

# Anatomy-guided domain adaptation for 3D in-bed human pose estimation

Alexander Bigalke<sup>a,\*</sup>, Lasse Hansen<sup>b</sup>, Jasper Diesel<sup>c</sup>, Carlotta Hennigs<sup>d</sup>, Philipp Rostalski<sup>d</sup>, Mattias P. Heinrich<sup>a</sup>

<sup>a</sup>*Institute of Medical Informatics, University of Lübeck, Ratzeburger Allee 160, 23538 Lübeck, Germany*

<sup>b</sup>*EchoScout GmbH, Maria-Goeppert-Str. 3, 23562 Lübeck, Germany*

<sup>c</sup>*Drägerwerk AG & Co. KGaA, Moislinger Allee 53-55, 23558 Lübeck, Germany*

<sup>d</sup>*Institute for Electrical Engineering in Medicine, University of Lübeck, Moislinger Allee 53-55, 23558 Lübeck, Germany*

## ARTICLE INFO

### Article history:

Received -

Received in final form -

Accepted -

Available online -

Communicated by -

**Keywords:** Domain adaptation, In-bed human pose estimation, Anatomy-constrained optimization, Anatomy-guided self-training, Point clouds

## ABSTRACT

3D human pose estimation is a key component of clinical monitoring systems. The clinical applicability of deep pose estimation models, however, is limited by their poor generalization under domain shifts along with their need for sufficient labeled training data. As a remedy, we present a novel domain adaptation method, adapting a model from a labeled source to a shifted unlabeled target domain. Our method comprises two complementary adaptation strategies based on prior knowledge about human anatomy. First, we guide the learning process in the target domain by constraining predictions to the space of anatomically plausible poses. To this end, we embed the prior knowledge into an anatomical loss function that penalizes asymmetric limb lengths, implausible bone lengths, and implausible joint angles. Second, we propose to filter pseudo labels for self-training according to their anatomical plausibility and incorporate the concept into the Mean Teacher paradigm. We unify both strategies in a point cloud-based framework applicable to unsupervised and source-free domain adaptation. Evaluation is performed for in-bed pose estimation under two adaptation scenarios, using the public SLP dataset and a newly created dataset. Our method consistently outperforms various state-of-the-art domain adaptation methods, surpasses the baseline model by 31%/66%, and reduces the domain gap by 65%/82%. Source code is available at <https://github.com/multimodallearning/da-3dhpe-anatomy>.

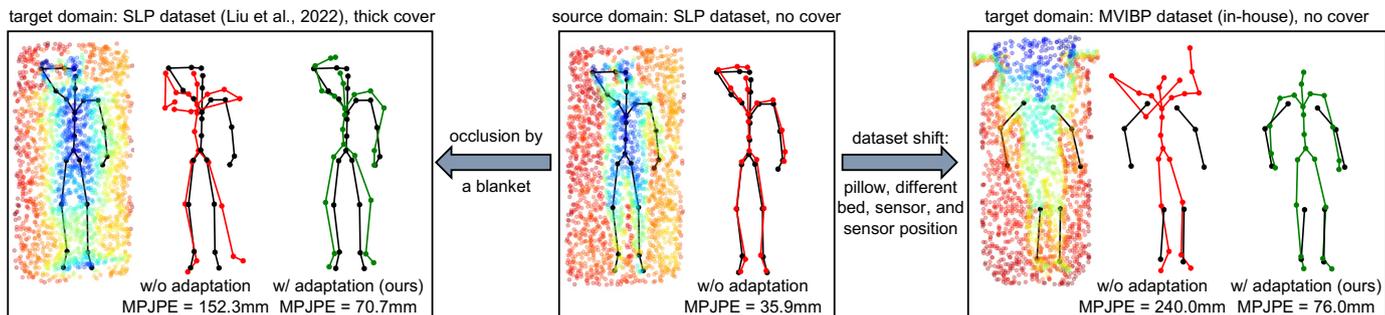
## 1. Introduction

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., 2020) and action recognition (Song et al., 2021). These tasks, in turn, open up a wide range of 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 and Padoy, 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 con-

\*Corresponding author: Tel.: +49 451 3101 5619;  
e-mail: alexander.bigalke@uni-luebeck.de (Alexander Bigalke)



**Fig. 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.**

ditions 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. 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 and Deng, 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 and Lempitsky, 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.

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. 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 and Valpola, 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 and Ostadabbas, 2019; Liu *et al.*, 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.

## 2. Related work

### 2.1. 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 Sec. 2.2), 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 and Ostadabbas (2019) estimate 2D poses from thermal images, and Casas *et al.* (2019); 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, 2022) train their models on synthetic depth or pressure maps of covered patients, and Afham *et al.* (2022); Chi *et al.* (2022) perform domain adaptation from labeled uncovered to unlabeled covered subjects based on thermal images (see Sec. 2.5).

### 2.2. 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.*, 2019; 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 and Lee (2019) 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).

### 2.3. 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.*, 2019; 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.*, 2017) 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 and Lempitsky, 2015; Tzeng *et al.*, 2017; Saito *et al.*, 2019), 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); 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 and Valpola, 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.*, 2020, 2021c; Yu *et al.*, 2019), as the predictive variance under input perturbations (Zhou *et al.*, 2022) and among different network heads (Zheng *et al.*, 2020; Zheng and Yang, 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); 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).

#### 2.4. 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.

#### 2.5. 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 and Lee (2021); 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); 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.* (2020); Zhou *et al.* (2017) to regularize supervised learning with weak 2D pose ground truth. Moreover, Cao and Zhao (2020); 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.

### 3. Methods

#### 3.1. Problem setup and notation

Point cloud-based 3D human pose estimation aims at predicting the 3D positions of  $K$  human joints of interest,  $\mathbf{Y} \in \mathbb{R}^{K \times 3}$ , from a 3D input point cloud  $\mathbf{X} \in \mathbb{R}^{N \times 3}$ . We address the task in a domain adaptation setting, where training data consists of a labeled source dataset  $\mathcal{S} = \{(\mathbf{X}_s, \mathbf{Y}_s)\}_{s=1}^{|\mathcal{S}|}$  and a shifted unlabeled target dataset  $\mathcal{T} = \{\mathbf{X}_t\}_{t=1}^{|\mathcal{T}|}$ . The goal is to learn a function  $f$  with parameters  $\theta_f$  that predicts human poses as  $\hat{\mathbf{Y}} = f(\mathbf{X}; \theta_f)$  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  $\mathbf{Y}$ , we indicate individual joints as  $\mathbf{y}_k \in \mathbb{R}^3$  and treat them as the nodes of a skeleton graph. We denote  $\mathcal{B} = \{\mathbf{b}_i\}_{i=1}^{N_\beta}$  as the set of all bone vectors  $\mathbf{b}_i \in \mathbb{R}^3$  that connect two joints in the skeleton graph, and  $\mathbf{b}_{t,i}$  indicates the  $i$ -th bone vector of the indexed pose  $\mathbf{Y}_t$ . We further indicate  $\mathcal{B}_\lambda \subset \mathcal{B}$  as the subset of  $N_\lambda$  bones  $\mathbf{b}_i^\lambda$  of the left body side that have a counterpart  $\mathbf{b}_i^\rho \in \mathcal{B}_\rho$  on the right body side. Finally, we term  $\mathcal{B}_\zeta = \{(\mathbf{b}_i, \mathbf{b}_j)\}$  as the set of all  $N_\zeta$  pairs of bone vectors that are connected by a joint and define  $\mathcal{I}_\zeta = \{(i, j)\}$  as the corresponding set of indices.

