Cross-Domain Self-Supervised Complete Geometric Representation Learning for Real-Scanned Point Cloud Based Pathological Gait Analysis

Xiao Gu, Yao Guo, Guang-Zhong Yang*, *Fellow, IEEE*, and Benny Lo*, *Senior Member, IEEE*

***Abstract*—Accurate lower-limb pose estimation is a prerequisite of skeleton based pathological gait analysis. To achieve this goal in free-living environments for long-term monitoring, single depth sensor has been proposed in research. However, the depth map acquired from a single viewpoint encodes only partial geometric information of the lower limbs and exhibits large variations across different viewpoints. Existing off-the-shelf three-dimensional (3D) pose tracking algorithms and public datasets for depth based human pose estimation are mainly targeted at activity recognition applications. They are relatively insensitive to skeleton estimation accuracy, especially at the foot segments. Furthermore, acquiring ground truth skeleton data for detailed biomechanics analysis also requires considerable efforts. To address these issues, we propose a novel cross-domain self-supervised complete geometric representation learning framework, with knowledge transfer from the unlabelled synthetic point clouds of full lower-limb surfaces. The proposed method can significantly reduce the number of ground truth skeletons (with only 1%) in the training phase, meanwhile ensuring accurate and precise pose estimation and capturing discriminative features across different pathological gait patterns compared to other methods.**

***Index Terms*—Gait Analysis, Pose Estimation, Self-Supervised Learning, Point Cloud Completion, Depth Images.**

I. Introduction

Gait analysis is an important tool for investigating the relationship between biomechanical parameters of lower limbs and their associated neurological/musculoskeletal disorders [1], [2]. For pathological gait analysis, accurate and precise three-dimensional (3D) skeleton extraction of lower limbs is a prerequisite for detecting subtle changes of gait abnormalities.

Hitherto, it has received considerable attention to develop unobtrusive gait analysis systems without complex laboratory settings. In practice, these systems have evolved from manual video annotations or complex infra-red motion capture (Mocap) systems, to pervasive wearable or vision sensors [3]. The advances in wearable sensors have enabled the lower limb movement capture outside laboratory-based contexts [4]. However, extracting reliable and unbiased kinematics information from raw sensory data is still challenging due to sensory/environmental noise, contextual/individual difference, etc. [5], [6].

Thanks to recent progresses in 2D human pose detection/tracking based on color images/videos [7], [8], [9], markerless human motion capture for gait analysis with a single vision sensor has received increasing interest [10]. However, the estimation of 3D skeletons form RGB images is still an open problem [11]. To this end, existing works have utilized anatomical, geometric, kinematic or temporal regularization to resolve the ambiguities in lifting 2D poses to 3D [11]. The prior knowledge underpinning these algorithms is derived from normal motion datasets (e.g., Human3.6M and COCO). They may probably fail in detecting subtle pathological gait changes because of the lack of generalization capability [12], [10].

Compared with RGB images, depth maps acquired from commercial depth sensors (RGB-D cameras), could capture additional, albeit noisy and incomplete, 3D geometric information from visible surfaces. Existing literature [13], [14] has reported preliminary attempts of applying off-the-shelf 3D pose trackers embedded in Kinect for clinical gait analysis. It was observed that, for some spatiotemporal parameters such as step length and width, high agreement with Mocap systems could be achieved at specific viewpoints [14]. However, researchers have emphasized the necessity of improved posture tracking for accurate joint angle estimation [14].

