, predicted from the input image, are directly lifted in the 3D space by a shallow neural network. [Katircioglu et al., 2018] pretrain an autoencoder on 3D skeleton data to learn a latent representation for human body poses. Afterwards, a DCNN is trained to directly regress from the input image to the latent space. They argue, that the latent representation constraints the network to output plausible poses.

3D Pose from Depth Data-driven methods, often based on random forests, predict 3D joints directly from single depth images, either by segmenting body parts [Shotton et al., 2011] or directly regressing the x, y, z coordinates [Girshick et al., 2011; Jung et al., 2016]. A DCNN approach is proposed by [Haque et al., 2016]. The network learns viewpoint invariant features enabling it to make predictions even from extreme views, e.g. a top-view. Alternatively, given the intrinsic parameters of the camera, a depth image can also be represented as a three dimensional point cloud. The model from [Moon et al., 2018] processes the voxelized point cloud with an adopted hourglass network [Newell et al., 2016] for 3D input and outputs per voxel likelihoods for each body joint. To compensate for the large memory footprint of processing volumes instead of images they adjusted the feature channels in each networks component.

Pose Estimation in Clinical Settings Finally, research on human pose estimation in clinical environments is briefly summarized. In-bed pose estimation is approached in several publications [Achilles et al., 2016; Chen et al., 2018a; Liu et al., 2019b]. Other methods deal with pose estimation of clinical staff in the operating room [Belagiannis et al., 2016; Kadkhodamohammadi et al., 2017, 2021].. [Belagiannis et al., 2016] propose a DCNN processing of RGB images from up to 5 views combined with 3D pictorial structures. Also on RGB input relies the approach of [Kadkhodamohammadi et al., 2021], where 2D joint detections from each view are fused to a 3D prediction via epipolar geometry. In our preliminary work [Hansen et al., 2019b], we demonstrated that depth images can reach the same accuracy for 2D pose estimation in clinical settings as RGB images, but did not infer 3D positions from the detected 2D joints.

3.2.1.2 Contributions

In this work, we aim to bring robust 3D human pose estimation to the clinical domain. Considering the challenges and conditions of a clinical environment, such as the operating room, we investigate a 3D joint estimation framework based on a network of multiple depth cameras. We consider a two-step architecture. 2D heatmaps are generated from depth images of each camera. These probabilities of joint positions are fused in a combined point cloud, which is then voxelised to a fixed grid. An autoencoder, pretrained to encode a latent presentation of human poses, processes the fused input to generate final 3D voxel likelihoods for each joint. Closest to our approach is the work of [Moon et al., 2018], but it uses only a single depth image and does not exploit 2D information. Our main contributions can be summarized as follows:

- To the best of our knowledge, our approach is the first that fuses 2D information from multiple depth sources in a single volume to predict voxel probabilities for 3D landmark localisation.
- Moreover, we show that encoding the 3D human pose in the latent space of the fully convolutional autoencoder boosts the performance of our pose estimation model.
- We validate our approach on the challenging MVOR dataset Srivastav et al., 2018 against several baseline methods and in different ablation experiments, demonstrating the advantages of its individual components

3.2.2 Methods

In this section, we present our approach for 3D human pose estimation from multiple depth cameras. Figure 3.3 illustrates the overall idea. In this work, we deal with the problem of single person pose estimation, meaning multiple people in an image are already annotated by individual bounding boxes and we process each cropped person instance by itself. Therefore, our method can easily be incorporated in a multi person

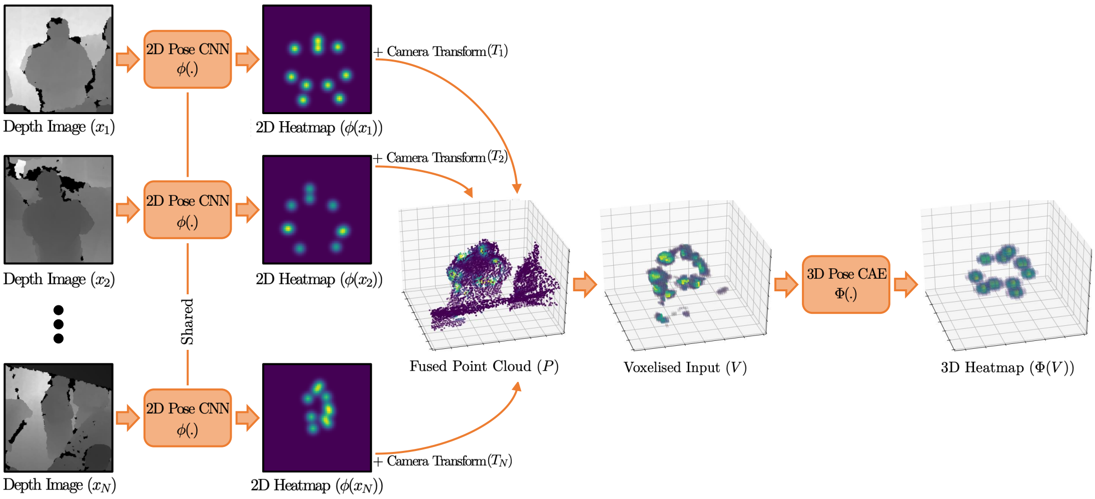


Fig. 3.3: The schematic of 2D information fusion for 3D human pose estimation. We predict 2d joint probabilities on each individual depth map. The heatmap information can be fused in a point cloud, which is voxelised and eventually processed by the the 3D pose convolutional autoencoder to obtain 3D voxel probabilities for each joint. For visualisation purposes, all joints are depicted in a single feature channel.

pose estimation framework in a top-down manner using a generic bounding box detector [Liu et al., 2016; Ren et al., 2015] beforehand.

Input to our method are cropped depth images $x_1, \ldots, x_N$ from 1 to N cameras. Additionally, intrinsic camera parameters and the transformation matrices $T_1, \ldots, T_N$, that transform 3D camera coordinates to 3D world coordinates, are given. The first step in our two-stage architecture is to predict probabilities of 2D joint positions from each depth frame. Here, we use the approach from [Xiao et al., 2018], explained in more detail in Section 3.2.2.1. As we have given the depth information as well as the intrinsics and extrinsics of each camera, we can generate a combined point cloud P from all views with the obtained joint heatmaps as additional features (see Section 3.2.2.2). However, for several reasons the feature point cloud only provides a rough estimate of the desired 3D joint positions. Firstly, as depth cameras only provide surface distances the detected probabilities of joint positions lie on the person's body surface, missing an additional (oriented) offset to the real 3D position. Another problem is the difference in spatial proximity in the depth image and the point cloud, respectively. While the predicted and ground truth joint may be very close to each other in the 2D image plane, they may be off by meters in the 3D point cloud. This can happen if the 2D detection is not accurately localized in the depth image and thus correspond to far off depth values. The same applies for partly and completely occluded joints. Lastly, in our setting probabilities of joints from up to N views are taken into account, that must be combined in a single prediction. Therefore, to solve the issues mentioned above, we make use of a second neural network, the proposed 3D pose convolutional autoencoder, that is further described in Section 3.2.2.3. Input to the network is the voxelized pointcloud V. The autoencoder is pretrained in a self-supervised fashion using ground truth 3D joint annotations from the public Human3.6m dataset [Ionescu et al., 2013] to encode the space of plausible human poses and outputs voxel likelihoods for each joint. Final pose coordinates are obtained by finding the maximum probability for each joint, whereby subvoxel accuracy can be obtained when considering neighbouring voxels for the final prediction. Therefore, the joint position is offset by a quarter of a voxel in the direction of the next highest neighbor (considering a fixed number of six neighbour voxels).

3.2.2.1 2D Pose CNN

The 2D pose CNN, denoted as ϕ in Figure 3.3, is adopted from Xiao et al., 2018. It uses a ResNet [He et al., 2016] as backbone network for image feature extraction. To generate heatmaps from the deep and low resolution feature space the authors employ several transposed convolutional layers. The regression target for the pose CNN is a multi-channel heatmap consisting of 2D Gaussians (with a given standard deviation) centered on each respective joint location. For implementation details and further insights, we refer to the original publication. Here, the CNN is pretrained with color images on the MPII human pose dataset [Andriluka et al., 2014] and then finetuned to adapt to the depth modality. The network is trained view-agnostic. Thus, we use the same model for 2D heatmap regression from each depth map x_i. We refer to an obtained heatmap as $\phi(x_i) = (\phi(x_i)_1, \ldots, \phi(x_i)_J)$, where J denotes the number of predicted joints.

3.2.2.2 2D-3D Information Fusion

In the next step, we aim to fuse the obtained 2D joint information from multiple views. Using the intrinsics of each camera N point clouds are generated from the depth images. With the given transformation matrices $T_1, \ldots, T_N$ each point cloud can be represented in the same reference system. Each point in the combined point cloud P consists of three coordinates and the corresponding J probability features. For further processing and voxelisation, the center of mass of the person is estimated from the point cloud coordinates that are equally weighted by all probability features. Note that in [Moon et al., 2018] an additional network is trained to predict the center, whereas in our case the rough 3D joint locations are sufficient to directly infer a decent estimation. The target person is then cropped from the point cloud by a cuboid filter with fixed edge length positioned at the estimated center of mass. For voxelisation, the 3D space is discretised based on a pre-defined voxel size and for each joint the probabilities of all points inside one voxel are averaged to obtain the corresponding voxel value. Thus, the dimension of the voxelised input $V = (V_1, \dots, V_J)$ for the 3D pose convolutional autoencoder is determined by the voxel size and the edge length of the cuboid filter used for cropping the combined point cloud. The number of feature channels corresponds to the number of joints J.

3.2.2.3 3D Pose CAE

The architecture of our 3D pose convolutional autoencoder (CAE) is illustrated in Figure 3.4. It is composed of three simple operations: non-strided 3D convolutions, 3D convolutions with stride 2 and 3D transposed convolutions with stride 2. The encoder part of the network starts with a non-strided convolution with a large kernel of size $7 \times 7 \times 7$. Afterwards, the obtained feature map is alternating processed by non-strided convolutions and convolutions with stride 2, that halves the feature maps resolution. At the same time the number of feature channels are increased with 128 features at the lowest resolution. Thus, the input is compressed by approximately $30 * J$ of its original size. To reconstruct the 3D heatmaps from the latent space three transposed convolutions with stride 2 and a kernel size of $4 \times 4 \times 4$ are employed before the final $1 \times 1 \times 1$ convolution. Rectified linear units (ReLUs) are used after every convolutional operation except the last one as non-linearities. The goal of the network is to predict voxel likelihoods for each joint, given the volumetric input V described in

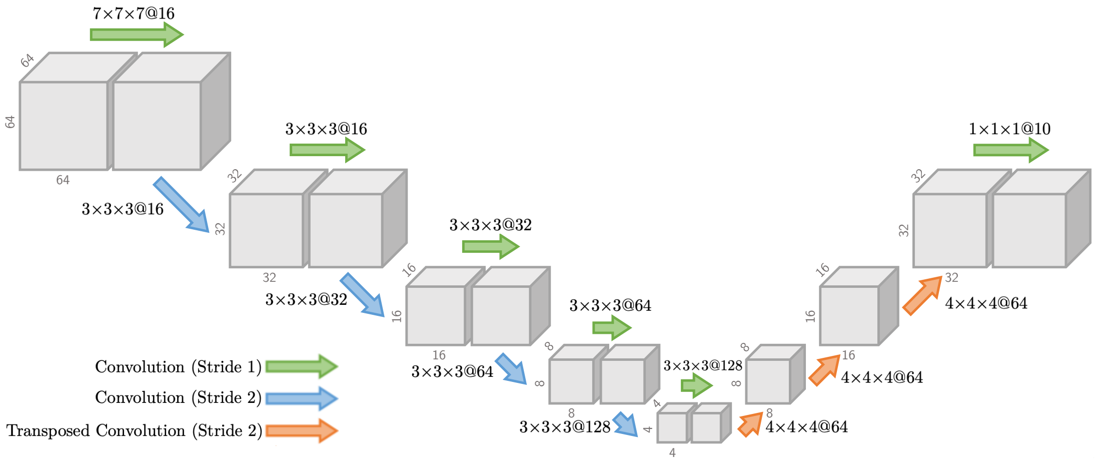


Fig. 3.4: A block diagram of the proposed 3D pose convolutional autoencoder. Convolutional operations are represented by the colored arrows and corresponding labels, specifying kernel sizes and output feature channels.

the previous section. The prediction target $V^* = (V_1^*, \dots, V_J^*)$ is generated from the ground truth 3D joint positions as follows:

$$V_j^*(u, v, w) = exp\left(-\frac{(u-u_j)^2 + (v-v_j)^2 + (w-w_j)^2}{2\sigma^2}\right), \tag{3.1}$$

where the j-th joints 3D ground truth position is (u_j, v_j, w_j), quantised in voxel coordinates, and σ denotes the standard deviation of the Gaussian kernel.

We aim to encode a space of plausible human poses in the CAEs latent space to regularise our predictions. Therefore, the network is pretrained in a self-supervised fashion. Input and target heatmap are generated from 3D ground truth poses from the Human3.6m dataset [Ionescu et al., 2013] as described above. Following the idea of denoising autoencoders [Vincent et al., 2010] to obtain *good* representations, all input channels of the input volume are randomly scaled and Gaussian noise is added. To even further reduce the risk of simply learning an identity mapping, we introduce several sources of noise, explicitly forcing the autoencoder to gain a deeper understanding of human body poses. Joints are either randomly added, removed and swapped or a large offset is added. The CAE needs to correct this perturbations to reach a low reconstruction error. In our human pose estimation framework, the CAE is trained in three stages: First, the network is pretrained as just described using only pertubated ground truth heatmaps (60 training epochs). Next, the decoder part is frozen and we regress from the voxelised input point cloud V into the latent space (30 training epochs) and finally, the complete network, including the decoder, is finetuned for another 30 epochs. This training strategy is also employed in [Katircioglu et al., 2018].

3.2.3 Experiments and Results

All experiments were conducted on the challenging MVOR dataset [Srivastav et al., 2018]. It consists of 732 multi-view frames from three RGB-D cameras providing registered depth and RGB images synchronised in time. The images were recorded in an operating room at the University Hospital of Strasbourg over a period of four days. For robust evaluation, each day was used as one fold in a 4-fold cross validation setting (57/330/223/122 images). For the task of 3D pose estimation, the ground truth annotations consist of 1061 upper-body poses, whereby the person is visible in one, two and three views in 132, 426 and 503 cases, respectively. An upper-body pose is defined by a total of ten joints (upper head, neck, shoulders, elbows, wrists and hips). As evaluation metric we use the 3D mean per joint position (MPJP) error in centimeters.

We compare our approach to the V2V-PoseNet [Moon et al., 2018], a state-of-the-art method for 3D pose estimation from voxelised input from single depth images. Here, for a fair comparison, we also use the fused voxelisation from all available depth images. Different ablation studies investigate the effectiveness of individual aspects of our approach. In the simplest baseline experiment, we use our 3D architecture as a direct pose estimator with the combined voxelised point cloud as input. The CNN is not pretrained to encode a representative pose space and no 2D information is used. To show the advantage of 2D-3D information fusion the next baseline utilize the probability of 2D joint positions but is still trained from scratch. The last comparison is between the 3D CAE with encoded pose space trained with and without the introduced noise. In a further experiment, we investigate the usefulness of having multiple cameras for 3D human pose estimation in clinical settings, here, the operating room. Therefore, we report MPJP error of our best model with respect to the number of supporting views for each annotation.

3.2.3.1 Implementation Details

We implement our pose estimation framework in PyTorch [Paszke et al., 2019]. All models are trained for 120 epochs. Adam optimization is used with an initial learning rate of 0.001. To stabilise training the learning rate is decayed by a multiplicative factor of 0.1 after 60 and 90 epochs. Additionally, we employ batch normalisation throughout all layers with a mini batch size of 16. The edge lengths of the cuboid filter are chosen as 200 cm each, which accommodates for the largest expected person. The combined point cloud is augmented with a random scale and a rotation around the z-axis. For the voxelisation of the point cloud, we use a pre-defined voxel size of approximately $3\,\mathrm{cm}^3$. The 3D heatmap targets are generated with a σ value of 2. Parameters are chosen based on hyperparameter optimisation on our method and kept fixed for all baseline experiments. For 2D pose estimation [Xiao et al., 2018] and the reference method V2V-PoseNet [Moon et al., 2018], we utilise the publicly available implementations[1].

3.2.3.2 Results

Quantitative results of our approach in comparison with the reference method and the baseline experiments are depicted in Table 3.1. A p-value of 0.0435 indicates statistical significance of the difference in joint position errors between the V2V-PoseNet and our proposed approach (calculated with the Wilcoxon signed rank test). The different baseline experiments show a constant increase in mean MPJP error, starting from 12.1 cm for the 3D network without pose encoding and without probabilities of 2D joint positions as input. Table 3.2 shows that increasing the number of supporting views, decreases the mean MPJP error (from 11.9 cm for one view to 7.1 cm for three views). With a mean MPJP error of 8.0 cm the localisation accuracy of the challenging wrist joints is more than halved (from 17.0 cm for one view). Qualitative results are illustrated in Figure 3.5. For better visualisation, the RGB images are shown instead

[1] [Xiao et al., 2018]: https://github.com/Microsoft/human-pose-estimation.pytorch
[Moon et al., 2018]: https://github.com/dragonbook/V2V-PoseNet-pytorch

Table 3.1: Comparison of 3D MPJP error (mean±std) in cm on the MVOR dataset for the reference method and several baselines. Head joints are omitted because they already yield extreme low errors and thus, are not suitable for comparison. The input modality is stated in brackets: 'img' for depth only and 'hm' for 2D joint heatmaps (generated from [Xiao et al., 2018]).

	shoulder	elbow	wrist	hip	mean
V2V-PoseNet (img)	**4.9 ± 5.3**	8.2 ± 7.0	12.6 ± 10.0	10.0 ± 7.5	8.9 ± 6.1
ours w/o pose enc. (img)	6.9 ± 7.7	12.1 ± 9.9	17.5 ± 11.6	11.9 ± 8.6	12.1 ± 7.6
ours w/o pose enc. (hm)	5.3 ± 6.0	7.9 ± 7.1	12.8 ± 9.4	10.3 ± 8.6	9.1 ± 6.3
ours w/o noise (hm)	5.5 ± 7.0	7.9 ± 7.0	11.3 ± 9.9	10.0 ± 9.0	8.6 ± 6.4
ours (hm)	5.2 ± 6.5	**7.6 ± 7.2**	**11.1 ± 10.1**	**9.5 ± 7.4**	**8.3 ± 6.3**

Table 3.2: 3D MPJP error (mean±std) in cm on the MVOR dataset with respect to the number of supporting views.

	shoulder	elbow	wrist	hip	mean
one view	6.0 ± 3.5	13.0 ± 6.2	17.0 ± 9.3	11.4 ± 7.5	11.9 ± 4.3
two views	5.6 ± 5.1	8.1 ± 6.3	13.5 ± 11.4	9.7 ± 5.0	9.2 ± 5.2
three views	**4.7 ± 7.9**	**6.4 ± 7.6**	**8.0 ± 7.9**	**9.2 ± 8.9**	**7.1 ± 7.2**

of the input depth frames. A visual inspection yields accurate and overall consistent predictions. Lastly, we report the runtime of our pipeline for a detection based on three views. We utilised a GTX 1070 and measured a time of 51 ms (3 × 7 ms for 2D joint detection, 10 ms for voxelisation and 20 ms for 3D joint detection).

3.2.4 Discussion and Conclusion

We proposed a 2D-3D information fusion approach for human pose estimation from multiple depth cameras. On the MVOR dataset, we demonstrated that our approach is able to robustly predict 3D human poses even in challenging scenarios, where body parts of nearby people or medical devices protrude in a bounding box and e.g. cause occlusions. We performed slightly better than a state-of-the-art pose estimator on the same dataset showing that fusion of 2D information and additional pose encoding is a feasible approach for 3D landmark localization. The desired accuracy in the joint position error clearly depends on the clinical application, however, reported average MPJP errors of $< 5\,\text{cm}$ for the RGB to 3D joints task of the 2018 Pose Track Challenge [Andriluka et al., 2018] indicate room for further improvement. While the introduced noise for explicitly learning the pose space gave slightly improved results, we believe

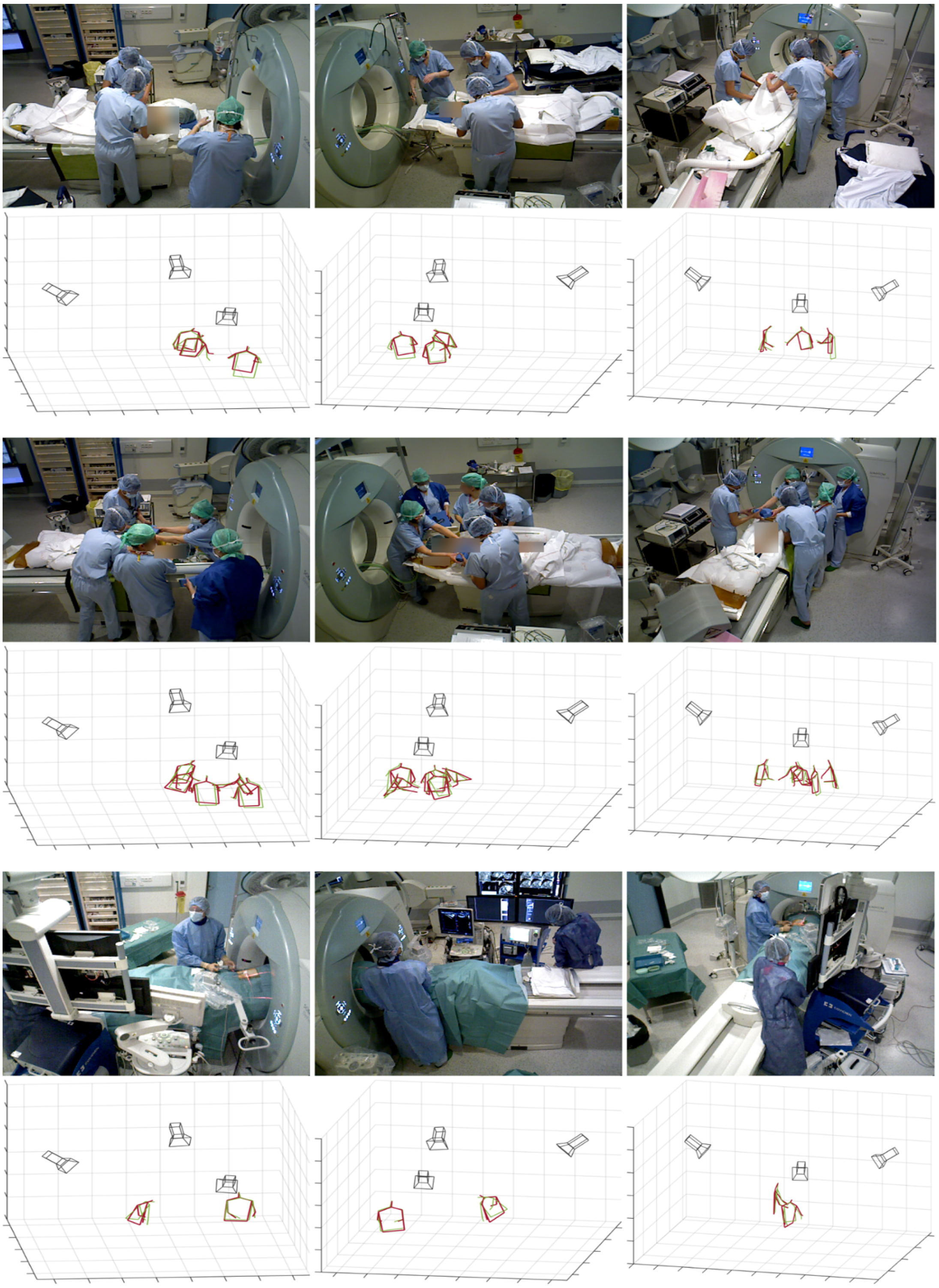


Fig. 3.5: Qualitative results of our approach on samples from the MVOR Dataset. We show the three RGB images and corresponding views of the 3D poses. Green poses depict the ground truth, whereas the red poses are generated from our method.

that the approach has further potential, e.g. by incorporating graphical models. The analysis of the localisation error with respect to the number of supporting views clearly shows that our method is capable of utilizing different views to increase accuracy. Finding an optimal arrangement for the cameras in the operating room may be a direction for future research. With a total running time of 51 ms on a mid-range GPU the framework is also suitable for real time applications. To this point, our method consists of two separated steps: The 2D pose estimation on the depth images and the 3D pose estimation from the voxelised input point cloud. Incorporating both steps in a single end-to-end framework may lead to superior performance and is clearly of high interest for further research.

Chapter 4

Semantic Segmentation on Anatomical Graphs

The following chapter describes the methodological transition from dense processing of the input data to graph-based approaches, exemplified by the medical imaging task of semantic segmentation. In the first part (Section 4.1), a method for deep learning on irregular domains published in [Hansen et al., 2019a] is presented, which is based on a combination of learnable 1x1 convolutions and feature propagation using multiple diffusion kernels. As an area of application, point clouds of people are again considered, but in contrast to the previous chapter, for pointwise segmentation instead of landmark detection. Section 4.2 then develops an approach to apply the idea of sparse processing to segmentation of dense medical CT and X-Ray scans. For this, image patches are sampled at sparse, descriptive key points and processed with a graph neural network before the spatial predictions are finally accumulated in a dense output array. This method is published in [Hansen et al., 2019d].

4.1 Multi-Kernel Diffusion CNNs for Graph-Based Learning

4.1.1 Introduction

The vast majority of image acquisition and analysis has so far focused on reconstructing and processing dense images or volumetric data. This is mainly motivated by the simplicity of representing data points and their spatial relationships on regular grids and storing or visualising them using arrays. In particular convolutional operators for feature extraction and pooling have seen increased importance for denoising, segmentation, registration and detection due to the rise of deep learning techniques. Learning spatial filter coefficients through backpropagation is well understood and computationally efficient due to highly optimised matrix multiplication routines for both CPUs and GPUs.

However, many alternative imaging devices such as time-of-flight based 3D scanners or ultrasound that is based on reflectance measurements are not necessarily optimally represented on dense 3D grids. Instead these sparse measurements can be stored and processed more naturally and effectively using point clouds that are connected by edges forming an irregular graph. Moreover, 3D data from multiple sources can be easily combined if represented as point clouds.

The supervised feature learning and further analyses on these irregular domains is a research area that is still in its early stage, in particular in the context of deep learning. The main limitations of previous approaches are their dependency on an equal number of nodes in all graphs (e.g. derived from point clouds) and the same topology, i.e. ordering of nodes and edge connections. Furthermore, some operations on irregular graphs are inefficient for parallel hardware, which limits their usefulness in real world scenarios.

4.1.1.1 Related Work

Of all hierarchical feature learning models, convolutional neural networks have shown to be one of the most successful approaches for a wide variety of tasks [Ren et al., 2015]. Attempts to transfer the concepts from the two dimensional image domain directly to a sparsely sampled 3D space include e.g. volumetric CNNs [Maturana et al., 2015] and multi-view CNNs [Su et al., 2015]. However, due to the sparseness of the observed space both techniques lack computational efficiency.

Another class of works addresses this problem more generally by studying the intrinsic structure of data on non-Euclidean and irregular domains. Noteworthy are in particular spectral descriptors that are based on the eigenfunctions and eigenvalues of the Laplace-Beltrami operator. The proposed methods include heat kernel signatures (HKS) [Sun et al., 2009], wave kernel signatures (WKS) [Aubry et al., 2011] and learnable optimal spectral descriptors (OSD) [Litman et al., 2014]. Spectral CNNs, defined on graphs, were first introduced in [Bruna et al., 2014]. The main drawback of this method is that it relies on prior knowledge of the graph structure to define a local neighbourhood for weight sharing. Consequently, the idea of graph convolutions has been extended in [Kipf et al., 2017; Mikael et al., 2015] by limiting the support size of the learned spectral filters, making them independent of graph topology. In [Boscaini et al., 2016; Masci et al., 2015; Monti et al., 2017] another approach is presented, which defines a new form of local intrinsic patches on point clouds and general graphs, where the weights parametrising the construction of patches are learned. Graph attention networks Velickovic et al., 2018 learn a functional mapping to define pairwise weights based on the concatenated features of the involved nodes. The localised spectral CNN (LSCNN) [Boscaini et al., 2015], which derives local patches from the windowed Fourier transform, can be seen as a combination of the spectral and the spatial method. [Bronstein et al., 2017] provides a comprehensive review of current research on this topic.

Deep learning applied directly on unordered point sets is considered in the PointNet framework [Qi et al., 2017a; Qi et al., 2017b]. The input point set is recursively partitioned into smaller subsets and max pooling is used as a symmetric function to aggregate information regardless of point ordering.

Closest to our approach is the work of [Atwood et al., 2016], that uses a power series of the transition matrix on a graph as diffusion operation to capture local node behaviour, while we additionally employ multiple diffusion constants to build the filter kernels based on different variants of the normalised Laplacian. Moreover, we found it critically to build our network in a multi-layer fashion which was not considered in Atwood et al., 2016.

4.1.1.2 Contributions

In this work, we propose a simplified architecture that helps to overcome the limitations stated above, i.e. it can be employed for both grid and irregular graphs, has a comparable or better computational performance than classic CNNs and is theoretically connected to research on mean field inference approaches for graphical models in computer vision. As detailed in Section 4.1.2, we propose multi-kernel diffusion convolutional neural networks (mkdCNNs) based on two simple building blocks: isotropic, rotationally-invariant graph diffusion operators that propagate information across edges (on the graph) and trainable 1×1 convolutions that manipulate features for each node individually. When employing multiple diffusion constants for the information propagation, which are linearly combined with the following 1×1 convolution, powerful regional features, e.g. curvature, can be learned. A random walk approach is considered to further simplify the diffusion process. In Section 4.1.3 we successfully validate the proposed multi-kernel diffusion convolutional network on the tasks of learning pointwise correspondences between point clouds of different human poses as well as segmenting body parts.

4.1.2 Methods

Input to our network is a matrix $P \in \mathbb{R}^{n \times f}$, where the i-th row corresponds to one of n points $p_i \in \mathbb{R}^f$ of a point cloud in an f-dimensional feature space.

4.1.2.1 Network Architecture

Figure 4.1 visualises our proposed mkdCNN composed of the building blocks described below in detail. The layer input is a feature map defined on a graph. The weighted edges of the graph determine the feature propagation between nodes implemented as diffusion operation. The feature learning step consists of 1×1 convolutions followed by non-linear activations. Therefore, its support is limited to each individual node. Stacked mkdCNN layers can be used in networks for global classification, with a final symmetric pooling function (e.g. max or average pooling), or for semantic node-wise segmentation in a fully convolutional manner.

4.1.2.2 Input Feature Graph

The simplest way to capture and represent local geometry in a point cloud is via a k-nearest neighbour graph G_k, where $\mathcal{N}_k(p_i)$ denotes the set of the k-nearest neighbours of a point p_i and edge weights are defined by a distance metric $dist_{ij}$ between two points p_i and p_j. An adjacency matrix A for the graph is constructed with entries

$$a_{ij} = \begin{cases} \exp(\frac{-dist_{ij}^2)}{2\cdot\sigma^2}), & \text{if } p_j \in \mathcal{N}_k(p_i) \\ 0, & \text{otherwise} \end{cases},$$

where σ denotes a scalar diffusion coefficient. In our work we employ multiple diffusion constants yielding different weighting schemes for the same graph. Spectral graph analysis Chung et al., 1997 allows us to extract further geometric properties from the point cloud, e.g. an intrinsic order of points, via the symmetric normalised graph Laplacian $L_{\text{sym}} = I - D^{-1/2}AD^{-1/2}$. I denotes the identity matrix. The degree Matrix D is solely defined by its diagonal elements $d_{ii} = \sum_j a_{ij}$. For large point clouds it may be necessary to approximate the highly sparse matrix L_{sym} to maintain the computational efficiency of deep networks on GPUs. For this purpose, we can perform an eigendecomposition using only the first $m \ll n$ eigenvalues, such that

$$L_{\text{sym}} = Q\Lambda Q^{\intercal},$$

where the diagonal matrix Λ holds the m eigenvalues and Q the corresponding eigenvectors. An alternative to the symmetric Laplacian is the random walk normalised Laplacian $L_{\text{rw}} = I - D^{-1}A$.

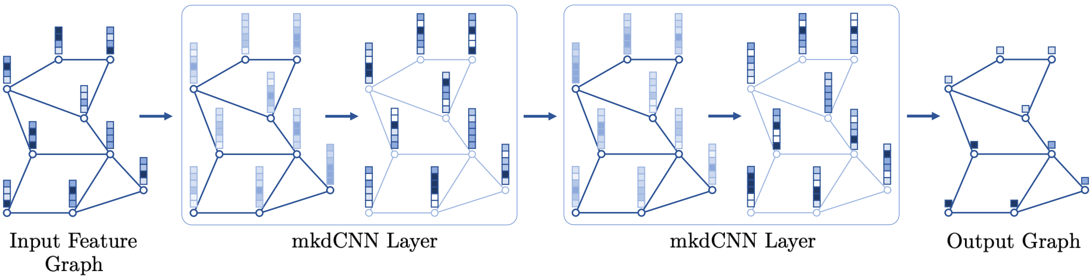


Fig. 4.1: Example of a two layer multi-kernel Diffusion CNN for node classification: Given an arbitrary input graph with f-dimensional features (left), we employ alternating layers of topology-independent diffusion operators with multiple isotropic kernels that propagate information across the graph, followed by 1×1 convolutions and activations that act on nodes individually and learn abstract representations of features (middle). In the end class predictions for each node a determined by a final 1×1 convolution (right).

Input point features can be arbitrarily defined depending on the application and additional given information. For graphs derived from or based on regular grids like 2D images and 3D volumes such features may be simple grayscale values/patches or more suitable approaches, e.g. extraction of BRIEF descriptors [Calonder et al., 2010]. Real world coordinates and surface normals can be extracted from 3D point clouds from stereo vision or time-of-flight systems. Once a graph is defined, the spectrum of the Laplacian itself can be used for feature extraction, e.g. B-spline based geometry vectors [Litman et al., 2014]. Furthermore, the construction of the mkdCNN makes it possible to learn meaningful information with no input features at all. In this case point features are simply initialized with ones.

4.1.2.3 mkdCNN Layer

Each of our proposed mkdCNN layers consists of two separated steps: the diffusion operation and the feature learning.