#### 3.2. Overview

An overview of our proposed method to solve the above problem is shown in Fig. 2. While supervised learning on labeled source data is performed by minimizing the task loss

$$\mathcal{L}_{\text{task}}(\theta_f; \mathcal{S}) = \frac{1}{|\mathcal{S}|} \sum_s \frac{1}{K} \|\mathbf{Y}_s - \hat{\mathbf{Y}}_s\|_1 \quad (1)$$

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. 3.3. 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. 3.4 for details).

#### 3.3. 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

$$\begin{aligned} \min_{\theta_f} \quad & \mathcal{L}_{\text{task}}(\theta_f; \mathcal{S}) \\ \text{s.t.} \quad & \hat{\mathbf{Y}}_t \text{ is a plausible human pose} \quad t = 1, \dots, |\mathcal{T}| \end{aligned} \quad (2)$$

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:

- **Symmetric limbs:** Corresponding limb pairs  $(\mathbf{b}_i^\lambda, \mathbf{b}_i^\rho)$  of the human body typically have roughly equal lengths, with a deviation  $|\|\mathbf{b}_i^\lambda\|_2 - \|\mathbf{b}_i^\rho\|_2| < \delta_i$  smaller than a limb-specific tolerance  $\delta_i$ . 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  $\mathbf{b}_i$  are constrained by bone-specific upper and lower bounds  $u_i^\beta$  and  $l_i^\beta$ , i.e.,  $l_i^\beta \leq \|\mathbf{b}_i\|_2 \leq u_i^\beta$ . Precise values for  $u_i^\beta$  and  $l_i^\beta$  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  $(\mathbf{b}_i, \mathbf{b}_j) \in \mathcal{B}_\zeta$  is constrained by joint-specific upper and lower bounds  $u_{ij}^\alpha$  and  $l_{ij}^\alpha$ , i.e.,  $l_{ij}^\alpha \leq \mathbf{b}_i / \|\mathbf{b}_i\|_2 \cdot \mathbf{b}_j / \|\mathbf{b}_j\|_2 \leq u_{ij}^\alpha$ . 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

$$\begin{aligned} \min_{\theta_f} \quad & \mathcal{L}_{\text{task}}(\theta_f; \mathcal{S}) \\ \text{s.t.} \quad & -\delta_i < \|\mathbf{b}_{t,i}^\lambda\|_2 - \|\mathbf{b}_{t,i}^\rho\|_2 < \delta_i \quad i = 1, \dots, N_\lambda; t = 1, \dots, |\mathcal{T}| \\ & l_i^\beta \leq \|\mathbf{b}_{t,i}\|_2 \leq u_i^\beta \quad i = 1, \dots, N_\beta; t = 1, \dots, |\mathcal{T}| \\ & l_{ij}^\alpha \leq \frac{\mathbf{b}_{t,i}}{\|\mathbf{b}_{t,i}\|_2} \cdot \frac{\mathbf{b}_{t,j}}{\|\mathbf{b}_{t,j}\|_2} \leq u_{ij}^\alpha \quad \forall (i, j) \in \mathcal{I}_\zeta; t = 1, \dots, |\mathcal{T}| \end{aligned} \quad (3)$$

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

$$\ell(x; l, u) = \begin{cases} |x - l| & x < l \\ |x - u| & x > u \\ 0 & l < x < u \end{cases} \quad (4)$$

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. 5.1.1). Given a human pose  $\mathbf{Y}$ ,

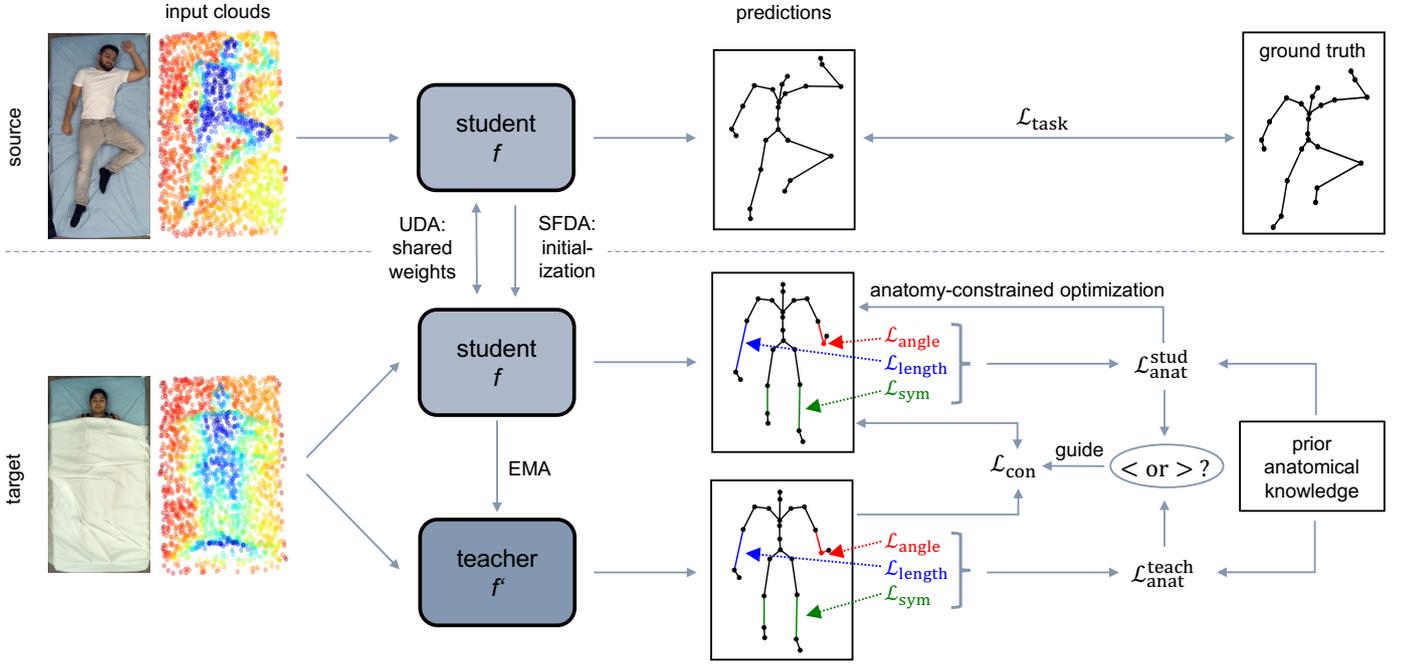


Fig. 2. Overview of our method for domain adaptive human pose estimation from point clouds (RGB images are only shown for better visualization). The framework comprises a learning student model and a teacher model, which represents the exponential moving average (EMA) of the student. While source training 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.