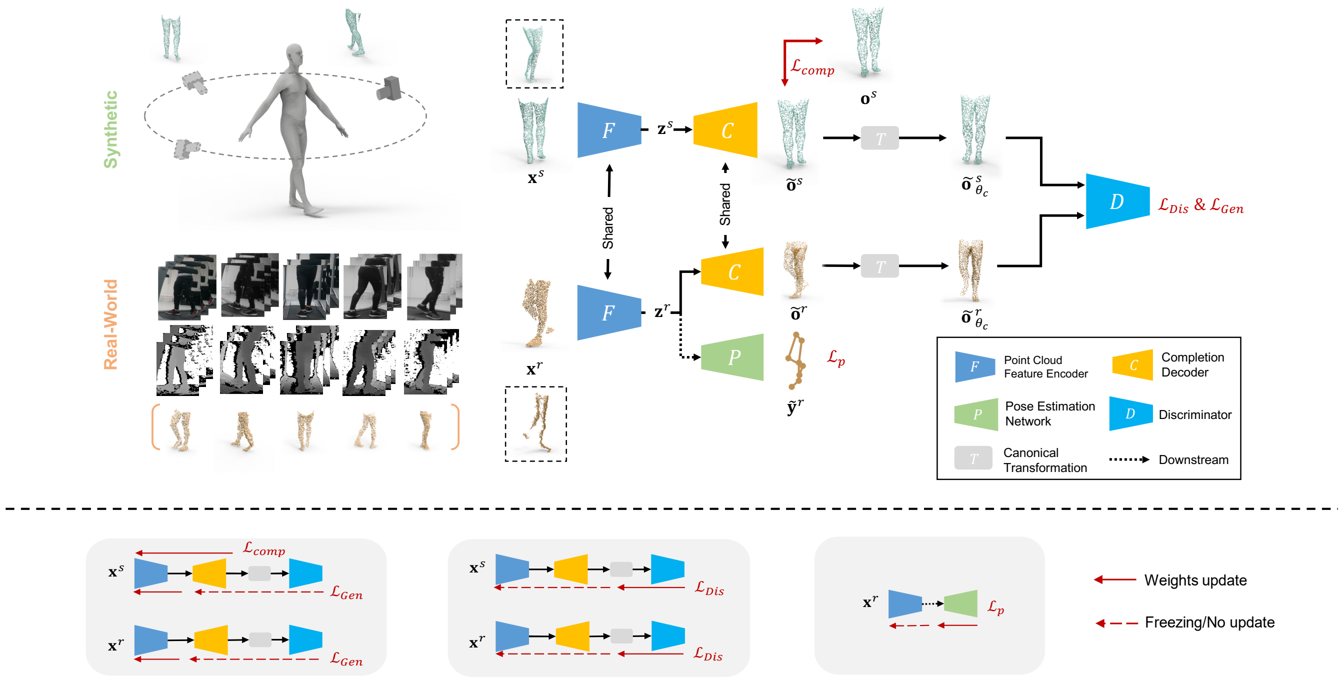
Thus far, numerous models and algorithms [15], [16], [17], combined with several public datasets [15], [18], have been developed to extract accurate 3D skeletons from depth maps/voxel grids/point clouds, paving the way for 3D human motion analysis. However, one of the remaining challenges is that only incomplete and noisy geometric information is directly encoded in depth maps, leading to large variations and incompleteness across views [19].

Furthermore, most existing datasets and algorithms for depth-based pose estimation are designed for activity recognition [15]. They pose less demanding requirements on the accuracy of biomechanical indices and pay less attention to lower limb joints. The ground truth skeletons in most datasets are either generated from (semi-)manual annotations or provided by the trajectories of attached markers. They may lack consistency across annotators or not satisfy the requirements for clinical diagnosis [20]. This can be potentially solved by training on large-scale clinical datasets collected from Mocap systems under standard clinical practices; however, it is

This work was supported by a Newton Fund grant ID:527636592 funded by the UK Department of Business, Energy and Industrial Strategy (BEIS) and delivered by the British Council, and also supported by the Science and Technology Commission of Shanghai Municipality under Grant 20DZ2220400, and the Interdisciplinary Program of Shanghai Jiao Tong University under Grant YG2021QN117. (Corresponding author: Guang-Zhong Yang and Benny Lo)

X. Gu and B. Lo are with the Hamlyn Centre, Imperial College London, London SW7 2AZ, UK (e-mail: {xiao.gu17, benny.lo}@imperial.ac.uk).

Y. Guo and G.-Z. Yang are with the Institute of Medical Robotics, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: {yao.guo, gzyang}@sjtu.edu.cn).



(a) Pretext-Generator (Section IV-B) (b) Pretext-Discriminator (Section IV-C) (c) Downstream (Section IV-D)

Fig. 1. Illustration of our proposed self-supervised learning framework. In the pretext task (a)(b), synthetic (partial and complete pairs) and realistic (only partial) data from multiple viewpoints are available for training. Firstly, feature encoder F and completion decoder C are applied to transform incomplete lower-limb point clouds $\mathbf{x}^s \& \mathbf{x}^r$ to the complete $\tilde{\mathbf{o}}^s \& \tilde{\mathbf{o}}^r$, respectively. Supervised completion loss $\mathcal{L}_{comp}$ is applied between $\tilde{\mathbf{o}}^s$ and $\mathbf{o}^s$. Meanwhile, to mitigate the domain shift between real-scanned and synthetic data, adversarial training is performed on the completed point clouds after canonical transformation by T, which are $\tilde{\mathbf{o}}^r_{\theta_c}$ and $\tilde{\mathbf{o}}^s_{\theta_c}$. Adversarial training is performed by implementing step (a) and (b) in an alternative way, where $\mathcal{L}_{Dis}$ enforces D to discriminate $\tilde{\mathbf{o}}^r_{\theta_c}$ and $\tilde{\mathbf{o}}^s_{\theta_c}$ in (b), whereas $\mathcal{L}_{Gen}$ enforces F to generate $\tilde{\mathbf{o}}^r_{\theta_c}$ and $\tilde{\mathbf{o}}^s_{\theta_c}$ that confuse D. Subsequently, in the downstream task (c), the derived geometric feature is fed into to a pose regression sub-network P to estimate lower-limb skeletons. The point cloud shown in the dashed rectangle is a rotated one from the original view for better visualization of the incompleteness. Please refer to Section IV for more details.

extremely expensive and laborious to perform such large-scale data collection [1]. Therefore, minimizing the efforts devoted to acquiring ground truth 3D skeleton data is critical in developing practical accessible gait analysis solutions.