To propagate features across the graph the Laplacian is used, thus making the propagation step for features independent of employed graph topologies and applicable to graph datasets with varying numbers of nodes. Essential to our mkdCNN layer is the use of multiple isotropic diffusion kernels as visualized in Figure 4.2. Together with the following node-wise feature learning, expressive regional features with different local support can be extracted from the non-linear combination of all kernels. The diffused point cloud values P' can be computed as the solution of the diffusion process

$$P' = (\lambda L_{\text{sym}} + I)^{-1} P,$$

where λ denotes the diffusion time Desbrun et al., 1999. Approximating L_{sym} with few eigenvectors as mentioned above yields an efficient computation, as

$$P' = Q(\lambda \Lambda + I)^{-1} Q^{\intercal} P.$$

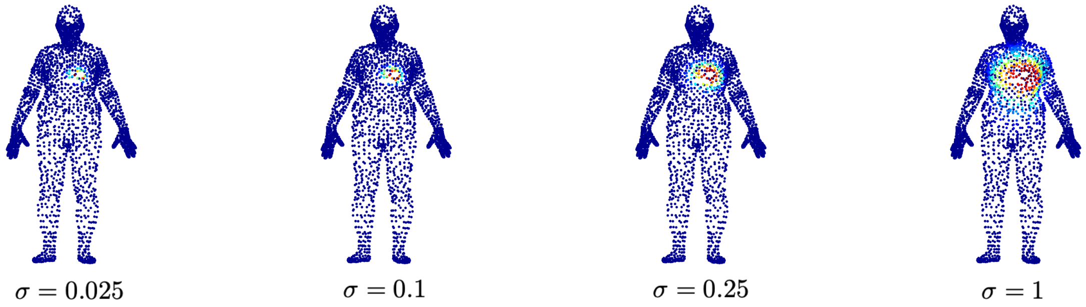


Fig. 4.2: Visualisation of multiple isotropic diffusion kernels (for one point on the chest of a subject) employed in a single feature propagation step of our mkdCNN Layer.

Therefore, diffusion is mainly affected by the parameters k, σ and λ, that give control over the locality of the feature propagation. As our network can be trained in an end-to-end manner those parameters can either be learned or determined on a holdout validation set. As an alternative diffusion operation, that does not involve the costly matrix inversion, we also considered a random walker, such that

$$P' = (I - L_{\mathrm{rw}})^t P.$$

In this case the diffusion parameters are k, σ and the number of diffusion steps t. Parallels to conditional random fields (CRFs) can be drawn. Our diffusion operation corresponds to one message passing step with the difference that the approximate mean and variance of features are propagated instead of an exact inference of all variables as in CRFs. In [Krähenbühl et al., 2011] a similar approach for efficient and approximate inference on grid-graphs is proposed that involves convolving a downsampled set of message variables with truncated Gaussian kernels.

In our proposed network, features are solely learned through 1×1 convolutions followed by a non-linearity. Besides adding depth to the network this choice is based on the analogy of our design with CRFs, where a label compatibility function is learned to penalize the assignment of different labels to nodes with similar properties [Krähenbühl et al., 2011]. Note that in CRFs the dimensionality of signals residing on each node is limited to the number of output labels and thus the compatibility function is restricted to only learn interactions across few classes, whereas in our approach the compatibility is established between feature maps. Furthermore, the exclusive use of 1×1 convolutions would make it conceptually easy to incorporate well studied building blocks from recent deep learning literature such as dense or residual connections into our network. Instance normalisation and dropout are used to stabilise training and we employ a block of two 1×1 kernels each.

4.1.3 Experiments and Results

Our new method is evaluated in two experiments: point descriptor learning and semantic body parts segmentation. We make use of the publicly available FAUST dataset [Bogo et al., 2014], which consists of 100 surface meshes of 10 different subjects, each scanned in 10 different poses. The 3D meshes have a resolution of 6890 vertices and point-wise correspondences between the shapes have been semi-automatically established for all points. As we are only interested in the scanned point clouds, we do not consider the given triangulations in our experiments. Following [Boscaini et al., 2015] we split the dataset in a disjoint training (subjects 1-7, 70 shapes), validation (subject 8, 10 shapes) and test set (subjects 9-10, 20 shapes).

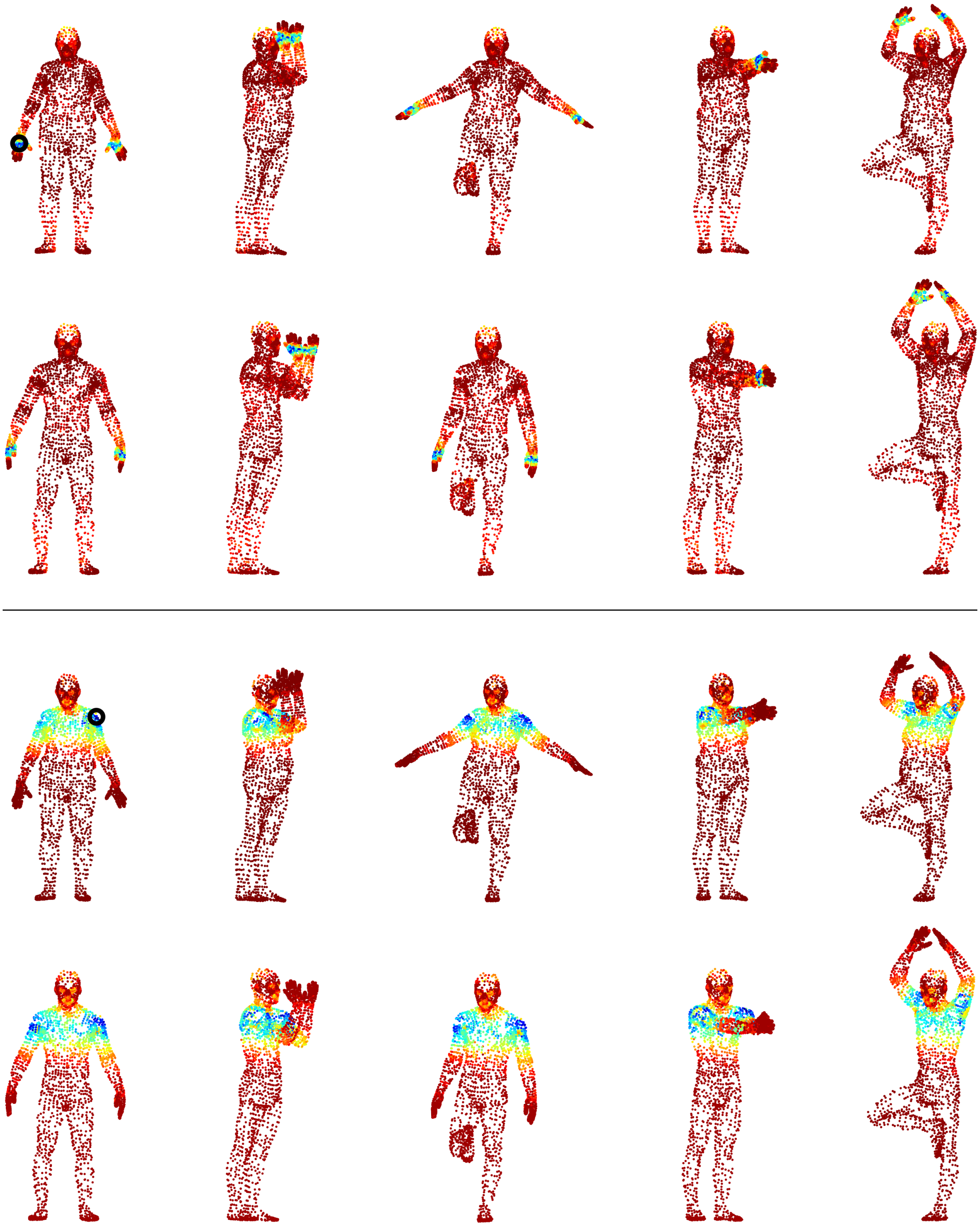


Fig. 4.3: Visualisation of distances in the descriptor space between all points on a selection of shapes from the FAUST test set and a single point ○ on a reference shape (upper left). Cold colors correspond to small distances. Distances are saturated at the median.

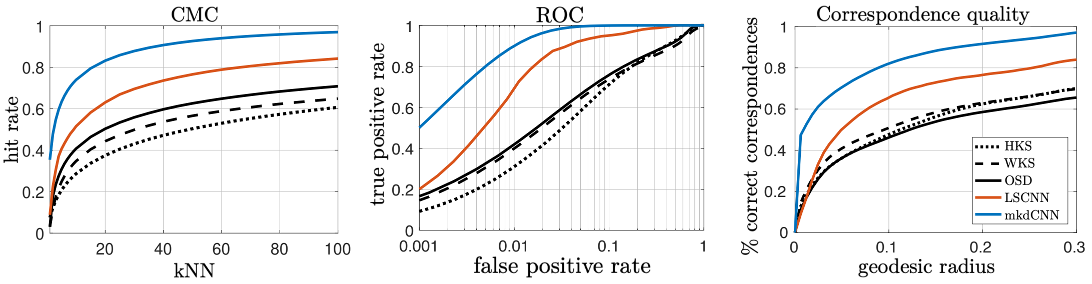


Fig. 4.4: Comparison of descriptor performances on the FAUST test set. For four comparison methods ■ 🟧 and our approach 🟦 we report the cumulative match characteristic (CMC), receiver operating characteristic (ROC) and the correspondence quality, which measures the distance between matched and ground truth points on the underlying mesh of the point cloud.

4.1.3.1 Point Descriptor Learning

The graph Laplacian for all point clouds was computed using $k = 100$ nearest neighbours. We employed a four layer mkdCNN using the random walk diffusion operation with parameters $\sigma = \{0.0125, 0.025, 0.05, 0.1, 0.125, 0.25, 0.5, 1\}$, $t = 7$ and no features on the input graph. All parameters were chosen according to automatic hyperparameter optimisation on the validation set. To train the descriptors we used a triplet hinge loss function - i.e. given a point on a randomly sampled shape its normalized Euclidean distance in the descriptor space to a non-corresponding point (on another random sampled shape) should be larger by a margin (here empirically set to 0.2) than its distance to a corresponding point (on another randomly sampled shape). The descriptor dimension was set to 16. Training was performed for 50 epochs with the Adam optimiser and an initial learning rate of 10^{-4}. For each optimisation step we considered 6890 triplets. We implemented our architecture in PyTorch [Paszke et al., 2019] and train a model (0.15 million free parameters) on a Nvidia GTX 1070 8GB in around five hours. At test time the extraction of all 6890 descriptors for one shape takes approximately 5 seconds. This time is dominated by the computation of the diffusion operation. Given precomputed diffusion operators our system is able to produce a throughput of 100k points per second.

We compare our mkdCNN to four other spectral descriptor approaches, namely HKS [Sun et al., 2009], WKS [Aubry et al., 2011], OSD [Litman et al., 2014] and LSCNN [Boscaini et al., 2015]. Publicly available implementations of the approaches were used and parameters (e.g. k for the computation of the graph Laplacian) optimised on the validation set.

Results Figure 4.4 shows different evaluation results for all approaches on the FAUST test set. First, the cumulative match characteristic (CMC) is shown. It evaluates the retrieval performance by testing if the correct corresponding point on one shape can be found inside the next k-nearest neighbours from the set of all points of another shape. The k-nearest neighbours are determined by Euclidean distances in the descriptor space and the mean over all points and over all shapes is reported. The hit rate @kNN = 10 improved from 0.29 (HKS), 0.35 (WKS), 0.41 (OSD) and 0.52 (LSCNN) to 0.73 (mkdCNN). The receiver operating characteristic (ROC) plots the true positive rate against the false positive rate of point pairs at several distance thresholds in the descriptor space. For a better distinction between the approaches, we plot the ROC curve in semilogarithmic scale. The measurements for the correspondence quality follow [Kim et al., 2011]. The ground truth meshes are used to compute the geodesic distances between all points on a shape and the percentage of point pair matches that are at most r-geodesically apart from their corresponding ground truth points are reported. For the mkdCNN this means that over 80% of point matches have a geodesic distance to their ground truth points of 10 cm or less.

Figure 4.3 visualises qualitative results of descriptors learned with the mkdCNN. A point is selected on a reference shape (on the right hand and left shoulder, respectively) and its distance in the descriptor space to all other points on the same and other shapes of the test set is computed. The distances are color-coded, where cold colors correspond to small distances. For most of the shapes distinct peaks around the ground truth are observable.

4.1.3.2 Semantic Body Parts Segmentation

The FAUST dataset does not include point-wise semantic labels for human body parts. Therefore, we labelled the points manually on a reference shape and transferred the labels to all other shapes via the known point correspondences. An exemplary ground truth labelling can be seen in Figure 4.7 (left). The 15 labels correspond to the head, thorax, abdomen, left hand, left lower arm, left upper arm, left foot, left lower leg, left upper leg, right hand, right lower arm, right upper arm, right foot, right lower leg and right upper leg. The semantic segmentation on the FAUST dataset was also investigated in [Kleiman et al., 2018], but only up to intrinsic symmetry (e.g. no distinction between right and left foot).

The default mkdCNN configuration for the semantic body parts segmentation is the same as for the descriptor learning task: the graph Laplacian is computed with $k = 100$ nearest neighbors; we use the random walk diffusion operation with diffusion parameters $\sigma = \{0.0125, 0.025, 0.05, 0.1, 0.125, 0.25, 0.5, 1\}$ and $t = 7$; no initial features are used on the input graph. For the classification task another 1×1 convolution is employed after the fourth mkdCNN layer producing softmax scores. The model is trained with a cross-entropy loss weighted with the root of inverse label frequencies and the Adam optimizer (initial learning rate: 10^{-4}). Training is stopped after 50 epochs.

General Results Figure 4.5 depicts segmentation results for a selection of point clouds from the FAUST test set. The mkdCNN produces accurate and precise point cloud labels, even for challenging poses (fourth column: touching hands, fifth column: right foot touches left knee). A Dice overlap of 0.95 ± 0.04 (averaged over all labels and all shapes of the test set) confirms the good visual impression.

Ablation Study Results To understand the effect of different parameter and architectural choices in the mkdCNN, we perform several ablation experiments on the segmentation task. The default configuration for all experiments is the one described above. Using the exact diffusion process instead of the random walk approach yields a slightly improved Dice (0.96 ± 0.03 vs 0.95 ± 0.04) at the cost of a much higher inference time (approximately 20 s and 5.5 s, respectively) due to the costly matrix inversion. Further evaluation results are shown as box-plots of Dice coefficients in Figure 4.6 (top row). In our first ablation experiment we study the effect of different number of diffusion kernels, i.e. the number of employed weighting schemes for the diffusion operation, on the segmentation results. Increasing the number of different

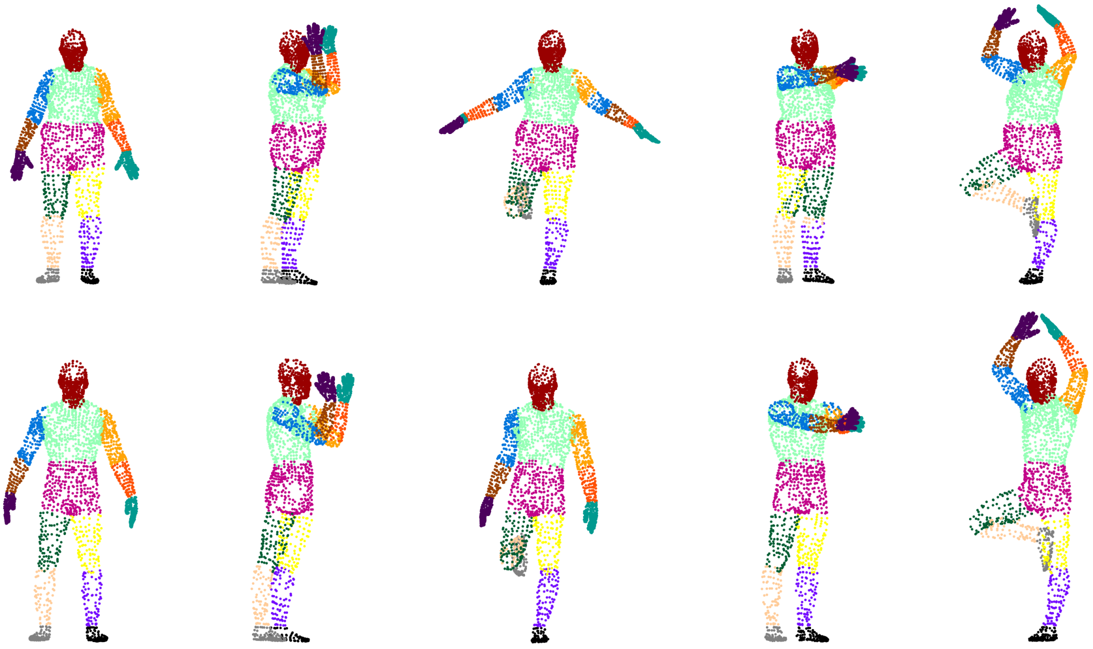


Fig. 4.5: Visualisation of segmentation results on a selection of shapes from the FAUST test set. While the segmentation is visually convincing for most 3D point clouds, small inconsistencies can be observed. On the third shape in the top row points of the right lower arm 🟫, right upper arm 🟦, left lower arm 🟥 and left upper arm 🟧 are not always assigned to the correct side of the body. The same applies for points of the right upper leg 🟩 and left upper leg 🟨 on the fourth shape in the top row.

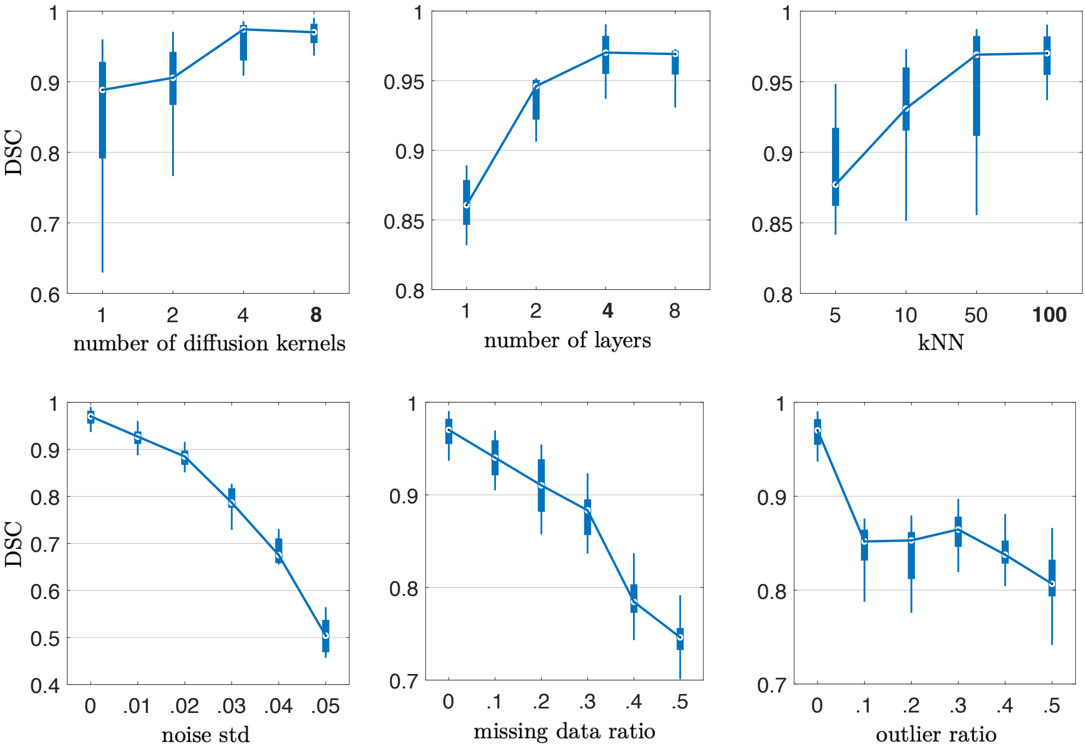
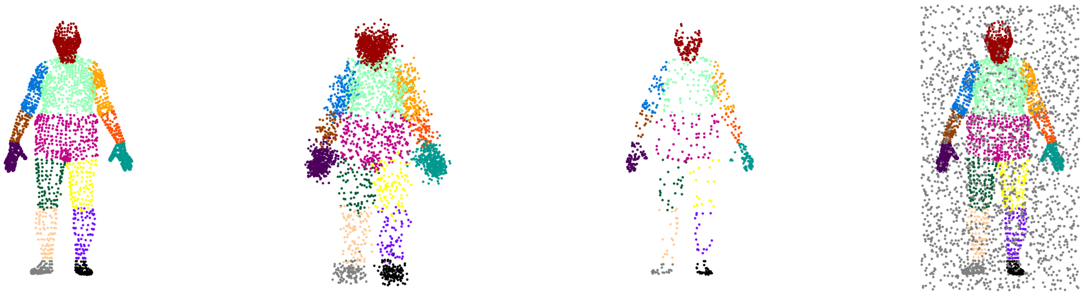


Fig. 4.6: Body parts segmentation results on the FAUST test set studying different parameter settings and data disturbances. Default configuration parameters are highlighted in bold.

σ values from one to eight (and thus also increasing the total number of trainable weights of subsequent 1×1 convolution layers from 40k to 150k) improves the mean Dice from 0.85 ± 0.11 to 0.95 ± 0.04. Particular interesting is the decreased standard deviation which implies a gain in robustness with respect to the variability between shapes. For the second experiment the number of mkdCNN Layers was set to 1, 2, 4 and 8, respectively. With an increased number of layers the size of the feature maps was reduced in order to keep the number of free parameters approximately the same for each configuration. Thus, the performance gain is introduced through a deeper mkdCNN architecture and not attributed to an increased capacity of the network. A difference in Dice overlap is especially recognizable between a one-layer and a two-layer mkdCNN. Another parameter that is not directly connected to the mkdCNN architecture but has a notable effect on the segmentation outcome is the number of nearest neighbors k for the creation of the graph Laplacian from a given point cloud. For the mkdCNN it seems to be an advantage to be build on top of a graph with many locally highly interconnected nodes. The mean Dice coefficient increases from 0.88 ± 0.04 ($k = 5$) to 0.95 ± 0.04 ($k = 100$).



ground truth noise missing data outlier

Fig. 4.7: Examples of our different data disturbance experiments. In this Figure Gaussian noise with a standard deviation of 0.03 is added to the ground truth points. The ratio for missing data points as well as added outliers is 0.3.

Robustness Rests Results A desirable property of a point cloud processing network is robustness against any disturbances of the input data. In a number of experiments we investigate the effect of different data disturbances on the segmentation results. Network parameters were not adapted for the robustness experiments. Figure 4.7 depicts the impact of the studied point perturbations on an exemplary ground truth point cloud. Robustness against noise is tested with random Gaussian noise added to the input cloud. We employ different standard deviations for the Gaussian (std $= \{0.01, 0.02, 0.03, 0.04, 0.05\}$). With a std of 0.02 the mean Dice coefficient is still approximately at 0.90. The results deteriorates with a std of 0.03 but Figure 4.7 shows that the noise is already at a very high level and unresolvable ambiguities exists in this synthetic ground truth. In the next experiment we remove points at random with a certain ratio. Even if every second point is removed the mkdCNN produces segmentations with a mean Dice of 0.75 without the need of adapting the network parameters. To investigate how the network can cope with outliers we randomly add points within the shapes bounding box. Added points are labelled as background. The results show that the segmentation task has become more difficult with the additional background class but is very robust against the ratio of outliers. Even for an outlier ratio of 0.5, i.e. half of all points belong to the background class, the Dice overlap is above 0.80. Figure 4.6 (bottom row) summarises the results of all data disturbance experiments.

4.1.4 Discussion and Conclusion

We have presented a new, simple architecture for descriptor learning and semantic segmentation on point clouds. By decoupling the graph propagation and feature learning step the mkdCNN overcomes the limitations of topology dependent approaches. Using the Laplacian and its approximation enables an efficient implementation of the diffusion of feature maps defined on sparse nodes that is transferable to different graphs and we showed that by providing multiple different kernels stronger features can be learned in each subsequent Layer. For the task of descriptor learning on point clouds from the FAUST dataset the mkdCNN (without any input features) shows better performance than a number of other spectral descriptors and learning approaches. Experiments on manually labeled body parts on the point clouds demonstrate the general feasibility of our approach for the task of semantic segmentation, even for highly noisy input (Gaussian noise, missing points, outliers). We validated several choices for our network architecture in ablation experiments and showed that a multi-layer mkdCNN with a high number of diffusion kernels build on top of a locally highly interconnected graph gives the best segmentation results in terms of Dice overlap. Visual inspection of the segmented point clouds expose rare failure cases due to ambiguities in the symmetry of the human body.

Overall the results for both tasks are very promising and demonstrate a substantial improvement over both hand-crafted spectral features and graph convolution approaches. Especially the proposed use of multiple kernels and the consistent employment of mkdCNN layers in a multi-layer fashion helped to decrease error rates for Dice coefficients in our investigated segmentation task by 66% when using eight instead of a single diffusion kernel and by 52% when increasing the depth from one to four layers. This is a significant improvement to the simple diffusion CNN in the work of Atwood et al., 2016, which is related to our mkdCNN in a configuration with only one kernel and a single layer. Despite using only topology-invariant and isotropic kernels, the learned non-linear combination in our proposed multi-kernel network help to create expressive and highly discriminative filters that enable accurate graph node classification. When visually inspecting the point descriptor similarity in Figure 4.3 it appears that the learned 16-dimensional feature vectors do not differentiate well between symmetric structures (e.g. left and right shoulder). However, the semantic labeling tasks demonstrated that the subtle global differences in the human pose are sufficient to correctly label and distinguish between the right and left half of the body. For some rare cases the evident errors are indeed the inconsistent assignment of points to the correct side of the body (see Figure 4.5 top row, third and fourth column).

Our mkdCNN framework provides some straightforward potential extensions for further improvements while maintaining its general design and inherent computational efficiency. Until now, we did not consider signals on our input point cloud but features like fast point feature histograms (FPFH) [Rusu et al., 2009], RGB values (acquired with real-world 3D scanners like the Kinect) or spectral features can potentially increase the networks performance. In this work we investigated the feasibility of the mkdCNN for learning on point clouds. As the diffusion operation is based on the graph Laplacian the network can be easily employed for general graphs. Testing our approach on graph datasets like Cora or PubMed Sen et al., 2008 may yield interesting new insights. An interesting research direction in general is to enable the possibility to not only learn features of a graph but also the connections (edge weights) between nodes and therefore incorporate mkdCNN into graph attention approaches [Velickovic et al., 2018].

4.2 Sparse Structured Prediction for Semantic Edge Detection

4.2.1 Introduction

The vast majority of medical image acquisition and analysis has so far focused on reconstructing and processing dense data. This is mainly motivated by the simplicity of representing data points and their spatial relationships on regular grids and storing or visualising them using arrays. In particular convolutional operators for feature extraction and pooling have seen increased importance for denoising, segmentation, registration and detection due to the rise of deep learning techniques. Learning spatial filter coefficients through backpropagation is well understood and computationally efficient due to highly optimized matrix multiplication routines for both CPUs and GPUs. However, for many computer vision tasks in medical image analysis such as landmark or edge detection it seems unnecessary and expensive (in terms of time and memory limitations) to process images end-to-end with dense methods, e.g. fully-convolutional networks or encoder-decoder architectures. Therefore, in this work, we aim to show new possibilities in the area of deep learning to process image data on sparse and irregular instead of dense grids. The feasibility of our suggested approach is demonstrated on the problem of semantic edge detection in CT and X-ray images.

4.2.1.1 Related Work

Of all hierarchical feature learning models, CNNs have shown to be one of the most successful approaches for a wide variety of tasks such as classification, bounding box regression and segmentation [He et al., 2017; Ronneberger et al., 2015]. Lately, another class of works (graph convolutional neural networks (GCNNs)) attempts to transfer these well-known concepts from the two dimensional image domain to non-Euclidean and irregular domains. Spectral CNNs, defined on graphs, were first introduced in [Bruna et al., 2014]. The main drawback of the proposed method is that it relies on prior knowledge of the graph structure to define a local neighbourhood for weight sharing. [Mikael et al., 2015] extended the ideas to graphs where no prior information on the structure is available. While [Bruna et al., 2014; Mikael et al., 2015] relied on splines for the formulation of their graph convolutional operators, [Kipf et al., 2017] uses truncated Chebyshev polynomials that allow for clear description of the support size of the learned spectral filters. [Bronstein et al., 2017] provides a comprehensive review of current research on this topic. In the medical domain GCNNs were successfully applied in a number of applications such as population-based disease prediction [Parisot et al., 2017], metric learning for brain connectivity graphs [Ktena et al., 2017] and survival analysis on pathological images [Li et al., 2018].

Edge detection is a key task in computer vision applications and is studied for decades [Canny, 1986]. [Dollár et al., 2013] chose a data-driven approach using random decision forests to predict structured labels from input image patches. This technique was successfully applied in the medical domain for multi-modal registration of ultrasound and CT/MRI images [Oktay et al., 2015]. [Xie et al., 2015] is the first deep learning method to explicitly learn edges. Features are extracted with a modified VGGNet and all layers are trained with deep supervision. In the end side outputs from different VGG Layers are fused to output a final edge map. State-of-the-art detectors for semantic edge detection mainly resemble encoder-decoder architectures that are trained with specialised loss terms [Liu et al., 2020; Yu et al., 2017].

4.2.1.2 Contributions

In this work we make a first step towards dense prediction from a few sparse sampling points using deep learning methods. We bring together the robustness of grid based CNNs and the flexibility of GCNNs in a single framework for pixel-level structured prediction. In this, our work differs from [Li et al., 2018], which used GCNNs for global context aggregation for image labelling. Our main contributions is the sparse structured prediction network (SSPNet). Furthermore, we successfully provide a first proof-of-concept for our new approach by evaluating it on the challenging task of semantic edge detection in medical images.

4.2.2 Methods

In this section, we present our proposed approach for sparse structured prediction for semantic edge detection. Figure 4.8 illustrates the general idea of our method. Input to our pipeline is an image x. A light-weight CNN ϕ, called sample CNN, extracts potentially informative locations from the image and outputs a single channel sample map $\phi(x)$. A fixed number N of sample coordinates $((x_1, y_1), \dots, (x_N, y_N))$ are drawn following a multinomial distribution with probabilities proportional to the sample map's values. Depending on the application many alternatives of extracting sampling locations are conceivable, e.g. for landmark detection one could initialize the sample map with the mean locations of the landmarks in the training set. At the given positions, patches $(p_1, \dots, p_N)$ are extracted from the input image x. Furthermore, a simple distance graph G_σ is generated. The adjacency matrix A of the graph G_σ is given by entries

$$a_{ij} = \exp\left(\frac{-d_{ij}^2}{2 \cdot \sigma^2}\right),$$

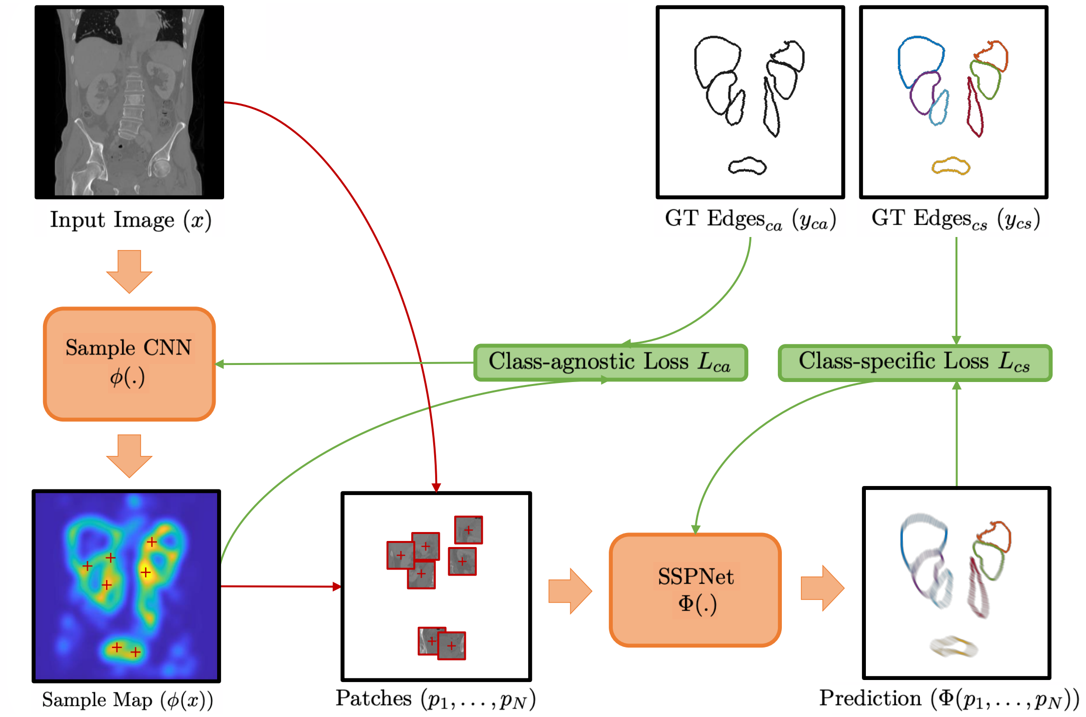

Input Image (x)
GT Edges$_{ca}$ (y_{ca})
GT Edges$_{cs}$ (y_{cs})
Sample CNN
$\phi(.)$
Class-agnostic Loss L_{ca}
Class-specific Loss L_{cs}
SSPNet
$\Phi(.)$
Sample Map ($\phi(x)$)
Patches ($p_1, \ldots, p_N$)
Prediction ($\Phi(p_1, \ldots, p_N)$)

Fig. 4.8: Our general idea for sparse semantic edge detection. We train a lightweight fully-convolutional CNN with a class-agnostic loss to output an informative heatmap from which samples are drawn with probabilities proportional to its values. Image patches are extracted around the chosen locations and our proposed SSPNet processes the generated patch graph to output a semantic edge activation for each sampling point. To recover a dense prediction all edges are accumulated in an array and the class-specific loss is applied to update the SSPNet's parameters.

where σ is a scalar diffusion coefficient and d_{ij} denotes the euclidean distance between two sampling locations (x_i, y_i) and (x_j, y_j). Again, depending on the application and given priors the graph may be initialized accordingly. Next, the extracted image patches $(p_1, \ldots, p_N)$ as well as the graph G_σ serve as input to our proposed SSPNet (explained in detail below), which predicts edges for each input image patch and accumulates all predictions on a dense grid weighted by their class-specific confidence. This semantic edge map is our final output. While the focus of this work is clearly on the SSPNet, in the following we also shortly describe the training of the sample CNN.