the violation of our anatomical constraints is then penalized by the loss functions

$$\begin{aligned} \mathcal{L}_{\text{sym}}(\mathbf{Y}) &= \frac{1}{N_\lambda} \sum_{i=1}^{N_\lambda} \ell(\|\mathbf{b}_i^\lambda\|_2 - \|\mathbf{b}_i^\rho\|_2, -\delta_i, \delta_i) \\ \mathcal{L}_{\text{length}}(\mathbf{Y}) &= \frac{1}{N_\beta} \sum_{i=1}^{N_\beta} \ell(\|\mathbf{b}_i\|_2, l_i^\beta, u_i^\beta) \\ \mathcal{L}_{\text{angle}}(\mathbf{Y}) &= \frac{1}{N_\zeta} \sum_{(i,j) \in \mathcal{I}_\zeta} \ell\left(\frac{\mathbf{b}_i}{\|\mathbf{b}_i\|_2} \cdot \frac{\mathbf{b}_j}{\|\mathbf{b}_j\|_2}; l_{ij}^\alpha, u_{ij}^\alpha\right) \end{aligned} \quad (5)$$

This enables us to replace the constrained optimization problem in Eq. (3) by the standard minimization of the joint loss function

$$\mathcal{L}(\theta_f; \mathcal{S}, \mathcal{T}) = \mathcal{L}_{\text{task}}(\theta_f; \mathcal{S}) + \lambda_1 \mathcal{L}_{\text{anat}}(\theta_f; \mathcal{T}) \quad (6)$$

with the anatomical loss

$$\mathcal{L}_{\text{anat}}(\theta_f, \mathcal{T}) = \frac{1}{|\mathcal{T}|} \sum_t \left[ \mathcal{L}_{\text{sym}}(\hat{\mathbf{Y}}_t) + \mathcal{L}_{\text{length}}(\hat{\mathbf{Y}}_t) + \mathcal{L}_{\text{angle}}(\hat{\mathbf{Y}}_t) \right] \quad (7)$$

and the weighting factor  $\lambda_1$ . Individual weighting factors for each loss were explored but did not yield an improvement.

### 3.3.1. 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  $\theta_f$ , 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  $\theta_g \subset \theta_f$  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 Batch-Norm 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. 3.4, 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.

### 3.4. Self-training with the Mean Teacher

The Mean Teacher framework (French et al., 2018; Tarvainen and Valpola, 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  $\theta'_f$  are not optimized by gradient descent but given as the exponential moving average (EMA) of the student's weights, updated as

$$\theta'_{f,i} = \mu\theta'_{f,i-1} + (1 - \mu)\theta_{f,i} \quad (8)$$

at iteration  $i$  with momentum  $\mu$ . 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

$$\mathcal{L}_{\text{con}}(\theta_f; \theta'_f, \mathcal{T}) = \frac{1}{|\mathcal{T}|} \sum_t \frac{1}{K} \|\hat{Y}_t - \hat{Y}'_t\|_1 \quad (9)$$

encouraging predictions  $\hat{Y}'_t = f'(X_t; \theta'_f)$  by the teacher and  $\hat{Y}_t = f(X_t; \theta_f)$  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.

#### 3.4.1. 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. (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. (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. 5.1.2).

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

$$h(\hat{Y}', \hat{Y}) = \mathbb{1}\left(\left[\mathbb{1}(\mathcal{L}_{\text{sym}}(\hat{Y}') < \mathcal{L}_{\text{sym}}(\hat{Y})) + \mathbb{1}(\mathcal{L}_{\text{length}}(\hat{Y}') < \mathcal{L}_{\text{length}}(\hat{Y})) + \mathbb{1}(\mathcal{L}_{\text{angle}}(\hat{Y}') < \mathcal{L}_{\text{angle}}(\hat{Y}))\right] \geq 2\right) \quad (10)$$

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. (9) as

$$\mathcal{L}_{\text{con}}(\theta_f; \theta'_f, \mathcal{T}) = \frac{1}{|\mathcal{T}|} \sum_t \frac{1}{K} h(\hat{Y}', \hat{Y}) \cdot \|\hat{Y}_t - \hat{Y}'_t\|_1 \quad (11)$$

Taking altogether, we integrate this consistency loss into our previous objective function from Eq. (6). To perform UDA, we thus minimize

$$\begin{aligned} \mathcal{L}(\theta_f; \theta'_f, \mathcal{S}, \mathcal{T}) = & \mathcal{L}_{\text{task}}(\theta_f; \mathcal{S}) \\ & + \lambda(\tau)\lambda_1 \mathcal{L}_{\text{anat}}(\theta_f; \mathcal{T}) \\ & + \lambda(\tau)\lambda_2 \mathcal{L}_{\text{con}}(\theta_f; \theta'_f, \mathcal{T}) \end{aligned} \quad (12)$$

Here,  $\lambda(\tau) = \exp(-5(1 - \min(\tau/T, 1)^2))$  depends on the current epoch  $\tau$  and continually increases from 0 to 1 during the first  $T$  epochs, as suggested by Tarvainen and Valpola (2017), while  $\lambda_2$  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

$$\mathcal{L}(\theta_f; \theta'_f, \mathcal{T}) = \lambda_1 \mathcal{L}_{\text{anat}}(\theta_f; \mathcal{T}) + \lambda_2 \mathcal{L}_{\text{con}}(\theta_f; \theta'_f, \mathcal{T}) \quad (13)$$

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. 3.3.1, we found it beneficial to minimize the loss in Eq. (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.

### 3.5. 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  $X \in \mathbb{R}^{N \times 3}$ , we estimate the associated 3D pose  $\hat{Y} \in \mathbb{R}^{K \times 3}$  as the weighted

sum over the  $N$  input points  $\mathbf{x}_i \in \mathbb{R}^3$ . To this end, we design  $f$  to output a stack of  $K$  softmax-normalized weight maps  $\mathbf{W} = f(\mathbf{X}; \theta_f) \in \mathbb{R}^{N \times K}$  over the input points. The  $k$ -th predicted joint is then given by  $\hat{\mathbf{y}}_k = \sum_{i=1}^N \mathbf{x}_i \cdot w_{ik}$ . 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. Experimental setup

### 4.1. 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 and Ostadabbas, 2019; Liu et al., 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  $\sim 1$  mm, thick cover  $\sim 3$  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<sup>1</sup>. 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 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.

### 4.2. 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 dataset are not identical (see Fig. 1, right), we restrict the evaluation to the matching joint pairs, namely feet, knees, shoulders, elbows, and hands.

<sup>1</sup>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.