In this paper we propose a novel cross-domain self-supervised complete geometric representation learning framework for 2.5-dimensional (2.5D) real point clouds converted from depth maps, with the help of unlabelled synthetic point clouds, as shown in Fig. 1. It uses synthetic partial-complete point cloud pairs to learn complete geometric representation of lower limbs. Simultaneously, a view-invariant adversarial training strategy is utilized, which aligns the representations between real and synthetic domains. The proposed method can reduce the number of ground truth real skeletons in the training phase, yet enable accurate and precise pose estimation for pathological gait recognition.

The main contributions of our paper are three-fold:

Self-Supervised Geometric Representation Learning: We develop a self-supervised method aimed at incomplete real-scanned point clouds. It can generate full lower-limb surface point sets from incomplete inputs, thus deriving complete geometric representations. This method can effectively reduce the need for labelled real data yet achieve good performance for pose estimation.

View-Invariant Domain Adaptation: We propose a view-invariant domain adaptation strategy between realistic and synthetic 2.5D point clouds. Compared to the vanilla adversarial training strategy [21], it is demonstrated that the heterogeneity gap across domains can be better handled after canonical transformation.

Discriminative Pathological Gait Analysis: The whole framework overcomes the noises inherent in low-cost RGB-D cameras and captures the subtle changes across different abnormal gait patterns. It is validated on our self-collected gait dataset and demonstrates promising recognition results.

II. Related Work

A. Depth Based Pose Estimation

Depth based 3D human pose estimation algorithms are either discriminative or generative [22]. Generative approaches are based on model-driven optimization, which aim to fit an explicit deformable body model to input depth images by minimizing specialized cost functions. They are either based on non-parameterized point cloud registration methods like iterative closest point algorithms [23], or parameterized ones like Gaussian mixture models [24].

Discriminative approaches directly infer human poses from depth maps, which optimizes a computational model by data-driven training. Among them, conventional methods applied machine learning models, mainly random forests, to learn the mapping from depth maps to key joints [25], [26], [27].

Recently, several deep learning architectures [28], [29], [22], [16] have been proposed based on different representations of depth data. Most of these proposed methods focus on hand pose estimation, with the potential of being applied to human body pose estimation as well. Among them, 2D convolution is performed on 2D depth images [28], [16] whereas 3D convolution is applied on the 3D volumetric representations [29],

[22]. Although Moon *et al.* [22] achieved superior performance for both hand and human pose estimation, the low resolution of voxelization would affect the estimation precision whereas it is computationally expensive to perform 3D convolution on high-resolution voxels [30]. In this paper, we target at point cloud based representations, which also differ from existing point-cloud-based pose estimation work as discussed in Section II-B.

B. Deep Learning on Point Clouds

Deep learning based point cloud analysis has received increasing attention due to the emerging solutions for unordered point sets, like PointNet [31] and its variants/extensions [32], [33], [34]. These architectures have been successfully applied to a variety of tasks, such as 3D shape recognition, shape completion [35], and object detection [34].

Recently, several methods have been proposed to estimate hand/human skeletons based on point sets [36], [37], [38], [19], [17]. Ge *et al.* [36] applied PointNet++ [32] based structure to model observable hand surfaces for pose regression. Li *et al.* [38] proposed a point-to-pose voting scheme to perform pose estimation from the weighted fusion of each point. Different from these papers which focused on exploring advanced network architectures, we focus on minimizing the needs for labelled data while achieving satisfactory pose estimation performance. Chen *et al.* [37] shared this motivation by proposing a semi-supervised training strategy to learn geometric representations with an autoencoder. Our work is inspired by this approach, yet is conceptually different. We aim to learn full geometric features via self-supervised learning and meanwhile leverage synthetic data to achieve adversarial training with real partial scans.

C. Domain Adaptation

Domain adaptation (DA) in homogeneous settings targets to mitigate the data distribution heterogeneity across different domains, where data from different sources are of the same feature space yet different distributions. It plays an important role in effective learning from synthetic data and subsequently addressing real-world tasks [39], [40]. The main categories of existing domain adaptation methods are discrepancy-, adversarial- or reconstruction-based, aligning the distribution in the embedded feature or low-level data space [21].

For point clouds, existing analysis methods are mostly focused on synthetic benchmarks (e.g., ShapeNet[1]) or real datasets only, and there is as yet a paucity of research focused on the adaptation between real and synthetic domains. In fact, there exist major differences between these two. The point sets from synthetic models are complete and clean, whereas those converted from depth scanning are incomplete and noisy. Recently, Chen *et al.* [41] proposed an adversarial strategy for shape completion of real-world depth scans. Different from its ultimate goal of getting complete point sets, we aim to derive clean and complete geometric representation from the latent space and it is expected to work consistently across views.

