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

Prior-Guided 3D Deep Learning for Point Cloud Analysis in Medicine Under Domain Shifts

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
Alexander Bigalke
from Witten

Lübeck, 2023

First referee: Prof. Dr. Mattias P. Heinrich
Second referee: Prof. Dr.-Ing. Katharina Breininger

Date of oral examination: 06.05.2024

Approved for printing. Lübeck, 17.07.2024

Abstract

Automated computational image analysis offers promising opportunities for healthcare, such as assisting clinicians in patient monitoring and medical scan assessment. Over the past decade, deep learning has significantly advanced the field, especially the analysis of dense intensity data on regular grids. By contrast, processing sparse 3D point clouds, captured by depth sensors in monitoring systems or extracted as descriptive keypoints from medical 3D scans, has received little attention yet. Given the potential benefits of the sparse geometric representation, including anonymity preservation, robustness to intensity variations, and computational efficiency, geometric deep learning on point clouds has great potential to play an increasingly important role in medical image analysis. However, various research questions, such as solving medicine-specific problems and investigating and tackling domain gaps, remain underexplored compared to dense image processing. This thesis addresses these shortcomings in three ways.

First, various point cloud-based deep learning methods for a heterogeneous set of medical tasks are developed, namely dynamic hand gesture recognition and in-bed body weight and pose estimation as monitoring tasks, and lung registration as a classical medical image analysis problem. The investigated tasks thus cover local detection and global regression/classification problems while considering single frames and temporal sequences as inputs. Promising experimental results across all these tasks demonstrate the versatile potential of point cloud-based approaches in medicine.

Second, the thesis investigates the impact of domain gaps, revealing a significant sensitivity of point cloud networks to geometric domain shifts, and, in response, develops multiple novel domain adaptation strategies. An anatomy-guided constrained optimization scheme is presented for pose estimation, and the Mean Teacher paradigm is adapted and significantly extended for domain adaptive point cloud registration. Both approaches improve upon the current state of the art in comprehensive evaluations.

Third, the thesis explores how solving the above problems can benefit from incorporating task-specific prior knowledge in various forms. On the one hand, the work derives an improved point cloud encoding scheme and a novel loss function from priors on the input and output distribution. On the other, it draws inspiration from human behavior and reasoning to develop novel learning paradigms and suitable two-stage and two-stream model architectures. Thorough experiments demonstrate significant advantages of these solutions over end-to-end approaches without explicit priors.

Overall, the developed geometric deep learning models and domain adaptation strategies successfully address diverse problems in medical image processing and thus significantly contribute to advancing point cloud analysis in medicine.

Zusammenfassung

Die automatisierte computergestützte Bildanalyse bietet vielversprechende Möglichkeiten für den Gesundheitssektor, wie z. B. die Unterstützung von Ärzten bei der Patientenüberwachung und Beurteilung medizinischer Scans. Deep Learning hat das Forschungsfeld in den letzten Jahren erheblich vorangebracht und insbesondere die Analyse dichter Intensitätsdaten mit regelmäßiger Gitterstruktur stark verbessert. Die Verarbeitung spärlicher 3D-Punktwolken, die von Tiefensensoren in Überwachungssystemen aufgenommen oder als deskriptive Keypoints aus medizinischen 3D-Scans extrahiert werden können, wurde dagegen bisher weniger beachtet. Angesichts der potenziellen Vorteile der spärlichen geometrischen Darstellung, die u. a. Anonymitätserhaltung, Robustheit gegenüber variierenden Intensitäten und höhere Recheneffizienz verspricht, hat geometrisches Deep Learning auf Punktwolken großes Potenzial, eine zunehmend wichtige Rolle in der medizinischen Bildanalyse zu spielen. Allerdings sind diverse Forschungsfragen, z. B. die Lösung von medizinspezifischen Problemen und die Untersuchung und Bewältigung von Domain-Shifts, im Vergleich zur Verarbeitung dichter Bilddaten noch wenig erforscht. Diese Arbeit geht diese offenen Probleme auf dreierlei Weise an.

Erstens werden punktwolkenbasierte Deep-Learning-Methoden für eine Reihe verschiedenartiger medizinischer Aufgaben entwickelt: die Schätzung von Körpergewicht und -pose im Bett liegender Patienten und die dynamische Handgestenerkennung im Kontext von Überwachungssystemen sowie Lungenregistrierung als klassisches medizinisches Bildverarbeitungsproblem. Die untersuchten Aufgaben decken somit lokale Detektions- und globale Regressions-/Klassifikationsprobleme ab, wobei sowohl Einzelbilder als auch zeitliche Sequenzen als Eingaben betrachtet werden. Vielversprechende experimentelle Ergebnisse für alle diese Aufgaben zeigen das vielseitige Potenzial von punktwolkenbasierten Ansätzen in der Medizin.

Zweitens untersucht die Arbeit die Auswirkungen von Domain-Gaps, zeigt dabei eine erhebliche Anfälligkeit von Punktwolkennetzwerken gegenüber geometrischen Domain-Shifts auf und entwickelt als Gegenmaßnahme mehrere neue Methoden zur Domain-Adaptation. Ein auf der menschlichen Anatomie basierendes Optimierungsverfahren wird für die Posenschätzung präsentiert und die Mean-Teacher-Strategie wird für die Punktwolkenregistrierung angepasst und signifikant erweitert. Beide Ansätze zeigen sich dem aktuellen Stand der Technik in umfassenden Evaluationen überlegen.

Drittens wird untersucht, wie die Lösung der obigen Probleme von der Einbeziehung aufgabenspezifischen Vorwissens in verschiedenen Formen profitieren kann. Zum einen werden ein verbessertes Enkodierungsschema für Punktwolken und eine neue Ver-

lustfunktion aus a priori Wissen über die Verteilung von Ein- und Ausgabedaten abgeleitet. Zum anderen dienen menschliche Verhaltens- und Denkweisen als Inspiration zur Entwicklung neuartiger Lernstrategien und geeigneter Zwei-Schritt- und Zwei-Stream-Modellarchitekturen. Die signifikanten Vorteile dieser Lösungen gegenüber Ende-zu-Ende-Ansätzen ohne explizites Vorwissen werden in ausgiebigen Experimenten demonstriert.

Insgesamt stellen die entwickelten geometrischen Deep-Learning-Modelle und Domain-Adaptation-Strategien erfolgreiche Lösungsansätze für diverse Probleme der medizinischen Bildverarbeitung dar und leisten somit einen wichtigen Beitrag zum Fortschritt der Punktwolkenanalyse in der Medizin.

Contents

Chapter 1

Introduction

1.1 Motivation

Automatic computer-assisted image processing offers promising opportunities in the healthcare sector and could become a key component in managing the increasing patient volumes of our aging population while maintaining or even improving the quality of care and treatment. On the one hand, algorithms for medical image analysis can support physicians in assessing medical imaging data, such as MRI, CT, ultrasound, or X-ray, to perform diagnosis [Rajpurkar et al., 2018], disease monitoring [Taylor et al., 2019], or treatment planning [Wong et al., 2020], among others. On the other hand, camera-based monitoring systems can relieve clinical staff of routine jobs and improve patient safety by automating documentation [Padoy, 2019], detecting critical events [Jähne-Raden et al., 2019], or diagnosing pathological behavior [Cunha et al., 2016].

Over the past decade, deep learning has revolutionized the field of automatic image analysis and, to date, dominates the state of the art for most tasks in medical image analysis [Chen et al., 2022b; Litjens et al., 2017]. Convolutional neural networks (CNNs) have played a crucial role in this development since their fundamental operation – the convolution – is naturally tailored to processing the predominant ordered 2D or 3D grid structure of medical imaging data and camera images. However, while enabling the success of CNNs, the intensity-based grid representation also involves several downsides, including memory- and computation-intensive dense data processing, privacy concerns, and a potential intensity/texture bias of trained networks.

3D point clouds are an alternative sparse and purely geometric input representation, promising to alleviate the deficiencies of grid data. Based on the discussed applications of medical image analysis, such point clouds can be generated in two different ways (see Fig. 1.1). On the one hand, they can represent distinctive keypoints extracted from volumetric medical scans. In the context of monitoring systems, on the other hand, modern depth sensors can directly capture the 3D geometric surfaces in an observed scene, which are, in turn, naturally represented as point clouds. Both forms of point clouds unify multiple benefits over grid data. 1) Point cloud-based algorithms have inherent robustness to variations in intensity distributions, as commonly encountered across clinical sites, while intensity-based CNNs often incur severe performance degra-

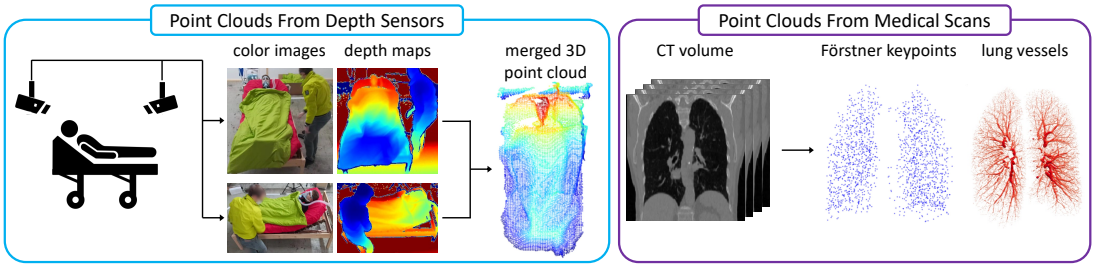


Fig. 1.1: Exemplary visualization of point cloud acquisition in medical image analysis. Left: Depth sensors in monitoring systems directly capture the 3D scene geometry, which is represented as a merged 3D point cloud in world coordinates. Right: There are different approaches to extracting sparse distinctive keypoints from dense medical imaging data, e.g., the Förstner operator or a lung vessel filter. The examples highlight several potential benefits of point clouds over the corresponding intensity/color-based grid data, including anonymity preservation, insensitivity to intensities, a computationally more efficient representation, and the natural fusion of data from multiple sensors.
Image sources: CT volume: [Castillo et al., 2013], lung vessel trees: [Shen et al., 2021].

dation. 2) The purely geometric representation can be considered anonymity-preserving [Silas et al., 2015], thus preventing privacy concerns of patients and clinical staff and facilitating data access and sharing. 3) The sparsity of point clouds makes them a more efficient data structure than dense grids. For instance, volumetric medical CT or MR scans with millions of voxels can be represented by point clouds with a few thousand keypoints. This reduction of data points can potentially translate to lower memory requirements and faster runtimes (depending on the available hardware). Hereafter focusing on monitoring systems, depth sensor-based point clouds offer further advantages over ordinary camera images. 1) Point clouds preserve the natural 3D structure of the original scene instead of projecting it to a 2D plane. 2) Depth sensors are virtually illumination-independent, while classical cameras require sufficient lighting and do not provide any meaningful information in the dark. 3) Point clouds from multiple depth sensors can naturally be fused in the world coordinate space, enabling the acquisition of complete scene geometries even under occlusions.

In light of these potential benefits, deep learning-based processing of point clouds, a branch of geometric deep learning [Bronstein et al., 2017], has been a popular research topic in recent years [Guo et al., 2020]. The particular challenge of this task resides in the irregular and unordered nature of point clouds, preventing the use of ordinary convolutions or fully connected layers. In response, researchers have developed generic convolution operators applicable to the irregular domain. Graph convolutions, for instance, overcome the irregularity of spatial neighborhoods through edge-wise feature extraction on the nearest neighbor graph and address the non-existent order of the input points through permutation-invariant aggregation functions [Bronstein et al., 2021].

These developments have been coupled with the design of new network architectures, culminating in performant deep learning approaches for point cloud analysis [Wei et al., 2023; Xiang et al., 2021; Xu et al., 2021].

Despite this progress, point cloud-based deep learning, with the seminal work by Qi et al. [2017a], is around five years behind CNN-based processing of grid data, where the famous AlexNet of Krizhevsky et al. [2012] accomplished a breakthrough. Consequently, several research questions are still considerably less explored than in dense image processing, which is, among others, noticeable in two decisive aspects. First, existing point cloud networks are primarily designed for and evaluated on standard computer vision tasks and datasets (e.g., 3D object classification [Wu et al., 2015], indoor scene segmentation [Armeni et al., 2016], or object detection in autonomous driving [Zhou et al., 2018]), while medical problems and applications remain largely unexplored. Second, these works mainly focus on fully-supervised learning with identically distributed training and test data from a single domain and rarely consider distribution shifts between different training and test domains as frequently encountered in practice. This is particularly critical since image-based studies have demonstrated poor cross-domain generalization of deep neural networks, suffering significant performance drops under domain shifts [Hendrycks et al., 2019]. Unlike image-based models, point cloud networks are insensitive to intensity shifts but can still be highly vulnerable to geometric shifts of the point distribution in 3D space [Ren et al., 2022], which can, e.g., result from varying viewpoints, sensor properties, or object shapes/geometries in different domains. Domain adaptation [Guan et al., 2021; Wang et al., 2018] is a promising approach to overcome this problem by adapting a model to a shifted target domain in an unsupervised manner. But while extensively explored for image-based methods, domain adaptation for point cloud-based neural networks is still at an early stage, again mainly limited to standard computer vision tasks and barely explored in the medical field.

In conclusion, point clouds feature several beneficial properties for 3D image analysis in medicine, rendering their deep learning-based analysis a promising research direction. Integrating the approach into clinical practice, however, still requires addressing multiple open challenges.

1.2 Objectives

The overall goal of this work is to advance deep learning-based point cloud analysis in medicine by addressing the previously identified open challenges. To this end, the thesis pursues three fundamental objectives: 1) Develop point cloud-based deep learning solutions for medical tasks and applications. 2) Investigate domain shifts and develop domain adaptation strategies. 3) Explore the incorporation of prior domain knowledge as a central theme in addressing the first two goals. These objectives are specified in more detail below.

Medical tasks and applications. To demonstrate the versatile potential of point cloud-based approaches, the first objective of this work is to develop deep learning solutions to a comprehensive set of heterogeneous medical imaging tasks, including global recognition and local detection tasks, single-frame and sequence analysis, and processing of both sensor- and image-based point clouds.

Since point clouds, in their most natural occurrence, originate from depth-sensing devices, as usable in clinical monitoring systems, the primary application focus of this work is the analysis of such point clouds, studied for the use case of in-bed patient monitoring. Among the broad range of sub-tasks in patient monitoring, the work focuses on two selected tasks with distinct characteristics: body weight estimation as a global regression task and body pose estimation as a local detection task. While these two tasks are addressed in a static (frame-by-frame) manner, medical imaging in general, and patient monitoring in particular, are often dynamic problems that can benefit from or even require the incorporation of temporal information. Therefore, point cloud-based sequence analysis is another focus of this work, again addressed at global and local levels by considering sequence classification and registration. These tasks are studied in a broader application context, namely for dynamic hand gesture recognition and CT-based lung registration. While the former, with applications like nurse calling and touchless gesture control of room lights, is still relevant to clinical monitoring systems, the latter extends the scope of the work from purely sensor-based point clouds as input data to keypoints from medical scans.

Domain shifts and domain adaptation. In practical clinical scenarios, training and test data frequently exhibit deviating data distributions, e.g., when captured at different clinical sites or in case of training on simulated or lab data. Therefore, to approach the practical deployment of the methods envisioned in the first objective, the second goal of this work is to address deep learning-based point cloud analysis under such domain shifts. The intended contribution is two-fold: First, it is essential to thoroughly analyze and quantify the effect of geometric domain shifts between training and test data on point cloud-based deep learning methods by evaluating the cross-domain performance of established models. For a comprehensive analysis, the goal is to consider diverse clinically relevant domain shifts with heterogeneous characteristics (synthetic-to-real and across clinical sites) and examine their impact on the various tasks discussed above (single-frame and sequence analysis of sensor- and image-based point clouds). Second, building on the previous findings, this work aims to develop novel domain adaptation methods to overcome the potential adverse effects of the studied domain shifts. This includes adjusting and optimizing existing image-based adaptation methods, such as adversarial feature and output space alignment and self-training, to the specific characteristics of the considered medical tasks and point cloud modalities but also devising completely novel strategies and learning paradigms.

Prior domain knowledge. Appropriate incorporation of prior domain knowledge is essential for developing deep learning methods. A prominent example is the incorporation of geometric characteristics of the input domain into neural architectures [Bronstein et al., 2017], commonly referred to as inductive bias. For instance, the shared weights of convolutional filters reflect a translational symmetry of the input data, and the permutation invariance of point clouds is integrated through suitable aggregation operators. These priors can be considered generic priors on the structure of the input domain, independent of the specific addressed task. Considering the previously discussed objectives of addressing specific novel tasks under different learning scenarios, this work focuses, in addition, on incorporating prior knowledge about the concrete problem to be solved. More precisely, the third objective of this work is to explore the incorporation of prior task-specific domain knowledge to develop novel point cloud-based deep learning models and training strategies under domain shifts. In this context, 'prior knowledge' is a rather broad term, motivating the investigation of various forms of priors throughout this work. In a narrow, more technical sense, explored priors include expected input and output distributions and well-known inductive biases of existing point cloud architectures. In a more far-reaching sense, human reasoning, procedures, and behavior in solving similar tasks serve as inspiration. On the deep learning side, these priors stimulate methodological research at diverse levels, including the model architecture (two-stage, two-stream), the form of supervision (novel loss function, intermediate supervision), and novel learning paradigms.

1.3 Organization and Contributions

The remainder of the thesis is divided into six chapters that successively elaborate on the above research objectives. First, Chapter 2 gives background information on point cloud acquisition and provides the methodological foundations of deep learning-based point cloud analysis and domain adaptation. Subsequently, as the core component of the thesis, Chapters 3–6 present new methodological developments. Finally, Chapter 7 summarizes the main findings, discusses them with respect to the three research objectives, and concludes the work with an outlook to future research.

The four methodological chapters follow a uniform structure. Each starts with a concise summary, outlining the chapter topic, contributions, and relevance to the thesis. The individual sections then present the developed methods in a self-contained manner, including 1) an introduction, which motivates the specific addressed problem, discusses related work, and highlights the scientific contributions, 2) a detailed formal description of the method, 3) a presentation of the performed experiments and observed results, 4) a discussion and conclusion.

A graphical overview of the four main chapters is shown in Fig. 1.2, schematically linking them to the principal research questions. On the one hand, the figure visualizes

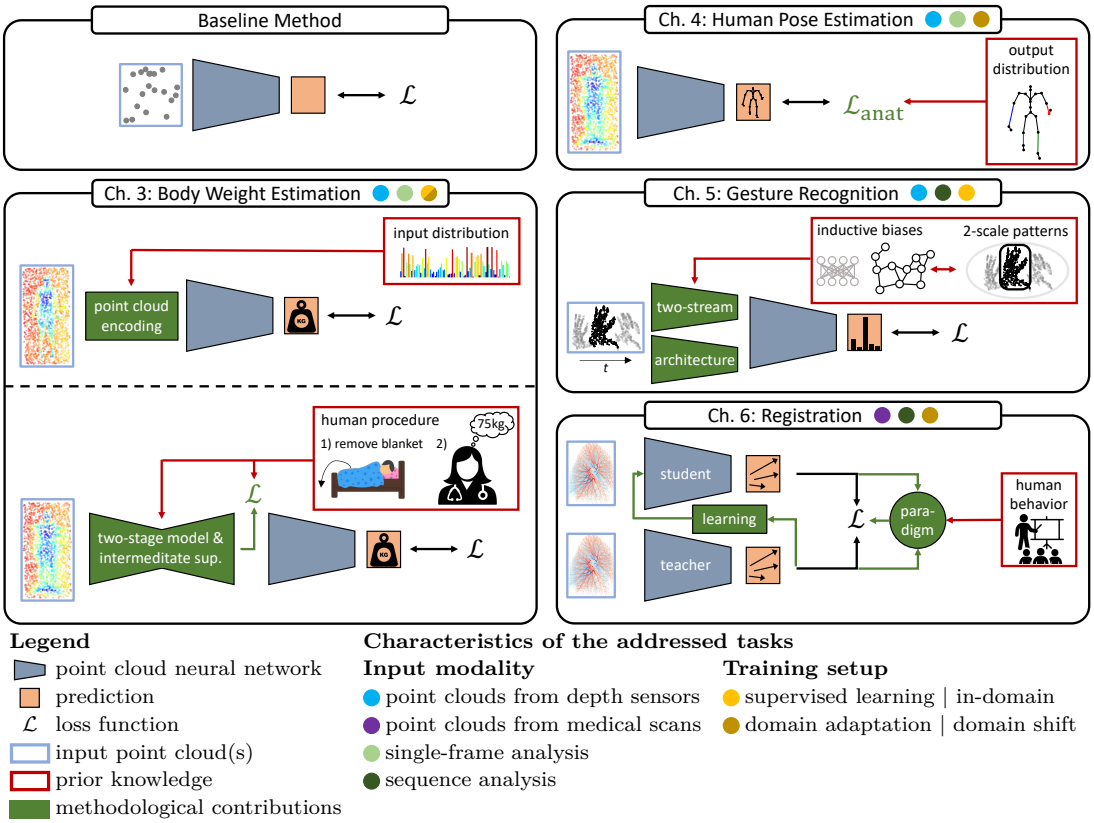


Fig. 1.2: Schematic overview of the four methodological chapters. The central overarching theme is the incorporation of prior knowledge to develop novel models, supervision strategies, and learning paradigms. The figure also highlights the transition from supervised single-frame processing to domain adaptive sequence analysis.

the explored prior knowledge (red) and resulting methodical developments (green). On the other hand, it highlights the transition from supervised learning on single frames in Chapter 3 over single-frame unsupervised domain adaptation (Chapter 4) and supervised sequence analysis (Chapter 5) to domain adaptive sequence analysis in Chapter 6. The main contributions of each chapter are outlined below.

- Chapter 3 deals with automated in-bed body weight estimation from point clouds, initially for fully-visible uncovered patients, and later for occluded patients under a blanket. As for the former, the chapter demonstrates, for the first time, the suitability of end-to-end deep learning solutions by adapting the concept of basis point sets as an efficient encoding scheme. The method is further optimized for the given problem by incorporating prior knowledge about the expected input distribution. As for the latter, inspired by the intuitive human approach to the task, a two-step method is developed, which virtually uncovers the patient before performing the

weight estimate, thus significantly reducing the complexity of the problem. Finally, the chapter analyzes weight estimation in a cross-domain setup, demonstrating the significance of geometric domain shifts and the consequent need for appropriate domain adaptation methods. The presented methods were published in:

[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 – BVM 2021.* 2021, pp. 254–259.
BVM Award for the third best scientific paper.

[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.
Invited BVM 2021 special issue paper – IF: 3.421.

- Chapter 4 shifts from supervised learning in Chapter 3 to unsupervised domain adaptation, addressed in the context of point cloud-based 3D in-bed human pose estimation. It investigates the exploitation of prior knowledge about the expected output distribution, i.e., the space of anatomically plausible human poses, to guide the adaptation process, yielding two complementary anatomy-guided adaptation strategies. First, a novel anatomical loss function is designed to explicitly constrain predictions to the plausible pose space by penalizing anatomically implausible poses. Second, pseudo labels for self-training under the Mean Teacher paradigm are filtered according to their anatomical plausibility. Experiments confirm the efficacy of both strategies, which outperform a comprehensive set of competing state-of-the-art adaptation methods adapted to the problem. The described methods were published in the following two works:

  [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.
  Oral presentation (12% of all submissions) – 50% overall acceptance rate.

  [Bigalke et al., 2023a] Bigalke, A., Hansen, L., Diesel, J., Hennigs, C., Rostalski, P., and Heinrich, M. P. "Anatomy-Guided Domain Adaptation for 3D In-Bed Human Pose Estimation". *Medical Image Analysis*, 2023, p. 102887.
  Invited MIDL 2022 special issue paper – IF: 13.828.

- Chapter 5 expands the focus from static frame-by-frame analysis in the previous chapters to temporal sequence analysis, studied at the example of point cloud-based dynamic hand gesture recognition. The chapter focuses on learning multi-scale

features, capturing global hand movements and local hand posture variations, which – following intuitive human reasoning – are the discriminative patterns in this task. This is achieved through a two-stream model, integrating two complementary 3D learning architectures with different characteristics and inductive biases, which thus inherently focus on the desired local/global features by network design. The approach is the top-performing method on two public benchmarks to date and was published in:

[Bigalke et al., 2021c] Bigalke, A. and Heinrich, M. P. "Fusing Posture and Position Representations for Point Cloud-Based Hand Gesture Recognition". In: *International Conference on 3D Vision – 3DV 2021.* 2021, pp. 617–626.
44% overall acceptance rate.

- Unifying the research topics of Chapters 4 and 5, Chapter 6 investigates sequence analysis under domain shifts at the example of point cloud-based medical image registration. The contributions of the chapter are two-fold. First, considering the lack of domain adaptive registration methods in the literature, the concept of self-training with the Mean Teacher is adapted to the specific characteristics of the registration task. In particular, this includes the innovative integration of combined feature learning and differentiable optimization into the Mean Teacher. The second part addresses the inherent noise in the pseudo labels from the teacher, which is a critical limitation of the method. Inspired by the human understanding of an optimal teacher-student relationship, two denoising learning paradigms are devised, including a novel filtering strategy and the dynamic synthesis of noise-free samples, which significantly boost performance and yield state-of-the-art performance for point cloud-based lung registration. The methods were published in:

  [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: *International Conference on Medical Image Computing and Computer Assisted Intervention – MICCAI 2022: Proceedings, Part VI.* 2022, pp. 280–290.
  31% overall acceptance rate.

  [Bigalke et al., 2023c] Bigalke, A. and Heinrich, M. P. "A Denoised Mean Teacher for Domain Adaptive Point Cloud Registration". In: *International Conference on Medical Image Computing and Computer Assisted Intervention – MICCAI 2023.* 2023, to appear.
  Early accepted (14% of all submissions) – 32% overall acceptance rate.

Chapter 2

Background

This chapter presents the fundamental concepts underlying the following methodological chapters. First, Sec. 2.1 deals with the acquisition of 3D point clouds and presents the different techniques used throughout this work. Next, Sec. 2.2 provides the essential concepts of deep learning, including a general introduction to machine learning and deep learning with CNNs and a more specific discussion of geometric deep learning for point cloud analysis. Finally, Sec. 2.3 formalizes the domain adaptation problem and gives an overview of different approaches to the problem.

2.1 Point Cloud Acquisition

Medical image analysis typically considers 3D intensity volumes $\boldsymbol{I} \in \mathbb{R}^{H \times W \times D}$ (e.g., from CT or MR scanners) or 2D intensity images $\boldsymbol{I} \in \mathbb{R}^{H \times W \times C}$ with $C = 1$ (e.g., X-ray or ultrasound) or $C = 3$ (color images from ordinary cameras) as input data. In other words, inputs represent observed intensity values on a dense regular 2D or 3D grid. By contrast, this work focuses on processing sparse, irregular, and purely geometric 3D point clouds $\mathcal{P} = \{\boldsymbol{p}_i | i = 1, ..., N\} \subset \mathbb{R}^3$, whose N points $\boldsymbol{p}_i = (x_i, y_i, z_i)$ are represented by their 3D coordinate vector and usually stored in a matrix $\boldsymbol{P} \in \mathbb{R}^{N \times 3}$. In Chapters 3, 4, and 5 of this work, the input clouds represent the 3D surfaces of objects in a scene as captured by a depth sensor. The underlying acquisition process is described in Sec. 2.1.1. The point clouds in Chapter 6, on the other hand, represent the voxel coordinates of distinctive 3D keypoints in dense image volumes. The algorithms used to extract these keypoints are presented in Sec. 2.1.2. Lastly, Sec. 2.1.3 introduces multiple employed algorithms to subsample point clouds.

2.1.1 Point Clouds From Depth Sensors

There are several technologies for depth measurements in a scene, such as stereo vision, structured light sensors, laser range finders/LiDAR, or time-of-flight (TOF) sensors. Point clouds in this work are predominantly based on TOF measurements. As the name suggests, TOF sensors perform depth measurements by emitting short pulses of infrared light and measuring the traveling time of the light from the sensor to a surface

and back. That way, the sensor captures a depth map $\boldsymbol{D} \in \mathbb{R}^{H \times W}$ with each pixel $D_{i,j}$ representing the distance from the camera to the surface at the corresponding location in the scene. Given the internal camera parameters, focal length (f_x, f_y) and coordinates of the principal point (c_x, c_y), each pixel can be lifted to a 3D point in the camera coordinate system as $\boldsymbol{p}^c = ((i - c_x) \cdot D_{i,j}/f_x, (j - c_y) \cdot D_{i,j}/f_y, D_{i,j})$. Moreover, given the external camera parameters, namely position $\boldsymbol{t} \in \mathbb{R}^3$ and rotation $\boldsymbol{R} \in \mathbb{R}^{3\times3}$ relating to a world coordinate system, a point in camera coordinates can be transformed into world coordinates as $\boldsymbol{p}^w = \boldsymbol{R}\boldsymbol{p}^c + \boldsymbol{t}$. Assuming multiple depth sensors to capture a scene from different viewpoints, it is then feasible to calibrate the sensors to the same world coordinate system and merge the captured point clouds in this shared space. Such a sensor setup can capture more complete surface geometries even under severe (self-)occlusions.

2.1.2 Image-Based Keypoint Extraction

Distinctive keypoints in 3D medical images can be derived in diverse ways. One intuitive approach is to sparsely sample points from surfaces of interest, as realized for abdominal organs in [Joutard et al., 2022] or the prostate in [Baum et al., 2021]. However, the segmentation of these structures is a non-trivial problem by itself, often requiring learning-based methods or manual annotations. Instead, keypoints in this work are derived with classical non-learning algorithms, focusing on lung CT images as the input. Exemplary visualizations of the extracted Förstner and lung vessel keypoints are shown in Fig. 1.1, while the functionality of the two underlying algorithms is described below.

2.1.2.1 Förstner Keypoints

The Förstner operator [Förstner et al., 1987] is an algorithm for detecting distinct points and corners in images based on image gradients. Sec. 6.1 of this work uses the implementation by Heinrich et al. [2015] who adapted the algorithm to lung CT scans. Given a 3D image volume $\boldsymbol{I}$ with voxel indices $\boldsymbol{i}$, a voxel-wise distinctiveness measure $D(\boldsymbol{i}) = 1/\text{trace}((\boldsymbol{G}_\sigma * (\nabla\boldsymbol{I}(\nabla\boldsymbol{I})^\top))^{-1})$ is calculated based on multiplied image gradients $\nabla\boldsymbol{I}$ that are smoothed with a 3D Gaussian kernel $\boldsymbol{G}_\sigma$. To ensure selected interest points are evenly distributed across the entire volume, the method does not extract the points with the overall highest scores but rather those representing local maxima, determined by non-maximum suppression.

2.1.2.2 Lung Vessel Trees

A different form of lung keypoints, used in Sec. 6.2, are sparse point clouds of the lung vessel trees. Shen et al. [2021] extracted them from lung CT scans in two steps. First, an initial lung vessel point cloud is generated with the Frangi filter [Frangi et al., 1998]. Second, the initial cloud is fit to the original CT volume by jointly minimizing

an inter-particle regularization energy and a particle-image fidelity, yielding roughly evenly distributed points along the centerlines of the vessels.

2.1.3 Subsampling of Point Clouds

Both depth sensor- and image-based point clouds often contain several 10k to 100k points. For computationally more efficient processing, subsampling algorithms reduce the clouds to a subset $\mathcal{P}' \subset \mathcal{P}$ with $N' \ll N$ points, ideally with a minimal loss of geometric information. A basic standard algorithm is uniform sampling, which randomly selects the points with equal probability. While sufficient for point clouds with relatively uniform density across space (as the sensor-based clouds in Chapters 3, 4, and 5), point clouds with non-uniform density will suffer a severe information loss in low-density areas. In this case, farthest point sampling and voxel-based downsampling ensure a more even space coverage. The former starts with a random point and then iteratively adds the point with the largest distance to the set of currently selected points. The latter divides the 3D space into equally-sized voxels and selects one representative point in each non-empty voxel. A downside of these two algorithms is that they still do not consider the density distribution and might thus ignore the most descriptive points (such as the bifurcations in lung vessel trees). For this reason, the lung vessel trees in Chapter 6 are downsampled with a density-aware algorithm. Based on point-wise Gaussian density estimates $d_i = 1/N \sum_{\boldsymbol{p_j} \in \boldsymbol{P}} \exp(-(\boldsymbol{p_i} - \boldsymbol{p_j})^2/\sigma^2)$ for all $\boldsymbol{p_i} \in \mathcal{P}$, non-maximum suppression selects those points representing local maxima in the density distribution.

2.2 Deep Learning

The general goal of supervised machine learning (ML) is to learn a function f^*, which maps an input $\boldsymbol{x} \in \mathbb{R}^d$ to an output $\boldsymbol{y} = f^*(\boldsymbol{x}) \in \mathbb{R}^k$, based on a training dataset $\mathcal{S} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{|\mathcal{S}|}$ of pairs of observed inputs $\boldsymbol{x}_i$ and desired outputs/ground truth $\boldsymbol{y}_i$. A standard ML approach to the problem approximates f^* by optimizing the parameters $\boldsymbol{\theta} \in \mathbb{R}^p$ of a flexible function $f(\cdot; \boldsymbol{\theta})$ to minimize a task-specific loss function

$$\mathcal{L}_{\text{task}}(\mathcal{S}; \boldsymbol{\theta}) = \sum_{i=1}^{|\mathcal{S}|} \text{dst}(f(\boldsymbol{x}_i; \boldsymbol{\theta}), \boldsymbol{y}_i) \tag{2.1}$$

where $\text{dst}(\cdot, \cdot)$ is a suitable distance measure between the predictions and ground truth. The current state of the art for many machine learning problems is dominated by deep learning (DL) approaches, which model the function f as a deep neural network. For a comprehensive introduction to this field, the reader is referred to Goodfellow et al. [2016], while the essential concepts are sketched throughout this section.

The fundamental component of neural networks is the artificial neuron. Characterized by a learnable weight vector $\boldsymbol{w} \in \mathbb{R}^d$ and bias $b \in \mathbb{R}$, it transforms an input vector $\boldsymbol{x}$

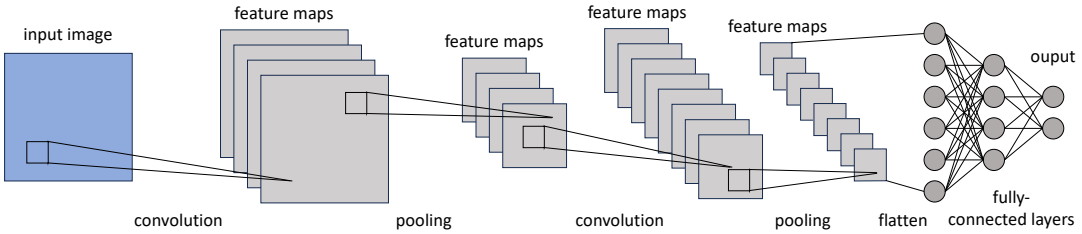


Fig. 2.1: Visualization of the typical architecture of a 2D CNN, mapping a 2D input image to a two-dimensional output vector. First, alternating convolution and pooling layers extract feature maps of decreasing resolution at an increasing depth. Second, the final feature maps are flattened to a vector, processed by the fully-connected layers.

with the function $\varphi(\boldsymbol{x}; \boldsymbol{w}, b) = \sigma(\boldsymbol{w}^\top \boldsymbol{x} + b)$, where σ is a non-linear activation function, such as the rectified linear unit (ReLU) with $\sigma(x) = \max(0, x)$. The simplest way to form a neural network from such neurons is the fully-connected network or multi-layer perceptron (MLP), where the neurons are arranged in multiple layers that are stacked on top of each other and connected by the learnable weights of the neurons, as visualized in the right part of Fig. 2.1. The distinct feature of this architecture is that each neuron has its own weight vector and is connected to all neurons in the previous layer. While such a network can – in theory – approximate arbitrary functions [Hornik et al., 1989], it cannot adequately handle unordered permutation-invariant inputs like point clouds, and its full connectivity leads to high parameter counts, in particular for high-dimensional inputs like images, complicating optimization. Therefore, the ideal design of neural networks has remained an active research field over the recent years and strongly depends on the modality and structure of the considered input data. Before discussing deep networks for point cloud analysis, the main focus of this work, in Sec. 2.2.2, Sec. 2.2.1 is dedicated to CNNs, which significantly advanced the state of the art in (medical) image analysis over the past decade.

2.2.1 Convolutional Neural Networks

CNNs are tailored to processing regular grid data $\boldsymbol{I}$ and model the function f by extracting feature maps with alternating convolutional and pooling layers while only using a few fully-connected layers at the very end for the final prediction, as exemplary illustrated in Fig. 2.1. Convolutional layers address the limitations of fully-connected layers by the concepts of local connectivity and weight sharing. More precisely, neurons of convolutional layers are only connected to a local patch of their input and share their weights with all other neurons in the layer, which substantially reduces the number of network parameters and directly builds the translational symmetry of visual problems into the network. From another perspective, a convolutional layer comprises multiple learnable local kernels/filters that are separately slid across the input to create

so-called feature maps, representing the kernel responses across space. Typically, these filters learn to detect low-level features like edges or corners at early layers and more abstract features in deeper layers. Formally, the operation corresponds to a discrete convolution, which, for a 2D input $\boldsymbol{I} \in \mathbb{R}^{H \times W \times C}$ with C channels and a local kernel $\boldsymbol{G} \in \mathbb{R}^{2K+1 \times 2K+1 \times C}$ with spatial size $2K+1$, is defined as

$$(\boldsymbol{I} * \boldsymbol{G})_{i,j} = \sum_{k,l=-K}^{K} \sum_{c=1}^{C} I_{i+k,j+l,c} G_{k,l,c} \tag{2.2}$$

The resulting response maps from all kernels are then stacked to a volume of feature maps. Every few convolutional layers are followed by pooling layers that downsample the feature maps to smaller resolutions through patch-wise max or average operations. That way, pooling improves computational efficiency, increases the receptive field, and makes the network invariant to small variations of feature locations. Modern CNNs also often include further components like Dropout [Srivastava et al., 2014], normalization layers [Ioffe et al., 2015; Wu et al., 2018], and residual [He et al., 2016] or dense [Huang et al., 2017] connections to ease optimization and/or improve model generalization. Given these building blocks, the parameters $\boldsymbol{\theta}$ of the CNN can be learned in a data-driven way by computing the partial derivatives $\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\text{task}}$ of the loss with respect to the parameters with the help of the backpropagation algorithm and moving parameters in the opposite direction of the gradients. Since the gradients are typically estimated on mini-batches of limited size and hence noisy, advanced optimization algorithms were developed to make the optimization more robust and accelerate convergence. Adam [Kingma et al., 2015], for instance, computes adaptive step sizes (learning rates) for each parameter based on the exponential moving averages of the first and second momentum of the gradients.

2.2.2 Geometric Deep Learning for Point Cloud Analysis

CNNs, as introduced in the previous section, require grid-structured data $\boldsymbol{I}$ as input and cannot directly process unordered and unstructured point clouds $\mathcal{P}$, for which the discrete convolution in Eq. (2.2) is not well defined. One option to still process point clouds with a CNN is to voxelize them into a regular 3D grid by discretizing the 3D space into fixed-size voxels and setting each voxel that contains at least one point of the cloud to one and all others to zero [Maturana et al., 2015]. While the resulting binary 3D volume can be processed by a 3D CNN, the representation involves a loss of spatial information due to quantization artifacts, and it is inefficient as computational and memory demands cubically increase with the resolution. Instead, it is desirable to directly process the raw point clouds by generalizing neural networks to irregular domains, as pursued by methods from geometric deep learning [Bronstein et al., 2017]. A general requirement for such networks is to respect the permutation invariance of point

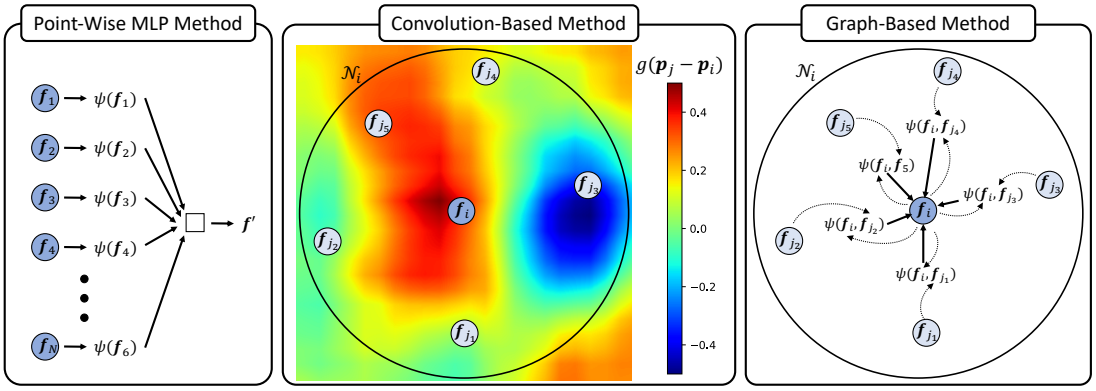


Fig. 2.2: Visual comparison of different deep learning approaches for point cloud analysis. The point-wise MLP applies a shared learnable function ψ to each point individually and aggregates the embeddings with a symmetric function $\square$. The convolution-based method applies a learnable continuous kernel g to all points in the local neighborhood $\mathcal{N}_i$. Graph-based methods use a learnable message passing function ψ to compute edge-wise features on the graph.

clouds, whose semantics remain unchanged when reordering the points, i.e., permuting the rows of the point matrix $\boldsymbol{P}$. Based on the categorization in [Guo et al., 2020], this section introduces three distinct approaches to the problem, namely point-wise MLP, convolution-based, and graph-based methods, as visualized in Fig. 2.2.

The presentation uses the following notation. All points $\boldsymbol{p}_i \in \mathcal{P}$ are associated with C-dimensional feature representations $\boldsymbol{f}_i \in \mathbb{R}^C$, stored in the feature matrix $\boldsymbol{F} \in \mathbb{R}^{N \times C}$, with $\boldsymbol{f}_{\cdot,c} \in \mathbb{R}^N$ denoting the c-th feature map over all points. At the input to the network, the features are typically identical to the point coordinates, potentially complemented by intensity information or surface normals. At hidden layers, the features are given as the output of the previous layer. Moreover, $\mathcal{N}_i$ denotes the set of indices of the points in the local neighborhood of $\boldsymbol{p}_i$, commonly represented by the k nearest neighbors or all points within a sphere of radius r centered at $\boldsymbol{p}_i$.

2.2.2.1 Point-Wise MLP Methods

As the most popular point-wise MLP method, the PointNet [Qi et al., 2017a] applies a shared MLP $\psi : \mathbb{R}^C \to \mathbb{R}^{C'}$ to the features of each point of the input cloud individually and aggregates the resulting point-wise output features with a channel-wise, symmetric function $\square$, such as maximum, average, or sum, yielding a global feature vector

$$\boldsymbol{f}' = \square_{\boldsymbol{f}_i \in \boldsymbol{F}} \psi(\boldsymbol{f}_i; \boldsymbol{\theta}_\psi) \tag{2.3}$$

Due to the weight-sharing mechanism, the point-wise features $\psi(\boldsymbol{f}_i; \boldsymbol{\theta}_\psi)$ are permutation-equivariant, and the symmetry of the aggregation function makes the global feature

vector $\boldsymbol{f}'$ permutation-invariant. When learning a global task like classification, a standard MLP can map the global feature vector to the output. For tasks requiring both local and global information, such as point segmentation, the global feature is concatenated to each point-wise local feature, and the resulting combined point features are again separately processed by another shared MLP.

A weakness of the PointNet architecture is that it completely ignores local structures. As a remedy, PointNet++ [Qi et al., 2017b] hierarchically groups the points in local clusters and separately processes them with a mini PointNet (similar to Eq. (2.3)), shared among all clusters at the same level. The spatial size of the clusters increases along the hierarchy, thus emulating the increasing receptive field and abstraction level of features in CNNs.

2.2.2.2 Convolution-Based Methods

Convolution-based methods follow the spirit of the regular grid convolution in Eq. (2.2) by formulating the convolution on a point cloud as the weighted sum over the input signal (point features) in local neighborhoods, with the weights being learnable parameters of a kernel function g. The essential challenge is the implementation of the weighted sum since the irregular structure of point clouds is incompatible with discrete kernels (used in grid convolutions). Here, two contrasting solutions are discussed, which address the problem on the side of the kernel and the input signal, respectively.

The first one replaces discrete kernels with continuous kernel functions $g_c(\boldsymbol{p}_j - \boldsymbol{p}_i)$, $g_c : \mathbb{R}^3 \to \mathbb{R}$, which depend on the offset vector between the center point $\boldsymbol{p}_i$ and the respective neighbor $\boldsymbol{p}_j$ and can thus provide the kernel weights for local neighborhoods with arbitrary spatial arrangements (Fig. 2.2, middle). Point-wise features $\boldsymbol{F}$ then transform as

$$(\boldsymbol{F} * g)_i = \sum_{j \in \mathcal{N}_i} \sum_{c=1}^{C} g_c(\boldsymbol{p}_j - \boldsymbol{p}_i) F_{j,c} \tag{2.4}$$

However, continuous kernels are more challenging to design and learn than discretized ones. A frequent solution is to model the kernel with a neural implicit function, i.e., a learnable MLP ψ, which predicts the kernel weights as $g_c(\boldsymbol{p}_j - \boldsymbol{p}_i) = \psi_c(\boldsymbol{p}_j, \boldsymbol{p}_i; \boldsymbol{\theta}_{\psi_c})$ based on low-level relations between $\boldsymbol{p}_i$ and $\boldsymbol{p}_j$. For instance, Liu et al. [2019b] condition ψ on the Euclidean distance and the offset vector between $\boldsymbol{p}_i$ and $\boldsymbol{p}_j$ and the individual coordinates, whereas the PointConv [Wu et al., 2019] exclusively uses the offset vector. In addition, to account for potentially non-uniform point densities in local neighborhoods, the kernel weights of PointConv are scaled with an inverse density estimate.

The second solution, by contrast, sticks to discretized kernels and, in exchange, maps the input signal $\boldsymbol{f}_j$ from the irregular neighborhood $\mathcal{N}_i$ of node i to a representation

in a regular local intrinsic coordinate system, where the transformed signal matches a discrete kernel. The mapping is performed with an L-dimensional patch operator

$$D_l(\boldsymbol{p}_i)\boldsymbol{F} = \sum_{j \in \mathcal{N}_i} w_l(\boldsymbol{u}(\boldsymbol{p}_i, \boldsymbol{p}_j))\boldsymbol{f}_j, \quad l = 1, ..., L \tag{2.5}$$

which computes the importance of feature $\boldsymbol{f}_j$ to the l-th dimension of the representation for the central node i by applying a parametric weighting function w_l to local pseudo-coordinates $\boldsymbol{u}(\boldsymbol{p}_i, \boldsymbol{p}_j)$ (e.g., local Cartesian coordinates or polar coordinates). While Masci et al. [2015] implement the weighting functions as Gaussian kernels with fixed hand-crafted means and variances, Monti et al. [2017] leave them as learnable parameters. The convolution is then defined as

$$(\boldsymbol{F} * \boldsymbol{g})_i = \sum_{l=1}^{L} g_l D_l(\boldsymbol{p}_i)\boldsymbol{F} \tag{2.6}$$

with $\boldsymbol{g} \in \mathbb{R}^L$ the learnable discrete kernel.

2.2.2.3 Graph-Based Methods

Graph-based methods treat the points of the cloud as the vertices $\mathcal{V} \subset \mathbb{R}^3$ of a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, connected by edges $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$, which can also be represented by the binary adjacency matrix $\boldsymbol{A} \in \{0, 1\}^{N \times N}$ with $A_{i,j} = \mathbb{1}_{(i,j) \in \mathcal{E}}$. The edges connect each vertex i with the vertices $j \in \mathcal{N}_i$ in its neighborhood.