### 4.3. 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. (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  $\mu$  for updating the teacher’s weights is set to 0.99 for UDA and 0.9996 for SFDA. Upper and lower bounds  $u_{ij}^\alpha$ ,  $u_i^\beta / l_{ij}^\alpha$ ,  $l_i^\beta$  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 (Sec. 4.4) were tuned on the validation set of the target domain under the uncover→cover scenario and kept fix 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).

### 4.4. 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); 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 and Lempitsky (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. 3.4). 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 repre-

**Table 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
$\mathcal{L}_{\text{angle}}$	✓		<b>106.7</b>
$\mathcal{L}_{\text{angle}}$		✓	119.3
$\mathcal{L}_{\text{sym}}$	✓		<b>105.9</b>
$\mathcal{L}_{\text{sym}}$		✓	108.6
$\mathcal{L}_{\text{length}}$	✓		<b>102.9</b>
$\mathcal{L}_{\text{length}}$		✓	104.1
$\mathcal{L}_{\text{anat}}$	✓		<b>96.6</b>

sent 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.

## 5. Results

### 5.1. Ablation study

We start by analyzing the two essential components of our method, namely anatomy-constrained optimization in Sec. 5.1.1 and anatomy-guided filtering of pseudo-labels in Sec. 5.1.2. The ablation experiments are performed under the uncover→cover setting on the SLP dataset.

#### 5.1.1. Anatomy-constrained optimization

In the first ablation experiment, we examine the effectiveness of the proposed anatomical loss functions from Eq. (5). We consider the UDA setting, discard the Mean Teacher, and minimize  $\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda_1 \mathcal{L}_x$  with  $\mathcal{L}_x \in \{\mathcal{L}_{\text{sym}}, \mathcal{L}_{\text{angle}}, \mathcal{L}_{\text{length}}, \mathcal{L}_{\text{anat}}\}$ .

**Table 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	-	-	<b>92.3</b>	<b>97.0</b>

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. 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 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%.

#### 5.1.2. 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. 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. 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 SFDA settings, discard anatomy-constrained optimization ( $\lambda_1 = 0$ ), and

**Table 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 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	<b>47.1</b>	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	<b>56.0</b>	47.2	33.3	<b>50.4</b>	112.5	188.4	119.0	96.6
ours, $\mathcal{L}_{\text{con}}$ only	✓		134.7	97.3	60.0	49.1	<b>33.2</b>	54.3	110.5	<b>163.9</b>	112.1	92.1
ours	✓		<b>120.4</b>	<b>97.0</b>	57.4	<b>47.1</b>	33.8	51.7	<b>109.1</b>	169.8	<b>109.6</b>	<b>89.6</b>
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	<b>32.9</b>	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	<b>178.1</b>	119.4	97.0
ours		✓	<b>132.8</b>	<b>102.5</b>	<b>62.3</b>	<b>49.8</b>	33.6	<b>53.1</b>	<b>118.3</b>	179.6	<b>117.3</b>	<b>95.6</b>

use different variants of Eq. (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 ( $h(\hat{Y}', \hat{Y}) = 1$ ), which is equivalent to the standard Mean Teacher. As another comparison method, similar to Ke et al. (2019); Mu et al. (2020); Zhou et al. (2022), 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 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.

## 5.2. Comparison to the state of the art

We compare our method to the comparison methods presented in Sec. 4.4 under the two adaptation scenarios uncover→cover (U→C) and SLP→MVIBP. Quantitative results are shown in Tab. 3, Tab. 4, and Fig. 5.2, 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

**Table 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	<b>61.5</b>	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	<b>75.6</b>	<b>103.0</b>	86.0
ours	✓		<b>62.6</b>	<b>70.2</b>	83.2	84.9	108.4	<b>81.8</b>
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	<b>75.3</b>	106.2	<b>156.8</b>	101.8
ours, $\mathcal{L}_{\text{con}}$ only		✓	<b>66.7</b>	<b>62.5</b>	83.7	98.4	182.9	98.8
ours		✓	68.8	68.3	79.3	<b>86.8</b>	174.4	<b>95.5</b>

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, 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 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—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 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.