D. Self-Supervised Learning

The advent of self-supervised learning has enabled effective feature learning over the course of training pretext tasks on self-generated labels. The learned feature representation can facilitate faster convergence or reduce over-fitting when labels are limited on downstream tasks. Till now, several methods for 2D visual feature learning have been developed, such as image inpainting/completion, jigsaw puzzle, or geometric transformation [42]. Especially, for human pose estimation from 2D RGB images, multiview consistency has been exploited as an effective self-supervised constraint during 2D-3D lifting [43]. Meanwhile, some recent work has developed strategies for point cloud analysis based on self-supervised reconstruction tasks [44], [45]. Using synthetic imagery to facilitate self-supervised learning for real-world tasks has been proposed in research like [39]. Different from it, our work proposes a solution to simultaneously handling the domain gap and view variations, when transferring knowledge learned from self-supervised learning in the synthetic domain.

III. Experiments

A. Real Pathological Gait Dataset

1) Experimental Settings: Our experiment was approved by Imperial College Research Ethics Committee (ICREC-18IC4915), following the standard biomechanics workflow. The experimental settings are shown in Fig. 2. The recruited 8 subjects were instructed to walk on a treadmill, with reflective markers attached to the anatomical landmarks of lower limbs. We applied the conventional gait model (CGM https://pycgm2.netlify.app/) where 28 reflective markers were used, as shown in Fig. 3(b). They were recorded by Vicon Mocap system with a sampling rate of 100Hz. Meanwhile, a RGB-D camera, RealSense D435 (Intel Corporation, California, US), was used to simultaneously take RGB and depth maps with a sampling rate of around 20Hz. It was placed on a tripod, and five viewpoints facing towards the subject were selected as tripod placement options, annotated in Fig. 2. We did not set strict requirements for the camera localization and orientation, which means that the camera extrinsic parameters would slightly differ across each trial of the same viewpoint. For each trial, only one RGB-D camera was used to avoid the mutual interference caused by multiple light projections. The RGB-D camera and the Mocap system were synchronized by the broadcast UDP signals from the Mocap system[2].

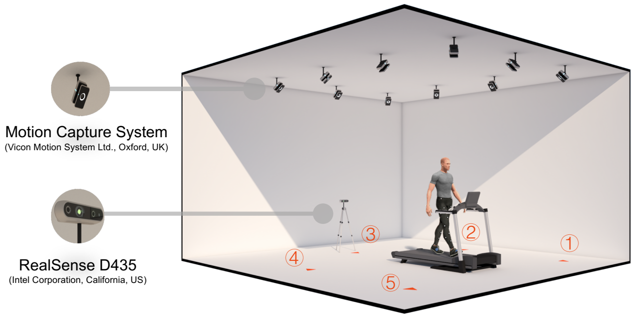
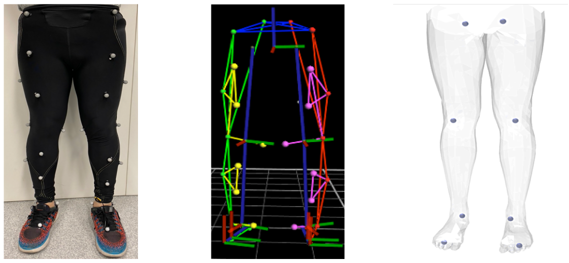

Motion Capture System
(Vicon Motion System Ltd., Oxford, UK)
RealSense D435
(Intel Corporation, California, US)
①
②
③
④
⑤

Fig. 2. Experimental settings for data collection. Reflective markers based on the conventional gait model were attached on the lower limbs of the subject. Motion capture system captures the 3D trajectories of these attached markers. Meanwhile, a tripod with a RGB-D camera embedded on the top was placed at one of the five annotated positions to simultaneously record RGB-D images. (Human animation is modified from https://optitrack.com/)

[1] https://shapenet.org/

[2] We followed the official procedure documented in https://docs.vicon.com/display/Nexus210/Automatically+start+and+stop+capture



(a) Subject (b) CGM (c) SMPL

Fig. 3. Illustration of (a) Real Subject (b) Conventional Gait Model (c) SMPL Model, and corresponding joint positions. There exists difference in the joint positions relative to the limb surface.

In our experiments, six different walking patterns were investigated. They are normal, supination, pronation, leg-length-discrepancy (LLD), toe-in and toe-out respectively, following the experimental settings of our previous research [6]. Supination/pronation was realized by edge-wedged orthotic insoles [46] reflecting outwarded/inwarded ankle positions. LLD was simulated by adding heel-lift insoles similar to [47]. Toe-in/toe-out refers to subjects walking with toes pointing inwards/outwards [48]. Totally, for each subject, around 30 trials (5 views×6 patterns) with a duration of 30s each were conducted.