Early graph-based methods are based on the spectrum of the graph, performing the convolution in the Fourier domain. Assuming an undirected graph with a symmetric adjacency matrix $\boldsymbol{A} = \boldsymbol{A}^\top$, the so-called normalized graph Laplacian $\boldsymbol{L} = \boldsymbol{1}_N - \boldsymbol{D}^{-1/2}\boldsymbol{A}\boldsymbol{D}^{-1/2}$, with $\boldsymbol{D} = \text{diag}(\sum_j A_{i,j})$, has an eigendecomposition $\boldsymbol{L} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$, where $\boldsymbol{U} = (\boldsymbol{u}_1, ..., \boldsymbol{u}_N)$ are the orthonormal eigenvectors and $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, ..., \lambda_N)$ is a diagonal matrix with the associated eigenvalues. Given these prerequisites, the Fourier graph transform of a feature map $\boldsymbol{f}_{\cdot,c}$ on the graph vertices is defined as $\hat{\boldsymbol{f}}_{\cdot,c} = \boldsymbol{U}^\top \boldsymbol{f}_{\cdot,c}$ with the inverse $\boldsymbol{f}_{\cdot,c} = \boldsymbol{U}\hat{\boldsymbol{f}}_{\cdot,c}$. Next leveraging the convolution theorem, the convolution $*_\mathcal{G}$ of a feature map on a graph with a kernel $\boldsymbol{g}$ becomes

$$\boldsymbol{f}_{\cdot,c} *_\mathcal{G} \boldsymbol{g} = \boldsymbol{U}((\boldsymbol{U}^\top \boldsymbol{f}_{\cdot,c}) \odot (\boldsymbol{U}^\top \boldsymbol{g})) = \boldsymbol{U}\text{diag}(\hat{g}_1, ..., \hat{g}_N)\hat{\boldsymbol{f}}_{\cdot,c} \tag{2.7}$$

where the filter is directly designed in the spectral domain. Bruna et al. [2014] leveraged this formulation to construct a spectral convolutional layer as the basis of a graph neural network, which, however, involves several problems. 1) The filter is not necessarily localized in the spatial domain, 2) its parameters are dependent on the specific topology of the graph, and 3) computing forward and backward Fourier transforms is computationally expensive. Therefore, Defferrard et al. [2016] consider

$\boldsymbol{g}$ as a function of $\boldsymbol{\Lambda}$ and approximate it by a K-th order expansion in Chebyshev polynomials $T_k(x)$, leaving them with

$$\boldsymbol{f}_{\cdot,c} *_{\mathcal{G}} \boldsymbol{g}_{\boldsymbol{\theta}} = \sum_{k=0}^{K-1} \theta_k T_k(\tilde{\boldsymbol{L}}) \boldsymbol{f}_{\cdot,c} \tag{2.8}$$

with the rescaled $\tilde{\boldsymbol{L}} = \frac{2}{\lambda_{\max}} \boldsymbol{L} - \mathbf{1}_N$. Now, the learnable parameters of the filter are a vector of Chebyshev coefficients $\boldsymbol{\theta} \in \mathbb{R}^K$. This formulation circumvents explicitly computing the Fourier transforms and is localized to nodes that are up to K steps away from the central node. To reduce network parameters and the risk of overfitting and ease building deeper networks, Kipf et al. [2017] truncate the expansion at $K = 1$ and set $\theta = \theta_0 = -\theta_1$, simplifying the convolution to

$$\boldsymbol{f}_{\cdot,c} *_{\mathcal{G}} \boldsymbol{g}_{\theta} = \theta(\mathbf{1}_N + \boldsymbol{D}^{-1/2} \boldsymbol{A} \boldsymbol{D}^{-1/2}) \boldsymbol{f}_{\cdot,c} \tag{2.9}$$

which is localized to the direct neighborhood.

Another group of graph-based methods directly operates in the spatial domain. Following Bronstein et al. [2021], many of these methods are specific implementations of the general concept of message-passing graph convolutions (Fig. 2.2, right), which compute the output features $\boldsymbol{f}_i'$ of a layer as

$$\boldsymbol{f}_i' = \phi(\boldsymbol{f}_i, \square_{j \in \mathcal{N}_i} \psi(\boldsymbol{f}_i, \boldsymbol{f}_j; \boldsymbol{\theta}_\psi); \boldsymbol{\theta}_\phi) \tag{2.10}$$

Here, the learnable message function $\psi : \mathbb{R}^C \times \mathbb{R}^C \to \mathbb{R}^{C''}$ computes so-called edge features for all neighbors, the symmetric aggregation function $\square$ performs the message passing, and $\phi : \mathbb{R}^C \times \mathbb{R}^{C''} \to \mathbb{R}^{C'}$ is another learnable function to update the features of vertex i with the help of a skip connection. Typically, ψ and ϕ are represented by MLPs. Applying the function to each node i independently yields a permutation-equivariant point-wise feature representation because ϕ is permutation-invariant due to the symmetric aggregation function. A popular form of message-passing graph convolutions is the edge convolution [Wang et al., 2019], which neglects ϕ, i.e., $\boldsymbol{f}_i' = \square_{j \in \mathcal{N}_i} \psi(\boldsymbol{f}_i, \boldsymbol{f}_j; \boldsymbol{\theta}_\psi)$. While there are different reasonable implementations of ψ, Wang et al. [2019] use an asymmetric message function

$$\psi(\boldsymbol{f}_i, \boldsymbol{f}_j; \boldsymbol{\theta}_\psi) = \psi(\boldsymbol{f}_i, \boldsymbol{f}_j - \boldsymbol{f}_i; \boldsymbol{\theta}_\psi) = \mathrm{ReLU}(\boldsymbol{\theta}_\psi^{(1)} \boldsymbol{f}_i + \boldsymbol{\theta}_\psi^{(2)} (\boldsymbol{f}_j - \boldsymbol{f}_i)) \tag{2.11}$$

to capture both global ($\boldsymbol{f}_i$) and local ($\boldsymbol{f}_j - \boldsymbol{f}_i$) information. This implementation of the edge convolution is extensively used throughout this thesis, constituting the basis operation of the networks in Chapters 4, 5, and 6.

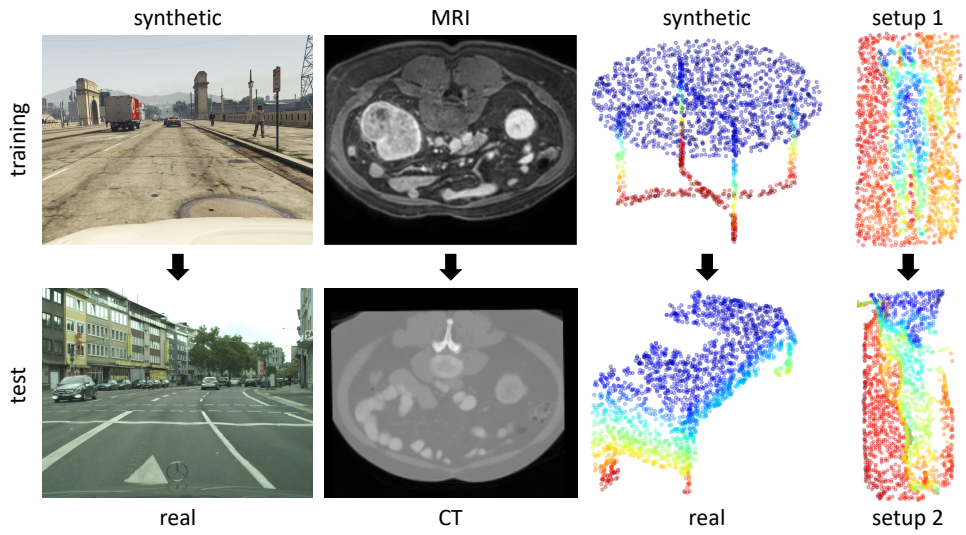


Fig. 2.3: Exemplary domain shifts encountered in (medical) image and point cloud analysis, including shifts between simulated and real data and between different modalities and environments. The image data primarily differ in the intensity distributions, while point clouds exhibit purely geometric domain shifts.
Image sources: upper left: Richter et al. [2016], lower left: Cordts et al. [2016], 2nd column: Hering et al. [2022], 3rd column: Huang et al. [2021], upper right: [Liu et al., 2022a], lower right: in-house data.

2.3 Domain Adaptation

Standard machine learning approaches, as introduced in Sec. 2.2, typically assume training and testing data to be identically distributed such that a properly trained model generalizes to the test data. In practical scenarios, however, training and test data often have different underlying probability distributions, also denoted as a domain shift (see Fig. 2.3 for visual examples), causing severe performance drops at test time. Overcoming the domain shift by adapting the model to the shifted target domain is commonly referred to as domain adaptation. This section first provides a more formal definition of the problem and then provides an overview of popular approaches to solve it.

2.3.1 Formal Definition

The formal definition follows the notation in the surveys by Pan et al. [2010] and Wang et al. [2018]. The authors define a domain $\mathcal{D} = \{\mathcal{X}, P(\boldsymbol{x})\}$ to consist of an input space $\mathcal{X} \subset \mathbb{R}^d$ with a marginal probability distribution $P(\boldsymbol{x})$ with $\boldsymbol{x} \in \mathcal{X}$. On a domain $\mathcal{D}$, a task $\mathfrak{T} = \{\mathcal{Y}, P(\boldsymbol{y}|\boldsymbol{x})\}$ is defined by a label space $\mathcal{Y} \subset \mathbb{R}^k$ and a conditional probability

distribution $P(\boldsymbol{y}|\boldsymbol{x})$ with $\boldsymbol{y} \in \mathcal{Y}$. The goal of standard machine learning is to learn $P(\boldsymbol{y}|\boldsymbol{x})$ from a labeled training dataset $\mathcal{S} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_i$.

One can now generally assume that training and test data stem from two domains, denoted as the source domain $\mathcal{D}^s = \{\mathcal{X}^s, P(\boldsymbol{x}^s)\}$ and target domain $\mathcal{D}^t = \{\mathcal{X}^t, P(\boldsymbol{x}^t)\}$, respectively, and associated with the tasks $\mathfrak{T}^s = \{\mathcal{Y}^s, P(\boldsymbol{y}^s|\boldsymbol{x}^s)\}$ and $\mathfrak{T}^t = \{\mathcal{Y}^t, P(\boldsymbol{y}^t|\boldsymbol{x}^t)\}$. Traditional machine learning deals with both equal domains $\mathcal{D}^s = \mathcal{D}^t$ and tasks $\mathfrak{T}^s = \mathfrak{T}^t$. This assumption contrasts with settings where training and test data differ due to domain divergence ($\mathcal{D}^s \neq \mathcal{D}^t$) and/or task divergence ($\mathfrak{T}^s \neq \mathfrak{T}^t$), which are addressed by transfer learning.

As a special form of transfer learning, domain adaptation treats domain divergence ($\mathcal{D}^s \neq \mathcal{D}^t$) with equal tasks ($\mathfrak{T}^s = \mathfrak{T}^t$). Domain divergence, in turn, can be caused by different input spaces ($\mathcal{X}^s \neq \mathcal{X}^t$) or distribution shifts ($P(\boldsymbol{x}^s) \neq P(\boldsymbol{x}^t)$). These scenarios are known as heterogeneous and homogeneous domain adaptation, respectively, with the latter the exclusive focus of this work.

Homogeneous domain adaptation problems can further be categorized according to the availability of training data:

- Classical unsupervised domain adaptation (UDA) assumes simultaneous access to sufficient labeled source data $\mathcal{S} = \{(\boldsymbol{x}_i^s, \boldsymbol{y}_i^s)\}_i$ and unlabeled target data $\mathcal{T} = \{\boldsymbol{x}_i^t\}_i$.
- In semi-supervised domain adaptation (SSDA), the unlabeled target data is complemented by a small set of labeled target data $\mathcal{T}^l = \{(\boldsymbol{x}_i^t, \boldsymbol{y}_i^t)\}_i$.
- Source-free domain adaptation (SFDA) considers the same data as UDA, but source and target datasets are only accessible in successive stages and not simultaneously.
- Test-time adaptation (TTA) is similar to SFDA with the difference that a full training set from the target domain is unavailable and the model is directly adapted to individual testing samples.

Chapter 4 of this work treats both UDA and SFDA while Chapter 6 focuses on UDA.

2.3.2 Domain Adaptation Methods

The goal of this section is to convey the essential concepts of popular domain adaptation methods, focusing on the classical UDA setting. For a comprehensive overview of the current literature, the reader is referred to Sec. 4.1.1.2 and the survey by Wang et al. [2018]. The presented techniques were mainly developed for computer vision tasks with images as inputs, but most underlying ideas are transferable to other modalities, including 3D point clouds. Therefore, the notation is kept generic with inputs designated as $\boldsymbol{x}$. As the starting point, assume a neural network f, which can be trained for a certain task on a source dataset $\mathcal{S}$ by minimizing the task loss $\mathcal{L}_{\text{task}}(\mathcal{S}; \boldsymbol{\theta})$ and generalizes well to test data from the same domain. Now, the essential question is of

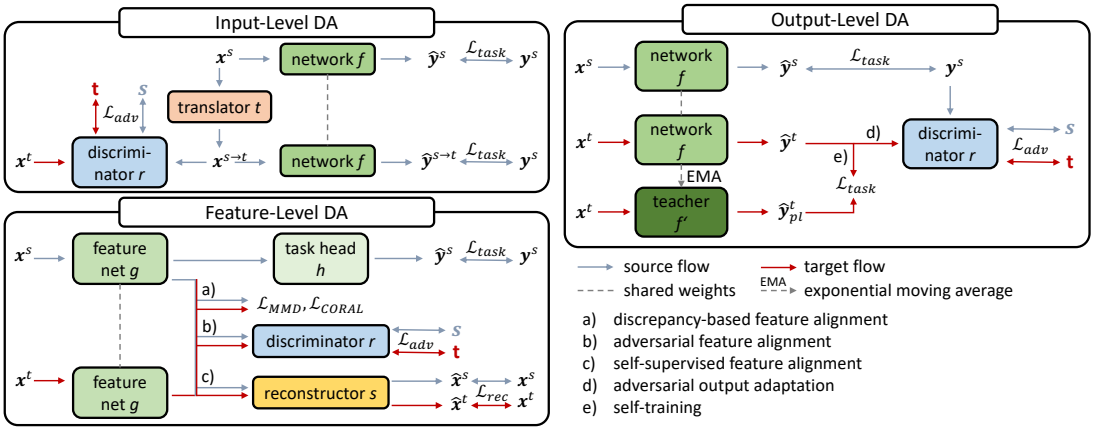


Fig. 2.4: Schematic visualization of different unsupervised domain adaptation approaches.

how to efficiently exploit the unlabeled target data $\mathcal{T}$ to adapt the model to the shifted target domain. In the following, existing approaches to the problem are categorized according to the level, where the adaptation takes place, namely the input space, the feature space, or the output space (see Fig. 2.4 for a schematic overview).

2.3.2.1 Input-Level Adaptation

The idea of input-level adaptation is to train an image translation model $t : \mathcal{X} \to \mathcal{X}$ that adapts the style of the source images $\boldsymbol{x}^s$ to the target domain as $\boldsymbol{x}^{s\to t} = t(\boldsymbol{x}^s; \boldsymbol{\theta}_t)$ while preserving the semantic content such that the corresponding source labels $\boldsymbol{y}^s$ remain valid. The result is a labeled dataset $\mathcal{T}^{\text{syn}} = \{(\boldsymbol{x}_i^{s\to t}, \boldsymbol{y}_i^s)\}_i$ in target style. Afterward, the task model f is trained on this translated dataset in a standard supervised manner by minimizing $\mathcal{L}_{\text{task}}(\mathcal{T}^{\text{syn}}; \boldsymbol{\theta})$.

The style transfer is typically achieved by embedding the translation model into a generative adversarial network (GAN)-based framework [Goodfellow et al., 2014], where it is trained in an adversarial manner against an auxiliary domain discriminator r. While the discriminator learns to distinguish real target images from translated source images by minimizing the binary cross entropy loss

$$\mathcal{L}_{\text{adv}}(\mathcal{S}, \mathcal{T}; \boldsymbol{\theta}_r, \boldsymbol{\theta}_t) = -\sum_{\boldsymbol{x}^t \in \mathcal{T}} \log(r(\boldsymbol{x}^t; \boldsymbol{\theta}_r)) - \sum_{(\boldsymbol{x}^s, \cdot) \in \mathcal{S}} \log(1 - r(t(\boldsymbol{x}^s; \boldsymbol{\theta}_t); \boldsymbol{\theta}_r)) \tag{2.12}$$

the translation model is trained to fool the discriminator by maximizing the second term of $\mathcal{L}_{\text{adv}}$. Since this technique primarily performs style transfer, it has been complemented by diverse strategies for content preservation, including a content-similarity loss [Bousmalis et al., 2017], a self-regularization loss [Shrivastava et al., 2017], and cycle-consistency constraints [Zhu et al., 2017b]. The latter approach

additionally trains a second translator-discriminator pair (t', r') for the backward mapping from target to source images and imposes the bi-directional cycle-consistency loss

$$\mathcal{L}_{cycle}(\mathcal{S}, \mathcal{T}; \boldsymbol{\theta}_t, \boldsymbol{\theta}_{t'}) = \sum_{\boldsymbol{x}^t \in \mathcal{T}} \|t(t'(\boldsymbol{x}^t; \boldsymbol{\theta}_{t'}), \boldsymbol{\theta}_t) - \boldsymbol{x}^t\|_1 + \sum_{(\boldsymbol{x}^s, \cdot) \in \mathcal{S}} \|t'(t(\boldsymbol{x}^s; \boldsymbol{\theta}_t); \boldsymbol{\theta}_{t'}) - \boldsymbol{x}^s\|_1 \tag{2.13}$$

which encourages the translators to restore the original images when successively applied and thus prevents the manipulation of semantic content.

2.3.2.2 Feature-Level Adaptation

Rather than aligning pixel-level distributions, feature-level domain adaptation aims to match the distributions of intermediate source and target features and, consequently, operates in the latent space of the model f, which is decomposed in a feature extractor g and task head h as $f(\boldsymbol{x}; \boldsymbol{\theta}) = h(g(\boldsymbol{x}; \boldsymbol{\theta}_g); \boldsymbol{\theta}_h)$. The underlying idea is that coupling supervised task learning on source data with the alignment of feature distributions makes the network learn both domain-invariant and semantically meaningful features. That way, the task head is expected to generalize well to target features even though exclusively trained on source features. Here, three implementations of feature matching are discussed.

Discrepancy-based methods minimize explicit distance measures between the feature distributions, such as the maximum mean discrepancy (MMD) [Gretton et al., 2006] or correlation alignment (CORAL) [Sun et al., 2016] loss. MMD measures the distance between the means of source and target features as

$$\mathcal{L}_{\text{MMD}}(\mathcal{S}, \mathcal{T}; \boldsymbol{\theta}_g) = \left\| \frac{1}{|\mathcal{S}|} \sum_{(\boldsymbol{x}^s, \cdot) \in \mathcal{S}} \phi(g(\boldsymbol{x}^s; \boldsymbol{\theta}_g)) - \frac{1}{|\mathcal{T}|} \sum_{\boldsymbol{x}^t \in \mathcal{T}} \phi(g(\boldsymbol{x}^t; \boldsymbol{\theta}_g)) \right\|^2 \tag{2.14}$$

Here, ϕ can either be an identity function or a mapping function to a kernel reproducing Hilbert space to not only align the means but also higher order statistics. The CORAL loss measures the distances between the second order statistics of the features as the Frobenius norm $\|\boldsymbol{C}^s - \boldsymbol{C}^t\|^2$ between the covariance matrices $\boldsymbol{C}^s / \boldsymbol{C}^t$ of the features $g(\boldsymbol{x}^s; \boldsymbol{\theta}_g) / g(\boldsymbol{x}^t; \boldsymbol{\theta}_g)$.

Adversarial-based feature alignment is, similar to image-to-image translation, based on the concept of GANs [Ganin et al., 2015; Tzeng et al., 2017]. Again, the approach includes an auxiliary domain discriminator r, which now operates in the feature space and is trained in an adversarial manner against the feature extractor. While the discriminator is optimized to distinguish source and target features by minimizing

$$\mathcal{L}_{\text{adv}}(\mathcal{S}, \mathcal{T}; \boldsymbol{\theta}_g, \boldsymbol{\theta}_r) = -\sum_{\boldsymbol{x}^t \in \mathcal{T}} \log(r(g(\boldsymbol{x}^t; \boldsymbol{\theta}_g); \boldsymbol{\theta}_r)) - \sum_{(\boldsymbol{x}^s, \cdot) \in \mathcal{S}} \log(1 - r(g(\boldsymbol{x}^s; \boldsymbol{\theta}_g); \boldsymbol{\theta}_r)) \tag{2.15}$$

the feature extractor maximizes the loss to deceive the discriminator and finally learn a domain-invariant latent space. Among different implementations of this adversarial optimization scheme, the gradient reversal layer, placed between g and r [Ganin et al., 2015] is a very elegant solution. Representing an identity function in the forward pass, it reverses the gradients during backpropagation such that g and r can be jointly optimized in a single forward-backward pass.

Finally, self-supervised approaches include an auxiliary network s to additionally learn one or multiple self-supervised tasks on the features from both domains. The underlying motivation is that, for a satisfying performance of s on both domains and all tasks, the feature extractor is required to extract domain-invariant and structure-preserving features, supposed the auxiliary tasks are carefully designed. Self-supervised tasks in the literature include image reconstruction [Bousmalis et al., 2016; Ghifary et al., 2016], aiming to reconstruct the original images from the feature representation by minimizing the reconstruction loss

$$\mathcal{L}_{\text{rec}}(\mathcal{S},\mathcal{T};\boldsymbol{\theta}_g,\boldsymbol{\theta}_s) = \sum_{(\boldsymbol{x}^s,\cdot)\in\mathcal{S}} \|s(g(\boldsymbol{x}^s;\boldsymbol{\theta}_g);\boldsymbol{\theta}_s)-\boldsymbol{x}^s\|^2 + \sum_{\boldsymbol{x}^t\in\mathcal{T}} \|s(g(\boldsymbol{x}^t;\boldsymbol{\theta}_g);\boldsymbol{\theta}_s)-\boldsymbol{x}^t\|^2 \quad (2.16)$$

and the prediction of image augmentations, such as rotation angles, vertical flips, or the original location of randomly sampled image patches [Sun et al., 2019b].

2.3.2.3 Output-Level Adaptation

Output-level adaptation guides the learning on unlabeled target data by directly supervising the network predictions $\hat{\boldsymbol{y}}^t = f(\boldsymbol{x}^t;\boldsymbol{\theta})$. Compared to feature- and input-level adaptation, the approach has the advantage of both training on real target images and explicitly optimizing the target features for the addressed task. The approach is commonly realized by two different techniques.

Adversarial output adaptation [Tsai et al., 2018] aligns the distributions of target predictions $\hat{\boldsymbol{y}}^t$ and source ground truth $\boldsymbol{y}^s$ by training a domain discriminator in an adversarial manner against the entire network f, implemented analogously to adversarial feature alignment. The underlying intuition is that, contrary to domain-variant input data, the abstracted representations in the output space are often inherently domain-invariant. In medical image segmentation, for instance, the organ segmentations of a human should have identical shapes regardless of the imaging modality (e.g., MR or CT). Hence, enforcing target predictions to follow the distribution of ground truth annotations in the source domain is a useful source of task-specific supervision for the network.

The other technique is self-training [Zou et al., 2018], which adapts the model in multiple steps. 1) Train the model on labeled source data by minimizing $\mathcal{L}_{\text{task}}(\mathcal{S};\boldsymbol{\theta})$. 2) Deploy the model on the unlabeled target data to generate a pseudo-labeled target dataset $\mathcal{T}^{\text{pl}} = \{(\boldsymbol{x}_i^t, \hat{\boldsymbol{y}}_i^t = f(\boldsymbol{x}_i^t;\boldsymbol{\theta}))\}_i$. 3) Re-train the model on the union of labeled

source and pseudo-labeled target data, minimizing $\mathcal{L}_{\text{task}}(\mathcal{S} \cup \mathcal{T}^{\text{pl}}; \boldsymbol{\theta})$. 4) Cyclically repeat steps 2) and 3). A special form of self-training is the Mean Teacher paradigm [Tarvainen et al., 2017], where pseudo labels are not cyclically updated but generated on the fly by a so-called teacher model, whose weights represent the exponential moving average of the weights of the learning model. As such, the teacher can be considered a temporal ensemble and its predictions are expected to be on average more accurate than those of the student, motivating their use for pseudo supervision.

Chapter 3

Deep Learning-Based Body Weight Estimation From Point Clouds

The first methodological chapter addresses body weight estimation of in-bed patients – a global regression problem – focusing on fully-supervised learning on single point clouds captured by a clinical monitoring system. The chapter explores the suitability of deep learning approaches to the problem and how to improve them by injecting prior knowledge about the task. Initially considering fully-visible patients without a blanket, Sec. 3.1 demonstrates that a basis point set encoding combined with a fully-connected network is a suitable end-to-end solution, which is further boosted by sampling the basis points according to the a priori known distribution of the input point clouds. Sec. 3.2 then investigates the weight estimation of occluded patients under a cover and develops a two-step method, emulating the intuitive human approach of uncovering the patients first and then estimating their weight in a separate step. It is shown that a 3D U-Net can virtually see under the blanket and reconstruct the patient's shape, thus accomplishing the first step and significantly simplifying the subsequent weight estimate by a customized 3D regression CNN. Finally, the chapter investigates the robustness of all developed methods to a changed room setup, revealing a significant susceptibility to the resulting geometric domain shift.

3.1 End-to-End Learning of Body Weight Prediction With Basis Point Sets

This section has been published in [Bigalke et al., 2021b]. According to the Contributor Roles Taxonomy (CRediT), the contributions of the author of this thesis to the publication are: Conceptualization (together with L.H., M.P.H.), Methodology, Software, Investigation, Writing – Original Draft, Writing – Review & Editing (together with L.H., M.P.H.), Visualization.

3.1.1 Introduction

The precise knowledge of a patient's body weight is a crucial requirement in several clinical scenarios, including anesthesia or drug dosage. In emergency situations, however, patients are often unable to communicate their weight due to unconsciousness, dementia or neurological disorder. Weighing the patient on-site with an ordinary scale is infeasible in case of severe injuries, and bed scales are expensive and not always available. For these reasons, weight is often estimated by clinical staff although this procedure has been shown to be error-prone in clinical studies [Menon et al., 2005].

To obtain more accurate weight estimates, several works use a multiple linear regression model to infer body weight from biometric measurements such as height, and waist and hip circumference [Lorenz et al., 2007]. Since manual measurements of these quantities are time-intensive and infeasible in case of certain injuries, it is difficult to integrate this approach into clinical routine. Instead, a fully automatic and contactless weight estimate is desirable.

This can be achieved by deriving the weight estimate from visual sensor data using methods from computer vision. For this purpose, the use of depth sensors is particularly suitable. Firstly, depth maps and corresponding point clouds carry rich geometric information, which is of eminent importance for accurate weight estimates. Secondly, patients are unidentifiable on depth maps which prevents any privacy concerns.

This work addresses the task of weight estimation of patients lying in bed from 3D point cloud data by means of deep learning techniques. Contrary to prior work in this field, we aim to learn weight prediction in an end-to-end fashion.

3.1.1.1 Related Work

More generic work in the field of weight estimation from visual data predicts the weight of free-standing subjects from RGB-D data [Velardo et al., 2012]. The proposed method segments the subject from the background and extrapolates biometric measures from the silhouette. The deduced measures are fed into a neural network to regress the subject's weight.

Several works address weight estimation from point clouds of lying patients in a clinical environment [Pfitzner et al., 2015, 2016]. Libra3D [Pfitzner et al., 2015] fits a mesh to the point cloud, whereby the patient's back is modeled with help of the bed plane. Based on the mesh, the volume of the patient is calculated and multiplied with a fixed empirically determined density to obtain a weight estimate. Pfitzner et al. [2016] extend this work by additionally extracting more abstract features from the point cloud, which are forwarded by a neural network for weight regression. All of these works rely on hand-crafted features and are not trained end-to-end.

In recent years, end-to-end learning from point clouds has become viable owing to deep learning architectures that directly operate on raw point sets. The pioneering PointNet

[Qi et al., 2017a] applies a shared multi-layer perceptron to each point individually and achieves permutation invariance through a symmetric max pooling operation.

3.1.1.2 Contributions

To our knowledge, this is the first work to learn weight prediction from 3D point clouds in an end-to-end fashion. Since learning weight regression directly from raw point clouds using a PointNet architecture [Qi et al., 2017a] is a complex task, we suggest to simplify the problem by considering input point clouds relative to a fixed reference. To achieve this, we adopt the idea of basis point sets (BPS) [Prokudin et al., 2019] to encode the input point cloud. The resulting feature vector is subsequently fed into a fully connected neural network to regress the weight. Based on the observation that the construction of the BPS in [Prokudin et al., 2019] is not ideal for our specific problem, we propose an adapted sampling scheme to incorporate prior knowledge about the distribution of input points. We experimentally validate our approach on the SLP dataset [Liu et al., 2019a] and significantly outperform several baselines, including a PointNet architecture [Qi et al., 2017a].

3.1.2 Methods

Our method receives a point cloud cropped around the bed as input and outputs the patient's weight in kg. The point cloud is initially pre-processed, subsequently encoded by means of a BPS and finally processed by a neural network for weight prediction (Fig. 3.1). We assume the patient to be uncovered and in a supine position, which can easily be realized in clinical workflow.

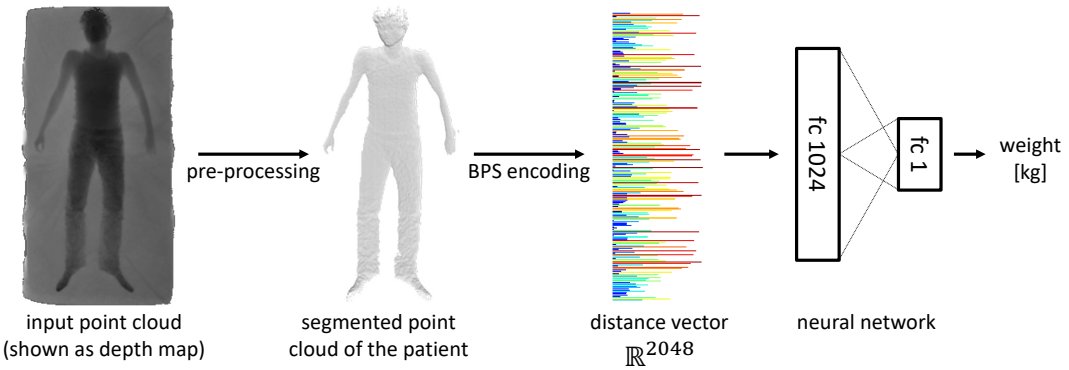


Fig. 3.1: Overview of our proposed pipeline for weight estimation from point clouds.

3.1.2.1 Pre-processing

In the pre-processing step, the patient needs to be segmented from the bed. First, we use the RANSAC algorithm [Fischler et al., 1981] to fit a plane to the mattress and keep only the points above the plane as it was done in [Pfitzner et al., 2015]. Most of the kept points belong to the patient, but there may remain point clusters belonging to other objects on the bed. To remove those points, we cluster the cloud using DBSCAN [Ester et al., 1996] and only keep the largest cluster.

3.1.2.2 Basis Point Set and Neural Network

After pre-processing, we are left with a set of patient point clouds $\boldsymbol{X}_i \in \mathbb{R}^{N_i \times 3}$, $(i = 1, ..., p)$, each comprising N_i points $\boldsymbol{x}_{ij} \in \mathbb{R}^3$. We encode the clouds with help of a BPS as elaborated in [Prokudin et al., 2019]. In [Prokudin et al., 2019], each cloud is initially normalized to fit a unit sphere which entails a loss of scale information. Since scale is indispensable for weight estimation, we only mean-center each cloud. Subsequently, a BPS

$$\boldsymbol{B} = [\boldsymbol{b}_1, ..., \boldsymbol{b}_k], \boldsymbol{b}_j \in \mathbb{R}^3, \|\boldsymbol{b}_j\| \leq r \tag{3.1}$$

is constructed by uniform sampling of k points from a sphere of radius r. This set is fixed for all point clouds in training and test set. We select $k = 2048$ and set the radius to the maximal radius of all point clouds in the training set, i.e., $r = \max_i(\max_j \|\boldsymbol{x}_{ij} - (\sum_k \boldsymbol{x}_{ik})/N_i\|)$.

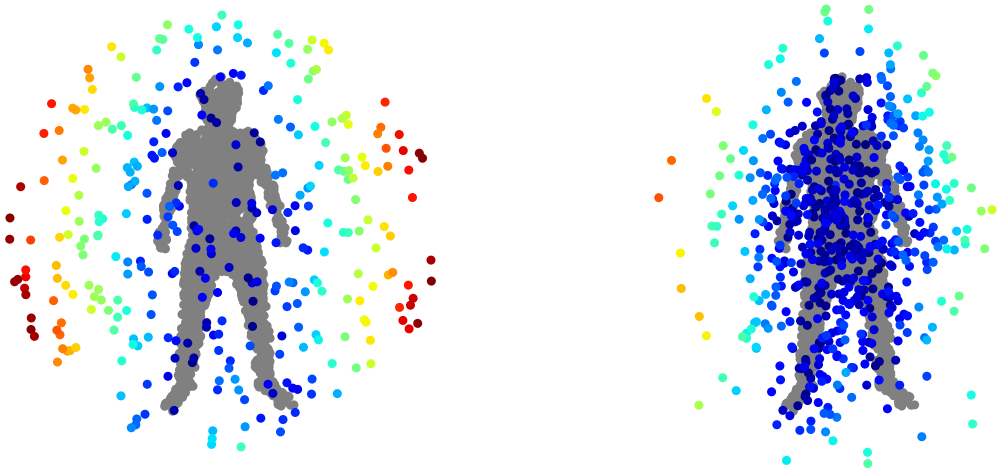
Given the BPS $\boldsymbol{B}$, an input point cloud $\boldsymbol{X}_i$ is encoded by computing the distance from each basis point to the nearest point in the input cloud, yielding a k-dimensional feature vector

$$\boldsymbol{f}_i^{\boldsymbol{B}} = [\min_{\boldsymbol{x}_{ij} \in \boldsymbol{X}_i} d(\boldsymbol{b}_1, \boldsymbol{x}_{ij}), ..., \min_{\boldsymbol{x}_{ij} \in \boldsymbol{X}_i} d(\boldsymbol{b}_k, \boldsymbol{x}_{ij})] \in \mathbb{R}^k \tag{3.2}$$

This feature vector is subsequently fed into a neural network, consisting of the following sequence of layers: BN, FC(1024), ReLU, BN, Dropout(p=0.8), FC(1). The network parameters are optimized by minimizing a mean squared error loss between predicted weight and ground truth.

3.1.2.3 Adapted Sampling of Basis Points

In Fig. 3.2a, a BPS obtained by uniform sampling in the sphere is shown relative to an input point cloud. We observe that many basis points are far away from the patient and thus encode less detailed information. As all patients have a similar orientation and occupy similar regions of the sphere, we conclude that the uniform distribution of basis points is not ideal for our specific problem. We believe that a more expressive basis can be constructed by incorporating prior knowledge about the distribution of input points. To achieve this, we propose to sample the basis points from a unified point cloud which comprises all clouds from the training set. As this basis is prone to



(a) Uniform sampling in the sphere.

(b) Sampling from training points.

Fig. 3.2: Comparison of two basis point sets constructed with different sampling schemes. We visualize a slice of the sphere around the input point cloud of a patient, which is shown in gray for reference. Basis points are shown in color to represent the distance to the closest input point. The basis points constructed by our sampling scheme (b) are substantially more concentrated around the patient.

overfitting, we subsequently add Gaussian noise with a standard deviation of $\sigma = 0.3$ to the sampled basis points. The resulting BPS is depicted in Fig. 3.2b.

3.1.3 Experiments and Results

3.1.3.1 Experimental Setup

Dataset. We evaluate our method on a subset of the SLP dataset [Liu et al., 2019a]. The subset comprises depth maps of 109 subjects which are lying in bed in a supine position without a cover. Each subject takes 15 different poses while staying in supine position, yielding an overall of 1635 frames. For each frame, a bounding box around the bed is obtained with the help of depth thresholding, and the corresponding image crop is transformed to a point cloud using the internal camera parameters. The weight of the subjects ranges from 43.7 to 105.1 kg with a mean of 68.0 kg and a standard deviation of 12.7 kg. We use the first 60 subjects for training, and results are reported for the remaining 49 subjects.

Implementation details. For pre-processing, we run RANSAC with a threshold of 1 cm for 1000 iterations. DBSCAN is used with $\epsilon = 2.5$ cm and minpts $= 5$. Network parameters are optimized with the ADAM optimizer. The initial learning rate is set to 0.001 and halved every 40 epochs. We use a batch size of 16 and train for 200 epochs. Each experiment is repeated ten times and we report mean and standard deviation.

Table 3.1: Results for weight estimation on the SLP dataset.

Method	MAE [kg]	MRE [%]	in 10% range [%]
Mean	9.46	14.6	40.8
Median	9.54	14.3	44.9
PointNet	5.42 ± 0.40	8.0 ± 0.5	70.7 ± 3.2
PointNet & median	4.84 ± 0.48	7.1 ± 0.6	74.7 ± 5.3
BPS random sampling	4.91 ± 0.09	7.5 ± 0.1	74.0 ± 0.8
BPS adapted sampling	4.69 ± 0.08	7.1 ± 0.1	76.1 ± 1.0
BPS adapted sampling & median	$\mathbf{4.19 \pm 0.12}$	$\mathbf{6.4 \pm 0.2}$	$\mathbf{78.6 \pm 2.9}$

Baselines. As baseline, we train a basic PointNet [Qi et al., 2017a] to directly regress the weight from the point cloud of the patient. Additionally, we estimate the weight of each test subject with a constant value which corresponds to the mean/median weight of all subjects from the training set of the same sex as the test subject.

3.1.3.2 Results

Results are presented in Tab. 3.1. We compare the baseline methods to three variants of our approach: 1) uniform sampling of basis points in the sphere, 2) sampling the basis points from training points, 3) same as 2), but for each subject, we take the median of the predicted weights for all 15 frames. For each method, we report the following metrics on the test set: mean absolute error (MAE), mean relative error (MRE), percentage of subjects within a relative error range of ±10%.

Results demonstrate that BPS with random sampling halves both MAE and MRE of the mean/median baselines and considerably improves on the PointNet architecture without median filtering as well. Applying our adapted sampling further reduces MAE by 0.22 kg and MRE by 0.4% points. Finally taking the median of 15 independent weight estimates for the same subject yields another improvement of 0.5 kg in MAE and 0.7% points in MRE. That way, we achieve an overall MAE of 4.19 kg and MRE of 6.4%, and the weight of 78.6% of the subjects is estimated within a 10% error range. This constitutes a relative performance gain in MAE of 56% compared to the mean/median baseline and of 13.4% in relation to the corresponding PointNet model.

3.1.4 Discussion and Conclusion

This work successfully applied the concept of BPS [Prokudin et al., 2019] to learn body weight prediction from point clouds of lying patients in an end-to-end fashion. We optimized the method for the specific problem at hand by introducing a customized sampling scheme for basis construction which takes the prior distribution of input

points into account and thus contributed a meaningful performance gain. Finally, the experiments showed that a further increase in accuracy can be achieved by statistical averaging over several independent weight estimates for the same subject. Altogether, our method achieves a higher accuracy (MAE=4.2 kg, MRE=6.4%) than weight estimates by clinical staff, which exhibit MAEs between 5.7 and 8.7 kg in [Lorenz et al., 2007] and MREs of 7.7 to 11.0% in [Menon et al., 2005]. That way, our work demonstrates the potential of end-to-end deep learning in the context of weight estimation and thus encourages further research in this direction. Future work could, for instance, incorporate semantic labels or point descriptors into the encoding or address the construction of an even more tailored basis set.

3.2 Weight Estimation Under the Cover With a 3D U-Net

This section has been published in [Bigalke et al., 2021a]. According to the Contributor Roles Taxonomy (CRediT), the contributions of the author of this thesis to the publication are: Conceptualization (together with L.H., J.D., M.P.H.), Methodology, Software, Investigation, Writing – Original Draft, Writing – Review & Editing (together with L.H., J.D., M.P.H.), Visualization. The source code is available at `https://github.com/multimodallearning/weight-estimation-under-cover`.

3.2.1 Introduction

Medical treatments often require the precise knowledge of a patient's body weight, e.g., for patient-adapted drug dosing. In emergency situations, however, a straightforward assessment of the patient's weight is often impossible. Unconsciousness of patients prevents a proper anamnesis, immobility impedes the usage of an ordinary scale, and bed scales are not always available. As a consequence, weight is often estimated by clinical staff although clinical studies have revealed the inaccuracy of these estimates [Fernandes et al., 1999; Menon et al., 2005]. Several works suggest that there is a possibility to increase accuracy by inferring weight estimates from biometric measurements [Buckley et al., 2012; Cattermole et al., 2017; Lorenz et al., 2007], but it is impractical to integrate the manual realization of these measurements into clinical routine. Instead, it is desirable to estimate the patient's weight in a fully automatic and contactless way based on visual sensor data.

Pfitzner et al. [2017] and our prior work [Bigalke et al., 2021b] already demonstrated that machine learning-based methods are capable of deriving precise weight estimates of lying patients on the basis of point cloud data. Point clouds carry rich geometric information while preserving the patient's data privacy [Silas et al., 2015] and are thus particularly suitable for the given problem. Whereas the proposed methods predict weight estimates with a promising accuracy, they involve a critical drawback: the methods are designed for and evaluated under highly controlled conditions. Patients

are expected to be uncovered and in supine position in [Bigalke et al., 2021b] and additionally even need to take a specific pose in [Pfitzner et al., 2017].

In clinical practice, however, these specific demands are not always fulfilled. Patients take arbitrary poses and might be covered by a blanket. Especially the occlusion of the patient by a blanket considerably complicates the weight estimation problem and poses new challenges. First, it is no longer possible to identify a clear boundary between patient and mattress. Second, it is difficult to distinguish volume that actually belongs to the patient from volume belonging to the blanket and hollow space under the cover. As a consequence, existing methods are either no longer applicable at all [Pfitzner et al., 2017] or suffer from a substantial degradation of accuracy [Bigalke et al., 2021b]. Specifically, we evaluated our prior work [Bigalke et al., 2021b] under occlusions by a blanket and observed an increase in the error of weight estimates by up to 58 %. In practice, the predicted weight estimates will thus either be less accurate or clinical staff needs to manually remove the cover. Both options are unsatisfactory with regard to the intended fully automatic solution. Instead, an ideal weight estimation framework would provide reliable estimates independent of the presence of a cover. In this work, we aim to bring vision-based weight assessment closer to this level and address point cloud-based weight estimation of patients which are covered by a blanket.

3.2.1.1 Related Work

General-purpose weight estimation. Body weight or body mass index estimation from full-body RGB, depth or RGB-D images has been addressed by numerous works, which predominantly rely on hand-crafted geometric or biometric features [Benalcazar et al., 2017; Jiang et al., 2019, 2020; Labati et al., 2012; Nguyen et al., 2014; Supranata et al., 2018; Velardo et al., 2012]. In a common approach, the subject is segmented from the background, features are subsequently extracted from the silhouette, and weight regression is performed by a neural network or support vector regression [Jiang et al., 2019; Labati et al., 2012; Nguyen et al., 2014; Velardo et al., 2012]. End-to-end learning of weight regression by means of deep convolutional neural networks has only been proposed by Nahavandi et al. [2017] and Altinigne et al. [2020], who utilize a U-Net [Ronneberger et al., 2015] and a ResNet [He et al., 2016] architecture, respectively.

Weight estimation in clinical settings. Most relevant to our work is weight estimation of patients lying in bed in a clinical environment. In an early work, Pirker et al. [2009] generate a merged point cloud from depth information of eight stereo camera pairs, which are placed around the bed, and compute body part-specific volumes by fitting a parametric human 3D model to the cloud. More recently, Pfitzner et al. [2015, 2016, 2017] predict the weight of a patient lying on a stretcher from a point cloud of a top-view depth camera. The authors start from a volume-based weight estimate in their initial work [Pfitzner et al., 2015] and gradually include PCA-based features

[Pfitzner et al., 2016] and contour-based features [Pfitzner et al., 2017], which are fed into a neural network for weight regression. Contrary to these feature-based methods, we proposed in our prior work [Bigalke et al., 2021b] to use basis point sets (BPS) [Prokudin et al., 2019] for end-to-end learning of weight estimation from point clouds. All of these methods assume the patient to be fully visible and not to be covered by a blanket.

Occlusion by a blanket. The occlusion of patients by a cover has been addressed by multiple works in the context of in-bed pose estimation. Achilles et al. [2016] train and evaluate their pose estimation framework on depth maps with simulated blankets. Other approaches aim to see through the blanket by means of particular sensors, namely thermal cameras [Liu et al., 2019a] or pressure mats [Casas et al., 2019]. In the context of weight estimation, however, such sensors appear less suitable than depth sensors, which capture richer geometric information. Multiple recent works estimate the patient's pose and 3D shape under blanket occlusions from multi-modal input data by fitting or predicting the parameters of a 3D human mesh model [Clever et al., 2020; Karanam et al., 2020; Singh et al., 2017; Yang et al., 2020; Yin et al., 2022]. Contrary to these works, our approach does not rely on a parametric model but explicitly addresses the occlusion problem in the input space.

Deep learning from point clouds. Deep learning from unstructured 3D point cloud data has attracted much attention in recent years [Guo et al., 2020]. In a pioneering work, the PointNet architecture [Qi et al., 2017a] applies a shared multi-layer perceptron to each input point individually and achieves a global permutation-invariant representation by max pooling. Subsequent works, such as PointNet++ [Qi et al., 2017b] and Dynamic Graph CNNs [Wang et al., 2019], incorporate the structure of local neighborhoods by means of hierarchical grouping and graph convolutions, respectively. In another line of work, point clouds are represented by 3D binary voxel grids, which are processed by 3D CNNs for shape classification [Maturana et al., 2015; Wu et al., 2015]. Beyond classification, the voxelized representation has been applied in the context of numerous other tasks, such as object detection [Song et al., 2016] or pose estimation [Moon et al., 2018].

3.2.1.2 Contributions

To our knowledge, this is the first work to learn weight prediction of covered patients. In light of the identified challenges in this setting, we regard weight estimation and occlusion by a blanket as two separate, independent problems and consequently propose a two-step solution. In a first step, we virtually remove the blanket by predicting the patient's shape without a blanket. For this purpose, we resort to a voxelized representation of point clouds and train a 3D U-Net [Çiçek et al., 2016] to accomplish

the task. This step is independent of the weight estimation problem and can be used as a pre-processing step for other tasks as well. In a second step, the actual weight regression is performed by a customized 3D CNN, which no longer needs to overcome the occlusion by a blanket. Thus, our proposed method essentially simplifies the overall problem and, as a beneficial by-product, provides a high degree of interpretability owing to the intermediate visualization of the uncovered patient.

The main contributions of this work can be summarized as follows:

- We introduce a novel two-stage pipeline of two 3D CNNs to predict the weight of covered patients from voxelized point cloud data.
- We propose to virtually uncover the patients to simplify the weight estimation problem and demonstrate the capability of a 3D U-Net to solve this task.

3.2.2 Methods

In this section, we present our approach for weight estimation of covered patients from point cloud data. We initially formalize the problem setup, subsequently give an overview of the proposed framework, and finally present its individual components in detail.

3.2.2.1 Problem Setup

Our goal is to develop a method that takes a 3D point cloud $\boldsymbol{X}^c \in \mathbb{R}^{N\times 3}$, which shows a covered patient lying in bed, as input and predicts the weight y of the patient. For this purpose, we assume access to a training dataset $\mathcal{T} = \{(\boldsymbol{X}_i^c, \boldsymbol{X}_i^{\setminus c}, y_i)\}_i$. It consists of pairwise point clouds $\boldsymbol{X}_i^c, \boldsymbol{X}_i^{\setminus c}$, which show a patient in unchanged pose with ($\boldsymbol{X}_i^c$) and without ($\boldsymbol{X}_i^{\setminus c}$) a cover, respectively, together with the ground truth weight y_i of the patient.

3.2.2.2 Framework

An overview of our proposed framework is visualized in Fig. 3.3. The core idea of our approach is to decouple the weight estimation problem itself from the occlusion problem caused by the blanket. To this end, we break the overall task down into two independent sub-problems, which are solved in two successive steps. In step one, we virtually remove the cover from the patient. By leveraging pairwise point clouds from $\mathcal{T}$, we learn to predict the patient's shape without a cover. This substantially simplifies the actual weight estimation performed in step two. The weight estimation problem is no longer complicated by a blanket and can thus be solved as for an uncovered patient.

In step one, we formally aim to learn a mapping from $\boldsymbol{X}^c$ to $\boldsymbol{X}^{\setminus c}$. Due to the inherent lack of point correspondences between two point clouds, however, it poses several technical challenges to properly define this problem for raw point clouds. Therefore, we

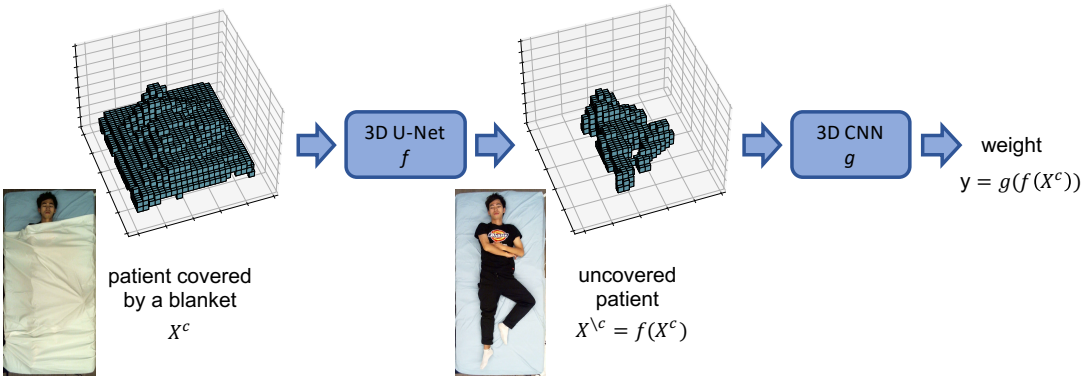


Fig. 3.3: Overview of our proposed two-stage pipeline for weight estimation of covered patients. Based on the voxelized input point cloud, we virtually uncover the patient with a 3D U-Net and perform the actual weight estimation based on the uncovered volume with a 3D CNN. Color images are only shown for better visualization and are not used in the pipeline.

resort to a voxelized representation such that two point clouds are naturally aligned. To voxelize a point cloud $\boldsymbol{X}$, a fixed-size cuboid volume around the cloud is discretized into a set of equally sized voxels, and a voxel is assigned the value 1 if it contains at least one point of the cloud and 0 otherwise. The resulting representation constitutes a binary 3D volume $\boldsymbol{X}^{vox} \in \{0,1\}^{h \times w \times d}$, where h, w and d denote the number of voxels in x-, y-, and z-direction, respectively. For ease of notation, we will omit the superscript from now on, and $\boldsymbol{X}$ refers to the voxelized representation of a point cloud.