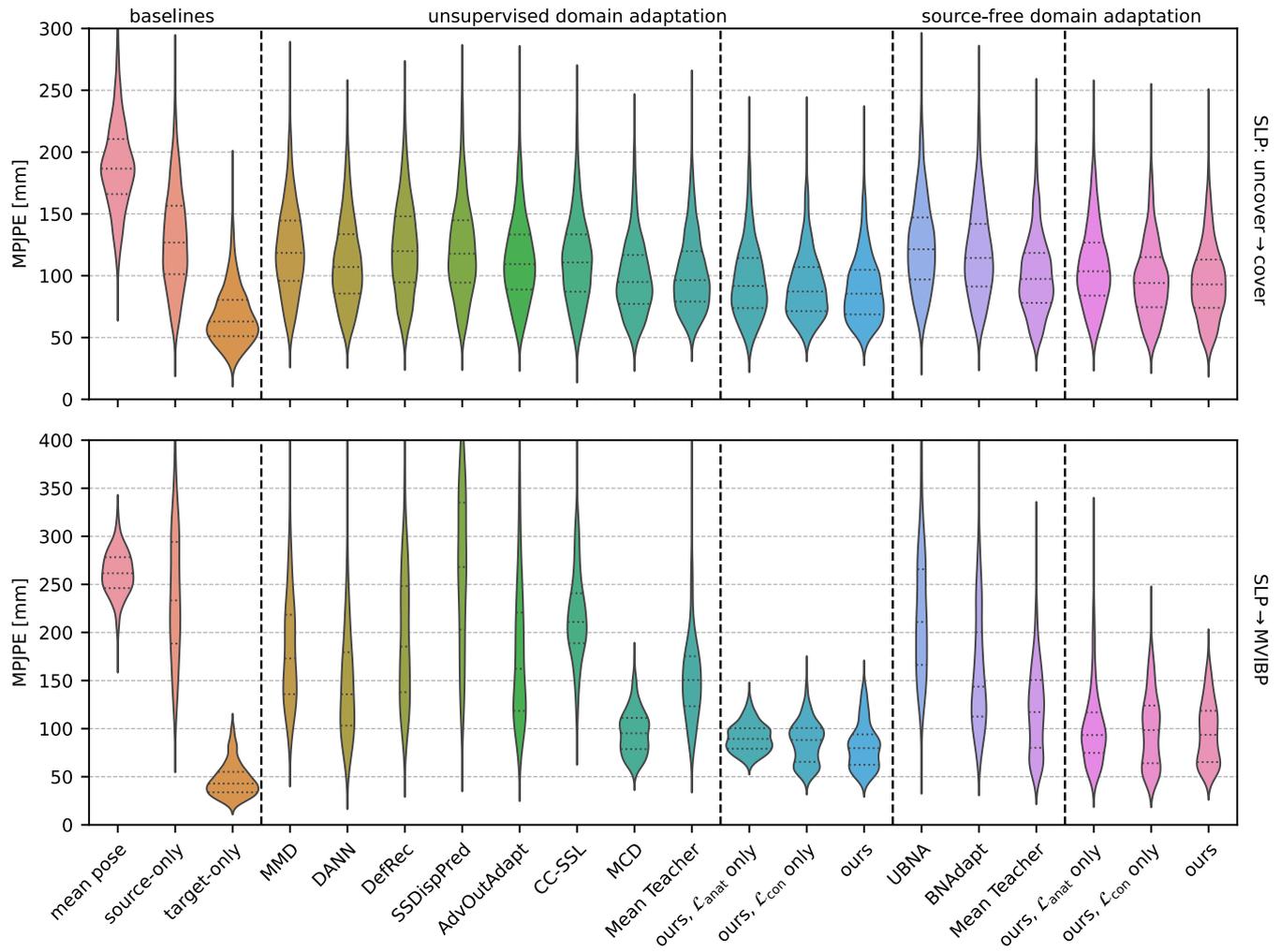


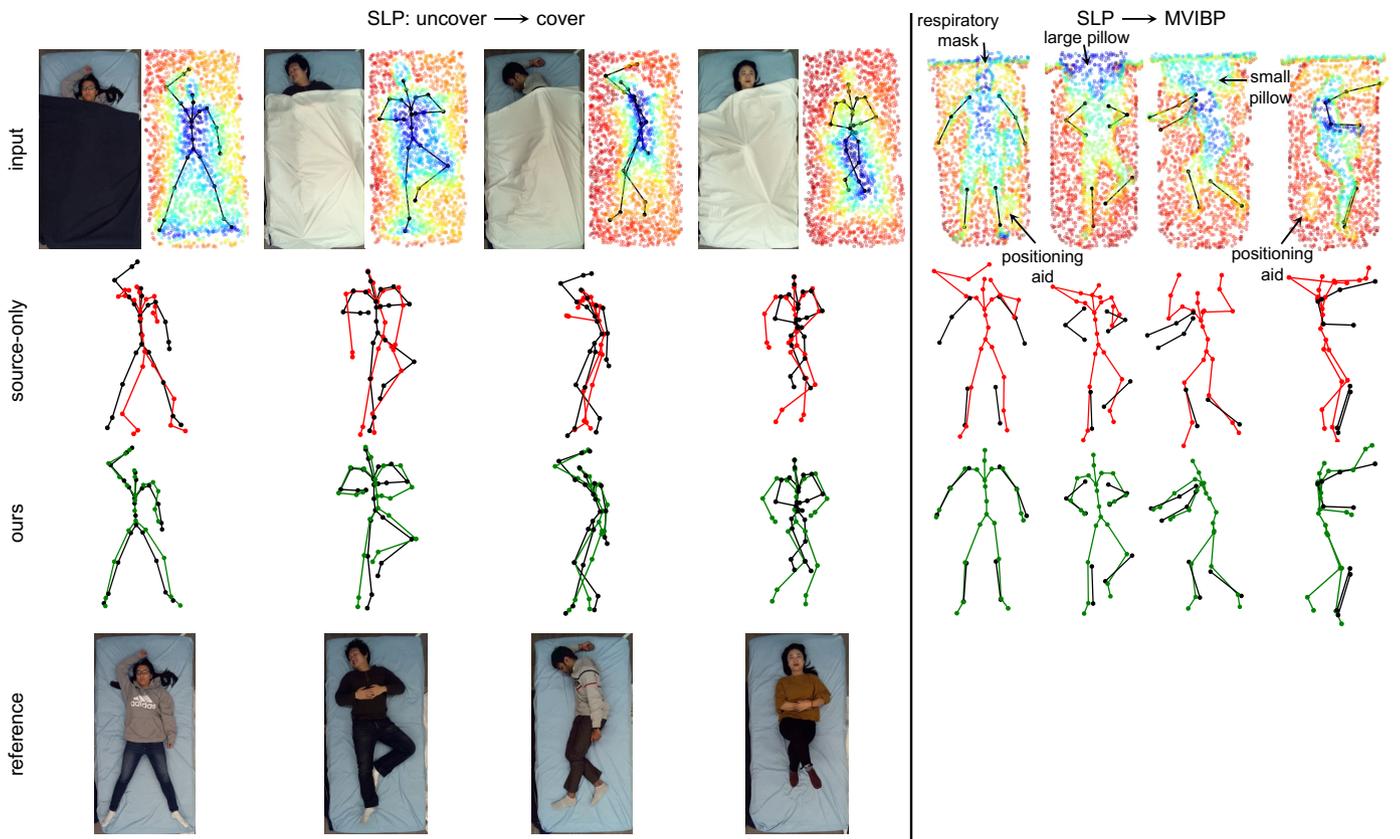
Fig. 3. Violin plots of the frame-averaged joint errors for all compared methods in the uncover→cover setting on the SLP dataset (top) and for SLP→MVIBP adaptation (bottom). Dashed lines inside the violins represent the 25th, 50th, and 75th percentiles.

## 6. 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 by 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



**Fig. 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.**

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 oppor-

tunities 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 and Black, 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 co-

herent 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).

## Acknowledgments

We gratefully acknowledge the financial support by the Federal Ministry for Economic Affairs and Climate Action of Germany (FKZ: 01MK20012B).