2) Data Preparation: With the recorded 3D positions of reflective markers, the embedded system can generate accurate joint localizations (Hip, Knee, Ankle and Knee) and kinematics via dynamic calibration, as shown in Fig. 3(b). The extrinsic camera parameters were derived by minimizing the reprojection error between the manually annotated 2D and extracted 3D localization of visible markers [49], [50]. Afterwards, the ground truth 3D keypoints for depth images under local camera frames can be acquired by the transformation.

Regarding preprocessing, we firstly applied the state-of-the-art human parsing algorithm Cross-Domain Complementary Learning (CDCL) [51] to generate lower-limb masks based on associated RGB and then lower-limb depth maps were converted to point clouds via pre-calibrated intrinsic parameters. They were downsampled to 2048 points each.The whole real dataset is composed of around 146,300 frames, evenly distributed per subject/condition/view. Further demographic details are given in the Supplementary Material.

B. Synthetic Gait Dataset

1) Synthetic Human Model: We applied SMPL (Skinned Multi-Person Linear Model) [52] to generate synthetic data based on the kinematics extracted from Mocap data. SMPL is a realistic articulated human model, parameterized by shape $\{\beta_i\}$ (body deformations) and pose $\{\theta_j\}$ (skeleton kinematics) parameters. Because of the realism and accessibility of this model, it has been widely used to facilitate RGB based human body part segmentation [53], human pose estimation [54] and human mesh recovery [55].

2) Synthetic Data Generation: The kinematics derived from our real-world training data based on CGM were transferred to SMPL $\{\theta_j\}$ to simulate different gait types. Meanwhile, $\{\beta_i\}$ parameters available from CAESAR dataset [56], [53] were adopted to generate varied but realistic body shapes. For each set of $\{\theta_j\}$, the camera was placed in random positions around the subject to generate point clouds based on the given kinematics. We applied Hidden Point Removal [57] to simulate the incompleteness. Subsequently, the point sets belonging to lower limbs were segmented as data of interest and sampled to a fixed number (2048). The corresponding complete point sets of lower limbs were also derived.

For each subject, we generated on average 4000 frames with corresponding kinematics available in the training/testing set and randomly selected shapes. It should be noted that there exists difference between our gait model (Fig. 3(b)) and the synthetic human model (Fig. 3(c)) in terms of their relative keypoint positions to the lower limb surface. In our proposed method, the keypoints in the synthetic dataset are not used for supervised training of pose estimation. Instead, a self-supervised strategy for synthetic data is exploited, which is further explained in Section IV-B.

IV. Methods

A. Objective and Method Overview

Extracted from depth maps, the observed point clouds $\mathbf{x} \in \mathbb{R}^{n\times 3}$ are incomplete/partial. We denote its corresponding complete point cloud (a point set sampled from the whole lower limb surface) as $\mathbf{o} \in \mathbb{R}^{n\times 3}$. Along with the changes of RGB-D camera positions and orientations, the observed incomplete point clouds would shape differently. Meanwhile, for each $\mathbf{x}$, its corresponding ground truth skeleton is denoted as $\mathbf{y} \in \mathbb{R}^{k\times 3}$, where k refers to the number of lower-limb keypoints involved. The superscription r or s of above symbols, if exists, refers to realistic or synthetic data respectively.

As mentioned above, it is extremely expensive and laborious to collect large-scale datasets with ground truth 3D skeletons. The objective of this paper is to derive a computational model $\mathcal{G}$: $\mathcal{X} \mapsto \mathcal{Y}$ that enables the estimation of lower-limb pose $\mathbf{y}^\mathbf{r}$ from the real-world incomplete point cloud $\mathbf{x}^\mathbf{r}$ captured with a single depth camera, and more importantly, minimizes the number of labels during training. To achieve this goal, a cross-domain self-supervised geometric representation learning framework is proposed by leveraging the unlabelled synthetic data, as shown in Fig. 1.

Firstly, with the access to synthetic partial-complete point cloud pairs, a self-supervised point cloud completion network, composed of feature encoder and completion decoder, is applied to learn full geometric representations. Subsequently, to deal with the heterogeneity gap between synthetic and real data as well as the variations caused by noises, a view-invariant adversarial network is cascaded with the canonical completed point clouds as input. The embedded latent feature can therefore be aligned between real and synthetic data through the optimization of the feature encoder. Afterwards, another pose estimation network is trained with the learned latent feature as the input for our downstream pose estimation task.

B. Point Cloud Completion: Self-Supervised Learning

2.5D point clouds only contain geometric information from the visible surfaces, which tend to shape differently under different gait kinematics and view angles. To deal with the limitations caused by incomplete point clouds, one ideal solution is to first derive the full geometric information. Inspired by existing self-supervised learning research on 2D images

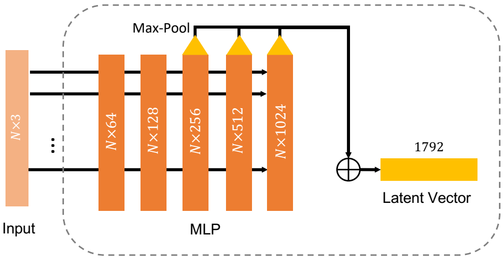


Fig. 4. The Combined Multi-Layer Perception (CMLP) module proposed in [35]. Compared to original MLP series layers, in our implementation, this architecture extracts the max-pooled values of the last three layers and subsequently combines them into a global latent vector. ⊕ refers to concatenation.