This representation enables us to formalize the task in step 1. Specifically, we intend to learn a function f with parameters $\boldsymbol{\theta}_f$ that takes the volume of a covered patient $\boldsymbol{X}^c$ as input and outputs the volume of the uncovered patient $\boldsymbol{X}^{\setminus c}$, i.e., $f(\boldsymbol{X}^c; \boldsymbol{\theta}_f) = \boldsymbol{X}^{\setminus c}$. In our pipeline, we implement f as a 3D U-Net [Çiçek et al., 2016] as detailed in Sec. 3.2.2.3 and optimize its parameters by minimizing the cross-entropy loss

$$\mathcal{L}(\boldsymbol{\theta}_f; \mathcal{T}) = \sum_{(\boldsymbol{X}_i^c, \boldsymbol{X}_i^{\setminus c}) \in \mathcal{T}} \mathrm{CE}(f(\boldsymbol{X}_i^c; \boldsymbol{\theta}_f), \boldsymbol{X}_i^{\setminus c}) \tag{3.3}$$

with respect to $\boldsymbol{\theta}_f$. Here, $\mathrm{CE}(\cdot, \cdot)$ denotes the element-wise binary cross-entropy loss function.

Once we have learned to uncover the patient in step 1, we subsequently learn a function g with parameters $\boldsymbol{\theta}_g$, which takes the volume of the uncovered patient predicted by f, namely $f(\boldsymbol{X}^c; \boldsymbol{\theta}_f)$, as input and outputs the patient's weight y, i.e.,

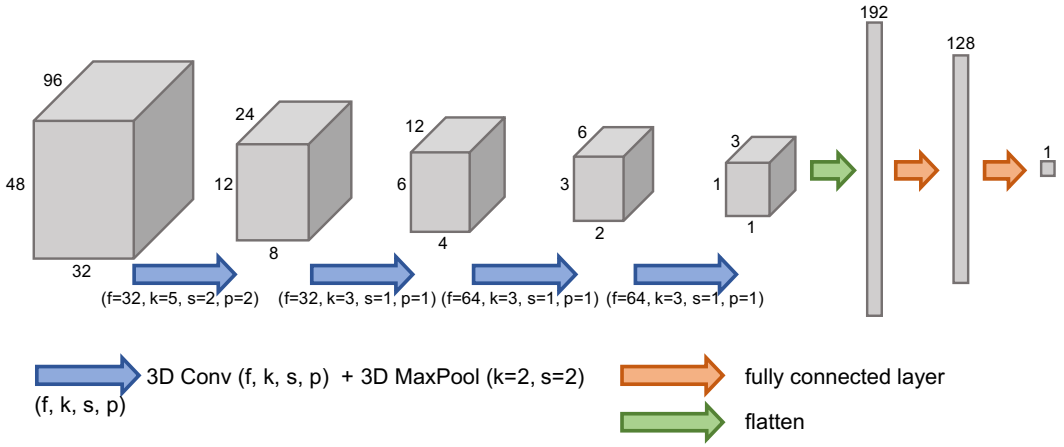


Fig. 3.4: Visualization of our proposed 3D CNN for weight estimation from a 3D volume. 3D convolutions are characterized by the number of output feature channels f, kernel size k, stride s and padding p. k, s and p are identical in all spatial dimensions.

$y = g(f(\boldsymbol{X}^c; \boldsymbol{\theta}_f); \boldsymbol{\theta}_g)$. We implement g as a 3D CNN introduced in Sec 3.2.2.4. The optimization is performed by minimizing the mean squared error loss

$$\mathcal{L}(\boldsymbol{\theta}_f, \boldsymbol{\theta}_g; \mathcal{T}) = \sum_{(\boldsymbol{X}_i^c, y_i) \in \mathcal{T}} \Big[g(f(\boldsymbol{X}_i^c; \boldsymbol{\theta}_f); \boldsymbol{\theta}_g) - y_i \Big]^2 \tag{3.4}$$

with respect to $\boldsymbol{\theta}_g$ while keeping $\boldsymbol{\theta}_f$ fixed.

3.2.2.3 3D U-Net

The architecture of the 3D U-Net strictly follows the original implementation in [Çiçek et al., 2016]. In short, the 3D U-Net comprises a contracting encoder path and an expanding decoder path, both including four levels of different resolutions. In the encoder path, features are extracted by means of 3D convolutions and downsampling is realized by max pooling operations. In the decoder path, low-resolution features are gradually upsampled by transposed convolutions and combined with high-resolution features of equal resolution from the encoder path. This is realized by skip connections and subsequent 3D convolutions, which merge the features.

3.2.2.4 3D CNN for Weight Regression

The architecture of our proposed 3D CNN for weight regression from 3D volumes is illustrated in Fig. 3.4. During optimization of the architecture, we found it crucial to downsample the input volume to very low resolution before final weight regression with the fully-connected network heads. To realize this, the architecture starts with

a 3D convolution with a kernel size of $5 \times 5 \times 5$ and a stride of 2, followed by a max pooling operation with kernel size $2 \times 2 \times 2$. Subsequently, the resulting feature map is alternately processed by 3D convolutions with kernel size $3 \times 3 \times 3$ and stride 1, and max pooling operations with kernel size $2 \times 2 \times 2$, which further reduce the spatial resolution. That way, the input volume is downsampled by a factor of 2^5 in each spatial dimension. At the same time, the number of feature channels is gradually increased to 64. Each convolutional layer is followed by a batch normalization layer and a ReLU non-linearity. After the last convolutional layer, the feature maps are flattened and forwarded by the fully connected network heads. These consist of a fully connected layer with 128 neurons, followed by ReLU non-linearity, dropout (p=0.8) and the output neuron with linear activation.

3.2.3 Experiments and Results

3.2.3.1 Experimental Setup

Dataset. We evaluate our method on the SLP dataset [Liu et al., 2022a]. The dataset consists of depth frames of 109 subjects, which take 45 different poses while lying in bed in supine and lateral (left and right) position. For each pose, three nearly identical depth frames are taken, which only differ in terms of the cover condition (no cover, thin cover, thick cover) and are thus ideal for learning the virtual removal of a blanket. Covers have a thickness of around 1 and 3 mm, respectively. For each frame, we detect all pixels belonging to bed and patient with the help of depth thresholding and clustering. We then transform these pixels to a point cloud using the internal camera parameters. Body weights of the subjects range from 43.7 to 105.1 kg and exhibit a mean of 68.0 kg and a standard deviation of 12.7 kg.

The dataset includes two different setups. 102 subjects were recorded in a lab setting, and the remaining seven subjects were recorded in a simulated hospital room. The two settings differ in terms of the used beds, mattresses, sheets, blankets, and sensor-to-bed distances, resulting in a substantial domain shift. We conduct the main experiments with the 102 subjects, training the model on the first 60 subjects and reporting results for the remaining 42 subjects. The seven subjects recorded in the hospital room are used for cross-domain evaluation. Generally, the model is jointly trained under both cover conditions (thin cover, thick cover) and positions (supine, lateral) while evaluation is performed for each cover type and position separately.

Implementation details. We implement our proposed framework in PyTorch [Paszke et al., 2019]. Network parameters are optimized with the ADAM optimizer. The initial learning rate is set to 0.001, and we use a batch size of 16. The 3D U-Net is trained for 50 epochs, whereby the learning rate is divided by 10 after 30 epochs. The 3D CNN for weight regression is trained for 120 epochs, whereby the learning rate is divided

by 10 at epoch 60 and 100. For voxelization of the raw point clouds, we discretize a cuboid volume of size $1.7\,\mathrm{m} \times 2.4\,\mathrm{m} \times 0.7\,\mathrm{m}$ into $48 \times 96 \times 32$ voxels with edge lengths of $3.5\,\mathrm{cm} \times 2.5\,\mathrm{cm} \times 2.2\,\mathrm{cm}$. The size of the cuboid volume has been chosen such that it covers all mean-centered clouds from the training set. When training the U-Net in step 1, we pre-process the target point cloud of the uncovered patient before voxelization to segment the patient from the bed [Bigalke et al., 2021b].

Baseline methods. We consider two baseline methods for weight estimation from covered patients. First, we train the plain 3D CNN without preceding U-Net for weight regression. Second, for a fair comparison regarding the number of model parameters, we train the composition of 3D U-Net and 3D CNN for weight regression in an end-to-end fashion without minimizing the intermediate loss in Eq. (3.3). As an upper bound, we train the plain 3D CNN to predict the weight of uncovered patients. To further investigate the effect of occlusions by a blanket on weight estimation performance, we additionally learn weight estimation of covered and uncovered patients with a PointNet architecture [Qi et al., 2017a] and the BPS-based fully connected network from [Bigalke et al., 2021b], which both operate on raw point cloud data instead of a voxel-based representation.

Metrics. As error metric, we use the mean absolute error (MAE) of the predicted weight estimates on the test set. Each experiment is repeated 5 times, and we report mean and standard deviation of the MAE.

3.2.3.2 Results

Quantitative results of our main experiments are presented in Tab. 3.2 and reveal four major insights.

First, we note that weight estimates in supine and lateral position have a similar accuracy under all cover conditions and for all models, whereby estimates in lateral position are in most cases slightly better.

Second, we observe that occlusions by both a thin and a thick cover lead to a substantial degradation of the performance of PointNet, BPS and 3D CNN. As expected, the performance for the thick cover is consistently slightly worse than for the thin cover. As an interesting side note, we notice that the voxel-based 3D CNN clearly outperforms both point cloud-based approaches under all three cover conditions. But even for the 3D CNN, we observe a relative increase in MAE of 39 % for the thin cover and of 44 % for the thick cover (averaged over supine and lateral position). This confirms the need for weight estimation methods that explicitly address the specific challenges caused by a partial covering of the patient by a blanket.

Third and most importantly, it can be seen that our proposed method successfully addresses these challenges and considerably outperforms the baseline models under

Table 3.2: Weight estimation results on the 42 subjects from the lab setting of the SLP dataset for different cover conditions and positions. We compare the MAE, measured in kg, of several baseline methods to our proposed framework.

Method	No cover	Thin cover	Thick cover	
PointNet	4.66 ± 0.23	5.97 ± 0.12	6.14 ± 0.11	supine position
BPS	4.33 ± 0.18	6.13 ± 0.11	6.68 ± 0.13	
3D CNN	$\mathbf{3.86 \pm 0.12}$	5.36 ± 0.06	5.56 ± 0.07	
3D U-Net + 3D CNN (weight regr. only e2e)	-	5.25 ± 0.03	5.44 ± 0.07	
3D U-Net + 3D CNN (ours)	-	$\mathbf{4.61 \pm 0.06}$	$\mathbf{4.71 \pm 0.13}$	
PointNet	4.13 ± 0.25	5.92 ± 0.06	6.00 ± 0.04	lateral position
BPS	4.07 ± 0.19	6.34 ± 0.17	6.64 ± 0.17	
3D CNN	$\mathbf{3.80 \pm 0.05}$	5.28 ± 0.06	5.45 ± 0.04	
3D U-Net + 3D CNN (weight regr. only e2e)	-	5.17 ± 0.04	5.24 ± 0.05	
3D U-Net + 3D CNN (ours)	-	$\mathbf{4.51 \pm 0.04}$	$\mathbf{4.54 \pm 0.10}$	

both cover conditions. Specifically, again averaged over supine and lateral position, the MAE is reduced by 14.3 % for the thin cover and by 16.0 % for the thick cover with respect to the baseline 3D CNN. Moreover, the gap between weight estimates with and without a cover achieved by the 3D CNN is reduced by 51.0 % for the thin cover and by 52.5 % for the thick cover.

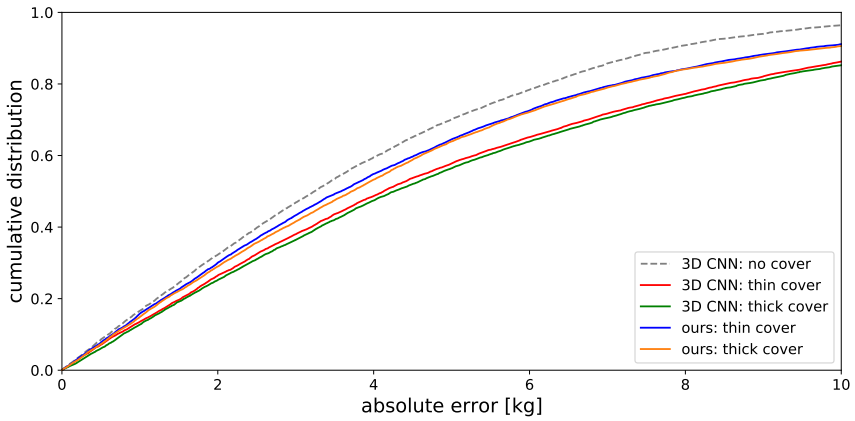


Fig. 3.5: Cumulative distribution of weight estimation errors for our method and the 3D CNN baseline. Our method clearly improves on the baseline model under both cover conditions and reduces the gap to weight estimates from uncovered patients.

Table 3.3: Quantitative results of the ablation experiment on the 42 subjects from the lab setting of the SLP dataset. We show the Dice overlap and the average surface distance in mm between the volume of the uncovered patient predicted by the 3D U-Net and the corresponding ground truth. For reference, we report the initial Dice overlap and surface distance between uncovered and covered patient volume.

		Thin cover		**Thick cover**	
Method	**Metric**	**Supine**	**Lateral**	**Supine**	**Lateral**
Initial	Dice	31.0	28.4	30.7	27.7
3D U-Net		76.1 ± 0.1	75.0 ± 0.2	76.1 ± 0.1	73.9 ± 0.1
Initial	Surface distance	12.1	12.4	11.8	12.6
3D U-Net		5.0 ± 0.1	4.9 ± 0.1	5.0 ± 0.1	5.1 ± 0.1

Fourth, we observe that the composition of 3D U-Net and 3D CNN, exclusively trained for weight regression in an end-to-end fashion, only slightly improves on the performance of the plain 3D CNN. We deduce that the superiority of our method is merely to a small extent due to increased model capacity. Rather, it is crucial to explicitly learn to virtually uncover the patient.

Finally, we provide a more detailed comparison of our method and the 3D CNN by plotting the cumulative distributions of absolute errors in Fig. 3.5. The observable trends are in line with the findings discussed above.

3.2.3.3 Ablation Study

In the ablation experiment, we intend to assess the capability of the U-net to uncover a patient in a more direct way. For quantitative evaluation, we compute the average surface distance from the target volumes $\boldsymbol{X}^{\setminus c}$ to the outputs of the U-Net $f(\boldsymbol{X}^c)$ as well as the Dice overlap between both volumes. For reference, we report the initial average surface distance and the initial Dice overlap between $\boldsymbol{X}^{\setminus c}$ and $\boldsymbol{X}^c$. Results of the experiment are shown in Tab. 3.3. Under both cover conditions and positions the U-Net more than doubles the initial Dice overlap and more than halves the initial surface distance. This demonstrates its capability to virtually uncover the patient with an adequate accuracy.

Regarding the Dice score, we observe a small gap between supine and lateral positions. This gap is most likely not due to less accurate predictions but due to the sensitivity of the Dice score to the size of ground truth volumes. Ground truth volumes in lateral positions are represented by less voxels than in supine positions such that errors have a larger negative impact on the Dice score. Referring to the average surface distance, which is less sensitive to the size of objects, the score is similar for both positions.

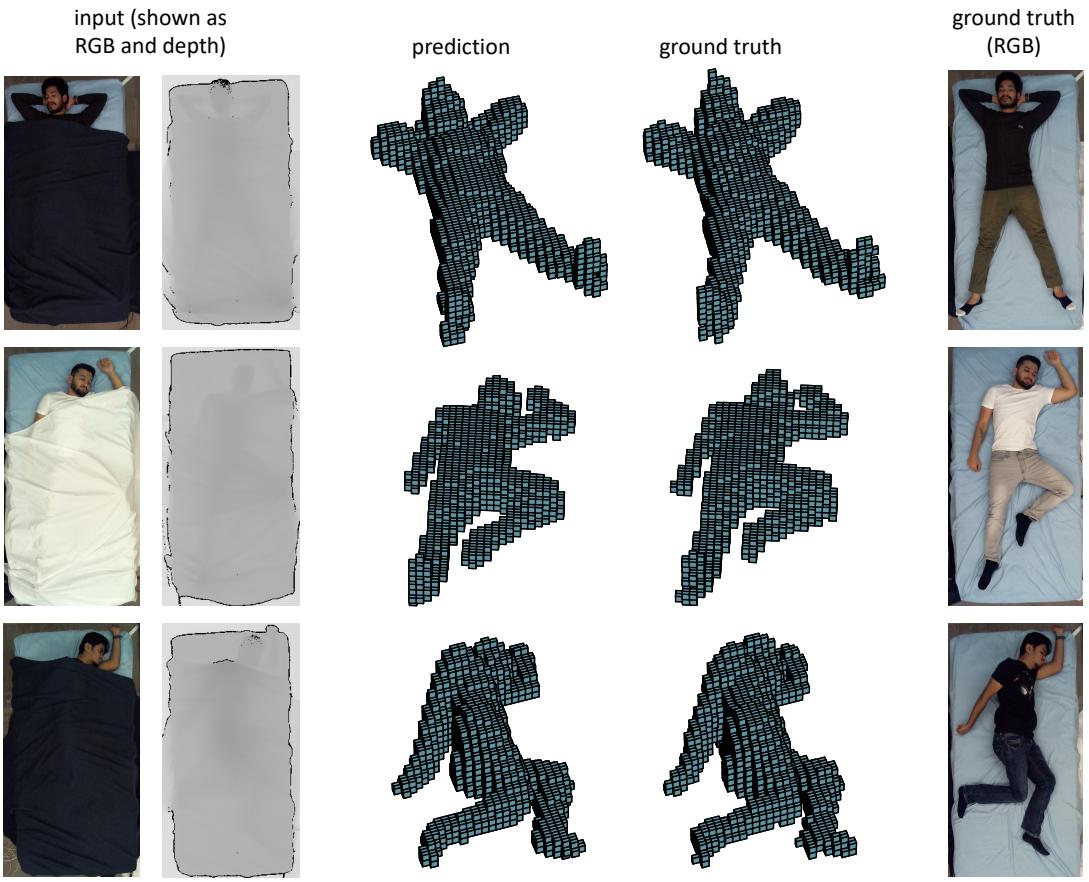


Fig. 3.6: Qualitative results of the ablation study on three samples from the SLP dataset. From left to right, each row shows RGB and depth image of the corresponding input volume, the prediction by our method, ground truth and the RGB image of the ground truth scene. RGB and depth images are shown for better visualization while our framework processes and outputs volumetric representations.

Qualitative results of the ablation study are presented in Fig. 3.6 and demonstrate that the predictions by the 3D U-Net are visually compelling as well. 3D volumes of the subjects are precisely recovered under varying poses, and even hollow space under the blanket is largely correctly classified as not being part of the human body.

3.2.3.4 Cover Detection for Full Automation

In practical applications, it is desirable to estimate the weight of both covered and uncovered patients with a single fully automatic pipeline. For this purpose, we initially need to classify if patients are covered or uncovered. Subsequently, the weight of

Table 3.4: Results of the cross-domain evaluation on the 7 subjects from the hospital room of the SLP dataset. We compare the MAE, measured in kg and averaged over supine and lateral position, of several baseline methods to our proposed framework.

Method	No cover	Thin cover	Thick cover
PointNet	6.99 ± 0.34	9.94 ± 0.80	10.91 ± 0.63
BPS	6.69 ± 0.57	9.57 ± 1.21	12.41 ± 2.41
3D CNN	$\mathbf{6.62 \pm 0.39}$	8.76 ± 0.54	9.85 ± 0.47
3D U-Net + 3D CNN (weight regr. only e2e)	-	8.93 ± 0.61	9.75 ± 0.40
3D U-Net + 3D CNN (ours)	-	$\mathbf{7.05 \pm 0.20}$	$\mathbf{8.51 \pm 0.40}$

uncovered patients can be estimated by the 3D CNN while the weight of covered patients is assessed by our entire framework of 3D U-Net and 3D CNN.

To automate cover detection, we train the baseline 3D CNN as a binary classifier. The network is trained for 10 epochs with an initial learning rate of 0.001, which is divided by 10 after 5 and 8 epochs. The 3D CNN achieves a classification accuracy of 100.0 %. Thus, subsequent weight estimation of a patient is virtually always performed by the appropriate framework and the accuracies reported in Tab. 3.2 remain unchanged in a fully automatic pipeline.

3.2.3.5 Cross-Domain Evaluation

To examine the cross-domain robustness of our method, we evaluate all models from Tab. 3.2 (trained on the 60 subjects from the lab setting) on the seven subjects from the simulated hospital room. Due to the substantial domain shift between training and test data, this is a challenging setting. Specifically, the varied sensor-to-bed distance leads to a different distribution of points in 3D space, the change of the bed alters the geometry of the entire scene, and mattresses and bed sheets might differ in terms of firmness and flexibility. Quantitative results of the experiment are shown in Tab. 3.4. On the one hand, our proposed method outperforms all baseline methods even in this more complicated setting. Compared to the baseline 3D CNN, the MAE is reduced by 19.5 % for the thin cover and by 13.6 % for the thick cover. This indicates that our two-step solution is more robust than end-to-end approaches. On the other hand, however, we observe that all methods exhibit a severe performance drop under all cover conditions compared to the in-domain evaluation (Tab. 3.2). For instance, the baseline 3D CNN deteriorates by 73 %, 65 % and 79 % for no cover, thin cover and thick cover, respectively, and our method degrades by 55 % for the thin cover and by 84 % for the thick cover. We conclude that the domain shift is a serious problem that needs to be addressed by methods from domain adaptation [Wang et al., 2018]. However, this is

beyond the scope of this work and is left to future research. Our results constitute an initial baseline for such work.

3.2.4 Discussion and Conclusion

We proposed a novel framework, consisting of a 3D U-Net and a 3D CNN, for weight estimation of covered patients from voxelized point clouds. In our experiments on the SLP dataset, we demonstrated that the 3D U-Net is capable of virtually uncovering a patient and to thus simplify the subsequent weight regression with the 3D CNN. Specifically, our method improved the weight estimation performance compared to baseline methods by up to 16 % and reduced the gap to weight estimates of uncovered patients by up to 52 %. Even in presence of a thick cover, our method achieves a higher accuracy (MAE=4.62 kg, corresponding to a mean relative error (MRE) of 7.0 %) than estimates by clinical staff, which exhibit MREs of 8.1 to 8.4 % in [Fernandes et al., 1999] and of 7.7 to 11.0 % in [Menon et al., 2005]. The accuracy of our method can even further be improved to an MAE of 3.8 kg and an MRE of 5.7 % by statistical averaging over multiple weight estimates for the same subject from different frames with varying poses as in [Bigalke et al., 2021b]. Altogether, our work constitutes an important step towards fully automatic weight estimation, which should ideally provide reliable weight estimates independent of any specific conditions. However, the occlusion of a patient by a cover is only one among multiple possible challenges which might occur in clinical practice. Another important problem, for instance, consists in the presence of a domain shift between training and test data, which lead to a substantial performance drop in our cross-domain experiment. To improve generalization to diverse settings (different room setups, beds, mattresses, viewpoints, etc.), future work could incorporate techniques from domain adaptation [Wang et al., 2018] or domain generalization [Zhou et al., 2022a] into the weight estimation framework.

Beyond the specific task of weight estimation, we believe that our approach to virtually uncover the patient constitutes a valuable tool for medical computer vision in general. Since the method is independent of the weight estimation problem, it can be seen as a generic pre-processing step with the potential to simplify any task that needs to overcome occlusions by a blanket. In particular, the integration of the approach into existing frameworks for in-bed pose and shape estimation [Achilles et al., 2016; Singh et al., 2017] appears promising and is of high interest for future work.

Chapter 4

3D In-Bed Human Pose Estimation Under Domain Shifts

The second methodological chapter adheres to the application of in-bed patient monitoring but now addresses the local detection problem of human pose estimation instead of the previously studied global weight regression. Rather than in a fully-supervised setting, the task is treated in an unsupervised domain adaptation scenario, aiming to overcome geometric domain shifts, which caused severe performance losses in the previous chapter. As the methodological contribution, the chapter presents two novel adaptation strategies that exploit prior knowledge about the output space, i.e., the human anatomy, to guide the learning process in the target domain. The underlying key idea is to embed the prior anatomical knowledge into a differentiable loss function that measures the violation of anatomical constraints on the human skeleton graph by predicted poses. It is then demonstrated that the loss function is both a powerful source of direct supervision by constraining predictions to the space of anatomically plausible poses and a valuable quality measure to filter pseudo labels for self-training under the Mean Teacher paradigm.

4.1 Anatomy-Guided Domain Adaptation

This section has been published in [Bigalke et al., 2023a]. According to the Contributor Roles Taxonomy (CRediT), the contributions of the author of this thesis to the publication are: Conceptualization (together with L.H., J.D., P.R., M.P.H.), Methodology, Software, Investigation, Resources (together with J.D., C.H.) Writing – Original Draft, Writing – Review & Editing (together with all co-authors), Visualization. The source code is available at `https://github.com/multimodallearning/da-3dhpe-anatomy`.

4.1.1 Introduction

3D human pose estimation is a fundamental problem in computer vision and the basis for various higher-level tasks, such as posture recognition [Liu et al., 2020b] and action recognition [Song et al., 2021]. These tasks, in turn, open up a wide range of

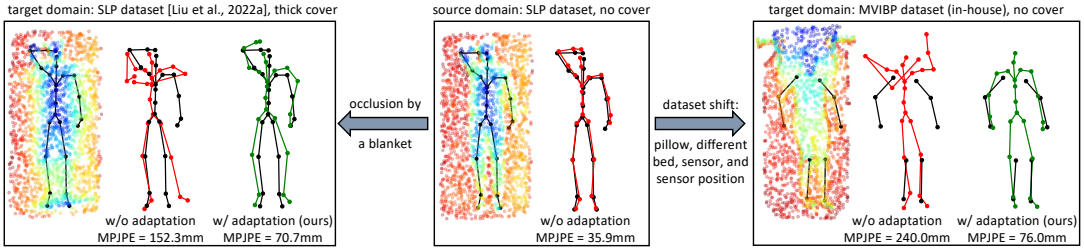


Fig. 4.1: Visualization of two domain shifts for point cloud-based in-bed pose estimation and their impact on model performance. We show input point clouds from the source domain and two different target domains (colors encode the depth in z-direction) alongside the ground truth poses in black, the predictions by a source-trained baseline model in red, and the predictions by our adaptation method in green. While the in-domain prediction of the baseline model is close to perfect, the predictions on the shifted domains are anatomically implausible and highly inaccurate (in terms of the mean per joint position error MPJPE). Adaptation with our anatomy-guided method substantially improves the accuracy and plausibility of the pose estimates.

applications in the field of human-computer interaction, which are in high demand in the automotive or gaming sectors, for instance [Chen et al., 2020]. The healthcare sector can also benefit from automatic pose estimation as pose-based assistance and monitoring systems promise to relieve clinical staff and improve patient safety and care. On the one hand, tracking the 3D joint positions of clinicians enables automated documentation, analysis, and optimization of clinical workflows [Mascagni et al., 2021; Rodrigues et al., 2022]. On the other hand, in-bed pose estimation, the application focus of this work, offers great potential for automatic patient monitoring: A pose-driven monitoring system could analyze movements [Chen et al., 2018], detect potentially critical events [Jähne-Raden et al., 2019], diagnose pathological movement patterns [Cunha et al., 2016], and prevent pressure ulcers [Ostadabbas et al., 2012].

In recent years, deep learning has substantially advanced the state of the art in general and clinical human pose estimation [Chen et al., 2020], making the deployment of the above systems more tangible. Nonetheless, several challenges remain, particularly in the clinical setting. First, data privacy and highly variable lighting conditions, including complete darkness, preclude the use of standard color images. As a remedy, we advocate the use of 3D point cloud data. Point clouds are not only anonymity-preserving [Silas et al., 2015] and insensitive to lighting conditions but also inherently preserve the 3D structure of the scene, making them a natural modality for 3D pose estimation. Second, the performance of deep learning-based methods strongly depends on access to large-scale labeled datasets [Ionescu et al., 2013]. The annotation of 3D poses, however, is generally laborious and even more involved in clinical settings: Data access is often restricted, and accurate annotations under severe occlusions, e.g., caused by blankets

in the case of patient monitoring, are only feasible under the controlled conditions of a lab study [Liu et al., 2022a]. Therefore, it is crucial to take full advantage of existing datasets [Liu et al., 2022a; Srivastav et al., 2018] as a training resource across diverse target domains. However, this is hampered by the poor generalization of deep models under domain shifts, resulting in severe performance drops when deploying a model in a shifted domain [Wang et al., 2021b]. In the clinical setting, such shifts can be due to varying room setups/environments in different hospitals/countries or changing visibility conditions (no blanket, blanket), as visualized in Fig. 4.1. While supervised fine-tuning on shifted data could alleviate the problem, it is often no viable solution given the high annotation costs. Instead, it is desirable to adapt a model from a labeled source to an unlabeled target domain in an unsupervised fashion. This can be realized by domain adaptation (DA) [Wang et al., 2018], the methodological focus of this work.

Classical unsupervised domain adaptation (UDA) methods approach the problem by jointly accessing data from both domains. Given the importance of data protection in the medical sector, however, this cannot always be guaranteed. Instead, it is a realistic scenario that the provider of a pose estimation model and its end-user are not able or willing to exchange their data. Consequently, the end-user needs to adapt the provided pretrained source model to the target domain without accessing the source data, denoted as source-free domain adaptation (SFDA) [Kundu et al., 2020]. With this in mind, it is desirable to have a universal DA method applicable to both UDA and SFDA as needed.

A popular branch of DA methods couples the supervised learning on labeled source data with the alignment of the distributions of source and target features, realized by discrepancy minimization [Tzeng et al., 2014] or adversarial learning [Ganin et al., 2015; Tzeng et al., 2017]. The learned target features, however, are not explicitly optimized for the actual task, and domain invariance does not guarantee task relevance. This problem can be addressed by performing the adaptation in the output space of the target domain, implemented by adversarial optimization [Tsai et al., 2018; Yang et al., 2018] and direct supervision with pseudo labels [Mu et al., 2020; Yang et al., 2021] in prior work. However, adversarial optimization is complex and unstable, and pseudo labels are noisy and can thus misguide the learning process. As an additional downside, adversarial methods are not applicable to SFDA since they require simultaneous access to both domains.

4.1.1.1 Contributions

We propose to overcome these problems by guiding the adaptation process with the aid of prior knowledge about human anatomy. Such prior knowledge contains valuable information about the expected pose distribution in the output space, which can be vastly restricted by excluding anatomically implausible poses that cannot be taken by a human. Notably, the prior knowledge is domain-independent and thus invariant under

the domain shifts discussed above. We propose two different strategies to exploit this knowledge (see Fig. 4.2 for an overview). First, we directly supervise predictions in the target domain by explicitly constraining them to the space of anatomically plausible poses. To this end, we derive three anatomical loss functions that penalize predictions with asymmetric limb lengths, implausible bone lengths, and implausible joint angles. Second, we filter noisy pseudo labels for self-training according to their anatomical plausibility, measured with our anatomical loss functions. Concretely, we incorporate this technique in the Mean Teacher paradigm [French et al., 2018; Tarvainen et al., 2017], where pseudo labels from the teacher are only used for supervision if they are more plausible than the current prediction of the learning student model. We unify these two strategies in a point cloud-based framework. It performs output adaptation without intricate adversarial optimization and mitigates noisy supervision through anatomical guidance. Moreover, it does not require simultaneous access to source and target domain and is thus applicable to both UDA and SFDA.

In summary, the main contributions of this work are:

1. We introduce an anatomy-guided domain adaptation method for point cloud-based 3D human pose estimation, including two complementary adaptation strategies based on prior anatomical knowledge.

2. We derive an anatomical loss function that constrains pose predictions in the target domain to the space of plausible poses by penalizing asymmetric limb lengths, implausible bone lengths, and implausible joint angles.

3. We propose to filter pseudo labels based on their anatomical plausibility and incorporate the concept into the Mean Teacher paradigm.

4. We demonstrate the efficacy of our method in the context of in-bed pose estimation for both UDA and SFDA under two different scenarios: the adaptation between the different environments of two datasets—the public SLP dataset [Liu et al., 2019a, 2022a] and a newly created dataset—and from uncovered to covered patients. Under all settings, our method is superior to a comprehensive set of state-of-the-art domain adaptation methods, which we adapted to the given problem.

A preliminary conference version of this work appeared at MIDL 2022 [Bigalke et al., 2022a]. In this journal version, we extend this work as follows: 1) We substantially extend the discussion of related works. 2) We give a more detailed description of the method and derive the anatomical loss function from a constrained optimization problem. 3) Extending the method, we use the anatomical loss not only for direct supervision but propose to use it as a criterion for filtering pseudo labels. 4) We formalize anatomy-constrained optimization and anatomy-guided filtering of pseudo labels in a unified framework applicable to UDA and SFDA. 5) We perform extensive additional experiments, demonstrating the efficacy of our method under a second

adaptation scenario (using our recently captured dataset) and in the challenging SFDA setting.

4.1.1.2 Related Work

Human pose estimation. 2D and 3D human pose estimation from regular 2D grid data is a widely studied problem, with most works focusing on RGB [Sun et al., 2019a; Xiao et al., 2018] and depth images [Haque et al., 2016; Moon et al., 2018] as the input modalities. Since our work treats point cloud-based pose estimation (see next paragraph), we refer the reader to Chen et al. [2020] for a comprehensive survey of grid-based methods and summarize works with clinical applications. The first line of such works addresses pose estimation of clinical staff in the operating room. While early methods rely on multi-view RGB [Belagiannis et al., 2016] and RGB-depth [Kadkhodamohammadi et al., 2017] images, more recent methods exploit multi-view [Hansen et al., 2019] and low-resolution [Srivastav et al., 2019] depth images to prevent privacy concerns by clinicians and patients. Another stream of methods treats in-bed patient pose estimation. Besides compliance with data protection, the primary challenge in this task consists of severe occlusions by blankets. Multiple works aim to see under the blanket with the help of suitable sensors. Liu et al. [2019a] estimate 2D poses from thermal images, and Casas et al. [2019] and Davoodnia et al. [2021] use pressure maps to estimate 3D and 2D poses, respectively. Alternatively, several methods learn to predict the pose and shape parameters of a human mesh model [Loper et al., 2015] under blanket occlusions by fusing multiple modalities, including thermal, pressure, depth, and RGB images [Karanam et al., 2020; Yang et al., 2020; Yin et al., 2022]. However, all the above methods require ground truth annotations under the blanket, which are difficult to obtain in a real-world application. As a remedy, Achilles et al. [2016], Clever et al. [2020], and Clever et al. [2022] train their models on synthetic depth or pressure maps of covered patients, and Afham et al. [2022] and Chi et al. [2022] perform domain adaptation from labeled uncovered to unlabeled covered subjects based on thermal images (see Sec. 4.1.1.2 – Domain adaptive pose estimation).

Point cloud-based pose estimation. Compared to all the above modalities, point clouds stand out by inherently preserving the 3D structure of the scene. Their unstructured nature, however, prevents the use of standard convolutions, complicating the processing with deep neural networks. The pioneering PointNet [Qi et al., 2017a] addressed the issue by extracting point-wise spatial representations, which are aggregated by max-pooling. To capture local geometric structures, various follow-up works proposed hierarchical grouping [Qi et al., 2017b] and generic convolutions [Li et al., 2018; Liu et al., 2019b; Wang et al., 2019; Wu et al., 2019; Xu et al., 2021] applicable to unstructured data.

Prior works on point cloud-based keypoint estimation primarily focus on hand pose estimation. The Hand PointNet [Ge et al., 2018a] employs the PointNet++ [Qi et al., 2017b] architecture for direct regression of the joint coordinates, followed by a refinement network for the fingertips. In another work, Ge et al. [2018b] extend PointNet++ to a stacked hourglass architecture [Newell et al., 2016] and estimate joint coordinates by combined regression of heatmaps and offset vectors. Li et al. [2019b] regress separate pose estimates from the representations of each input point, which are aggregated in a final estimate. Hermes et al. [2022] reduce the complexity of the regression problem by predicting joint coordinates as the weighted sum over the input points, complemented by a set of support points. In our work, we employ the Dynamic Graph CNN (DGCNN) by Wang et al. [2019] as the backbone architecture and formulate human pose regression similar to Hermes et al. [2022].

Domain adaptation. Classical UDA assumes joint access to a labeled source and a shifted unlabeled target domain. We broadly classify UDA methods according to the level where the adaptation is performed: the input level, the feature level, and the output level. The idea of input-level adaptation [Hoffman et al., 2018; Li et al., 2019d; Murez et al., 2018] is to align the image styles or pixel-level distributions of source and target data through image-to-image translation modules like CycleGAN [Zhu et al., 2017b] or CUT [Park et al., 2020]. By contrast, feature-level adaptation aims at aligning intermediate feature distributions from the source and target domain. This was realized by minimizing explicit distance measures between both distributions [Rozantsev et al., 2018; Sun et al., 2016; Tzeng et al., 2014], by adversarial learning with a domain discriminator [Ganin et al., 2015; Saito et al., 2019; Tzeng et al., 2017], and by simultaneously learning an auxiliary self-supervised task in both domains [Bousmalis et al., 2016; Ghifary et al., 2016; Sun et al., 2019b]. Finally, Luo et al. [2019] and Tsai et al. [2018] proposed to align source and target distributions in the output space by training the entire task network in an adversarial manner against a discriminator.

An alternative technique for output-level adaptation is self-training with pseudo labels [Zou et al., 2018]. The basic idea is to alternately generate pseudo labels on unlabeled target data with the current model and to re-train the model using these labels. A specific form of self-training is the Mean Teacher paradigm [Tarvainen et al., 2017], where pseudo labels are continuously generated by a teacher model, whose weights are given as the exponential moving average of the weights of the learning student network. Initially introduced for semi-supervised classification, the concept was transferred to domain adaptation by French et al. [2018] and subsequently adapted to diverse tasks, including object detection [Cai et al., 2019; Deng et al., 2021], medical image segmentation [Li et al., 2020; Perone et al., 2019], and medical registration [Bigalke et al., 2022b]. However, pseudo labels are typically noisy, which can hamper the adaptation process. Therefore, multiple works guide the supervision with pseudo

labels through uncertainty estimates, computed by Monte Carlo Dropout [Wang et al., 2020b; Wang et al., 2021c; Yu et al., 2019], as the predictive variance under input perturbations [Zhou et al., 2022b] and among different network heads [Zheng et al., 2020; Zheng et al., 2021], and as the reconstruction error of a denoising autoencoder [Adiga Vasudeva et al., 2022].

Unlike UDA, SFDA aims to adapt a pre-trained source model to the target domain without accessing source data. Thus, the explicit alignment of both domains is no longer feasible. To overcome this problem, Kurmi et al. [2021] and Liu et al. [2021b] generate synthetic source data by exploiting the pre-trained source model. In a different approach, the source model is directly adapted to the target domain by entropy minimization [Wang et al., 2021a], entropy minimization guided by shape priors [Bateson et al., 2020], and information maximization [Liang et al., 2020]. Similar to UDA, self-training with reliable [Kundu et al., 2021] or denoised [Chen et al., 2021] pseudo labels and the Mean Teacher [Hegde et al., 2021; Wang et al., 2022] were also deployed in source-free settings. Another line of works achieved SFDA by progressively adapting the statistics of the BatchNorm layers to the target domain [Klingner et al., 2022; Liu et al., 2021a; Zhang et al., 2022].

Point cloud-based domain adaptation. The vast majority of point cloud-based DA methods perform feature-level adaptation through self-supervision and mainly differ by the pretext tasks. The proposed tasks include the reconstruction of a deformed point cloud [Achituve et al., 2021], solving 3D puzzles [Alliegro et al., 2021], and learning the implicit function that represents the underlying shape model [Shen et al., 2022]. Some works suggested multi-level self-supervised learning at global and local scales [Fan et al., 2022; Zou et al., 2021]: global tasks are scale and rotation prediction, while local tasks consist in the reconstruction of local areas and the localization of local distortions. Besides self-supervised DA, Qin et al. [2019] proposed multi-level alignment of local and global features, and Cardace et al. [2021] introduced a point cloud-specific self-training strategy with pseudo label refinement.

Domain adaptive pose estimation. Many of the introduced concepts for domain adaptation were adapted to general human/animal pose estimation and clinical human pose estimation. Martínez-González et al. [2018] performed adversarial feature alignment for 2D human pose estimation from depth maps. Liu et al. [2022c] proposed semantically aware feature alignment coupled with a skeleton-aware pose refinement module for 3D human pose estimation from RGB images. Yang et al. [2018] addressed the same task through adversarial output adaptation. Cao et al. [2019], Li et al. [2021], and Mu et al. [2020] suggested different forms of self-training for 2D animal pose estimation. Kim et al. [2022] proposed a multi-level adaptation method for 2D human pose estimation, comprising style transfer at the input level and self-training with the Mean Teacher at

the output level. In the clinical context, Srivastav et al. [2022] presented a self-training framework with domain-specific normalization layers [Chang et al., 2019] for 2D clinician pose estimation and instance segmentation in the operating room. Two multi-level adaptation strategies for 2D in-bed pose estimation, adapting from uncovered to covered patients on thermal images, were presented by Afham et al. [2022] and Chi et al. [2022]. The authors combined image-to-image translation at the input level with extreme augmentations and knowledge distillation [Afham et al., 2022] and with adversarial feature alignment and self-training [Chi et al., 2022], respectively.

Compared to all discussed works, our method includes three essential methodical novelties. First, it is the first approach to domain adaptive human pose estimation from 3D point clouds. Second, unlike guiding self-training with pseudo labels through uncertainty estimates, we filter pseudo labels based on plausibility constraints derived from prior knowledge about the output space distribution. Third, unlike adversarial and self-training-based output adaptation, we perform output space adaptation through anatomy-constrained optimization, realized by embedding anatomical constraints into a loss function. The latter contribution is technically related to constrained optimization for medical image segmentation, introduced by Kervadec et al. [2019] for weakly-supervised learning and adapted to domain adaptation by Bateson et al. [2021]. However, their proposed constraints on the sizes of target structures do not apply to human pose estimation, which requires specifically tailored constraints on the human skeleton graph. Few works used such anatomical losses for 3D human pose estimation. A geometric constraint on the ratio of bone lengths was proposed by Zhang et al. [2020b] and Zhou et al. [2017] to regularize supervised learning with weak 2D pose ground truth. Moreover, Cao et al. [2020] and Sun et al. [2017] introduced bone and symmetry losses as additional penalties in a fully supervised setting, where accurate ground truth poses, including precise bone lengths, are available. These scenarios are substantially different from our unsupervised setting, where the anatomical loss functions are the only source of supervision on unlabeled target data and are derived from weaker constraints.

4.1.2 Methods

4.1.2.1 Problem Setup and Notation

Point cloud-based 3D human pose estimation aims at predicting the 3D positions of K human joints of interest, $\boldsymbol{Y} \in \mathbb{R}^{K\times 3}$, from a 3D input point cloud $\boldsymbol{X} \in \mathbb{R}^{N\times 3}$. We address the task in a domain adaptation setting, where training data consists of a labeled source dataset $\mathcal{S} = \{(\boldsymbol{X}_s, \boldsymbol{Y}_s)\}_{s=1}^{|\mathcal{S}|}$ and a shifted unlabeled target dataset $\mathcal{T} = \{\boldsymbol{X}_t\}_{t=1}^{|\mathcal{T}|}$. The goal is to learn a function f with parameters $\boldsymbol{\theta}_f$ that predicts human poses as $\hat{\boldsymbol{Y}} = f(\boldsymbol{X}; \boldsymbol{\theta}_f)$ and achieves optimal performance on target data at test time. We aim to solve the problem both in the UDA and SFDA setting. UDA assumes simultaneous access to source and target data. In SFDA, by contrast, source and target

data are only accessible in successive stages. The model is initially trained on source data and subsequently adapted to unlabeled target data without access to source data.

Notation. For a human pose $\boldsymbol{Y}$, we indicate individual joints as $\boldsymbol{y}_k \in \mathbb{R}^3$ and treat them as the nodes of a skeleton graph. We denote $\mathcal{B} = \{\boldsymbol{b}_i\}_{i=1}^{N_\beta}$ as the set of all bone vectors $\boldsymbol{b}_i \in \mathbb{R}^3$ that connect two joints in the skeleton graph, and $\boldsymbol{b}_{t,i}$ indicates the i-th bone vector of the indexed pose $\boldsymbol{Y}_t$. We further indicate $\mathcal{B}_\lambda \subset \mathcal{B}$ as the subset of N_λ bones $\boldsymbol{b}_i^\lambda$ of the left body side that have a counterpart $\boldsymbol{b}_i^\rho \in \mathcal{B}_\rho$ on the right body side. Finally, we term $\mathcal{B}_\zeta = \{(\boldsymbol{b}_i, \boldsymbol{b}_j)\}$ as the set of all N_ζ pairs of bone vectors that are connected by a joint and define $\mathcal{I}_\zeta = \{(i, j)\}$ as the corresponding set of indices.

4.1.2.2 Overview

An overview of our proposed method to solve the above problem is shown in Fig. 4.2. While supervised learning on labeled source data is performed by minimizing the task loss

$$\mathcal{L}_{\text{task}}(\boldsymbol{\theta}_f; \mathcal{S}) = \frac{1}{|\mathcal{S}|} \sum_s \frac{1}{K} \left\| \boldsymbol{Y}_s - \hat{\boldsymbol{Y}}_s \right\|_1 \tag{4.1}$$

we aim to bridge the domain gap by exploiting domain-invariant prior knowledge about human anatomy. To this end, we introduce two complementary anatomy-based training strategies that guide the learning process in the unlabeled target domain. On the one hand, we directly embed the prior knowledge into an anatomical loss function ($\mathcal{L}_{\text{anat}}$) to penalize anatomically implausible predictions. We derive the loss from an anatomically constrained optimization problem in Sec. 4.1.2.3. On the other hand, we leverage prior anatomical knowledge to filter pseudo labels for self-training with the Mean Teacher, realized by $\mathcal{L}_{\text{con}}$ (see Sec. 4.1.2.4 for details).

4.1.2.3 Anatomy-Constrained Optimization

We start our discussion for UDA. Our goal is to guide the learning on unlabeled target data by constraining predictions to the space of anatomically plausible poses. To this end, we formulate network training as the constrained optimization problem

$$\begin{aligned} \min_{\boldsymbol{\theta}_f} \quad & \mathcal{L}_{\text{task}}(\boldsymbol{\theta}_f; \mathcal{S}) \\ \text{s.t.} \quad & \hat{\boldsymbol{Y}}_t \text{ is a plausible human pose} \quad t = 1, ..., |\mathcal{T}| \end{aligned} \tag{4.2}$$

At this stage, the essential question is how to formalize the plausibility constraint. Given the high complexity of the human pose space, we approximate it by means of explicit prior knowledge about human anatomy. Specifically, we combine three simpler constraints on the human skeleton graph that are strong indicators for the plausibility of a pose:

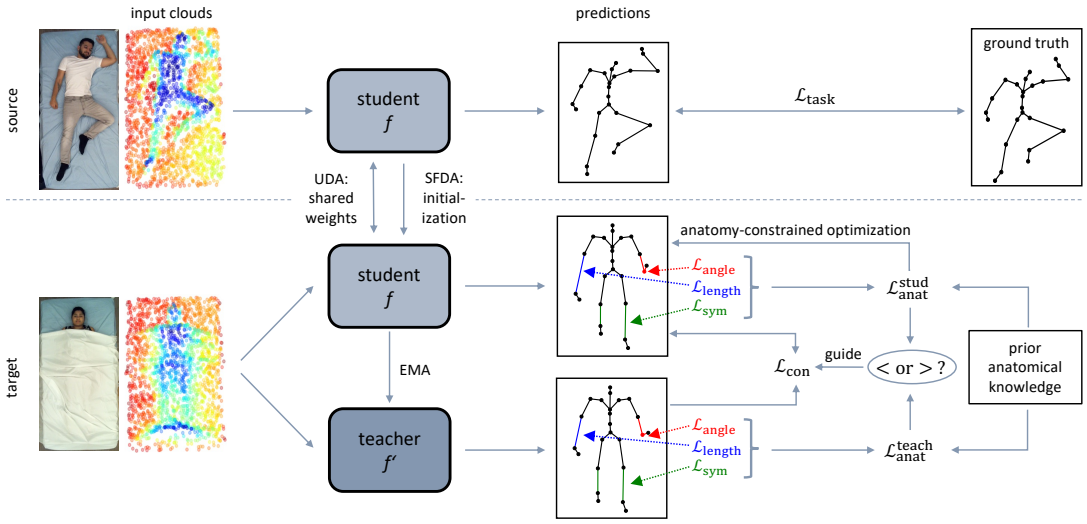


Fig. 4.2: Overview of our method for domain adaptive human pose estimation from point clouds (RGB images are only shown for better visualization). The framework comprises a learning student model and a teacher model, which represents the exponential moving average (EMA) of the student. While source training of the student consists in minimizing a supervised task loss, we perform anatomy-guided learning in the unlabeled target domain. Based on prior knowledge about human anatomy, we formulate an anatomical loss that measures the violation of symmetry, bone lengths, and joint angle constraints. We use the loss to 1) explicitly constrain the student predictions to the space of plausible human poses and 2) filter pseudo labels from the teacher network for self-training according to their anatomical plausibility. As such, the method is applicable to unsupervised domain adaptation (UDA), where the model is jointly trained on the source and target data, and source-free domain adaptation (SFDA), which accesses the domains in two successive steps.