## References

- Achilles, F., Ichim, A.E., Coskun, H., Tombari, F., Noachtar, S., Navab, N., 2016. Patient mocap: Human pose estimation under blanket occlusion for hospital monitoring applications, in: International conference on medical image computing and computer-assisted intervention, Springer. pp. 491–499.
- Achituve, I., Maron, H., Chechik, G., 2021. Self-supervised learning for domain adaptation on point clouds, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 123–133.
- Adiga Vasudeva, S., Dolz, J., Lombaert, H., 2022. Leveraging labeling representations in uncertainty-based semi-supervised segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 265–275.
- Afham, M., Haputhanthri, U., Pradeepkumar, J., Anandakumar, M., De Silva, A., Edussooriya, C.U., 2022. Towards accurate cross-domain in-bed human pose estimation, in: ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE. pp. 2664–2668.
- Akhter, I., Black, M.J., 2015. Pose-conditioned joint angle limits for 3d human pose reconstruction, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1446–1455.
- Alliegro, A., Boscaini, D., Tommasi, T., 2021. Joint supervised and self-supervised learning for 3d real world challenges, in: 2020 25th International Conference on Pattern Recognition (ICPR), IEEE. pp. 6718–6725.
- Bateson, M., Dolz, J., Kervadec, H., Lombaert, H., Ayed, I.B., 2021. Constrained domain adaptation for image segmentation. *IEEE Transactions on Medical Imaging* 40, 1875–1887.
- Bateson, M., Kervadec, H., Dolz, J., Lombaert, H., Ben Ayed, I., 2020. Source-relaxed domain adaptation for image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 490–499.
- Bateson, M., Lombaert, H., Ben Ayed, I., 2022. Test-time adaptation with shape moments for image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 736–745.
- Belagiannis, V., Wang, X., Shitrit, H.B.B., Hashimoto, K., Stauder, R., Aoki, Y., Kranzfelder, M., Schneider, A., Fua, P., Ilic, S., Feussner, H., Navab, N., 2016. Parsing human skeletons in an operating room. *Machine Vision and Applications* 27, 1035–1046.
- Bertsekas, D.P., 1997. Nonlinear programming. *Journal of the Operational Research Society* 48, 334–334.
- Bigalke, A., Hansen, L., Diesel, J., Heinrich, M.P., 2022a. Domain adaptation through anatomical constraints for 3d human pose estimation under the cover, in: International Conference on Medical Imaging with Deep Learning, PMLR. pp. 173–187.
- Bigalke, A., Hansen, L., Heinrich, M.P., 2022b. Adapting the mean teacher for keypoint-based lung registration under geometric domain shifts, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 280–290.
- Bousmalis, K., Trigeorgis, G., Silberman, N., Krishnan, D., Erhan, D., 2016. Domain separation networks. *Advances in neural information processing systems* 29.
- Cai, Q., Pan, Y., Ngo, C.W., Tian, X., Duan, L., Yao, T., 2019. Exploring object relation in mean teacher for cross-domain detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11457–11466.
- Cao, J., Tang, H., Fang, H.S., Shen, X., Lu, C., Tai, Y.W., 2019. Cross-domain adaptation for animal pose estimation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9498–9507.
- Cao, X., Zhao, X., 2020. Anatomy and geometry constrained one-stage framework for 3d human pose estimation, in: Proceedings of the Asian Conference on Computer Vision.
- Cardace, A., Spezialetti, R., Ramirez, P.Z., Salti, S., Di Stefano, L., 2021. Refrec: Pseudo-labels refinement via shape reconstruction for unsupervised 3d domain adaptation, in: 2021 International Conference on 3D Vision (3DV), IEEE. pp. 331–341.
- Casas, L., Navab, N., Demirci, S., 2019. Patient 3d body pose estimation from pressure imaging. *International journal of computer assisted radiology and surgery* 14, 517–524.
- Chang, W.G., You, T., Seo, S., Kwak, S., Han, B., 2019. Domain-specific batch normalization for unsupervised domain adaptation, in: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition, pp. 7354–7362.
- Chen, C., Liu, Q., Jin, Y., Dou, Q., Heng, P.A., 2021. Source-free domain adaptive fundus image segmentation with denoised pseudo-labeling, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 225–235.
- 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., Gilja, V., 2018. Patient-specific pose estimation in clinical environments. *IEEE journal of translational engineering in health and medicine* 6, 1–11.
- Chen, Y., Tian, Y., He, M., 2020. Monocular human pose estimation: A survey of deep learning-based methods. *Computer Vision and Image Understanding* 192, 102897.
- Chi, Z., Wang, S., Li, X., Chang, C.T., Islam, M., Holkar, A., Pronger, S., Liu, T., Lam, K.M., He, X., 2022. Multi-level unsupervised domain adaption for privacy-protected in-bed pose estimation, in: International Workshop on Advanced Imaging Technology (IWAIT) 2022, SPIE. pp. 431–436.
- Clever, H.M., Erickson, Z., Kapusta, A., Turk, G., Liu, K., Kemp, C.C., 2020. Bodies at rest: 3d human pose and shape estimation from a pressure image using synthetic data, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 6215–6224.
- Clever, H.M., Grady, P., Turk, G., Kemp, C.C., 2022. Bodypressure-inferring body pose and contact pressure from a depth image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Cunha, J.P.S., Choupina, H.M.P., Rocha, A.P., Fernandes, J.M., Achilles, F., Loesch, A.M., Vollmar, C., Hartl, E., Noachtar, S., 2016. Neurokinect: a novel low-cost 3dvideo-eeeg system for epileptic seizure motion quantification. *PloS one* 11, e0145669.
- Davoodnia, V., Ghorbani, S., Etemad, A., 2021. In-bed pressure-based pose estimation using image space representation learning, in: ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE. pp. 3965–3969.
- Deng, J., Li, W., Chen, Y., Duan, L., 2021. Unbiased mean teacher for cross-domain object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4091–4101.
- Doersch, C., Gupta, A., Efros, A.A., 2015. Unsupervised visual representation learning by context prediction, in: Proceedings of the IEEE international conference on computer vision, pp. 1422–1430.
- Fan, H., Chang, X., Zhang, W., Cheng, Y., Sun, Y., Kankanhalli, M., 2022. Self-supervised global-local structure modeling for point cloud domain adaptation with reliable voted pseudo labels, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6377–6386.
- French, G., Mackiewicz, M., Fisher, M., 2018. Self-ensembling for visual domain adaptation, in: 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings.
- Ganin, Y., Lempitsky, V., 2015. Unsupervised domain adaptation by back-propagation, in: International conference on machine learning, PMLR. pp. 1180–1189.
- Ge, L., Cai, Y., Weng, J., Yuan, J., 2018a. Hand pointnet: 3d hand pose estimation