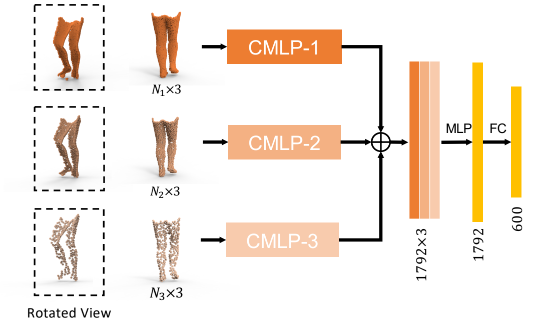


Fig. 5. The detailed architecture for feature encoder F. The point sets are sampled to three resolutions $N_1 = 2048, N_2 = 1024, N_3 = 512$, and put into each CMLP branch to generate its corresponding latent vector. Three latent vectors are subsequently concatenated together to form a 1792×3 feature map. Subsequently, a MLP [3-1] is applied to convert the feature map to a 1792 latent vector, with a fully-connected layer cascaded to generate the final latent vector with a length of 600. Point clouds in dashed rectangles are the rotated ones from original views for better visualizations of incompleteness. ⊕ refers to concatenation.

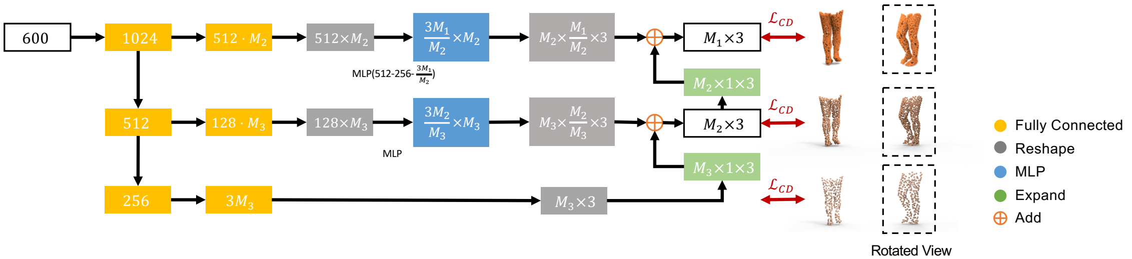


Fig. 6. The detailed architecture for completion C or reconstruction R. Firstly, the bottom branch generates sparse point sets composed of M_3 points based on a series of fully connected layers. Subsequently, the middle branch predicts $\frac{M2}{M3}$ relative coordinates to each point in the generated M_3 point sets, thus deriving totally M_2 points. Then, denser M_1 points are generated in the top branch by using the same strategy. ⊕ refers to adding instead of concatenation.

like image inpainting/completion [58], we build a function $\mathcal{F} : \mathcal{X} \mapsto \mathcal{Z}$ that can derive the full representation $\mathbf{z}$ in the latent space from the input $\mathbf{x}$. Subsequently another function $\mathcal{C} : \mathcal{Z} \mapsto \mathcal{O}$ is utilized to recover the full point sets $\mathbf{o}$. Once the full geometric information is extracted by $\mathcal{F}$, another pose estimation function $\mathcal{P}$ can be applied for pose regression from $\mathbf{z}$. Therefore, we have $\mathcal{G} = \mathcal{F} \circ \mathcal{P}$.

However, in real world applications, it is challenging to acquire $\mathbf{o}^r$ for dynamic lower limbs via existing 3D reconstruction techniques [59]. Even with multiple cameras positioned around, some surface areas cannot be captured due to self-occlusion and the noises in depth sensors tend to have profound effects on the reconstruction quality. Fortunately, this issue does not hold for the synthetic model, since it is easy to acquire paired $\{\mathbf{x}^s, \mathbf{o}^s\}$. The full representation $\mathbf{z}^s$ of $\mathbf{x}^s$ can be extracted after the optimization of $\mathcal{F} \circ \mathcal{C}$.

1) Network Architecture: In the practical deployment, an autoencoder-resembling architecture composed of a feature encoder F and a completion decoder C is applied. Based on the state-of-the-art point cloud completion network PF-Net [35], we adopt a hierarchical feature learning architecture consisting of multiple branches with point sets sampled from multiple resolutions as input, as shown in Fig. 7. In detail, the point cloud is sampled with iterative farthest point sampling (IFPS) [32]. The branch for feature extraction of each resolution adopts a structure modified from PointNet [31], as shown in Fig. 4. It extracts the maxpooled features from the last several Multi-Layer Perceptron (MLP) layers and merges them into a global latent feature vector, referred to as a Combined MLP (CMLP). Afterwards, the global latent vectors from different resolutions are concatenated and go through another MLP layer to form the final latent vector. Such an architecture has demonstrated the superior performance in deriving geometric structures from incomplete point clouds [35].