- **Symmetric limbs:** Corresponding limb pairs $(\boldsymbol{b}_i^{\lambda}, \boldsymbol{b}_i^{\rho})$ of the human body typically have roughly equal lengths, with a deviation $\left|\|\boldsymbol{b}_i^{\lambda}\|_2 - \|\boldsymbol{b}_i^{\rho}\|_2\right| < \delta_i$ smaller than a limb-specific tolerance δ_i. We set $\delta_i = 0$ by default but retain the option for an adjustment when dealing with pathologically asymmetric limbs.

- **Plausible bone lengths:** The lengths of human bones $\boldsymbol{b}_i$ are constrained by bone-specific upper and lower bounds u_i^{β} and l_i^{β}, i.e., $l_i^{\beta} \leq \|\boldsymbol{b}_i\|_2 \leq u_i^{\beta}$. Precise values for u_i^{β} and l_i^{β} can be looked up in an anatomical textbook or inferred from the statistics of the training set.

- **Plausible joint angles:** Human joints cannot freely rotate in 3D space, but the range of angles that can be taken is limited. More formally, the normalized dot product of two connected bone vectors $(\boldsymbol{b}_i, \boldsymbol{b}_j) \in \mathcal{B}_{\zeta}$ is constrained by joint-specific upper and lower bounds u_{ij}^{α} and l_{ij}^{α}, i.e., $l_{ij}^{\alpha} \leq \boldsymbol{b}_i/\|\boldsymbol{b}_i\|_2 \cdot \boldsymbol{b}_j/\|\boldsymbol{b}_j\|_2 \leq u_{ij}^{\alpha}$. Again, the

precise determination of upper and lower bounds can be based on an anatomical textbook or the statistics of the training set.

Altogether, this yields the novel optimization problem

$$\begin{aligned} &\min_{\boldsymbol{\theta}_f} \ \mathcal{L}_{\text{task}}(\boldsymbol{\theta}_f; \mathcal{S}) \\ &\text{s.t. } -\delta_i < \|\boldsymbol{b}_{t,i}^{\lambda}\|_2 - \|\boldsymbol{b}_{t,i}^{\rho}\|_2 < \delta_i \qquad i = 1, ..., N_\lambda; t = 1, ..., |\mathcal{T}| \\ &\qquad l_i^{\beta} \leq \|\boldsymbol{b}_{t,i}\|_2 \leq u_i^{\beta} \qquad\qquad i = 1, ..., N_\beta; t = 1, ..., |\mathcal{T}| \\ &\qquad l_{ij}^{\alpha} \leq \frac{\boldsymbol{b}_{t,i}}{\|\boldsymbol{b}_{t,i}\|_2} \cdot \frac{\boldsymbol{b}_{t,j}}{\|\boldsymbol{b}_{t,j}\|_2} \leq u_{ij}^{\alpha} \quad \forall (i,j) \in \mathcal{I}_\zeta; t = 1, ..., |\mathcal{T}| \end{aligned} \tag{4.3}$$

As discussed in prior work [Bateson et al., 2021; Kervadec et al., 2019], a known method to solve such a problem requires the minimization of the Lagrangian dual [Bertsekas, 1997]. However, this technique becomes unstable and computationally intractable when deep neural networks are involved. Alternatively, the problem can be approximated by relaxing the hard constraints to soft constraints in the form of differentiable loss functions that augment the original objective and penalize violations of the constraints. To implement this, we define the base penalty function

$$\ell(x; l, u) = \begin{cases} |x - l| & x < l \\ |x - u| & x > u \\ 0 & l < x < u \end{cases} \tag{4.4}$$

which outputs 0 if the input x lies inside the lower and upper bounds and penalizes inputs outside this range with a linear L1 loss. We also experimented with a quadratic penalty, which performed slightly worse (Sec. 4.1.3.2). Given a human pose $\boldsymbol{Y}$, the violation of our anatomical constraints is then penalized by the loss functions

$$\begin{aligned} \mathcal{L}_{\text{sym}}(\boldsymbol{Y}) &= \frac{1}{N_\lambda} \sum_{i=1}^{N_\lambda} \ell\left(\|\boldsymbol{b}_i^{\lambda}\|_2 - \|\boldsymbol{b}_i^{\rho}\|_2, -\delta_i, \delta_i\right) \\ \mathcal{L}_{\text{length}}(\boldsymbol{Y}) &= \frac{1}{N_\beta} \sum_{i=1}^{N_\beta} \ell\left(\|\boldsymbol{b}_i\|_2, l_i^{\beta}, u_i^{\beta}\right) \\ \mathcal{L}_{\text{angle}}(\boldsymbol{Y}) &= \frac{1}{N_\zeta} \sum_{(i,j) \in \mathcal{I}_\zeta} \ell\left(\frac{\boldsymbol{b}_i}{\|\boldsymbol{b}_i\|_2} \cdot \frac{\boldsymbol{b}_j}{\|\boldsymbol{b}_j\|_2}; l_{ij}^{\alpha}, u_{ij}^{\alpha}\right) \end{aligned} \tag{4.5}$$

This enables us to replace the constrained optimization problem in Eq. (4.3) by the standard minimization of the joint loss function

$$\mathcal{L}(\boldsymbol{\theta}_f; \mathcal{S}, \mathcal{T}) = \mathcal{L}_{\text{task}}(\boldsymbol{\theta}_f; \mathcal{S}) + \lambda_1 \mathcal{L}_{\text{anat}}(\boldsymbol{\theta}_f; \mathcal{T}) \tag{4.6}$$

with the anatomical loss

$$\mathcal{L}_{\text{anat}}(\boldsymbol{\theta}_f, \mathcal{T}) = \frac{1}{|\mathcal{T}|} \sum_t \Big[\mathcal{L}_{\text{sym}}(\hat{\boldsymbol{Y}}_t) + \mathcal{L}_{\text{length}}(\hat{\boldsymbol{Y}}_t) + \mathcal{L}_{\text{angle}}(\hat{\boldsymbol{Y}}_t) \Big] \tag{4.7}$$

and the weighting factor λ_1. Individual weighting factors for each loss were explored but did not yield an improvement.

Optimization and SFDA Since $\mathcal{L}_{\text{task}}(\mathcal{S})$ and $\mathcal{L}_{\text{anat}}(\mathcal{T})$ each only depend on a single domain, the above method is technically also applicable to SFDA by separately minimizing the two losses in successive stages. (Strictly speaking, when deriving upper and lower bounds from the training set, the anatomical loss still accesses labels from the source domain. But unlike visual input data, the upper and lower bounds of the label distribution do not represent sensitive information in terms of data protection, and sharing them among institutions is uncritical.) However, for both UDA and SFDA, when minimizing $\mathcal{L}_{\text{anat}}(\mathcal{T})$ over all model parameters $\boldsymbol{\theta}_f$, we observed a mode collapse in the target domain, where the model predicted a roughly fixed anatomically plausible pose independent of the input. The phenomenon was particularly prominent in SFDA as the absence of joint supervision on source data caused the model to forget that the predicted pose should match the given input. As suggested in our preceding work [Bigalke et al., 2022a], an intuitive solution to this problem is to minimize $\mathcal{L}_{\text{anat}}(\mathcal{T})$ over a restricted subset of network parameters $\boldsymbol{\theta}_g \subset \boldsymbol{\theta}_f$ while minimizing $\mathcal{L}_{\text{task}}$ over all parameters. We experimentally found that only optimizing the feature extractor g of f yields excellent results in UDA, whereas SFDA required a further restriction to the parameters of the BatchNorm layers of g to achieve decent results. While this technique successfully prevents the mode collapse, it also limits the adaptation capacity of the network. As an alternative, we therefore propose to combine anatomy-constrained optimization with supervision through pseudo labels, which can prevent the mode collapse without restricting the adaptability of the network. In our prior work, we already experimentally demonstrated that anatomy-constrained optimization works particularly well in combination with pseudo labels provided by the Mean Teacher [French et al., 2018]. In the following Sec. 4.1.2.4, we formalize the Mean Teacher framework in the context of our problem and extend the standard version by filtering the provided pseudo labels according to their anatomical plausibility.

4.1.2.4 Self-training With the Mean Teacher

The Mean Teacher framework [French et al., 2018; Tarvainen et al., 2017] extends the learning model f, from now on denoted as the student model, by a second so-called teacher model f' with identical architecture. Unlike the student model, the weights

of the teacher $\boldsymbol{\theta}'_f$ are not optimized by gradient descent but given as the exponential moving average (EMA) of the student's weights, updated as

$$\boldsymbol{\theta}'_{f,i} = \mu\boldsymbol{\theta}'_{f,i-1} + (1-\mu)\boldsymbol{\theta}_{f,i} \tag{4.8}$$

at iteration i with momentum μ. Thus, the teacher can be seen as a temporal ensemble of the student and is therefore expected to provide—on average—more stable and accurate predictions than the student. The essential idea of the framework is to leverage this superiority of the teacher by supervising student predictions on unlabeled target data with pseudo labels provided by the teacher. This is implemented by a consistency loss

$$\mathcal{L}_{\text{con}}(\boldsymbol{\theta}_f; \boldsymbol{\theta}'_f, \mathcal{T}) = \frac{1}{|\mathcal{T}|} \sum_t \frac{1}{K} \left\| \hat{\boldsymbol{Y}}_t - \hat{\boldsymbol{Y}}'_t \right\|_1 \tag{4.9}$$

encouraging predictions $\hat{\boldsymbol{Y}}'_t = f'(\boldsymbol{X}_t; \boldsymbol{\theta}'_f)$ by the teacher and $\hat{\boldsymbol{Y}}_t = f(\boldsymbol{X}_t; \boldsymbol{\theta}_f)$ by the student to be consistent. To prevent trivial solutions and vanishing gradients, teacher and student operate on different augmentations of the same input sample that are reversed in the output space to align the predicted poses. In our point cloud-based framework, augmentations consist of random global translation, rotation, and subsampling of input points.

Anatomy-guided filtering of pseudo labels. For the consistency loss to efficiently guide the learning process on target data, predictions by the teacher should be more accurate than those of the student. While this is expected on average, there will be samples where the teacher prediction is inferior to the student prediction. In such cases, the consistency loss in Eq. (4.9) drives the student towards a worse solution and thus hampers the learning process. Instead, we would ideally filter the pseudo labels provided by the teacher and only use those labels for supervision that are more accurate than the current predictions of the student. Since accuracy itself can obviously not be measured in the absence of ground truth, another criterion for filtering pseudo labels is needed.

We propose to filter pseudo labels based on their anatomical plausibility. Specifically, we argue that anatomically plausible poses are more likely to be correct than implausible poses. Consequently, we assess pseudo labels by the teacher and predictions by the student by measuring their plausibility with our three anatomical loss functions in Eq. (4.5). Given the comparisons of the three loss functions, we use only those pseudo labels for supervision, for which at least two out of three anatomical losses indicate a higher plausibility (smaller value) than for the corresponding student predictions. Note that we could alternatively select pseudo labels by comparing the sum of all three losses ($\mathcal{L}_{\text{anat}}$) or just a single loss, but the above criterion gave the best results in the ablation study (Sec. 4.1.3.2).

To formalize the approach, we define the boolean function $\mathbb{1}(\textit{condition})$, which is equal to 1 if the condition is fulfilled and 0 otherwise. Given teacher and student predictions $\hat{\boldsymbol{Y}}'$ and $\hat{\boldsymbol{Y}}$, we then define the function

$$\begin{aligned} h(\hat{\boldsymbol{Y}}', \hat{\boldsymbol{Y}}) = \mathbb{1}\Big(\Big[& \mathbb{1}\left(\mathcal{L}_{\text{sym}}(\hat{\boldsymbol{Y}}') < \mathcal{L}_{\text{sym}}(\hat{\boldsymbol{Y}})\right) \\ & + \mathbb{1}\left(\mathcal{L}_{\text{length}}(\hat{\boldsymbol{Y}}') < \mathcal{L}_{\text{length}}(\hat{\boldsymbol{Y}})\right) \\ & + \mathbb{1}\left(\mathcal{L}_{\text{angle}}(\hat{\boldsymbol{Y}}') < \mathcal{L}_{\text{angle}}(\hat{\boldsymbol{Y}})\right) \Big] \geq 2 \Big) \end{aligned} \tag{4.10}$$

which outputs 1 if our criterion affirms the use of the teacher prediction for supervision and 0 otherwise. We finally reformulate the consistency loss from Eq. (4.9) as

$$\mathcal{L}_{\text{con}}(\boldsymbol{\theta}_f; \boldsymbol{\theta}'_f, \mathcal{T}) = \frac{1}{|\mathcal{T}|} \sum_t \frac{1}{K} h(\hat{\boldsymbol{Y}}', \hat{\boldsymbol{Y}}) \cdot \left\| \hat{\boldsymbol{Y}}_t - \hat{\boldsymbol{Y}}'_t \right\|_1 \tag{4.11}$$

Taking altogether, we integrate this consistency loss into our previous objective function from Eq. (4.6). To perform UDA, we thus minimize

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}_f; \boldsymbol{\theta}'_f, \mathcal{S}, \mathcal{T}) = & \mathcal{L}_{\text{task}}(\boldsymbol{\theta}_f; \mathcal{S}) \\ & + \lambda(\tau)\lambda_1 \mathcal{L}_{\text{anat}}(\boldsymbol{\theta}_f; \mathcal{T}) \\ & + \lambda(\tau)\lambda_2 \mathcal{L}_{\text{con}}(\boldsymbol{\theta}_f; \boldsymbol{\theta}'_f, \mathcal{T}) \end{aligned} \tag{4.12}$$

Here, $\lambda(\tau) = \exp(-5(1 - \min(\tau/T, 1)^2))$ depends on the current epoch τ and continually increases from 0 to 1 during the first T epochs, as suggested by Tarvainen et al. [2017], while λ_2 is a fixed weighting factor. Time dependency is needed to suppress noisy gradients from $\mathcal{L}_{\text{con}}$ and $\mathcal{L}_{\text{anat}}$ at early epochs when the weights of the student and the teacher model are still close to initialization.

For SFDA, we adapt the model pre-trained on source data by minimizing

$$\mathcal{L}(\boldsymbol{\theta}_f; \boldsymbol{\theta}'_f, \mathcal{T}) = \lambda_1 \mathcal{L}_{\text{anat}}(\boldsymbol{\theta}_f; \mathcal{T}) + \lambda_2 \mathcal{L}_{\text{con}}(\boldsymbol{\theta}_f; \boldsymbol{\theta}'_f, \mathcal{T}) \tag{4.13}$$

Time-dependent weighting is not required because pre-training avoids noisy gradients. Note that, for SFDA, the student and the teacher are initialized with the same weights of the pre-trained source model. This is in contrast to the initialization with different random weights in UDA. Furthermore, related to our discussion in Sec. 4.1.2.3, we found it beneficial to minimize the loss in Eq. (4.13) just with respect to the weights of the feature extractor of f while freezing the network heads. This reduces the risk of the model forgetting source knowledge, which is constantly present when dealing with SFDA.

4.1.2.5 Point Cloud-Based 3D Pose Estimation

While our formulation is agnostic to the specific implementation of the function f, we realize point cloud-based 3D pose estimation as follows. Given an input point cloud $\boldsymbol{X} \in \mathbb{R}^{N\times 3}$, we estimate the associated 3D pose $\hat{\boldsymbol{Y}} \in \mathbb{R}^{K\times 3}$ as the weighted sum over the N input points $\boldsymbol{x}_i \in \mathbb{R}^3$. To this end, we design f to output a stack of K softmax-normalized weight maps $\boldsymbol{W} = f(\boldsymbol{X}; \boldsymbol{\theta_f}) \in \mathbb{R}^{N\times K}$ over the input points. The k-th predicted joint is then given by $\hat{\boldsymbol{y}}_k = \sum_{i=1}^{N} \boldsymbol{x}_i \cdot w_{ik}$. In our work, we implement f as the segmentation architecture of DGCNN [Wang et al., 2019] with 40 neighbors in the neighborhood graph. The network comprises a feature extractor with six convolutional layers and network heads with a shared MLP of three fully-connected layers, yielding 986k model parameters.

4.1.3 Experiments and Results

4.1.3.1 Experimental Setup

Datasets. We evaluate our method for the use case of in-bed patient monitoring, using two in-bed human pose datasets: the public SLP dataset [Liu et al., 2019a, 2022a] and an in-house dataset denoted as MVIBP (multi-view in-bed pose) dataset.

SLP. The SLP dataset comprises single-view depth frames of 109 subjects, captured with a Kinect v2 mounted centrally above the bed. Each subject takes 45 arbitrary resting poses, evenly distributed across supine and lateral (left, right) positions. For each pose, the subjects do not move until three frames with varying cover conditions (no cover, thin cover ∼1 mm, thick cover ∼3 mm) are captured. That way, pose annotations for frames without a cover are also valid for frames with cover. While the original dataset includes 2D joints, Clever et al. [2022] provided the 24 joints of the SMPL model [Loper et al., 2015] as 3D ground truth for the first 102 subjects. We restrict our experiments to these subjects. The first 70 subjects are used for training, subjects 71-80 for validation, and subjects 81-102 for testing. As pre-processing, we transformed depth frames to point clouds using the internal camera parameters and removed all points outside a predefined box around the bed.

MVIBP. The MVIBP dataset comprises multi-view depth frames of 13 subjects captured by three synchronized Azure Kinect cameras on the left and right sides and at the foot of the bed. We recorded video data of the subjects, which were asked to freely move while staying in either supine, left, or right position[1]. Subjects remained permanently uncovered, but—contrary to the SLP dataset—we occasionally bedded them on a small or large pillow. To further simulate a clinically realistic scenario, we used positioning aids, and subjects sometimes wore a respiratory mask (*not* used for active ventilation). Given the video data, we extracted discrete frames at fixed time

[1]The conduct of our study was approved by the ethical review board of Lübeck University. Only healthy adults were included, and all subjects gave their informed consent.

intervals. After removing visually similar frames, we processed the remaining ones in four steps. First, we transformed the depth frames from all three cameras to a point cloud using the internal camera parameters. Second, using the external calibration among the cameras, we rotated each cloud to world coordinates and merged the three clouds. Third, we removed all points outside a predefined box around the bed. Fourth, we downsampled the cloud with a voxel filter with an edge length of 2 cm. For each resulting cloud, we manually annotated the ground truth positions of ten joints (feet, knees, shoulders, elbows, and hands) according to the location of the corresponding SMPL joints. To eliminate duplicate poses from the dataset, we only kept those frames where at least one joint moved by more than a threshold of 10 cm compared to the previous extracted frame. This results in a total of 2408 frames, 1165 showing a supine and 1243 a lateral position. Regarding the data split, we use three subjects (361 frames, 177 with supine and 184 with lateral position) for testing and the remaining subjects for training. A validation set is not required because hyper-parameters are not tuned on this dataset.

Adaptation scenarios. Given the two datasets, we consider two adaptation scenarios, featuring domain shifts with different characteristics.

Uncover→cover. Using only the SLP dataset, we consider uncovered subjects as the labeled source and covered subjects as the unlabeled target domain. Thus, the domain shift consists in the occlusion of the subjects by a cover. The scenario is relevant in practical applications because the annotation of uncovered subjects is viable, while it is virtually infeasible for covered patients in practice. (The same adaptation problem for thermal image data was addressed in the IEEE VIP Cup 2021 [Liu et al., 2022b].) For our experiments, we randomly divide the training data by subject into three splits with 30, 20, and 20 subjects. For each split, we use only one cover condition—uncover, thin cover, and thick cover, respectively—while the remaining data is discarded. This yields 30 subjects as the source and 40 subjects as the target domain. For validation and test set, we use both the thin and the thick cover for all frames of all subjects.

SLP→MVIBP. We focus on uncovered subjects and consider SLP as the labeled source and MVIBP as the unlabeled target dataset. The domain shift results from a broad range of factors: 1) different sensors (Kinect v2 vs. Azure Kinect), 2) different camera perspectives and camera-to-bed distances (yielding differing distributions of points in 3D space), 3) different geometry of the used beds (the bed in MVIBP has a headboard), 4) pillows, positioning aids, and respiration masks are only used in MVIBP, 5) cropped point clouds from MVIBP may contain persons walking around the bed. This scenario is relevant in clinical practice as it simulates the deployment of a model in a different environment, e.g., in another hospital. In our experiments, we use the training set from the SLP dataset (70 uncovered subjects) as the labeled source dataset and the training set from MVIBP (10 subjects) as the unlabeled target dataset. Results

are reported on the test set of MVIBP (3 subjects). Since the annotated pose skeletons in the two datasets are not identical (see Fig. 4.1, right), we restrict the evaluation to the matching joint pairs, namely feet, knees, shoulders, elbows, and hands.

Implementation details. We implement our method in PyTorch and use the Adam optimizer for training. We train for 100 epochs for UDA and for 80 epochs for SFDA with a constant learning rate of 0.001. Batches are composed of 8 source and 8 target samples for UDA and of 8 target samples only for SFDA. The weighting factors in Eq. (4.12) are set to $\lambda_1 = 0.1$ and $\lambda_2 = 1$, and the ramp-up length T is set to 40 epochs. The momentum μ for updating the teacher's weights is set to 0.99 for UDA and 0.9996 for SFDA. Upper and lower bounds u_{ij}^{α}, u_i^{β} / l_{ij}^{α}, l_i^{β} of our anatomical constraints are set to the max/min values from the training set of the source domain. For regularization, we use a weight decay of 1e-5 and augment the input point clouds by random rotation around the z-axis, translation, and subsampling to 2048 points. For further details, we refer to our public code at `https://github.com/multimodallearning/da-3dhpe-anatomy`. The above hyper-parameters of our method and the hyper-parameters of all comparison methods (see next paragraph) were tuned on the validation set of the target domain under the uncover→cover scenario and kept fixed for adaptation from SLP to MVIBP. Final results are reported on the test sets of the target domain in terms of the mean per joint position error (MPJPE).

Comparison methods. In this section, we describe the comparison methods used in the experiments. We start by describing the lower and upper bounds.

1) Mean pose. For each sample from the test set, we estimate the pose as the mean pose over all training samples. To construct this mean pose, we anchor the root joint of all training poses at the origin and compute the mean over these centered poses. For evaluation, we apply the same anchoring to the test pose and then compare it to the mean pose. We use this trivial baseline to assess the variability of the used datasets. Note, however, that this baseline accesses ground truth information (location of root joint) at inference time.

2) Source-only. The source-only model is exclusively trained on labeled source data without adaptation techniques and represents a lower bound.

3) Target-only. The target-only model (oracle) is trained on labeled data from the target domain and thus constitutes an upper bound.

To our knowledge, there is no prior work for domain adaptive 3D human pose estimation from point clouds. Therefore, we adapt a comprehensive set of state-of-the-art DA methods to the problem. We primarily describe UDA methods.

4) MMD. Similar to the methods by Rozantsev et al. [2018] and Tzeng et al. [2014], the distributions of source and target features are aligned by minimizing the Maximum Mean Discrepancy (MMD) loss [Gretton et al., 2006], computed for the global feature

vector after conv6 in the DGCNN. We explored a linear and an exponential kernel, with the former yielding slightly better results.

5) DANN. Ganin et al. [2015] proposed to learn domain-invariant features by adversarial learning: a domain discriminator learns to distinguish source and target features while the feature extractor is trained to fool the discriminator. Adversarial optimization is realized by a gradient reversal layer after the feature extractor. We implement the discriminator as a fully-connected network with three layers and apply it to the global feature vector after conv6 in the DGCNN.

6) DefRec. The method by Achituve et al. [2021] performs point cloud-based DA through self-supervised learning. The pretext task is to reconstruct the original input point cloud from a deformed version, where a subset of points is replaced by new points sampled from an isotropic Gaussian distribution with small standard deviation.

7) SSDispPred. Inspired by the method of Doersch et al. [2015], we design a novel pretext task for self-supervised DA, which consists in predicting the displacement vector between two randomly sampled patches from an input cloud.

8) AdvOutAdapt. We adopt the adversarial output adaptation method by Yang et al. [2018]. A discriminator learns to distinguish predicted poses on target data from ground truth poses in the source domain. Meanwhile, the pose estimation network is trained to fool the discriminator by predicting poses that match the distribution of ground truth poses. As for the implementation of the discriminator, we explored diverse architectures of fully-connected and graph neural networks, with the former yielding better results. This method is related to our anatomy-constrained optimization since the discriminator could theoretically learn to penalize implausible predictions similar to our anatomical losses.

9) CC-SSL. Mu et al. [2020] proposed a consistency-constrained curriculum learning strategy for efficient self-training with pseudo labels. First, the confidence for initial pseudo labels from the source-only model is assessed by measuring the consistency under input perturbations. The most confident pseudo labels are then selected for supervised training. After some epochs, the pseudo labels are updated, their confidence is reassessed, and a larger proportion of pseudo labels is selected for the next stage of supervised training. This procedure is repeated several times.

10) MCD. Inspired by the concept of Maximum Classifier Discrepancy [Saito et al., 2018], we extend the pose estimation model by a second network head with a different weight initialization. DA is realized by performing two sequential optimization steps at each iteration. First, the feature extractor and the network heads are jointly optimized on labeled source data. Second, the feature extractor only is optimized on unlabeled target data by minimizing the discrepancy between the predictions of the network heads for the same input sample.

11) Mean Teacher. An extension of the Mean Teacher [French et al., 2018; Srivastav et al., 2022] is already part of our method (Sec. 4.1.2.4). The original Mean Teacher thus

corresponds to an ablated version of our method, excluding anatomy-guided filtering of pseudo labels and anatomy-constrained optimization.

We further describe three state-of-the-art comparison methods for SFDA.

12) UBNA. Klingner et al. [2022] perform SFDA by partially adapting the statistics of the BatchNorm layers to the target domain. The authors use an exponentially decaying momentum factor for the adaptation such that the updated statistics represent a mix of the statistics from the source and target domain.

12) BNAdapt. Zhang et al. [2022] also tackled SFDA by adapting the statistics of the BatchNorm layers. Specifically, they proposed to use the statistics of the test batch itself at inference time instead of the running mean and variance captured during training. Note, however, that this requires sufficiently large batches at test time, which are not always available. We found a batch size of 64 to be a good trade-off between memory consumption and performance. To minimize random effects due to the composition of the test batches, we repeat each experiment five times and report average scores.

13) Mean Teacher. Wang et al. [2022] extended the Mean Teacher to continual test time adaptation, which is closely related to SFDA. The authors proposed to improve the quality of the pseudo labels from the Mean Teacher by averaging over multiple predictions under different input augmentations. Moreover, they addressed catastrophic forgetting by stochastically resetting a small ratio of weights to the original pre-trained weights after each iteration. In our experiments, however, neither augmentation-averaged pseudo labels nor stochastic weight restoration brought any benefits. Therefore, we use the standard Mean Teacher with frozen network heads, which is identical to the ablated version of our method.

4.1.3.2 Ablation Study

We start by analyzing the two essential components of our method, namely anatomy-constrained optimization and anatomy-guided filtering of pseudo-labels. The ablation experiments are performed under the uncover→cover setting on the SLP dataset.

Anatomy-constrained optimization. In the first ablation experiment, we examine the effectiveness of the proposed anatomical loss functions from Eq. (4.5). We consider the UDA setting, discard the Mean Teacher, and minimize $\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda_1 \mathcal{L}_{\text{x}}$ with $\mathcal{L}_{\text{x}} \in \{\mathcal{L}_{\text{sym}}, \mathcal{L}_{\text{angle}}, \mathcal{L}_{\text{length}}, \mathcal{L}_{\text{anat}}\}$. For the three individual losses ($\mathcal{L}_{\text{sym}}, \mathcal{L}_{\text{angle}}, \mathcal{L}_{\text{length}}$), we examine L1 and L2 penalties.

Results of the experiment are shown in Tab. 4.1. Our insights are three-fold. First, each of the three individual loss functions alone substantially reduces the error of the source-only baseline—irrespective of the used penalty function. Second, for all three constraints, the L1 penalty is superior to the L2 penalty, whereby the gap is particularly notable for the angle constraint. Third, aggregating the individual losses in $\mathcal{L}_{\text{anat}}$ further improves performance. This indicates that the three proposed constraints

Table 4.1: Mean per joint position error (MPJPE) for domain adaption with different anatomical loss functions compared to the source-only and target-only models. For each of the three anatomical constraints, we compare a linear L1 against a quadratic L2 penalty. The evaluation is performed for UDA under the uncover→cover adaptation scenario on the SLP dataset.

Method	**L1**	**L2**	**MPJPE [mm]**
Source-only			130.4
Target-only			67.7
$\mathcal{L}_{\text{angle}}$	✓		**106.7**
$\mathcal{L}_{\text{angle}}$		✓	119.3
$\mathcal{L}_{\text{sym}}$	✓		**105.9**
$\mathcal{L}_{\text{sym}}$		✓	108.6
$\mathcal{L}_{\text{length}}$	✓		**102.9**
$\mathcal{L}_{\text{length}}$		✓	104.1
$\mathcal{L}_{\text{anat}}$	✓		**96.6**

effectively complement each other, thus better approximating the space of plausible poses than any of the constraints alone. Overall, our anatomy-constrained optimization reduces the error of the source-only model by 26% and the gap between the source-only and the target-only model by 54%.

Anatomy-guided filtering of pseudo labels. Next, we examine the effect of anatomy-guided filtering of pseudo labels. We start by verifying our hypothesis that anatomically plausible pose estimates are more likely to be correct than implausible ones. To this end, we use the source-only model for inference on the validation set of the target domain. For each predicted pose, we compute the pose error (MPJPE) and the anatomical losses $\mathcal{L}_{\text{sym}}, \mathcal{L}_{\text{angle}}, \mathcal{L}_{\text{length}}$, and $\mathcal{L}_{\text{anat}}$. We then compute the Pearson coefficient R and the corresponding p-value between pose errors and each of the losses (see Tab. 4.2, columns 2,3). For all loss functions, p-values smaller than 0.001 prove a significant correlation, confirming our hypothesis. Comparing the Pearson coefficients among the individual loss functions, we obtain—from weakest to strongest correlation—$\mathcal{L}_{\text{angle}}, \mathcal{L}_{\text{sym}}$, and $\mathcal{L}_{\text{length}}$. Interestingly, this order is identical to model performance when using the loss functions for direct supervision in anatomy-constrained optimization (see Tab. 4.1). The Pearson coefficient for $\mathcal{L}_{\text{anat}}$ ranges between those for $\mathcal{L}_{\text{sym}}$ and $\mathcal{L}_{\text{length}}$.

Given the confirmation of our initial hypothesis, we now explore the suitability of the loss functions for filtering pseudo labels. To this end, we consider both UDA and

Table 4.2: Mean per joint position error (MPJPE) for different techniques to filter pseudo labels under the Mean Teacher paradigm. The evaluation was performed for UDA and SFDA under the uncover→cover adaptation scenario on the SLP dataset. Pearson coefficient R and corresponding significance value p indicate the correlation between the anatomical loss functions and the MPJPE, measured on predictions of the source-only model on the validation set of the target domain.

Method	R	p	MPJPE [mm] (UDA)	MPJPE [mm] (SFDA)
No filtering	-	-	102.3	100.8
Consistency	-	-	102.1	100.5
$\mathcal{L}_{\text{angle}}$	0.20	$< 10^{-3}$	100.1	102.8
$\mathcal{L}_{\text{sym}}$	0.36	$< 10^{-3}$	99.6	98.4
$\mathcal{L}_{\text{length}}$	0.56	$< 10^{-3}$	92.9	98.7
$\mathcal{L}_{\text{anat}}$	0.45	$< 10^{-3}$	93.1	97.9
2 out of 3	-	-	**92.3**	**97.0**

SFDA settings, discard anatomy-constrained optimization ($\lambda_1 = 0$), and use different variants of Eq. (4.10) to guide the consistency training. Besides our proposed method (denoted as '2 out of 3'), we filter pseudo labels by directly comparing each of the losses $\mathcal{L}_{\text{sym}}, \mathcal{L}_{\text{angle}}, \mathcal{L}_{\text{length}}$, and $\mathcal{L}_{\text{anat}}$. As the baseline, we perform no filtering ($h(\hat{\boldsymbol{Y}}', \hat{\boldsymbol{Y}}) = 1$), which is equivalent to the standard Mean Teacher. As another comparison method, similar to Ke et al. [2019], Mu et al. [2020], and Zhou et al. [2022b], we filter pseudo labels based on their consistency under input augmentations. Specifically, we forward two augmented versions of the input through both the student and the teacher model and compute a consistency loss between the two student predictions and the two teacher predictions. On this basis, the teacher predictions are only used for supervision if they are more consistent than the student predictions.

Results of the experiment are shown in Tab 4.2, columns 4 and 5. Our insights are four-fold. First, consistency-based filtering yields only a minor improvement compared to the baseline without filtering. Second, as intuitively expected, we observe a rough trend that a higher correlation between the anatomical loss functions and the pose error comes along with improved performance when using the losses for filtering pseudo labels. Specifically, filtering based on $\mathcal{L}_{\text{angle}}$ yields a minor improvement for UDA and even a slight degradation for SFDA. Moderate improvements under both scenarios are realized by $\mathcal{L}_{\text{sym}}$, while $\mathcal{L}_{\text{length}}$ and $\mathcal{L}_{\text{anat}}$ achieve the top performance among the loss functions. Third, our proposed '2 out of 3' method further improves on $\mathcal{L}_{\text{length}}$ and $\mathcal{L}_{\text{anat}}$. This indicates that our proposed ensembling strategy of the three individual losses is superior to simple aggregation in $\mathcal{L}_{\text{anat}}$, where different scales of the losses are neglected. Fourth, our method surpasses the baseline method (no filtering) by 10% for

UDA and by 4% for SFDA. Thus, our anatomy-based filtering strategy considerably improves the efficiency of self-training with pseudo labels under the Mean Teacher paradigm.

4.1.3.3 Comparison to the State of the Art

We compare our method to the comparison methods presented in Sec. 4.1.3.1 under the two adaptation scenarios uncover→cover (U→C) and SLP→MVIBP. Quantitative results are shown in Tab. 4.3, Tab. 4.4, and Fig. 4.3, revealing mostly consistent findings.

First, we note that the mean pose baseline yields an insufficient accuracy under both scenarios, with a similar mean error when averaged over the same set of joints. This indicates a comparable difficulty and variability of poses across the SLP and MVIBP datasets. Note that the low error for hip and core joints for U→C is due to their spatial proximity to the root joint whose ground truth position was used at inference time.

Second, the source-only baseline is far superior to the mean pose estimate but still substantially worse than the target-only oracle. Specifically, the MPJPE of the target-only model is increased by 93% for U→C (100% when averaged over the joints shared with MVIBP) and by even 413% for SLP→MVIBP. This confirms that both considered domain shifts pose severe problems for deep learning-based pose estimation models. Interestingly, the domain shift due to the occlusion by a cover, which intuitively appears more severe to humans than the shift between the two datasets, has a substantially less negative impact on model performance. We also observe that the MPJPE for shoulders, elbows, and hands of the source-only model is higher for SLP→MVIBP than for U→C. The reason presumably is that the domain shift for SLP→MVIBP is partially caused by the presence of a headboard and pillows, which mainly complicate the localization of joints in the upper body (see Fig. 4.4, rows 5-7). Meanwhile, the MPJPE for feet and knees of the source-only model is lower for SLP→MVIBP, and the oracle achieves lower errors for all joints for SLP→MVIBP. These two observations, in turn, are likely due to the absence of a blanket in this scenario, simplifying the pose estimation problem, especially for joints of the lower body.

Third, we assess the performance of the state-of-the-art comparison methods and our method for UDA. All comparison methods improve the source-only model under both scenarios, except for SSDispPred, which fails for SLP→MVIBP. The ranking of the methods is also similar under both domain shifts (only CC-SSL is less effective for SLP→MVIBP), with MCD achieving the lowest error. Most importantly, the results show that both of our proposed methods alone, i.e., anatomy-constrained optimization ($\mathcal{L}_{\text{anat}}$ only) and anatomy-guided filtering of pseudo labels ($\mathcal{L}_{\text{con}}$ only), already outperform all comparison methods under both settings, with $\mathcal{L}_{\text{con}}$ only being slightly superior to $\mathcal{L}_{\text{anat}}$ only. Notably, our anatomy-constrained optimization surpasses adversarial output adaptation, highlighting the effectiveness of explicit constraints contrary to adversarial optimization. The results further show that our two methods

Table 4.3: Results for adaptation in the uncover→cover setting on the SLP dataset for both UDA and SFDA methods. We compare the MPJPE [mm] of our method to diverse competing methods. Results are averaged over the thin and thick cover as the scores are almost identical. Mean* indicates the average over the joints shared with the MVIBP dataset, namely feet, knees, shoulders, elbows, and hands.

Method	UDA	SFDA	Feet	Knees	Hips	Core	Head	Shoul	Elb	Hands	Mean*	Mean
Mean pose			239.4	240.0	56.1	31.8	102.7	134.6	292.6	383.0	257.9	189.1
Source-only			174.1	148.1	74.5	56.5	34.8	65.7	168.2	273.2	165.9	130.4
Target-only			86.4	64.8	36.7	31.6	29.4	42.3	80.6	140.0	82.8	67.7
MMD	✓		164.6	124.6	68.5	56.9	35.3	62.8	177.1	243.0	154.4	121.7
DANN	✓		168.8	114.5	60.9	50.3	33.3	55.0	144.8	218.8	140.4	111.6
DefRec	✓		161.0	130.6	68.1	51.4	34.5	63.6	175.3	255.0	157.1	122.6
SSDispPred	✓		168.4	122.7	65.7	51.0	33.9	59.9	165.1	258.4	154.9	121.9
AdvOutAdapt	✓		181.4	128.6	62.9	**47.1**	35.5	59.3	136.8	207.9	142.8	112.9
CC-SSL	✓		144.9	134.1	71.7	54.9	33.6	59.7	145.4	222.3	141.3	112.4
MCD	✓		151.8	116.8	63.7	52.6	33.6	53.1	120.4	171.4	122.7	99.4
Mean Teacher	✓		155.9	109.8	73.6	57.4	35.0	56.1	118.6	175.9	123.3	102.3
Ours, $\mathcal{L}_{\text{anat}}$ only	✓		141.5	102.2	**56.0**	47.2	33.3	**50.4**	112.5	188.4	119.0	96.6
Ours, $\mathcal{L}_{\text{con}}$ only	✓		134.7	97.3	60.0	49.1	**33.2**	54.3	110.5	**163.9**	112.1	92.1
Ours	✓		**120.4**	**97.0**	57.4	**47.1**	33.8	51.7	**109.1**	169.8	**109.6**	**89.6**
UBNA		✓	172.9	136.1	71.3	57.2	37.3	60.4	149.1	259.8	155.7	124.5
BNAdapt		✓	167.5	125.0	69.4	59.0	35.1	63.5	154.0	229.5	147.9	118.5
Mean Teacher		✓	137.4	110.6	66.2	51.6	**32.9**	56.8	134.7	186.8	125.3	100.8
Ours, $\mathcal{L}_{\text{anat}}$ only		✓	155.6	120.0	64.2	54.6	37.0	55.7	126.0	206.7	132.8	107.7
Ours, $\mathcal{L}_{\text{con}}$ only		✓	133.1	102.8	65.1	50.9	33.2	55.5	127.4	**178.1**	119.4	97.0
Ours		✓	**132.8**	**102.5**	**62.3**	**49.8**	33.6	**53.1**	**118.3**	179.6	**117.3**	**95.6**

Table 4.4: Results for SLP→MVIBP adaptation for both UDA and SFDA. We compare the MPJPE [mm] of our method to diverse competing methods.

Method	**UDA**	**SFDA**	**Feet**	**Knees**	**Shoul**	**Elb**	**Hands**	**Mean**
Mean pose			272.4	228.8	128.3	229.4	449.8	261.7
Source-only			117.2	104.5	114.0	347.0	517.3	240.0
Target-only			55.8	37.3	32.0	36.0	73.1	46.8
MMD	✓		132.4	120.0	130.1	265.0	273.6	184.2
DANN	✓		111.2	103.4	118.6	206.6	206.4	149.2
DefRec	✓		112.6	82.8	112.1	370.9	296.3	194.9
SSDispPred	✓		148.1	99.8	132.4	355.2	600.7	267.2
AdvOutAdapt	✓		157.3	141.6	101.4	189.7	297.3	177.4
CC-SSL	✓		85.7	73.4	90.1	321.1	515.7	217.2
MCD	✓		89.6	77.1	**61.5**	92.2	161.9	96.4
Mean Teacher	✓		93.4	77.3	110.6	291.6	197.0	154.0
Ours, $\mathcal{L}_{\text{anat}}$ only	✓		86.6	85.2	70.7	85.6	126.0	90.8
Ours, $\mathcal{L}_{\text{con}}$ only	✓		63.0	70.5	117.9	**75.6**	**103.0**	86.0
Ours	✓		**62.6**	**70.2**	83.2	84.9	108.4	**81.8**
UBNA		✓	101.0	102.5	130.2	371.3	386.1	218.2
BNAdapt		✓	96.3	108.5	142.2	223.7	231.1	160.4
Mean Teacher		✓	84.4	72.2	109.1	131.8	194.6	118.4
Ours, $\mathcal{L}_{\text{anat}}$ only		✓	81.7	88.8	**75.3**	106.2	**156.8**	101.8
Ours, $\mathcal{L}_{\text{con}}$ only		✓	**66.7**	**62.5**	83.7	98.4	182.9	98.8
Ours		✓	68.8	68.3	79.3	**86.8**	174.4	**95.5**

are complementary as their combination further reduces the MPJPE to 89.6 mm for U→C and 81.8 mm for SLP→MVIBP. This corresponds to a relative improvement of 31% and 66% over the source-only model and a reduction of the gap between the source-only and the target-only model of 65% and 82%, respectively.

Finally, we compare the SFDA methods. Again, each comparison method reduces the domain gap under both settings. Among these methods, the Mean Teacher achieves the highest performance, surprisingly outperforming its counterpart for UDA. As possible reasons, we suspect the frozen weights of the network heads and a better adaptation of the BatchNorm statistics in SFDA. Regarding our proposed methods, we make the expected observation that all three versions perform slightly worse than in the UDA setting. Nevertheless, they are still superior to the comparison methods for SFDA (except $\mathcal{L}_{\text{anat}}$ only, which is inferior to the Mean Teacher for U→C), and—

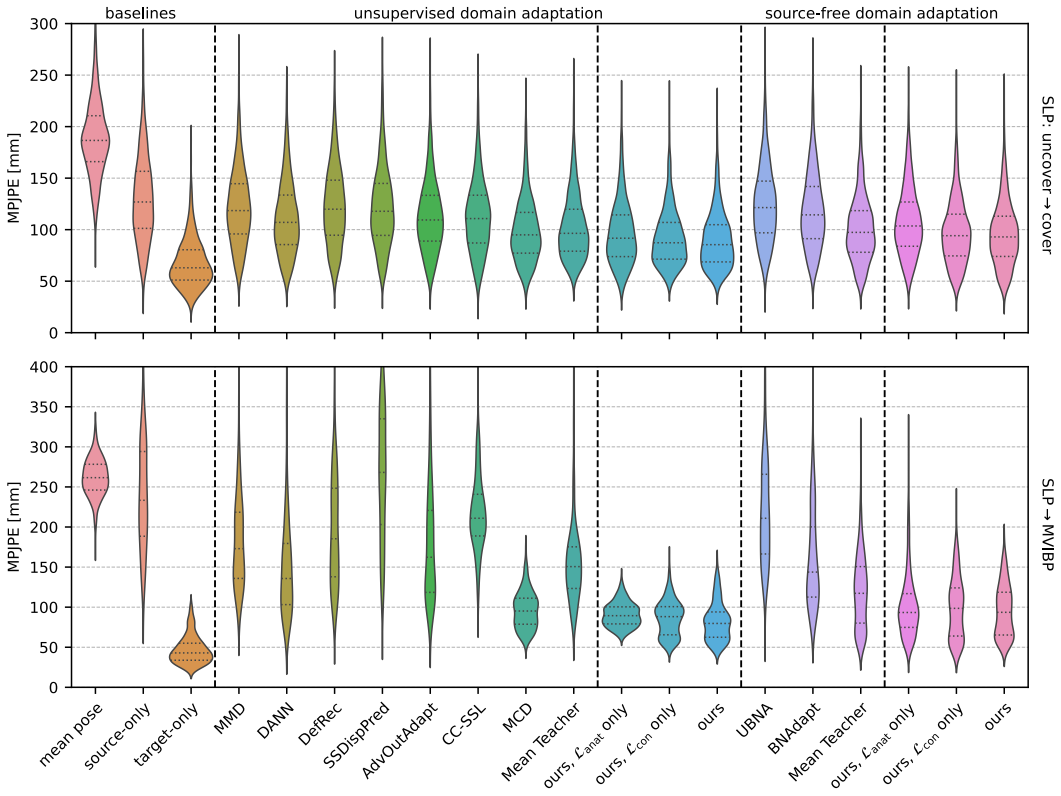


Fig. 4.3: Violin plots of the frame-averaged joint errors for all compared methods in the uncover→cover setting on the SLP dataset (top) and for SLP→MVIBP adaptation (bottom). Dashed lines inside the violins represent the 25th, 50th, and 75th percentiles.

importantly—the combined method is even superior to all competing UDA methods under both domain shifts. This demonstrates the high efficiency of our method under the challenging SFDA setting.

Qualitative results are shown in Fig. 4.4 and are consistent with the quantitative findings. Both the occlusion by a blanket (columns 1-4) and the presence of medical/bed utils (positioning aid, pillow, respiratory mask; columns 5-8) confuse the source-only model, which predicts inaccurate and anatomically implausible poses. By contrast, the predictions by our anatomy-guided adaptation method are more accurate and anatomically more plausible. In particular, our method prevents implausible bone lengths in arms (columns 1,2,3,5,6,7) and legs (columns 1,2) and implausible angles in the shoulder, elbow, and wrist joints (columns 1,3,5,6). Two failure cases of our method are shown in columns 4 and 8, where the predicted poses appear plausible but are inconsistent with the actual pose.

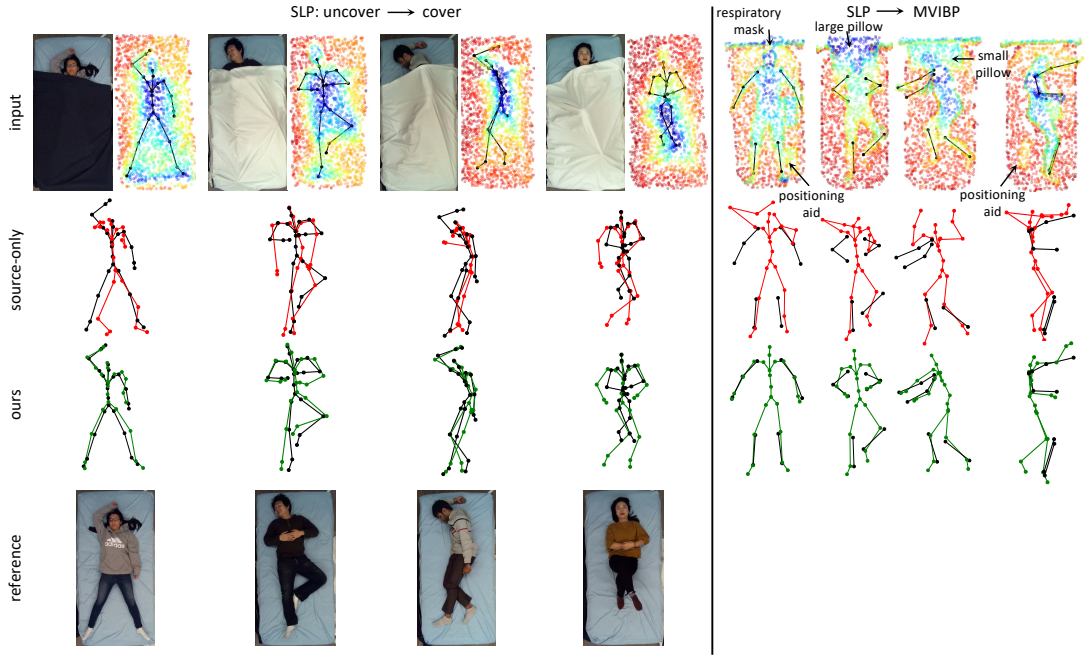


Fig. 4.4: Qualitative results on test samples from the target domain for uncover→cover adaptation (columns 1-4) and for SLP→MVIBP adaptation (columns 5-8). We show input point clouds (upper row), predictions by the source-only model (second row, red), and predictions by our method (third row, green) each together with the ground truth pose in black. For the samples from the SLP dataset, we also show the color images belonging to the point clouds (first row) for better visualization and the corresponding color images without a cover (fourth row) for reference. Regarding the MVIBP dataset, we must not show any color images due to data privacy.