4.2.2.1 Sample CNN

The sample CNN ϕ is based on a lightweight version of the holistically-nested architecture in Xie et al., 2015. We significantly cut the networks capacity by removing deeper layers and use reduced numbers of filters. In total the network consists of only three

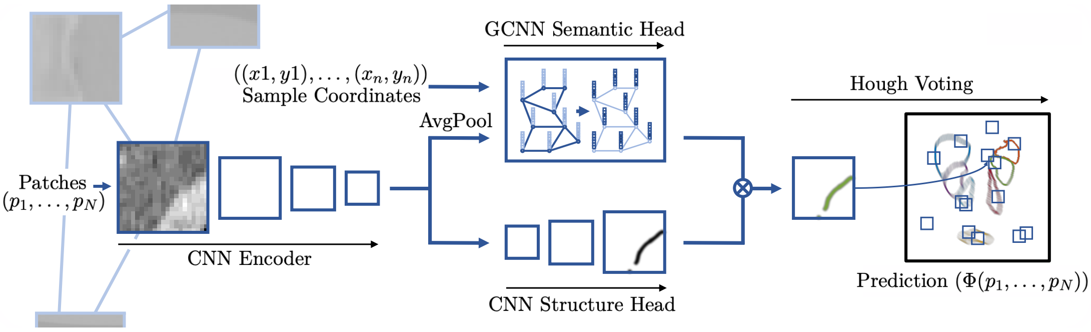


Fig. 4.9: The proposed sparse structured prediction net (SSPNet) expects a graph of patches sampled at informative image locations. For each patch a CNN encoder extracts a set of feature maps, which are further processed by 1) the structure head that predicts local edge activations and 2) the semantic head where global context is aggregated by a GCNN. We perform a weighted Hough voting to accumulate all predictions and recover a dense semantic edge map.

layers (each with two 3 × 3 convolutions + relu activation). Layer 2 and 3 start with convolutions with stride 2 resulting in the network's receptive field size of 23. After each layer a side output $\hat{y}_i$ is generated by a further 1 × 1 convolution and sigmoid activation. Side outputs are concatenated and fused to form a final prediction $\hat{y}_0$ by a 1 × 1 convolution and sigmoid activation. The sample CNN is trained with deep supervision on all outputs using the loss function from Deng et al., 2018 which combines the binary cross-entropy (BCE) and Dice loss ($DICE$) to

$$L_{ca} = \sum_{i=0}^{3} \alpha BCE(\hat{y}_0, y_{ca}) + \beta DICE(\hat{y}_0, y_{ca}).$$

In the loss term y_{ca} depicts the class-agnostic version of the ground truth edge map and α and β control the weighting of the two losses. Trading robustness for precise localization the final sample map is obtained from prediction y_0 after multiple average pooling steps with stride 1.

4.2.2.2 SSPNet

As stated above input for our SSPNet Φ are the extracted image patches $(p_1, \ldots, p_N)$ as well as the graph G_σ. The network itself consists of a CNN encoder part, a structure and semantic head and a final Hough voting step to recover a dense prediction from the single patches. An overview of our proposed SSPNet is given in Figure 4.9. The CNN encoder applies four convolutions (kernel sizes: 5, 3, 3, 3) with relu activations on each image patch. The resulting feature maps are further processed by two network heads. The structure head consists of three transposed convolutions (kernel sizes 3, 3, 3) and relu activations. The final structured prediction is obtained by a 1 × 1 convolution with sigmoid activation. The semantic head aggregates global context information and is modeled with a GCNN. Input features on our graph are the average pooled feature maps from the CNN encoder. We also experiment with explicitly adding the sampling coordinates as additional informative features. For this part of our work we decided to strive for simplicity and use a simple random walk diffusion with a single σ kernel to pool features across our input graph G_σ [Atwood et al., 2016; Hansen et al., 2019a]. The diffusion process can be described by the diffusion matrix

$$L = I - D^{-1}A,$$

where I denotes the identity matrix and the degree matrix D is solely defined by its diagonal elements $d_{ii} = \sum_j a_{ij}$. By matrix multiplication with the input feature vector a weighted average pooling across edges of the graph is employed. The diffusion pooling is followed by two 1 × 1 convolutions with relu activations. In total we employ two of the described graph convolutions. Final semantic confidence scores for each node (sampling locations) on the graph are obtained by a 1 × 1 convolution with sigmoid activation. As Hough voting has been proven to be effective for locating shapes in images [Ballard, 1981; Lindner et al., 2015] we accumulate the structured predictions from all image patches on a dense grid (with C channels, where C corresponds to the number of semantic classes) and weight each prediction with the corresponding semantic confidence score. Furthermore, each grid point is normalized by the number of predictions made for this point. Note that by construction the final semantic edge map holds values between 0 and 1 and we can apply our class-specific similar to our class-agnostic loss as

$$L_{cs} = \sum_{i=0}^{C-1} w_i(\alpha BCE(\hat{y}^{(i)}, y_{cs}^{(i)}) + \beta DICE(\hat{y}^{(i)}, y_{cs}^{(i)})),$$

where y_{cs} depicts the one-hot encoded semantic ground truth edges, such that pixels can belong to multiple labels. Classes may be weighted by the parameters w_i.

4.2.3 Experiments and Results

We validate the feasibility of our approach on two different datasets for the task of semantic edge detection. The first dataset consists of 10 2D coronal slices of abdominal CT scans from VISCERAL [Jimenez-del-Toro et al., 2016] and the second dataset is the JSRT dabase of 247 chest X-ray images [Shiraishi et al., 2000]. As validation metric we use the F-score on the fixed contour threshold (ODS), where the threshold is determined from all images in the test dataset. Before evaluation the thresholded predictions are thinned and spurious detections (< 10 pixels) are removed. ODS metrics are computed for each semantic class individually and we report the mean value. We compare our approach against three different 5-Layer UNet implementations (UNet-S, UNet-M, UNet-L). The UNet-S has a comparable capacity in terms of learnable parameters as our SSPNet, whereas the UNet-L has almost 2.5 as many parameters.

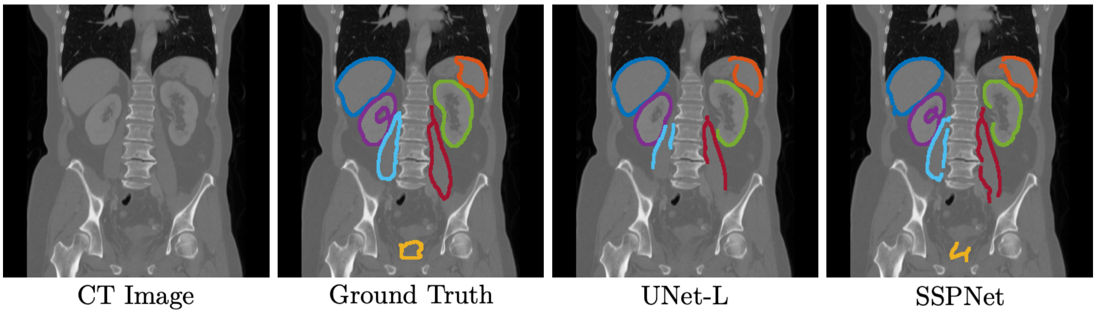
4.2.3.1 Implementation Details

All models were trained for 300 and 100 epochs for VISCERAL and JSRT, respectively. ADAM optimisation was used with an initial learning rate of .02. We employ batch normalization with a mini batch size of 4 and an exponential learning rate schedule with a multiplicative factor of 0.99 to stabilize training. The images are augmented with a random affine transformation. The graph for the SSPNet is computed with a σ value of 0.1 and normalized coordinates. We set the α and β parameters of the loss terms to .001 and 1, respectively. Class weights were applied corresponding to the organ label occurrences for all experiments with the UNet variants. During training and at test time we sample patches at 500 and 2000 locations, respectively. All hyperparameters were determined by grid search for our simplest baseline method and kept fixed for all further experiments.

4.2.3.2 VISCERAL

We perform initial experiments on 10 2D coronal slices of abdominal CT scans from VISCERAL in a leave-one-out fashion. The images are resampled to an isotropic pixelsize of 1.5mm^2 and cropped to dimensions of 320×312 without any guidance. We consider ground truth labels for seven anatomical structures: liver, spleen, bladder, left kidney, right kidney, left psoas major muscle (pmm) and right pmm. Besides our described architecture we test three other baselines of the approach: A SSPNet employing only 1×1 convolutions instead of the GCNN, the GCNN without sampling coordinates as additional input features and the network of 1×1 convolutions with sampling coordinates as additional features.

Results Qualitative and quantitative results are depicted in Figure 4.10. The GCNN outperforms the network with only 1×1 convolutions in both cases with and without sampling coordinates as additional features, although the result is much clearer in the second case (ODS of .690 against .786). The best SSPNet with an ODS of .827 yields a higher score than the UNet-S and Unet-M and performs only slightly worse than the UNet-L (ODS of .769, .791 and .834 respectively). Without class-weighting the UNet variants perform worse with ODS values of .763, .773 and .817, respectively. In contrast, the SSPNet showed similar results with and without class weighting. The visual comparison shows a clearer outline of the psosas muscels and a better detection of the unary bladder in favor of the SSPNet.



	Parameters	Samples	ODS
UNet-S	∼500k	dense	.769
UNet-M	∼900k	dense	.791
UNet-L	∼1300k	dense	.834
SSPNet – 1 × 1 conv.	∼500k	2000	.690
SSPNet – GCNN	∼500k	2000	.786
SSPNet – 1 × 1 conv. + coords.	∼500k	2000	.801
SSPNet – GCNN + coords.	∼500k	2000	.827

Fig. 4.10: Qualitative and quantitative results on VISCERAL. The images show edge overlays from seven anatomical structures: liver 🟦, spleen 🟧, bladder 🟨, left kidney 🟪, right kidney 🟩, left psoas major muscle (pmm) 🟦 and right pmm 🟥. Our approach outlines edges of the psoas muscles much clearer and also detects the urinary bladder.

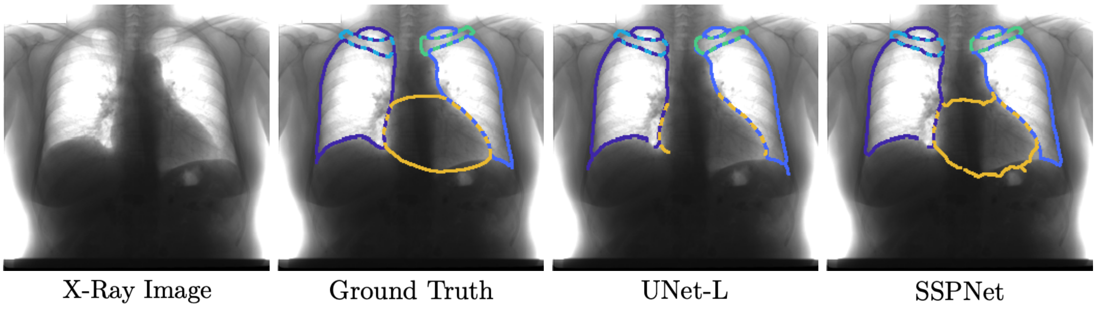
4.2.3.3 JSRT

The JSRT database consists of 247 chest X-ray images that were downsampled to dimensions of 256 × 256. A four-fold cross validation was employed to compute the results. We test the SSPNet with additional sampling coordinates as input features against the three UNet implementations UNet-S, UNet-M and UNet-L. Ground truth labels are generated from the provided landmarks for five anatomical structures: left lung, right lung, left clavicle, right clavicle and heart.

Results Qualitative and quantitative results are depicted in Figure 4.11. The SSPNet yields a slightly higher OSD score of .900 than the UNet-L with .884, though visual results are mostly comparable. However, in some cases the UNet misses parts of the edges of the heart whereas the SSPNet can follow informative gradients along its outline.

4.2.4 Discussion and Conclusion

In this work we proposed a new approach for structured prediction for semantic edge detection from a few sparse sampling locations on an image. To the best of our



	Parameters	Samples	ODS
UNet-S	~500k	dense	.874
UNet-M	~900k	dense	.878
UNet-L	~1300k	dense	.884
SSPNet − GCNN + coords.	~500k	2000	.900

Fig. 4.11: Qualitative and quantitative results on the JSRT chest X-ray database. The images (from left to right: original X-ray, ground truth, Unet-L, SSPNet) show edge overlays from five anatomical structures: left lung 🟦, right lung 🟦, left clavicle 🟦, right clavicle 🟩 and heart 🟨. The UNet misses parts of the edges of the heart whereas our approach successfully follows informative gradients along its outline.

knowledge the SSPNet is the first deep learning network that combines structured prediction with CNNs and global context aggregation with graph convolutions to recover a dense output. In our experiments on VISCERAL and JSRT we showed that the SSPNet performed better or on par with several UNet variants while having the lowest number of trainable parameters.

For future work, incorporating the SSPNet in an end-to-end learning framework instead of working with an explicitly trained sample CNN is clearly of high interest. This may be achieved by using a more complex GCNN model with attention mechanisms, e.g. [Monti et al., 2018], which could lead the selection of sampling locations. With an extension to 3D volumes, our approach can be evaluated on medical datasets with stronger memory and computational limitations. While in this work the focus was on edge detection, other tasks for structured prediction, such as landmark detection in medical images, may also be suited well for our approach.

In conclusion, we showed that our SSPNet is feasible for semantic edge detection in medical images and we believe that it can be used as a potential alternative to dense encoder-decoder architectures for general pixel-level image tasks in deep learning.

Chapter 5

Decoupled Feature Learning and Sparse Graphical Optimisation for Medical Image Registration

This chapter presents a total of three graph-based approaches for the task of medical image registration. The main methodological focus and difference to the previous chapters is the decoupling of deep learning based feature extraction and (sparse graphical) optimisation. The first contribution (Section 5.1), published in [Hansen et al., 2021c], deals with a sparse loopy belief propagation algorithm for keypoint based abdominal CT and CT-MR registration, that is extended by an efficient displacement candidates sampling scheme. The algorithm is further boosted by additionally processing semantic label maps, that were separately trained with a state-of-the-art UNet architecture. The following methods presented in Sections 5.2 and 5.3 are both combining learned geometric features from anatomical keypoint graphs extracted from lung CT scans with conventional point cloud registration algorithms, highlighting the feasibility and benefits of decoupled graph learning for medical image registration. Both approaches are separately published in [Hansen et al., 2019c] and [Hansen et al., 2021a], respectively.

5.1 Revisiting Iterative Highly Efficient Optimisation Schemes

5.1.1 Introduction

While the main focus of this work is on investigating the general problem of efficient medical image registration, we here specifically deal with the clinical task of inter- and intra-patient alignment of abdominal CT/MR scans. Enabling deformable multimodal fusion (of thorax and abdomen) has numerous medical applications, e.g. for aligning pre-interventional scans for image-guided (radio)therapy and multimodal diagnostic. Inter-subject CT registration can enable statistical modelling of variations of abdominal organs for abnormality detection and to provide a canonical atlas space.

Medical image registration is often considered a task with substantial computation times that may prevent its application in clinical workflows. While fully-convolutional segmentation networks can produce contours almost instantly, a classical iterative registration algorithm may often take minutes to converge. Hence, deep learning based image registration has been proposed to improve run times, so far however often with a degradation in alignment quality. The recent comprehensive medical registration challenge Learn2Reg showed that especially for large internal motion and multimodal fusion tasks, conventional methods are more robust and accurate than learning based approaches at the cost of longer runtimes. Yet little effort has recently been devoted in designing new iterative optimisation strategies for registration. GPU-accelerated optimisation routines have been explored in the AirLab framework [Sandkühler et al., 2020] and as instance refinement in the DL-based VoxelMorph and PDD [Balakrishnan et al., 2019; Heinrich, 2019]. Discrete optimisation may inherently reduce the number of iterations but has large memory requirements and either limited accuracy or a fixed capture range.

We hypothesis that a suitable combination of discrete and iterative schemes has been mostly overlooked in previous research but could provide new perspectives for medical registration without learning. We also note, that the limited availability of GPUs in clinical setups is often disregarded for run times - hence speeding up CPU computation bring real benefits. Since registration algorithms are often used as versatile tools to handle a variety of tasks an evaluation on complementary applications is desirable: here we consider monomodal inter-subject abdominal CT registration and multimodal intra-patient CT/MR fusion both in affine/rigid and nonlinear settings.

5.1.1.1 Related Work

Learning-based registration can be roughly subdivided in metric- and label-supervised approaches or combinations of them, where most algorithms employ a fully-convolutional multi-scale CNN architecture with (potentially multiple) spatial transformer layers [Balakrishnan et al., 2019; Mok et al., 2020]. Since, the alignment target can often not be determined alone by a limited number of manually segmented structures - e.g. fissures and lobes are not sufficient to guide intra patient lung registration - metric-supervised methods may be considered more general. In conventional image registration a generic similarity metric (possible based on hand-crafted features) is optimised together with a regulariser in a multi-scale and usually iterative fashion.

Two popular MRF-based discrete optimisation approaches, drop [Glocker et al., 2008] and deeds Heinrich et al., 2013a are related to our work. Drop employs Fast-PD as optimisation backend, limits the complexity by sampling discrete displacements only along the 3 principal axes (31 possible vectors in 3D) and iteratively refines the matching. It uses a B-spline transformation, in a multi-resolution setting and local cross-correlation as metric. Deeds relies on the faster MST-BP optimisation

[Felzenszwalb et al., 2005], and uses a dense discrete displacement search (up to 5000 vectors each). It employs a multi-scale approach with up to 5 warps, MIND features [Heinrich et al., 2013b] for similarity and symmetry constraints. Both methods run in $\approx$ 1 minute on multi-thread CPUs.

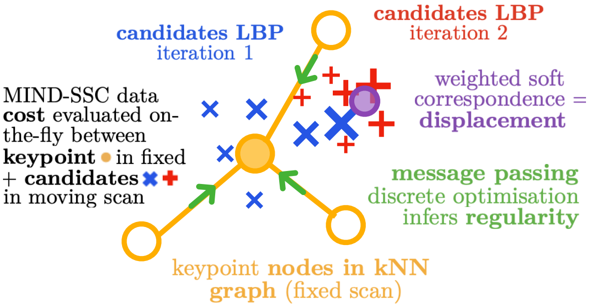
The closest work to our method in learning based registration is linearised multi-sampling [Jiang et al., 2019], which was proposed as an extension for spatial transformer networks to tackle the noisy gradient estimation in differentiable trilinear interpolation. Different to other DL-registration approaches the gradient with respect to a sub-pixel displacement is not determined by differentiating the interpolation coefficients directly, but instead a hyperplane is fitted at a small number of random sampling points in the proximity of the current displacement. Using the slope of this plane as gradient estimation was shown to be more stable especially for difficult transformations.

5.1.1.2 Contributions

We propose an efficient strategy for dynamically solving a regularised cost function that is founded in graph-based optimisation and surpasses most conventional and learning-based registration algorithms in terms of accuracy and speed. Similar to multi-sampling we employ a randomised set of displacement candidates nearby the previous solution at each iteration. But instead of directly computing a metric-based displacement gradient, we perform a probabilistic discrete optimisation using a sparse variant of loopy belief propagation (LBP) [Felzenszwalb et al., 2006] for the joint regularised cost function. In contrast to dense discrete optimisation, the space of considered displacements is reduced by orders of magnitudes and no fixed range of motion has to be pre-defined. The methods improves runtimes - with <1 seconds GPU or <5 seconds CPU - compared to DL registration, while matching or exceeding accuracy on three demanding tasks for inter- and intra-patient registration of CT-CT and MR-CT of the abdomen.

5.1.2 Methods

Our method comprises a spatially randomly distributed sampling of control points (keypoints) and a conventional feature extraction step using hand-crafted contrast-invariant descriptors followed by the proposed iterative (dynamic) loopy belief propagation optimisation of a regularised registration cost function. The registration is performed in a small number of outer iterations in which the displacement search is dynamically changed based on the probabilistic estimate of the inner optimisation iterations for LBP. The feature descriptors can be efficiently evaluated on-line through indexing or nearest neighbour interpolation for each keypoint in the fixed image and each candidate in the moving image. Finally, either a trimmed least square or thin-plate spline fitting is employed to obtain either a linear transform or extrapolate a nonlinear displacement field.



Algorithm: iterative LBP registration

```
1) Extract MIND for fixed and moving scan
2) Compute sparse keypoint graph (symmetric kNN) on fixed scan (storing edge indices)
3) initialise with zero displacement and iterate
    1) draw new displacement candidates
    2) compute MIND data term (sampling)
    3) iterate for loopy belief propagation
        1) gather incoming messages
        2) subtract reverse messages
        3) min-sum regularisation
        4) scatter outgoing messages
    4) compute soft(max) correspondences and update displacements
4) fit linear (least squares) or nonlinear (thin-plate spline) transformation for dense displacements → optionally second transform
```

Fig. 5.1: Schematic overview and pseudo-code of our proposed iterative optimisation approach for keypoints based 3D medical image registration using loopy belief (LBP) message passing.

We employ MIND with self-similarity context with 12 channels [Heinrich et al., 2013b] as state-of-the-art descriptor and enhance the capturing of spatial context by sampling a small local patch of size 3x3x3 of these descriptors at each keypoint positions p_{f_i}, resulting in a 324-dimensional vector. We use a symmetric k-nearest neighbour (kNN) graph on the set of keypoints in the fixed scan P_f with edges $(ij) \in E$ that connect keypoints p_{f_i} and p_{f_j}. To optimise a regularised cost function that minimises the dissimilarity between feature vectors in fixed and - at a displaced position - in the moving scan we use the sum of squared difference metric and define a diffusion like regulariser. In discrete optimisation the regularisation term is considered for each pair of possible candidate displacement for each edge using $r_{ij}^{pq} = \left|(c_i^p - p_{f_i}) - (c_j^q - p_{f_j})\right|_2^2$. In contrast to most common discrete optimisation schemes, the set of candidates c_j^q is not pre-determined or even equally spaced for each keypoint but dynamically adapted throughout iterations. The actual optimisation is performed in a few inner iterations, where messages are passed in parallel that enable the exchange of information across the graph and a more accurate deformation estimation. The algorithm is described in detail in [Felzenszwalb et al., 2006] and uses the following equation to compute outgoing messages $m_{i\to j}^t$ from p_{f_i} to p_{f_j} at iteration t: $m_{i\to j}^t = \min_{1,\ldots,q,\ldots l} \left(d_i + \alpha r_{ij}^q - m_{j\to i}^{t-1} + \sum_{(h,i)\in E} m_{h\to i}^{t-1}\right)$.

Once each LBP optimisation has converged, new candidates are drawn from a local neighbourhood (uniformly spatially distributed) and the search region can gradually adapt to the true optimum. To extrapolate from the final correspondences to a dense displacement either thin-plate splines or affine least-squares fitting is employed. A visual overview of the concept with a brief pseudo-code is given in Fig. 5.1.

Compared to conventional stochastic gradient descent methods our proposed optimisation scheme benefits from the discrete setting of the search that finds the combinatorial minimum of a wide range of displacements. In contrast to commonly used loopy belief propagation (which iterates only over steps of message passing) we introduce a second and important outer iteration over the capture ranges of displacements. This approach is to the best of our knowledge the first that couples a discrete belief propagation optimisation with a sparsely distributed and locally adaptive solution space of potential displacements for each control point.

5.1.3 Experiments and Results

To demonstrate the effectiveness and robustness of our iterative optimisation approach, we perform a comprehensive evaluation on three challenging abdominal datasets covering inter- and intra-patient, multimodal (CT/MR) as well as pre- and non-pre-aligned registration tasks. The datasets are described in detail in Section 5.1.3.1. The evaluation metrics are outlined below and closely follow the evaluation design of the L2R challenge, assessing accuracy and smoothness of the displacement field as well as the runtime of the algorithm. Finally, we describe implementation and configuration details of our proposed and comparison methods. All methods and experiments (with exception of drop2) were implemented and evaluated using the latest version of PyTorch (v1.7.1).

5.1.3.1 Datasets

CT-CT (L2R 2020) For our experiments on inter-patient alignment of abdominal CT scans we use the corresponding dataset (Task 3) from the Learn2Reg (L2R) challenge [Hering et al., 2022]. It contains 30 abdominal CT scans with 13 manually labelled anatomical structures: spleen, right kidney, left kidney, gall bladder, esophagus, liver, stomach, aorta, inferior vena cava, portal and splenic vein, pancreas, left adrenal gland, and right adrenal gland. The image data and labels themselves stem from [Xu et al., 2016] and were pre-registered to a canonical space and resampled to same voxel resolution and spatial dimensions (192x160x256). We report evaluation results on the predefined validation set, which describes a subset of 10 CT scans and 45 registration pairs.

CT-CT To study the effect of non-pre-aligned data and thus more complex deformations on the different registration methods we also consider the original data from [Xu et al., 2016]. The segmentation labels, resampling procedure of the CT scans and validation pairs remain the same as described above.

MR-CT Inter-patient alignment of abdominal CT to MR is investigated on 16 selected scan pairs from different trials in The Cancer Imaging Archive (TCIA) project

[Akin et al., 2016; Clark et al., 2013; Erickson et al., 2016; Linehan et al., 2016]. Four different sized organs (liver, spleen, left and right kidney) were manually labelled by us for each MR-CT scan pair to assess the registration accuracy. The data (with exception of a withheld test set) will be made publicly available as part of a larger dataset for abdominal registration later this year. Pre-processing comprises resampling the scan pairs to an isotropic resolution of 2 mm and cropping/padding to consistent voxel dimensions of 192 × 160 × 192. We report evaluation results over all 16 MR-CT pairs.

5.1.3.2 Evaluation Metrics

The used evaluation metrics cover the three most important aspects of medical image registration: 1) registration accuracy, 2) smoothness of the displacement field and 3) runtime of the algorithm. We assess the accuracy by means of alignment of the manually labelled organ segmentations using the Dice similarity metric (F1 score). In addition, the plausibility of the deformation field is of great clinical relevance and can be estimated by the standard deviation of the (logarithmic) Jacobian determinant [Kabus et al., 2009; Leow et al., 2007] (SDlogJ). Finally, we report the runtime of the registration methods, both on CPU (Apple M1) and GPU (NVIDIA RTX 2080 Ti), including all necessary computations after reading the images until predicting the final dense displacement field.

5.1.3.3 Comparison Methods and Hyper-Parameters

We employ a number of different related state-of-the-art learning- and optimisation-based registration methods in comparison to our proposed iterative LBP approach. Hyperparameters for ours and all comparison methods were selected on a subset of training scans for CT-CT and a small number of validation scans for MR-CT. Note, that iterative registration methods are much less sensitive to manual parameter choices than deep learning approaches. First, we substantially extended and improved upon the original VoxelMorph method [Balakrishnan et al., 2019] by replacing the simplistic U-Net with a two-stream architecture and a custom MIND-based metric-loss. Despite these advancements VoxelMorph+ (VM+) does not yield satisfactory results for very large misalignments (CT-CT w/o pre-align and MR-CT). Second, the overall runner-up in the Learn2Reg challenge, PDD-net [Heinrich et al., 2020], again with MIND loss, is used with all available extensions: multiple warps and instance optimisation for CT-CT, as well as affine least squares fitting for MR-CT. Third, drop2 a recent re-implementation of [Glocker et al., 2008] is used as a related baseline for discrete optimisation. We explored a wide range of settings and found that diffusion regularisation and an initial B-spline spacing of 80mm in combination with the following settings worked best: cross-correlation (weight: 0.25) and entropy correlation (weight: 0.5) for CT-CT and

MR-CT respectively with three pyramid levels each: 12mm, 8mm, 4mm for CT-CT and 8mm, 6mm, 4mm for MR-CT with a doubling of the iterations for the latter.

Furthermore, we evaluate two alternatives to our proposed method, all with the same metric patch-based MIND-SSC and one (nonlinear, for CT-CT (L2R 2020)) or two warps (affine + nonlinear, for CT-CT and MR-CT): dense 3D displacement sampling with LBP and continuous Adam optimisation with diffusion regularisation. For our proposed method, we chose 20 outer and 3 inner (LBP) iterations, where 12 (CT-CT) or 16 (MR-CT) displacement candidates are drawn from a uniform distribution for each of 2048 (or 8192 for nnUNet features) keypoints using capture ranges that start from ±15 voxels and linearly decrease to ±3 voxels. For Adam we employed a learning rate of 0.1 and 100 iterations. For the dense variant of LBP, we considered 11^3 displacements in parallel with a range of ±30 voxels. The regularisation is performed on a kNN graph with 10 nearest neighbours for the LBP variants and 16 neighbours for Adam. The regularisation weight (alpha) is set to 2.5 (LBP) and 0.2 (Adam), respectively. For all methods (with exception of Voxelmorph+) a coarse and convex binary mask is used to restrict registration to the region of interest (which includes approximately 50% of the original number of image voxels).

5.1.3.4 Results

Quantitative and qualitative results of our experiments are shown in Table 5.1 and Figure 5.2, respectively. For the pre-aligned CT-CT dataset our proposed iterative LBP approach yields a Dice score of 40.1%, which is competitive to the best performing deep learning based comparison method, PDD-Net, with 41.5%. The sparse candidate sampling strategy of our method enables fast runtimes of <1 second on GPU and approximately 5 seconds on CPU. Making use of label information (weakly supervised learning for Voxelmorph+, using nnUNet [Isensee et al., 2021] softmax features for optimisation based approaches) consistently and significantly improves the registration results by 8.5% points (VoxelMorph+), 18% points (Adam), 22% points (dense LBP) and 25.1% (iterative LBP). The inference time of the network is included in the total runtimes of the optimisation based methods. Our methods final Dice score of 65.2% is also comparable to the reported score of 67% of the winning entry (LapIRN Mok et al., 2020) of the L2R challenge on the hidden test set (validation and test results are generally comparable). Results on the non-pre-aligned CT-CT data show only a moderate decrease in Dice score of 37.7% for our method (-2.4% points). Other comparison methods are less robust, e.g. drop2 or Adam, decrease by 5.5% points and 5.7% points Dice score, respectively. In this experiment, in addition to the Dice metric, we evaluate the 95th percentile of the Hausdorff Distance (HD95) to further highlight the differences between the state-of-the-art continuous optimiser (Adam) and our proposed iterative LBP approach. We find initial values of 30.14 voxels, which are reduced by Adam to 26.1 voxels. Our proposed method can substantially and statistically significantly (p < 0.0002 using a Wilcoxon signed rank test) improve upon this with an error of only 17.3 voxels. For the inter-patient MR-CT dataset our iterative LBP method clearly yields the best Dice score of 76.4% with runtimes of 1.0 seconds on GPU and approximately 13 seconds on CPU.

5.1.4 Discussion and Conclusion

Differences to drop2, the iterative MRF registration technique proposed in [Glocker et al., 2008]: while the general idea of solving registration as iterative discrete optimisation is similar our method differs in a number of important design choices that greatly improve accuracy and runtime. First, our model uses a sparse keypoint graph and computes the similarity cost of displacements separately for each node, while [Glocker et al., 2008] uses a grid-based model. We employ a feature-based metric and found that spatially sampling one high-dimensional vector per keypoint sufficiently captures similarity and no spline interpolant has to be evaluated. Loopy belief propagation is used as discrete optimisation backend, which has a simpler implementation, provides probabilistic estimates and lends itself to parallelism. Finally, our displacement sampling

Table 5.1: Quantitative evaluation results of proposed and comparison methods on all three abdominal datasets considered. Runtimes are given for both GPU and CPU (GPU/CPU). Experiments in the first half (after the first double rule) are based on similarity metrics (MIND, NCC), while experiments in the second half (after the second double rule) make use of label information.

	CT-CT (L2R 2020)			CT-CT			MR-CT		
	Dice [%]	SDlogJ	Time [s]	Dice [%]	SDlogJ	Time [s]	Dice [%]	SDlogJ	Time [s]
Initial	25.1			11.6			26.5		
VM+	35.4	.134	**0.2**/13						
PDD-Net	**41.5**	.129	1.4/73	35.1	.131	1.4/73	63.6	.210	**1.0**/37
drop2	37.0	.147	-/43	31.5	.132	-/43	56.2	.109	-/68
Adam	36.6	.080	1.6/27	30.9	.087	2.7/46	71.1	.074	2.7/46
dense LBP	38.6	.119	0.8/14	36.4	.098	1.3/23	71.7	.085	1.3/23
iter LBP	40.1	.093	0.6/**5**	**37.7**	.092	**1.0/13**	**76.4**	.075	**1.0/13**
VM+	43.9	.162	**0.2/13**						
Adam	54.6	.046	3.4/52						
dense LBP	62.1	.170	2.2/33						
iter LBP	**65.2**	.088	1.9/22						

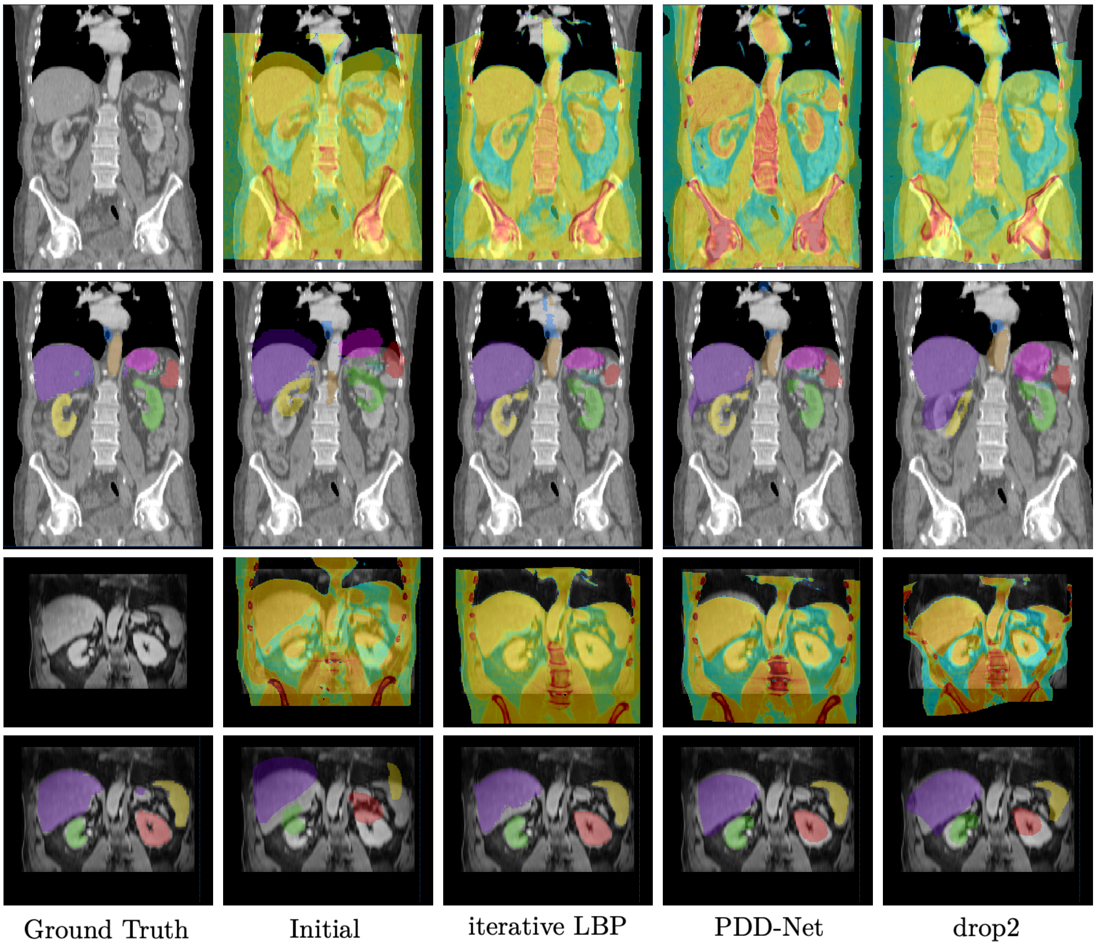


Fig. 5.2: Qualitative results of selected methods (coronal view). Row 1-2: Overlayed warped moving CT scan and warped segmentation labels for inter-patient CT-CT (L2R 2020). Row 3-4: intra-patient MR-CT. PDD and our proposed iterative LBP achieve more regular transformations than drop2 and are both very close to the ground truth for inter-subject alignment. Our method yields the highest accuracy for multimodal fusion especially at organ boundaries (lungs, kidneys).

injects more randomness by avoiding an axis parallel selection of motion vectors and thus converges faster.