On the other hand, for the completion decoder C, the completed point cloud is generated in a hierarchical multi-resolution fashion as proposed in [35]. Instead of directly using fully-connected layers or folding-based layers [60], the adopted architecture is based on "add" and "expansion" operations, progressively generating points from sparse to dense (from M3 to M1). The expected output of C is the full point set $\mathbf{o}$. Details can be found in the Supplementary Material.

2) Loss Functions: We apply the Chamfer Distance Loss $\mathcal{L}_{CD}$ to measure the difference between the recovered $\tilde{\mathbf{o}}^s$ and ground truth $\mathbf{o}^s$, which is formulated as below,

$$\begin{aligned}\mathcal{L}_{CD}(\mathbf{o_1}, \mathbf{o_2}) = &\frac{1}{|\mathbf{o_1}|} \sum_{p_1 \in \mathbf{o_1}} \min_{p_2 \in \mathbf{o_2}} \|p_1 - p_2\|_2^2 \\ &+ \frac{1}{|\mathbf{o_2}|} \sum_{p_2 \in \mathbf{o_2}} \min_{p_1 \in \mathbf{o_1}} \|p_1 - p_2\|_2^2\end{aligned} \tag{1}$$

The hierarchical multi-resolution point cloud completion loss $\mathcal{L}_{comp}$ is defined as below,

$$\begin{aligned}\mathcal{L}_{comp} = &\mathcal{L}_{CD}(C(F(\mathbf{x}^s))_{M_1}, \mathbf{o}^s_{M_1}) \\ &+ \lambda_{M_2}\mathcal{L}_{CD}(C(F(\mathbf{x}^s))_{M_2}, \mathbf{o}^s_{M_2}) \\ &+ \lambda_{M_3}\mathcal{L}_{CD}(C(F(\mathbf{x}^s))_{M_3}, \mathbf{o}^s_{M_3})\end{aligned} \tag{2}$$

where λ_{M_2} and λ_{M_3} are weights of the Chamfer Distance loss for the intermediate sparser point clouds (point sets with point number M2 and M3 respectively). The optimization goal is to minimize $\mathcal{L}_{comp}$.

C. View-Invariant Real-Synthetic Domain Adaptation

Although F & C trained on synthetic $\{(\mathbf{x}^s, \mathbf{o}^s)\}$ pairs allow full geometric feature extraction, there still exist gaps when handling real world point clouds, which originate from the noise, clothing deformation, motion artifacts, etc. We aim to diminish the real-synthetic domain shift, thus enabling a good generalization capability of trained models F & C on real partial scans $\mathbf{x}^r$. Therefore, domain adversarial training is introduced to align distributions across real and synthetic domains. We do not directly attempt to align the embedded feature $\mathbf{z}^s$ and $\mathbf{z}^r$, which would be largely varied due to different angled viewpoints in both real and synthetic datasets. Performing adversarial training on this latent space would cause unnecessary disturbance due to viewpoint heterogeneity, especially in our case that $\mathbf{z}$ is finally used for pose regression rather than a certain classification task [61].

Instead, we perform adversarial training on the canonical form of the completed point cloud $\tilde{\mathbf{o}}^r = C(F(\mathbf{x}^r))$ and $\tilde{\mathbf{o}}^s = C(F(\mathbf{x}^s))$, as shown in Fig. 1. Based on the ground truth[3] body orientations under local camera frames, a transformation matrix ${}^{\theta_c}T_\theta$ is applied, where θ denotes the orientation of lower limbs under current camera frames, while θ_c denotes the pre-defined canonical frame. To derive a clean as well as complete point cloud representation $\mathbf{z}^r$ from $\mathbf{x}^r$, only the parameters of feature encoder F are optimized during the adversarial training in case that the domain adaptation capability would be mastered by C. Over the course of adversarial training, $\mathcal{L}_{comp}$ of synthetic data is also added to ensure the completion task.

1) Network Architecture: The discriminator is designed based on PointNet, where the completed point clouds encompass a series of MLP layers and the features from the last three layers are extracted by maxpooling, forming a global latent feature vector. Subsequently, shallow fully connected layers are applied to discriminate between real and synthetic.