4.1.4 Discussion and Conclusion

We introduced a novel domain adaptation method for point cloud-based 3D human pose estimation. Our main methodological contribution is to bridge the domain gap with the aid of prior anatomical knowledge, accomplished by two complementary anatomy-based adaptation strategies. First, we directly supervise target predictions by imposing explicit anatomical constraints on the output space. Second, we filter pseudo labels for self-training according to their anatomical plausibility. Our experiments for in-bed pose estimation confirm the efficacy of both approaches and allow the following conclusions: 1) Anatomical constraints are a powerful source of weak supervision to guide the learning process in the absence of ground truth. 2) Anatomy-based filtering of pseudo labels substantially improves the efficiency of self-training. Specifically, we evaluated our method under two different domain shifts, adapting from uncovered to covered subjects and between the different environments of two datasets. In both

settings, our method outperformed diverse comparison methods, surpassed the baseline model by 31%/66%, and reduced the domain gap by 65%/82%. In absolute terms, it reduced the mean error of pose estimates to less than 9 cm for covered patients and to almost 8 cm for uncovered patients. At the same time, our method proved efficient for both UDA and SFDA, thus enabling adaptation even in case of restricted data access. In summary, our method can avoid the need for costly manual annotations in novel target domains, which is a significant obstacle to the flexible use of pose estimation models. Thus, it could become an essential factor in advancing the practical deployment of clinical monitoring systems.

Considering this intended application in a realistic clinical setting, a more detailed discussion of the outcomes of our study is needed. First, while the reported results of our method for pose estimation are promising, in practice, we are interested in the performance of higher-level downstream tasks like action or posture recognition. In consequence, the following open questions still need to be analyzed in future clinical validation studies: To what extent do the improvements by our method enhance the performance of different downstream tasks? Is the pose accuracy by our method sufficient, or are there any downstream tasks that require higher accuracy? Second, the evaluation in our work is restricted to healthy subjects in both domains. However, when adapting a model from a lab dataset (source) to clinical data (target), we might face a population shift, with clinical patients showing pathologically induced anatomical abnormalities, such as asymmetric or deformed limbs. Our symmetry and bone length losses in their original form ($\delta_i = 0$, bounds derived from the source data) would then provide incorrect supervision and no longer be a suitable criterion for filtering pseudo labels. This, in turn, might hamper the general adaptation process and degrade pose estimates for pathological patients. A similar problem would occur when adapting from adults in the source to children (at the pediatric ward, for instance) in the target domain. Advantageously, the formulation of our method is flexible enough to prevent such problems by carefully adjusting the upper and lower bounds of symmetry and bone length constraints according to the target population. The bone lengths of children could be looked up in an anatomical textbook, and patient-specific bounds would enable the incorporation of patient-specific anatomical abnormalities.

As a methodological outlook, we see multiple further opportunities for the beneficial use of anatomical priors. First, our anatomical constraints only approximate the space of plausible poses, still permitting implausible poses. This is mainly caused by our realization of the constraint on the joint angles: 1) Joints are considered in isolation, ignoring the pose dependency of joint limits [Akhter et al., 2015], and 2) the used scalar product does not uniquely represent 3D angles. Incorporating a kinematic model could alleviate these problems and help enforce a globally plausible joint-angle configuration. Second, imposing anatomical constraints during training does not preclude implausible pose estimates at inference time. Embedding the constraints in the model architecture itself, instead, would eliminate implausible predictions and could thus increase model

robustness. Third, in the context of patient monitoring, we have access to a continuous stream of input data instead of isolated frames, opening up further options for using anatomical priors. On the one hand, we can exploit confident pose estimates to derive approximate patient-specific bone lengths. These could serve as prior knowledge to guide the pose estimation on subsequent frames of the same patients, for instance, by conditioning the model on their specific anatomy. On the other hand, a model that operates on a sequence of successive frames could be constrained to predict anatomically coherent poses across time.

Finally, beyond the specific task of point cloud-based human pose estimation, our method might also be beneficial for domain adaptation in general medical imaging tasks. On the one hand, anatomy-constrained optimization could be adapted to 3D landmark detection tasks. On the other hand, the filtering of pseudo labels according to explicit prior knowledge about the structure of the output space is—to the best of our knowledge—a novel concept transferable to other tasks. Pseudo labels in medical segmentation, for instance, could be filtered according to prior knowledge about shape descriptors [Bateson et al., 2022; Kervadec et al., 2021].

Chapter 5

Sequential Point Cloud Analysis for Hand Gesture Recognition

This chapter represents the transition from single-frame analysis in the previous chapters to the analysis of temporal point cloud sequences, studied for a global recognition problem, namely dynamic hand gesture classification, in a fully-supervised setting. Given that hand gestures are characterized by two-scale patterns – local hand posture variations and global hand movements – the methodological focus is on learning expressive multi-scale features. Hence, the chapter develops a dual-stream model to decouple the learning of local and global representations and thoroughly analyzes the suitability of different geometric learning architectures with distinct inductive biases to capture the desired complementary features in the two streams. In-depth ablation experiments confirm the selected graph CNN for fine-grained geometric feature extraction in the local stream and a basis point set encoding combined with a fully-connected DenseNet in the global stream as the optimal design choices. The method achieves state-of-the-art performance on two public benchmarks.

5.1 Learning Multi-Scale Features With Complementary Geometric Architectures

This section has been published in [Bigalke et al., 2021c]. According to the Contributor Roles Taxonomy (CRediT), the contributions of the author of this thesis to the publication are: Conceptualization (together with M.P.H.), Methodology, Software, Investigation, Writing – Original Draft, Writing – Review & Editing (together with M.P.H.), Visualization. The source code is available at `https://github.com/multimodallearning/hand-gesture-posture-position`.

5.1.1 Introduction

Automatic hand gesture recognition is an important problem in computer vision with numerous applications in human-computer interaction, such as sign language translation or touchless gesture control. In solving the problem, two fundamental challenges have

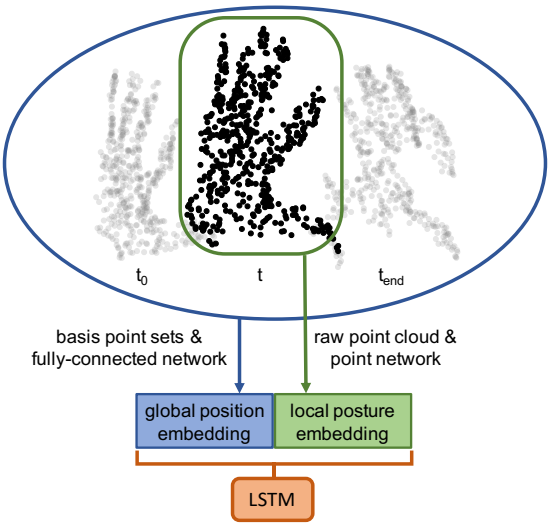


Fig. 5.1: The key idea of our method: We use different point cloud representations (basis point sets & raw point clouds) and network types (fully-connected network & point network) in a two-stream model to capture both global position and local posture features.

been identified in the literature [Liu et al., 2020a; Min et al., 2019]. First, gestures are characterized by spatial variations along the temporal dimension, which necessitates the effective extraction of spatio-temporal features. Second, discriminative patterns for gesture recognition occur at two different scales. Whereas movements of the entire hand occur at a global scale, fine variations of the hand posture, e.g., tiny finger movements, appear at a local scale.

The vast majority of approaches to solving the problem are based on video data [Chai et al., 2016; Lin et al., 2018; Miao et al., 2017; Molchanov et al., 2016; Zhang et al., 2017; Zhu et al., 2017a] or hand skeleton sequences [Chen et al., 2019; Guo et al., 2021; Hou et al., 2018; Li et al., 2019c; Shi et al., 2020; Zhang et al., 2020a]. Only recently, few works considered sequences of 3D point clouds as an alternative input modality and revealed several beneficial properties of point clouds [Kingkan et al., 2018; Min et al., 2019, 2020]. Most importantly, point clouds represent the 3D geometric structure of the entire visible hand surface, which is crucial for accurate gesture recognition. Beyond, point clouds are robust to varying camera perspectives and insensitive to illumination and background clutter. Altogether, point clouds have great potential to advance automatic gesture recognition.

Returning to the identified main challenges of hand gesture recognition, we note that existing point cloud-based methods primarily address the extraction of spatio-temporal features [Kingkan et al., 2018; Min et al., 2019, 2020]. Capturing multi-scale features, on

the contrary, is not explicitly addressed and - if at all - only implicitly achieved through end-to-end optimization. Our starting hypothesis is that this procedure alongside the used single-stream network architectures does not sufficiently account for patterns at both local and global scale. Indeed, our preliminary experiments suggest that the state-of-the-art PointLSTM [Min et al., 2020] is biased towards fine-structural posture features and that global movements of the whole hand are insufficiently considered in the decision process.

5.1.1.1 Contributions

Based on the above observation, in this work we explicitly address the learning of multi-scale features, which capture both fine-structural cues about the hand posture and information about global position and orientation of the entire hand. To this end, we propose a two-stream architecture to extract two separate spatial embeddings from each frame (Fig. 5.1). These are subsequently fused and jointly processed by an LSTM [Hochreiter et al., 1997] for temporal modelling.

To learn the desired features in each stream, we take advantage of the complementary benefits of different point cloud representations. Point cloud networks such as PointNet++ [Qi et al., 2017b] or DGCNN [Wang et al., 2019] are specially tailored to capture fine geometric structures in raw point clouds. Thus, they are ideally suited for hand posture analysis in the local stream. For the extraction of global position features, however, the inductive bias of these networks towards geometric structures is suboptimal. Moreover, raw point clouds contain only implicit information about the relative position of the hand in the entire sphere of action because empty regions in 3D space are not explicitly represented. Therefore, we resort to an alternative representation of point clouds, namely the concept of residual basis point sets (BPS) [Prokudin et al., 2019]. While BPS have previously been used for static object-level tasks [Prokudin et al., 2019], we highlight their suitability for capturing global position features over time in the context of gesture recognition. Specifically, their compact encoding can be seen as an activity map of the entire sphere of action and can be processed by ordinary fully-connected networks, which do not exhibit the above-mentioned inductive bias.

Altogether, the combination of these complementary representations enables us to learn discriminative features that capture both posture and position. At the same time, our method is computationally efficient as it forgoes the expensive extraction of low-level spatio-temporal features required by previous methods [Min et al., 2019, 2020]. In summary, the main contributions of this work are:

- We introduce a novel two-stream framework for point cloud-based hand gesture recognition.

- Our framework is tailored to explicitly learn global position and local posture features by leveraging the complementary benefits of raw point clouds and BPS-based representations.
- We set new state-of-the-art accuracy on the DHG and Shrec'17 dataset as demonstrated by extensive experiments.

5.1.1.2 Related Work

Gesture from video. Early works for video-based hand gesture recognition rely on hand-crafted features [Freeman et al., 1995; Wang et al., 2016]. Since the growing success of deep learning, 3D CNNs are widely used for spatio-temporal feature extraction [Lin et al., 2018; Miao et al., 2017; Molchanov et al., 2016; Zhang et al., 2017; Zhu et al., 2017a]. For improved long-term modelling, multiple works incorporate RNNs [Molchanov et al., 2016], LSTMs [Chai et al., 2016; Rastgoo et al., 2020] or convolutional LSTMs [Wang et al., 2017; Zhang et al., 2017; Zhu et al., 2017a]. To reduce the negative impact of background clutter in video data, two-stage pipelines including hand or skeleton detectors have been proposed [Chai et al., 2016; Lin et al., 2018; Narayana et al., 2018; Rastgoo et al., 2020; Wang et al., 2017]. Attention blocks are used as a more implicit method to focus on most relevant information [Dhingra et al., 2019]. Moreover, multi-stream architectures unify complementary information from multiple modalities such as RGB, depth, IR or optical flow [Abavisani et al., 2019; Miao et al., 2017; Narayana et al., 2018; Zhu et al., 2017a].

Gesture from skeleton. The availability of real-time methods for hand pose estimation [Li et al., 2019a] enables gesture recognition on the basis of skeleton sequences. Beyond early feature-based methods [De Smedt et al., 2016], several deep learning-based approaches process raw skeleton sequences with RNNs [Chen et al., 2017] or GRUs [Maghoumi et al., 2019]. Alternatively, skeletons are represented as graphs to directly capture spatial relations among joints, which are then processed by graph CNNs [Chen et al., 2019; Guo et al., 2021; Li et al., 2019c; Zhang et al., 2020a]. Another widespread methodology are spatio-temporal attention mechanisms [Chen et al., 2019; Hou et al., 2018; Shi et al., 2020; Zhang et al., 2020a]. In a different approach, Liu et al. [2020a] decouple the hand gesture into hand movement and hand posture, which are modelled in two separate streams. In spirit, our approach is similar to this work, but we operate on 3D point clouds instead of skeleton data, requiring a different methodological procedure.

Gesture from 3D point cloud. Unlike video data, point clouds preserve the original 3D geometric structure. At the same time, they are more robust than estimated hand skeletons, which can be error-prone in practice. Due to their unstructured nature, however, 3D point clouds are challenging to process by end-to-end deep networks [Guo

et al., 2020]. For analysis of individual point clouds, the pioneering PointNet [Qi et al., 2017a] extracts spatial embeddings for each point individually, which are aggregated by permutation-invariant max pooling. A series of works incorporate the geometric structure of local neighborhoods by means of hierarchical grouping [Qi et al., 2017b] and generic convolutions that are applicable to irregular point clouds [Li et al., 2018; Liu et al., 2019b; Wang et al., 2019; Wu et al., 2019; Xu et al., 2021]. Based on [Qi et al., 2017b], Ge et al. [2018a] proposed a framework for hand pose estimation from static point clouds. For gesture recognition based on point cloud sequences, Salami et al. [2020] rely on [Qi et al., 2017b] to extract frame-wise spatial features and perform temporal modelling with an LSTM. To learn spatio-temporal features at the point level, Kingkan et al. [2018] fuse all frames in a single point cloud, which is processed by their point attention network. To preserve the temporal order of successive frames, Min et al. [2019] extend the grouping operation from [Qi et al., 2017b] to the temporal domain for spatio-temporal feature extraction. Min et al. [2020] also propose PointLSTM, a novel LSTM cell for unordered point clouds, to capture long-term relationships at the point level. All these works primarily focus on spatio-temporal feature extraction and do not explicitly consider multiple scales as it is done in our work.

5.1.2 Methods

In this section, we present our approach to point cloud-based hand gesture recognition. We initially give an overview of the proposed framework and subsequently present its individual components in detail.

5.1.2.1 Framework

An overview of our proposed method is shown in Fig. 5.2. Input to the method is a sequence of 3D point clouds $\mathbb{P}^T = (\boldsymbol{P}^{(1)}, \boldsymbol{P}^{(2)}, ..., \boldsymbol{P}^{(T)})$ of length T. Each point cloud $\boldsymbol{P}^{(t)} = \{\boldsymbol{p}_i^{(t)} | i = 1, ..., n_t\} \in \mathbb{R}^{n_t \times 3}$ displays the hand at time t and comprises n_t points $\boldsymbol{p}_i^{(t)} = (x_i^{(t)}, y_i^{(t)}, z_i^{(t)})$, which are represented by their 3D coordinate vector. Given C different gesture categories, our goal is to learn a function that maps the input sequence to the correct class $c \in \{1, ..., C\}$.

The key idea of our approach is to explicitly enforce the learning of both local fine-structural posture and global position features by extracting them in separate streams that use different representations and network types. Specifically, given the input sequence $\mathbb{P}^T$, we extract a spatial representation $\boldsymbol{x}^{(t)} = [\boldsymbol{x}_{glob}^{(t)}, \boldsymbol{x}_{loc}^{(t)}]$ from each frame $\boldsymbol{P}^{(t)}$, which is a concatenation of global position features $\boldsymbol{x}_{glob}^{(t)}$ and local posture features $\boldsymbol{x}_{loc}^{(t)}$. To ensure the desired complementary properties, $\boldsymbol{x}_{glob}^{(t)} = f(\boldsymbol{P}^{(t)}; \boldsymbol{\theta}_f)$ and $\boldsymbol{x}_{loc}^{(t)} = g(\boldsymbol{P}^{(t)}; \boldsymbol{\theta}_g)$ are extracted by two separate networks f and g with parameters $\boldsymbol{\theta}_f$ and $\boldsymbol{\theta}_g$, respectively. Subsequently, temporal modelling is performed by an LSTM [Hochreiter et al., 1997] l whose last hidden state $\boldsymbol{h}^{(T)}$ is used for final classification.

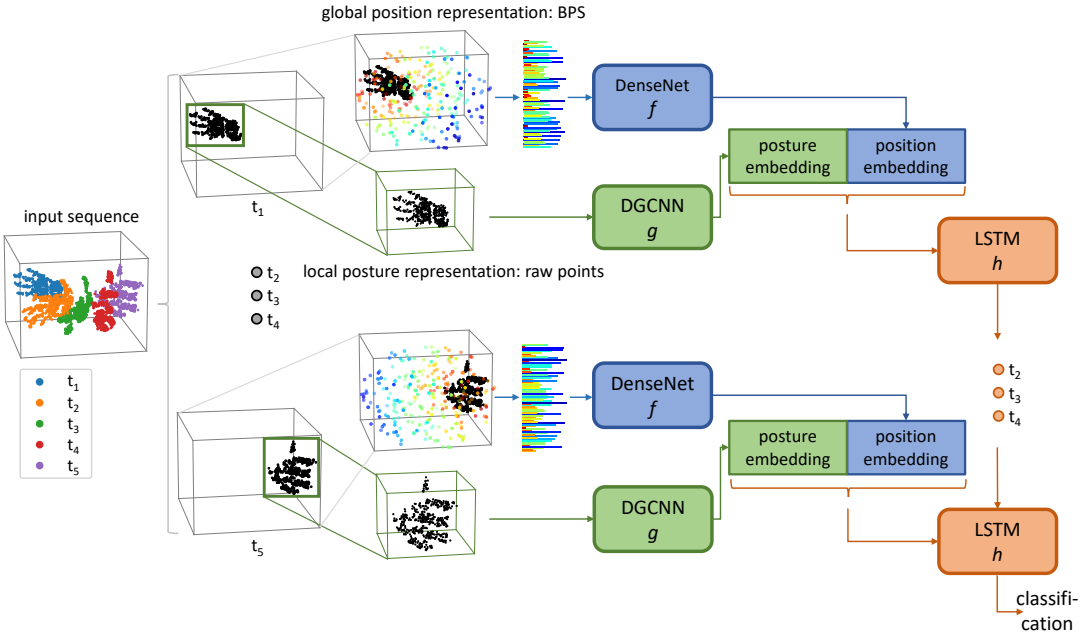


Fig. 5.2: Overview of our proposed framework for point cloud-based gesture recognition. For each frame, we extract a posture embedding from the raw point cloud using a point cloud network (DGCNN) and a position embedding from the BPS encoding with help of a fully-connected DenseNet. The features are concatenated, and the temporal evolution is modelled by an LSTM for final gesture classification.

5.1.2.2 Local Posture Features

The features $\boldsymbol{x}_{loc}^{(t)}$ extracted in the local stream by the network g are supposed to focus on fine variations of the hand posture. To induce the network to learn such features, we rely on two concepts. First, before forwarding frames $\boldsymbol{P}^{(t)}$ through the network g, we remove any information about global hand movement and position by separately mean-centering each frame as $\boldsymbol{P}_{mc}^{(t)} = \boldsymbol{P}^{(t)} - \sum_{i=1}^{n_t} \boldsymbol{p}_i^{(t)} / n_t$. That way, the network is enforced to focus on local hand posture variations. Second, we take advantage of the architectural bias of state-of-the-art point cloud networks such as PointNet++ [Qi et al., 2017b] or DGCNN [Wang et al., 2019]. They are explicitly designed to capture fine geometric structures in local regions and are thus ideal for capturing hand posture variations on the basis of the given raw point clouds. In our framework, we use the DGCNN, which models local neighborhoods by means of dynamic graph convolutions. It provides fast computations and has been proven to extract expressive features for classification and segmentation tasks. Specifically, we implement g as the original classification architecture of DGCNN with 16 nearest neighbors and use the features at fc-512.

5.1.2.3 Global Position Features

Contrary to the local posture features $\boldsymbol{x}_{loc}^{(t)}$, the features $\boldsymbol{x}_{glob}^{(t)}$ in the global stream extracted by the network f are supposed to represent the global position and orientation of a point cloud $\boldsymbol{P}^{(t)}$ within the overall sphere of action. Our approach to implementing this is driven by two major motivations. First, we intend to implement the function f as a general-purpose neural network, which does not exhibit the above-mentioned architectural bias towards fine geometric structures. Second, we seek for an alternative representation of point clouds that explicitly encodes both occupied and empty space of the global scene. Contrary to such a representation, raw point clouds only represent visible surfaces and thus only implicitly represent empty space.

To satisfy both requirements, we propose to extract global position features by means of BPS [Prokudin et al., 2019] in combination with a fully-connected DenseNet [Huang et al., 2017]. BPS are a residual representation of point clouds, which encode an input point cloud $\boldsymbol{P}^{(t)}$ relative to a fixed set of k basis points

$$\boldsymbol{B} = [\boldsymbol{b}_1, ..., \boldsymbol{b}_k], \boldsymbol{b}_j \in \mathbb{R}^3, \|\boldsymbol{b}_j\| \leq r \tag{5.1}$$

which are uniformly sampled in a sphere with radius r in the default setting. Note that this basis is fixed for all input point clouds from both training and test set at any time step t. Given the basis $\boldsymbol{B}$, a point cloud $\boldsymbol{P}^{(t)}$ is encoded by computing the Euclidean distance $d(.,.)$ from each basis point to its nearest point in the input cloud, resulting in a feature vector

$$\boldsymbol{x}_{bps}^{(t)} = [\min_{\boldsymbol{p}_i^{(t)} \in \boldsymbol{P}^{(t)}} d(\boldsymbol{b}_1, \boldsymbol{p}_i^{(t)}), ..., \min_{\boldsymbol{p}_i^{(t)} \in \boldsymbol{P}^{(t)}} d(\boldsymbol{b}_k, \boldsymbol{p}_i^{(t)})] \in \mathbb{R}^k \tag{5.2}$$

To ensure a meaningful global position encoding, it is crucial that the basis points cover the entire sphere of action to be observed. In other words, the radius r of the BPS and the scale of all input sequences need to match. To achieve this, we set $r = 1$ and scale mean-centered input sequences $\mathbb{P}^T$ by a hyper-parameter R, which is set based on the distribution of spheres of action in the training set. Specifically, we ignore outliers and choose R around the maximum radius of the remaining samples. In experiments, our method was robust against the exact choice of R.

A visualization of BPS encodings at different time steps t is shown in Fig. 5.2. Note how the basis points can be seen as a sparse 3D heat map with small values (warm colors) close to the current position of the hand and large values (cold colors) in empty regions. That way, BPS constitute a suitable representation to track global hand movements across the entire sphere of action.

Given the frame-wise BPS encoding $\boldsymbol{x}_{bps}^{(t)}$, we extract more abstract global position features $\boldsymbol{x}_{glob}^{(t)} = f(\boldsymbol{x}_{bps}^{(t)}; \boldsymbol{\theta}_f)$ by means of a fully-connected DenseNet f. Due to the fully-connected architecture, it does not exhibit a bias towards local structures but rather

extracts a global representation. We resort to the same architecture as in [Prokudin et al., 2019], which consists of two blocks with two fully-connected layers of 256 neurons and a single output layer with 1024 neurons. Fully-connected layers are followed by ReLU activation and batch normalization. Furthermore, batch normalization is applied to $\boldsymbol{x}_{bps}^{(t)}$ before the first layer for normalization.

5.1.2.4 Temporal Modelling and Classification

Once frame-wise local posture features $\boldsymbol{x}_{loc}^{(t)}$ and global position features $\boldsymbol{x}_{glob}^{(t)}$ are extracted, we model temporal evolution by means of an LSTM l. Generally, there are multiple options to combine both features. For instance, Liu et al. [2020a] model posture and movement in two independent streams, which are only fused after the final classification layer. In contrast, we argue that a full separation of both streams is not ideal because the interplay between motion and posture variations contains essential cues for gesture recognition and should thus be included in temporal modelling. Consequently, we fuse posture and position features before temporal modelling and feed the concatenated feature vectors $\boldsymbol{x}^{(t)} = [\boldsymbol{x}_{glob}^{(t)}, \boldsymbol{x}_{loc}^{(t)}]$ $(t = 1, ..., T)$ into the LSTM. We opt for a simple LSTM with one hidden layer with 256 neurons. The output of the LSTM consists of its T hidden states $\boldsymbol{h}^{(1)}, ..., \boldsymbol{h}^{(T)}$. For classification, we apply a fully-connected layer followed by Softmax normalization to $\boldsymbol{h}^{(T)}$, yielding a C-dimensional vector with class probabilities. At training time, we apply Dropout (p=0.5) before the classification layer for regularization and perform classification based on the last $T/4$ hidden states to accelerate convergence.

5.1.3 Experiments and Results

5.1.3.1 Experimental Setup

Datasets. We evaluate our proposed method on two public hand gesture datasets, namely the Shrec'17 dataset and the DHG 14/28 dataset.

Shrec'17 [De Smedt et al., 2017]. The Shrec'17 dataset comprises depth map sequences of 14 different hand gesture categories, which can be assigned to two supercategories. Coarse gestures are mainly defined by global movements of the entire hand while fine gestures are characterized by the evolution of hand posture. All gestures are performed by 28 subjects in two different ways, namely with one finger or with the whole hand. This yields two different protocols for evaluation. We refer to the 14 gestures (14G) protocol if one-finger and whole-hand gestures are summarized as one class, and otherwise to the 28 gestures (28G) protocol. Subjects perform each gesture one to ten times, yielding an overall of 2800 sequences, which are split by subject into 1960 training and 840 test sequences. In both training and test set, the class distribution is roughly balanced. Beyond depth maps, the dataset comprises 3D skeleton sequences of 22 hand joints and has extensively been used for skeleton-based gesture recognition.

DHG 14/28 [De Smedt et al., 2016]. The DHG dataset has been recorded in the same way as the Shrec'17 dataset regarding gesture classes and data modalities. The only difference is that the gestures have been performed by another set of 20 subjects. As there is no official data split, evaluation is commonly performed by leave-one-subject-out cross-validation [De Smedt et al., 2016].

Implementation details. We implement our framework in PyTorch [Paszke et al., 2019]. All network parameters are jointly optimized in an end-to-end fashion by minimizing an ordinary cross-entropy loss with the Adam optimizer. The initial learning rate is set to 0.0001, weight decay is 0.005, and batch size is 8. We use Gaussian weight noise with a variance of 0.02 for the weights of the LSTM [Graves et al., 2013]. The network is trained for 100 epochs, whereby the learning rate is divided by a factor of 10 at epoch 50, 80 and 90. We use $k = 256$ basis points in the global branch, and the scale parameter is set to $R = 60\,\text{cm}$ for both datasets.

Pre-processing. To generate point cloud sequences from original depth map sequences, we follow the procedure in [Min et al., 2020]. 32 frames are uniformly sampled along the temporal axis. In each sampled depth frame, the hand silhouette is extracted and transformed into a 3D point cloud from which 512 points are sampled. For further details about these steps, we refer the reader to the supplementary material of [Min et al., 2020]. During training, we perform data augmentation in terms of random rotation by up to $\pm 10°$ and random scaling by up to $\pm 10\,\%$. Moreover, each point cloud is randomly subsampled to 128 points at each training epoch, while a fixed subset of 128 points is used at test time.

Baseline methods. For our baseline models, we extract spatial features $\boldsymbol{x}^{(t)}$ from each frame $\boldsymbol{P}^{(t)}$ in a single stream, whereby global information is preserved (no frame-wise mean-centering as in the local branch). Subsequent temporal modelling is performed with the same LSTM as in our framework. We consider multiple different representations and network architectures for this baseline: PointNet [Qi et al., 2017a] and DGCNN [Wang et al., 2019] are applied to raw point clouds, a 3D occupancy grid with 32^3 voxels is processed by a 3D CNN [Maturana et al., 2015], and the BPS encoding is processed by our DenseNet. Note that the latter model is identical to our framework without the local stream and hence denoted as 'ours w/o local'.

5.1.3.2 Preliminary Experiment

When motivating the composition of our framework, we claimed that point cloud-based networks tend to be biased towards local geometric structures (fine hand posture variations) and do not sufficiently consider global patterns (hand movements). In the preliminary experiment, we intend to verify this claim. For this purpose, we initially

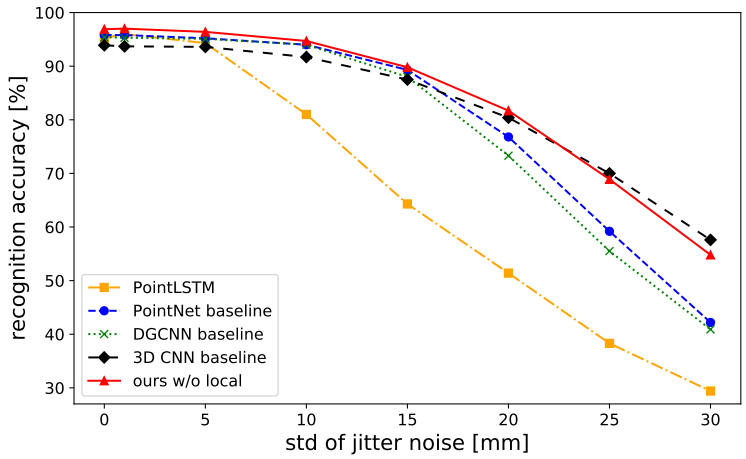


Fig. 5.3: Recognition accuracy on coarse gestures of the Shrec dataset under the 14G protocol as a function of the standard deviation of jitter noise applied at test time.

train our baseline models and the state-of-the-art PointLSTM [Min et al., 2020] on the 28G protocol of the clean Shrec dataset. At test time, we add jitter noise of varying scale and only consider coarse gestures within the 14G protocol. The underlying idea is that jitter noise corrupts fine local structures while information about global movements remains and should be sufficient to distinguish coarse gestures in the 14G protocol. Thus, the observed robustness against jitter noise is an indicator of the extent to which the networks rely on global motion. Results of the experiment are shown in Fig. 5.3. We observe that the performance of PointLSTM severely drops with an increasing scale of jitter noise. This indicates that global movement information is insufficiently incorporated into the decision process and motivates our approach to explicitly learn both global and local features. Point cloud-based PointNet and DGCNN baselines are more robust than PointLSTM, but they are clearly inferior to the 3D CNN baseline and ours w/o local (-10 % points at std=25 mm), which are based on volumetric and residual BPS representations, respectively. Thus, the latter two methods appear to have the strongest focus on global motions and seem particularly suitable for feature extraction in the global stream of our framework.

5.1.3.3 Ablation Study

In this section, we analyze different components of our proposed framework, namely the combination of different inputs, the backbone architectures for local and global stream, the procedure for fusing global and local features, and the number of basis points k. All ablation experiments are conducted on the Shrec'17 dataset for 14G and

Table 5.1: Recognition accuracy (%) for different combinations of global (g) and local (l) streams on the Shrec dataset for 14G and 28G protocol.

Method	Streams	14G	28G
PointNet baseline	g	94.2 ± 0.4	91.1 ± 0.3
DGCNN baseline	g	94.3 ± 0.3	93.3 ± 0.5
3D CNN baseline	g	92.9 ± 1.4	91.1 ± 0.4
Ours w/o local	g	$\mathbf{96.2 \pm 0.4}$	92.7 ± 0.2
Ours w/o global	l	87.4 ± 0.7	86.8 ± 0.7
Ours w/o mean-centering	g+g	95.9 ± 0.5	94.6 ± 0.6
Ours	g+l	96.1 ± 0.4	$\mathbf{95.2 \pm 0.4}$

28G protocol if not stated otherwise. Each experiment is repeated three times, and we report mean and standard deviation of the final accuracies.

Input combinations. In the first ablation experiment, we compare the effectiveness of different combinations of global and local streams in Tab. 5.1. In this context, global refers to inputs with preserved information about global movements whereas this information is removed in local streams. First, we note that the isolated local stream of our framework ('ours w/o global') performs substantially worse than all baseline models under both protocols. This is expected because the local stream does not dispose of any global position information such that the recognition of coarse gestures such as swipes is more difficult. Second, it can be seen that the isolated global stream of our framework ('ours w/o local') is the top performing method for the 14G protocol. This demonstrates the capability of the BPS encoding to extract discriminative position features. For the 28G protocol, however, which contains more fine-grained gestures, the model exhibits a performance drop of 3.5 % points and is inferior to the DGCNN baseline, which excels at capturing fine variations in hand posture. Third, we observe that complementing the BPS-based global stream by the local DGCNN stream ('ours') maintains the performance for the 14G protocol while the performance for the 28G protocol is enhanced by 2.5 % points such that the DGCNN baseline is outperformed by 1.9 % points. This demonstrates the effectiveness of fusing complementary information from two scales and representations. Finally, we investigate the effect of mean-centering the inputs to the local stream in a frame-wise fashion. We exclude this step in another experiment ('ours w/o mean-centering') such that movement information is preserved in the local stream. Our framework surpasses this method by 0.6 % points under the 28G protocol. This indicates that the removal of global position information in the

Table 5.2: Recognition accuracy (%) for different global backbone architectures with **fixed DGCNN in the local stream** on the Shrec dataset for 14G and 28G protocol.

Global backbone	**14G**	**28G**
Translation vector	94.4 ± 0.4	93.4 ± 0.4
3D CNN	93.5 ± 0.5	92.8 ± 0.5
PointNet	95.3 ± 0.5	94.2 ± 0.5
DGCNN	95.6 ± 0.4	94.7 ± 0.4
BPS+DenseNet (ours)	$\mathbf{96.1 \pm 0.4}$	$\mathbf{95.2 \pm 0.4}$

local stream induces the DGCNN to put a stronger focus on fine structures and to learn more expressive posture features.

Global backbone architecture. Besides the combination of BPS encoding and DenseNet, we evaluate the following representations / network architectures for the extraction of global features while keeping the local branch fixed: 1) global information is encoded by a translation vector that represents the difference between the mean coordinates of two consecutive frames, 2) a 3D CNN is applied to a 3D occupancy grid with 32^3 voxels, 3) and 4) PointNet and DGCNN are applied to raw point clouds. Results are shown in Tab. 5.2. For all backbones, the performance is clearly superior to our model without global branch (see Tab. 5.1). Improvements amount to 6.1-8.7 % points and 6.0-8.4 % points for the 14G and 28G protocol, respectively. This indicates that all backbones enrich the local features $\boldsymbol{x}_{loc}$ by complementary position features. Among the different architectures, the proposed BPS-based encoding achieves the best results under both protocols. This underlines the particular suitability of BPS for extracting global position features. Beyond, we note that the 3D CNN performs worst under both protocols although it showed promising properties for global feature extraction in the preliminary experiment. Presumably, the reason is that the 3D grid needs to cover the entire sphere of action, which leads to a very low resolution and extreme sparsity.

Local backbone architecture. We compare three different point cloud networks (PointNet, PointNet++ [Qi et al., 2017b], DGCNN) and BPS+DenseNet for the extraction of local posture features while keeping the global branch fix. Results of the experiment are provided in Tab. 5.3. BPS are clearly inferior to all point-based methods and are thus less suitable for the extraction of fine posture features. Regarding the point cloud networks, all architectures perform similarly under the 14G protocol, whereby the basic PointNet is even slightly superior to PointNet++ and DGCNN. Under the 28G protocol, where fine local structures are of greater importance, the advanced PointNet++ and DGCNN outperform PointNet by 0.4 % points. In our

Table 5.3: Recognition accuracy (%) for different local backbone architectures with **fixed BPS+DenseNet in the global stream** on the Shrec dataset for 14G and 28G protocol.

Local backbone	**14G**	**28G**
BPS+DenseNet	94.8 ± 0.1	93.3 ± 0.3
PointNet	$\mathbf{96.3 \pm 0.2}$	94.8 ± 0.7
PointNet++	96.2 ± 0.5	$\mathbf{95.2 \pm 0.5}$
DGCNN (ours)	96.1 ± 0.4	$\mathbf{95.2 \pm 0.4}$

framework, we opt for DGCNN due to its higher computational efficiency compared to PointNet++.

Fusion type. There are multiple options to fuse global position and local posture features for temporal modelling. Beyond our proposed strategy, which we denote as early fusion, we consider three further strategies: 1) no fusion: global and local features are processed by separate LSTMs and classification heads whose softmax scores are averaged (as in [Liu et al., 2020a]). 2) late fusion: as in 1), but the last hidden states of both LSTMs are concatenated before being fed into a shared classification head. 3) intermediate fusion: global and local features are processed by two separate LSTMs, whose hidden states are concatenated and processed by another LSTM before classification. This approach is similar to [Gammulle et al., 2017]. The performance of these strategies is compared in Tab. 5.4. Our proposed early fusion approach clearly outperforms all other strategies by 1.2-1.5 % points under the 14G protocol and by 1.1-1.6 % points under the 28G protocol. Compared to no fusion and late fusion, early fusion benefits from joint temporal modelling of posture and position features. Meanwhile, intermediate fusion might suffer from overfitting and complicated optimization due to increased model complexity.

Table 5.4: Recognition accuracy (%) for different fusion types of local and global features on the Shrec dataset for 14G and 28G protocol.

Fusion type	**14G**	**28G**
No fusion	94.6 ± 0.4	94.1 ± 0.4
Late fusion	94.6 ± 0.7	93.6 ± 0.7
Intermediate fusion	94.9 ± 0.6	93.8 ± 0.6
Early fusion (ours)	$\mathbf{96.1 \pm 0.4}$	$\mathbf{95.2 \pm 0.4}$

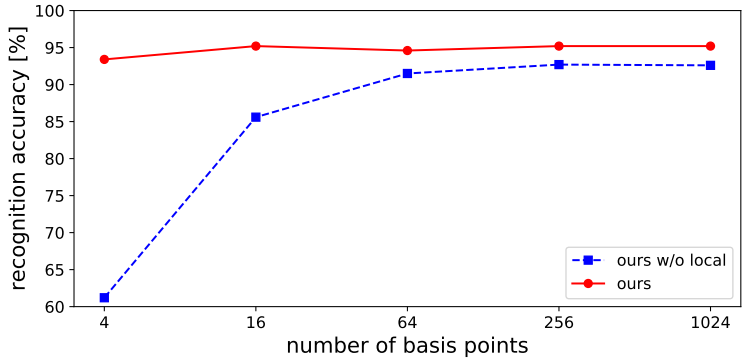


Fig. 5.4: Recognition accuracy of our full model and our model without local stream on the 28G protocol of the Shrec dataset as a function of the number of basis points.

Number of basis points. Finally, we investigate the influence of the number of basis points k on model performance. Changing k has a similar influence as modifying the resolution of a grid because larger k allows to capture finer details of the input point cloud. We train and evaluate both our full model and our model without local branch for a varying number of basis points under the 28G protocol of the Shrec dataset. Results are visualized in Fig. 5.4. Regarding the full framework, the performance is insensitive to the number of basis points and only slightly decreases for very few basis points. This behavior is expected because fine structures are captured in the local branch, and the global branch only needs to capture complementary coarse position features. For the isolated global branch, the performance substantially drops for a decreasing amount of basis points. The reason is that the BPS encoding needs to capture fine local structures in addition to the global position, which is hardly possible at low resolution.

5.1.3.4 Comparison to the State of the Art

We compare our approach to a comprehensive set of competing state-of-the-art methods, including Key frames [De Smedt et al., 2017], SoCJ+HoHD+HoWR [De Smedt et al., 2016], CNN+LSTM [Nunez et al., 2018], Res-TCN/STA-Res-TCN [Hou et al., 2018], ST-GCN [Yan et al., 2018], ST-TS-HGR-NET [Nguyen et al., 2019], DG-STA [Chen et al., 2019], HPEV+HMM+FRPV [Liu et al., 2020a], DSTA-Net [Shi et al., 2020], and PointLSTM [Min et al., 2020]. The comparison is performed on the Shrec'17 and DHG dataset, and results are reported in Tab. 5.5. On the Shrec'17 dataset, our method achieves an accuracy of 96.1 % for the 14G protocol and of 95.2 % for the 28G protocol and thus surpasses the state-of-the-art point cloud-based method (PointLSTM [Min et al., 2020]) by 0.2 and 0.5 % points, respectively. Compared to the skeleton-based

Table 5.5: Recognition accuracy (%) of our approach compared to the state of the art on the Shrec and DHG datasets for 14G and 28G protocol.

		Shrec		**DHG**	
Method	**Modality**	**14G**	**28G**	**14G**	**28G**
Key frames	Depth	82.9	71.9	N/A	N/A
SoCJ+HoHD+HoWR	Skeleton	88.2	81.9	83.1	80.0
CNN+LSTM	Skeleton	N/A	N/A	85.6	81.1
Res-TCN	Skeleton	91.1	87.3	86.9	83.6
STA-Res-TCN	Skeleton	93.6	90.7	89.2	85.0
ST-GCN	Skeleton	92.7	87.7	91.2	87.1
ST-TS-HGR-NET	Skeleton	94.3	89.4	87.3	83.4
DG-STA	Skeleton	94.4	90.7	91.9	88.0
HPEV+HMM+FRPV	Skeleton	94.9	92.3	92.5	88.9
DSTA-Net	Skeleton	**97.0**	93.9	**93.8**	90.9
PointLSTM-middle	Point cloud	95.9	94.7	N/A	N/A
Ours	Point cloud	96.1	**95.2**	92.0	**91.7**

methods, our method achieves top performance on the 28G protocol. While the most recent DSTA-Net [Shi et al., 2020] is outperformed by 1.3 % points, most preceding methods are exceeded by a clear margin of more than 4 % points. Under the 14G protocol, our method still achieves competitive results and is superior to all methods but DSTA-Net, which is 0.9 % points better.

Qualitatively, these observations can be transferred to the DHG dataset. As a first general remark, we note that all methods achieve a worse accuracy on the DHG dataset than on the Shrec dataset although the datasets share the same gesture categories. Beyond, we are the first to benchmark a point cloud-based method on the DHG dataset and hence only compare to skeleton-based methods. Under the 28G protocol, our model achieves an accuracy of 91.7 % and is thus superior to all skeleton-based methods. On the 14G protocol, our method achieves a competitive accuracy of 92.0 % but lags behind HPEV+HMM+FRPV [Liu et al., 2020a] and DSTA-Net [Shi et al., 2020] by 0.5 and 1.8 % points, respectively.

In summary, our method achieves new state-of-the-art performance under the challenging 28G protocol on both the Shrec and DHG dataset while being competitive under the 14G protocol of both datasets. We firstly conclude that our method constitutes an efficient alternative to PointLSTM for point cloud-based gesture recognition. Secondly, the results suggest that the use of point clouds can be generally advantageous over skeleton data when dealing with challenging settings where a high level of detail needs to

Table 5.6: Number of model parameters and inference times of our proposed framework and the state-of-the-art PointLSTM.

Model	#Paras	Time [ms]
PointLSTM-middle	1.3M	25.0
Ours w/o local	2.4M	1.4
Ours local=PointNet	3.6M	1.8
Ours	4.6M	7.8

be captured (28G protocol). Whereas hand skeletons constitute a sparse representation of the hand, point clouds represent the entire visible surface of the hand and thus naturally preserve finer details of the hand posture.

5.1.3.5 Model Size and Inference Time

In Tab. 5.6, we compare the number of parameters and inference times of multiple models. Inference times have been measured on a Titan RTX for a single forward pass of a sequence with length $T = 32$. Regarding the number of model parameters, all our models exceed PointLSTM. This is primarily attributable to fully-connected layers, which are used in all components of our framework (DGCNN, DenseNet, LSTM). But as forward passes through fully-connected layers are fast, the increased number of parameters has no negative impact on inference time. On the contrary, our proposed framework is more than three times faster than PointLSTM. Replacing the DGCNN in the local branch by a more efficient PointNet, which still achieves state-of-the-art performance (Tab. 5.3), reduces the inference time by another factor of 4 and is more than 13 times faster than PointLSTM. Our model without local stream, which is on par with the state of the art under the 14G protocol (Tab. 5.1), further reduces the inference time by around 20 %.

5.1.4 Discussion and Conclusion

In this work, we propose a novel framework for point cloud-based gesture recognition. Our key idea is to learn two separate spatial embeddings in order to explicitly capture both hand position and hand posture evolution. This is achieved by combining the complementary strengths of two different representations (BPS encoding and raw point cloud) and associated network types (fully-connected network and point network) in a two-stream architecture. In experiments, we demonstrate that our approach is competitive with state-of-the-art while being computationally more efficient. Contrary to previous methods [Min et al., 2019, 2020], our framework does not include any modules for low-level spatio-temporal feature extraction. In future work, we intend

to explore how to efficiently combine our two-stream approach with spatio-temporal feature extraction on the point level.

Chapter 6

Domain Adaptive Lung Point Cloud Registration

The final methodological chapter unites the research topics from Chapters 4 (domain adaptation) and 5 (sequence analysis) by addressing domain adaptive point cloud registration. The considered application is exhale-to-inhale lung registration, i.e., unlike in all previous chapters, the point clouds now represent keypoints in medical scans. As the first methodological contribution, Sec. 6.1 presents a novel self-training scheme, which adapts the Mean Teacher paradigm to the registration problem. The adaptation involves adjusting the geometric augmentation scheme and integrating the combination of a learnable feature extraction module with a differentiable optimizer, whose regularizing effect implicitly stabilizes the training under inherently noisy pseudo supervision from the teacher. Subsequently, Sec. 6.2 presents two methods to explicitly denoise the supervision by the teacher, including a novel strategy to filter out detrimental pseudo labels and a generative approach that dynamically synthesizes new training samples with precise displacement fields for supervision.

6.1 Adapting the Mean Teacher for Point Cloud Registration

This section has been published in [Bigalke et al., 2022b]. According to the Contributor Roles Taxonomy (CRediT), the contributions of the author of this thesis to the publication are: Conceptualization (together with L.H., M.P.H.), Methodology, Software, Investigation, Writing – Original Draft, Writing – Review & Editing (together with L.H., M.P.H.), Visualization. The source code is available at `https://github.com/multimodallearning/registration-da-mean-teacher`.

6.1.1 Introduction

Image registration is a fundamental task in medical image analysis, for instance required for multi-modal data fusion or patient monitoring over time. For a long time, the state of the art for image registration was dominated by conventional optimization methods

[Sotiras et al., 2013], whose high accuracy comes at the cost of high run times. In recent years, learning-based methods—driven by deep neural networks—achieved competitive performances [Haskins et al., 2020]. These methods substantially reduce inference times, but they involve two other significant drawbacks. First, high performance is strongly dependent on the availability of labeled training data, which are costly to collect. Second, deep neural networks often generalize poorly to shifted domains. Once trained in a labeled source domain, the models are likely to suffer a performance drop when deployed on data from a shifted target domain. While shifts in intensity distributions—typical for medical imaging—can be mitigated by keypoint-based registration [Hansen et al., 2021], such methods still suffer from geometric domain shifts, for instance, due to varying fields of view under different imaging protocols. Fine-tuning or re-training on data from the shifted domain could alleviate the performance drop but is often impractical due to high labeling costs. Alternatively, domain adaptation [Wang et al., 2018] is a promising technique to adapt a model from a labeled source to an unlabeled target domain. While extensively explored for medical classification and segmentation tasks [Guan et al., 2021], domain adaptation for image registration has rarely been studied in the literature [Kruse et al., 2021; Mahapatra et al., 2020] and will be the focus of this work.