We have developed an efficient discrete optimisation framework for versatile medical image registration that contrary to recent trends does not rely on learned convolutions and improves both runtime and accuracy compared to deep learning registration. This avoids lengthy preparations for supervised training and makes our method generally applicable. In addition, separating label prediction and optimisation-based registration offers improved explainability for clinicians compared to end-to-end learning methods (black box approach). The extensive experimental validation on three different abdominal registration tasks demonstrate substantial advantages over both unsupervised learning based and previous (iterative) discrete optimisers. We believe this is due to the following reasons: 1) our method avoids regular grids and dynamically adapts the shape of the solution space, finding an optimal balance between coverage and accuracy. 2) the combination of locally discriminative image features and globally regular graph optimisation can more efficiently address the computational demands on medical image registration than convolutional network architectures. A further surprising outcome is the competitiveness of Adam optimisation using regularisation across keypoints and MIND as loss term. This shows that iterative optimisation for registration should always be considered as a strong baseline when discussing new learning-based approaches. Future work could further reduce the complexity of supervised nnUNet feature extraction through quantisation and pruning.

5.2 Deformable Point Set Registration with Regularised GNNs

5.2.1 Introduction

Registration, i.e. determining a spatial transformation that aligns two images or point sets, is a fundamental task in medical image and shape analysis and a prerequisite for numerous clinical applications. It is widely used for image-guided intervention, motion compensation in radiation therapy, atlas-based segmentation or monitoring of disease progression. Non-rigid registration is ill-posed and thus a non-convex optimisation problem with a very high number of degrees of freedom. In addition, the medical domain poses particular challenges on the registration task, e.g. non-linear intensity differences in multi-modal images or high inter-patient variations in anatomical shape and appearance.

5.2.1.1 Related Work

Iconic registration Voxel-based intensity-driven medical image registration has been an active area of research, which can e.g. be solved using discrete [Glocker et al., 2008] optimisation of a similarity metric and a regularisation constraint on the smoothness of the deformation field. Data driven deep learning methods based on convolutional neural networks (CNNs), have only recently been used in the field of medical image registration. In [Vos et al., 2019] an iconic and unsupervised learning approach is introduced that learns features to drive a registration and replaces the iterative optimisation with a feed-forward CNN. While achieving impressive runtimes of under a second on a GPU the accuracy for CT lung motion estimation is inferior to conventional methods. Weak supervision in the form of landmarks or multi-label segmentations was used in the CNN framework of [Hu et al., 2018], where the similarity measure is based on the alignment of the registered labels.

Geometric Registration To capture large deformations, e.g. present in intra-patient inhale-exhale examinations of COPD patients [Castillo et al., 2013] or vessel-guided brain shift compensation [Bayer et al., 2018], geometric registration models - based on keypoints or surfaces - offer a promising solution. Point-based registration has not yet profited from the advantages of deep feature learning due to the restriction of conventional CNNs to densely gridded input. Many current geometric methods (e.g. [Bayer et al., 2018] and [Ravikumar et al., 2019]) are based on the well-established coherent point drift (CPD) algorithm [Myronenko et al., 2010]. In addition to 3D coordinates, they incorporate further image or segmentation-derived features, such as point orientations or scalar fractional anisotropy (FA) values [Ravikumar et al., 2019].

Geometric Deep Learning While these hand-crafted features clearly improved on the results of the CPD, recent methods from the field of geometric deep learning [Bronstein et al., 2017] would enable a data-driven feature extraction directly from point sets. The PointNet framework [Qi et al., 2017a] was one of the first approaches to apply deep learning methods to unordered point sets. A limitation of the approach is that is does not consider local neighbourhood information, which was addressed in [Wang et al., 2019a] by dynamically building a k-nearest-neighbour graph on the point set and thus also enabling feature propagation along edges in that graph. Combining convolutional feature learning with a differentiable and robustly regularised fitting process has first been proposed for multi-camera scene reconstruction in [Brachmann et al., 2017] (DSAC), but has so far been limited to rigid alignment.

Large Deformation Lung Registration Both iconic and geometric approaches have often been found to yield relative large residual errors for large motion lung registration (forced inhale-to-exhale): e.g. 4.68 mm for the discrete optimisation algorithm in [Glocker et al., 2008] applied to the DIR-lab COPD data [Castillo et al., 2013] and 3.61 mm (on the inhale-exhale pairs of the EMPIRE10 challenge) for [Ehrhardt et al., 2010], which used both keypoint- and intensity-based information. Learning the alignment of such difficult data appears to be so far impossible with intensity-driven CNN approaches that already struggle with more shallow breathing in 4D-CT [Vos et al., 2019]. Thus being able to directly match vessel- and airway trees based on geometric features alone can provide a valuable pre-alignment for further intensity-based registration (cf. [Heinrich et al., 2015]) or be directly used in clinical applications to perform atlas-based labelling of anatomical segments and branchpoints for physiological studies [Tschirren et al., 2005].

5.2.1.2 Contributions

Our work contributes two important steps towards data-driven point set registration that enables the incorporation of deep feature learning into a regularised CPD fitting algorithm. First, we utilise dynamic graph CNNs [Wang et al., 2019a] in an auxiliary metric learning task to establish robust correspondences between a moving and a fixed point set. These learned features are shown to yield an improved modelling of prior probabilities in the CPD algorithm. Since all operations of the CPD algorithm are differentiable, we secondly show that it is possible to further optimise the parameters of the feature extraction network directly on the registration task. To evaluate our method we register keypoints extracted from inhale and exhale states in lung CT-scans from the challenging DIR-Lab COPD dataset [Castillo et al., 2013] showing the general feasibility of a deep learning point set registration framework in an end-to-end manner and with only geometric information.

5.2.2 Methods

In this section, we introduce our proposed method for deformable point set registration with deeply learned features. Figure 5.3 summarises the methods general idea. Input to our method are the fixed point set P_F and the moving point set P_M. While we make no assumptions on the number of points or correspondences in the input point sets, we assume a further set of keypoint correspondences with P_F for the supervised learning task, which is denoted as P_C. We compute geometric features from P_F and P_M with a shared dynamic graph CNN (DGCNN [Wang et al., 2019a]) ϕ. The spatial positions together with the extracted descriptors are input to the feature based CPD algorithm that produces displacement vectors for all points in P_M. We then employ thin-plate splines (TPS) [Bookstein, 1989] as a scattered data interpolation method to compute the displacements for P_C, which yields the transformed point set $P_C^{'}$. Finally, we can compute the mean squared error (MSE) of the Euclidean distance between correspondences in P_F and $P_C^{'}$ as a loss L for the optimisation of the feature extraction network ϕ. In the following, we describe the descriptor learning with the DGCNN as well as the extensions to the CPD algorithm to exploit point features as prior probabilities.

5.2.2.1 Descriptor Learning on Point Sets with Dynamic Graph CNNs

Our proposed network architecture for geometric feature extraction is illustrated in Figure 5.4. A key component is the edge convolution introduced in [Wang et al., 2019a], that dynamically builds a k-Nearest-Neighbor (kNN) graph from the points in the input feature space and then aggregates information from neighbouring points to output a final feature map. We employ several edge convolutions with DenseNet style feature concatenation to efficiently capture both local and global geometry. The final feature

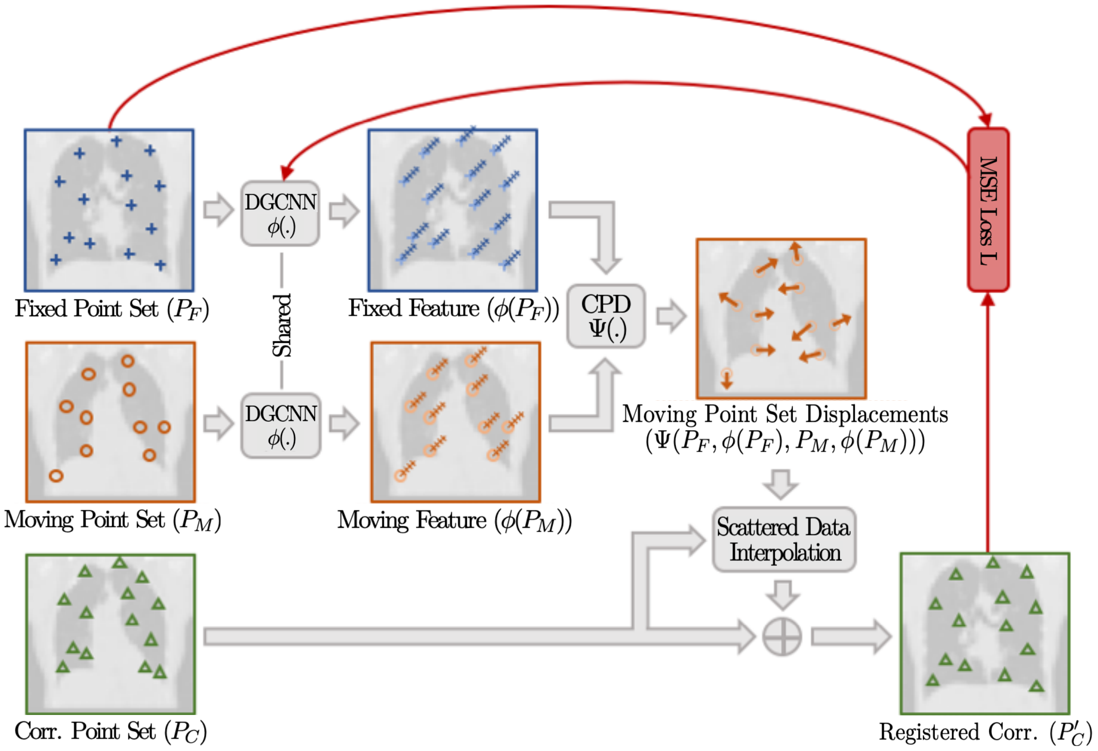

Fixed Point Set (P_F)
DGCNN $\phi(.)$
Shared
Fixed Feature ($\phi(P_F)$)
CPD $\Psi(.)$
MSE Loss L
Moving Point Set (P_M)
DGCNN $\phi(.)$
Moving Feature ($\phi(P_M)$)
Moving Point Set Displacements ($\Psi(P_F, \phi(P_F), P_M, \phi(P_M))$)
Scattered Data Interpolation
Corr. Point Set (P_C)
Registered Corr. (P'_C)

Fig. 5.3: Illustration of our proposed method for supervised non-rigid point set registration. While we investigate the problem of 3D registration, here, point sets are depicted in two dimensions for simplicity. Also, point sets are underlaid with coronal lung CT slices as visualization aides. No image information is used in our registration pipeline.

descriptor is obtained by fully connected layers that reduce the point information to a given dimensionality. We restrict the output descriptor space by L_2 normalisation to enable constant parametrisation of subsequent operations in the registration pipeline which stabilizes network training. To establish robust initial correspondences between the moving and fixed point set the model is pretrained in an auxiliary metric learning task using a triplet loss.

5.2.2.2 Feature-Based Coherent Point Drift

The CPD algorithm formulates the alignment of two point sets as a probability density estimation problem. The points in the moving point set P_M are described as centroids of gaussian mixture models (GMMs) and are fitted to the points in the fixed point set P_F by maximizing the likelihood. To find the displacements for P_M the Expectation Maximization (EM) algorithm is used, where in the E-step point correspondence probabilities C are computed and in the M-step the displacement vectors are updated.

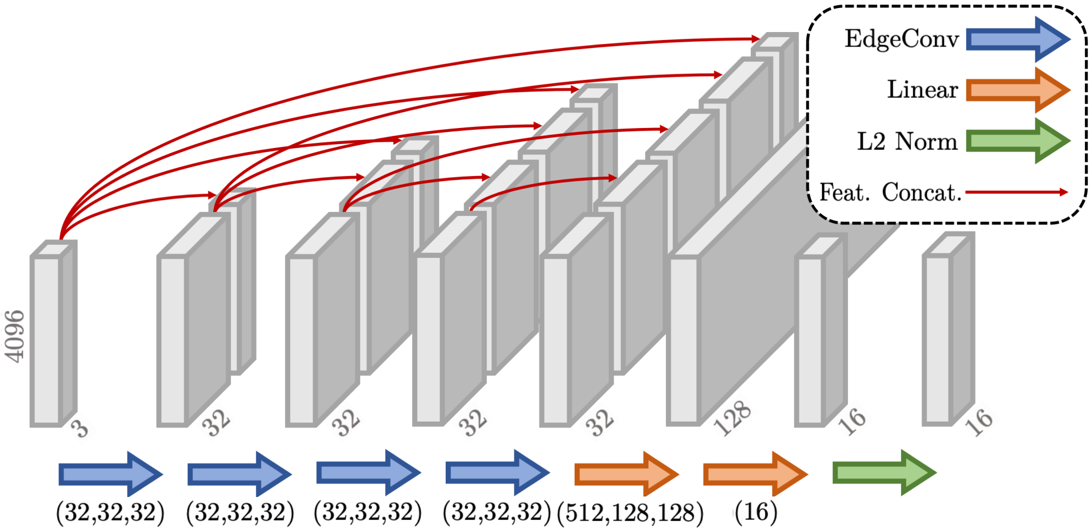


Fig. 5.4: Proposed network architecture for geometric feature extraction from the fixed and moving point set. Input is a three-dimensional point set and the network computes a 16-dimensional geometric descriptor for each of the 4096 points. The number of layer neurons for each operation is specified in the corresponding brackets.

We incorporate the learned geometric feature descriptors $\phi(P_F)$ and $\phi(P_M)$ as additional prior probabilites with

$$C(P_F, \phi(P_F), P_M, \phi(P_M)) = C_{pos}(P_F, P_M) + \alpha \cdot C_{feat}(\phi(P_F), \phi(P_M)), \qquad (5.1)$$

where C_{pos} denotes the spatial point correspondence described in [Myronenko et al., 2010], α is a trade-off and scaling parameter and

$$C_{feat_{mn}}(\phi(P_F)_n, \phi(P_M)_m) = \exp(-\frac{1}{2 \cdot \rho^2} \|\phi(P_F)_n - \phi(P_M)_m\|^2) \qquad (5.2)$$

with $n = 1 \ldots N$ and $m = 1 \ldots M$. N and M denote the number of points in P_F and P_M, respectively. In addition to the parameter ρ in Eq. 5.2, that controls the width of the Gaussian, the CPD algorithm includes three more free parameters: w, λ and β. Parameter w models the amount of noise and outliers in the point sets, while parameters λ and β control the smoothness of the deformation field.

5.2.3 Experiments and Results

Registering the fully inflated to exhaled lungs is considered one of the most demanding tasks in medical image registration, which is important for analysing e.g. local ventilation defects in COPD patients. We use the DIR-Lab COPD data set [Castillo et al., 2013] with 10 inhale-exhale pairs of 3D CT scans for all our experiments. The thorax volumes are resampled to isotropic voxel-sizes of 1 mm and a few thousands keypoints are extracted from inner lung structures with the Foerstner operator. Automatic correspondences to supervise the learning of our DGCNN are established using the discrete and intensity-based registration algorithm of [Heinrich et al., 2015], which has an accuracy of ∼1 mm. In all experiments, no CT-based intensity information is used and all processing relies entirely on the geometric keypoint locations.

In our first experiment, we learn point descriptors directly in a supervised metric learning task. Therefore, a triplet loss is employed forcing feature similarity between corresponding keypoint regions in point set pairs. The inhale and exhale point set form the positive pair, while points from the permuted exhale point set yield as negative examples. These learned features can be directly used in a kNN registration. We then investigate the combination of spatial positions and learned descriptors in the feature-based CPD algorithm. Finally, in our concluding experiment, the feature network is trained in an end-to-end manner as described in Section 5.2.2 to further optimize the pretrained geometric features.

5.2.3.1 Implementation Details

Due to the limited number of instances in the used dataset we perform a leave-one-out validation, where we evaluate on one inhale and exhale point set and train our network with the remaining nine pairs. During training we use farthest point sampling to obtain 4096 points from the inhale and exhale point set, respectively. Each evaluation is run ten times and results are averaged to account for the effect of the sampling step. The employed network parameters are specified in Figure 5.4. For the CPD algorithm (250 iterations) we use following parameters: $\alpha = 0.05$, $\rho = 0.5$ $w = 0.1$, $\lambda = 5$ and $\beta = 1$. For the end-to-end training we relax parameters ρ and β to 0.25 and 0.5, respectively, to allow for further optimization of input features.

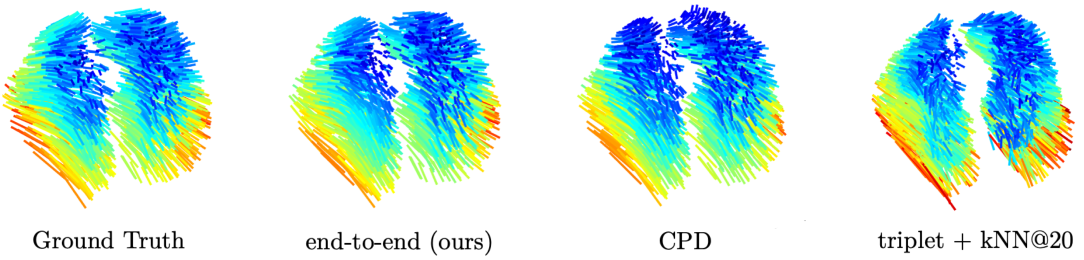


Fig. 5.5: Qualitative results in terms of 3D motion vectors on test case #5. The magnitude is color coded from blue (small motion) to red (large motion).

Table 5.2: Results for the 10 inhale and exhale CT scan pairs of the DIR-Lab COPD data set Castillo et al., 2013. The mean target registration error (TRE) in mm is computed on the 300 expert annotated landmark pairs per case. The p-values are obtained by a rank-sum test over all 3000 landmark errors with respect to our best performing approach.

Case #	initial	center-aligned	triplet + kNN@20	CPD	triplet + CPD (ours)	end-to-end (ours)
1	26.3	17.8	8.1	5.5	4.2	**3.4**
2	21.8	14.7	15.6	**8.4**	9.3	8.9
3	12.6	10.6	6.4	2.7	2.5	**2.4**
4	29.6	19.0	8.3	4.8	3.4	**3.2**
5	30.1	18.4	7.8	8.4	5.2	**4.6**
6	28.5	16.2	7.5	14.0	5.1	**4.3**
7	21.6	10.2	6.3	3.0	2.6	**2.5**
8	26.5	17.4	6.3	6.8	4.3	**3.9**
9	14.9	14.1	9.0	3.5	**3.1**	3.6
10	21.8	19.6	14.9	**7.4**	7.5	**7.4**
mean	23.4	15.7	9.0	6.4	4.7	**4.3**
std	11.9	7.0	5.5	5.2	4.1	**3.6**
p-val	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$1.4 \cdot 10^{-2}$	-

5.2.3.2 Results

Qualitative results are shown in Figure 5.5 where our approach demonstrates a good trade-off between the very smooth motion of the CPD and the potential for large correspondences of the features from triplet-learning. Our quantitative results that are evaluated on 300 independent expert landmark pairs for each patient demonstrate that registering the point clouds directly with CPD (3D coordinates as input) yield a relatively large target registration error (TRE) of 6.4±5.2 mm (see Table 5.2). Employing kNN registration based on a DGCNN trained with keypoint correspondences to extract geometric features without regularization is still inferior with a TRE of 9.0±5.5 mm highlighting the challenges of this point-based registration task and the difficulties of addressing the deformable alignment with one-to-one correspondence search. Combining the geometric features of a pre-trained DGCNN with the regularizing CPD that is extended to use 19-dimensional inputs (16 features + 3 coordinates) yields a substantial improvement over each individual method with a TRE of 4.7±4.1 mm. Finally, using end-to-end learning to back-propagate the regularized alignment errors through the iterative point drift layers to further improve the feature learning shows another small but significant improvement to 4.3±3.6 mm. These alignment errors cannot be directly compared to the large variety of image- and feature-based registration algorithms that reached 3.6 mm [Ehrhardt et al., 2010], 4.7 mm [Glocker et al., 2008] or 1.1 mm [Heinrich et al., 2015] for similar datasets, but were based on intensity information, while our comparison is restricted to purely geometric approaches without intensity. In addition, a better outcome would be expected by extending the keypoint extraction to focus on vessel- or airway-based nodes and to include anatomical tree-based edges in the graph model. Nevertheless, the results clearly showed that our models are already able to directly learn semantic geometric features in a data-driven manner based on the inherent correspondence information.

5.2.4 Discussion and Conclusion

We have presented a new method for deformable point set registration that learns geometric features from irregular point sets using a dynamic graph CNN (DGCNN) together with a regularizing and fully differentiable high-dimensional coherent point drift (CPD) model. Our results clearly indicate that geometric feature learning, even from relatively uninformative point clouds, is possible with DGCNNs and can be further enhanced when incorporating the CPD model into the optimization. Evaluated on challenging inhale-exhale lung registration of COPD patients we achieve an improvement of 2.1 mm over the classical CPD method and are competitive with many classical image-based registration algorithms despite the fact that no intensity information is used. In addition to these encouraging findings, we believe that alternative regularization models to the CPD, that require fewer iteration steps could have potential to further improve this approach. In future works, many more applications, e.g. surface point shape alignment and analysis, could benefit from deep point registration.

5.3 Deep Learning Based Geometric Registration without Visual Features

5.3.1 Introduction

Current learning approaches for medical image analysis predominantly consider the processing of volumetric scans as a dense voxel-based task. However, the underlying anatomy could in many cases be modelled more efficiently using only a sparse subset of relevant geometric keypoints. When sufficient amounts of labelled training data are available and the region of interest can be robustly initialised, sparse surface segmentation models have been largely outperformed by dense fully-convolutional networks in the past few years [Isensee et al., 2021]. However, dense learning based image registration has not yet reached the accuracy of conventional methods for the estimation of large deformations where geometry matters - e.g. for inspiration-expiration lung CT alignment. The combination of iconic (image-based) and geometric registration approaches have excelled in deformable lung registration but they are often time-consuming and rely on multiple steps of pre-alignment, mask-registration, graph-based optimisation and multi-level continuous refinement with different image-based metrics [Rühaak et al., 2017]. In this work, we aim to address 3D lung registration as a purely geometric alignment of two point clouds (a few thousand 3D points for inhale and exhale lungs each). While this certainly reduces the complexity of the dense deformable 3D registration task, it may also reduce the accuracy since intensity- and edge-based clues are no longer present. Yet, we demonstrate in our experimental validation that even this limited search range for potential displacements leads to huge and significant gains compared to dense learning based registration frameworks - mainly stemming from the robustness of our framework to implicitly learn the geometric alignment of vessel and airway trees.

5.3.1.1 Related Work

Point Cloud Learning Conventional point cloud registration (iterative closest point, coherent point drift [Besl et al., 1992; Myronenko et al., 2010]) often focused on the direct alignment of unstructured 3D points based on their coordinates. Newer work on graph convolutional learning has demonstrated that relevant geometric features can be extracted from point clouds with neighbourhood relations defined on kNN graphs and enable semantic labelling or global classification of shapes, object parts and human poses and gestures [Bronstein et al., 2017; Qi et al., 2017a]. Graph Convolutional Networks (GCN) [Kipf et al., 2017] define localised filter and use a polynomial series of the graph Laplacian (Tschebyscheff polynomials) further simplified to the immediate neighbourhood of each node. The graph attention networks introduced in [Velickovic et al., 2018] are a promising extension based on attention mechanism. Similarly, dynamic edge convolutions [Wang et al., 2019a] achieve information propagation by learning a function that predicts pairwise edge weights based on previous features of both considered nodes.

Learning Based Image Registration In image registration, learning based methods have surpassed their untrained optimisation-based counterparts in terms of accuracy and speed for 2D optical flow estimation, where millions of realistic ground truth displacement fields can be generated [Sun et al., 2018a]. Advantages have also been found for certain 3D medical registration tasks, for which thousands of scans with pixel-level expert annotations are available and the complexity of deformations is well represented in the training dataset [Balakrishnan et al., 2019; Mok et al., 2020; Xu et al., 2019]. As evident from a recent medical registration challenge [Hering et al., 2022], deep learning has not yet reached the accuracy and robustness for inspiration to expiration CT lung registration, where detailed anatomical labels are scarce (learning lobe alignment might not directly translate into low registration errors [Hering et al., 2021]) and the motion is large and complex. Even for the simpler case of shallow breathing in 4D CT, few learning-based works have come close to the best conventional methods (e.g. [Rühaak et al., 2017]) despite increasingly complex network pipelines [Fu et al., 2020].

Learning Graphical Registration More recent research in computer vision has also explored geometric learning for 3D scene flow [Liu et al., 2019c] that aims to register two 3D point clouds by finding soft correspondences. The challenge stems from the difficulty of jointly embedding two irregular point cloud (sub-)sets to enable end-to-end learning of geometric features and correspondence scores. Other recent approaches in point set registration/matching combine deep feature learning with GCNs and classical optimisation techniques, to solve the optimal transport [Puy et al., 2020] or reformulate traditional matching algorithms into deep network modules [Sarlin et al., 2020]. In the medical domain, combining sparse MRF-based registration [Sotiras et al., 2010] and multi-level continuous refinement [Rühaak et al., 2017] yielded the highest accuracy for two 3D lung benchmarks comprising inspiration and expiration [Castillo et al., 2013; Murphy et al., 2011].

We strongly believe that geometry can be a key element in advancing learning based registration and that the focus on visual features and fully-convolutional networks has for certain applications diverted research from mathematically proven graphical concepts that can excel within geometric networks.

5.3.1.2 Contributions

We propose a novel geometric learning method for large motion estimation across lung respiration that combines graph convolutional networks on keypoint clouds with sparse message passing. Our method considers geometric registration as soft correspondence search between two keypoint clouds with a restricted set of candidates from the moving point cloud for each fixed keypoint. 1) We are the first to combine edge convolutions as end-to-end geometric feature learning from sparse keypoints with differentiable loopy belief propagation (discrete optimisation) for regularisation of displacements on a kNN graph adapted to irregular sets of candidates for each node. 2) Our compact yet elegant networks, demonstrate surprisingly large gains in accuracy and outperform deep learning approaches that make use of additional visual clues by more than 50% reduced target registration errors for lung scans of COPD patients. 3) We present a further novel variant of our approach that discretises the sparse correspondence probabilities using differentiable extrapolation for a further six fold gain in computational efficiency and with similar accuracy.

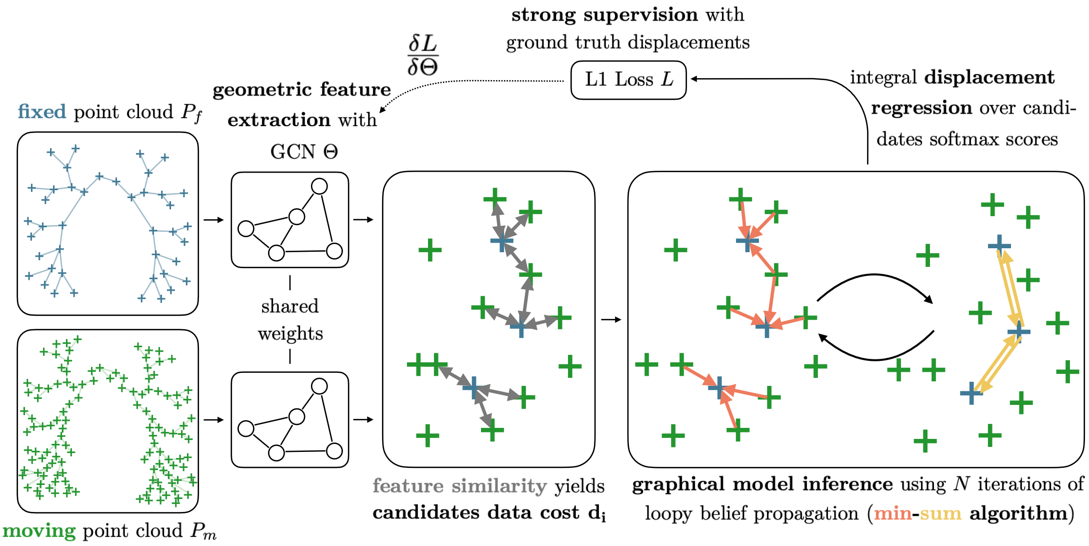

strong supervision with
ground truth displacements
$\frac{\delta L}{\delta \Theta}$
L1 Loss L
integral displacement regression over candidates softmax scores
geometric feature extraction with GCN Θ
fixed point cloud P_f
shared weights
moving point cloud P_m
feature similarity yields candidates data cost $\mathbf{d_i}$
graphical model inference using N iterations of loopy belief propagation (min-sum algorithm)

Fig. 5.6: Overview of our proposed method for accurate point cloud alignment using geometric features combined with loopy belief propagation in an end-to-end trainable deep learning registration framework.

5.3.2 Methods

5.3.2.1 Loopy Belief Propagation for Regularised Registration of Keypoint Graphs

We aim to align two point clouds, a fixed point cloud P_f ($|P_f| = N_f$) and a moving point cloud P_m ($|P_m| = N_m$). They consist of distinctive keypoints $p_{f_i} \in P_f$ and $p_{m_i} \in P_m$. We further define a symmetric k-nearest neighbour (kNN) graph on P_f with edges $(ij) \in E$ that connect keypoints p_{f_i} and p_{f_j}. A displacement vector $v_i \in V$ for each fixed keypoint p_{f_i} is derived from soft correspondences from a restricted set of possible candidates $c_i^p \in C_i$ (determined by l-nearest neighbour search ($|C_i| = l$) in the moving point cloud P_m). The regularised motion vector field V is inferred using loopy belief propagation enforcing spatial coherence of motion vectors. The data cost d_i^p ($d_i = (d_i^1, \ldots, d_i^p, \ldots, d_i^l)$) for a fixed point p_{f_i} and a single candidate c_i^p is modeled as

$$d_i^p = \|\theta(p_{f_i}) - \theta(c_i^p)\|_2^2\,, \tag{5.3}$$

where $\theta(.)$ denotes a general feature transformation of the input point (e.g. deep learning based geometric features, see Section 5.3.2.2). Especially in this case of sparse to sparse inference, missing or noisy correspondences can lead to severe registration errors. Therefore, a robust regularisation between neighbouring fixed keypoints (defined by edges $(ij) \in E$) is enforced by penalizing the deviation of relative displacements.

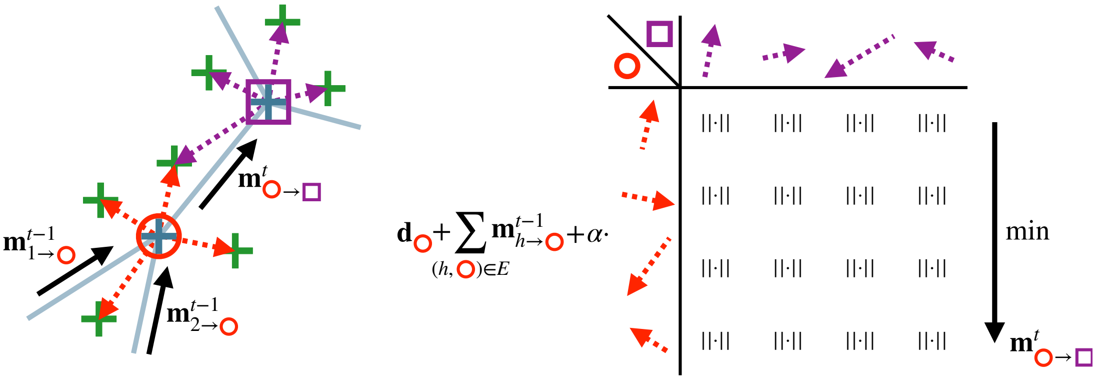

$\mathbf{m}^{t-1}_{1\to○}$
$\mathbf{m}^{t}_{○\to□}$
$\mathbf{m}^{t-1}_{2\to○}$
$\mathbf{d}_{○}+\sum_{(h,○)\in E}\mathbf{m}^{t-1}_{h\to○}+\alpha\cdot$
$\|\cdot\|$
min
$\mathbf{m}^{t}_{○\to□}$

Fig. 5.7: Illustration of proposed message passing scheme for keypoint registration. The current outgoing message for the considered keypoint is composed of the candidates data cost and incoming messages from the previous iteration. In addition, the squared deviation (weighted by α) of candidate displacements is minimised for a coherent motion across the kNN graph. Reverse messages are not shown for visual clarity.

The regularisation cost r_{ij}^{pq} ($r_{ij}^{q} = (r_{ij}^{1q}, \ldots, r_{ij}^{pq}, \ldots, r_{ij}^{lq})$) for two fixed keypoints p_{f_i}, p_{f_j} and candidates c_i^p, c_j^q can then be described as

$$r_{ij}^{pq} = \left\|(c_i^p - p_{f_i}) - (c_j^q - p_{f_j})\right\|_2^2. \tag{5.4}$$

To compute the marginal distributions of soft correspondences over the fixed kNN graph we employ N iterations of loopy belief propagation (min-sum algorithm) with outgoing messages $m_{i\to j}^{t}$ from p_{f_i} to p_{f_j} at iteration t defined as

$$m_{i\to j}^{t} = \min_{1,\ldots,q,\ldots l}\left(d_i + \alpha r_{ij}^{q} - m_{j\to i}^{t-1} + \sum_{(h,i)\in E} m_{h\to i}^{t-1}\right). \tag{5.5}$$

The hyperparameter α weights the displacement deviation penalty and thus controls the smoothness of the motion vector field V. Initial messages $m_{i\to j}^{0}$ are set to 0. A graphical description of the presented message passing scheme is also shown in Figure 5.7 and for further in-depth details on efficient belief propagation the reader is referred to [Felzenszwalb et al., 2006].

Fast Approximation Using a Discretised Candidates Space While the proposed message passing approach is easily parallelisable, it still lacks some efficiency as the number of messages to compute for each keypoint is dependent on the number of neighbours k. We propose to reduce the number of message computations per node to 1 by discretising the sparse candidates cost d_i in a dense cost volume D_i with fixed grid resolution r. Voxelisation of sparse input has been used in point cloud learning to speed up computation [Liu et al., 2019d]. D_i can be efficiently populated using nearest neighbour interpolation at (normalised) relative displacement locations $o_i^p = (o_{i_x}^p, o_{i_y}^p, o_{i_z}^p) = c_i^p - p_{f_i}$, evaluating

$$D_i(u, v, w) = \frac{1}{N_{u,v,w}} \sum_{p=1}^{l} \mathbb{I}[\lfloor o_{i_x}^p r \rfloor = u, \lfloor o_{i_y}^p r \rfloor = v, \lfloor o_{i_z}^p r \rfloor = w] d_i^p, \tag{5.6}$$

where (following notations in [Liu et al., 2019d]) $\mathbb{I}[\cdot]$ denotes a binary indicator that specifies whether the location o_i^p belongs to the voxel grid (u, v, w) and $N_{u,v,w}$ is a normalisation factor (in case multiple displacements end up in the same voxel grid). By operating on the dense displacement space D_i, we can employ an efficient quadratic diffusion regularisation using min convolutions [Felzenszwalb et al., 2006] that are separable in dimensions and also avoid the costly computation of k different messages per node. Approximation errors stem solely from the discretisation step.