- tion using point sets, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8417–8426.
- Ge, L., Ren, Z., Yuan, J., 2018b. Point-to-point regression pointnet for 3d hand pose estimation, in: Proceedings of the European conference on computer vision (ECCV), pp. 475–491.
- Ghifary, M., Kleijn, W.B., Zhang, M., Balduzzi, D., Li, W., 2016. Deep reconstruction-classification networks for unsupervised domain adaptation, in: European conference on computer vision, Springer. pp. 597–613.
- Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., Smola, A., 2006. A kernel method for the two-sample-problem. *Advances in neural information processing systems* 19.
- Hansen, L., Siebert, M., Diesel, J., Heinrich, M.P., 2019. 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, 1871–1879.
- Haque, A., Peng, B., Luo, Z., Alahi, A., Yeung, S., Fei-Fei, L., 2016. Towards viewpoint invariant 3d human pose estimation, in: European conference on computer vision, Springer. pp. 160–177.
- Hegde, D., Sindagi, V., Kilic, V., Cooper, A.B., Foster, M., Patel, V., 2021. Uncertainty-aware mean teacher for source-free unsupervised domain adaptive 3d object detection. *arXiv preprint arXiv:2109.14651*.
- Hermes, N., Hansen, L., Bigalke, A., Heinrich, M.P., 2022. Support point sets for improving contactless interaction in geometric learning for hand pose estimation, in: *Bildverarbeitung für die Medizin 2022*. Springer, pp. 89–94.
- Hoffman, J., Tzeng, E., Park, T., Zhu, J.Y., Isola, P., Saenko, K., Efros, A., Darrell, T., 2018. Cycada: Cycle-consistent adversarial domain adaptation, in: *International conference on machine learning*, Pmlr. pp. 1989–1998.
- Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C., 2013. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence* 36, 1325–1339.
- Jähne-Raden, N., Kulau, U., Marscholke, M., Wolf, K.H., 2019. Inbed: a highly specialized system for bed-exit-detection and fall prevention on a geriatric ward. *Sensors* 19, 1017.
- Kadkhodamohammadi, A., Gangi, A., de Mathelin, M., Padoy, N., 2017. A multi-view rgb-d approach for human pose estimation in operating rooms, in: 2017 IEEE winter conference on applications of computer vision (WACV), IEEE. pp. 363–372.
- Karanam, S., Li, R., Yang, F., Hu, W., Chen, T., Wu, Z., 2020. Towards contactless patient positioning. *IEEE transactions on medical imaging* 39, 2701–2710.
- Ke, Z., Wang, D., Yan, Q., Ren, J., Lau, R.W., 2019. Dual student: Breaking the limits of the teacher in semi-supervised learning, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6728–6736.
- Kervadec, H., Bahig, H., Letourneau-Guillon, L., Dolz, J., Ayed, I.B., 2021. Beyond pixel-wise supervision: semantic segmentation with higher-order shape descriptors.
- Kervadec, H., Dolz, J., Tang, M., Granger, E., Boykov, Y., Ayed, I.B., 2019. Constrained-cnn losses for weakly supervised segmentation. *Medical image analysis* 54, 88–99.
- Kim, D., Wang, K., Saenko, K., Betke, M., Sclaroff, S., 2022. A unified framework for domain adaptive pose estimation, in: *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIII*, Springer. pp. 603–620.
- Klingner, M., Termöhlen, J.A., Ritterbach, J., Fingscheidt, T., 2022. Unsupervised batchnorm adaptation (ubna): A domain adaptation method for semantic segmentation without using source domain representations, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 210–220.
- Kundu, J.N., Kulkarni, A., Singh, A., Jampani, V., Babu, R.V., 2021. Generalize then adapt: Source-free domain adaptive semantic segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7046–7056.
- Kundu, J.N., Venkat, N., Babu, R.V., 2020. Universal source-free domain adaptation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4544–4553.
- Kurmi, V.K., Subramanian, V.K., Nambodiri, V.P., 2021. Domain impression: A source data free domain adaptation method, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 615–625.
- Li, C., Lee, G.H., 2021. From synthetic to real: Unsupervised domain adaptation for animal pose estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1482–1491.
- Li, K., Wang, S., Yu, L., Heng, P.A., 2020. 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*, Springer. pp. 418–427.
- Li, S., Lee, D., 2019. Point-to-pose voting based hand pose estimation using residual permutation equivariant layer, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11927–11936.
- Li, Y., Bu, R., Sun, M., Wu, W., Di, X., Chen, B., 2018. Pointcnn: Convolution on x-transformed points. *Advances in neural information processing systems* 31, 820–830.
- Li, Y., Yuan, L., Vasconcelos, N., 2019. Bidirectional learning for domain adaptation of semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6936–6945.
- Liang, J., Hu, D., Feng, J., 2020. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation, in: *International Conference on Machine Learning*, PMLR. pp. 6028–6039.
- Liu, J., Wang, Y., Liu, Y., Xiang, S., Pan, C., 2020. 3d posturennet: A unified framework for skeleton-based posture recognition. *Pattern Recognition Letters* 140, 143–149.
- Liu, S., Huang, X., Fu, N., Li, C., Su, Z., Ostadabbas, S., 2022a. Simultaneously-collected multimodal lying pose dataset: Enabling in-bed human pose monitoring. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Liu, S., Huang, X., Marcenaro, L., Ostadabbas, S., 2022b. 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, 121–129.
- Liu, S., Ostadabbas, S., 2019. 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*, Springer. pp. 236–245.
- Liu, S., Sehgal, N., Ostadabbas, S., 2022c. Adapted human pose: monocular 3d human pose estimation with zero real 3d pose data. *Applied Intelligence*, 1–16.
- Liu, X., Xing, F., Yang, C., El Fakhri, G., Woo, J., 2021a. Adapting off-the-shelf source segmenter for target medical image segmentation, in: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer. pp. 549–559.
- Liu, Y., Fan, B., Xiang, S., Pan, C., 2019. Relation-shape convolutional neural network for point cloud analysis, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8895–8904.
- Liu, Y., Zhang, W., Wang, J., 2021b. Source-free domain adaptation for semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1215–1224.
- Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J., 2015. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)* 34, 1–16.
- Luo, Y., Zheng, L., Guan, T., Yu, J., Yang, Y., 2019. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2507–2516.
- Martínez-González, A., Villamizar, M., Canévet, O., Odobez, J.M., 2018. Investigating depth domain adaptation for efficient human pose estimation, in: Proceedings of the European Conference on Computer Vision (ECCV) Workshops, pp. 0–0.
- Mascagni, P., Padoy, N., 2021. Or black box and surgical control tower: recording and streaming data and analytics to improve surgical care. *Journal of Visceral Surgery* 158, S18–S25.
- Moon, G., Chang, J.Y., Lee, K.M., 2018. V2v-posenet: Voxel-to-voxel prediction network for accurate 3d hand and human pose estimation from a single depth map, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5079–5088.
- Mu, J., Qiu, W., Hager, G.D., Yuille, A.L., 2020. Learning from synthetic animals, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12386–12395.
- Murez, Z., Kolouri, S., Kriegman, D., Ramamoorthi, R., Kim, K., 2018. Image to image translation for domain adaptation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4500–4509.
- Newell, A., Yang, K., Deng, J., 2016. Stacked hourglass networks for human pose estimation, in: European conference on computer vision, Springer. pp. 483–499.