6.1.1.1 Related Work

Existing works on domain adaptive registration rely on two different concepts. Mahapatra et al. [2020] increase the invariance of a generative registration model to the type of input images by encoding the images to the latent space of an autoencoder. In a different approach, Kruse et al. [2021] adapt the concept of maximum classifier discrepancy [Saito et al., 2018] to multi-modal registration. However, this requires the quantization of displacement fields and the use of a classification architecture instead of state-of-the-art registration models. Generally, domain adaptive registration is challenging because concepts established for other tasks are often unsuitable for registration. The mainstream approach of domain-invariant feature learning through adversarial learning [Ganin et al., 2015; Tzeng et al., 2017] or reconstruction [Bousmalis et al., 2016; Ghifary et al., 2016], for instance, was primarily used in classification tasks. In consequence, the methods focus on the alignment of global feature vectors. This is insufficient for registration, which highly depends on the identification of local correspondences. Alternatively, Tsai et al. [2018] proposed domain adaptation in the output space, where the distributions of predictions in source and target domain are aligned through adversarial learning. This concept was successfully applied to semantic segmentation [Tsai et al., 2018] and human pose estimation [Yang et al., 2018]. However, the raw displacement fields output by registration models are less structured than segmentation masks or human poses such that aligning their distributions might be ineffective. Instead of distribution matching, self-ensembling [French et al., 2018] addresses

the domain shift by imposing consistency constraints in the output space. The method is based on the framework of the Mean Teacher [Tarvainen et al., 2017], comprising a learning student model and a so-called teacher model, which represents a temporal ensemble with its weights corresponding to the exponential moving average (EMA) of the student model. Supervision on unlabeled target data is then provided by enforcing consistent predictions of student and teacher model. Beyond classification [French et al., 2018], this concept has successfully been adapted to diverse tasks, including medical image segmentation [Perone et al., 2019; Yu et al., 2019] and clinical human pose estimation [Srivastav et al., 2022].

6.1.1.2 Contributions

In this work, we extend the concept of self-ensembling to geometric domain adaptation for image registration, embedding a keypoint-based registration model into the Mean Teacher paradigm (see Fig. 6.1). Our framework is built upon the keypoint-based method by Hansen et al. [2021], which aligns two point clouds (extracted from the input scan pair) by combining a Graph Convolutional Network (GCN) for learned geometric feature extraction with loopy belief propagation (LBP) for alignment. To incorporate the method into the Mean Teacher framework, we present two crucial modifications of the standard Mean Teacher. First, we adapt the stochastic augmentation scheme to the specific characteristics of the registration task by incorporating inverse geometric transformations. Second, we present the first Mean Teacher that combines learned feature extraction (GCN) with differentiable optimization (LBP). Notably, the differentiability of LBP enables us to impose the consistency constraint between student and teacher model on the final predicted displacement fields after LBP. That way, the adaptation of the GCN can benefit from the regularizing effect of LBP. Overall, our method offers a simple and robust adaptation procedure. It involves only few hyper-parameters and does not rely on intricate adversarial optimization as previous works. We experimentally evaluate our method for exhale-to-inhale lung CT registration, considering two different geometric domain shifts (varying field of view and breathing type). Results demonstrate substantial improvements compared to the baseline model and even competitive performance with fully-supervised models.

6.1.2 Methods

6.1.2.1 Problem Statement

Given fixed and moving input scans $\boldsymbol{F}, \boldsymbol{M} \in \mathbb{R}^{H \times W \times D}$, the goal of registration is to predict a displacement field $\boldsymbol{\varphi}$ that spatially aligns $\boldsymbol{M}$ to $\boldsymbol{F}$ within a given region of interest. We address this task in the classical domain adaptation setting, where training data comprises a labeled source dataset $\mathcal{S}$ and an unlabeled target dataset $\mathcal{T}$. $\mathcal{S}$ consists of triplets $(\boldsymbol{F}^s, \boldsymbol{M}^s, \boldsymbol{\varphi}^s)$ of scan pairs $(\boldsymbol{F}^s, \boldsymbol{M}^s)$ with corresponding ground

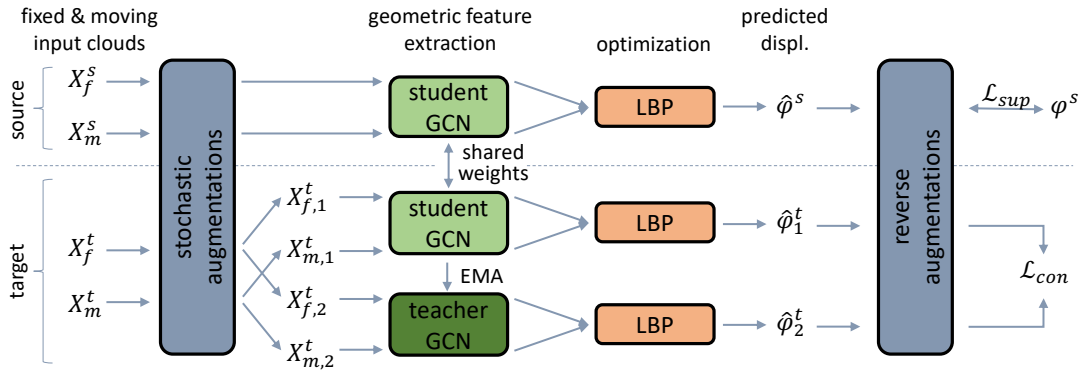


Fig. 6.1: Overview of our proposed self-ensembling framework for domain adaptive keypoint-based image registration.

truth displacement fields $\boldsymbol{\varphi}^s$. $\mathcal{T}$ contains scan pairs $(\boldsymbol{F}^t, \boldsymbol{M}^t)$ without any ground truth. Given the training data, we aim to learn a function f with parameters $\boldsymbol{\theta}$ that predicts displacement fields $\hat{\boldsymbol{\varphi}} = f(\boldsymbol{F}, \boldsymbol{M}; \boldsymbol{\theta})$ and achieves the best possible performance on the target domain. In the following, we first introduce our baseline model f for standard supervised learning on source data (Sec. 6.1.2.2) and subsequently present our self-ensembling framework for domain adaptation (Sec. 6.1.2.3).

6.1.2.2 Baseline Model

Most learning-based methods perform registration based on intensities in dense voxel space. Instead, we formulate registration as the pure geometric alignment of two point clouds $\boldsymbol{X}_f \in \mathbb{R}^{N_f \times 3}, \boldsymbol{X}_m \in \mathbb{R}^{N_m \times 3}$, representing distinct keypoints of the input scans. The underlying motivation is that this reduces the vulnerability to intensity shifts between source and target domain, enabling us to focus on remaining geometric shifts. The general efficacy of learning-based registration of point clouds in a fully-supervised setting has previously been demonstrated by Hansen et al. [2021]. We adopt their model as our baseline and briefly summarize its major components.

First, fixed and moving point clouds $\boldsymbol{X}_f$ and $\boldsymbol{X}_m$ are extracted from the input scans using the Förstner algorithm and non-maximum suppression [Heinrich et al., 2015]. Second, a GCN g extracts point-wise geometric features $g(\boldsymbol{X}_f; \boldsymbol{\theta}_g)$, $g(\boldsymbol{X}_m; \boldsymbol{\theta}_g)$ from the raw coordinates of both clouds. The GCN is based on the edge convolution from [Wang et al., 2018], operating on the knn-graph of the clouds to account for neighborhood relations. Third, the extracted features are used to guide the inference of final displacement vectors with LBP. Specifically, correspondence probabilities between fixed keypoints and candidate sets from the moving cloud are computed via LBP by jointly minimizing a data cost and a regularization cost. While the latter enforces smoothness of the predicted displacement field, the data cost is defined as the distance

between geometric features and induces high correspondence probabilities between points with similar features. Thus, effective feature extraction with the GCN is crucial for the assignment of accurate correspondences. Finally, displacement vectors are inferred by integrating the probabilities over the corresponding displacements between fixed and candidate points, allowing for more accurate displacement vectors than hard assignments. We formally summarize the combination of GCN and LBP as

$$\hat{\boldsymbol{\varphi}} = LBP(g(\boldsymbol{X}_f;\boldsymbol{\theta}_g), g(\boldsymbol{X}_m;\boldsymbol{\theta}_g)) =: f(\boldsymbol{X}_f, \boldsymbol{X}_m;\boldsymbol{\theta}), \quad \hat{\boldsymbol{\varphi}} \in \mathbb{R}^{N_f \times 3} \tag{6.1}$$

and refer the reader to [Hansen et al., 2021] for more details. As such, f is fully differentiable, enabling end-to-end learning of parameters $\boldsymbol{\theta}$ by minimizing the supervised loss

$$\mathcal{L}_{sup} = \|\hat{\boldsymbol{\varphi}} - \boldsymbol{\varphi}\|_1 \tag{6.2}$$

6.1.2.3 Domain-Adaptive Registration with the Mean Teacher

To adapt the baseline model to a shifted target domain, we propose a novel self-ensembling framework, embedding the model into the Mean Teacher paradigm [French et al., 2018; Tarvainen et al., 2017]. An overview of the method is shown in Fig. 6.1. The framework extends the baseline model, and now includes two GCNs for feature extraction, namely a student GCN with weights $\boldsymbol{\theta}$ and a teacher GCN with weights $\boldsymbol{\theta}'$. The student network is optimized via gradient descent, whereas the weights of the teacher are the exponential moving average of the student's weights, updated as

$$\boldsymbol{\theta}'_i = \alpha\boldsymbol{\theta}'_{i-1} + (1-\alpha)\boldsymbol{\theta}_i \tag{6.3}$$

at iteration i with the momentum α. The student is optimized by minimizing the joint loss function

$$\mathcal{L}(\boldsymbol{\theta};\boldsymbol{\theta}',\mathcal{S},\mathcal{T}) = \mathcal{L}_{sup}(\boldsymbol{\theta};\mathcal{S}) + \lambda(t)\mathcal{L}_{con}(\boldsymbol{\theta};\boldsymbol{\theta}',\mathcal{T}) \tag{6.4}$$

composed of the supervised loss $\mathcal{L}_{sup}$ (cf. Eq. 6.2) on labeled source data and a consistency loss $\mathcal{L}_{con}$ on unlabeled target data, weighted by a time-dependent factor $\lambda(t)$. The consistency loss penalizes different predictions by student and teacher model. Implementing $\mathcal{L}_{con}$ for the given registration problem requires two subtle but decisive adaptations of the standard Mean Teacher for classification.

First, the standard Mean Teacher imposes the consistency constraint at the output of the learning network, which usually coincides with the final prediction. In our case, however, the output of the learning GCN is an intermediate representation, and we observed that applying consistency constraints at this level hampers the learning process. Instead, we propose to align predicted displacement vectors after LBP. That way, the adaptation process can benefit from the regularizing effect of LBP.

Second, an important component of the Mean Teacher is the application of different stochastic augmentations to the input of teacher and student streams. Unlike classification, however, registration requires the augmentation of pairs of inputs instead of single inputs. Moreover, the associated displacement fields in the output space are not invariant to the input transformation (as is the case for classification), but they are transformed together with the input. For $\mathcal{L}_{con}$ to be meaningful, transformations applied at the input level need to be reversed at the output level such that predictions in teacher and student streams are aligned. To ensure reversibility at the output level, we sample one transformation for each stream and apply it to both fixed and moving cloud, yielding two augmented pairs $(\boldsymbol{X}^t_{f,1}, \boldsymbol{X}^t_{m,1})$,$(\boldsymbol{X}^t_{f,2}, \boldsymbol{X}^t_{m,2})$. In practice, we transform point clouds by random rotation, translation, and scaling. The reverse transformation of the displacement vectors is then given by inverse scaling and rotation (displacement vectors are invariant to the synchronous translation of both inputs). Denoting reverse augmentations as $\text{aug}^{-1}(.)$, we formalize the consistency loss as

$$\mathcal{L}_{con} = \left\| \text{aug}_1^{-1}(f(\boldsymbol{X}^t_{f,1}, \boldsymbol{X}^t_{m,1}; \boldsymbol{\theta})) - \text{aug}_2^{-1}(f(\boldsymbol{X}^t_{f,2}, \boldsymbol{X}^t_{m,2}; \boldsymbol{\theta}')) \right\|_1 \tag{6.5}$$

At early epochs, the weights of the teacher model are still close to the random initialization, inducing noisy gradients from $\mathcal{L}_{con}$. Therefore, the weighting factor $\lambda(t) = \lambda_0 \cdot \exp(-5(1 - \min(t/T, 1)^2))$ depends on the current epoch t and steadily increases from 0 to λ_0 during the first T epochs, as suggested by [Tarvainen et al., 2017].

6.1.3 Experiments and Results

6.1.3.1 Experimental Setup

Datasets and pre-processing. We evaluate our method for exhale-to-inhale lung CT registration under two adaptation scenarios using three public datasets. First, we consider the DIR-Lab COPD dataset [Castillo et al., 2013] as source and the Learn2Reg (L2R) Task 2 dataset [Hering et al., 2020] as target domain. The domain shift consists in exhale scans from the target domain exhibiting a cropped field of view such that upper and lower parts of the lungs are partially cut off. Our training data comprise 10 labeled scan pairs from COPD and 12 unlabeled pairs from the train/val split of L2R. Evaluation is performed on the official test split of L2R (10 scan pairs). Second, we perform adaptation from the DIR-Lab 4D CT dataset [Castillo et al., 2009] as the source to the COPD dataset as the target domain. While scans from 4D CT were acquired from patients with shallow resting breathing, scan pairs from COPD show actively forced full inhalation and exhalation. Thus, the domain gap consists in the breathing type, yielding deformations with different characteristics. Here, evaluation is performed via 5-fold cross-validation on the 10 scans from COPD. For each fold, training data include the 10 labeled scan pairs from 4D CT and 8 unlabeled scan

pairs from COPD, and we evaluate the trained model on the remaining 2 scan pairs from COPD. In both adaptation scenarios, labels for source data are available in the form of landmark correspondences. To supervise predicted displacement vectors for the keypoints in the fixed cloud, we interpolate displacements of the landmarks to the entire volume and grid-sample at the keypoint locations. Scans from all domains are pre-processed in an identical way. We resample inhale scans to $1.75 \times 1.25 \times 1.75\,\text{mm}$ and exhale scans to $1.75 \times 1.00 \times 1.25\,\text{mm}$ to compensate for a different volume scaling during inspiration and expiration. Subsequently, we crop fixed-size regions of interest with $192 \times 192 \times 208\,\text{vx}$ around the center of automatically generated lung masks.

Implementation details. We implement our framework in PyTorch and optimize parameters of the student GCN with the Adam optimizer. Batches are composed of 4 source and 4 target samples. Training is performed for 100 epochs with a constant learning rate of 0.01. Hyper-parameters of the baseline model (GCN architecture and LBP parameters) are adopted from [Hansen et al., 2021]. Hyper-parameters of the self-ensembling framework are determined via cross-validation on the training samples from Learn2Reg and set to $\lambda_0 = 1$, $\alpha = 0.98$, and $T = 20$ epochs. Stochastic augmentations include random rotations around all axes by angles from $[-10°, 10°]$, random scaling by a global factor from $[0.9, 1.1]$, and random translation by a vector from $[-0.1, 0.1]^3$.

Baseline methods. 1) As a lower bound (*source-only*), we train the baseline model on source data without any adaptation techniques. 2) As an upper bound (*target-only*), we train the baseline model on labeled training data from the target domain. 3) As an alternative domain adaptation technique, we experimented with DANN [Ganin et al., 2015]. Specifically, we used max pooling to condense geometric features into a global feature vector and aligned the source and target distributions of these global features by adversarial learning. As expected, this approach turned out to be unsuitable for registration and lead to divergence. 4) As an intensity-based baseline method, we use the recent Voxelmorph++ (*VM++*) [Heinrich et al., 2022], which combines an extension of Voxelmorph with instance optimization through Adam. We train the model once on source and once on target data.

Metrics. We interpolate predicted displacement vectors at sparse keypoints to the entire volume and report the mean target registration error (TRE) at available landmark correspondences. We also inspected the regularity of the interpolated deformation fields. However, due to the regularizing effect of LBP, all methods predict smooth deformation fields (percentages of folding voxels $|J_\varphi|_{<0} < 10^{-4}$) such that a quantitative comparison does not provide additional insights.

Table 6.1: Results for both adaptation scenarios, reported as TRE in mm.

Method	COPD→L2R	4D CT→COPD
Initial	10.24	11.99
VM++ source	4.34	3.55
VM++ target	3.09	2.46
Source-only	4.96	4.02
Target-only	2.61	2.26
Ours	2.64	2.01
Ours + Adam	**2.38**	**1.93**

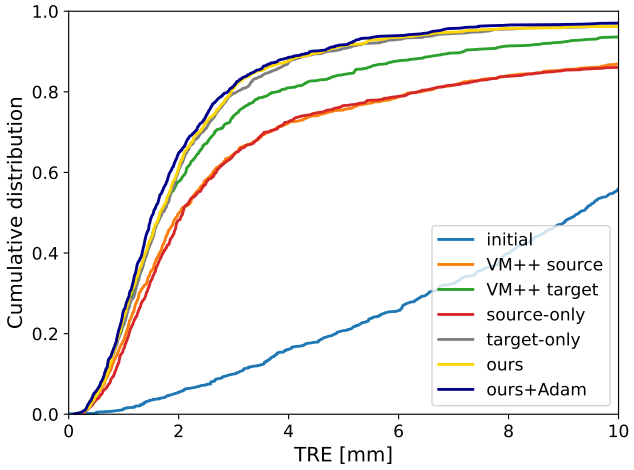


Fig. 6.2: Cumulative target registration errors for adaptation from COPD to L2R.

6.1.3.2 Results

Quantitative results of our experiments are shown in Tab. 6.1 and reveal consistent findings under both adaptation scenarios. First, both our source-only model and the VM++ source model perform substantially worse than their respective counterparts trained on target data, increasing the TRE by 40 to 90 %. This demonstrates that both considered domain shifts pose severe problems for intensity- and keypoint-based methods and need to be addressed by effective adaptation methods. Second, and most importantly, our proposed method effectively adapts the baseline model to the target domain. Specifically, it achieves mean TREs of 2.64 mm and 2.01 mm, respectively, thus improving on the source-only model by 47 %/50 % while matching or even surpassing the performance of the target-only model. This demonstrates that our self-ensembling framework effectively leverages unlabeled target data to close the domain gap. Third,

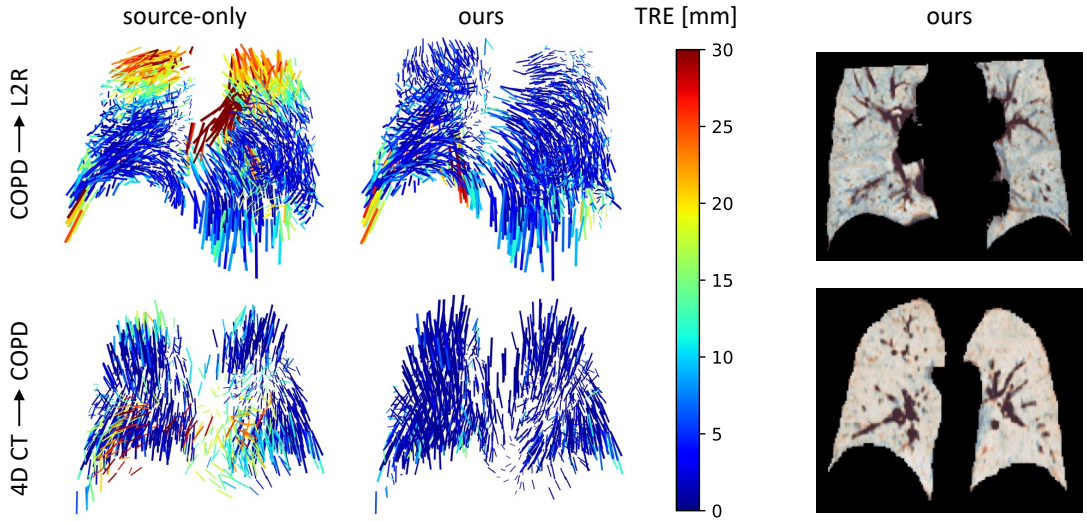


Fig. 3: Qualitative results for one sample case from each adaptation scenario. The first two columns show predicted displacement vectors by the source-only model and by our model. The linewidth is proportional to the distance of displacements, and colors encode the TRE (clamped to 30 mm). The last column shows overlaid CT slices after registration by our method. Inhale and exhale scans are shown in orange and blue shades, respectively, adding up to grayscale in case of alignment.

we investigate how far we can push the accuracy of our method by fine-tuning the displacement fields with MIND-based instance optimization with Adam [Siebert et al., 2021]. This further reduces the TRE to 2.38 mm and 1.93 mm. A detailed comparison of all discussed methods for COPD→L2R is shown in Fig. 6.2. Finally, it is notable that our method compares favorably to the state-of-the-art LapIRN [Mok et al., 2021], the winner of the recent Learn2Reg challenge, which is slightly superior on the L2R test set (TRE=1.98 mm) but clearly inferior to our method on COPD (TRE=3.83 mm).

Qualitative results are presented in Fig. 3. The first two columns visualize the effect of the studied domain gaps. For COPD→L2R, errors of the source-only model primarily occur in the superior part of the lung, which is partially outside the scanning region in exhale scans of the target domain while fully visible on source data. For 4DCT→COPD, the source-only model mainly fails in the anterior region of the lung, which is strongly deformed by full inspiration-expiration (target domain) but relatively static during shallow breathing (source domain). Our method substantially reduces these errors, highlighting the efficient adaptation to the target domain. Finally, the CT overlays (last column) show the largely accurate alignment of inner lung structures by our method.

6.1.4 Discussion and Conclusion

In this work, we addressed geometric domain adaptation in the context of image registration and proposed a novel self-ensembling framework. Specifically, we embedded a keypoint-based registration model into the Mean Teacher paradigm and thus guided the learning process in the target domain by consistency-based supervision. In our experiments for exhale-to-inhale registration of lung CT scans, we demonstrated that our method successfully reduced the domain gap under two challenging adaptation settings, including different breathing types and imaging protocols. Specifically, it surpassed the baseline model by 50 %/47 % and even matched the performance of a supervised model trained on labeled target data. These results indicate great potential of the Mean Teacher framework for medical image registration, demonstrating its capability to improve feature learning in the absence of labels. While our use case of keypoint-based registration under geometric shifts is rather specific, our method is flexible and can easily be adapted to diverse scenarios, including domain adaptation under intensity shifts with classical 3D CNNs and semi-supervised learning. However, experimental evaluation under these settings is needed and will be the subject of future work.

6.2 Denoising the Mean Teacher for Point Cloud Registration

This section has been published in [Bigalke et al., 2022b]. According to the Contributor Roles Taxonomy (CRediT), the contributions of the author of this thesis to the publication are: Conceptualization (together with M.P.H.), Methodology, Software, Investigation, Writing – Original Draft, Writing – Review & Editing (together with M.P.H.), Visualization. The source code is available at `https://github.com/multimodallearning/denoised_mt_pcd_reg`.

6.2.1 Introduction

Recent deep learning-based registration methods have shown great potential in solving medical image registration problems [Fu et al., 2020; Haskins et al., 2020]. Most of these methods perform the registration based on the raw volumetric intensity images, e.g., [Balakrishnan et al., 2019; Chen et al., 2022a; De Vos et al., 2019; Mok et al., 2020; Zhao et al., 2019]. By contrast, only a few recent works [Hansen et al., 2021; Shen et al., 2021] operate on sparse, purely geometric point clouds extracted from the images, even though this representation promises multiple potential benefits, including computational efficiency, robustness against intensity shifts in the image domain, and anonymity preservation. The latter, for instance, can facilitate public data access and federated learning, as exemplified by a recently released point cloud dataset of lung

vessels [Shen et al., 2021] whose underlying CT scans are not publicly accessible. On the other hand, the sparsity of point clouds and the absence of intensity information make the registration problem more challenging. In particular, unsupervised learning with similarity metrics – as established for dense image registration [Chen et al., 2022a; Mok et al., 2020] – was shown ineffective for deformable point cloud registration [Shen et al., 2021], as confirmed by our experiments. Since manual annotations for supervised learning are prohibitively costly, an alternative consists of training on synthetic deformations with known displacements [Shen et al., 2021], as known from dense registration [Eppenhof et al., 2018; Uzunova et al., 2017]. The inevitable domain gap between synthetic and real deformations, however, involves the risk of suboptimal performance on real data. In this work, we aim to bridge this gap through domain adaptation (DA).

6.2.1.1 Related Work

DA has widely been studied for classification and segmentation tasks [Guan et al., 2021], with popular techniques ranging from adversarial feature [Ganin et al., 2015; Tzeng et al., 2017] or output [Tsai et al., 2018] alignment to self-supervised feature learning [Sun et al., 2019b]. However, these methods are insufficient for the specific characteristics of the registration problem, involving a more complex output space and requiring the detection of local correspondences. Instead, recent works adapted the Mean Teacher paradigm [Tarvainen et al., 2017], previously established for domain adaptive classification [French et al., 2018] and segmentation [Perone et al., 2019], to the registration problem [Bigalke et al., 2022b; Jin et al., 2022; Xu et al., 2022]. The basic idea is to supervise the learning student model with displacement fields (pseudo labels) provided by a teacher model, whose weights represent the exponential moving average of the student's weights. A significant limitation of this method, however, is the inevitable noise in the pseudo labels, potentially misguiding the adaptation process. Prior works addressed this problem by refining the pseudo labels [Jin et al., 2022] or weighting them according to model uncertainty, estimated through Monte Carlo dropout [Xu et al., 2022; Yu et al., 2019]. However, even refined pseudo labels remain inaccurate, and the proposed refinement strategy [Jin et al., 2022] assumes piecewise rigid motions of 3D objects and does not apply to complex deformations in medical applications. And weighting pseudo labels according to teacher uncertainty [Xu et al., 2022; Yu et al., 2019] does not explicitly consider the quality of the actual registrations, completely ignores the quality and certainty of the current student predictions, and can, therefore, not prevent detrimental supervision of the student through inferior teacher predictions.

6.2.1.2 Contributions

We introduce two complementary strategies to denoise the Mean Teacher for domain adaptive point cloud registration, addressing the above limitations (see Fig. 6.4). Both strategies are based on our understanding of an optimal student-teacher relationship. First, if the student's solution to a problem is superior to that of the teacher, good teachers should not insist on their solution but accept the student's approach. To implement this, inspired by a recent technique to filter pseudo labels for human pose estimation [Bigalke et al., 2023a], we propose to assess the quality of both the teacher and student registrations with the Chamfer distance and to provide only those registrations of the teacher as supervision to the student that are more accurate. This approach differs from previous uncertainty-based methods [Xu et al., 2022; Yu et al., 2019] in two decisive aspects: 1) It explicitly assesses the quality of final registrations, using a model-free and objective measure with little computational overhead compared to multiple forward passes in Monte Carlo dropout. 2) The selection process considers both teacher and student predictions and can thus prevent detrimental supervision by the teacher. Our second strategy follows the intuition that good teachers should not pose problems to which they do not know the solution. Instead, they should come up with novel tasks with precisely known solutions. Consequently, we propose a completely novel teacher paradigm, where predicted deformations by the teacher are used to synthesize new training pairs for the student, consisting of the original moving inputs and their warps. These input pairs come with precise noise-free displacement labels and significantly differ from static hand-crafted synthetic deformations [Shen et al., 2021]. 1) The deformations are based on a real data pair that the teacher aims to align. 2) The deformations are dynamic and become more realistic as the teacher improves. Finally, we unify both strategies in a joint framework for domain adaptive point cloud registration. It is compatible with arbitrary geometric registration models, stable to train, and involves only a few hyper-parameters. We experimentally evaluate the method for inhale-to-exhale registration of lung vessel point clouds on the public PVT dataset [Shen et al., 2021], demonstrating substantial improvements over diverse competing methods and state-of-the-art performance.

6.2.2 Methods

6.2.2.1 Problem Setup and Standard Mean Teacher

In point cloud registration, we are given fixed and moving point clouds $\boldsymbol{F} \in \mathbb{R}^{N_\mathrm{F} \times 3}, \boldsymbol{M} \in \mathbb{R}^{N_\mathrm{M} \times 3}$ and aim to predict a displacement vector field $\boldsymbol{\varphi} \in \mathbb{R}^{N_\mathrm{M} \times 3}$ that spatially aligns $\boldsymbol{M}$ to $\boldsymbol{F}$ as $\boldsymbol{M} + \boldsymbol{\varphi}$. We address the task in a domain adaptation setting with training data comprising a labeled source dataset $\mathcal{S}$ of triplets $(\boldsymbol{M}_s, \boldsymbol{F}_s, \boldsymbol{\varphi}_s)$ and a shifted unlabeled target dataset $\mathcal{T}$ of tuples $(\boldsymbol{M}_t, \boldsymbol{F}_t)$. While the formulation of our method is agnostic to the specific domain shift between $\mathcal{S}$ and $\mathcal{T}$, in this work, we generate

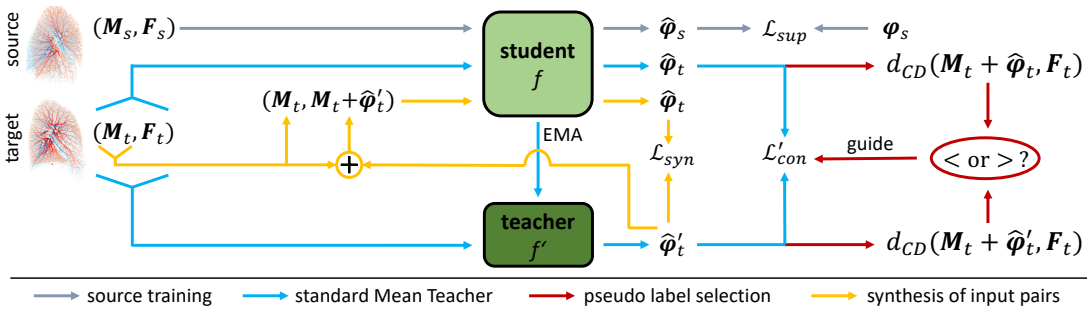


Fig. 6.4: Overview of our denoised Mean Teacher for domain adaptive registration. We overcome noisy supervision by the teacher with a novel pseudo label selection strategy and the synthesis of new training pairs with precisely known displacements.

the source samples on the fly as random synthetic deformations of the target clouds using a fixed hand-crafted deformation function $def : \mathbb{R}^{N\times 3} \to \mathbb{R}^{N\times 3}$, i.e., source triplets are given as $(def(\boldsymbol{F}_t), \boldsymbol{F}_t, \boldsymbol{F}_t - def(\boldsymbol{F}_t))$ or $(def(\boldsymbol{M}_t), \boldsymbol{M}_t, \boldsymbol{M}_t - def(\boldsymbol{M}_t))$. Note that def preserves point correspondences enabling ground truth computation through point-wise subtraction. Given the training data, we aim to learn a function f that predicts deformation vector fields as $\hat{\boldsymbol{\varphi}} = f(\boldsymbol{M}, \boldsymbol{F})$ with optimal performance in the target domain.

Baseline Mean Teacher. To solve the problem, the standard Mean Teacher framework [Bigalke et al., 2022b; French et al., 2018; Tarvainen et al., 2017] employs two identical networks, denoted as the student f and teacher f', with parameters $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$. While the student's weights $\boldsymbol{\theta}$ are optimized through gradient descent, the teacher's weights correspond to the exponential moving average (EMA) of the student and are updated as $\boldsymbol{\theta}'_i = \alpha\boldsymbol{\theta}'_{i-1} + (1-\alpha)\boldsymbol{\theta}_i$ at iteration i with momentum α. Meanwhile, the student is trained by minimizing

$$\mathcal{L}(\boldsymbol{\theta}) = \lambda_1 \underbrace{\|f(\boldsymbol{M}_s, \boldsymbol{F}_s) - \boldsymbol{\varphi}_s\|_2^2}_{\mathcal{L}_{\text{sup}}} + \lambda_2 \underbrace{\|f(\boldsymbol{M}_t, \boldsymbol{F}_t) - f'(\boldsymbol{M}_t, \boldsymbol{F}_t)\|_2^2}_{\mathcal{L}_{\text{con}}} \tag{6.6}$$

consisting of the supervised loss $\mathcal{L}_{\text{sup}}$ on source data and the consistency loss $\mathcal{L}_{\text{con}}$ on target data, weighted by λ_1 and λ_2. $\mathcal{L}_{\text{con}}$ guides the learning of the student in the target domain with pseudo-supervision from the teacher, which, as a temporal ensemble, is expected to be superior to the student. Nonetheless, predictions by the teacher can still be noisy and inaccurate, limiting the efficacy of the adaptation process.

6.2.2.2 Chamfer Distance-Based Filtering of Pseudo Labels

In a worst-case scenario, the student might predict an accurate displacement field $\hat{\boldsymbol{\varphi}}_t$, which is strongly penalized by the consistency loss due to an inaccurate teacher predic-

tion $\hat{\boldsymbol{\varphi}}'_t$. To prevent such detrimental supervision, we aim to select only those teacher predictions for supervision that are superior to the corresponding student predictions, which, however, is complicated by the absence of ground truth. We, therefore, propose to assess the quality of student and teacher registrations by measuring the similarity/distance between fixed and warped moving clouds, with higher similarities/lower distances indicating more accurate registrations. Among existing similarity measures, we opt for the symmetric Chamfer distance [Wu et al., 2020], which computes the distance between two point clouds $\boldsymbol{X}, \boldsymbol{Y}$ as

$$d_{\mathrm{CD}}(\boldsymbol{X}, \boldsymbol{Y}) = \sum_{\boldsymbol{x} \in \boldsymbol{X}} \min_{\boldsymbol{y} \in \boldsymbol{Y}} \|\boldsymbol{x} - \boldsymbol{y}\|_2^2 + \sum_{\boldsymbol{y} \in \boldsymbol{Y}} \min_{\boldsymbol{x} \in \boldsymbol{X}} \|\boldsymbol{x} - \boldsymbol{y}\|_2^2 \tag{6.7}$$

While we experimentally found the Chamfer distance insufficient as a direct loss function – presumably due to sparse differentiability and susceptibility to local minima, we still observed a strong correlation between Chamfer distance and actual registration error, making it a suitable choice for our purposes. We also explored other measures (Laplacian curvature [Wu et al., 2020], Gaussian MMD [Feydy, 2020]), which proved slightly inferior (Supp., Tab. 2). Formally, we thus measure the quality of the student prediction $\hat{\boldsymbol{\varphi}}_t = f(\boldsymbol{M}_t, \boldsymbol{F}_t)$ as $d_{\mathrm{CD}}(\boldsymbol{M}_t + \hat{\boldsymbol{\varphi}}_t, \boldsymbol{F}_t)$ and analogously for the teacher prediction $\hat{\boldsymbol{\varphi}}'_t$. We then define our indicator function

$$I(\hat{\boldsymbol{\varphi}}_t, \hat{\boldsymbol{\varphi}}'_t) = \begin{cases} 1 & d_{\mathrm{CD}}(\boldsymbol{M}_t + \hat{\boldsymbol{\varphi}}'_t, \boldsymbol{F}_t) < d_{\mathrm{CD}}(\boldsymbol{M}_t + \hat{\boldsymbol{\varphi}}_t, \boldsymbol{F}_t) \\ 0 & \text{else} \end{cases} \tag{6.8}$$

and reformulate the consistency loss in Eq. (6.6) as

$$\mathcal{L}'_{\mathrm{con}} = I(\hat{\boldsymbol{\varphi}}_t, \hat{\boldsymbol{\varphi}}'_t) \cdot \|f(\boldsymbol{M}_t, \boldsymbol{F}_t) - f'(\boldsymbol{M}_t, \boldsymbol{F}_t)\|_2^2 \tag{6.9}$$

6.2.2.3 Synthesizing Inputs With Noise-Free Supervision

While the above filtering strategy mitigates detrimental supervision, the selected pseudo labels are still inaccurate. Therefore, we complement the strategy with a novel teacher paradigm, where the teacher dynamically synthesizes new training pairs with precisely known displacements for supervision. Specifically, given a teacher prediction $\hat{\boldsymbol{\varphi}}'_t = f'(\boldsymbol{M}_t, \boldsymbol{F}_t)$, we do not only use it to supervise the student on the same input pair but also generate a new input sample $(\boldsymbol{M}_t, \boldsymbol{M}_t + \hat{\boldsymbol{\varphi}}'_t)$ by warping $\boldsymbol{M}_t$ with $\hat{\boldsymbol{\varphi}}'_t$. The underlying displacement field is naturally precisely known, enabling noise-free training of the student by minimizing

$$\mathcal{L}_{\mathrm{syn}} = \|f(\boldsymbol{M}_t, \boldsymbol{M}_t + \hat{\boldsymbol{\varphi}}'_t) - \hat{\boldsymbol{\varphi}}'_t\|_2^2 \tag{6.10}$$

To our knowledge, there is no prior work with a similarly "generative" teacher model. Altogether, we train the student network by minimizing the loss

$$\mathcal{L}(\boldsymbol{\theta}) = \lambda_1 \mathcal{L}_{\text{sup}} + \lambda_2 \mathcal{L}'_{\text{con}} + \lambda_3 \mathcal{L}_{\text{syn}} \tag{6.11}$$

Technical details. The synthesized input pairs $(\boldsymbol{M}_t, \boldsymbol{M}_t + \hat{\boldsymbol{\varphi}}'_t)$ exhibit exact point correspondence, i.e., for each point in $\boldsymbol{M}_t$ exists a corresponding point in $\boldsymbol{M}_t + \hat{\boldsymbol{\varphi}}'_t$. That is usually not the case for real data pairs and thus introduces another domain shift, which prevented proper convergence in our initial experiments. To overcome the problem, we exploit that the original point clouds in a dataset, denoted as $\boldsymbol{M}_{t*}$, usually comprise more points than the subsampled clouds $\boldsymbol{M}_t$ that are fed to the network. Given predicted displacements $\hat{\boldsymbol{\varphi}}'_t$ for $\boldsymbol{M}_t$, we interpolate the displacement vectors to $\boldsymbol{M}_{t*}$ with an isotropic Gaussian kernel, yielding $\hat{\boldsymbol{\varphi}}'_{t*}$. The final input pair is then obtained by sampling disjoint point subsets from $(\boldsymbol{M}_{t*}, \boldsymbol{M}_{t*} + \hat{\boldsymbol{\varphi}}'_{t*})$, excluding one-to-one correspondences.

6.2.3 Experiments and Results

6.2.3.1 Experimental Setup

Datasets. We evaluate our method for inhale-to-exhale registration of lung vessel point clouds on the public PVT dataset [Shen et al., 2021] (`https://github.com/uncbiag/robot`, License: CC BY-NC-SA 3.0). The dataset comprises 1,010 such data pairs, which were extracted from lung CT scans as part of the IRB-approved COPDGene study (NCT00608764). Ten of these scan pairs are cases from the DIR-Lab COPDGene dataset [Castillo et al., 2013] and thus annotated with 300 landmark correspondences. We use these cases as the test set and split the remaining unlabeled pairs into 800 cases for training and 200 for validation (on synthetic deformations only). The original point clouds in the dataset have a very high resolution (∼100k points), making the processing with deep networks computationally costly. Therefore, we extract distinctive keypoints by local density estimation followed by non-maximum suppression. We extract two sets of such keypoints for each cloud: one with the ∼8k most distinctive points for inference, and another with ∼16k points, from which we randomly sample subsets during training for increased variability (see Sec. 6.2.2.3, technical details). Finally, we pre-align each pair by matching the mean and standard deviation of the coordinates.

Implementation details. The registration network f is implemented as the default 4-scale architecture of PointPWC-Net [Wu et al., 2020], operating on 8192 points per cloud. Following [Wu et al., 2020], we implement $\mathcal{L}_{\text{sup}}$, $\mathcal{L}'_{\text{con}}$, and $\mathcal{L}_{\text{syn}}$ as multi-scale losses. Optimization is performed with the Adam optimizer. We first pre-train the network on source data (batch size 4) for 160 epochs and subsequently minimize the joint loss in Eq. (6.11) for 140 epochs, both with a constant learning rate of 0.001,

which requires up to 11 GB and 13/23 h on an RTX2080. For joint optimization, we use mixed batches of 4 source and 4 target samples, set $\lambda_1 = \lambda_2 = \lambda_3 = 10$, and the EMA-parameter to $\alpha = 0.996$. While the original PVT data pairs represent the target domain in all experiments, we consider two variants of the function def to synthesize source data pairs: a realistic task-specific 2-scale random field similar to [Shen et al., 2021] and a simple rigid transformation. This enables us to evaluate our method under two differently severe domain shifts. Since real validation data are unavailable, hyper-parameters of all compared methods were tuned in a synthetic adaptation scenario, with the rigid deformations in the source and the 2-scale random field deformations in the target domain. For further implementation details, we refer to our public code.

Comparison methods. 1) The source-only model is exclusively trained on source data without DA. 2) We adopt the standard Mean Teacher [Bigalke et al., 2022b]. 3) An uncertainty-aware Mean Teacher (UA-MT), similar to [Xu et al., 2022; Yu et al., 2019]. 4) As proposed in [Wu et al., 2020], we performed purely unsupervised training on target data with a Chamfer loss. However, consistent with the findings in [Shen et al., 2021], this approach could not converge for complex geometric lung structures. Instead, we use the Chamfer loss on target data as an additional loss to complement supervised source training. 5) We guide the learning on target data with the cycle-consistency method from [Mittal et al., 2020]. 6) As a classical algorithm, we adapt sLBP [Hansen et al., 2021]. 7) We collect the results of two current SOTA methods, S-Robot and D-Robot, from [Shen et al., 2021], which combine deep networks (Point U-Net, PointPWC-Net), trained on synthetic deformations, with optimal transport modules. Note, however, that the experimental setup in [Shen et al., 2021] slightly differs from our setting in terms of more input points (60k vs. 8k) and additional input features (vessel radii), thus accessing more information.

Metrics. We interpolate the predicted displacements from the moving input cloud to the annotated moving landmarks with an isotropic Gaussian kernel ($\sigma = 5\,\text{mm}$) and measure the target registration error (TRE) with respect to the fixed landmarks. To assess the smoothness of the predictions, we interpolate the sparse displacement fields to the underlying image grid and measure the standard deviation of the logarithm of the Jacobian determinant (SDlogJ).

6.2.3.2 Results

Quantitative results are shown in Tab. 6.2 and reveal the following insights: 1) The source-only model benefits from realistic synthetic deformations in the source domain, yielding a 40.9% lower TRE. 2) The standard Mean Teacher proves effective under the weaker domain shift (−40.7% TRE compared to source-only) but only achieves

Table 6.2: Quantitative results on the PVT dataset, reported as mean TRE and 25/75% percentiles in mm and SDlogJ. † indicates a deviating experimental setup (Sec. 6.2.3.1).

	def = **2-scale rnd. field**				def = **rigid**			
Method	**TRE**	**25%**	**75%**	**SDlogJ**	**TRE**	**25%**	**75%**	**SDlogJ**
Initial	23.32	13.22	31.61	-	23.32	13.22	31.61	-
Pre-align	12.83	8.25	16.68	-	12.83	8.25	16.68	-
sLBP	3.62	1.24	3.29	0.038	3.62	1.24	3.29	0.038
S-Robot†	5.48	2.86	7.14	N/A	N/A	N/A	N/A	N/A
D-Robot†	2.86	1.25	3.11	N/A	N/A	N/A	N/A	N/A
Source-only	4.50	1.62	5.49	0.034	7.62	3.12	11.15	0.019
Chamfer loss	3.96	1.47	4.43	0.036	4.18	1.54	5.27	0.043
cycle-consistency	3.93	1.48	4.36	0.035	6.47	2.43	9.08	0.029
Mean Teacher	2.67	1.33	3.12	**0.028**	6.40	2.42	9.50	**0.013**
UA-MT	2.58	1.28	3.04	0.029	5.71	1.83	8.77	0.015
Ours w/o $\mathcal{L}_{\text{syn}}$	2.49	1.23	2.88	0.030	2.57	1.22	2.93	0.027
Ours w/o $\mathcal{L}'_{\text{con}}$	2.96	1.27	3.29	0.035	3.00	1.21	3.39	0.034
Ours	**2.31**	**1.16**	**2.66**	0.034	**2.38**	**1.12**	**2.66**	0.033

a slight improvement of 16.0% in the more challenging scenario, where pseudo labels by the teacher are naturally noisier, in turn limiting the efficacy of the adaptation process. 3) Our proposed strategy to filter pseudo labels (ours w/o $\mathcal{L}_{\text{syn}}$) improves the standard teacher and its uncertainty-aware extension, particularly notable under the more severe domain shift ($-59.8/-55.0\%$ TRE). 4) Synthesizing novel data pairs with the teacher (ours w/o $\mathcal{L}'_{\text{con}}$) alone is slightly inferior to the standard teacher for realistic deformations in the source domain but substantially superior for simple rigid transformations. 5) Combining our two strategies yields further considerable improvements to TREs of 2.31 and 2.38 mm, demonstrating their complementarity. Thus, our method improves the standard Mean Teacher by 13.5/62.8%, outperforms all competitors by statistically significant margins ($p < 0.001$ in a Wilcoxon signed-rank test), and sets a new state-of-the-art accuracy. Remarkably, our method achieves almost the same accuracy for simple rigid transformations in the source domain as for complex, realistic deformations. Thus, it eliminates the need for designing task-specific deformation models, which requires strong domain knowledge. Qualitative results are presented in Fig. 6.5 and Fig. 6.6, demonstrating accurate and smooth deformation fields by our method, as confirmed by the SDlogJ in Tab. 6.2, which takes small values for all methods.

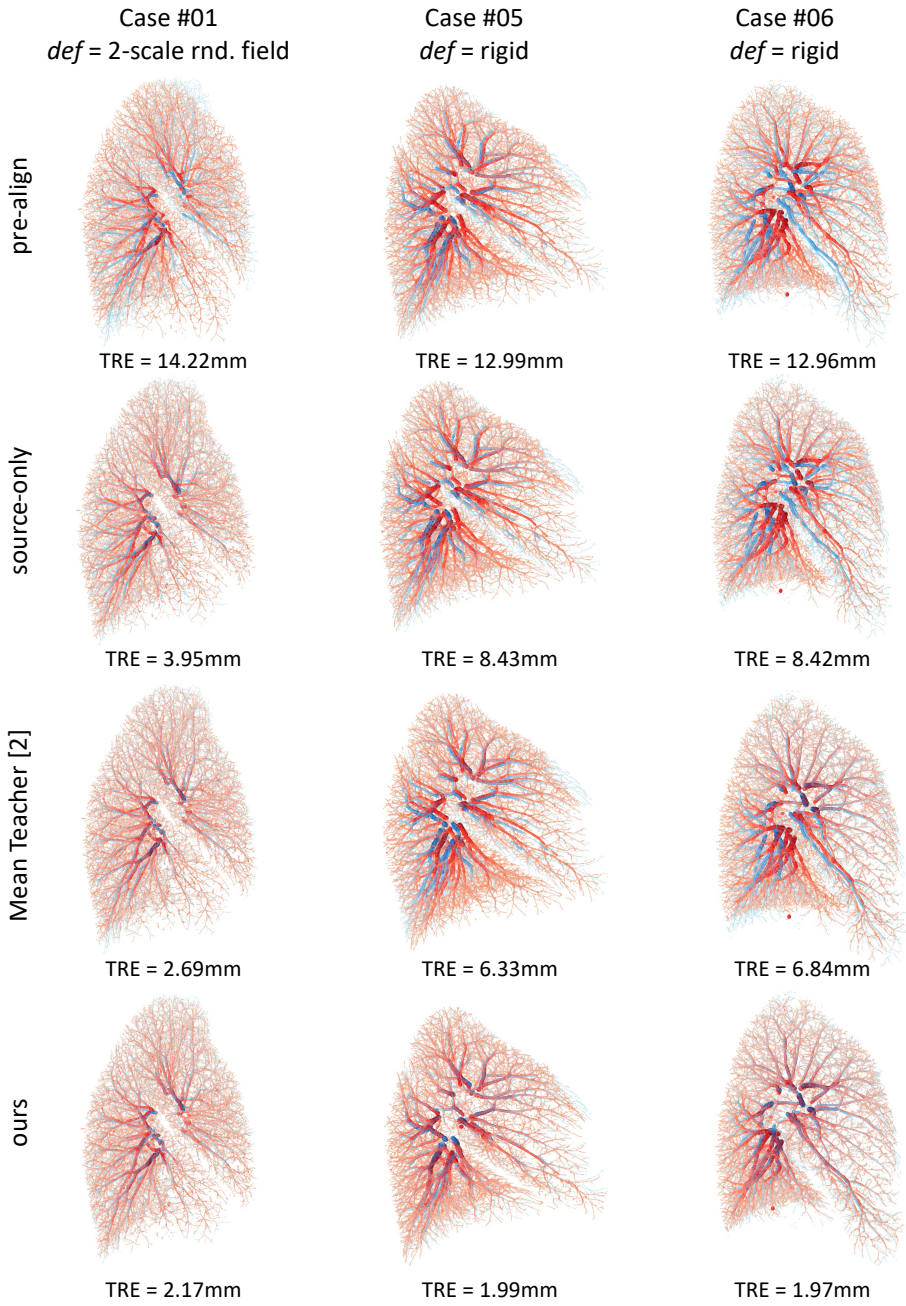


Fig. 6.5: Qualitative results on three cases of the PVT dataset. We show overlays of the original high-resolution point clouds of the fixed exhale and warped inhale structures in blue and red, respectively. To this end, we interpolated predicted displacement fields to the high-resolution inhale clouds with an isotropic Gaussian kernel. From top to bottom, each column shows the pre-aligned point clouds and the registrations by the source-only model, the standard Mean Teacher, and our method.