5.3.2.2 Geometric Feature Extraction with Graph Convolutional Neural Networks

Distinctive keypoint graphs that describe plausible shapes contain inherent geometric information. These include local features such as curvature but also global semantics of the graph (e.g. surface or structure connectivity). Recent work on data-driven graph convolutional learning has shown that descriptive geometric features can be extracted from point clouds with neighbourhood relations defined based on kNN graphs. Edge convolutions [Wang et al., 2019a] can be interpreted as irregular equivalents to dense convolutional kernels. Following notations in [Wang et al., 2019a] we define edge features $e_{ij} = h_\theta(f_i, f_j - f_i)$, where f_i denote F-dimensionsal features on points $p_i \in P$ (first feature layer given as $f_i = p_i$). The edge function h_θ computes the Euclidean inner product of the learnable parameters $\theta = (\theta_1, \ldots, \theta'_F)$ with f_i (keypoint information) and $f_j - f_i$ (local neigbourhood information). The F'-dimensional feature output f'_i of an edge convolution is then given by

$$f_i^{'} = \max_{(i,j)\in E} e_{ij}, \tag{5.7}$$

where the max operation is to be understood as a dimension-wise aggregation function. Employing multiple layers of edge convolutions in a graph neural network and applying it to the fixed and moving point clouds (P_f, P_m) yields descriptive geometric features, which can be directly used to compute candidate data costs (see Equation 5.3).

5.3.2.3 Deep Learning Based End-to-End Geometric Registration Framework

Having described the methodological details, we now summarise the full end-to-end registration framework (see Figure 5.6 for an overview). Input to the registration framework are the fixed P_f and moving P_m point cloud. In a first step, descriptive geometric features are extracted from P_f and P_m with a graph convolutional network θ (shared weights). The network consists of three edge convolutional layers, whereby edge functions are implemented as three layers of 1×1 convolutions, instance normalisation and leaky ReLUs. Feature channels are increased from 3 to 64. Two 1×1 convolutions output the final 64-dimensional point feature embeddings. Thus, the total number of free trainable parameters of the network is 26880. In general, the moving cloud will contain more points than the fixed cloud (to enable an accurate correspondence search). To account for this higher density of P_m, the GCN θ acts on the kNN graph for P_f and on the $3k$NN graph for P_m. As described in Section 5.3.2.1 the geometric features $\theta(P_f)$ and $\theta(P_m)$ are used to compute the candidates cost and final marginal distributions are obtained from N iterations of (sparse or discretised) loopy belief propagation. As all operations in our optimisation step are differentiable the network parameters can be trained end-to-end. The training is supervised with ground truth motion vectors $\hat{\mathbf{v}}_i \in \hat{V}$ (based on 300 available manual annotated and corresponding landmark pairs) using an L1 loss (details on integral regression of the predicted motion vectors V from the marginals in Section 5.3.2.4).

5.3.2.4 Implementation Details: Keypoints, Visual Features and Integral Loss

While our method is generally applicable to a variety of point cloud tasks, we adapted parts of our implementation to keypoint registration of lung CT.

Keypoints We extract Förstner keypoints with non-maximum suppression as described in [Heinrich et al., 2015]. A corner score (distinctiveness volume) is computed using $D(x) = 1/\operatorname{trace}\left((G_\sigma * (\nabla F \nabla F^T))^{-1}\right)$, where G_σ describes a Gaussian kernel and ∇F spatial gradients of the fixed/moving scans computed with a seven-point stencil. Additionally, we modify the extraction to allow for a higher spatial density of keypoints in the moving scan by means of trilinear upsampling of the volume before non-maximum suppression. Only points within the available lung masks are considered.

Visusal Features To enable a fair comparison to state-of-the-art methods that are based on image intensities, we also evaluate variants of all geometric registration approaches with local MIND-SSC features Heinrich et al., 2013b. These use a 12-channel representation of local self-similarity and are extracted as small patches of size $3 \times 3 \times 3$

with stride=2. The dimensionality is then further reduced from 324 to 64 using a PCA (computed on each scan pair independently).

Integral Loss As motivated before, we aim to find soft correspondences that enable the estimation of relative displacements, without directly matching a moving keypoint location, but rather a probability for each candidate. A softmax operator over all candidates is applied to the negated costs after loopy belief propagation (multiplied by a heuristic scalar factor). These normalised predictions are integrated over the corresponding relative displacements. When considering a discretised search space (the dLBP variant), final displacements are obtained via integration over the fully quantised 3D displacement space. To obtain a dense displacement field for evaluation (landmarks do not necessarily coincide with keypoints), all displacement vectors of the sparse keypoints are accumulated in a displacement field tensor using trilinear extrapolation and spatial smoothing. This differentiable dense extrapolation enables the use of an L1 loss on (arbitrary) ground truth correspondences.

5.3.3 Experiments and Results

To demonstrate the effectiveness of our novel learning-based geometric 3D registration method, we perform extensive experimental validation on the DIR-Lab COPDgene data [Castillo et al., 2013] that consists of 10 lung CT scan pairs at full inspiration (fixed) and full expiration (moving), annotated with 300 expert landmarks each. Our focus lies in evaluating point cloud registration without visual clues and we extract a limited number of keypoints (point clouds) in fixed (≈2000 each) and moving scans (≈6000 each) within the lungs. Since, learning benefits from a variability of data, we add 25 additional 3D scan pairs showing inhale-exhale CT from the EMPIRE10 [Murphy et al., 2011] challenge, for which no landmarks are publicly available and we only include automatic correspondences generated using [Heinrich et al., 2015] for supervision. We performed leave-one-out cross validation on the 10 COPD scans with sparse-to-dense extrapolation for landmark evaluation. Training was performed with a batch size of 4 and an initial learning rate of 0.01 for 150 epochs. All additional hyperparamters for baselines and our proposed methods (regularisation cost weighting α, scalar factor for integral loss, etc.) were tuned on case #4 of the COPDgene dataset and left unaltered for the remaining folds.

Overall, we compare five different algorithms that work purely on geometric information, five further methods that use visual input features and one deep-learning baseline for dense intensity registration (the winner of the Learn2Reg 2020 challenge LapIRN [Mok et al., 2020]). Firstly, we compare our proposed sparse-LBP regularisation with geometric feature learning (sLBP+GF) to a version without geometric learning (sLBP) and coherent point drift [Myronenko et al., 2010] without (CPD) and with geometric feature learning (CPD+GF). The non-learning based methods directly use the keypoint

Table 5.3: Results of methods based on geometric features and optimisation on the COPDgene dataset [Castillo et al., 2013]. We report the target registration error (TRE) in millimeters for individual cases as well as the average distance and standard deviation over all landmarks. The average GPU runtime in seconds is listed in the last row.

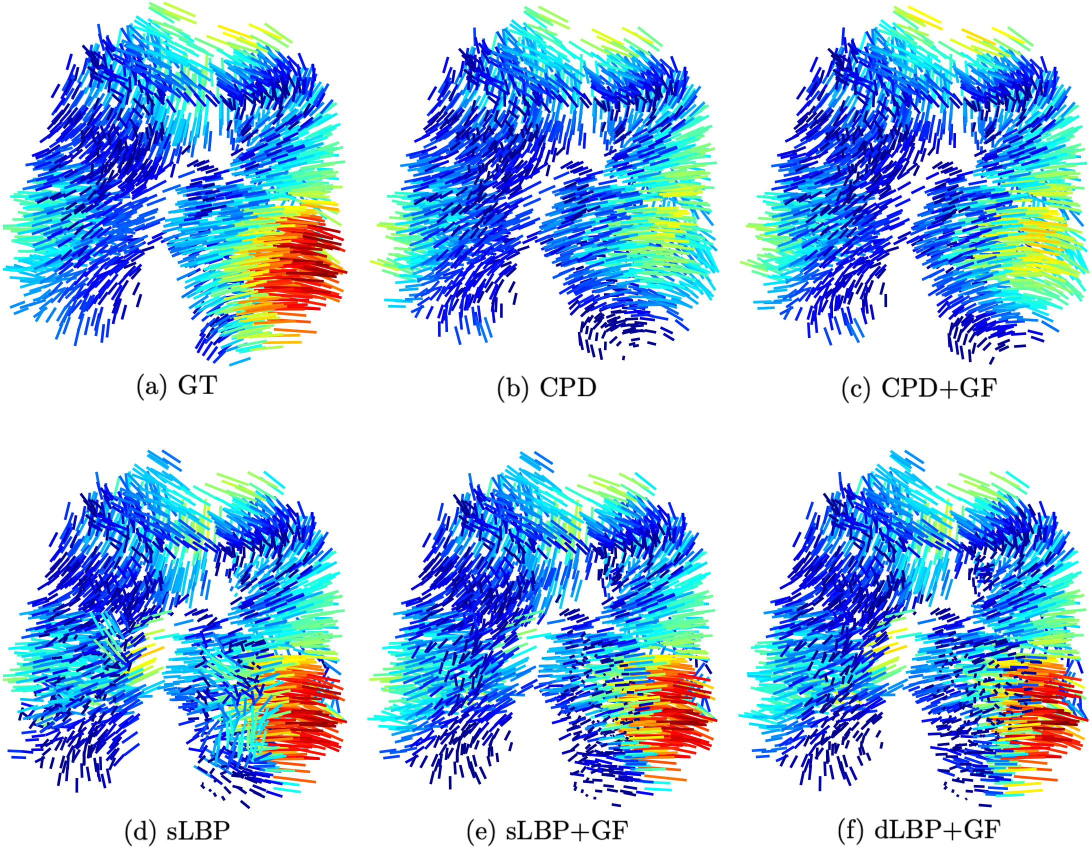
Case \#	initial	CPD	CPD+GF	sLBP	sLBP+GF (ours)	dLBP+GF (ours)
1	26.33	3.02	2.75	2.55	**1.88**	2.14
2	21.79	10.83	**5.96**	8.69	6.22	6.69
3	12.64	1.94	1.88	1.56	**1.53**	1.68
4	29.58	2.89	2.84	3.57	**2.63**	3.01
5	30.08	3.01	2.70	3.01	**2.02**	2.42
6	28.46	3.22	3.65	2.85	**2.21**	2.69
7	21.60	2.52	2.44	1.87	**1.64**	1.83
8	26.46	3.85	3.58	2.08	**1.93**	2.14
9	14.86	2.83	2.58	1.53	**1.55**	1.82
10	21.81	3.57	5.57	3.15	**2.79**	3.72
mean	23.36	3.77	3.40	3.08	**2.44**	2.81
std	11.86	2.54	1.35	2.09	1.40	1.50
time		7.63	7.66	2.91	3.05	**0.49**

coordinates (x,y,z) as input features. In addition, we evaluate the novel discretisation of sparse candidates that is again integrated into an end-to-end geometric learning with differentiable LBP regularisation (dLBP+GF) and leads to substantial efficiency gains. The results clearly demonstrate the great potential of keypoint based registration for the complex task of large deformable lung registration. Numerical and qualitative results are shown in Table 5.3 and Figure 5.8, respectively. Even the baseline methods using no features at all, CPD and sLBP, where inference is based only on optimisation on the extracted keypoint graphs, achieve convincing target registration errors of 3.77 *mm* and 3.08 *mm*. Adding learned geometric features within our proposed geometric registration framework leads to relative improvements of 10% (CPD+GF) and 20% (sLBP+GF), respectively. For the efficient approximation of our proposed appraoch (dLBP+GF) the TRE increases by approximate 0.35 *mm* but at the same time the average runtime is improved six fold to just below 0.5 seconds (which is competitive with dense visual deep learning methods such as LapIRN (see Table 5.4)). A statistical test (Wilcoxon signed-rank test calculated over all landmark pairs) with respect to our proposed method (sLBP+GF) shows that improvements on all other comparison methods are highly significant ($p < 0.001$). We made great efforts to use state-of-the-art

Table 5.4: Results of methods based on visual features on the COPDgene dataset [Castillo et al., 2013]. We report the average TRE and standard deviation in millimeters over all landmarks. The average GPU runtime in seconds is listed in the last column. For easier comparison we also add the results of our "geometry only" approaches.

	mean	std	time
Initial	23.36	11.86	
FLOT+MIND	5.87	1.30	1.63
LapIRN	4.99	1.98	1.08
FE+MIND	3.83	1.21	16.71
sPDD+MIND	3.16	0.69	2.17
CPD+MIND	2.40	0.81	13.12
sLBP+MIND (ours)	**1.74**	0.38	4.65
sLBP+GF (ours)	2.44	1.40	3.05
dLBP+GF (ours)	2.81	1.50	**0.49**

learning-based 3D scene flow registration methods and obtained only meaningful results when incorporating the visual MIND features for FLOT [Puy et al., 2020] and heavily adapting the FlowNet3d embedding strategy [Liu et al., 2019c] (denoted as FE+MIND). FlowNet3d aims to learn a flow embeddings (FE) using a concatenation of two candidate sets (from connected graph nodes), which does not lead to satisfactory results due to the permutation invariant nature of these sparse candidates. Hence, we designed a layer that captures all pairwise combinations and leads to a higher dimensional intermediate tensor that is fed into 1×1 convolutions and is projected (with max-pooling) to a meaningful message vector. For FLOT, we replaced the feature extraction with the handcrafted MIND-PCA embeddings and also removed the refinement convolutions after the optimal transport block (we observed severe overfitting in our training setting when employing the refinement). The sPDD method is based on the probabilistic dense displacement (PDD) network and was modified to operate on the sparse fixed keypoints (instead of a regular grid as in the original published work [Heinrich, 2019]). Results for the state-of-the-art learning based 3D scene flow registration methods and further comparison experiments using visual input features can be found in Table 5.4. Our proposed sparse registration approach using visual MIND features (sLBP+MIND) achieves a TRE well below 2 *mm* and thus, improves on the geometry based equivalent (sLBP+GF) by 0.7 *mm*. However, the extraction of visual features slows down the inference time by 1.6 and 4.1 (dLBP+GF) seconds, respectively. Notably, all proposed geometric registration methods achieve results on par with or significantly better (e.g. more than 50% gain in target registration error w.r.t the dense multi-scale network LapIRN) than the deep learning based comparison methods with additional visual


(a) GT

(b) CPD

(c) CPD+GF

(d) sLBP

(e) sLBP+GF

(f) dLBP+GF

Fig. 5.8: Qualitative results of different geometric methods ((b)-(f)) on case #1 of the COPDgene dataset [Castillo et al., 2013]. The ground truth motion vector field is shown in (a). Different colors encode small (blue) and large motion (red).

features. Conventional registration methods achieve TREs around 1 to 1.5 *mm* with runtimes of 3 to 30 minutes [Avants et al., 2008; Heinrich et al., 2015; Rühaak et al., 2017].

5.3.4 Discussion and Conclusion

We believe our concept clearly demonstrates the advantages of decoupling feature extraction and optimisation by combining parallelisable differentiable message passing for sparse correspondence finding with graph convolutions for geometric feature learning. Our method enables effective graph-based pairwise regularisation and compact networks for robustly capturing geometric context for large deformation estimation. It is much more capable for 3D medical image registration as adaptations of scene flow approaches, which indicates that these methods may be primarily suited for aligning objects with repetitive semantic object/shape parts that are well represented in large training databases.

We demonstrated that even without using visual features, the proposed geometric registration substantially outperforms very recent deep convolutional registration networks that excelled in other medical tasks. The reason for this large performance gap can firstly lie in the complexity of aligning locally ambiguous structures (vessels, airways) that undergo large deformations and that focusing on relatively few relevant 3D keypoints is a decisive factor in learning meaningful geometric transformations. Our new idea to discretise the sparse candidate displacements into a dense embedding using differentiable extrapolation yields immensive computational gains by reducing the number of message computations (from $k = 9$ to 1 per node) and thereby also enabling future use within alternative regularisation algorithms.

While our experimental analysis was so far restricted to lung anatomies, we strongly believe that graph-based regularisation models combined with geometric learning will play an important role for tackling other large motion estimation tasks, the alignment of anatomies across subjects for studying shape variations and tracking in image-guided interventions. Being able to work independently of visual features opens new possibilities for multimodal registration, where our method only requires comparable keypoints to be found, e.g. using probabilistic edge maps [Oktay et al., 2015]. In addition, the avoidance of highly parameterised CNNs can establish new concepts to gain a better interpretability of deep learning models.

Chapter 6

Learning Sparse Graphical Optimisation

The fourth, concluding methodological chapter deals with the question of the extent to which not only feature extraction, as in the previous chapter, but also sparse optimisation can be learned in a data-driven manner. A combination of a local CNN and global GNN is trained to solve a sparse 3D MRF on a keypoint graph extracted from lung CT scans in an unsupervised manner. In addition to the novel deep-learning architecture, the form of supervision, a dense spatial transformer integrated into the end-to-end training for the sparse key points, is of methodological interest. The framework is comprehensively described and evaluated in Section 6.1 as published in [Hansen et al., 2021b].

6.1 Deep Graph Regularisation Networks on Sparse Keypoints

6.1.1 Introduction

The automated analysis of multiple thoracic CT scans plays an important role for diagnosis and treatment planning of pulmonary diseases including, lung cancer [Flampouri et al., 2006], COPD [Galbán et al., 2012], emphysema or pneumonia [Pan et al., 2020]. 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 [Beek et al., 2004]). Accurate deformable intra-patient registration between different respiratory levels is vital for localised ventilation measurements [Reinhardt et al., 2008]. 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 [Mok et al., 2020] showed great success at a multi-task registration challenge [Hering et al., 2022] (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.

6.1.1.1 Related Work

Discrete graphical optimisation models or Markov Random Fields (MRF), that include graph cuts [Glocker et al., 2008], message passing [Felzenszwalb et al., 2006; Heinrich et al., 2013a] 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. [Heinrich et al., 2015]). Hence, besides requiring additional fine-tuning stages for subvoxel accuracy [Rühaak et al., 2017] 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) [Kamnitsas et al., 2017; Zheng et al., 2015], which used approximate mean-field inference of conditional random fields (CRF). In [Knobelreiter et al., 2017] an architecture with unary and pairwise convolutional neural networks (CNNs) was proposed for two-view stereo estimation. Our own prior work [Heinrich et al., 2020; Heinrich, 2019], 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 [Girshick et al., 2015], 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 [Johnson et al., 2016] 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 [Fu et al., 2020; Hering et al., 2019] that have a large number of trainable parameters, but still fall short of the accuracy of efficient conventional registration techniques [König et al., 2018]. Recurrent networks in combination with parametrised transformation models are investigated in [Sandkühler et al., 2019]. 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. [Ilg et al., 2017; Sun et al., 2018a]. However, as discussed in [Heinrich, 2019] 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 [Bronstein et al., 2017]. Most relevant in our context are the PointNet [Qi et al., 2017a], which ignores the local connectivity of point clouds for simplicity, the diffusion based graph convolution network that is restricted to isotropic graph filters [Duvenaud et al., 2015] and newer edge weighted graph networks [Wang et al., 2019a]. Graph attention networks introduced in [Velickovic et al., 2018] are another related and promising research direction based on the attention mechanism that was popularised in machine translation and medical image analysis [Schlemper et al., 2019]. 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 [Huang et al., 2018; Tchapmi et al., 2017].

6.1.1.2 Contributions

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 [Hansen et al., 2020]. 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 regularised displacement vectors on a sparsely sampled irregular grid.

2. In contrast to our preliminary work in [Hansen et al., 2020], 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 [Castillo et al., 2009] and COPDgene Castillo et al., 2013 datasets to average TREs of 1.39 mm and 1.34 mm, respectively.

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 6.1.2. 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 [Heinrich et al., 2013b]). 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 6.1.3 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 6.1.4.

6.1.2 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 6.1 shows an overview of our proposed deep learning registration framework. Individual steps and details of our method are described in the next sections.

6.1.2.1 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 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.

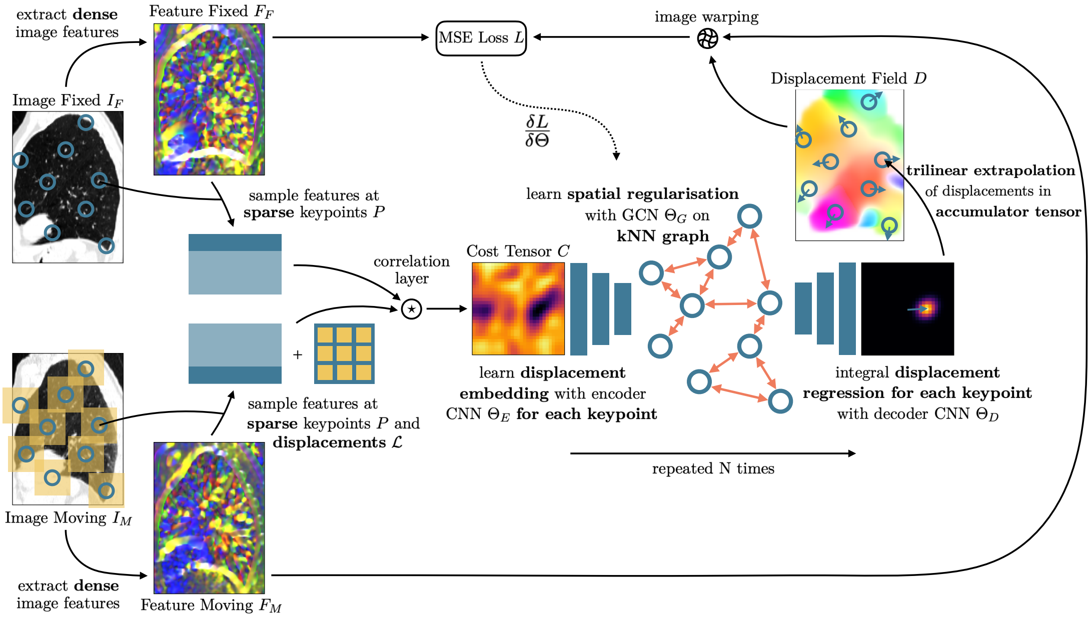

extract **dense** image features
Feature Fixed F_F
MSE Loss L
image warping
Image Fixed I_F
Displacement Field D
$\frac{\delta L}{\delta \Theta}$
sample features at **sparse** keypoints P
learn **spatial regularisation** with GCN Θ_G on **kNN graph**
trilinear extrapolation of displacements in **accumulator tensor**
correlation layer
Cost Tensor C
+
learn **displacement embedding** with encoder CNN Θ_E **for each keypoint**
integral **displacement regression for each keypoint** with decoder CNN Θ_D
sample features at **sparse** keypoints P and **displacements** $\mathcal{L}$
repeated N times
Image Moving I_M
extract **dense** image features
Feature Moving F_M

Fig. 6.1: Overview of our novel deep learning framework for keypoint-based dense deformable image registration. Feature maps F_F and F_M (here: MIND features [Heinrich et al., 2013b]) are extracted from both, the fixed (I_F) and moving (I_M) image (Section 6.1.2.3). Additionally, a set of sparse keypoints P is identified at distinctive locations in the fixed image using the Foerstner operator [Förstner et al., 1987] (Section 6.1.2.2). Correlating the sampled MIND features at the keypoints in the fixed image and dense displaced locations $\mathcal{L}$ in the moving image, yields a cost tensor C for each keypoint (Section 6.1.2.4). We then predict displacement vectors with our proposed GraphRegNet θ, that consists of three neural network modules. First, an encoder CNN θ_E learns a low-dimensional displacement embedding for each cost tensor, then we employ a GCN θ_G that distributes the learned embeddings across the kNN graph of the keypoints to achieve spatial regularisation (Section 6.1.2.5). Final displacement vectors are obtained via integration over the predefined displacement space $\mathcal{L}$, weighted by probabilities of the predicted softmax map of the decoder CNN θ_D. All displacment vectors of the sparse keypoints are accumulated in the displacement field tensor D using trilinear extrapolation (+ spatial smoothing), which makes this densification operation fully differentiable and enables the use of an MSE loss L on the fixed and warped moving MIND image, guiding the training process in an unsupervised fashion (Section 6.1.2.6).

6.1.2.2 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 [Heinrich et al., 2015; Polzin et al., 2013], employing the Foerstner interest operator [Förstner et al., 1987], which is run-time efficient and led to state-of-the-art results in lung registration for conventional methods [Rühaak et al., 2017]. Alternatively, a keypoint graph can be constructed from vessel segmentations (cf. [Fu et al., 2020]). 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

$$S(I_F) = \frac{1}{\mathrm{Tr}((G_{\sigma_1} * (\nabla I_F \nabla I_F^T))^{-1})}. \tag{6.1}$$

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.

6.1.2.3 Image Feature Extraction

Since in the context of this work we are strongly focusing on the prediction and regularisation 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 [Haber et al., 2006]) as a future task. The MIND descriptor was first proposed in [Heinrich et al., 2011] for the task of multi-modal image registration and extended in [Heinrich et al., 2012; Heinrich et al., 2013b]. 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 $F_F, F_M : \mathbb{R}^3 \to \mathbb{R}^{12}$. In addition to the subsequent feature correlation, the extracted MIND descriptors also serve as signals for the unsupervised warping loss (see Section 6.1.2.6).

6.1.2.4 Feature Correlation

Following previous work on discrete 3D medical registration [Heinrich et al., 2013a, 2015] and the successful methods for 2D optical flow estimation [Ilg et al., 2017; Sun et al., 2018a] we perform a similarity search for all fixed features $F_F(p)$ sampled at keypoints $p \in P$ and moving features $F_M(p+l)$ sampled at potential displacement locations $l = (l_x, l_y, l_z) \in \mathcal{L} = q \ \cdot \ \{-l_{max}, \dots, -1, 0, 1, \dots, l_{max}\}^3$. 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

$$C(p,l) = \frac{1}{12} \sum_{i=0}^{11} (F_F^i(p) - F_M^i(p+l))^2, \tag{6.2}$$

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. This means that for the evaluation of $F_M(p+l)$ all displacements l for each keypoint p have to be considered, which greatly reduces the risk of missing correspondences (in contrast to sparse point cloud registration). To account for noisy similarities we smooth the cost tensor C with a Gaussian kernel G_{σ_2} along the displacement dimensions. However, this does not impose an explicit regularisation on the cost tensors and they may still have a large proportion of poor correspondences or multiple local optima (as can be observed in Figure 6.1), which brings with it the need of robust global regularisation as proposed and described in the following section.

6.1.2.5 GraphRegNet

We now aim to predict a sparse displacement field D_S that assigns a displacement vector $d = (d_x, d_y, d_z)$ to each keypoint $p \in P$. 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 [Wang et al., 2019a]) in a DenseNet [Huang et al., 2017] 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

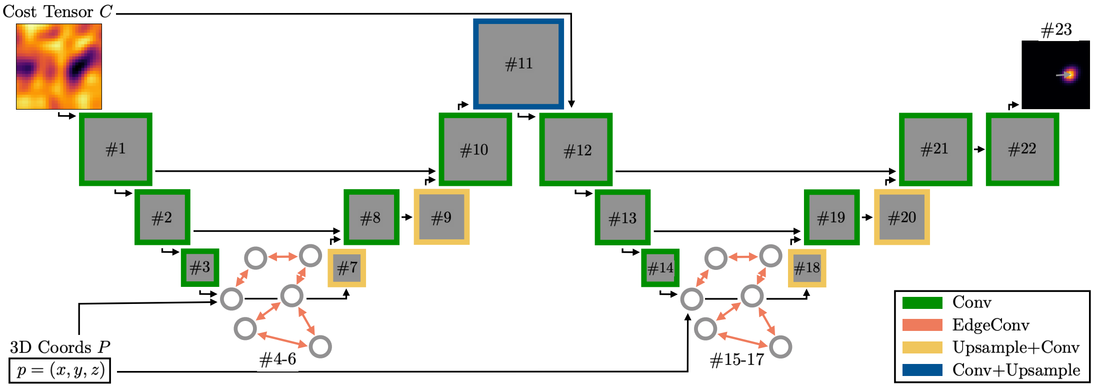
$$f_i' = ReLU(\operatorname*{avg}_{(i,j)\in E} e_{ij}) \tag{6.3}$$

following notations in [Wang et al., 2019a] and describes the non-linear transformation of a feature vector f_i at point $p_i \in P$. The edge features $e_{ij} = h_\theta(f_i, f_j - f_i)$ 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 $\theta = (\theta_1, \ldots, \theta_{|f_i'|})$ with the concatenated keypoint (f_i) and local neigbourhood features $(f_j - f_i)$. The decoder CNN θ_D predicts the displacement vectors d using an integral regression approach [Sun et al., 2018b], 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_p for each keypoint. The final displacement vector d can now be determined by integration over the displacement search region $\mathcal{L}$ weighted by the (softmax) normalized predictions $\tilde{H}_p$ as

$$d = \sum_{l\in\mathcal{L}} l \cdot \tilde{H}_p(l). \tag{6.4}$$

Training with direct regression of displacement vectors was also tested but could not converge. To stabilise 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 normalisation layer and a Leaky ReLU.

The detailed network architecture of the GraphRegNet with a total number of ∼33.000 trainable parameters is summarised in Figure 6.2. Subindices E1, G1, D1, etc. represent the corresponding layer of the encoder, 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.



			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

Fig. 6.2: Block diagram of the GraphRegNet architecture used in our experiments. Detailed information on the individual layers can be found in the corresponding Table.

The integral displacement regression layer is abbreviated as *IntReg* and represents the transition from a discrete to a continuous displacement space.

To summarise, 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. Therefore, 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.

6.1.2.6 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 $d \in D_S$ are accumulated in a dense, low resolution tensor (initialized with zeroes) at respective keypoints 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 6.1 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 regularisation on the sparse keypoints. An explicit regularisation 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.

6.1.3 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 [Castillo et al., 2009] and COPDgene [Castillo et al., 2013] 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

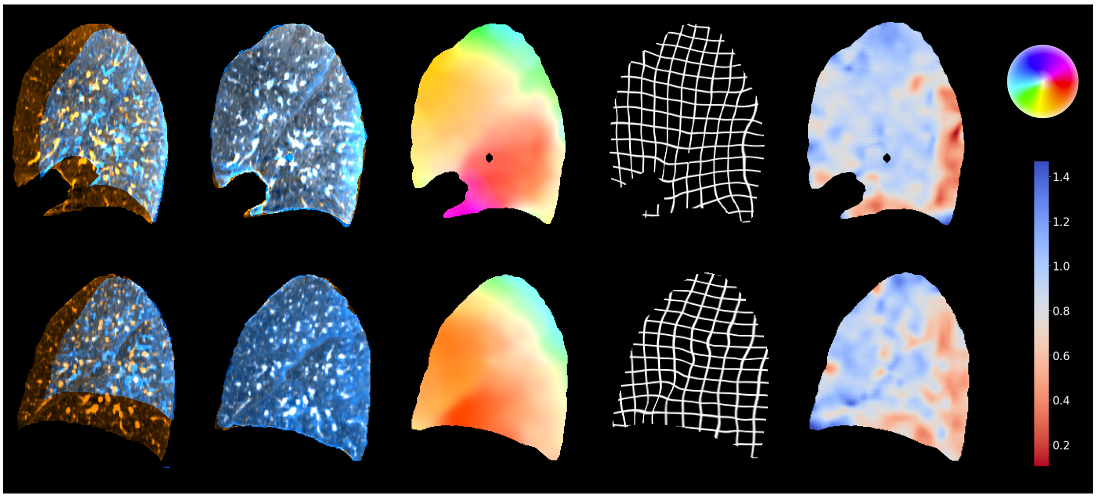


Fig. 6.3: Qualitative results of our proposed GraphRegNet registration framework. Each row visualizes registration results for one scan pair from the COPDgene dataset in saggital view. The first and second column show color overlays (orange: inhale scan, blue: exhale scan) for the initial and final alignment, 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.

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 [Murphy et al., 2011], POPI [Vandemeulebroucke et al., 2007]) and all experiments were carried out using a 5-fold cross validation. Figure 6.3 shows qualitative results for two scan pairs from the COPDgene dataset.

6.1.3.1 Implementation Details

We implemented our proposed registration framework using the deep learning library PyTorch [Paszke et al., 2019] running on a NVIDIA Titan RTX. During preprocessing all scan pairs are resampled to a fixed volume size of D×H×W = 192×160×192 and clipped at HU values of -1000 and 1500. All hyperparameters 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 regularisation 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 regularisation 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 approximately 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).

6.1.3.2 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 (`https://github.com/voxelmorph/voxelmorph`) of the widely used Voxelmorph registration framework [Balakrishnan et al., 2019] with few extensions and evaluate it on both DIR-Lab datasets.

DLIR [Vos et al., 2019] 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.

Ep18 [Eppenhof et al., 2018] 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.

OSL [Fechter et al., 2020] 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 could neither reach the registration accuracy of conventional approaches nor could benefit from the fast runtimes deep learning networks offer in inference.

LRN [Fu et al., 2020] 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.

mlVN [Hering et al., 2019] Hering et al. proposed a deep learning based multi-level variational image registration network (mlVIRNET). It uses three resolution stages with progressively trained CNNs (initialised 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 regularisation and additional manual lobe segmentations.

BMRF [Blendowski et al., 2019] 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.

VM+ [Balakrishnan et al., 2019] 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.

LIRN [Mok et al., 2020] LIRN is a multi-resolution pyramid registration network. In contrast to [Hering et al., 2019] the CNNs at different levels are trained end-to-end

Table 6.1: Registration results on the DIR-Lab 4D CT [Castillo et al., 2009] and COPDgene [Castillo et al., 2013] 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	Ep18	OSL	LRN	mlVN	BMRF	VM+	LIRN	FE+	PDD+	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	***							***	***	***	***	***	***	***	***	***	***	

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 [Hering et al., 2022] and its implementation is publicly accessible (`https://github.com/cwmok/LapIRN`).

FE+ [Liu et al., 2019c] The FlowNet3D [Liu et al., 2019c] 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.

PDD+ [Heinrich, 2019] Own previous work of Heinrich, the PDD net, uses approximate minconvolutions and mean field inference to regularise 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 (`https://github.com/multimodallearning/pdd_net`) to operate on the irregular keypoint graph, use the same dense warping loss and employ a second (refinement) stage.

MST A non-deep learning method that is a close GPU reimplemenation of the CorrField method [Heinrich et al., 2015] with a focus on fast computation time. While the computation of the final displacement field took more than a minute using the C++ implementation (`http://www.mpheinrich.de/software.html`), 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 regularise the displacement costs.

6.1.3.3 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.

RW In this ablation experiment we replace the GraphRegNet with a non-deep learning approach, a random walk (RW) [Aldous, 1989] 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.

NoReg A baseline that highlights the problem of large proportion of poor initial correspondences from the feature similarity computation and thus, the importance of explicit spatial regularisation. In this experiment the GCN θ_G is removed from the architecture. Additionally, we do not add keypoint coordinates to the displacement embeddings. The network is only able to learn on the local displacement dimensions and smooth across the displacements of the cost tensor.