- Ostadabbas, S., Yousefi, R., Nourani, M., Faezipour, M., Tamil, L., Pompeo, M.Q., 2012. A resource-efficient planning for pressure ulcer prevention. *IEEE Transactions on Information Technology in Biomedicine* 16, 1265–1273.
- Park, T., Efros, A.A., Zhang, R., Zhu, J.Y., 2020. Contrastive learning for unpaired image-to-image translation, in: *European conference on computer vision*, Springer. pp. 319–345.
- Perone, C.S., Ballester, P., Barros, R.C., Cohen-Adad, J., 2019. Unsupervised domain adaptation for medical imaging segmentation with self-ensembling. *NeuroImage* 194, 1–11.
- Qi, C.R., Su, H., Mo, K., Guibas, L.J., 2017a. Pointnet: Deep learning on point sets for 3d classification and segmentation, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 652–660.
- Qi, C.R., Yi, L., Su, H., Guibas, L.J., 2017b. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems* 30.
- Qin, C., You, H., Wang, L., Kuo, C.C.J., Fu, Y., 2019. Pointdan: A multi-scale 3d domain adaption network for point cloud representation. *Advances in Neural Information Processing Systems* 32.
- Rodrigues, V.F., Antunes, R.S., Seewald, L.A., Bazo, R., dos Reis, E.S., dos Santos, U.J., Righi, R.d.R., Junior, L.G.d.S., da Costa, C.A., Bertollo, F.L., Maier, A., Eskofier, B., Horz, T., Pfister, M., Fahrig, R., 2022. A multi-sensor architecture combining human pose estimation and real-time location systems for workflow monitoring on hybrid operating suites. *Future Generation Computer Systems*.
- Rozantsev, A., Salzmann, M., Fua, P., 2018. Beyond sharing weights for deep domain adaptation. *IEEE transactions on pattern analysis and machine intelligence* 41, 801–814.
- Saito, K., Ushiku, Y., Harada, T., Saenko, K., 2019. Strong-weak distribution alignment for adaptive object detection, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6956–6965.
- Saito, K., Watanabe, K., Ushiku, Y., Harada, T., 2018. Maximum classifier discrepancy for unsupervised domain adaptation, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3723–3732.
- Shen, Y., Yang, Y., Yan, M., Wang, H., Zheng, Y., Guibas, L.J., 2022. Domain adaptation on point clouds via geometry-aware implicits, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7223–7232.
- Silas, M.R., Grassia, P., Langerman, A., 2015. Video recording of the operating room—is anonymity possible? *Journal of Surgical Research* 197, 272–276.
- Song, L., Yu, G., Yuan, J., Liu, Z., 2021. Human pose estimation and its application to action recognition: A survey. *Journal of Visual Communication and Image Representation* 76, 103055.
- Srivastav, V., Gangi, A., Padoy, N., 2019. Human pose estimation on privacy-preserving low-resolution depth images, in: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer. pp. 583–591.
- Srivastav, V., Gangi, A., Padoy, N., 2022. Unsupervised domain adaptation for clinician pose estimation and instance segmentation in the operating room. *Medical Image Analysis* 80, 102525.
- Srivastav, V., Issenhuth, T., Abdolrahim, K., de Mathelin, M., Gangi, A., Padoy, N., 2018. 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*.
- Sun, B., Feng, J., Saenko, K., 2016. Return of frustratingly easy domain adaptation, in: *Proceedings of the AAAI Conference on Artificial Intelligence*.
- Sun, K., Xiao, B., Liu, D., Wang, J., 2019a. Deep high-resolution representation learning for human pose estimation, in: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5693–5703.
- Sun, X., Shang, J., Liang, S., Wei, Y., 2017. Compositional human pose regression, in: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2602–2611.
- Sun, Y., Tzeng, E., Darrell, T., Efros, A.A., 2019b. Unsupervised domain adaptation through self-supervision. *arXiv preprint arXiv:1909.11825*.
- Tarvainen, A., Valpola, H., 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems* 30.
- Tsai, Y.H., Hung, W.C., Schuster, S., Sohn, K., Yang, M.H., Chandraker, M., 2018. Learning to adapt structured output space for semantic segmentation, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7472–7481.
- Tzeng, E., Hoffman, J., Saenko, K., Darrell, T., 2017. Adversarial discriminative domain adaptation, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7167–7176.
- Tzeng, E., Hoffman, J., Zhang, N., Saenko, K., Darrell, T., 2014. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*.
- Wang, D., Shelhamer, E., Liu, S., Olshausen, B., Darrell, T., 2021a. Tent: Fully test-time adaptation by entropy minimization, in: *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*.
- Wang, J., Jin, S., Liu, W., Liu, W., Qian, C., Luo, P., 2021b. When human pose estimation meets robustness: Adversarial algorithms and benchmarks, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11855–11864.
- Wang, K., Zhan, B., Zu, C., Wu, X., Zhou, J., Zhou, L., Wang, Y., 2021c. Triple-uncertainty guided mean teacher model for semi-supervised medical image segmentation, in: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer. pp. 450–460.
- Wang, M., Deng, W., 2018. Deep visual domain adaptation: A survey. *Neurocomputing* 312, 135–153.
- Wang, Q., Fink, O., Van Gool, L., Dai, D., 2022. Continual test-time domain adaptation, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7201–7211.
- Wang, Y., Sun, Y., Liu, Z., Sarma, S.E., Bronstein, M.M., Solomon, J.M., 2019. Dynamic graph cnn for learning on point clouds. *Acm Transactions On Graphics (tog)* 38, 1–12.
- Wang, Y., Zhang, Y., Tian, J., Zhong, C., Shi, Z., Zhang, Y., He, Z., 2020. Double-uncertainty weighted method for semi-supervised learning, in: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer. pp. 542–551.
- Wu, W., Qi, Z., Fuxin, L., 2019. Pointconv: Deep convolutional networks on 3d point clouds, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9621–9630.
- Xiao, B., Wu, H., Wei, Y., 2018. Simple baselines for human pose estimation and tracking, in: *Proceedings of the European conference on computer vision (ECCV)*, pp. 466–481.
- Xu, M., Ding, R., Zhao, H., Qi, X., 2021. Paconv: Position adaptive convolution with dynamic kernel assembling on point clouds, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3173–3182.
- Yang, F., Li, R., Georgakis, G., Karanam, S., Chen, T., Ling, H., Wu, Z., 2020. Robust multi-modal 3d patient body modeling, in: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer. pp. 86–95.
- Yang, J., Shi, S., Wang, Z., Li, H., Qi, X., 2021. St3d: Self-training for unsupervised domain adaptation on 3d object detection, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10368–10378.
- Yang, W., Ouyang, W., Wang, X., Ren, J., Li, H., Wang, X., 2018. 3d human pose estimation in the wild by adversarial learning, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5255–5264.
- Yin, Y., Robinson, J.P., Fu, Y., 2022. Multimodal in-bed pose and shape estimation under the blankets, in: *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 2411–2419.
- Yu, L., Wang, S., Li, X., Fu, C.W., Heng, P.A., 2019. Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation, in: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer. pp. 605–613.
- Zhang, J., Qi, L., Shi, Y., Gao, Y., 2022. Generalizable model-agnostic semantic segmentation via target-specific normalization. *Pattern Recognition* 122, 108292.
- Zhang, Z., Hu, L., Deng, X., Xia, S., 2020. Weakly supervised adversarial learning for 3d human pose estimation from point clouds. *IEEE transactions on visualization and computer graphics* 26, 1851–1859.
- Zheng, H., Motch Perrine, S.M., Pitirri, M.K., Kawasaki, K., Wang, C., Richtsmeier, J.T., Chen, D.Z., 2020. 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*, Springer. pp. 802–812.
- Zheng, Z., Yang, Y., 2021. Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation. *International Journal of Computer Vision* 129, 1106–1120.

- Zhou, Q., Feng, Z., Gu, Q., Cheng, G., Lu, X., Shi, J., Ma, L., 2022. Uncertainty-aware consistency regularization for cross-domain semantic segmentation. *Computer Vision and Image Understanding*, 103448.
- Zhou, X., Huang, Q., Sun, X., Xue, X., Wei, Y., 2017. Towards 3d human pose estimation in the wild: a weakly-supervised approach, in: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 398–407.
- Zhu, J.Y., Park, T., Isola, P., Efros, A.A., 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks, in: *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232.
- Zou, L., Tang, H., Chen, K., Jia, K., 2021. Geometry-aware self-training for unsupervised domain adaptation on object point clouds, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6403–6412.
- Zou, Y., Yu, Z., Kumar, B., Wang, J., 2018. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training, in: *Proceedings of the European conference on computer vision (ECCV)*, pp. 289–305.