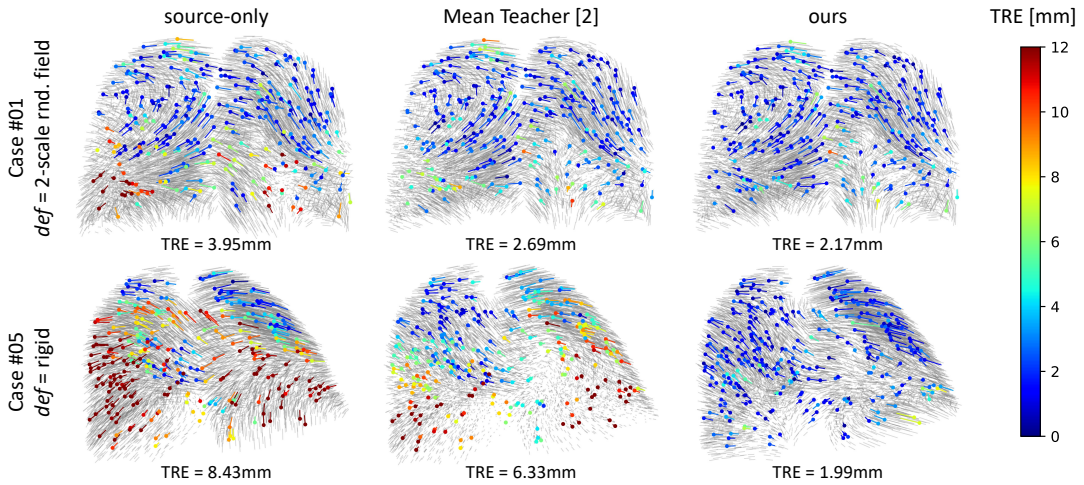


Fig. 6.6: Qualitative results on two sample cases of the PVT dataset. Predicted displacement fields are shown in gray. Colored dots and lines represent moving landmarks and their interpolated flow, with colors encoding the TRE (clamped to 12 mm).

Ablation study. In an additional ablation experiment, we assess different similarity measures between point clouds for filtering pseudo labels. More precisely, we replace the Chamfer distance in Eq. (6.8) by the Gaussian MMD [Feydy, 2020] and the Laplacian curvature [Wu et al., 2020]. We perform the experiments for rigid deformations in the source domain and discard the synthesis of novel training samples, i.e., $\lambda_3 = 0$. Results are presented in Tab. 6.3, showing that all considered similarity measures are suitable for filtering pseudo labels with the Chamfer distance achieving the top performance.

6.2.4 Discussion and Conclusion

Our work addressed domain adaptive point cloud registration to bridge the gap between synthetic source and real target deformations. Starting from the established Mean Teacher paradigm, we presented two novel strategies to tackle the noise of pseudo

Table 6.3: Quantitative results for different similarity measures to filter pseudo labels. As metrics, we report the mean TRE and 25%/75% percentiles in mm. We performed the experiment for rigid deformations in the source domain (def = rigid).

Similarity measure	**TRE**	**25%**	**75%**
None (standard Mean Teacher)	6.40	2.42	9.50
Gaussian MMD	2.72	1.28	3.19
Laplacian curvature	2.86	1.34	3.43
Chamfer distance (ours w/o $\mathcal{L}_{\text{syn}}$)	**2.57**	**1.22**	**2.93**

labels from the teacher model, which is a persistent, significant limitation of the method. Specifically, we 1) proposed to prevent detrimental supervision through the teacher by filtering pseudo labels according to Chamfer distances of student and teacher registrations and 2) introduced a novel teacher-student paradigm, where the teacher synthesizes novel training data pairs with perfect noise-free displacement labels. Our experiments for lung vessel registration on the PVT dataset demonstrated the efficacy of our method under two scenarios, outperforming the standard Mean Teacher by up to 62.8% and setting a new state-of-the-art accuracy (TRE=2.31 mm). As such, our method even favorably compares to popular image-based deep learning methods (VoxelMorph [Balakrishnan et al., 2019] and LapIRN [Mok et al., 2020], e.g., achieve TREs of 7.98 and 4.99 mm on the original DIR-Lab CT images) but lags behind conventional image-based optimization methods [Rühaak et al., 2017] with 0.83 mm TRE. But while the latter require run times of several minutes to process the dense intensity scans with 30M+ voxels, our method processes sparse, purely geometric point clouds with 8k points only, enabling anonymity-preservation and extremely fast inference within 0.2 s. In this light, we see two significant potential impacts of our work: First, our method could generally advance purely geometric keypoint-based medical registration, previously limited by the inefficacy of unsupervised learning with similarity metrics. In particular, medical point cloud registration, currently primarily focusing on lung anatomies, still needs to be investigated for other anatomical structures (abdomen, brain) in future work, which might benefit from our generic approach. Second, our method is conceptually transferable to dense image registration (e.g., intensity-based similarity metrics [Heinrich et al., 2012; Vos et al., 2020] can replace the Chamfer distance). In this context, it appears of great interest to revisit learning from synthetic deformations [Eppenhof et al., 2018] within a DA setting or to combine our method with unsupervised learning under metric supervision.

Chapter 7

Summary and Conclusion

This work developed various geometric deep learning methods for point cloud analysis in the medical context. The presented methods contribute to achieving three overarching goals: 1) opening up novel applications for point cloud-based medical image analysis, 2) explicitly analyzing and tackling distribution shifts between different domains, and 3) founding methodological developments on the incorporation of task-specific prior knowledge.

This concluding chapter provides an overview and discussion of the research findings of the thesis. First, Sec. 7.1 summarizes the scientific contributions of the methodological chapters. Then, Sec. 7.2 elaborates on their relevance to the research questions elaborated at the outset of the thesis. Finally, Sec. 7.3 discusses open problems and limitations of the presented approaches together with recent related work and points out emerging directions for future research.

7.1 Contributions

Opening up in-bed body weight and shape analysis as a novel application of point cloud-based neural networks. Starting with analyzing depth sensor-based point clouds acquired by clinical monitoring systems, Chapter 3 investigated body weight estimation of in-bed patients and, for the first time, demonstrated the suitability of deep learning approaches for the problem. For the case of fully-visible uncovered patients, point clouds were encoded with a basis point set, yielding a compact representation to be efficiently processed by a fully-connected network for weight regression. The expressiveness of the encoding was further enhanced by sampling the basis points according to the expected distribution of the input points, thus capturing finer details of the patient's shape and further improving weight estimates, eventually reaching a clinically satisfying accuracy. As another novel application, the second part of Chapter 3 addressed the weight estimation of partially occluded patients under a blanket. The core idea was to simplify the overall problem by treating weight estimation and occlusion as independent sub-problems to be solved by separate networks one after the other. Hence, a two-stage method was developed, virtually removing the blanket in the first step, thus allowing weight estimation for uncovered patients in the second step. As for the former,

the chapter demonstrated the capability of a 3D U-Net to accurately reconstruct the patient's shape, which translated to more than halving the accuracy gap between the weight estimates of covered and uncovered patients in standard in-domain experiments. Beyond, the experiments also included a cross-domain evaluation for weight estimation with and without a blanket, showing a significant sensitivity to distribution shifts of all compared methods.

To sum up, Chapter 3 demonstrated the capability of geometric deep learning approaches to accurately solve the highly challenging, previously underexplored problems of point cloud-based body weight estimation and body shape reconstruction under the blanket but also revealed the urgent need for methods that can cope with domain shifts.

Guiding domain adaptation with prior knowledge about human anatomy. Sticking to the use case of point cloud-based in-bed patient monitoring, Chapter 4 explored human pose estimation, i.e., the localization of human joints of interest. Motivated by the outcome of the cross-domain experiments in Chapter 3, the task was addressed in a domain adaptation scenario, aiming to adapt a model from a labeled source to a shifted unlabeled target domain. Compared to standard supervised learning, unsupervised domain adaptation promises improved performance across shifted domains without labor-intensive manual annotations. Throughout the chapter, two adaptation strategies were developed, which guide the learning process in the target domain with the aid of prior knowledge about human anatomy. First, network training was formulated as a constrained optimization problem to restrict model predictions to the anatomically plausible pose space, approximated by three constraints on the human skeleton graph (symmetric limb lengths, plausible bone lengths, and plausible joint angles). These hard constraints were relaxed to soft anatomical loss functions, providing supervision for training by penalizing violations of the anatomical constraints. By contrast, the second strategy implicitly drives the network toward anatomically more correct solutions by filtering pseudo labels for self-training under the Mean Teacher paradigm according to their plausibility, as measured with the anatomical loss functions. The two strategies were unified in a flexible, network-agnostic adaptation framework, which features more stable optimization than intricate adversarial methods while mitigating noisy pseudo labels in standard self-training approaches. Experiments under different distribution shifts and adaptation scenarios proved the complementarity of both strategies and their advantages over the prior state of the art.

In summary, Chapter 4 exploited task-specific prior knowledge to derive constraints on the output space, which, in the context of unsupervised domain adaptation, proved both a powerful source of weak supervision and a valuable quality measure for assessing pseudo labels.

Exploiting complementary characteristics of geometric learning architectures to capture multi-scale representations. Unlike single-frame analysis performed in Chapters 3 and 4, Chapter 5 focused on analyzing sequences of point clouds, concretely aiming to recognize dynamic hand gestures. Starting from the observation that discriminative patterns of hand gestures occur at fine-grained local (hand posture) and coarse global (position) scales, early experiments showed that current state-of-the-art single-stream approaches do not sufficiently capture the required multi-scale features. Consequently, the chapter developed a dual-stream model to decouple the learning of local and global spatial representations, which are eventually fused in an LSTM for temporal modeling. Crucially, the designed dual-stream model combines complementary geometric network architectures with different inductive biases to enforce an inherent focus on the required features in each stream. More precisely, a graph CNN with local connectivity was anticipated as the ideal extractor of fine local geometric features, while global movements were captured by combining a basis point set encoding with a fully-connected DenseNet. Extensive ablation experiments confirmed the advantages of these design choices over various alternatives, enabling state-of-the-art performance on two popular public benchmark datasets.

More generally, Chapter 5 has shown that different geometric network architectures, with their heterogeneous inductive biases, have different foci in feature extraction, which should be explicitly considered when designing new models to exploit potential complementary characteristics and achieve optimal performance.

Adjusting and denoising the Mean Teacher for domain adaptive registration. Chapter 6 operated at the intersection of the topics of Chapters 4 and 5 by addressing domain adaptation for the registration of sequential frames. Contrary to the previous chapters, the task was not addressed for depth sensor-based point clouds but for keypoints extracted from exhale and inhale lung CT images, namely Förstner keypoints in the first and lung vessel trees in the second part. The first part introduced a novel self-ensembling strategy for domain adaptive registration by adapting the Mean Teacher paradigm to the registration problem. Of particular interest was the innovative combination of learnable feature extraction with a differentiable optimizer within the Mean Teacher paradigm, which could benefit from the regularizing effect of the optimizer. While this approach virtually closed the domain gap, it required manually annotated correspondences in the source domain and did not explicitly address the persistent noise in the pseudo labels from the teacher model. Therefore, the second part of the chapter avoided manual labels by considering synthetic deformations in the source domain and proposed two novel learning paradigms to denoise the Mean Teacher. The latter include filtering pseudo labels according to the quality of the final registration of both teacher and student predictions and the dynamic synthesis of new training samples with noise-free labels by the teacher. Experiments proved

the effectiveness of both strategies, enabling state-of-the-art performance for point cloud-based lung registration.

In summary, Chapter 6 demonstrated the great potential of Mean Teacher-based self-training for the registration task and contributed two registration-specific strategies to mitigate the adverse effect of persistent noisy pseudo labels.

7.2 Research Findings

This section discusses the relevance of the previously summarized contributions of this thesis to the overarching research objectives: the exploration of deep learning-based point cloud analysis for novel clinical tasks and applications, the investigation of distribution shifts and domain adaptation, and the incorporation of task-specific prior knowledge.

Medical tasks and applications. The point cloud-based deep learning methods developed throughout this thesis address various medical image analysis tasks, thus opening up a wide range of potential clinical applications.

Chapter 3 achieved accurate point cloud-based in-bed body weight estimation, which is a crucial requirement for patient-specific treatments like proper drug dosing or adjusting ventilation parameters. The superiority of the presented deep learning-based methods (average error under 4 kg/6%) over the estimates by clinical staff (average error of 8-11% in clinical studies [Menon et al., 2005]) promises improved quality of such medical treatments. Beyond body weight as the specific regression target, the presented methods might also apply to further related clinically relevant regression tasks, such as estimating body measures and proportions, which can be similarly beneficial to patient-specific treatments.

The methods presented in Chapter 4 can infer the body pose of an in-bed patient from a point cloud, achieving accurate predictions (average error under 9 cm) even under challenging blanket occlusions. Such precise knowledge of the human pose enables diverse clinical downstream tasks, including behavior monitoring, detection of critical events, diagnosis of pathological movements, and position classification for automated care documentation. Moreover, it is valuable for clinical studies, which, for instance, have already proven connections between sleeping poses and sleep apnea [Lee et al., 2015] or carpal tunnel syndrome [McCabe et al., 2010]. Since the developed methods are not specifically tailored to the in-bed use case, they would also apply to the pose estimation of clinical staff, as required for automated workflow documentation and analysis and context-aware assistance systems.

The third studied task was point cloud-based dynamic hand gesture recognition in Chapter 5. The developed method classifies gestures from 28 categories with state-of-the-art accuracies of over 90%. In the clinical context, accurate gesture classification

enables, among others, the recognition and documentation of surgical work steps during an operation and touchless gesture control, e.g., of room lights or electronic devices, by clinical staff, which can significantly simplify work procedures, especially in sterile environments. While hand gesture recognition is rather a narrow field, the findings of the chapter might partially transfer to the classification of general point cloud sequences, which, in turn, finds diverse applications in patient monitoring, such as respiration type recognition or action classification.

Finally, Chapter 6 addressed the inhale-to-exhale registration of Förstner keypoints and lung vessel point clouds extracted from lung CT scans. With target registration errors (TREs) around 2 mm, the developed methods achieve state-of-the-art performance among point-based methods and even competitive accuracy with state-of-the-art intensity-based deep learning approaches. The methods still lag behind the best conventional intensity-based methods [Rühaak et al., 2017] with TREs under 1 mm but offer, in exchange, higher computational efficiency and anonymity preservation. As such, they might become a valuable tool for lung registration in clinical applications, which is essential for diagnosing pulmonary diseases and planning their treatment. In a broader view, accurate point cloud registration also opens up novel opportunities in patient monitoring, e.g., the generation of a complete 3D patient model by registering multiple views of the same patient in different positions.

Altogether, the considered problems exhibit a wide range of heterogeneous characteristics: Inputs are given as individual frames and temporal sequences and represent both depth sensor-based point clouds and keypoints from medical scans, and addressed tasks include global regression (body weight) and classification (hand gesture) and the local detection of landmarks (body pose) and correspondences (registration). The presented geometric deep learning solutions achieved promising results across all these tasks, demonstrating the versatile potential of the research field for diverse medical applications.

Domain shifts and domain adaptation. The contribution of this thesis to the topic of domain adaptation in point cloud analysis is two-fold. First, diverse experiments examined the generalization performance of point cloud-based deep learning methods under previously unexplored geometric domain shifts in clinical settings. Second, the thesis developed multiple domain adaptation methods to tackle such distribution shifts.

As for the former, Chapter 3 assessed the robustness of several in-bed weight estimation models to a changed room setup, including a varied bed, mattress, sheet, blanket, and sensor position. Such distribution shifts occur regularly in real-world applications, which include model deployment in various rooms and hospitals across different cities and countries. However, all considered methods showed severe error increases of 55 to 86%, for instance, from 3.8 to 6.7 kg for the top-performing model.

Chapter 4 examined a similar environmental shift for in-bed human pose estimation, causing an even more severe increase in the prediction error by a factor of five. More precisely, the source-trained baseline model achieved a mean localization error over 12 joints of 24.0 cm in the shifted target domain compared to 4.7 cm of the in-domain model trained on target data (oracle/upper bound). Another experiment evaluated the pose model under a shift from uncovered to covered patients, which is relevant to avoid challenging manual annotations under the blanket. The results revealed a lower but still significant error increase of 91% from 6.8 cm of the oracle to 13.0 cm of the baseline model (averaged over 24 joints).

Finally, Chapter 6 explored keypoint-based exhale-to-inhale lung registration of medical scans under multiple domain shifts, including different breathing types in the source and target domain, varied fields of view due to data acquisition in different clinical centers, and synthetically generated source vs. real-world target data. Again, the source-trained baseline models generalized insufficiently to the shifted target domains, as quantified by TREs of 4.0 to 5.0 mm, which is a 78 to 90% increase of the TREs of the oracles (2.3 to 2.6 mm).

Summarizing the cross-domain generalization experiments, point cloud networks are highly vulnerable to various geometric domain shifts, as demonstrated for multiple clinical applications. In response, this work developed several unsupervised domain adaptation methods to overcome such domain gaps.

Chapter 4 introduced two adaptation strategies for human pose estimation. The first one guides network training in the unlabeled target domain by constraining predictions to the space of plausible poses, implemented with a novel anatomical loss function that penalizes anatomically implausible predictions. The second technique builds upon the well-known Mean Teacher paradigm, where the learning student model is supervised with pseudo labels from a weight-averaged teacher model. The proposed method filters the inherently noisy pseudo labels from the teacher according to their anatomical plausibility, thus stabilizing the adaptation process and preventing detrimental supervision by the teacher. Notably, both methods do not require simultaneous access to source and target data and hence apply to both classical unsupervised and source-free domain adaptation, which is particularly beneficial when dealing with sensitive clinical data. The two strategies proved highly effective in the experiments under two domain shifts (uncover to cover, different environment), overall improving the baseline model by 31/66% (13.0→9.0 cm/24.0→8.2 cm pose error) and reducing the domain gap by 65/82%.

The methodological developments for domain adaptive registration in Chapter 6 all build upon the Mean Teacher paradigm and comprise three main contributions. First, given the general lack of domain adaptation methods for registration, the Mean Teacher was, for the first time, adapted to the registration problem, including two decisive modifications: extending the augmentation scheme by inverse geometric transformations and integrating the combination of learnable feature extraction with differentiable optimization. This approach already surpassed the baseline model by

47/50% (5.0→2.6 mm/4.0→2.0 mm TRE) under two adaptation scenarios (clinical sites, breathing types), but the method remains generally limited by the unavoidable noise of the pseudo-label supervision from the teacher. Therefore, the second part of the chapter developed two strategies to denoise the teacher. First, conceptually similar to the filtering technique in Chapter 4, pseudo registration fields by the teacher were only selected for supervision if more accurate than the corresponding student prediction, as measured by a similarity metric between the fixed and deformed moving input. Second, the displacement fields predicted by the teacher were used to dynamically synthesize novel training data pairs, whose underlying displacements are precisely known by design and provide noise-free supervision to the student. When evaluated under two differently severe synthetic-to-real domain shifts, the unified method improved the standard Mean Teacher by 13/63% and outperformed the baseline model by 49/69%.

To sum up, the thesis presented multiple novel domain adaptation strategies, including a prior-guided constrained optimization scheme and the adaptation, optimization, and extension of the Mean Teacher paradigm to point cloud-based medical imaging tasks. The developed methods proved effective under various domain shifts and for tasks with heterogeneous characteristics, namely the single-frame analysis of sensor-based point clouds for human pose estimation and sequence analysis of image-based keypoints for registration.

Prior domain knowledge. A common feature of all previously discussed methodological developments is their foundation on diverse types of prior knowledge.

The basis point set encoding used for weight estimation in the first part of Chapter 3 was originally designed for encoding arbitrarily shaped and oriented 3D objects. By contrast, point clouds of in-bed patients have roughly consistent shapes and orientations. Adapting the sampling scheme of the basis point set to this expected input distribution could improve the accuracy of weight estimates by 4.5%.

For the more complex task of weight estimation under the cover, addressed in the second part of Chapter 3, end-to-end learning was replaced by a two-stage method following an intuitive human approach to the problem, which first removes the blanket and estimates the weight in a second independent step. This procedure surpassed the corresponding end-to-end approach by 13%.

The developed domain adaptation methods for human pose estimation in Chapter 4 rely on prior knowledge about human anatomy. More specifically, such prior knowledge enables humans to assess the validity of predicted poses by verifying the satisfaction of anatomical constraints on the human skeleton graph. This prior knowledge was translated to an anatomical loss function that quantified the violation of the constraints in a differentiable manner and was, in turn, used to provide direct supervision and quantitatively assess the quality of pseudo labels. The performed experiments demonstrated clear advantages of this explicit incorporation of anatomical priors over

implicitly learning the human anatomy in adversarial output space adaptation (-14% pose error) and uncertainty-based assessment of pseudo labels (-10% pose error).

The model for hand gesture recognition presented in Chapter 5 unifies two distinct priors. First, human reasoning provides the insight that discriminative patterns of hand gestures occur at two scales, motivating the design of a dual-stream model for multi-scale feature extraction. Second, it is well-known that different neural networks have distinct inductive biases and, thus, inherently focus on different patterns. Based on this knowledge, two geometric learning architectures with suitable complementary characteristics were combined in the dual-stream model to extract the desired features in each network stream.

Finally, the denoising strategies for the Mean Teacher in Chapter 6 were strongly inspired by the human understanding of an optimal relationship between students and their teacher. First, good teachers should not insist on their own solutions if their students have an equivalent or even better approach. Second, responsible teachers should not pose problems with unknown solutions but rather construct novel tasks with precisely known solutions. These two principles are reflected in the developed denoising learning strategies, including pseudo-label filtering and the synthesis of new training pairs. In the most challenging experimental setting, both techniques improved the standard Mean Teacher by 60 and 53%, respectively.

Overall, the thesis thus explored a comprehensive set of various priors, including input and output distributions, the well-known inductive biases of existing networks, and human procedures, reasoning, and behavior. In all cases, extensive experiments demonstrated significant benefits over the corresponding end-to-end approaches without explicit task-specific priors.

7.3 Limitations and Outlook

The methodological developments in this thesis advanced deep learning-based solutions to diverse point cloud analysis tasks in medical imaging, as evidenced by extensive experimental validations. Nonetheless, some limitations remain, raising follow-up research questions and pointing to promising research directions for future work.

Clinical validation. All proposed methods achieved excellent results on public benchmark datasets, measured as weight, pose, classification, or registration accuracy/error. While promising, these measures, however, do not directly reflect the value of the methods in clinical practice, which is highly relevant given the intended clinical deployment and needs to be analyzed in future clinical validation studies by answering, for instance, the following open questions: How does the accuracy of weight estimates translate to the quality of weight-based treatments such as drug dosing and ventilation? To what extent does the performance of pose-based downstream tasks

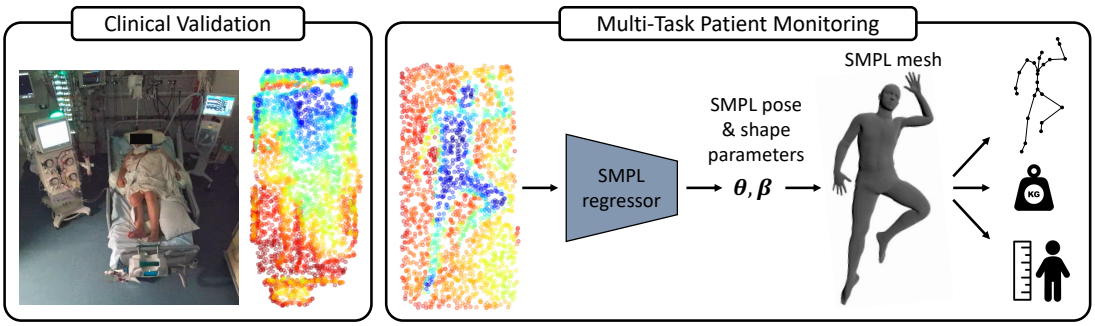


Fig. 7.1: Visualization of ongoing and future work. Left: Visual sample data from the ongoing clinical study at the university hospital Hamburg-Eppendorf, showing the color image of the hospital room and the associated cropped point cloud of the bed. Right: SMPL regression. Regressing the pose and shape parameters of the SMPL model provides a 3D mesh model of the patient from which 3D pose, body weight, and body measures, among others, can be inferred.

(position classification, movement analysis) depend on the accuracy of pose estimates? Is the achieved accuracy in gesture recognition sufficient to facilitate clinical workflows, or do the (few) errors cause practical inconveniences? Which registration accuracy is required for reliable diagnosis and treatment planning? As a first step in this direction, the developed methods for patient monitoring are currently being integrated into a live demonstrator that generates recommendations for ventilation parameters based on weight estimates and will be evaluated on real-world clinical data from an ongoing study (see Fig. 7.1, left).

Patient monitoring with a single multi-task model. The in-bed monitoring tasks addressed in this work (weight and pose estimation) were solved with independent task-specific models. Considering the amount of other relevant monitoring tasks, such as body measure, proportion, and shape analysis, position classification, body part segmentation, etc., further pursuing this task-specific approach would result in uncountable models to be trained separately and deployed in parallel, which is not only computationally inefficient but also neglects potentially beneficial interrelationships between tasks. Instead, future work could explore a multi-task model with a single shared feature extractor and distinct network heads for separate tasks. This approach would significantly reduce the computational burden, and the regularizing effect of parameter sharing could even improve the (cross-domain) generalization performance, as shown in prior works on multi-task learning in different contexts [Carlucci et al., 2019; Wang et al., 2020a; Zhang et al., 2021].

Directly regressing the parameters of a parametric human mesh model, such as the SMPL model [Loper et al., 2015], is an appealing alternative to explicit multi-task learning. In this formulation, one single-task model provides a complete 3D patient

model from which various pose and shape features can be derived (Fig. 7.1, right). The approach was increasingly explored in recent work, e.g., by Clever et al. [2022], but the achieved performance on individual tasks (pose and weight estimation) was inferior to that of the separate single-task models presented in this thesis. A presumable reason is that the pose parameters of the SMPL model represent 3D rotations whose regression is a highly non-linear and challenging task requiring further investigation. Moreover, even though existing methods leveraged data from depth sensors, they eventually operated on gridded 2D depth maps, thus discarding the inherent 3D structure of point clouds that might benefit the task and still needs to be explored in this context in future work.

Incorporating temporal information. The weight and pose estimation models presented in Chapters 3 and 4 operate on individual frames but would likely benefit from processing temporal sequence data, as ubiquitous in monitoring systems. While Chapters 5 and 6 have already investigated point cloud-based sequence analysis, transferring the findings to pose and weight estimation is non-trivial and requires further research.

In weight estimation, accumulating multiple views of the same patient in different positions would provide increasingly comprehensive and detailed information about the patient's shape. An end-to-end model could learn to fuse the views in a unified representation, or registration methods could explicitly construct a complete unified point cloud of the patient. As for the latter, the methods from Chapter 6 would be a reasonable starting point.

Pose estimation could benefit from both short- and long-range temporal information. As for the latter, considering monitoring over several hours or days, patient-specific bone lengths derived from high-confident pose estimates (for basic positions without a blanket) constitute a valuable prior for all following estimates. Promising implementations include, for instance, conditioning a model on the specific anatomy of a patient and refining model predictions through anatomy-guided fitting in a post-processing step. On a shorter timescale, jointly processing successive video frames, e.g., with spatio-temporal convolutions [Fan et al., 2021], can capture fine-grained motions and provide more detailed cues about the patient's pose, which could improve and stabilize pose estimates.

From domain adaptation to domain generalization. All cross-domain experiments in this work demonstrated poor model generalization under geometric domain shifts, which was successfully addressed by developing effective domain adaptation methods. But while unsupervised domain adaptation avoids costly manual annotations, it involves several practical issues, particularly prominent in the medical context. First, standard domain adaptation methods require simultaneous access to source and target data, which is often infeasible due to the protection of sensitive data. Concretely, the

provider (e.g., a company) and end-user (e.g., a hospital) of a deep learning model might often be unable to share their data, prohibiting standard domain adaptation. Source-free domain adaptation as performed in Chapter 4 overcomes this problem but, as the second issue, still requires computationally demanding model adaptation to each novel domain, which is impractical or even infeasible due to limited computational resources and technical knowledge of the end-user.

Instead, the provider would ideally develop a robust model that generalizes to arbitrary domains, as is the goal of domain generalization [Zhou et al., 2022a]. Beyond the previously discussed multi-task approach, elaborated augmentation techniques were recently shown to be an effective strategy in image-based segmentation [Ouyang et al., 2022; Su et al., 2023], but the designed transformations obviously do not apply to point clouds. Huang et al. [2021] embedded point cloud-specific augmentations in a meta-learning framework, but this approach primarily addresses the generalization from synthetic to real 3D objects, which significantly differs from the heterogeneous domain shifts in clinical practice. Thus, domain generalization for point cloud analysis in medicine remains a highly relevant but unexplored research field to be investigated in future work.

From lung registration to arbitrary anatomies through anatomy-agnostic keypoint extraction from medical scans. The point cloud registration methods from Chapter 6 achieved accurate results for lung registration, but their suitability for other anatomies – even though highly desirable – has not been investigated yet. In this context, it is worth noting that the used registration models and developed adaptation methods are not specifically tailored to lung registration and should, therefore, generalize well to other anatomies. Meanwhile, the supposedly critical step lies in extracting distinctive keypoints from the original medical scans, realized with a lung vessel filter and the Förstner operator in this work. While the former is highly anatomy-specific, the latter theoretically applies to other anatomies, but the geometric registration of Förstner keypoints outside the lung is not sufficiently investigated yet. Prior work already demonstrated accurate prostate and abdominal registration based on surface point clouds of organs [Baum et al., 2021; Joutard et al., 2022], which, however, requires a performant (anatomy-specific) segmentation algorithm, again limiting flexibility. Instead, an anatomy-agnostic keypoint detector would enable flexible deployment of the proposed registration methods across diverse anatomies. While keypoints based on a Sobel filter or Canny edge detector could be a reasonable starting point, its concrete implementation needs further investigation.

From point cloud analysis to a universal modality-agnostic framework. While all methods developed throughout this thesis were designed for point cloud analysis, it is worth highlighting that some of the methodological concepts also apply to analyzing

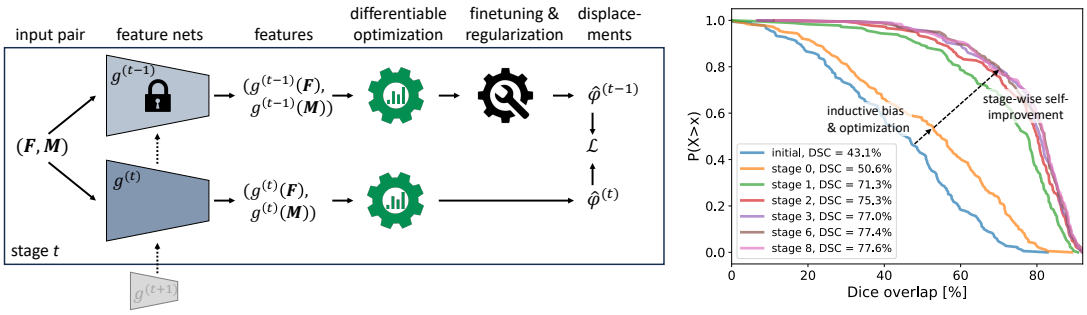


Fig. 7.2: Overview of the modality-agnostic framework for unsupervised registration presented in [Bigalke et al., 2023b]. Left: Schematic visualization of the cyclical self-training scheme. The modular registration model comprises a deep network for feature extraction and a differentiable optimizer to predict displacements, which – depending on the input modality – are implemented as a graph CNN and loopy belief propagation for point clouds (see Sec. 6.1) and as a CNN and coupled convex optimization for images [Siebert et al., 2022]. At stage t, the learning network $g^{(t)}$ is supervised with finetuned and regularized pseudo labels generated based on the features from the frozen network $g^{(t-1)}$ from the previous stage. Right: Results for image-based abdomen CT registration shown as the "opposite" cumulative distribution of Dice overlaps after different stages of self-training.

gridded data. In particular, the proposed learning strategies for domain adaptation (constrained optimization, denoised self-training) are easily adaptable to dense image-based settings and may advance domain adaptive 3D pose estimation and registration in this context. Going one step further, our recent work published at MICCAI 2023 [Bigalke et al., 2023b] explicitly developed a universal modality-agnostic registration framework (Fig. 7.2). More precisely, the work proposed optimization-guided cyclical self-training as a novel learning paradigm for unsupervised registration, implemented in a flexible, modular framework that achieves state-of-the-art performance for both image-based and point cloud-based registration.

References

[Abavisani et al., 2019] Abavisani, M., Joze, H. R. V., and Patel, V. M. "Improving the Performance of Unimodal Dynamic Hand-Gesture Recognition With Multimodal Training". In: *Conference on Computer Vision and Pattern Recognition – CVPR 2019*. 2019, pp. 1165–1174.

[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: *International Conference on Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016: Proceedings, Part I*. 2016, pp. 491–499.

[Achituve et al., 2021] Achituve, I., Maron, H., and Chechik, G. "Self-Supervised Learning for Domain Adaptation on Point Clouds". In: *Winter Conference on Applications of Computer Vision – WACV 2021*. 2021, pp. 123–133.

[Adiga Vasudeva et al., 2022] Adiga Vasudeva, S., Dolz, J., and Lombaert, H. "Leveraging Labeling Representations in Uncertainty-Based Semi-Supervised Segmentation". In: *International Conference on Medical Image Computing and Computer Assisted Intervention – MICCAI 2022: Proceedings, Part VIII*. 2022, pp. 265–275.

[Afham et al., 2022] Afham, M., Haputhanthri, U., Pradeepkumar, J., Anandakumar, M., De Silva, A., and Edussooriya, C. U. "Towards Accurate Cross-Domain In-Bed Human Pose Estimation". In: *International Conference on Acoustics, Speech and Signal Processing – ICASSP 2022*. 2022, pp. 2664–2668.

[Akhter et al., 2015] Akhter, I. and Black, M. J. "Pose-Conditioned Joint Angle Limits for 3D Human Pose Reconstruction". In: *Conference on Computer Vision and Pattern Recognition – CVPR 2015*. 2015, pp. 1446–1455.

[Alliegro et al., 2021] Alliegro, A., Boscaini, D., and Tommasi, T. "Joint Supervised and Self-Supervised Learning for 3D Real World Challenges". In: *International Conference on Pattern Recognition – ICPR 2020*. 2021, pp. 6718–6725.

[Altinigne et al., 2020] Altinigne, C. Y., Thanou, D., and Achanta, R. "Height and Weight Estimation From Unconstrained Images". In: *International Conference on Acoustics, Speech and Signal Processing – ICASSP 2020*. 2020, pp. 2298–2302.

[Armeni et al., 2016] Armeni, I., Sener, O., Zamir, A. R., Jiang, H., Brilakis, I., Fischer, M., and Savarese, S. "3D Semantic Parsing of Large-Scale Indoor Spaces". In:

Conference on Computer Vision and Pattern Recognition – CVPR 2016. 2016, pp. 1534–1543.

[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.

[Bateson et al., 2020] Bateson, M., Kervadec, H., Dolz, J., Lombaert, H., and Ben Ayed, I. "Source-Relaxed Domain Adaptation for Image Segmentation". In: *International Conference on Medical Image Computing and Computer Assisted Intervention – MICCAI 2020: Proceedings, Part I*. 2020, pp. 490–499.

[Bateson et al., 2021] Bateson, M., Dolz, J., Kervadec, H., Lombaert, H., and Ben Ayed, I. "Constrained Domain Adaptation for Image Segmentation". *IEEE Transactions on Medical Imaging* 40 (7), 2021, pp. 1875–1887.

[Bateson et al., 2022] Bateson, M., Lombaert, H., and Ben Ayed, I. "Test-Time Adaptation with Shape Moments for Image Segmentation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2022, pp. 736–745.

[Baum et al., 2021] Baum, Z. M., Hu, Y., and Barratt, D. C. "Real-Time Multimodal Image Registration With Partial Intraoperative Point-Set Data". *Medical Image Analysis* 74, 2021, p. 102231.

[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.

[Benalcazar et al., 2017] Benalcazar, D., Benalcazar, D., and Erazo, A. "Artificial Neural Networks and Digital Image Processing: An Approach for Indirect Weight Measurement". In: *Ecuador Technical Chapters Meeting – ETCM 2017*. 2017, pp. 1–6.

[Bertsekas, 1997] Bertsekas, D. P. "Nonlinear Programming". *Journal of the Operational Research Society* 48 (3), 1997, pp. 334–334.

[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 – BVM 2021*. 2021, pp. 254–259.

[Bigalke et al., 2021c] Bigalke, A. and Heinrich, M. P. “Fusing Posture and Position Representations for Point Cloud-Based Hand Gesture Recognition”. In: *International Conference on 3D Vision – 3DV 2021*. 2021, pp. 617–626.

[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: *International Conference on Medical Image Computing and Computer Assisted Intervention – MICCAI 2022: Proceedings, Part VI*. 2022, pp. 280–290.

[Bigalke et al., 2023a] Bigalke, A., Hansen, L., Diesel, J., Hennigs, C., Rostalski, P., and Heinrich, M. P. “Anatomy-Guided Domain Adaptation for 3D In-Bed Human Pose Estimation”. *Medical Image Analysis*, 2023, p. 102887.

[Bigalke et al., 2023b] Bigalke, A., Hansen, L., Mok, T. C., and Heinrich, M. P. “Unsupervised 3D Registration Through Optimization-Guided Cyclical Self-Training”. In: *International Conference on Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*. 2023, to appear.

[Bigalke et al., 2023c] Bigalke, A. and Heinrich, M. P. “A Denoised Mean Teacher for Domain Adaptive Point Cloud Registration”. In: *International Conference on Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*. 2023, to appear.

[Bousmalis et al., 2016] Bousmalis, K., Trigeorgis, G., Silberman, N., Krishnan, D., and Erhan, D. “Domain Separation Networks”. *Advances in Neural Information Processing Systems – NeurIPS 2016* 29, 2016, pp. 343–351.

[Bousmalis et al., 2017] Bousmalis, K., Silberman, N., Dohan, D., Erhan, D., and Krishnan, D. “Unsupervised Pixel-Level Domain Adaptation With Generative Adversarial Networks”. In: *Conference on Computer Vision and Pattern Recognition – CVPR 2017*. 2017, pp. 3722–3731.

[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.

[Bronstein et al., 2021] Bronstein, M. M., Bruna, J., Cohen, T., and Veličković, P. “Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges”. *arXiv preprint arXiv:2104.13478*, 2021.

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

[Buckley et al., 2012] Buckley, R. G., Stehman, C. R., Dos Santos, F. L., Riffenburgh, R. H., Swenson, A., Mjos, N., Brewer, M., and Mulligan, S. "Bedside Method to Estimate Actual Body Weight in the Emergency Department". *The Journal of Emergency Medicine* 42 (1), 2012, pp. 100–104.

[Cai et al., 2019] Cai, Q., Pan, Y., Ngo, C.-W., Tian, X., Duan, L., and Yao, T. "Exploring Object Relation in Mean Teacher for Cross-Domain Detection". In: *Conference on Computer Vision and Pattern Recognition – CVPR 2019*. 2019, pp. 11457–11466.

[Cao et al., 2019] Cao, J., Tang, H., Fang, H.-S., Shen, X., Lu, C., and Tai, Y.-W. "Cross-Domain Adaptation for Animal Pose Estimation". In: *International Conference on Computer Vision – ICCV 2019*. 2019, pp. 9498–9507.

[Cao et al., 2020] Cao, X. and Zhao, X. "Anatomy and Geometry Constrained One-Stage Framework for 3D Human Pose Estimation". In: *Asian Conference on Computer Vision – ACCV 2020*. 2020.

[Carlucci et al., 2019] Carlucci, F. M., D'Innocente, A., Bucci, S., Caputo, B., and Tommasi, T. "Domain Generalization by Solving Jigsaw Puzzles". In: *Conference on Computer Vision and Pattern Recognition – CVPR 2019*. 2019, pp. 2229–2238.

[Cardace et al., 2021] Cardace, A., Spezialetti, R., Ramirez, P. Z., Salti, S., and Di Stefano, L. "RefRec: Pseudo-Labels Refinement via Shape Reconstruction for Unsupervised 3D Domain Adaptation". In: *International Conference on 3D Vision – 3DV 2021*. 2021, pp. 331–341.

[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.

[Casas et al., 2019] Casas, L., Navab, N., and Demirci, S. "Patient 3D Body Pose Estimation From Pressure Imaging". *International Journal of Computer Assisted Radiology and Surgery* 14 (3), 2019, pp. 517–524.

[Cattermole et al., 2017] Cattermole, G. N., Graham, C. A., and Rainer, T. H. "Mid-Arm Circumference Can Be Used to Estimate Weight of Adult and Adolescent Patients". *Emergency Medicine Journal* 34 (4), 2017, pp. 231–236.

[Chai et al., 2016] Chai, X., Liu, Z., Yin, F., Liu, Z., and Chen, X. "Two Streams Recurrent Neural Networks for Large-Scale Continuous Gesture Recognition". In: *International Conference on Pattern Recognition – ICPR 2016*. 2016, pp. 31–36.

[Chang et al., 2019] Chang, W.-G., You, T., Seo, S., Kwak, S., and Han, B. "Domain-Specific Batch Normalization for Unsupervised Domain Adaptation". In: *Conference on Computer Vision and Pattern Recognition – CVPR 2019*. 2019, pp. 7354–7362.

[Chen et al., 2017] Chen, X., Guo, H., Wang, G., and Zhang, L. "Motion Feature Augmented Recurrent Neural Network for Skeleton-Based Dynamic Hand Gesture Recognition". In: *International Conference on Image Processing – ICIP 2017*. 2017, pp. 2881–2885.

[Chen et al., 2018] 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". *IEEE Journal of Translational Engineering in Health and Medicine* 6, 2018, pp. 1–11.

[Chen et al., 2019] Chen, Y., Zhao, L., Peng, X., Yuan, J., and Metaxas, D. N. "Construct Dynamic Graphs for Hand Gesture Recognition via Spatial-Temporal Attention". In: *British Machine Vision Conference – BMVC 2019*. 2019, p. 103.

[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.

[Chen et al., 2021] Chen, C., Liu, Q., Jin, Y., Dou, Q., and Heng, P.-A. "Source-Free Domain Adaptive Fundus Image Segmentation With Denoised Pseudo-Labeling". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention – MICCAI 2021: Proceedings, Part V*. 2021, pp. 225–235.

[Chen et al., 2022a] Chen, J., Frey, E. C., He, Y., Segars, W. P., Li, Y., and Du, Y. "Transmorph: Transformer for Unsupervised Medical Image Registration". *Medical Image Analysis* 82, 2022, p. 102615.

[Chen et al., 2022b] Chen, X., Wang, X., Zhang, K., Fung, K.-M., Thai, T. C., Moore, K., Mannel, R. S., Liu, H., Zheng, B., and Qiu, Y. "Recent Advances and Clinical Applications of Deep Learning in Medical Image Analysis". *Medical Image Analysis*, 2022, p. 102444.

[Chi et al., 2022] Chi, Z., Wang, S., Li, X., Chang, C.-T., Islam, M., Holkar, A., Pronger, S., Liu, T., Lam, K.-M., and He, X. "Multi-Level Unsupervised Domain Adaption for Privacy-Protected In-Bed Pose Estimation". In: *International Workshop on Advanced Imaging Technology – IWAIT 2022*. Vol. 12177. 2022, pp. 431–436.

[Çiçek et al., 2016] Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., and Ronneberger, O. "3D U-Net: Learning Dense Volumetric Segmentation From Sparse Annotation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016: Proceedings, Part II*. 2016, pp. 424–432.

[Clever et al., 2020] Clever, H. M., Erickson, Z., Kapusta, A., Turk, G., Liu, K., and Kemp, C. C. "Bodies at Rest: 3D Human Pose and Shape Estimation From a Pressure Image Using Synthetic Data". In: *Conference on Computer Vision and Pattern Recognition – CVPR 2020*. 2020, pp. 6215–6224.

[Clever et al., 2022] Clever, H. M., Grady, P., Turk, G., and Kemp, C. C. "BodyPressure-Inferring Body Pose and Contact Pressure from a Depth Image". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[Cordts et al., 2016] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., and Schiele, B. "The cityscapes dataset for semantic urban scene understanding". In: *Conference on Computer Vision and Pattern Recognition – CVPR 2016*. 2016, pp. 3213–3223.

[Cunha et al., 2016] Cunha, J. P. S., Choupina, H. M. P., Rocha, A. P., Fernandes, J. M., Achilles, F., Loesch, A. M., Vollmar, C., Hartl, E., and Noachtar, S. "NeuroKinect: A Novel Low-Cost 3DVideo-EEG System for Epileptic Seizure Motion Quantification". *PloS one* 11 (1), 2016, e0145669.

[Davoodnia et al., 2021] Davoodnia, V., Ghorbani, S., and Etemad, A. "In-Bed Pressure-Based Pose Estimation Using Image Space Representation Learning". In: *International Conference on Acoustics, Speech and Signal Processing – ICASSP 2021*. 2021, pp. 3965–3969.

[De Smedt et al., 2016] De Smedt, Q., Wannous, H., and Vandeborre, J.-P. "Skeleton-Based Dynamic Hand Gesture Recognition". In: *Conference on Computer Vision and Pattern Recognition Workshops – CVPR-W 2016*. 2016, pp. 1–9.

[De Smedt et al., 2017] De Smedt, Q., Wannous, H., Vandeborre, J.-P., Guerry, J., Le Saux, B., and Filliat, D. "Shrec'17 Track: 3D Hand Gesture Recognition Using a Depth and Skeletal Dataset". In: *Eurographics Workshop on 3D Object Retrieval – 3DOR 2017*. 2017, pp. 1–6.

[De Vos et al., 2019] De 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.

[Defferrard et al., 2016] Defferrard, M., Bresson, X., and Vandergheynst, P. "Convolutional Neural Networks on Graphs With Fast Localized Spectral Filtering". In: *Advances in Neural Information Processing Systems – NeurIPS 2016*. Vol. 29. 2016.

[Deng et al., 2021] Deng, J., Li, W., Chen, Y., and Duan, L. "Unbiased Mean Teacher for Cross-Domain Object Detection". In: *Conference on Computer Vision and Pattern Recognition – CVPR 2021*. 2021, pp. 4091–4101.

[Dhingra et al., 2019] Dhingra, N. and Kunz, A. “Res3ATN-Deep 3D Residual Attention Network for Hand Gesture Recognition in Videos”. In: *International Conference on 3D Vision – 3DV 2019*. 2019, pp. 491–501.

[Doersch et al., 2015] Doersch, C., Gupta, A., and Efros, A. A. “Unsupervised Visual Representation Learning by Context Prediction”. In: *International Conference on Computer Vision – ICCV 2015*. 2015, pp. 1422–1430.

[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.

[Ester et al., 1996] Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases With Noise.” In: *International Conference on Knowledge Discovery and Data Mining – KDD 1996)*. Vol. 96. 34. 1996, pp. 226–231.

[Fan et al., 2021] Fan, H., Yu, X., Ding, Y., Yang, Y., and Kankanhalli, M. “PSTNet: Point Spatio-Temporal Convolution on Point Cloud Sequences”. In: *International Conference on Learning Representations – ICLR 2021*. 2021.

[Fan et al., 2022] Fan, H., Chang, X., Zhang, W., Cheng, Y., Sun, Y., and Kankanhalli, M. “Self-Supervised Global-Local Structure Modeling for Point Cloud Domain Adaptation With Reliable Voted Pseudo Labels”. In: *Conference on Computer Vision and Pattern Recognition – CVPR 2022*. 2022, pp. 6377–6386.

[Fernandes et al., 1999] Fernandes, C., Clark, S., Price, A., and Innes, G. “How Accurately Do We Wstimate Patients’ Weight in Emergency Departments?” *Canadian Family Physician* 45, 1999, p. 2373.

[Feydy, 2020] Feydy, J. “Geometric Data Analysis, Beyond Convolutions”. PhD thesis. Université Paris-Saclay Gif-sur-Yvette, France, 2020.

[Fischler et al., 1981] Fischler, M. A. and Bolles, R. C. “Random Sample Consensus: A Paradigm for Model Fitting With Applications to Image Analysis and Automated Cartography”. *Communications of the ACM* 24 (6), 1981, pp. 381–395.

[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: *ISPRS Intercommission Conference on Fast Processing of Photogrammetric Data*. Vol. 6. 1987, pp. 281–305.

[Frangi et al., 1998] Frangi, A. F., Niessen, W. J., Vincken, K. L., and Viergever, M. A. “Multiscale Vessel Enhancement Filtering”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention -— MICCAI 1998: Proceedings 1*. 1998, pp. 130–137.

[French et al., 2018] French, G., Mackiewicz, M., and Fisher, M. "Self-Ensembling for Visual Domain Adaptation". In: *International Conference on Learning Representations – ICLR 2018*. 2018.

[Freeman et al., 1995] Freeman, W. T. and Roth, M. "Orientation Histograms for Hand Gesture Recognition". In: *International Workshop on Automatic Face and Gesture Recognition*. Vol. 12. 1995, pp. 296–301.

[Fu et al., 2020] Fu, Y., Lei, Y., Wang, T., Curran, W. J., Liu, T., and Yang, X. "Deep Learning in Medical Image Registration: A Review". *Physics in Medicine & Biology* 65 (20), 2020, 20TR01.

[Gammulle et al., 2017] Gammulle, H., Denman, S., Sridharan, S., and Fookes, C. "Two Stream LSTM: A Deep Fusion Framework for Human Action Recognition". In: *Winter Conference on Applications of Computer Vision – WACV 2017*. 2017, pp. 177–186.