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 Qi et al., 2017a architecture.

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

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.

6.1.3.4 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 6.1 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 *BMRF* and our own experiments for *VM+*, *LapIRN*, *FE+*, *PDD+* and *MST* that

Table 6.2: 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 on GPU and/or CPU. For a detailed run-time analysis of our method we refer to Section 6.1.3.1.

	TRE	avg. computation time (GPU/CPU)
DIS-CO [Rühaak et al., 2017]	0.82 ± 0.97	- / 5 minutes
ANTs [Song et al., 2010]	1.79 ± 2.10	- / 3 hours
Elastix [Klein et al., 2009]	1.32 ± 1.24	- / 14 minutes
NiftyReg [Modat et al., 2010]	2.19 ± 2.00	- / 9 minutes
ours	1.34 ± 1.44	2 seconds / 30 seconds

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 6.4 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.

6.1.3.5 Jacobian Determinant

For the assessment of realistic and well regularised 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 6.3.

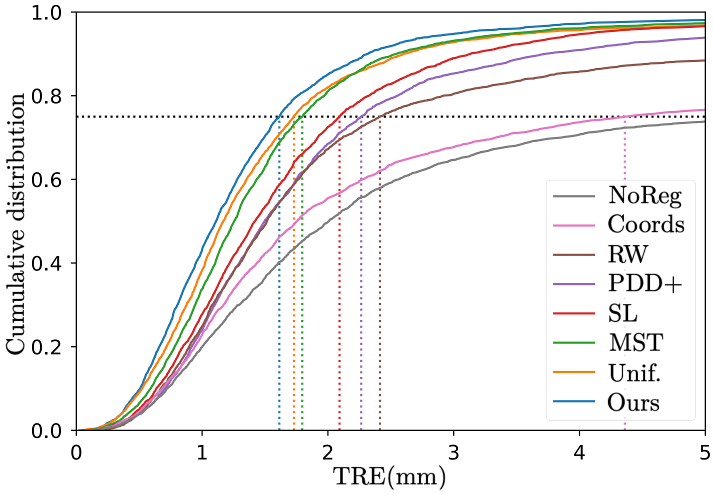


Fig. 6.4: Cumulative distribution of target registration errors in millimeters for all keypoint based methods on all landmark pairs of the COPDgene [Castillo et al., 2013] dataset. In addition, the dotted lines visualize the 75th percentiles of the TRE, which are 1.61 mm (Ours), 1.73 mm (Uniform), 1.80 mm (MST), 2.09 mm (SL), 2.27 mm (PDD+), 2.43 mm (RW), 4.38 mm (Coords) and 5.25 mm (NoReg).

6.1.4 Discussion and Conclusion

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 [Castillo et al., 2009] dataset. It surpasses the registration accuracy of the best performing method, the LungRegNet [Fu et al., 2020] 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 [Castillo et al., 2013] 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. Rühaak et al., 2017 (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 [Rühaak et al., 2017], which reports a computation time of 5 minutes for a single registration. Well established registration software such as *ANTs* [Song et al., 2010], *Elastix* [Klein et al., 2009] or *NiftyReg* [Modat et al., 2010] perform worse or on par with our approach (1.79 mm (boosted [Muenzing et al., 2014]), 1.32 mm and 2.19 mm (boosted Muenzing et al., 2014), respectively) while taking from 9 minutes to 3 hours for a single scan pair (values taken from [Eppenhof et al., 2018; Muenzing et al., 2014]), 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[Ilg et al., 2017] or PWC-Net[Sun et al., 2018a]) and uses graph convolutions as a data-driven, trainable regulariser (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 optimisation.

In this work, we have presented a novel deep network architecture for learning well-regularised 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.

Chapter 7

Summary

In the previous four chapters, various novel methods for deep learning based medical image analysis tasks have been developed and presented with a common focus on the processing and exploitation of sparse representations (point clouds and keypoint graphs). The proposed methods cover a wide range of different input modalities, clinical tasks and target anatomies, and thus, each contributes to answering the exploratory research question regarding the benefits of point clouds and graph representations for medical image analysis tasks formulated in detail at the outset of this work.

In this concluding part, the main contributions of the approaches presented in the methodological chapters are first summarised (Section 7.1) and then put in the context of the research question (Section 7.2). Ongoing research in the field of graph learning in medical imaging is outlined, as well as promising future research directions that can be derived from the approaches presented in this work (Section 7.3).

7.1 Contributions

Exploitation of Pretrained CNNs for Processing Point Clouds and Implicit Learning of Relations in Graphical Models With the emergence of affordable time-of-flight cameras, point clouds have entered commercial and scientific applications of scene understanding and can be considered as a novel type of image modality for clinical analysis tasks. Chapter 3 investigated the task of clinical context monitoring using point clouds from a multi-camera setup in the operating room. For this purpose, the pose of the surgeons (joint landmarks such as head, shoulder, wrists etc.) should be detected reliably and accurately. Here, the point clouds were considered as 2D projections (depth images) and dense voxelisations, which enabled their processing with conventional CNNs. An initial pose estimation on depth images led to significantly better detection results than direct predictions from 3D representations, since the used 2D CNNs could be pretrained on large image data sets (a common approach in computer vision), which is generally infeasible for point clouds due to the lack of available data. By sparse fusion of the 2D detection results via the point cloud representations from multiple cameras, this advantage could also be exploited for 3D pose estimation. While this method has already yielded state-of-the art results, there have been occasional predictions of generally implausible poses (e.g. partial confusion of left and right body parts). To mitigate this behaviour, a self-supervised denoising Convolutional Autoencoder was employed in a post-processing step that was trained to recover poses altered by artificial semantic noise (swapping, removing or displacing landmarks). The experiments confirmed the anticipated regularising effect of the implicit pose graph embedding that restricts the output space to anatomical plausible predictions.

In summary, this chapter has demonstrated that for certain clinical applications grid-based representations (projections and voxelisations) are a suitable approach in handling sparse point clouds, in particular because they enable to employ pre-trained CNNs and to implicitly learn relations in graphical models with ease.

Demonstrating Feasibility and Efficiency of Graph Neural Networks for Segmentation Learning on Point Clouds and Dense Medical Scans Transitioning from dense processing of sparse point clouds to direct exploitation of the graph structure, two exemplary applications for semantic segmentation were investigated in Chapter 4. First, sticking to point clouds from time-of-flight cameras and the human pose, a supervised graph-based learning method for segmenting body parts was developed, and second, an approach examining how sparse representation learning can be applied in the conventional field of medical image segmentation on dense CT or X-Ray scans was presented. In the first part, it was shown that a repeated interleaving of isotropic diffusion operations (with multiple fixed diffusion coefficients to account for different local/global context) and learnable 1×1 convolutions suffices to extract semantic labels from raw point clouds with only point coordinates as input features, highlighting the rich inherent information of graph representations. A limitation of the approach is that the diffusion process is static and robust hyperparameters (diffusion coefficients) need to be determined, which was addressed in [Wang et al., 2019a] by replacing the diffusion operation with learnable edge convolutions (adopted in own subsequently developed methods). When applying the concept of sparse graph-based segmentation to dense medical scans, the first task is to create a meaningful anatomical graph from the image data. One option that has been shown in the experiments to be a robust approach across different datasets is to use a very shallow and light-weight CNN that estimates a coarse probability map for the target anatomy from which keypoints can then be sampled. A novel combination of an Encoder CNN (for patch based feature extraction at keypoint locations), a GNN (for global geometric feature propagation) and a Decoder CNN (for spatial decoding of segmentation labels) enables efficient end-to-end learning at high resolutions. In all experiments conducted, this method provided better or on par segmentation results than a UNet with a much lower complexity (far fewer trainable parameters and thus reduced risk of overfitting).

Overall, the methods presented in Chapter 4 have made an important contribution to demonstrating the feasibility of graph-based learning methods for clinical imaging tasks (here: semantic segmentation on both point clouds and dense medical scans).

Sparse Graphical Optimisation with Decoupled Geometric Feature Extraction Enables Accurate and Runtime Efficient Keypoint Registration While the previous chapters described end-to-end learning methods, Chapter 5 focused on the question of how the advantages of decoupled deep learning feature extraction and conventional keypoint-based graphical optimisation can be combined. A direct methodological advantage is that the approaches presented in this chapter, which strictly decouple feature extraction and optimisation, are no black-box algorithms (unlike end-to-end learning methods) and may thus achieve greater clinical acceptance. Furthermore, applied to the task of medical image registration, it has been shown that both the efficiency and runtime of deep learning methods as well as the accuracy of conventional methods have been matched and in many cases exceeded, in particular for images with large displacements such as those encountered in inter- and intra-patient registration for thoracic and abdominal scans. In the first part, a registration algorithm based on sparse graphical message passing was extended with an iterative local adaptation of the solution space, resulting in very fast runtimes of less than a second on GPU and less than 15 seconds on CPU (comparable to the fastest end-to-end deep learning frameworks). Using handcrafted image features as metric, this approach has already achieved state-of-the-art target registration errors. The experiments further showed, that the method is easily adaptable to a supervised setting (availability of anatomical segmentation labels for a training subset) with greatly improved results by training a CNN segmentation model and using the predictions as additional input features. This strategy differs from popular deep learning registration approaches, which use anatomical labels only as supervision during training (direct vs. indirect loss), and, as demonstrated, offers the flexibility to employ any optimisation method with minimal runtime overhead. The second part of the chapter examined methods for exhale to inhale lung CT registration based solely on sparse keypoints extracted from the CT scans with an interest point detector (corresponding for the most part to pulmonary bifurcations and airway trees). The Coherent Point Drift algorithm and Loopy Belief Propagation were chosen as conventional optimisation baselines, for both of which it could be shown that additional learned keypoint features with GNNs significantly boosted the final alignment. Using only the learned features from the keypoint graphs, both methods could already outperform dense state-of-the-art deep learning registration frameworks. This shows convincingly again how much information is embedded in the geometries of the anatomical graphs alone and challenges design choices of popular deep learning registration architectures. [Hering et al., 2021] also suggested that the incorporation of a keypoint loss helps to mitigate the problem, however, only cases with shallow breathing were considered.

All methods described in this Chapter 5 exploit graphical displacement optimisation over sparse rather than dense locations, which makes them runtime and memory efficient and thus enable a comprehensive exploration of the displacement search space, which has been shown to can be further constrained by decoupled feature extraction using CNNs and GNNs.

Combination of CNNs and GNNs to solve Sparse 3D MRFs for Medical Image Registration Chapter 6 complements Chapter 5 in that it replaced the conventional graphical optimisation step by a novel deep learning architecture, that learns to solve a 3D MRF on a keypoint graph for estimating well-regularised dense displacement fields. CNN and GNN layers, that act on the displacement and spatial dimensions, respectively, are combined in a single end-to-end network with similarity cost tensors evaluated at sparse keypoint location as input and predicted displacement vectors as output. Extrapolation of the sparse displacements to a dense field enables an image-based similarity metric to be employed for training the network. Optimisation over sparse keypoints rather than a regular grid was on the one hand necessary to deal with the large memory footprint (in particular during training), but the experiments also found significant improvements in registration accuracy.

The approach presented sets the current state-of-the-art for breath-hold exhale to inhale lung CT registration for deep learning methods and reduces the gap to the most accurate conventional methods to less than 0.5 mm with orders of magnitude faster runtimes. Besides the progress in the application of learning-based registration, methods derived in this chapter are also an important contribution toward the development of a general deep learning network architecture for solving 3D MRFs on irregular graphs.

7.2 Research Findings

In this section, the contributions of this thesis summarised before will be put in the context of the research question of *which, how, and to what extent medical image analysis tasks can benefit from point cloud and graph representations in combination with deep learning methods* and the associated objectives formulated in Chapter 1. To ensure a presentation coherent with the introduction, the discussion is again guided by the following three aspects:

Which... The methods presented in this thesis have broadly covered the tasks of landmark detection, semantic segmentation and image registration, three of the most relevant topics in medical image processing. The focus for each task was on investigating the processing and integration of point clouds and graph-based approaches.

Point clouds from time-of-flight cameras enable completely novel clinical applications, as demonstrated in Chapter 3 with a framework for 3D human pose estimation in the operating room, enabling clinical context-aware assistance and monitoring systems (automatic analysis and improvement of surgical workflows, safe human–robot interaction, etc.).

Similar to landmark estimation in point clouds, full semantic segmentation of the human body as presented in Chapter 4 is feasible and of broad clinical interest for patient monitoring applications, such as estimation of body weight and height or contactless measurement of pulse and temperature from different body parts. For conventional medical imaging input modalities such as CT and X-Ray scans, deep learning for semantic surface segmentation on extracted keypoints graphs also showed convincing results (see Chapter 4). Application areas include time-constrained settings and mobile/embedded systems due to the memory and runtime efficiency reached by the sparseness of the data and processing method. The findings are also of interest for research on radiation dose reduction, as learning on keypoint graphs may make the acquisition of sparsely sampled medical images conceivable.

The medical image registration methods presented in Chapter 5 and Chapter 6 have comprehensively addressed key point-based approaches for different image modalities (CT/MR), inter- and intra-patient registration, and two challenging (potentially highly displaced) anatomical regions of interest, namely the thorax and abdomen. Since the methods not only demonstrate state-of-the-art accuracy, but also very fast runtimes, they can be employed in various application areas, such as aligning preoperative scans for image-guided therapy and multimodal diagnostics, nodule tracking, statistical evaluation of variations of organs for pathology detection or to generate a canonical atlas space.

Although, for obvious reasons, not all areas and applications of medical image analysis have been covered, a comprehensive overview of different tasks that benefit from point cloud processing and the exploitation of keypoint graphs has been provided. Not only have new areas emerged for applications such as context analysis using point clouds, but it has also been possible to incorporate graph learning and optimisation successfully in many conventional medical imaging tasks like segmentation and registration.

How. . . Various novel methods for handling point clouds and exploiting sparse keypoint graphs in medical applications were developed and presented in this thesis.

One viable option is to ignore the sparseness of point clouds and treat the data as dense projections (depth images) or voxelisations, as it has been investigated in Chapter 3. Although this approach certainly has its shortcomings, such as occlusion in projection images or loss of information through discretisation, several advantages have been identified in this particular application of 3D human body pose estimation for scene understanding. Depth images have many structural similarities (edges, proportions, etc.) to natural colour images, which enabled the use of CNNs pre-trained on publicly available very large datasets of natural images for feature extraction, significantly improving prediction accuracy. It has also been demonstrated that 3D CAEs can, through self-supervised training, encode an implicit graph model for pose determination thus constraining the valid output pose space and correct implausible predictions.

The direct processing of point clouds for learning-based tasks has been addressed in the methodological development of multi-kernel diffusion CNNs in Chapter 4, a graph neural network approach that combines diffusion operations on nearest neighbour graphs of different locality with learnable 1×1 convolutions. Although similar in many aspects, edge convolutions developed in parallel in [Wang et al., 2019a], which employ an aggregation mechanism and edgewise convolutions instead of the diffusion operation, offer more flexibility and have been adopted in subsequent methods. Chapter 4 also described how graph learning can be applied to dense medical scans for structured prediction tasks like segmentation. By sampling a distinctive keypoint graph from the image, employing a neural graph network for feature propagation and a dense voting scheme to densify the predictions, an efficient alternative to conventional dense encoder-decoder architectures has been presented that in addition to intensity information can also exploit the inherent structural information of the underlying anatomical graph.

The methods for medical image registration presented in Chapter 5 have in common a clear disentanglement of learning based feature extraction and differentiable sparse graphical optimisation of displacements. For feature extraction, it has been shown that both, the dense processing with UNet architectures (and subsequent evaluation at sparse sampled locations), but also the training of GNNs directly on keypoint graphs are suitable and efficient approaches. However, for sparse graphical optimisation, message passing algorithms have clearly prevailed over a deformable point cloud registration approach (coherent point drift) in the experiments. In particular, since through an adaptive iterative extension of the loopy belief propagation method, the runtime efficiency could be greatly improved.

In Chapter 6, it was demonstrated that optimisation on keypoint graphs in the context of registration can also be learned. Using a combination of CNNs that filter similarity tensors at the local level in the displacement dimension and GNNs that learn a globally optimal solution, an MRF-like registration problem can be solved within an end-to-end training framework.

In summary, following approaches for incorporating point clouds and graphs into medical imaging tasks have been thoroughly investigated: processing dense representations of point clouds with regular CNNs, trainable diffusion networks, dense structured prediction from keypoint graphs, disentangling feature extraction and sparse graphical optimisation, and solving MRFs with a CNN/GNN combination.

To what extent... All the approaches presented were comprehensively compared in experiments with the respective state-of-the-art methods, to not only prove feasibility but also show advances in the field.

The pose estimation framework introduced in Chapter 3 significantly improves on the landmark distance error of previous approaches through the 2D to 3D fusion (from approximately 9 cm to 8 cm), and further prevents the prediction of completely implausible poses through the pose constraints of the CAE. The largest errors are made in the exact localisation of the extremities, especially the wrist joints (average landmark distance to the ground truth of 11 cm). For a final assessment of accuracy, an inter-rater error would be beneficial, which, however, is currently not available for the dataset.

In Chapter 4, it has been shown that a diffusion network with only 1×1 convolutions as learnable function is capable of segmenting body parts of point clouds with F1 scores of 0.95 (slightly worse results when introducing different types of noise). Wrongly segmented predictions were often global inaccuracies such as confusing left and right parts of the body. When applying graph learning to dense segmentation on medical scans (also Chapter 4), better or results on-par with those of UNet models could be achieved. At the same time, the AI models used are less than half the size (500k versus 1300k trainable parameters) and are evaluated at only 2000 keypoint patches, making the method appealing for memory-constrained applications.

The keypoint-based registration methods presented in Chapter 5 have significantly improved the runtime compared to conventional methods (from minutes to subseconds) and have shown that registration accuracy can benefit greatly from decoupled deep learning feature extraction, especially in supervised settings (increase of Dice overlap from 40% to 65% for abdominal CT registration with anatomical segmentations). However, it should be noted that particularly the precision in the respective ROIs improves and the risk of label bias, i.e. overfitting towards certain anatomies, is introduced. Based on the application example of breath-hold lung CT registration, it could also be demonstrated that recent dense deep learning registration frameworks are only marginally suitable for the prediction of large displacements and are easily outperformed by graph learning methods, even when processing just the anatomical graph without any image information (>50% reduced target registration error).

The learning method for graphical optimisation presented in Chapter 6 represents the deep learning based state-of-the-art for this task. It closes the gap in accuracy to conventional methods to about 0.3 mm with significantly shorter runtimes.

For all medical imaging tasks considered, the inclusion of graph-based approaches has either improved accuracy metrics, reduced runtime and memory requirements, or both.

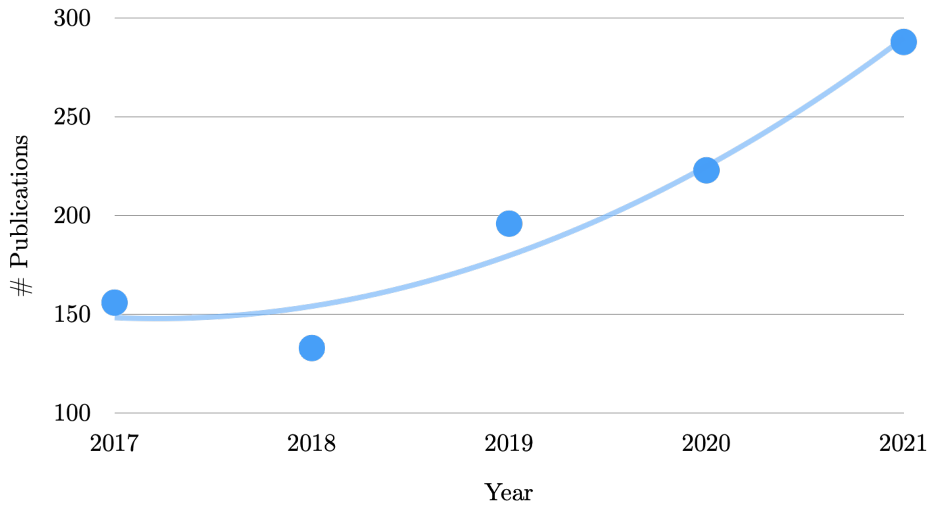


Fig. 7.1: Trend over the last five years in medical image analysis research on the topic of point clouds and keypoint graphs (search query: (point clouds OR keypoints OR graphs)). Publications in the top conferences and journals in the field (MICCAI, IPMI, MIDL, TMI, MedIA) were considered.

7.3 Recent Developments and Outlook

The objective of this work has been to investigate graph-based methods for solving medical image analysis tasks. Figure 7.1 shows the increase in publications of thematically related scientific works at medical imaging conferences and journals over the last five years, which highlights the importance of this topic as an active area of research. To better position the presented work in this context a brief summary of recent developments and applications of graph learning in medical image processing is given below.

Current Related Work Point clouds continue to play a vital role in scene understanding for health-related applications. In addition to environment tracking in operating rooms during surgery, as addressed in this work, ambulant or home-based fall detection systems in elderly care [Gutiérrez et al., 2021] as well as in-bed context monitoring of ICU or IMCU patients [Liu et al., 2019a] are of great clinical interest. Point clouds and derived graphical representations are also increasingly exploited in medical segmentation and reconstruction tasks either to directly infer solid surface models while imposing shape constraints [Gaggion et al., 2021; Kong et al., 2021; Nakao et al., 2021]. They are also used to generally improve segmentation results through graph features [Garcia-Uceda Juarez et al., 2019; Soberanis-Mukul et al., 2020; Tan et al., 2021]. Graph approaches enable learning-based classification/grading of whole-slide histology images, which is only possible with patch processing when using dense CNNs due to the enormous memory demands [Wojciechowska et al., 2021; Zhou et al., 2019]. Great progress has also been made in the field of disease diagnosis on MRI and fMRI scans, as GNNs show promise for a successfully modelling of functional and structural connectivity of the brain [Li et al., 2021; Ma et al., 2020; Yan et al., 2021; Yao et al., 2021]. Furthermore, graph relations can be established not only at the image level but also at the population level, thus leveraging relationships between patients to improve downstream classification tasks [Cosmo et al., 2020; Parisot et al., 2017]. For the task of medical image registration multiple works find, similar to the solutions proposed here, that incorporating descriptive keypoints can substantially reduce the registration error [Evan et al., 2022; Hering et al., 2021; Shen et al., 2021].

Limitations and On-going Work Regarding the own methods presented, despite the overall favourable outcomes, for all approaches certain limitations exist and additional research questions arise, which could be answered in future investigations. Several of them are in fact already studied in our own on-going work (see List of Publications).

The approaches presented in Chapter 3 improved the state of the art in 3D human pose estimation in a clinical setting, but the landmark errors at more difficult joints such as wrists remained relatively large. The VAE used to build a restrictive pose embedding space was able to prevent most but not all estimates of implausible poses and does not guarantee anatomically correct predictions. Further anatomical constraints such as bone lengths, angles or symmetries can be used in objective functions to improve the accuracy of the predicted pose graph [Bigalke et al., 2022a]. If bone lengths are known in advance (for example, from reliable predictions in unobstructed views), the landmark detection could also be replaced by an explicit regularised fitting of a fixed pose graph to the point cloud data. It is also very likely that a temporal analysis instead of the currently employed single-frame approach would lead to better results with fewer outliers. In addition to efforts to improve the accuracy of the method, its application in other areas, in particular for in-bed pose estimation of patients, is of great clinical interest [Bigalke et al., 2021a,b].

The graph-based segmentation network introduced in Chapter 4 yielded partly better partly on-par Dice scores compared to a dense U-Net on different datasets. At the same time, it has a significantly smaller number of trainable parameters and evaluates the scans at only 2000 distinctive keypoints. However, this lower complexity does not directly translate to a faster runtime, due in particular to the random memory access on the GPU for the irregular graph structure [Liu et al., 2019d]. To circumvent this and still benefit from the reduced complexity, it would be conceivable to integrate graph convolutions directly into a dense regular multiscale U-net architecture by replacing the expensive 3x3x3 kernels with rotation-invariant message passing mechanisms [Weihsbach et al., 2022]. A further limitation of the method is the two-step approach, where first a

CNN is trained to predict distinctive regions for sampling keypoints, which are then exploited in the actual segmentation network. An end-to-end approach could further increase segmentation accuracy, as key point sampling and feature learning would then mutually benefit from one another [Evan et al., 2022; Sarlin et al., 2020].

Incorporation of keypoints and graph based optimisation in deep learning image registration led to accurate registration results as shown in Chapter 5. The main focus was on point cloud based registration, so it is of interest to investigate whether dense deep learning networks, such as the widely used Voxelmorph framework, could also benefit from keypoints, e.g. by replacing the commonly used spatial transformers with keypoint supervision [Heinrich et al., 2022]. The keypoint graphs employed in the presented methods are either derived from heuristic interest point detectors such as the Förstner operator or randomly sampled from regions of interest. Extraction of more semantically descriptive anatomical graphs, such as airway trees in CT lung registration or organ surfaces, has the potential of further improving target registration quality [Falta et al., 2022b]. An open research question remains the data-driven end-to-end learning of sampling locations for deformable keypoint based image registration [Evan et al., 2022; Sarlin et al., 2020]. Furthermore, the decoupling of structure and appearance inherent in the graph-based approach, as well as the reduced network complexity, offer potential for research in the fields of domain adaptation and interpretability [Bigalke et al., 2022b].

The trainable graph network to solve 3D MRFs presented in Chapter 6 yielded state-of-the-art results for deep learning based registration of large motion breath-hold lung CT scans. Alternative approaches to address learning-based optimisation are likewise conceivable. In particular, recurrent approaches that more closely resemble conventional gradient descent methods could benefit in terms of accuracy as well as runtime (in particular on keypoints) from data-driven deep learning as a form of population-based instance optimisation [Falta et al., 2022a].

In general, for all learning methods presented in this thesis, the results depend on the quality and quantity of the underlying data sets. In addition to the relatively small public data sets used here it would therefore be desirable to train the models on additional (annotated) images in order to observe convergence with respect to the data. Finally, although all algorithms have been made publicly available (https://github.com/multimodallearning) when possible (restrictions apply for industrial cooperations), it would be an advantage for the future usage of the methods to make them accessible as a modular and self-configuring framework, similar to the efforts for the nnUNet [Isensee et al., 2021]. The objective would be to configure the registration algorithm using both task-agnostic rule-based design/architecture decisions (based on expert knowledge) and automatic hyperparameter optimisation (evaluation of a target metric on validation data) in such a way that user intervention is no longer necessary and optimal results are achieved even on diverse datasets, such as those included in the current Learn2Reg challenge [Hering et al., 2022].

References

[Achilles et al., 2016] Achilles, F., Ichim, A.-E., Coskun, H., Tombari, F., Noachtar, S., and Navab, N. “Patient MoCap: Human Pose Estimation under Blanket Occlusion for Hospital Monitoring Applications”. In: *Medical Image Computing and Computer Assisted Intervention –MICCAI 2016.* 2016, pp. 491–499.

[Adams et al., 1994] Adams, R. and Bischof, L. “Seeded Region Growing”. *IEEE Transactions on pattern analysis and machine intelligence* 16 (6), 1994, pp. 641–647.

[Akin et al., 2016] Akin, O., Elnajjar, P., Heller, M., Jarosz, R., Erickson, B., Kirk, S., and Filippini, J. “Radiology Data from the Cancer Genome Atlas Kidney Renal Clear Cell Carcinoma [TCGA-KIRC] Collection”. *The Cancer Imaging Archive*, 2016.

[Alansary et al., 2019] Alansary, A., Oktay, O., Li, Y., Le Folgoc, L., Hou, B., Vaillant, G., Kamnitsas, K., Vlontzos, A., Glocker, B., Kainz, B., and Rückert, D. “Evaluating Reinforcement Learning Agents for Anatomical Landmark Detection”. *Medical Image Analysis* 53, 2019, pp. 156–164.

[Aldous, 1989] Aldous, D. J. “Lower Bounds for Covering Times for Reversible Markov Chains and Random Walks on Graphs”. *Journal of Theoretical Probability* 2 (1), 1989, pp. 91–100.

[Andriluka et al., 2009] Andriluka, M., Roth, S., and Schiele, B. “Pictorial Structures Revisited: People Detection and Articulated Pose Estimation”. In: *Conference on Computer Vision and Pattern Recognition –CVPR 2009.* 2009, pp. 1014–1021.

[Andriluka et al., 2014] Andriluka, M., Pishchulin, L., Gehler, P., and Schiele, B. “2D Human Pose Estimation: New Benchmark and State of the Art Analysis”. In: *Conference on Computer Vision and Pattern Recognition –CVPR 2014.* 2014, pp. 3686–3693.

[Andriluka et al., 2018] Andriluka, M., Iqbal, U., Insafutdinov, E., Pishchulin, L., Milan, A., Gall, J., and Schiele, B. “Posetrack: A benchmark for human pose estimation and tracking”. In: *Conference on Computer Vision and Pattern Recognition –CVPR 2018.* 2018, pp. 5167–5176.

[Atwood et al., 2016] Atwood, J. and Towsley, D. “Diffusion-Convolutional Neural Networks”. In: *Advances in Neural Information Processing Systems –NeurIPS 2016.* 2016, pp. 1993–2001.

[Aubry et al., 2011] Aubry, M., Schlickewei, U., and Cremers, D. “The Wave Kernel Signature: A Quantum Mechanical Approach to Shape Analysis”. In: *Workshop on Dynamic Shape Capture and Analysis –ICCV 2011 Workshops.* 2011, pp. 1626–1633.

[Avants et al., 2008] Avants, B. B., Epstein, C. L., Grossman, M., and Gee, J. C. “Symmetric Diffeomorphic Image Registration with Cross-Correlation: Evaluating Automated Labeling of Elderly and Neurodegenerative Brain”. *Medical Image Analysis* 12 (1), 2008, pp. 26–41.

[Balakrishnan et al., 2019] Balakrishnan, G., Zhao, A., Sabuncu, M. R., Guttag, J., and Dalca, A. V. “Voxelmorph: A Learning Framework for Deformable Medical Image Registration”. *IEEE Transactions on Medical Imaging* 38 (8), 2019, pp. 1788–1800.

[Ballard, 1981] Ballard, D. H. “Generalizing the Hough Transform to Detect Arbitrary Shapes”. *Pattern Recognition* 13 (2), 1981, pp. 111–122.

[Barrett et al., 1997] Barrett, W. A. and Mortensen, E. N. “Interactive Live-wire Boundary Extraction”. *Medical image analysis* 1 (4), 1997, pp. 331–341.

[Bauer et al., 2013] Bauer, S., Seitel, A., Hofmann, H., Blum, T., Wasza, J., Balda, M., Meinzer, H.-P., Navab, N., Hornegger, J., and Maier-Hein, L. “Real-Time Range Imaging in Health Care: A Survey”. In: *Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications.* 2013, pp. 228–254.

[Bayer et al., 2018] Bayer, S., Ravikumar, N., Strumia, M., Tong, X., Gao, Y., Ostermeier, M., Fahrig, R., and Maier, A. “Intraoperative Brain Shift Compensation using a Hybrid Mixture Model”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention –MICCAI 2018.* 2018, pp. 116–124.

[Beek et al., 2004] Beek, E. J., Wild, J. M., Kauczor, H.-U., Schreiber, W., Mugler III, J. P., and Lange, E. E. “Functional MRI of the Lung using Hyperpolarized 3-Helium Gas”. *Journal of Magnetic Resonance Imaging* 20 (4), 2004, pp. 540–554.

[Belagiannis et al., 2016] Belagiannis, V., Wang, X., Shitrit, H. B. B., Hashimoto, K., Stauder, R., Aoki, Y., Kranzfelder, M., Schneider, A., Fua, P., Ilic, S., Feussner, H., and Navab, N. “Parsing Human Skeletons in an Operating Room”. *Machine Vision and Applications* 27 (7), 2016, pp. 1035–1046.

[Besl et al., 1992] Besl, P. J. and McKay, N. D. “Method for Registration of 3D Shapes”. In: *Sensor Fusion IV: Control Paradigms and Data Structures.* Vol. 1611. 1992, pp. 586–606.

[Bigalke et al., 2021a] Bigalke, A., Hansen, L., Diesel, J., and Heinrich, M. P. “Seeing Under the Cover with a 3D U-Net: Point Cloud-Based Weight Estimation of

Covered Patients". *International Journal of Computer Assisted Radiology and Surgery* 16 (12), 2021, pp. 2079–2087.

[Bigalke et al., 2021b] Bigalke, A., Hansen, L., and Heinrich, M. P. "End-to-end Learning of Body Weight Prediction from Point Clouds with Basis Point Sets". In: *Bildverarbeitung für die Medizin 2021 –BVM 2021*. 2021, pp. 254–259.

[Bigalke et al., 2022a] Bigalke, A., Hansen, L., Diesel, J., and Heinrich, M. P. "Domain Adaptation through Anatomical Constraints for 3D Human Pose Estimation under the Cover". In: *International Conference on Medical Imaging with Deep Learning –MIDL 2022*. 2022, pp. 173–187.

[Bigalke et al., 2022b] Bigalke, A., Hansen, L., and Heinrich, M. P. "Adapting the Mean Teacher for Keypoint-based Lung Registration under Geometric Domain Shifts". In: *Medical Image Computing and Computer Assisted Intervention –MICCAI 2022*. 2022, pp. 280–290.

[Bishop et al., 2006] Bishop, C. M. and Nasrabadi, N. M. *Pattern Recognition and Machine Learning*. Vol. 4. 4. 2006.

[Blendowski et al., 2019] Blendowski, M. and Heinrich, M. P. "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* 14 (1), 2019, pp. 43–52.

[Bogo et al., 2014] Bogo, F., Romero, J., Loper, M., and Black, M. J. "FAUST: Dataset and Evaluation for 3D Mesh Registration". In: *Conference on Computer Vision and Pattern Recognition –CVPR 2014*. 2014, pp. 3794–3801.

[Bookstein, 1989] Bookstein, F. L. "Principal Warps: Thin-Plate Splines and the Decomposition of Deformations". *IEEE Transactions on Pattern Analysis and Machine Intelligence* 11 (6), 1989, pp. 567–585.

[Boscaini et al., 2015] Boscaini, D., Masci, J., Melzi, S., Bronstein, M. M., Castellani, U., and Vandergheynst, P. "Learning Class-Specific Descriptors for Deformable Shapes using Localized Spectral Convolutional Networks". *Computer Graphics Forum* 34 (5), 2015, pp. 13–23.

[Boscaini et al., 2016] Boscaini, D., Masci, J., Rodolà, E., and Bronstein, M. "Learning Shape Correspondence with Anisotropic Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems –NeurIPS 2016*. 2016, pp. 3189–3197.

[Brachmann et al., 2017] Brachmann, E., Krull, A., Nowozin, S., Shotton, J., Michel, F., Gumhold, S., and Rother, C. "DSAC-Differentiable RANSAC for Camera Localization". In: *Conference on Computer Vision and Pattern Recognition –CVPR 2017*. 2017, pp. 6684–6692.

[Bronstein et al., 2017] Bronstein, M. M., Bruna, J., LeCun, Y., Szlam, A., and Vandergheynst, P. “Geometric Deep Learning: Going Beyond Euclidean Data”. *IEEE Signal Processing Magazine* 34 (4), 2017, pp. 18–42.

[Bruna et al., 2014] Bruna, J., Zaremba, W., Szlam, A., and LeCun, Y. “Spectral Networks and Deep Locally Connected Networks on Graphs”. In: *International Conference on Learning Representations –ICLR 2014*. 2014.

[Calonder et al., 2010] Calonder, M., Lepetit, V., Strecha, C., and Fua, P. “BRIEF: Binary Robust Independent Elementary Features”. In: *European Conference on Computer Vision –ECCV 2010*. 2010, pp. 778–792.

[Canny, 1986] Canny, J. “A Computational Approach to Edge Detection”. *Transactions on Pattern Analysis and Machine Intelligence* (6), 1986, pp. 679–698.

[Cao et al., 2017] Cao, Z., Simon, T., Wei, S.-E., and Sheikh, Y. “Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields”. In: *Conference on Computer Vision and Pattern Recognition –CVPR 2017*. 2017, pp. 7291–7299.

[Castillo et al., 2009] Castillo, R., Castillo, E., Guerra, R., Johnson, V. E., McPhail, T., Garg, A. K., and Guerrero, T. “A Framework for Evaluation of Deformable Image Registration Spatial Accuracy using Large Landmark Point Sets”. *Physics in Medicine & Biology* 54 (7), 2009, p. 1849.

[Castillo et al., 2013] Castillo, R., Castillo, E., Fuentes, D., Ahmad, M., Wood, A. M., Ludwig, M. S., and Guerrero, T. “A Reference Dataset for Deformable Image Registration Spatial Accuracy Evaluation using the COPDgene Study Archive”. *Physics in Medicine & Biology* 58 (9), 2013, p. 2861.

[Chan et al., 2001] Chan, T. F. and Vese, L. A. “Active Contours without Edges”. *IEEE Transactions on Image Processing* 10 (2), 2001, pp. 266–277.

[Chen et al., 2018a] Chen, K., Gabriel, P., Alasfour, A., Gong, C., Doyle, W. K., Devinsky, O., Friedman, D., Dugan, P., Melloni, L., Thesen, T., Gonda, D., Sattar, S., Wang, S., and Gilja, V. “Patient-Specific Pose Estimation in Clinical Environments”. *Journal of Translational Engineering in Health and Medicine* 6, 2018, pp. 1–11.

[Chen et al., 2018b] Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G., and Sun, J. “Cascaded Pyramid Network for Multi-Person Pose Estimation”. In: *Conference on Computer Vision and Pattern Recognition –CVPR 2018*. 2018, pp. 7103–7112.

[Chen et al., 2020] Chen, Y., Tian, Y., and He, M. “Monocular human pose estimation: A survey of deep learning-based methods”. *Computer Vision and Image Understanding* 192, 2020, p. 102897.

[Chung et al., 1997] Chung, F. R. and Graham, F. C. *Spectral Graph Theory*. 92. American Mathematical Society, 1997.

[Clark et al., 2013] Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M., et al. “The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository”. *Journal of Digital Imaging* 26 (6), 2013, pp. 1045–1057.

[Cosmo et al., 2020] Cosmo, L., Kazi, A., Ahmadi, S.-A., Navab, N., and Bronstein, M. “Latent-graph Learning for Disease Prediction”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention –MICCAI 2020.* 2020, pp. 643–653.

[De Fauw et al., 2018] De Fauw, J., Ledsam, J. R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., Askham, H., Glorot, X., O’Donoghue, B., Visentin, D., Driessche, G., Lakshminarayanan, B., Meyer, C., Mackinder, F., Bouton, S., Ayoub, K., Chopra, R., King, D., Karthikesalingam, A., Hughes, C. O., Raine, R., Hughes, J., Sim, D. A., Egan, C., Tufail, A., Montgomery, H., Hassabis, D., Rees, G., Back, T., Khaw, P. T., Suleyman, M., Cornebise, J., Keane, P. A., and Ronneberger, O. “Clinically Applicable Deep Learning for Diagnosis and Referral in Retinal Disease”. *Nature Medicine* 24 (9), 2018, pp. 1342–1350.

[Deng et al., 2009] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. “ImageNet: A Large-scale Hierarchical Image Database”. In: *Conference on Computer Vision and Pattern Recognition –CVPR 2009.* 2009, pp. 248–255.

[Deng et al., 2018] Deng, R., Shen, C., Liu, S., Wang, H., and Liu, X. “Learning to Predict Crisp Boundaries”. In: *European Conference on Computer Vision –ECCV 2018.* 2018, pp. 562–578.

[Desbrun et al., 1999] Desbrun, M., Meyer, M., Schröder, P., and Barr, A. H. “Implicit Fairing of Irregular Meshes using Diffusion and Curvature Flow”. In: *Conference on Computer Graphics and Interactive Techniques –SIGGRAPH 1999.* 1999, pp. 317–324.

[Dietz et al., 2016] Dietz, A., Schröder, S., Pösch, A., Frank, K., and Reithmeier, E. “Contactless Surgery Light Control based on 3D Gesture Recognition”. In: *Global Conference on Artificial Intelligence –GCAI 2016.* 2016, pp. 138–146.

[Dollár et al., 2013] Dollár, P. and Zitnick, C. L. “Structured Forests for Fast Edge Detection”. In: *International Conference on Computer Vision and Pattern Recognition –CVPR 2013.* 2013, pp. 1841–1848.

[Duvenaud et al., 2015] Duvenaud, D. K., Maclaurin, D., Iparraguirre, J., Bombarell, R., Hirzel, T., Aspuru-Guzik, A., and Adams, R. P. “Convolutional Networks on Graphs for Learning Molecular Fingerprints”. In: *Advances in Neural Information Processing Systems –NeurIPS 2015.* 2015, pp. 2224–2232.

[Ehrhardt et al., 2010] Ehrhardt, J., Werner, R., Schmidt-Richberg, A., and Handels, H. “Automatic Landmark Detection and Non-Linear Landmark-and Surface-Based Registration of Lung CT Images”. In: 2010, pp. 165–174.

[Eppenhof et al., 2018] Eppenhof, K. A. and Pluim, J. P. “Pulmonary CT Registration through Supervised Learning with Convolutional Neural Networks”. *IEEE Transactions on Medical Imaging* 38 (5), 2018, pp. 1097–1105.

[Erickson et al., 2016] Erickson, B., Kirk, S., Lee, Y., Bathe, O., Kearns, M., Gerdes, C., and Lemmerman, J. “Radiology Data from The Cancer Genome Atlas Liver Hepatocellular Carcinoma [TCGA-LIHC] Collection”. *The Cancer Imaging Archive*, 2016.

[Evan et al., 2022] Evan, M. Y., Wang, A. Q., Dalca, A. V., and Sabuncu, M. R. “KeypointMorph: Robust Multi-modal Affine Registration via Unsupervised Keypoint Detection”. In: *Medical Imaging with Deep Learning –MIDL 2022.* 2022.

[Falta et al., 2022a] Falta, F., Hansen, L., and Heinrich, M. P. “Learning Iterative Optimisation for Deformable Image Registration of Lung CT with Recurrent Convolutional Networks”. In: *Medical Image Computing and Computer Assisted Intervention –MICCAI 2022.* 2022, pp. 301–309.

[Falta et al., 2022b] Falta, F., Hansen, L., Himstedt, M., and Heinrich, M. P. “Learning an Airway Atlas from Lung CT using Semantic Inter-Patient Deformable Registration”. In: *Bildverarbeitung für die Medizin 2022 –BVM 2022.* 2022, pp. 75–80.

[Fang et al., 2017] Fang, H.-S., Xie, S., Tai, Y.-W., and Lu, C. “RMPE: Regional Multi-Person Pose Estimation”. In: *International Conference on Computer Vision –ICCV 2017.* 2017, pp. 2334–2343.

[Fauser et al., 2019] Fauser, J., Stenin, I., Bauer, M., Hsu, W.-H., Kristin, J., Klenzner, T., Schipper, J., and Mukhopadhyay, A. “Toward an Automatic Preoperative Pipeline for Image-guided Temporal Bone Surgery”. *International Journal of Computer Assisted Radiology and Surgery* 14 (6), 2019, pp. 967–976.

[Fechter et al., 2020] Fechter, T. and Baltas, D. “One-Shot Learning for Deformable Medical Image Registration and Periodic Motion Tracking”. *IEEE Transactions on Medical Imaging* 39 (7), 2020, pp. 2506–2517.

[Felzenszwalb et al., 2005] Felzenszwalb, P. F. and Huttenlocher, D. P. “Pictorial Structures for Object Recognition”. *International Journal of Computer Vision* 61 (1), 2005, pp. 55–79.

[Felzenszwalb et al., 2006] Felzenszwalb, P. F. and Huttenlocher, D. P. “Efficient Belief Propagation for Early Vision”. *International Journal of Computer Vision* 70 (1), 2006, pp. 41–54.

[Felzenszwalb et al., 2008] Felzenszwalb, P., McAllester, D., and Ramanan, D. “A Discriminatively Trained, Multiscale, Deformable Part Model”. In: *Conference on Computer Vision and Pattern Recognition –CVPR 2008.* 2008, pp. 1–8.

[Feng et al., 2020] Feng, D., Haase-Schütz, C., Rosenbaum, L., Hertlein, H., Glaeser, C., Timm, F., Wiesbeck, W., and Dietmayer, K. "Deep Multi-Modal Object Detection and Semantic Segmentation for Autonomous Driving: Datasets, Methods, and Challenges". *IEEE Transactions on Intelligent Transportation Systems* 22 (3), 2020, pp. 1341–1360.

[Flampouri et al., 2006] Flampouri, S., Jiang, S. B., Sharp, G. C., Wolfgang, J., Patel, A. A., and Choi, N. C. "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* 51 (11), 2006, p. 2763.

[Förstner et al., 1987] Förstner, W. and Gülch, E. "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.

[Fu et al., 2020] Fu, Y., Lei, Y., Wang, T., Higgins, K., Bradley, J. D., Curran, W. J., Liu, T., and Yang, X. "LungRegNet: An Unsupervised Deformable Image Registration Method for 4D-CT Lung". *Medical Physics* 47 (4), 2020, pp. 1763–1774.

[Gaggion et al., 2021] Gaggion, N., Mansilla, L., Milone, D. H., and Ferrante, E. "Hybrid Graph Convolutional Neural Networks for Landmark-based Anatomical Segmentation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention.* 2021, pp. 600–610.

[Galbán et al., 2012] Galbán, C. J., Han, M. K., Boes, J. L., Chughtai, K. A., Meyer, C. R., Johnson, T. D., Galbán, S., Rehemtulla, A., Kazerooni, E. A., Martinez, F. J., et al. "Computed Tomography–based Biomarker Provides Unique Signature for Diagnosis of COPD Phenotypes and Disease Progression". *Nature Medicine* 18 (11), 2012, p. 1711.

[Garcia-Uceda Juarez et al., 2019] Garcia-Uceda Juarez, A., Selvan, R., Saghir, Z., and Bruijne, M. d. "A Joint 3D UNet-Graph Neural Network-based Method for Airway Segmentation from Chest CTs". In: *Machine Learning in Medical Imaging –MICCAI 2019 Workshops.* 2019, pp. 583–591.

[Ghavami et al., 2019] Ghavami, N., Hu, Y., Gibson, E., Bonmati, E., Emberton, M., Moore, C. M., and Barratt, D. C. "Automatic Segmentation of Prostate MRI using Convolutional Neural Networks: Investigating the Impact of Network Architecture on the Accuracy of Volume Measurement and MRI-ultrasound Registration". *Medical Image Analysis* 58, 2019, p. 101558.

[Girshick et al., 2011] Girshick, R., Shotton, J., Kohli, P., Criminisi, A., and Fitzgibbon, A. "Efficient Regression of General-Activity Human Poses from Depth Images". In: *International Conference on Computer Vision –ICCV 2011.* 2011, pp. 415–422.

[Girshick et al., 2015] Girshick, R., Iandola, F., Darrell, T., and Malik, J. “Deformable Part Models are Convolutional Neural Networks”. In: *Conference on Computer Vision and Pattern Recognition –CVPR 2015*. 2015, pp. 437–446.

[Glocker et al., 2008] Glocker, B., Komodakis, N., Tziritas, G., Navab, N., and Paragios, N. “Dense Image Registration through MRFs and Efficient Linear Programming”. *Medical Image Analysis* 12 (6), 2008, pp. 731–741.

[Glocker et al., 2011] Glocker, B., Sotiras, A., Komodakis, N., and Paragios, N. “Deformable Medical Image Registration: Setting the State of the Art with Discrete Methods”. *Annual rReview of Biomedical Engineering* 13, 2011, pp. 219–244.

[Goodfellow et al., 2016] Goodfellow, I., Bengio, Y., and Courville, A. *Deep Learning*. 2016.

[Gutiérrez et al., 2021] Gutiérrez, J., Rodríguez, V., and Martin, S. “Comprehensive Review of Vision-based Fall Detection Systems”. *Sensors* 21 (3), 2021, p. 947.

[Haber et al., 2006] Haber, E. and Modersitzki, J. “Intensity Gradient based Registration and Fusion of Multi-Modal Images”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention –MICCAI 2006*. 2006, pp. 726–733.

[Hansen et al., 2019a] Hansen, L., Diesel, J., and Heinrich, M. P. “Multi-Kernel Diffusion CNNs for Graph-Based Learning on Point Clouds”. In: *Geometry Meets Deep Learning –ECCV 2018 Workshops*. 2019, pp. 456–469.

[Hansen et al., 2019b] Hansen, L., Diesel, J., and Heinrich, M. P. “Regularised Landmark Detection with CAEs for Human Pose Estimation in the Operating Room”. In: *Bildverarbeitung für die Medizin 2019 –BVM 2019*. 2019, pp. 178–183.

[Hansen et al., 2019c] Hansen, L., Dittmer, D., and Heinrich, M. P. “Learning Deformable Point Set Registration with Regularized Dynamic Graph CNNs for Large Lung Motion in COPD Patients”. In: *Graph Learning in Medical Imaging –MICCAI 2019 Workshops*. 2019, pp. 53–61.

[Hansen et al., 2019d] Hansen, L. and Heinrich, M. P. “Sparse Structured Prediction for Semantic Edge Detection in Medical Images”. In: *International Conference on Medical Imaging with Deep Learning –MIDL 2019*. 2019, pp. 250–259.

[Hansen et al., 2019e] Hansen, L., Siebert, M., Diesel, J., and Heinrich, M. P. “Fusing Information From Multiple 2D Depth Cameras for 3D Human Pose Estimation in the Operating Room”. *International Journal of Computer Assisted Radiology and Surgery* 14 (11), 2019, pp. 1871–1879.

[Hansen et al., 2020] Hansen, L. and Heinrich, M. P. “Tackling the Problem of Large Deformations in Deep Learning Based Medical Image Registration Using

Displacement Embeddings". In: *International Conference on Medical Imaging with Deep Learning –Extended Abstract Track –MIDL 2020.* 2020.

[Hansen et al., 2021a] Hansen, L. and Heinrich, M. P. "Deep Learning Based Geometric Registration for Medical Images: How Accurate Can We Get Without Visual Features?" In: *Information Processing in Medical Imaging –IPMI 2021.* 2021, pp. 18–30.

[Hansen et al., 2021b] Hansen, L. and Heinrich, M. P. "GraphRegNet: Deep Graph Regularisation Networks on Sparse Keypoints for Dense Registration of 3D Lung CTs". *IEEE Transactions on Medical Imaging* 40 (9), 2021, pp. 2246–2257.

[Hansen et al., 2021c] Hansen, L. and Heinrich, M. P. "Revisiting Iterative Highly Efficient Optimisation Schemes in Medical Image Registration". In: *Medical Image Computing and Computer Assisted Intervention –MICCAI 2021.* 2021, pp. 203–212.

[Haque et al., 2016] Haque, A., Peng, B., Luo, Z., Alahi, A., Yeung, S., and Fei-Fei, L. "Towards Viewpoint Invariant 3D Human Pose Estimation". In: *European Conference on Computer Vision –ECCV 2016.* 2016, pp. 160–177.

[He et al., 2016] He, K., Zhang, X., Ren, S., and Sun, J. "Deep Residual Learning for Image Recognition". In: *Conference on Computer Vision and Pattern Recognition –CVPR 2016.* 2016, pp. 770–778.

[He et al., 2017] He, K., Gkioxari, G., Dollár, P., and Girshick, R. "Mask R-CNN". In: *International Conference on Computer Vision –ICCV 2017.* 2017, pp. 2980–2988.

[Heimann et al., 2009] Heimann, T. and Meinzer, H.-P. "Statistical shape models for 3D medical image segmentation: a review". *Medical image analysis* 13 (4), 2009, pp. 543–563.

[Heinrich et al., 2011] Heinrich, M. P., Jenkinson, M., Bhushan, M., Matin, T., Gleeson, F. V., Brady, J. M., and Schnabel, J. A. "Non-Local Shape Descriptor: A New Similarity Metric for Deformable Multi-Modal Registration". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention –MICCAI 2011.* 2011, pp. 541–548.

[Heinrich et al., 2012] Heinrich, M. P., Jenkinson, M., Bhushan, M., Matin, T., Gleeson, F. V., Brady, M., and Schnabel, J. A. "MIND: Modality Independent Neighbourhood Descriptor for Multi-Modal Deformable Registration". *Medical Image Analysis* 16 (7), 2012, pp. 1423–1435.

[Heinrich et al., 2013a] Heinrich, M. P., Jenkinson, M., Brady, S. M., and Schnabel, J. A. "MRF-based Deformable Registration and Ventilation Estimation of Lung CT". *IEEE Transaction on Medical Imaging* 32 (7), 2013, pp. 1239–48.

[Heinrich et al., 2013b] Heinrich, M. P., Jenkinson, M., Papież, B. W., Brady, M., and Schnabel, J. A. "Towards Realtime Multimodal Fusion for Image-Guided

Interventions using Self-Similarities". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention –MICCAI 2013*. 2013, pp. 187–194.

[Heinrich et al., 2015] Heinrich, M. P., Handels, H., and Simpson, I. J. "Estimating Large Lung Motion in COPD Patients by Symmetric Regularised Correspondence Fields". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention –MICCAI 2015*. 2015, pp. 338–345.

[Heinrich et al., 2020] Heinrich, M. P. and Hansen, L. "Highly Accurate and Memory Efficient Unsupervised Learning-Based Discrete CT Registration Using 2.5D Displacement Search". In: *Medical Image Computing and Computer Assisted Intervention –MICCAI 2020*. 2020, pp. 190–200.

[Heinrich et al., 2022] Heinrich, M. P. and Hansen, L. "Voxelmorph++ Going Beyond the Cranial Vault with Keypoint Supervision and Multi-Channel Instance Optimisation". In: *International Workshop on Biomedical Image Registration –WBIR 2022*. 2022.

[Heinrich, 2019] Heinrich, M. P. "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 –MICCAI 2021*. 2019, pp. 50–58.

[Hering et al., 2019] Hering, A., Ginneken, B., and Heldmann, S. "mlVirnet: Multilevel Variational Image Registration Network". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention –MICCAI 2019*. 2019, pp. 257–265.

[Hering et al., 2021] Hering, A., Häger, S., Moltz, J., Lessmann, N., Heldmann, S., and van Ginneken, B. "CNN-based lung CT Registration with Multiple Anatomical Constraints". *Medical Image Analysis* 72, 2021, p. 102139.

[Hering et al., 2022] Hering, A., Hansen, L., Mok, T. C. W., Chung, A. C. S., Siebert, H., Häger, S., Lange, A., Kuckertz, S., Heldmann, S., Shao, W., Vesal, S., Rusu, M., Sonn, G., Estienne, T., Vakalopoulou, M., Han, L., Huang, Y., Yap, P.-T., Brudfors, M., Balbastre, Y., Joutard, S., Modat, M., Lifshitz, G., Raviv, D., Lv, J., Li, Q., Jaouen, V., Visvikis, D., Fourcade, C., Rubeaux, M., Pan, W., Xu, Z., Jian, B., Benetti, F. D., Wodzinski, M., Gunnarsson, N., Sjölund, J., Grzech, D., Qiu, H., Li, Z., Großbröhmer, C., Hoopes, A., Reinertsen, I., Xiao, Y., Landman, B., Huo, Y., Murphy, K., Ginneken, B., Dalca, A., and Heinrich, M. P. "Learn2Reg: Comprehensive Multi-Task Medical Image Registration Challenge, Dataset and Evaluation in the Era of Deep Learning". *arXiv preprint arXiv:2112.04489*, 2022.

[Hu et al., 2018] Hu, Y., Modat, M., Gibson, E., Li, W., Ghavami, N., Bonmati, E., Wang, G., Bandula, S., Moore, C. M., Emberton, M., et al. "Weakly-Supervised

Convolutional Neural Networks for Multimodal Image Registration”. *Medical Image Analysis* 49, 2018, pp. 1–13.

[Huang et al., 2017] Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. “Densely Connected Convolutional Networks”. In: *Conference on Computer Vision and Pattern Recognition –CVPR 2017*. 2017, pp. 4700–4708.

[Huang et al., 2018] Huang, Q., Wang, W., and Neumann, U. “Recurrent Slice Networks for 3D Segmentation of Point Clouds”. In: *Conference on Computer Vision and Pattern Recognition –CVPR 2018*. 2018, pp. 2626–2635.

[Ilg et al., 2017] Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., and Brox, T. “Flownet 2.0: Evolution of Optical Flow Estimation with Deep Networks”. In: *Conference on Computer Vision and Pattern Recognition –CVPR 2017*. 2017, pp. 2462–2470.

[Ioffe et al., 2015] Ioffe, S. and Szegedy, C. “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift”. In: *International Conference on Machine Learning*. 2015, pp. 448–456.

[Ionescu et al., 2013] Ionescu, C., Papava, D., Olaru, V., and Sminchisescu, C. “Human3.6m: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments”. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36 (7), 2013, pp. 1325–1339.

[Isensee et al., 2021] Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J., and Maier-Hein, K. H. “nnU-Net: A Self-Configuring Method for Deep Learning-Based Biomedical Image Segmentation”. *Nature Methods* 18 (2), 2021, pp. 203–211.

[Jacob et al., 2013] Jacob, M. G., Li, Y.-T., Akingba, G. A., and Wachs, J. P. “Collaboration with a Robotic Scrub Nurse”. *Communications of the ACM* 56 (5), 2013, pp. 68–75.

[Jiang et al., 2019] Jiang, W., Sun, W., Tagliasacchi, A., Trulls, E., and Yi, K. M. “Linearized Multi-Sampling for Differentiable Image Transformation”. In: *International Conference on Computer Vision –ICCV 2019*. 2019, pp. 2988–2997.

[Jimenez-del-Toro et al., 2016] Jimenez-del-Toro, O., Müller, H., Krenn, M., Gruenberg, K., Taha, A. A., Winterstein, M., Eggel, I., Foncubierta-Rodriguez, A., Goksel, O., Jakab, A., et al. “Cloud-Based Evaluation of Anatomical Structure Segmentation and Landmark Detection Algorithms: VISCERAL Anatomy Benchmarks”. *IEEE Transactions on Medical Imaging* 35 (11), 2016, pp. 2459–2475.

[Johnson et al., 2016] Johnson, M. J., Duvenaud, D. K., Wiltschko, A., Adams, R. P., and Datta, S. R. “Composing Graphical Models with Neural Networks for Structured Representations and Fast Inference”. In: *Advances in Neural Information Processing Systems –NeurIPS 2016*. 2016, pp. 2946–2954.

[Jumper et al., 2021] Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., and Hassabis, D. "Highly Accurate Protein Structure Prediction with AlphaFold". *Nature* 596 (7873), 2021, pp. 583–589.

[Jung et al., 2016] Jung, H. Y., Suh, Y., Moon, G., and Lee, K. M. "A Sequential Approach to 3D Human Pose Estimation: Separation of Localization and Identification of Body Joints". In: *European Conference on Computer Vision –ECCV 2016*. 2016, pp. 747–761.

[Kabus et al., 2009] Kabus, S., Klinder, T., Murphy, K., Ginneken, B., Lorenz, C., and Pluim, J. P. "Evaluation of 4D-CT Lung Registration". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention –MICCAI 2009*. 2009, pp. 747–754.

[Kadkhodamohammadi et al., 2017] Kadkhodamohammadi, A., Gangi, A., Mathelin, M., and Padoy, N. "A Multi-View RGB-D Approach for Human Pose Estimation in Operating Rooms". In: *Winter Conference on Applications of Computer Vision –WACV 2017*. 2017, pp. 363–372.

[Kadkhodamohammadi et al., 2021] Kadkhodamohammadi, A. and Padoy, N. "A Generalizable Approach for Multi-View 3D Human Pose Regression". *Machine Vision and Applications* 32 (1), 2021, pp. 1–14.

[Kamnitsas et al., 2017] Kamnitsas, K., Ledig, C., Newcombe, V. F., Simpson, J. P., Kane, A. D., Menon, D. K., Rueckert, D., and Glocker, B. "Efficient Multi-Scale 3D CNN with Fully Connected CRF for Accurate Brain Lesion Segmentation". *Medical Image Analysis* 36, 2017, pp. 61–78.

[Katircioglu et al., 2018] Katircioglu, I., Tekin, B., Salzmann, M., Lepetit, V., and Fua, P. "Learning Latent Representations of 3D Human Pose with Deep Neural Networks". *International Journal of Computer Vision* 126 (12), 2018, pp. 1326–1341.

[Kim et al., 2011] Kim, V. G., Lipman, Y., and Funkhouser, T. "Blended Intrinsic Maps". In: *ACM Transactions on Graphics*. Vol. 30. 4. 2011, p. 79.

[Kipf et al., 2017] Kipf, T. N. and Welling, M. "Semi-Supervised Classification with Graph Convolutional Networks". In: *International Conference on Learning Representations –ICLR 2017*. 2017.

[Klein et al., 2009] Klein, S., Staring, M., Murphy, K., Viergever, M. A., and Pluim, J. P. "Elastix: A Toolbox for Intensity-based Medical Image Registration". *IEEE Transactions on Medical Imaging* 29 (1), 2009, pp. 196–205.

[Kleiman et al., 2018] Kleiman, Y. and Ovsjanikov, M. “Robust Structure-based Shape Correspondence”. *Computer Graphics Forum* 20, 2018, pp. 1–13.

[Knobelreiter et al., 2017] Knobelreiter, P., Reinbacher, C., Shekhovtsov, A., and Pock, T. “End-to-End Training of Hybrid CNN-CRF Models for Stereo”. In: *Conference on Computer Vision and Pattern Recognition –CVPR 2017*. 2017, pp. 2339–2348.

[König et al., 2018] König, L., Rühaak, J., Derksen, A., and Lellmann, J. “A Matrix-Free Approach to Parallel and Memory-Efficient Deformable Image Registration”. *SIAM Journal on Scientific Computing* 40 (3), 2018, pp. 858–888.

[Kong et al., 2021] Kong, F., Wilson, N., and Shadden, S. “A Deep-learning Approach for Direct Whole-Heart Mesh Reconstruction”. *Medical Image Analysis* 74, 2021, p. 102222.

[Krähenbühl et al., 2011] Krähenbühl, P. and Koltun, V. “Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials”. In: *Advances in Neural Information Processing Systems –NeurIPS 2011*. 2011, pp. 109–117.

[Krizhevsky et al., 2012] Krizhevsky, A., Sutskever, I., and Hinton, G. E. “ImageNet Classification with Deep Convolutional Neural Networks”. *Advances in Neural Information Processing Systems –NeurIPS 2012* 25, 2012.

[Ktena et al., 2017] Ktena, S. I., Parisot, S., Ferrante, E., Rajchl, M., Lee, M., Glocker, B., and Rueckert, D. “Distance Metric Learning using Graph Convolutional Networks: Application to Functional Brain Networks”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention –MICCAI 2017*. 2017, pp. 469–477.

[Le et al., 2017] Le, M., Lieman-Sifry, J., Lau, F., Sall, S., Hsiao, A., and Golden, D. “Computationally Efficient Cardiac Views Projection using 3D Convolutional Neural Networks”. In: *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support –MICCAI 2017 Workshops*. 2017, pp. 109–116.

[LeCun et al., 1989] LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. “Backpropagation applied to Handwritten ZIP Code Recognition”. *Neural computation* 1 (4), 1989, pp. 541–551.

[Leow et al., 2007] Leow, A. D., Yanovsky, I., Chiang, M.-C., Lee, A. D., Klunder, A. D., Lu, A., Becker, J. T., Davis, S. W., Toga, A. W., and Thompson, P. M. “Statistical Properties of Jacobian Maps and the Realization of Unbiased Large-Deformation Nonlinear Image Registration”. *IEEE Transactions on Medical Imaging* 26 (6), 2007, pp. 822–832.

[Li et al., 2018] Li, R., Yao, J., Zhu, X., Li, Y., and Huang, J. “Graph CNN for Survival Analysis on Whole Slide Pathological Images”. In: *International Conference on*

Medical Image Computing and Computer-Assisted Intervention –MICCAI 2018. 2018, pp. 174–182.

[Li et al., 2021] Li, X., Zhou, Y., Dvornek, N., Zhang, M., Gao, S., Zhuang, J., Scheinost, D., Staib, L. H., Ventola, P., and Duncan, J. S. “BrainGNN: Interpretable Brain Graph Neural Network for fMRI Analysis”. *Medical Image Analysis* 74, 2021, p. 102233.

[Lindner et al., 2015] Lindner, C., Bromiley, P. A., Ionita, M. C., and Cootes, T. F. “Robust and Accurate Shape Model Matching using Random Forest Regression-Voting”. *Transactions on Pattern Analysis and Machine Intelligence* 37 (9), 2015, pp. 1862–1874.

[Linehan et al., 2016] Linehan, M., Gautam, R., Kirk, S., Lee, Y., Roche, C., Bonaccio, E., and Jarosz, R. “Radiology Data from the Cancer Genome Atlas Cervical Kidney Renal Papillary Cell Carcinoma [KIRP] Collection”. *The Cancer Imaging Archive*, 2016.

[Litman et al., 2014] Litman, R. and Bronstein, A. M. “Learning Spectral Descriptors for Deformable Shape Correspondence”. *Transactions on Pattern Analysis and Machine Intelligence* 36 (1), 2014, pp. 171–180.

[Litjens et al., 2017] Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., Van Der Laak, J. A., Van Ginneken, B., and Sánchez, C. I. “A Survey on Deep Learning in Medical Image Analysis”. *Medical Image Analysis* 42, 2017, pp. 60–88.

[Liu et al., 2016] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., and Berg, A. C. “SSD: Single Shot Multibox Detector”. In: *European Conference on Computer Vision –ECCV 2016*. 2016, pp. 21–37.

[Liu et al., 2019a] Liu, S. and Ostadabbas, S. “Seeing Under the Cover: A Physics Guided Learning Approach for In-bed Pose Estimation”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention –MICCAI 2019*. 2019, pp. 236–245.

[Liu et al., 2019b] Liu, S., Yin, Y., and Ostadabbas, S. “In-Bed Pose Estimation: Deep Learning with Shallow Dataset”. *Journal of Translational Engineering in Health and Medicine* 7, 2019, pp. 1–12.

[Liu et al., 2019c] Liu, X., Qi, C. R., and Guibas, L. J. “Flownet3D: Learning Scene Flow in 3D Point Clouds”. In: *International Conference on Computer Vision and Pattern Recognition –CVPR 2019*. 2019, pp. 529–537.

[Liu et al., 2019d] Liu, Z., Tang, H., Lin, Y., and Han, S. “Point-Voxel CNN for Efficient 3D Deep Learning”. In: *Advances in Neural Information Processing Systems –NeurIPS 2019*. 2019, pp. 965–975.

[Liu et al., 2020] Liu, Y., Cheng, M.-M., Fan, D.-P., Zhang, L., Bian, J., and Tao, D. *Semantic Edge Detection with Diverse Deep Supervision.* 2020. arXiv: 1804.02864.

[Lou et al., 2019] Lou, B., Doken, S., Zhuang, T., Wingerter, D., Gidwani, M., Mistry, N., Ladic, L., Kamen, A., and Abazeed, M. E. "An Image-based Deep Learning Framework for Individualising Radiotherapy Dose: A Retrospective Analysis of Outcome Prediction". *The Lancet Digital Health* 1 (3), 2019, e136–e147.

[Ma et al., 2018] Ma, J., Jiang, J., Zhou, H., Zhao, J., and Guo, X. "Guided Locality Preserving Feature Matching for Remote Sensing Image Registration". *IEEE transactions on geoscience and remote sensing* 56 (8), 2018, pp. 4435–4447.

[Ma et al., 2020] Ma, J., Zhu, X., Yang, D., Chen, J., and Wu, G. "Attention-guided Deep Graph Neural Network for Longitudinal Alzheimer's Disease Analysis". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention –MICCAI 2020.* 2020, pp. 387–396.

[Marstal et al., 2016] Marstal, K., Berendsen, F., Staring, M., and Klein, S. "SimpleElastix: A User-friendly, Multi-lingual Library for Medical Image Registration". In: *Conference on Computer Vision and Pattern Recognition Workshops –CVPR Workshops 2016.* 2016, pp. 134–142.

[Masci et al., 2011] Masci, J., Meier, U., Cireşan, D., and Schmidhuber, J. "Stacked Convolutional Auto-Encoders for Hierarchical Feature Extraction". In: *International Conference on Artificial Neural Networks -ICANN 2011.* 2011, pp. 52–59.

[Masci et al., 2015] Masci, J., Boscaini, D., Bronstein, M., and Vandergheynst, P. "Geodesic Convolutional Neural Networks on Riemannian Manifolds". In: *Workshop on 3D Representation and Recognition –ICCV 2015 Workshops.* 2015, pp. 37–45.

[Maturana et al., 2015] Maturana, D. and Scherer, S. "VoxNet: A 3D Convolutional Neural Network for Real-Time Object Recognition". In: *International Conference on Intelligent Robots and Systems –IROS 2015.* 2015, pp. 922–928.

[McCoy et al., 2018] McCoy, T. H. and Perlis, R. H. "Temporal Trends and Characteristics of Reportable Health Data Breaches, 2010-2017". *Journal of the American Medical Association* 320 (12), 2018, pp. 1282–1284.

[Mikael et al., 2015] Mikael, H., Joan, B., and Yann, L. *Deep Convolutional Networks on Graph-Structured Data.* 2015. arXiv: 1506.05163.

[Modat et al., 2010] Modat, M., Ridgway, G. R., Taylor, Z. A., Lehmann, M., Barnes, J., Hawkes, D. J., Fox, N. C., and Ourselin, S. "Fast Free-Form Deformation using Graphics Processing Units". *Computer Methods and Programs in Biomedicine* 98 (3), 2010, pp. 278–284.

[Mohanty et al., 2020] Mohanty, S. P., Czakon, J., Kaczmarek, K. A., Pyskir, A., Tarasiewicz, P., Kunwar, S., Rohrbach, J., Luo, D., Prasad, M., Fleer, S., et al. “Deep Learning for Understanding Satellite Imagery: An Experimental Survey”. *Frontiers in Artificial Intelligence*, 2020, p. 85.