[Ganin et al., 2015] Ganin, Y. and Lempitsky, V. "Unsupervised Domain Adaptation by Backpropagation". In: *International Conference on Machine Learning – ICML 2015*. 2015, pp. 1180–1189.

[Ge et al., 2018a] Ge, L., Cai, Y., Weng, J., and Yuan, J. "Hand Pointnet: 3D Hand Pose Estimation Using Point Sets". In: *Conference on Computer Vision and Pattern Recognition – CVPR2018*. 2018, pp. 8417–8426.

[Ge et al., 2018b] Ge, L., Ren, Z., and Yuan, J. "Point-to-Point Regression Pointnet for 3D Hand Pose Estimation". In: *European Conference on Computer Vision – ECCV 2018*. 2018, pp. 475–491.

[Ghifary et al., 2016] Ghifary, M., Kleijn, W. B., Zhang, M., Balduzzi, D., and Li, W. "Deep Reconstruction-Classification Networks for Unsupervised Domain Adaptation". In: *European Conference on Computer Vision – ECCV 2016*. 2016, pp. 597–613.

[Goodfellow et al., 2014] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. "Generative Adversarial Networks". *Advances in Neural Information Processing Systems – NeurIPS 2014* 27, 2014.

[Goodfellow et al., 2016] Goodfellow, I., Bengio, Y., and Courville, A. *Deep Learning*. `http://www.deeplearningbook.org`. MIT Press, 2016.

[Graves et al., 2013] Graves, A., Mohamed, A.-r., and Hinton, G. "Speech Recognition With Deep Recurrent Neural Networks". In: *International Conference on Acoustics, Speech and Signal Processing – ICASSP 2013*. 2013, pp. 6645–6649.

[Gretton et al., 2006] Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., and Smola, A. "A Kernel Method for the Two-Sample-Problem". *Advances in Neural Information Processing Systems – NeurIPS 2006* 19, 2006.

[Guan et al., 2021] Guan, H. and Liu, M. “Domain Adaptation for Medical Image Analysis: A Survey”. *IEEE Transactions on Biomedical Engineering*, 2021.

[Guo et al., 2020] Guo, Y., Wang, H., Hu, Q., Liu, H., Liu, L., and Bennamoun, M. “Deep Learning for 3D Point Clouds: A Survey”. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43 (12), 2020, pp. 4338–4364.

[Guo et al., 2021] Guo, F., He, Z., Zhang, S., Zhao, X., Fang, J., and Tan, J. “Normalized Edge Convolutional Networks for Skeleton-based Hand Gesture Recognition”. *Pattern Recognition*, 2021, p. 108044.

[Hansen et al., 2019] 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., 2021] Hansen, L. and Heinrich, M. P. “Deep Learning Based Geometric Registration for Medical Images: How Accurate Can We Get Without Visual Features?” In: *International Conference on Information Processing in Medical Imaging – IPMI 2021*. 2021, pp. 18–30.

[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.

[Haskins et al., 2020] Haskins, G., Kruger, U., and Yan, P. “Deep Learning in Medical Image Registration: A Survey”. *Machine Vision and Applications* 31 (1), 2020, pp. 1–18.

[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.

[Hegde et al., 2021] Hegde, D., Sindagi, V., Kilic, V., Cooper, A. B., Foster, M., and Patel, V. “Uncertainty-Aware Mean Teacher for Source-Free Unsupervised Domain Adaptive 3D Object Detection”. *arXiv preprint arXiv:2109.14651*, 2021.

[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., 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: Proceedings, Part II*. 2015, pp. 338–345.

[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: *Workshop on Biomedical Image Registration – WBIR 2022*. 2022, pp. 85–95.

[Hendrycks et al., 2019] Hendrycks, D. and Dietterich, T. "Benchmarking Neural Network Robustness to Common Corruptions and Perturbations". In: *International Conference on Learning Representations – ICLR 2019*. 2019.

[Hering et al., 2020] Hering, A., Murphy, K., and Ginneken, B. *Learn2Reg Challenge: CT Lung Registration - Training Data*. 2020.

[Hering et al., 2022] Hering, A., Hansen, L., Mok, T. C., Chung, A. C., Siebert, H., Häger, S., Lange, A., Kuckertz, S., Heldmann, S., Shao, W., et al. "Learn2Reg: Comprehensive Multi-Task Medical Image Registration Challenge, Dataset and Evaluation in the Era of Deep Learning". *IEEE Transactions on Medical Imaging*, 2022.

[Hermes et al., 2022] Hermes, N., Hansen, L., 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 – BVM 2022*. 2022, pp. 89–94.

[Hochreiter et al., 1997] Hochreiter, S. and Schmidhuber, J. "Long Short-Term Memory". *Neural Computation* 9 (8), 1997, pp. 1735–1780.

[Hoffman et al., 2018] Hoffman, J., Tzeng, E., Park, T., Zhu, J.-Y., Isola, P., Saenko, K., Efros, A., and Darrell, T. "Cycada: Cycle-Consistent Adversarial Domain Adaptation". In: *International Conference on Machine Learning – ICML 2018*. 2018, pp. 1989–1998.

[Hornik et al., 1989] Hornik, K., Stinchcombe, M., and White, H. "Multilayer Feedforward Networks Are Universal Approximators". *Neural Networks* 2 (5), 1989, pp. 359–366.

[Hou et al., 2018] Hou, J., Wang, G., Chen, X., Xue, J.-H., Zhu, R., and Yang, H. "Spatial-Temporal Attention Res-TCN for Skeleton-Based Dynamic Hand Gesture Recognition". In: *European Conference on Computer Vision Workshops – ECCV-W 2018*. 2018, pp. 273–286.

[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., 2021] Huang, C., Cao, Z., Wang, Y., Wang, J., and Long, M. "Metasets: Meta-Learning on Point Sets for Generalizable Representations". In: *Conference on Computer Vision and Pattern Recognition – CVPR 2021*. 2021, pp. 8863–8872.

[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 – ICML 2015*. 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.

[Jähne-Raden et al., 2019] Jähne-Raden, N., Kulau, U., Marschollek, M., and Wolf, K.-H. “INBED: A Highly Specialized System for Bed-Exit-Detection and Fall Prevention on a Geriatric Ward”. *Sensors* 19 (5), 2019, p. 1017.

[Jiang et al., 2019] Jiang, M. and Guo, G. “Body Weight Analysis From Human Body Images”. *IEEE Transactions on Information Forensics and Security* 14 (10), 2019, pp. 2676–2688.

[Jiang et al., 2020] Jiang, M., Shang, Y., and Guo, G. “Computational Approach to Body Mass Index Estimation From Dressed People in 3D Space”. *IET Image Processing* 14 (7), 2020, pp. 1248–1256.

[Jin et al., 2022] Jin, Z., Lei, Y., Akhtar, N., Li, H., and Hayat, M. “Deformation and Correspondence Aware Unsupervised Synthetic-to-Real Scene Flow Estimation for Point Clouds”. In: *Conference on Computer Vision and Pattern Recognition – CVPR 2022*. 2022, pp. 7233–7243.

[Joutard et al., 2022] Joutard, S., Pheiffer, T., Audigier, C., Wohlfahrt, P., Dorent, R., Piat, S., Vercauteren, T., Modat, M., and Mansi, T. “A Multi-Organ Point Cloud Registration Algorithm for Abdominal CT Registration”. In: *Workshop on Biomedical Image Registration – WBIR 2022*. 2022, pp. 75–84.

[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.

[Karanam et al., 2020] Karanam, S., Li, R., Yang, F., Hu, W., Chen, T., and Wu, Z. “Towards Contactless Patient Positioning”. *IEEE Transactions on Medical Imaging* 39 (8), 2020, pp. 2701–2710.

[Ke et al., 2019] Ke, Z., Wang, D., Yan, Q., Ren, J., and Lau, R. W. “Dual Student: Breaking the Limits of the Teacher in Semi-Supervised Learning”. In: *International Conference on Computer Vision – ICCV 2019*. 2019, pp. 6728–6736.

[Kervadec et al., 2019] Kervadec, H., Dolz, J., Tang, M., Granger, E., Boykov, Y., and Ayed, I. B. “Constrained-CNN Losses for Weakly Supervised Segmentation”. *Medical Image Analysis* 54, 2019, pp. 88–99.

[Kervadec et al., 2021] Kervadec, H., Bahig, H., Letourneau-Guillon, L., Dolz, J., and Ayed, I. B. "Beyond Pixel-Wise Supervision: Semantic Segmentation With Higher-Order Shape Descriptors". In: *Medical Imaging with Deep Learning – MIDL 2021*. 2021.

[Kim et al., 2022] Kim, D., Wang, K., Saenko, K., Betke, M., and Sclaroff, S. "A Unified Framework for Domain Adaptive Pose Estimation". In: *European Conference on Computer Vision – ECCV 2022*. 2022, pp. 603–620.

[Kingma et al., 2015] Kingma, D. P. and Ba, J. "Adam: A Method for Stochastic Optimization". In: *International Conference on Learning Representations – ICLR 2015*. 2015.

[Kingkan et al., 2018] Kingkan, C., Owoyemi, J., and Hashimoto, K. "Point Attention Network for Gesture Recognition Using Point Cloud Data". In: *British Machine Vision Conference – BMVC 2018*. 2018, p. 118.

[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.

[Klingner et al., 2022] Klingner, M., Termöhlen, J.-A., Ritterbach, J., and Fingscheidt, T. "Unsupervised Batchnorm Adaptation (UBNA): A Domain Adaptation Method for Semantic Segmentation Without Using Source Domain Representations". In: *Winter Conference on Applications of Computer Vision – WACV 2022*. 2022, pp. 210–220.

[Krizhevsky et al., 2012] Krizhevsky, A., Sutskever, I., and Hinton, G. E. "ImageNet Classification With Deep Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems – NeurIPS 2019*. Vol. 25. 2012, pp. 1106–1114.

[Kruse et al., 2021] 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 – BVM 2021*. 2021, pp. 192–197.

[Kundu et al., 2020] Kundu, J. N., Venkat, N., and Babu, R. V. "Universal Source-Free Domain Adaptation". In: *Conference on Computer Vision and Pattern Recognition – CVPR 2020*. 2020, pp. 4544–4553.

[Kundu et al., 2021] Kundu, J. N., Kulkarni, A., Singh, A., Jampani, V., and Babu, R. V. "Generalize Then Adapt: Source-Free Domain Adaptive Semantic Segmentation". In: *International Conference on Computer Vision – ICCV 2021*. 2021, pp. 7046–7056.

[Kurmi et al., 2021] Kurmi, V. K., Subramanian, V. K., and Namboodiri, V. P. "Domain Impression: A Source Data Free Domain Adaptation Method". In:

Winter Conference on Applications of Computer Vision – WACV 2021. 2021, pp. 615–625.

[Labati et al., 2012] Labati, R. D., Genovese, A., Piuri, V., and Scotti, F. "Weight Estimation From Frame Sequences Using Computational Intelligence Techniques". In: *International Conference on Computational Intelligence for Measurement Systems and Applications – CIMSA 2012*. 2012, pp. 29–34.

[Lee et al., 2015] Lee, C. H., Kim, D. K., Kim, S. Y., Rhee, C.-S., and Won, T.-B. "Changes in Site of Obstruction in Obstructive Sleep Apnea Patients According to Sleep Position: A DISE Study". *The Laryngoscope* 125 (1), 2015, pp. 248–254.

[Li et al., 2018] Li, Y., Bu, R., Sun, M., Wu, W., Di, X., and Chen, B. "PointCNN: Convolution on X-transformed points". *Advances in Neural Information Processing Systems – NeurIPS 2018* 31, 2018, pp. 820–830.

[Li et al., 2019a] Li, R., Liu, Z., and Tan, J. "A Survey on 3D Hand Pose Estimation: Cameras, Methods, and Datasets". *Pattern Recognition* 93, 2019, pp. 251–272.

[Li et al., 2019b] Li, S. and Lee, D. "Point-to-Pose Voting Based Hand Pose Estimation Using Residual Permutation Equivariant Layer". In: *Conference on Computer Vision and Pattern Recognition - CVPR 2019*. 2019, pp. 11927–11936.

[Li et al., 2019c] Li, Y., He, Z., Ye, X., He, Z., and Han, K. "Spatial Temporal Graph Convolutional Networks for Skeleton-Based Dynamic Hand Gesture Recognition". *EURASIP Journal on Image and Video Processing* 2019 (1), 2019, pp. 1–7.

[Li et al., 2019d] Li, Y., Yuan, L., and Vasconcelos, N. "Bidirectional Learning for Domain Adaptation of Semantic Segmentation". In: *Conference on Computer Vision and Pattern Recognition – CVPR 2019*. 2019, pp. 6936–6945.

[Li et al., 2020] Li, K., Wang, S., Yu, L., and Heng, P.-A. "Dual-Teacher: Integrating Intra-Domain and Inter-Domain Teachers for Annotation-Efficient Cardiac Segmentation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention – MICCAI 2020: Proceedings, Part I*. 2020, pp. 418–427.

[Li et al., 2021] Li, C. and Lee, G. H. "From Synthetic to Real: Unsupervised Domain Adaptation for Animal Pose Estimation". In: *Conference on Computer Vision and Pattern Recognition – CVPR 2021*. 2021, pp. 1482–1491.

[Liang et al., 2020] Liang, J., Hu, D., and Feng, J. "Do We Really Need to Access the Source Data? Source Hypothesis Transfer for Unsupervised Domain Adaptation". In: *International Conference on Machine Learning – ICML 2020*. 2020, pp. 6028–6039.

[Lin et al., 2018] Lin, C., Wan, J., Liang, Y., and Li, S. Z. "Large-Scale Isolated Gesture Recognition Using a Refined Fused Model Based on Mmasked Res-C3D

Network and Skeleton LSTM". In: *International Conference on Automatic Face & Gesture Recognition – FG 2018*. 2018, pp. 52–58.

[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., 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: Proceedings, Part I*. 2019, pp. 236–245.

[Liu et al., 2019b] Liu, Y., Fan, B., Xiang, S., and Pan, C. "Relation-Shape Convolutional Neural Network for Point Cloud Analysis". In: *Conference on Computer Vision and Pattern Recognition – CVPR 2019*. 2019, pp. 8895–8904.

[Liu et al., 2020a] Liu, J., Liu, Y., Wang, Y., Prinet, V., Xiang, S., and Pan, C. "Decoupled Representation Learning for Skeleton-Based Gesture Recognition". In: *Conference on Computer Vision and Pattern Recognition – CVPR 2020*. 2020, pp. 5751–5760.

[Liu et al., 2020b] Liu, J., Wang, Y., Liu, Y., Xiang, S., and Pan, C. "3D PostureNet: A Unified Framework for Skeleton-Based Posture Recognition". *Pattern Recognition Letters* 140, 2020, pp. 143–149.

[Liu et al., 2021a] Liu, X., Xing, F., Yang, C., El Fakhri, G., and Woo, J. "Adapting Off-the-Shelf Source Segmenter for Target Medical Image Segmentation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention – MICCAI 2021: Proceedings, Part II*. 2021, pp. 549–559.

[Liu et al., 2021b] Liu, Y., Zhang, W., and Wang, J. "Source-Free Domain Adaptation for Semantic Segmentation". In: *Conference on Computer Vision and Pattern Recognition – CVPR 2021*. 2021, pp. 1215–1224.

[Liu et al., 2022a] Liu, S., Huang, X., Fu, N., Li, C., Su, Z., and Ostadabbas, S. "Simultaneously-Collected Multimodal Lying Pose Dataset: Enabling In-Bed Human Pose Monitoring". *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45 (1), 2022, pp. 1106–1118.

[Liu et al., 2022b] Liu, S., Huang, X., Marcenaro, L., and Ostadabbas, S. "Privacy-Preserving In-Bed Human Pose Estimation: Highlights from the IEEE Video and Image Processing Cup 2021 Student Competition [SP Competitions]". *IEEE Signal Processing Magazine* 39 (3), 2022, pp. 121–129.

[Liu et al., 2022c] Liu, S., Sehgal, N., and Ostadabbas, S. "Adapted Human Pose: Monocular 3D Human Pose Estimation With Zero Real 3D Pose Data". *Applied Intelligence*, 2022, pp. 1–16.

[Loper et al., 2015] Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., and Black, M. J. "SMPL: A Skinned Multi-Person Linear Model". *ACM transactions on graphics (TOG)* 34 (6), 2015, pp. 1–16.

[Lorenz et al., 2007] Lorenz, M. W., Graf, M., Henke, C., Hermans, M., Ziemann, U., Sitzer, M., and Foerch, C. "Anthropometric Approximation of Body Weight in Unresponsive Stroke Patients". *Journal of Neurology, Neurosurgery & Psychiatry* 78 (12), 2007, pp. 1331–1336.

[Luo et al., 2019] Luo, Y., Zheng, L., Guan, T., Yu, J., and Yang, Y. "Taking a Closer Look at Domain Shift: Category-Level Adversaries for Semantics Consistent Domain Adaptation". In: *Conference on Computer Vision and Pattern Recognition – CVPR 2019*. 2019, pp. 2507–2516.

[Maghoumi et al., 2019] Maghoumi, M. and LaViola, J. J. "DeepGRU: Deep Gesture Recognition Utility". In: *International Symposium on Visual Computing*. 2019, pp. 16–31.

[Mahapatra et al., 2020] Mahapatra, D. and Ge, Z. "Training Data Independent Image Registration Using Generative Adversarial Networks and Domain Adaptation". *Pattern Recognition* 100, 2020, p. 107109.

[Martínez-González et al., 2018] Martínez-González, A., Villamizar, M., Canévet, O., and Odobez, J.-M. "Investigating Depth Domain Adaptation for Efficient Human Pose Estimation". In: *European Conference on Computer Vision Workshops – ECCV-W 2018*. 2018, pp. 0–0.

[Masci et al., 2015] Masci, J., Boscaini, D., Bronstein, M., and Vandergheynst, P. "Geodesic Convolutional Neural Networks on Riemannian Manifolds". In: *International Conference on Computer Vision Workshops – ICCV-W 2015*. 2015, pp. 37–45.

[Mascagni et al., 2021] Mascagni, P. and Padoy, N. "OR Black Box and Surgical Control Tower: Recording and Streaming Data and Analytics to Improve Surgical Care". *Journal of Visceral Surgery* 158 (3), 2021, S18–S25.

[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.

[McCabe et al., 2010] McCabe, S. J. and Xue, Y. "Evaluation of Sleep Position as a Potential Cause of Carpal Tunnel Syndrome: Preferred Sleep Position on the Side is Associated with Age and Gender". *Hand* 5 (4), 2010, pp. 361–363.

[Menon et al., 2005] Menon, S. and Kelly, A.-M. "How Accurate is Weight Estimation in the Emergency Department?" *Emergency Medicine Australasia* 17 (2), 2005, pp. 113–116.

[Miao et al., 2017] Miao, Q., Li, Y., Ouyang, W., Ma, Z., Xu, X., Shi, W., and Cao, X. "Multimodal Gesture Recognition Based on the ResC3D Network". In: *International Conference on Computer Vision Workshops – ICCV-W 2017*. 2017, pp. 3047–3055.

[Min et al., 2019] Min, Y., Chai, X., Zhao, L., and Chen, X. "FlickerNet: Adaptive 3D Gesture Recognition From Sparse Point Clouds". In: *British Machine Vision Conference – BMVC 2019*. 2019, p. 105.

[Min et al., 2020] Min, Y., Zhang, Y., Chai, X., and Chen, X. "An Efficient PointLSTM for Point Clouds Based Gesture Recognition". In: *Conference on Computer Vision and Pattern Recognition – CVPR 2020*. 2020, pp. 5761–5770.

[Mittal et al., 2020] Mittal, H., Okorn, B., and Held, D. "Just Go With the Flow: Self-Supervised Scene Flow Estimation". In: *Conference on Computer Vision and Pattern Recognition – CVPR 2020*. 2020, pp. 11177–11185.

[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 2020: Proceedings, Part III*. 2020, pp. 211–221.

[Mok et al., 2021] Mok, T. C. and Chung, A. "Conditional Deformable Image Registration With Convolutional Neural Network". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention – MICCAI 2021: Proceedings, Part IV*. 2021, pp. 35–45.

[Molchanov et al., 2016] Molchanov, P., Yang, X., Gupta, S., Kim, K., Tyree, S., and Kautz, J. "Online Detection and Classification of Dynamic Hand Gestures With Recurrent 3D Convolutional Neural Network". In: *Conference on Computer Vision and Pattern Recognition – CVPR 2016*. 2016, pp. 4207–4215.

[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.

[Moon et al., 2018] Moon, G., Chang, J. Y., and Lee, K. M. "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.

[Mu et al., 2020] Mu, J., Qiu, W., Hager, G. D., and Yuille, A. L. "Learning From Synthetic Animals". In: *Conference on Computer Vision and Pattern Recognition – CVPR 2020*. 2020, pp. 12386–12395.

[Murez et al., 2018] Murez, Z., Kolouri, S., Kriegman, D., Ramamoorthi, R., and Kim, K. "Image to Image Translation for Domain Adaptation". In: *Conference on Computer Vision and Pattern Recognition – CVPR 2018*. 2018, pp. 4500–4509.

[Nahavandi et al., 2017] Nahavandi, D., Abobakr, A., Haggag, H., Hossny, M., Nahavandi, S., and Filippidis, D. "A Skeleton-Free Kinect System for Body Mass Index Assessment Using Deep Neural Networks". In: *International Systems Engineering Symposium – ISSE 2017*. 2017, pp. 1–6.

[Narayana et al., 2018] Narayana, P., Beveridge, R., and Draper, B. A. "Gesture Recognition: Focus on the Hands". In: *Conference on Computer Vision and Pattern Recognition – CVPR 2018*. 2018, pp. 5235–5244.

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

[Nguyen et al., 2014] Nguyen, T. V., Feng, J., and Yan, S. "Seeing Human Weight From a Single RGB-D Image". *Journal of Computer Science and Technology* 29 (5), 2014, pp. 777–784.

[Nguyen et al., 2019] Nguyen, X. S., Brun, L., Lézoray, O., and Bougleux, S. "A Neural Network Based on SPD Manifold Learning for Skeleton-Based Hand Gesture Recognition". In: *Conference on Computer Vision and Pattern Recognition – CVPR 2019*. 2019, pp. 12036–12045.

[Nunez et al., 2018] Nunez, J. C., Cabido, R., Pantrigo, J. J., Montemayor, A. S., and Velez, J. F. "Convolutional Neural Networks and Long Short-Term Memory for Skeleton-Based Human Activity and Hand Gesture Recognition". *Pattern Recognition* 76, 2018, pp. 80–94.

[Ostadabbas et al., 2012] Ostadabbas, S., Yousefi, R., Nourani, M., Faezipour, M., Tamil, L., and Pompeo, M. Q. "A Resource-Efficient Planning for Pressure Ulcer Prevention". *IEEE Transactions on Information Technology in Biomedicine* 16 (6), 2012, pp. 1265–1273.

[Ouyang et al., 2022] Ouyang, C., Chen, C., Li, S., Li, Z., Qin, C., Bai, W., and Rueckert, D. "Causality-Inspired Single-Source Domain Generalization for Medical Image Segmentation". *IEEE Transactions on Medical Imaging* 42 (4), 2022, pp. 1095–1106.

[Padoy, 2019] Padoy, N. "Machine and Deep Learning for Workflow Recognition During Surgery". *Minimally Invasive Therapy & Allied Technologies* 28 (2), 2019, pp. 82–90.

[Pan et al., 2010] Pan, S. J. and Yang, Q. "A Survey on Transfer Learning". *IEEE Transactions on Knowledge and Data Engineering* 22 (10), 2010, pp. 1345–1359.

[Park et al., 2020] Park, T., Efros, A. A., Zhang, R., and Zhu, J.-Y. "Contrastive Learning for Unpaired Image-to-Image Translation". In: *European Conference on Computer Vision – ECCV 2020*. 2020, pp. 319–345.

[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 2019*. 2019, pp. 8024–8035.

[Perone et al., 2019] Perone, C. S., Ballester, P., Barros, R. C., and Cohen-Adad, J. "Unsupervised Domain Adaptation for Medical Imaging Segmentation With Self-Ensembling". *NeuroImage* 194, 2019, pp. 1–11.

[Pfitzner et al., 2015] Pfitzner, C., May, S., Merkl, C., Breuer, L., Köhrmann, M., Braun, J., Dirauf, F., and Nüchter, A. "Libra3d: Body Weight Estimation for Emergency Patients in Clinical Environments With a 3D Structured Light Sensor". In: *International Conference on Robotics and Automation – ICRA 2015*. 2015, pp. 2888–2893.

[Pfitzner et al., 2016] Pfitzner, C., May, S., and Nüchter, A. "Neural Network-Based Visual Body Weight Estimation for Drug Dosage Finding". In: *Medical Imaging 2016: Image Processing*. Vol. 9784. 2016, pp. 524–532.

[Pfitzner et al., 2017] Pfitzner, C., May, S., and Nüchter, A. "Evaluation of Features from RGB-D Data for Human Body Weight Estimation". *IFAC-PapersOnLine* 50 (1), 2017, pp. 10148–10153.

[Pirker et al., 2009] Pirker, K., Rüther, M., Bischof, H., Skrabal, F., and Pichler, G. "Human Body Volume Estimation in a Clinical Environment". *AAPR/OAGM: Challenges in the Biosciences: Image Analysis and Pattern Recognition Aspects*, 2009.

[Prokudin et al., 2019] Prokudin, S., Lassner, C., and Romero, J. "Efficient Learning on Point Clouds With Basis Point Sets". In: *International Conference on Computer Vision – ICCV 2019*. 2019, pp. 4332–4341.

[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 – CVPR 2017*. 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". *Advances in Neural Information Processing Systems – NeurIPS 2017* 30, 2017.

[Qin et al., 2019] Qin, C., You, H., Wang, L., Kuo, C.-C. J., and Fu, Y. "PointDAN: A Multi-Scale 3D Domain Adaption Network for Point Cloud Representation". *Advances in Neural Information Processing Systems – NeurIPS 2019* 32, 2019.

[Rajpurkar et al., 2018] Rajpurkar, P., Irvin, J., Ball, R. L., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C. P., et al. "Deep Learning for

Chest Radiograph Diagnosis: A Retrospective Comparison of the CheXNeXt Algorithm to Practicing Radiologists". *PLoS Medicine* 15 (11), 2018, e1002686.

[Rastgoo et al., 2020] Rastgoo, R., Kiani, K., and Escalera, S. "Video-Based Isolated Hand Sign Language Recognition Using a Deep Cascaded Model". *Multimedia Tools and Applications* 79, 2020, pp. 22965–22987.

[Ren et al., 2022] Ren, J., Pan, L., and Liu, Z. "Benchmarking and Analyzing Point Cloud Classification Under Corruptions". In: *International Conference on Machine Learning – ICML 2022*. 2022, pp. 18559–18575.

[Richter et al., 2016] Richter, S. R., Vineet, V., Roth, S., and Koltun, V. "Playing for Data: Ground Truth From Computer ames". In: *European Conference on Computer Vision – ECCV 2016*. 2016, pp. 102–118.

[Rodrigues et al., 2022] Rodrigues, V. F., Antunes, R. S., Seewald, L. A., Bazo, R., Reis, E. S., Santos, U. J., Righi, R. d. R., Junior, L. G. d. S., Costa, C. A., Bertollo, F. L., Maier, A., Eskofier, B., Horz, T., Pfister, M., and Fahrig, R. "A Multi-Sensor Architecture Combining Human Pose Estimation and Real-Time Location Systems for Workflow Monitoring on Hybrid Operating Suites". *Future Generation Computer Systems*, 2022.

[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: Proceedings, Part III*. 2015, pp. 234–241.

[Rozantsev et al., 2018] Rozantsev, A., Salzmann, M., and Fua, P. "Beyond Sharing Weights for Deep Domain Adaptation". *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41 (4), 2018, pp. 801–814.

[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.

[Saito et al., 2018] Saito, K., Watanabe, K., Ushiku, Y., and Harada, T. "Maximum Classifier Discrepancy for Unsupervised Domain Adaptation". In: *Conference on Computer Vision and Pattern Recognition – CVPR 2018*. 2018, pp. 3723–3732.

[Saito et al., 2019] Saito, K., Ushiku, Y., Harada, T., and Saenko, K. "Strong-Weak Distribution Alignment for Adaptive Object Detection". In: *Conference on Computer Vision and Pattern Recognition – CVPR 2019*. 2019, pp. 6956–6965.

[Salami et al., 2020] Salami, D., Palipana, S., Kodali, M., and Sigg, S. "Motion Pattern Recognition in 4D Point Clouds". In: *International Workshop on Machine Learning for Signal Processing – MLSP 2020*. 2020, pp. 1–6.

[Shen et al., 2021] Shen, Z., Feydy, J., Liu, P., Curiale, A. H., 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, pp. 5373–5389.

[Shen et al., 2022] Shen, Y., Yang, Y., Yan, M., Wang, H., Zheng, Y., and Guibas, L. J. "Domain Adaptation on Point Clouds via Geometry-Aware Implicits". In: *Conference on Computer Vision and Pattern Recognition – CVPR 2022*. 2022, pp. 7223–7232.

[Shi et al., 2020] Shi, L., Zhang, Y., Cheng, J., and Lu, H. "Decoupled Spatial-Temporal Attention Network for Skeleton-Based Action-Gesture Recognition". In: *Asian Conference on Computer Vision – ACCV 2020*. 2020.

[Shrivastava et al., 2017] Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W., and Webb, R. "Learning From Simulated and Unsupervised Images Through Adversarial Training". In: *Conference on Computer Vision and Pattern Recognition – CVPR 2017*. 2017, pp. 2107–2116.

[Siebert et al., 2021] 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*. 2021.

[Siebert et al., 2022] Siebert, H. and Heinrich, M. P. "Learn to Fuse Input Features for Large-Deformation Registration With Differentiable Convex-Discrete Optimisation". In: *Workshop on Biomedical Image Registration – WBIR 2022*. 2022, pp. 119–123.

[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.

[Singh et al., 2017] Singh, V., Ma, K., Tamersoy, B., Chang, Y.-J., Wimmer, A., O'Donnell, T., and Chen, T. "DARWIN: Deformable Patient Avatar Representation With Deep Image Network". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention – MICCAI 2017: Proceedings, Part II*. 2017, pp. 497–504.

[Song et al., 2016] Song, S. and Xiao, J. "Deep Sliding Shapes for Amodal 3D Object Detection in RGB-D Images". In: *Conference on Computer Vision and Pattern Recognition – CVPR 2016*. 2016, pp. 808–816.

[Song et al., 2021] Song, L., Yu, G., Yuan, J., and Liu, Z. "Human Pose Estimation and its Application to Action Recognition: A Survey". *Journal of Visual Communication and Image Representation* 76, 2021, p. 103055.

[Sotiras et al., 2013] Sotiras, A., Davatzikos, C., and Paragios, N. "Deformable Medical Image Registration: A Survey". *IEEE Transactions on Medical Imaging* 32 (7), 2013, pp. 1153–1190.

[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., Kadkhodamohammadi, A., 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.

[Srivastav et al., 2019] Srivastav, V., Gangi, A., and Padoy, N. "Human Pose Estimation on Privacy-Preserving Low-Resolution Depth Images". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention – MICCAI 2019*. 2019, pp. 583–591.

[Srivastav et al., 2022] Srivastav, V., Gangi, A., and Padoy, N. "Unsupervised Domain Adaptation for Clinician Pose Estimation and Instance Segmentation in the Operating Room". *Medical Image Analysis* 80, 2022, p. 102525.

[Su et al., 2023] Su, Z., Yao, K., Yang, X., Wang, Q., Sun, J., and Huang, K. "Rethinking Data Augmentation for Single-source Domain Generalization in Medical Image Segmentation". In: *Conference on Artificial Intelligence – AAAI 2023*. 2023.

[Sun et al., 2016] Sun, B., Feng, J., and Saenko, K. "Return of Frustratingly Easy Domain Adaptation". In: *Conference on Artificial Intelligence – AAAI 2016*. Vol. 30. 1. 2016.

[Sun et al., 2017] Sun, X., Shang, J., Liang, S., and Wei, Y. "Compositional Human Pose Regression". In: *International Conference on Computer Vision – ICCV 2017*. 2017, pp. 2602–2611.

[Sun et al., 2019a] Sun, K., Xiao, B., Liu, D., and Wang, J. "Deep High-Resolution Representation Learning for Human Pose Estimation". In: *Conference on Computer Vision and Pattern Recognition – CVPR 2019*. 2019, pp. 5693–5703.

[Sun et al., 2019b] Sun, Y., Tzeng, E., Darrell, T., and Efros, A. A. "Unsupervised Domain Adaptation Through Self-Supervision". *arXiv preprint arXiv:1909.11825*, 2019.

[Supranata et al., 2018] Supranata, T. H., Davin, P. S. S., Jeremy, D. K., Pratiwi, A. E., and Wulandari, M. "Body Weight Measurement Using Image Processing Based on Body Surface Area and Elliptical Tube Volume". In: *International Conference on Information Technology and Electrical Engineering – ICITEE 2018*. 2018, pp. 290–294.

[Tarvainen et al., 2017] Tarvainen, A. and Valpola, H. "Mean Teachers Are Better Role Models: Weight-Averaged Consistency Targets Improve Semi-Supervised Deep Learning Results". *Advances in Neural Information Processing Systems – NeurIPS 2017* 30, 2017, pp. 1195–1204.

[Taylor et al., 2019] Taylor, S., Brown, J. M., Gupta, K., Campbell, J. P., Ostmo, S., Chan, R. P., Dy, J., Erdogmus, D., Ioannidis, S., Kim, S. J., et al. "Monitoring Disease Progression With a Quantitative Severity Scale for Retinopathy of Prematurity Using Deep Learning". *JAMA Ophthalmology* 137 (9), 2019, pp. 1022–1028.

[Tsai et al., 2018] Tsai, Y.-H., Hung, W.-C., Schulter, S., Sohn, K., Yang, M.-H., and Chandraker, M. "Learning to Adapt Structured Output Space for Semantic Segmentation". In: *Conference on Computer Vision and Pattern Recognition – CVPR 2018*. 2018, pp. 7472–7481.

[Tzeng et al., 2014] Tzeng, E., Hoffman, J., Zhang, N., Saenko, K., and Darrell, T. "Deep Domain Confusion: Maximizing for Domain Invariance". *arXiv preprint arXiv:1412.3474*, 2014.

[Tzeng et al., 2017] Tzeng, E., Hoffman, J., Saenko, K., and Darrell, T. "Adversarial Discriminative Domain Adaptation". In: *Conference on Computer Vision and Pattern Recognition – CVPR 2017*. 2017, pp. 7167–7176.

[Uzunova et al., 2017] Uzunova, H., Wilms, M., Handels, H., and Ehrhardt, J. "Training CNNs for Image Registration From Few Samples With Model-Based Data Augmentation". In: *International Conference on Medical Image Computing and Computer Assisted Intervention – MICCAI 2017: Proceedings, Part I*. 2017, pp. 223–231.

[Velardo et al., 2012] Velardo, C. and Dugelay, J.-L. "What Can Computer Vision Tell You About Your Weight?" In: *European Signal Processing Conference – EUSIPCO 2012*. 2012, pp. 1980–1984.

[Vos et al., 2020] Vos, B. D., Velden, B. H., Sander, J., Gilhuijs, K. G., Staring, M., and Išgum, I. "Mutual Information for Unsupervised Deep Learning Image Registration". In: *Medical Imaging 2020: Image Processing*. Vol. 11313. 2020, pp. 155–161.

[Wang et al., 2016] Wang, H., Chai, X., Hong, X., Zhao, G., and Chen, X. "Isolated Sign Language Recognition With Grassmann Covariance Matrices". *ACM Transactions on Accessible Computing* 8 (4), 2016, pp. 1–21.

[Wang et al., 2017] Wang, H., Wang, P., Song, Z., and Li, W. "Large-Scale Multimodal Gesture Recognition Using Heterogeneous Networks". In: *International Conference on Computer Vision Workshops – ICCV-W 2017*. 2017, pp. 3129–3137.

[Wang et al., 2018] Wang, M. and Deng, W. “Deep Visual Domain Adaptation: A Survey”. *Neurocomputing* 312, 2018, pp. 135–153.

[Wang et al., 2019] 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, pp. 1–12.

[Wang et al., 2020a] Wang, S., Yu, L., Li, C., Fu, C.-W., and Heng, P.-A. “Learning From Extrinsic and Intrinsic Supervisions for Domain Generalization”. In: *European Conference on Computer Vision – ECCV 2020*. 2020, pp. 159–176.

[Wang et al., 2020b] Wang, Y., Zhang, Y., Tian, J., Zhong, C., Shi, Z., Zhang, Y., and He, Z. “Double-Uncertainty Weighted Method for Semi-Supervised Learning”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention – MICCAI 2020: Proceedings, Part I*. 2020, pp. 542–551.

[Wang et al., 2021a] Wang, D., Shelhamer, E., Liu, S., Olshausen, B., and Darrell, T. “Tent: Fully Test-Time Adaptation by Entropy Minimization”. In: *International Conference on Learning Representations – ICLR 2021*. 2021.

[Wang et al., 2021b] Wang, J., Jin, S., Liu, W., Liu, W., Qian, C., and Luo, P. “When Human Pose Estimation Meets Robustness: Adversarial Algorithms and Benchmarks”. In: *Conference on Computer Vision and Pattern Recognition – CVPR 2021*. 2021, pp. 11855–11864.

[Wang et al., 2021c] Wang, K., Zhan, B., Zu, C., Wu, X., Zhou, J., Zhou, L., and Wang, Y. “Tripled-Uncertainty Guided Mean Teacher Model for Semi-Supervised Medical Image Segmentation”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention – MICCAI 2021: Proceedings, Part II*. 2021, pp. 450–460.

[Wang et al., 2022] Wang, Q., Fink, O., Van Gool, L., and Dai, D. “Continual Test-Time Domain Adaptation”. In: *Conference on Computer Vision and Pattern Recognition – CVPR 2022*. 2022, pp. 7201–7211.

[Wei et al., 2023] Wei, M., Wei, Z., Zhou, H., Hu, F., Si, H., Chen, Z., Zhu, Z., Qiu, J., Yan, X., Guo, Y., et al. “AGConv: Adaptive Graph Convolution on 3D Point Clouds”. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[Wong et al., 2020] Wong, J., Fong, A., McVicar, N., Smith, S., Giambattista, J., Wells, D., Kolbeck, C., Giambattista, J., Gondara, L., and Alexander, A. “Comparing Deep Learning-Based Auto-Segmentation of Organs at Risk and Clinical Target Volumes to Expert Inter-Observer Variability in Radiotherapy Planning”. *Radiotherapy and Oncology* 144, 2020, pp. 152–158.

[Wu et al., 2015] Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., and Xiao, J. “3D Shapenets: A Deep Representation for Volumetric Shapes”. In:

Conference on Computer Vision and Pattern Recognition – CVPR 2015. 2015, pp. 1912–1920.

[Wu et al., 2018] Wu, Y. and He, K. "Group Normalization". In: *European Conference on Computer Vision – ECCV 2018*. 2018, pp. 3–19.

[Wu et al., 2019] Wu, W., Qi, Z., and Fuxin, L. "PointConv: Deep Convolutional Networks on 3D Point Clouds". In: *Conference on Computer Vision and Pattern Recognition – CVPR 2019*. 2019, pp. 9621–9630.

[Wu et al., 2020] Wu, W., Wang, Z. Y., Li, Z., Liu, W., and Fuxin, L. "PointPWC-Net: Cost Volume on Point Clouds for (Self-)Supervised Scene Flow Estimation". In: *European Conference on Computer Vision – ECCV 2020: Proceedings, Part V*. 2020, pp. 88–107.

[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.

[Xiang et al., 2021] Xiang, T., Zhang, C., Song, Y., Yu, J., and Cai, W. "Walk in the Cloud: Learning Curves for Point Clouds Shape Analysis". In: *International Conference on Computer Vision – ICCV 2021*. 2021, pp. 915–924.

[Xu et al., 2021] Xu, M., Ding, R., Zhao, H., and Qi, X. "PAConv: Position Adaptive Convolution with Dynamic Kernel Assembling on Point Clouds". In: *Conference on Computer Vision and Pattern Recognition – CVPR 2021*. 2021, pp. 3173–3182.

[Xu et al., 2022] Xu, Z., Luo, J., Lu, D., Yan, J., Frisken, S., Jagadeesan, J., Wells III, W. M., Li, X., Zheng, Y., and Tong, R. K.-y. "Double-Uncertainty Guided Spatial and Temporal Consistency Regularization Weighting for Learning-Based Abdominal Registration". In: *International Conference on Medical Image Computing and Computer Assisted Intervention – MICCAI 2022: Proceedings, Part VI*. 2022, pp. 14–24.

[Yan et al., 2018] Yan, S., Xiong, Y., and Lin, D. "Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition". In: *Conference on Artificial Intelligence – AAAI 2018*. 2018, pp. 7444–7452.

[Yang et al., 2018] Yang, W., Ouyang, W., Wang, X., Ren, J., Li, H., and Wang, X. "3D Human Pose Estimation in the Wild by Adversarial Learning". In: *Conference on Computer Vision and Pattern Recognition – CVPR 2018*. 2018, pp. 5255–5264.

[Yang et al., 2020] Yang, F., Li, R., Georgakis, G., Karanam, S., Chen, T., Ling, H., and Wu, Z. "Robust Multi-Modal 3D Patient Body Modeling". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention – MICCAI 2020: Proceedings, Part III*. 2020, pp. 86–95.

[Yang et al., 2021] Yang, J., Shi, S., Wang, Z., Li, H., and Qi, X. "ST3D: Self-Training for Unsupervised Domain Adaptation on 3D Object Detection". In: *Conference on Computer Vision and Pattern Recognition – CVPR 2021*. 2021, pp. 10368–10378.

[Yin et al., 2022] Yin, Y., Robinson, J. P., and Fu, Y. "Multimodal In-Bed Pose and Shape Estimation Under the Blankets". In: *International Conference on Multimedia*. 2022, pp. 2411–2419.

[Yu et al., 2019] Yu, L., Wang, S., Li, X., Fu, C.-W., and Heng, P.-A. "Uncertainty-Aware Self-Ensembling Model for Semi-Supervised 3D Left Atrium Segmentation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention – MICCAI 2019: Proceedings, Part II*. 2019, pp. 605–613.

[Zhang et al., 2017] Zhang, L., Zhu, G., Shen, P., Song, J., Afaq Shah, S., and Bennamoun, M. "Learning Spatiotemporal Features Using 3DCNN and Convolutional LSTM for Gesture Recognition". In: *International Conference on Computer Vision Workshops – ICCV-W 2017*. 2017, pp. 3120–3128.

[Zhao et al., 2019] Zhao, S., Dong, Y., Chang, E. I., Xu, Y., et al. "Recursive Cascaded Networks for Unsupervised Medical Image Registration". In: *International Conference on Computer Vision – ICCV 2019*. 2019, pp. 10600–10610.

[Zhang et al., 2020a] Zhang, W., Lin, Z., Cheng, J., Ma, C., Deng, X., and Wang, H. "STA-GCN: Two-Stream Graph Convolutional Network With Spatial-Temporal Attention for Hand Gesture Recognition". *The Visual Computer* 36 (10), 2020, pp. 2433–2444.

[Zhang et al., 2020b] Zhang, Z., Hu, L., Deng, X., and Xia, S. "Weakly Supervised Adversarial Learning for 3D Human Pose Estimation From Point Clouds". *IEEE Transactions on Visualization and Computer Graphics* 26 (5), 2020, pp. 1851–1859.

[Zhang et al., 2021] Zhang, Y. and Yang, Q. "A Survey on Multi-Task Learning". *IEEE Transactions on Knowledge and Data Engineering* 34 (12), 2021, pp. 5586–5609.

[Zhang et al., 2022] Zhang, J., Qi, L., Shi, Y., and Gao, Y. "Generalizable Model-Agnostic Semantic Segmentation via Target-specific Normalization". *Pattern Recognition* 122, 2022, p. 108292.

[Zheng et al., 2020] Zheng, H., Motch Perrine, S. M., Pitirri, M. K., Kawasaki, K., Wang, C., Richtsmeier, J. T., and Chen, D. Z. "Cartilage Segmentation in High-Resolution 3D Micro-CT Images via Uncertainty-Guided Self-Training With Very Sparse Annotation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention – MICCAI 2020: Proceedings, Part I*. 2020, pp. 802–812.

[Zheng et al., 2021] Zheng, Z. and Yang, Y. "Rectifying Pseudo Label Learning via Uncertainty Estimation for Domain Adaptive Semantic Segmentation". *International Journal of Computer Vision* 129 (4), 2021, pp. 1106–1120.

[Zhou et al., 2017] Zhou, X., Huang, Q., Sun, X., Xue, X., and Wei, Y. "Towards 3D Human Pose Estimation in the Wild: A Weakly-Supervised Approach". In: *International Conference on Computer Vision – ICCV 2017*. 2017, pp. 398–407.

[Zhou et al., 2018] Zhou, Y. and Tuzel, O. "VoxelNet: End-to-End Learning for Point Cloud Based 3D Object Detection". In: *Conference on Computer Vision and Pattern Recognition – CVPR 2018*. 2018, pp. 4490–4499.

[Zhou et al., 2022a] Zhou, K., Liu, Z., Qiao, Y., Xiang, T., and Loy, C. C. "Domain Generalization: A Survey". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[Zhou et al., 2022b] Zhou, Q., Feng, Z., Gu, Q., Cheng, G., Lu, X., Shi, J., and Ma, L. "Uncertainty-Aware Consistency Regularization for Cross-Domain Semantic Segmentation". *Computer Vision and Image Understanding*, 2022, p. 103448.

[Zhu et al., 2017a] Zhu, G., Zhang, L., Shen, P., and Song, J. "Multimodal Gesture Recognition Using 3-D Convolution and Convolutional LSTM". *IEEE Access* 5, 2017, pp. 4517–4524.

[Zhu et al., 2017b] Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. "Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks". In: *International Conference on Computer Vision – ICCV 2017*. 2017, pp. 2223–2232.

[Zou et al., 2018] Zou, Y., Yu, Z., Kumar, B., and Wang, J. "Unsupervised Domain Adaptation for Semantic Segmentation via Class-Balanced Self-Training". In: *European Conference on Computer Vision – ECCV 2018*. 2018, pp. 289–305.

[Zou et al., 2021] Zou, L., Tang, H., Chen, K., and Jia, K. "Geometry-Aware Self-Training for Unsupervised Domain Adaptation on Object Point Clouds". In: *International Conference on Computer Vision – ICCV 2021*. 2021, pp. 6403–6412.

List of Publications

This list contains journal articles and conference papers published during the work on this dissertation.

Journal Articles as First Author

- Bigalke, A., Hansen, L., Diesel, J., Hennigs, C., Rostalski, P., and Heinrich, M. P. "Anatomy-Guided Domain Adaptation for 3D In-Bed Human Pose Estimation". *Medical Image Analysis*, 2023, p. 102887.
- 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.

Conference Papers as First Author

- Bigalke, A., Hansen, L., Mok, T. C., and Heinrich, M. P. "Unsupervised 3D Registration Through Optimization-Guided Cyclical Self-Training". In: *International Conference on Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*. 2023, to appear.
- Bigalke, A. and Heinrich, M. P. "A Denoised Mean Teacher for Domain Adaptive Point Cloud Registration". In: *International Conference on Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*. 2023, to appear.
- Bigalke, A., Hansen, L., and Heinrich, M. P. "Adapting the Mean Teacher for Keypoint-Based Lung Registration Under Geometric Domain Shifts". In: *International Conference on Medical Image Computing and Computer Assisted Intervention – MICCAI 2022: Proceedings, Part VI*. 2022, pp. 280–290.
- 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, A. and Heinrich, M. P. "Fusing Posture and Position Representations for Point Cloud-Based Hand Gesture Recognition". In: *International Conference on 3D Vision – 3DV 2021*. 2021, pp. 617–626.

- 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 – BVM 2021*. 2021, pp. 254–259.

Journal Articles and Conference Papers as Co-Author

- Heinrich, M. P., Bigalke, A., Großbröhmer, C., and Hansen, L. "Chasing Clouds: Differentiable Volumetric Rasterisation of Point Clouds as a Highly Efficient and Accurate Loss for Large-Scale Deformable 3D Registration". In: *International Conference on Computer Vision – ICCV 2023*. 2023, to appear.
- Hermes, N., Bigalke, A., and Heinrich, M. P. "Point Cloud-Based Scene Flow Estimation on Realistically Deformable Objects: A Benchmark of Deep Learning-Based Methods". *Journal of Visual Communication and Image Representation*, 2023, p. 103893.
- Weihsbach, C., Bigalke, A., Kruse, C. N., Hempe, H., and Heinrich, M. P. "Deep-STAPLE: Learning to Predict Multimodal Registration Quality for Unsupervised Domain Adaptation". In: *Workshop on Biomedical Image Registration – WBIR 2022*. 2022, pp. 37–46.
- Hermes, N., Hansen, L., 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 – BVM 2022*. 2022, pp. 89–94.