[Mok et al., 2020] Mok, T. C. and Chung, A. C. “Large Deformation Diffeomorphic Image Registration with Laplacian Pyramid Networks”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention –MICCAI 2021.* 2020, pp. 211–221.

[Monti et al., 2017] Monti, F., Boscaini, D., Masci, J., Rodola, E., Svoboda, J., and Bronstein, M. M. “Geometric Deep Learning on Graphs and Manifolds using Mixture Model CNNs”. In: *Conference on Computer Vision and Pattern Recognition –CVPR 2017.* 2017, pp. 5115–5124.

[Monti et al., 2018] Monti, F., Shchur, O., Bojchevski, A., Litany, O., Günnemann, S., and Bronstein, M. M. *Dual-Primal Graph Convolutional Networks.* 2018. arXiv: `1806.00770`.

[Moon et al., 2018] Moon, G., Yong Chang, J., and Mu Lee, K. “V2V-PoseNet: Voxel-to-Voxel Prediction Network for Accurate 3D Hand and Human Pose Estimation from a Single Depth Map”. In: *Conference on Computer Vision and Pattern Recognition –CVPR 2018.* 2018, pp. 5079–5088.

[Mori et al., 2004] Mori, G., Ren, X., Efros, A. A., and Malik, J. “Recovering Human Body Configurations: Combining Segmentation and Recognition”. In: *Conference on Computer Vision and Pattern Recognition –CVPR 2004.* Vol. 2. 2004, pp. 326–333.

[Muenzing et al., 2014] Muenzing, S. E., Ginneken, B., Viergever, M. A., and Pluim, J. P. “DIRBoost–An Algorithm for Boosting Deformable Image Registration: Application to Lung CT Intra-Subject Registration”. *Medical Image Analysis* 18 (3), 2014, pp. 449–459.

[Murphy et al., 2011] 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. “Evaluation of Registration Methods on Thoracic CT: The EMPIRE10 Challenge”. *IEEE Transactions on Medical Imaging* 30 (11), 2011, pp. 1901–1920.

[Myronenko et al., 2010] Myronenko, A. and Song, X. “Point Set Registration: Coherent Point Drift”. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32 (12), 2010, pp. 2262–2275.

[Myronenko, 2018] Myronenko, A. “3D MRI Brain Tumor Segmentation using Autoencoder Regularization”. In: *Brainles –MICCAI 2018 Workshops.* 2018, pp. 311–320.

[Nakao et al., 2021] Nakao, M., Tong, F., Nakamura, M., and Matsuda, T. “Image-to-Graph Convolutional Network for Deformable Shape Reconstruction from a Single Projection Image". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention –MICCAI 2021*. 2021, pp. 259–268.

[Nekolla et al., 2017] Nekolla, E., Schegerer, A., Griebel, J., and Brix, G. "Frequency and Doses of Diagnostic and Interventional X-Ray Applications: Trends between 2007 and 2014". *Der Radiologe* 57 (7), 2017, pp. 555–562.

[Newell et al., 2016] Newell, A., Yang, K., and Deng, J. "Stacked Hourglass Networks for Human Pose Estimation". In: *European Conference on Computer Vision –ECCV2016*. 2016, pp. 483–499.

[Newell et al., 2017] Newell, A., Huang, Z., and Deng, J. "Associative Embedding: End-to-End Learning for Joint Detection and Grouping". In: *Advances in Neural Information Processing Systems –NeurIPS 2017*. 2017, pp. 2277–2287.

[Nguyen et al., 2020] Nguyen, M., He, T., An, L., Alexander, D. C., Feng, J., and Yeo, B. T. "Predicting Alzheimer's Disease Progression using Deep Recurrent Neural Networks". *NeuroImage* 222, 2020, p. 117203.

[OECD, 2021] OECD, *Health at a Glance 2021*. 2021, p. 274.

[Oktay et al., 2015] Oktay, O., Schuh, A., Rajchl, M., Keraudren, K., Gomez, A., Heinrich, M. P., Penney, G., and Rueckert, D. "Structured Decision Forests for Multi-Modal Ultrasound Image Registration". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention –MICCAI 2015*. 2015, pp. 363–371.

[Oktay et al., 2017] Oktay, O., Ferrante, E., Kamnitsas, K., Heinrich, M. P., Bai, W., Caballero, J., Cook, S. A., De Marvao, A., Dawes, T., O'Regan, D. P., et al. "Anatomically Constrained Neural Networks (ACNNs): Application to Cardiac Image Enhancement and Segmentation". *IEEE Transactions on Medical Imaging* 37 (2), 2017, pp. 384–395.

[Padoy et al., 2012] Padoy, N., Blum, T., Ahmadi, S.-A., Feussner, H., Berger, M.-O., and Navab, N. "Statistical Modeling and Recognition of Surgical Workflow". *Medical Image Analysis* 16 (3), 2012, pp. 632–641.

[Pan et al., 2020] Pan, F., Ye, T., Sun, P., Gui, S., Liang, B., Li, L., Zheng, D., Wang, J., Hesketh, R. L., Yang, L., et al. "Time Course of Lung Changes on Chest CT during recovery from 2019 Novel Coronavirus (COVID-19) Pneumonia". *Radiology*, 2020, p. 200370.

[Parisot et al., 2017] Parisot, S., Ktena, S. I., Ferrante, E., Lee, M., Moreno, R. G., Glocker, B., and Rueckert, D. "Spectral Graph Convolutions for Population-Based Disease Prediction". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention –MICCAI 2017*. 2017, pp. 177–185.

[Paszke et al., 2019] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. “PyTorch: An Imperative Style, High-Performance Deep Learning Library”. In: *Advances in Neural Information Processing Systems –NeurIPS 2015.* 2019, pp. 8024–8035.

[Pavlakos et al., 2017] Pavlakos, G., Zhou, X., Derpanis, K. G., and Daniilidis, K. “Coarse-to-Fine Volumetric Prediction for Single-Image 3D Human Pose”. In: *Conference on Computer Vision and Pattern Recognition –CVPR 2017.* 2017, pp. 1263–1272.

[Polzin et al., 2013] Polzin, T., Rühaak, J., Werner, R., Strehlow, J., Heldmann, S., Handels, H., and Modersitzki, J. “Combining Automatic Landmark Detection and Variational Methods for Lung CT Registration”. In: *International Workshop on Pulmonary Image Analysis –MICCAI 2013 Workshops.* 2013, pp. 85–96.

[Puy et al., 2020] Puy, G., Boulch, A., and Marlet, R. “FLOT: Scene Flow on Point Clouds Guided by Optimal Transport”. In: *European Conference on Computer Vision –ECCV 2020.* 2020, pp. 527–544.

[Qi et al., 2017a] Qi, C. R., Su, H., Mo, K., and Guibas, L. J. “PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation”. In: *Conference on Computer Vision and Pattern Recognition.* 2017, pp. 652–660.

[Qi et al., 2017b] Qi, C. R., Yi, L., Su, H., and Guibas, L. J. “PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space”. In: 2017.

[Ravikumar et al., 2019] Ravikumar, N., Gooya, A., Beltrachini, L., Frangi, A. F., and Taylor, Z. A. “Generalised Coherent Point Drift for Group-Wise Multi-Dimensional Analysis of Diffusion Brain MRI Data”. *Medical Image Analysis* 53, 2019, pp. 47–63.

[Reinhardt et al., 2008] Reinhardt, J. M., Ding, K., Cao, K., Christensen, G. E., Hoffman, E. A., and Bodas, S. V. “Registration-based Estimates of Local Lung Tissue Expansion Compared to Xenon CT Measures of Specific Ventilation”. *Medical Image Analysis* 12 (6), 2008, pp. 752–763.

[Ren et al., 2015] Ren, S., He, K., Girshick, R., and Sun, J. “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks”. In: *Advances in Neural Information Processing Systems –NeurIPS 2015.* 2015, pp. 91–99.

[Ronneberger et al., 2015] Ronneberger, O., Fischer, P., and Brox, T. “U-Net: Convolutional Networks for Biomedical Image Segmentation”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention –MICCAI 2015.* 2015, pp. 234–241.

[Rueckert et al., 2019] Rueckert, D. and Schnabel, J. A. “Model-based and Data-driven Strategies in Medical Image Computing”. *Proceedings of the IEEE* 108 (1), 2019, pp. 110–124.

[Rühaak et al., 2017] Rühaak, J., Polzin, T., Heldmann, S., Simpson, I. J., Handels, H., Modersitzki, J., and Heinrich, M. P. “Estimation of Large Motion in Lung CT by Integrating Regularized Keypoint Correspondences Into Dense Deformable Registration”. *IEEE Transactions on Medical Imaging* 36 (8), 2017, pp. 1746–1757.

[Rusu et al., 2009] Rusu, R. B., Blodow, N., and Beetz, M. “Fast Point Feature Histograms (FPFH) for 3D Registration”. In: *International Conference on Robotics and Automation –ICRA 2009.* 2009, pp. 3212–3217.

[Salomon et al., 2019] Salomon, L., Alfirevic, Z., Da Silva Costa, F., Deter, R., Figueras, F., Ghi, T. a., Glanc, P., Khalil, A., Lee, W., Napolitano, R., et al. “ISUOG Practice Guidelines: Ultrasound Assessment of Fetal Biometry and Growth”. *Ultrasound in Obstetrics & Gynecology* 53 (6), 2019, pp. 715–723.

[Sandkühler et al., 2019] Sandkühler, R., Andermatt, S., Bauman, G., Nyilas, S., Jud, C., and Cattin, P. C. “Recurrent Registration Neural Networks for Deformable Image Registration”. In: *Advances in Neural Information Processing Systems –NeurIPS.* 2019, pp. 8758–8768.

[Sandkühler et al., 2020] Sandkühler, R., Jud, C., Andermatt, S., and Cattin, P. C. *AirLab: Autograd Image Registration Laboratory.* 2020. arXiv: 1806.09907.

[Sardar et al., 2019] Sardar, P., Abbott, J. D., Kundu, A., Aronow, H. D., Granada, J. F., and Giri, J. “Impact of Artificial Intelligence on Interventional Cardiology: From Decision-Making Aid to Advanced Interventional Procedure Assistance”. *JACC: Cardiovascular Interventions* 12 (14), 2019, pp. 1293–1303.

[Sarlin et al., 2020] Sarlin, P.-E., DeTone, D., Malisiewicz, T., and Rabinovich, A. “Superglue: Learning Feature Matching with Graph Neural Networks”. In: *International Conference on Computer Vision and Pattern Recognition –CVPR 2020.* 2020, pp. 4938–4947.

[Schlemper et al., 2019] Schlemper, J., Oktay, O., Schaap, M., Heinrich, M., Kainz, B., Glocker, B., and Rueckert, D. “Attention Gated Networks: Learning to Leverage Salient Regions in Medical Images”. *Medical Image Analysis* 53, 2019, pp. 197–207.

[Sen et al., 2008] Sen, P., Namata, G., Bilgic, M., Getoor, L., Galligher, B., and Eliassi-Rad, T. “Collective Classification in Network Data”. *AI Magazine* 29 (3), 2008, pp. 93–106.

[Seregni et al., 2017] Seregni, M., Paganelli, C., Summers, P., Bellomi, M., Baroni, G., and Riboldi, M. “A Hybrid Image Registration and Matching Framework for

Real-Time Motion Tracking in MRI-guided Radiotherapy". *IEEE Transactions on Biomedical Engineering* 65 (1), 2017, pp. 131–139.

[Shen et al., 2021] Shen, Z., Feydy, J., Liu, P., Curiale, A., San Jose Estepar, R., San Jose Estepar, R., and Niethammer, M. "Accurate Point Cloud Registration with Robust Optimal Transport". *Advances in Neural Information Processing Systems –NeuRIPS 2021* 34, 2021.

[Shiraishi et al., 2000] Shiraishi, J., Katsuragawa, S., Ikezoe, J., Matsumoto, T., Kobayashi, T., Komatsu, K.-i., Matsui, M., Fujita, H., Kodera, Y., and Doi, K. "Development of a Digital Image Database for Chest Radiographs with and without a Lung Nodule: Receiver Operating Characteristic Analysis of Radiologists' Detection of Pulmonary Nodules". *American Journal of Roentgenology* 174 (1), 2000, pp. 71–74.

[Shotton et al., 2011] Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., and Blake, A. "Real-Time Human Pose Recognition in Parts from Single Depth Images". In: *Conference on Computer Vision and Pattern Recognition –CVPR 2011.* 2011, pp. 1297–1304.

[Silas et al., 2015] Silas, M. R., Grassia, P., and Langerman, A. "Video Recording of the Operating Room — Is Anonymity Possible?" *Journal of Surgical Research* 197 (2), 2015, pp. 272–276.

[Soberanis-Mukul et al., 2020] Soberanis-Mukul, R. D., Navab, N., and Albarqouni, S. "Uncertainty-based Graph Convolutional Networks for Organ Segmentation Refinement". In: *Medical Imaging with Deep Learning –MIDL 2020.* 2020, pp. 755–769.

[Song et al., 2010] Song, G., Tustison, N., Avants, B., and Gee, J. C. "Lung CT Image Registration using Diffeomorphic Transformation Models". In: 2010, pp. 23–32.

[Sotiras et al., 2010] Sotiras, A., Ou, Y., Glocker, B., Davatzikos, C., and Paragios, N. "Simultaneous Geometric-Iconic Registration". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention –MICCAI 2010.* 2010, pp. 676–683.

[Srivastava et al., 2014] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. "Dropout: A Simple Way to Prevent Neural Networks from Overfitting". *The Journal of Machine Learning Research* 15 (1), 2014, pp. 1929–1958.

[Srivastav et al., 2018] Srivastav, V., Issenhuth, T., Abdolrahim, K., Mathelin, M., Gangi, A., and Padoy, N. "MVOR: A Multi-view RGB-D Operating Room Dataset for 2D and 3D Human Pose Estimation". In: *Large-Scale Annotation of Biomedical Data and Expert Label Synthesis –MICCAI 2018 Workshops.* 2018, pp. 1–10.

[Su et al., 2015] Su, H., Maji, S., Kalogerakis, E., and Learned-Miller, E. “Multi-View Vonvolutional Neural Networks for 3D Shape Recognition”. In: *International Conference on Computer Vision –ICCV 2015*. 2015, pp. 945–953.

[Sun et al., 2009] Sun, J., Ovsjanikov, M., and Guibas, L. “A Concise and Provably Informative Multi-Scale Signature Based on Heat Diffusion”. *Computer Graphics Forum* 28 (5), 2009, pp. 1383–1392.

[Sun et al., 2018a] Sun, D., Yang, X., Liu, M.-Y., and Kautz, J. “PWC-Net: CNNs for Optical Flow using Pyramid, Warping, and Cost Volume”. In: *International Conference on Computer Vision and Pattern Recognition –CVPR 2018*. 2018, pp. 8934–8943.

[Sun et al., 2018b] Sun, X., Xiao, B., Wei, F., Liang, S., and Wei, Y. “Integral Human Pose Regression”. In: *European Conference on Computer Vision –ECCV 2018*. 2018, pp. 529–545.

[Tan et al., 2021] Tan, Z., Feng, J., and Zhou, J. “SGNet: Structure-Aware Graph-Based Network for Airway Semantic Segmentation”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention –MICCAI 2021*. 2021, pp. 153–163.

[Tchapmi et al., 2017] Tchapmi, L., Choy, C., Armeni, I., Gwak, J., and Savarese, S. “SEGCloud: Semantic Segmentation of 3D Point Clouds”. In: *International Conference on 3D Vision –3DV 2017*. 2017, pp. 537–547.

[Tekin et al., 2016] Tekin, B., Katircioglu, I., Salzmann, M., Lepetit, V., and Fua, P. “Structured Prediction of 3D Human Pose with Deep Neural Networks”. In: *British Machine Vision Conference –BMVC 2016*. 2016, pp. 1–11.

[Toshev et al., 2014] Toshev, A. and Szegedy, C. “DeepPose: Human Pose Estimation via Deep Neural Networks”. In: *Conference on Computer Vision and Pattern Recognition –CVPR 2014*. 2014, pp. 1653–1660.

[Tschirren et al., 2005] Tschirren, J., McLennan, G., Palágyi, K., Hoffman, E. A., and Sonka, M. “Matching and Anatomical Labeling of Human Airway Tree”. *IEEE Transactions on Medical Imaging* 24 (12), 2005, pp. 1540–1547.

[Van Ginneken et al., 2006] Van Ginneken, B., Stegmann, M. B., and Loog, M. “Segmentation of Anatomical Structures in Chest Radiographs using Supervised Methods: A Comparative Study on a Public Database”. *Medical image analysis* 10 (1), 2006, pp. 19–40.

[Vandemeulebroucke et al., 2007] Vandemeulebroucke, J., Sarrut, D., Clarysse, P., et al. “The POPI-Model, a Point-Validated Pixel-based Breathing Thorax Model”. In: *International Conference on the Use of Computers in Radiation Therapy –ICCR 2007*. Vol. 2. 2007, pp. 195–199.

[Velickovic et al., 2018] Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., and Bengio, Y. “Graph Attention Networks”. In: *International Conference on Learning Representations –ICLR 2018.* 2018.

[Vincent et al., 2010] Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., and Manzagol, P.-A. “Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion”. *Journal of Machine Learning Research* 11 (12), 2010, pp. 3371–3408.

[Vos et al., 2019] Vos, B. D., Berendsen, F. F., Viergever, M. A., Sokooti, H., Staring, M., and Išgum, I. “A Deep Learning Framework for Unsupervised Affine and Deformable Image Registration”. *Medical Image Analysis* 52, 2019, pp. 128–143.

[Wang et al., 2019a] Wang, Y., Sun, Y., Liu, Z., Sarma, S. E., Bronstein, M. M., and Solomon, J. M. “Dynamic Graph CNN for Learning on Point Clouds”. *ACM Transactions on Graphics* 38 (5), 2019.

[Wang et al., 2019b] Wang, Y., Zhao, L., Wang, M., and Song, Z. “Organ at Risk Segmentation in Head and Neck CT Images using a Two-stage Segmentation Framework based on 3D U-Net”. *IEEE Access* 7, 2019, pp. 144591–144602.

[Wei et al., 2016] Wei, S.-E., Ramakrishna, V., Kanade, T., and Sheikh, Y. “Convolutional Pose Machines”. In: *Conference on Computer Vision and Pattern Recognition –CVPR 2016.* 2016, pp. 4724–4732.

[Weihsbach et al., 2022] Weihsbach, C., Hansen, L., and Heinrich, M. P. “XEdgeConv-Net: Leveraging Graph Convolutions for Efficient, Permutation- and Rotation-invariant Dense 3D Medical Image Segmentation”. In: *Geometric Deep Learning in Medical Image Analysis –GeoMedIA 2022.* 2022, pp. 61–71.

[Wojciechowska et al., 2021] Wojciechowska, M., Malacrino, S., Garcia Martin, N., Fehri, H., and Rittscher, J. “Early Detection of Liver Fibrosis Using Graph Convolutional Networks”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention –MICCAI 2021.* 2021, pp. 217–226.

[Wu et al., 2019] Wu, Y. and Ji, Q. “Facial Landmark Detection: A Literature Survey”. *International Journal of Computer Vision* 127 (2), 2019, pp. 115–142.

[Wu et al., 2020] Wu, X., Qiu, L., Gu, X., and Long, Z. “Deep Learning-based Generic Automatic Surface Defect Inspection (ASDI) with Pixelwise Segmentation”. *IEEE Transactions on Instrumentation and Measurement* 70, 2020, pp. 1–10.

[Xiao et al., 2018] Xiao, B., Wu, H., and Wei, Y. “Simple Baselines for Human Pose Estimation and Tracking”. In: *European Conference on Computer Vision –ECCV 2018.* 2018, pp. 466–481.

[Xie et al., 2015] Xie, S. and Tu, Z. “Holistically-Nested Edge Detection”. In: *International Conference on Computer Vision and Pattern Recognition –CVPR 2015.* 2015, pp. 1395–1403.

[Xu et al., 2016] Xu, Z., Lee, C. P., Heinrich, M. P., Modat, M., Rueckert, D., Ourselin, S., Abramson, R. G., and Landman, B. A. "Evaluation of Six registration Methods for the Human Abdomen on Clinically Acquired CT". *IEEE Transactions on Biomedical Engineering* 63 (8), 2016, pp. 1563–1572.

[Xu et al., 2019] Xu, Z. and Niethammer, M. "DeepAtlas: Joint Semi-Supervised Learning of Image Registration and Segmentation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention –MICCAI 2019*. 2019, pp. 420–429.

[Yan et al., 2021] Yan, J., Chen, Y., Yang, S., Zhang, S., Jiang, M., Zhao, Z., Zhang, T., Zhao, Y., Becker, B., Liu, T., et al. "Multi-head GAGNN: A Multi-head Guided Attention Graph Neural Network for Modeling Spatio-temporal Patterns of Holistic Brain Functional Networks". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention –MICCAI 2021*. 2021, pp. 564–573.

[Yao et al., 2012] Yao, A., Gall, J., and Van Gool, L. "Coupled Action Recognition and Pose Estimation from Multiple Views". *International Journal of Computer Vision* 100 (1), 2012, pp. 16–37.

[Yao et al., 2021] Yao, D., Sui, J., Wang, M., Yang, E., Jiaerken, Y., Luo, N., Yap, P.-T., Liu, M., and Shen, D. "A Mutual Multi-scale Triplet Graph Convolutional Network for Classification of Brain Disorders using Functional or Structural Connectivity". *IEEE Transactions on Medical Imaging* 40 (4), 2021, pp. 1279–1289.

[Yu et al., 2017] Yu, Z., Feng, C., Liu, M.-Y., and Ramalingam, S. "CaseNet: Deep Category-Aware Semantic Edge Detection". In: *Conference on Computer Vision and Pattern Recognition –CVPR 2018*. 2017, pp. 21–26.

[Yusoff et al., 2013] Yusoff, Y. A., Basori, A. H., and Mohamed, F. "Interactive Hand and Arm Gesture Control for 2D Medical Image and 3D Volumetric Medical Visualization". *Procedia-Social and Behavioral Sciences* 97, 2013, pp. 723–729.

[Zheng et al., 2007] Zheng, Y., Steiner, K., Bauer, T., Yu, J., Shen, D., and Kambhamettu, C. "Lung Nodule Growth Analysis from 3D CT Data with a Coupled Segmentation and Registration Framework". In: *International Conference on Computer Vision –ICCV 2007*. 2007, pp. 1–8.

[Zheng et al., 2015] Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., Huang, C., and Torr, P. H. "Conditional Random Fields as Recurrent Neural Networks". In: *International Conference on Computer Vision –ICCV 2015*. 2015, pp. 1529–1537.

[Zhou et al., 2019] Zhou, Y., Graham, S., Alemi Koohbanani, N., Shaban, M., Heng, P.-A., and Rajpoot, N. "CGC-Net: Cell Graph Convolutional Network for Grading of Colorectal Cancer Histology Images". In: *Visual Recognition for Medical Images –ICCV 2019 Workshops*. 2019.

List of Publications

This list contains journal articles, conference papers and abstracts published or submitted during the work on this dissertation. An asterisk (*) indicates co-first authorship.

Journal articles as first author

- Hering*, A., Hansen*, L., Mok, T. C. W., Chung, A. C. S., Siebert, H., Häger, S., Lange, A., Kuckertz, S., Heldmann, S., Shao, W., Vesal, S., Rusu, M., Sonn, G., Estienne, T., Vakalopoulou, M., Han, L., Huang, Y., Yap, P.-T., Brudfors, M., Balbastre, Y., Joutard, S., Modat, M., Lifshitz, G., Raviv, D., Lv, J., Li, Q., Jaouen, V., Visvikis, D., Fourcade, C., Rubeaux, M., Pan, W., Xu, Z., Jian, B., Benetti, F. D., Wodzinski, M., Gunnarsson, N., Sjölund, J., Grzech, D., Qiu, H., Li, Z., Großbröhmer, C., Hoopes, A., Reinertsen, I., Xiao, Y., Landman, B., Huo, Y., Murphy, K., Ginneken, B., Dalca, A., and Heinrich, M. P. "Learn2Reg: Comprehensive Multi-Task Medical Image Registration Challenge, Dataset and Evaluation in the Era of Deep Learning". *IEEE Transactions on Medical Imaging*, 2022.
- Hansen, L. and Heinrich, M. P. "GraphRegNet: Deep Graph Regularisation Networks on Sparse Keypoints for Dense Registration of 3D Lung CTs". *IEEE Transactions on Medical Imaging* 40 (9), 2021, pp. 2246–2257.
- Hansen, L., Siebert, M., Diesel, J., and Heinrich, M. P. "Fusing Information From Multiple 2D Depth Cameras for 3D Human Pose Estimation in the Operating Room". *International Journal of Computer Assisted Radiology and Surgery* 14 (11), 2019, pp. 1871–1879.

Conference papers as first author

- Hansen, L. and Heinrich, M. P. "Revisiting Iterative Highly Efficient Optimisation Schemes in Medical Image Registration". In: *Medical Image Computing and Computer Assisted Intervention –MICCAI 2021*. 2021, pp. 203–212.
- Hansen, L. and Heinrich, M. P. "Deep Learning Based Geometric Registration for Medical Images: How Accurate Can We Get Without Visual Features?" In: *Information Processing in Medical Imaging –IPMI 2021*. 2021, pp. 18–30.

- Hansen, L. and Heinrich, M. P. “Discrete Unsupervised 3D Registration Methods for the Learn2Reg Challenge”. In: *Segmentation, Classification, and Registration of Multi-modality Medical Imaging Data –MICCAI 2020 Challenges*. 2021, pp. 68–73.
- Hansen, L., Dittmer, D., and Heinrich, M. P. “Learning Deformable Point Set Registration with Regularized Dynamic Graph CNNs for Large Lung Motion in COPD Patients”. In: *Graph Learning in Medical Imaging –MICCAI 2019 Workshops*. 2019, pp. 53–61.
- Hansen, L. and Heinrich, M. P. “Sparse Structured Prediction for Semantic Edge Detection in Medical Images”. In: *International Conference on Medical Imaging with Deep Learning –MIDL 2019*. 2019, pp. 250–259.
- Hansen, L., Diesel, J., and Heinrich, M. P. “Regularised Landmark Detection with CAEs for Human Pose Estimation in the Operating Room”. In: *Bildverarbeitung für die Medizin 2019 –BVM 2019*. 2019, pp. 178–183.
- Hansen, L., Diesel, J., and Heinrich, M. P. “Multi-Kernel Diffusion CNNs for Graph-Based Learning on Point Clouds”. In: *Geometry Meets Deep Learning –ECCV 2018 Workshops*. 2019, pp. 456–469.

Abstracts as first author

- Hansen, L., Hering, A., Großbröhmer, C., and Heinrich, M. P. “Continuous Benchmarking in Medical Image Registration - Review of the Current State of the Learn2Reg Challenge”. In: *International Conference on Medical Imaging with Deep Learning –Extended Abstract Track –MIDL 2022*. 2022.
- Hansen, L., Sieren, M., Hobe, M., Saalbach, A., Schulz, H., Barkhausen, J., and Heinrich, M. P. “Radiographic Assessment of CVC Malpositioning: How can AI best support clinicians?” In: *International Conference on Medical Imaging with Deep Learning –Extended Abstract Track –MIDL 2021*. 2021.
- Siebert*, H., Hansen*, L., and Heinrich, M. P. “Learning a Metric without Supervision: Multimodal Registration using Synthetic Cycle Discrepancy”. In: *International Conference on Medical Imaging with Deep Learning –Extended Abstract Track –MIDL 2021*. 2021.
- Hansen, L. and Heinrich, M. P. “Probabilistic Dense Displacement Networks for Medical Image Registration - Contributions to the Learn2Reg Challenge”. In: *Bildverarbeitung für die Medizin 2021 –Abstract Track –BVM 2021*. 2021, pp. 125–126.
- Hansen, L. and Heinrich, M. P. “Tackling the Problem of Large Deformations in Deep Learning Based Medical Image Registration Using Displacement Em-

beddings”. In: *International Conference on Medical Imaging with Deep Learning –Extended Abstract Track –MIDL 2020.* 2020.

- Hansen, L., Blendowski, M., and Heinrich, M. P. “In Defence of Mathematical Models for Deep Learning based Registration”. In: *Bildverarbeitung für die Medizin 2020 –Abstract Track –BVM 2021.* 2020, p. 32.

Journal articles as co-author

- Siebert, H., Hansen, L., and Heinrich, M. P. “Learning a Metric for Multimodal Medical Image Registration without Supervision Based on Cycle Constraints”. *Sensors* 22 (3), 2022, p. 1107.
- Bigalke, A., Hansen, L., Diesel, J., and Heinrich, M. P. “Seeing Under the Cover with a 3D U-Net: Point Cloud-Based Weight Estimation of Covered Patients”. *International Journal of Computer Assisted Radiology and Surgery* 16 (12), 2021, pp. 2079–2087.
- Blendowski, M., Hansen, L., and Heinrich, M. P. “Weakly-Supervised Learning of Multi-Modal Features for Regularised Iterative Descent in 3D Image Registration”. *Medical Image Analysis* 67, 2021, p. 101822.
- Bockelmann, N., Graßhoff, J., Hansen, L., Bellani, G., Heinrich, M. P., and Rostalski, P. “Deep Learning for Prediction of Diaphragm Activity from the Surface Electromyogram”. *Current Directions in Biomedical Engineering* 5 (1), 2019, pp. 17–20.

Conference papers as co-author

- Weihsbach, C., Hansen, L., and Heinrich, M. P. “XEdgeConv-Net: Leveraging Graph Convolutions for Efficient, Permutation- and Rotation-invariant Dense 3D Medical Image Segmentation”. In: *Geometric Deep Learning in Medical Image Analysis –GeoMedIA 2022.* 2022, pp. 61–71.
- Bigalke, A., Hansen, L., and Heinrich, M. P. “Adapting the Mean Teacher for Keypoint-based Lung Registration under Geometric Domain Shifts”. In: *Medical Image Computing and Computer Assisted Intervention –MICCAI 2022.* 2022, pp. 280–290.
- Falta, F., Hansen, L., and Heinrich, M. P. “Learning Iterative Optimisation for Deformable Image Registration of Lung CT with Recurrent Convolutional Networks”. In: *Medical Image Computing and Computer Assisted Intervention –MICCAI 2022.* 2022, pp. 301–309.

- Heinrich, M. P. and Hansen, L. “Voxelmorph++ Going Beyond the Cranial Vault with Keypoint Supervision and Multi-Channel Instance Optimisation”. In: *International Workshop on Biomedical Image Registration –WBIR 2022.* 2022.
- Bigalke, A., Hansen, L., Diesel, J., and Heinrich, M. P. “Domain Adaptation through Anatomical Constraints for 3D Human Pose Estimation under the Cover”. In: *International Conference on Medical Imaging with Deep Learning –MIDL 2022.* 2022, pp. 173–187.
- Siebert, H., Hansen, L., and Heinrich, M. P. “Fast 3D Registration with Accurate Optimisation and Little Learning for Learn2Reg 2021”. In: *Biomedical Image Registration, Domain Generalisation and Out-of-Distribution Analysis –MICCAI 2021 Challenges.* 2022, pp. 174–179.
- Falta, F., Hansen, L., Himstedt, M., and Heinrich, M. P. “Learning an Airway Atlas from Lung CT using Semantic Inter-Patient Deformable Registration”. In: *Bildverarbeitung für die Medizin 2022 –BVM 2022.* 2022, pp. 75–80.
- Graf, L., Mischkewitz, S., Hansen, L., and Heinrich, M. P. “Spatiotemporal Attention for Realtime Segmentation of Corrupted Sequential Ultrasound Data”. In: *Bildverarbeitung für die Medizin 2022 –BVM 2022.* 2022, pp. 235–240.
- Hermes, N., Hansen, L., Himstedt, M., Bigalke, A., and Heinrich, M. P. “Support Point Sets for Improving Contactless Interaction in Geometric Learning for Hand Pose Estimation”. In: *Bildverarbeitung für die Medizin 2022 –BVM 2022.* 2022, pp. 89–94.
- Bigalke, A., Hansen, L., and Heinrich, M. P. “End-to-end Learning of Body Weight Prediction from Point Clouds with Basis Point Sets”. In: *Bildverarbeitung für die Medizin 2021 –BVM 2021.* 2021, pp. 254–259.
- Kruse, C. N., Hansen, L., and Heinrich, M. P. “Multi-modal Unsupervised Domain Adaptation for Deformable Registration Based on Maximum Classifier Discrepancy”. In: *Bildverarbeitung für die Medizin 2021 –BVM 2021.* 2021, pp. 192–197.
- Siebert, H., Hansen, L., and Heinrich, M. P. “Architecture Matters: Evaluating Design Choices for Deep Learning Registration Networks”. In: *Bildverarbeitung für die Medizin 2021 –BVM 2021.* 2021, pp. 111–116.
- Heinrich, M. P. and Hansen, L. “Highly Accurate and Memory Efficient Unsupervised Learning-Based Discrete CT Registration Using 2.5D Displacement Search”. In: *Medical Image Computing and Computer Assisted Intervention –MICCAI 2020.* 2020, pp. 190–200.
- Wattenberg, M., Hansen, L., Klein, P., Heinrich, M. P., Stille, M., and Buzug, T. M. “Reconstruction of Blood Vessel Paths from Sparse Cone Beam Projection Images using Neural Networks and Ray-Tracing”. In: *International Conference*

on Image Formation in X-Ray Computed Tomography –CT Meeting 2020. 2020, pp. 482–485.

- Keuth, R., Hansen, L., and Heinrich, M. P. “Der Einfluss von Segmentierung auf die Genauigkeit eines CNN-Klassifikators zur Mimik-Steuerung”. In: *Bildverarbeitung für die Medizin 2020 –BVM 2020.* 2020, pp. 294–300.

Abstracts as co-author

- Bigalke, A., Hansen, L., and Heinrich, M. P. “A Novel Mean Teacher Framework for Domain Adaptive Lung Registration”. In: *International Workshop on Biomedical Image Registration –Extended Abstract Track –WBIR 2022.* 2022.
- Falta, F., Hansen, L., and Heinrich, M. P. “Learning Iterative Optimisation for Deformable Image Registration with Recurrent Convolutional Networks”. In: *International Workshop on Biomedical Image Registration –Extended Abstract Track –WBIR 2022.* 2022.
- Heinrich, M. P. and Hansen, L. “Unsupervised learning of multimodal image registration using domain adaptation with projected Earth Move’s discrepancies”. In: *International Conference on Medical Imaging with Deep Learning –Extended Abstract Track –MIDL 2020.* 2020.
- Ha, I. Y., Hansen, L., Wilms, M., and Heinrich, M. P. “Geometric Deep Learning and Heatmap Prediction for Large Deformation Registration of Abdominal and Thoracic CT”. In: *International Conference on Medical Imaging with Deep Learning –Extended Abstract Track –MIDL 2019.* 2019.