2) Loss Functions: In the adversarial training, Wasserstein GAN with gradient penalty (WGAN-GP) [62], [63] is applied, the loss functions of which are formulated as below,

$$\begin{cases} \mathcal{L}_{Dis} = \mathbb{E}_{\mathbf{x}^r}[D({}^{\theta_c}T_\theta \cdot C(F(\mathbf{x}_\theta))] \\ \qquad - \mathbb{E}_{\mathbf{x}^s}[D({}^{\theta_c}T_{\theta'} \cdot C(F(\mathbf{x}_{\theta'}))] \\ \qquad + \lambda_{gp}\mathbb{E}_{\hat{\mathbf{x}}}[(\|\nabla_{\hat{\mathbf{x}}} D(\hat{\mathbf{x}})\|_2 - 1)^2] \\ \mathcal{L}_{Gen} = -\mathbb{E}_{\mathbf{x}^r}[D({}^{\theta_c}T_\theta \cdot C(F(\mathbf{x}_\theta))] \end{cases} \tag{3}$$

where $\hat{\mathbf{x}}$ are sampled from points of the canonical synthetic and realistic completed point clouds, and λ_{gp} is the weight of gradient penalty during the optimization of D.

Overall, the parameters of F is optimized by the minimization of $\mathcal{L}_{Gen}+\lambda_{comp}\mathcal{L}_{comp}$, while D is optimized alternatively by the minimization of $\mathcal{L}_{Dis}$.

D. Pose Estimation Supervised Training

To this end, the feature extractor F is able to handle real partial noisy point clouds and derive clean full geometric representations. Therefore, a pose regression model P consisting of three fully-connected layers is applied to perform pose estimation from $\mathbf{z}$ after optimizing $F\&C$.

$$\mathcal{L}_p = \mathbb{E}[\|P(F(\mathbf{x}^r)) - \mathbf{y}^r\|_2^2] \tag{4}$$

V. Results and Discussion

A. Implementation Details

In the experiment, we applied the leave-one-subject-out (LOSO) cross validation paradigm, where in each session 7 subjects were selected for training and validation (90% & 10%), and the held-out 1 subject testing. To simulate the real-world scenario where raw data can be easily acquired with RGB-D camera while ground truth (GT) skeleton is difficult to obtain, we evaluate the performance of our algorithms when labelled data is scarce. In practice, we only used the GT skeleton of 1% real data (around 1%×125k=1.25k). On the other hand, as it is easy to collect synthetic data and obtain its labels, we used all the ground truth skeleton when needed (4k×7=28k). It is also noteworthy that the synthetic data was generated with varied shapes and viewpoints based on the kinematics from the aforementioned 1% real data. Other details can be found in the Supplementary Material.

In summary, in each LOSO session, extracted from the 7 training subjects, 1% Mocap GT, 100% raw real data, as well as synthetic data based on 1% Mocap GT were used for training. The fully trained model was then applied to test the held-out 1 subject. Since the Mocap data of the held-out subject was only used to evaluate the performance, this experiment setting can validate the feasibility of conducting gait analysis outside complex laboratory settings, as well as minimizing the efforts of acquiring Mocap data for model pre-training.

B. Self-Supervised Point Cloud Completion

1) Qualitative Results: The completed point clouds from $\mathbf{x}^r$ are visualized in Fig. 7. It should be noted that because of the variance (e.g. cloth, shoe, noise, etc.) between real and synthetic data, the "complete" point clouds cannot recover some realistic details. However, it can capture the complete geometric information for pose estimation as discussed later.

2) JSD Metric Comparison: To compare the similarity between different point clouds, we utilized the Jensen-Shannon Divergence (JSD) metric as in [64], [63], which measures the JSD of their marginal distributions in the Euclidean 3D space. All the completed point clouds are firstly transformed to the canonical frame before measuring. The results are displayed in Table I. The JSD between $\tilde{\mathbf{o}}^s$ and $\tilde{\mathbf{o}}^r$ is decreased from

TABLE I
COMPARISON OF JSD METRICS FOR DIFFERENT COMPLETE POINT CLOUDS. $\tilde{\mathbf{o}}^{r*}$ DENOTES THE COMPLETED REAL POINT SETS WITHOUT DOMAIN ADAPTATION. HIGHER SIMILARITY HAS SMALLER JSD VALUES.

	$\tilde{\mathbf{o}}^{r*}$	$\tilde{\mathbf{o}}^r$	$\tilde{\mathbf{o}}^s$	$\mathbf{o}^s$
$\tilde{\mathbf{o}}^{r*}$	-	0.0221	0.0362	0.2710
$\tilde{\mathbf{o}}^r$	0.0221	-	0.0197	0.2546
$\tilde{\mathbf{o}}^s$	0.0362	0.0197	-	0.2543
$\mathbf{o}^s$	0.2710	0.2546	0.2543	-

[3] For synthetic data, the ground truth orientation can be easily derived; for the real gait dataset, the orientation is roughly provided by the orientation of the treadmill by pre-calibrated camera localization relative to the treadmill with reflective markers.

TABLE II
COMPARISON OF DIFFERENT METHODS WITH ONLY 1% GROUND TRUTH OF REAL DATA IS UTILIZED DURING TRAINING.

Methods	3D Euclidean Distance Error (cm)↓				3D Angle Error (degree)↓			Classification↑			
	Hip	Knee	Ankle	Toe	Knee	Ankle	Foot-Progress	Prec	Rec	F1	Acc
V2V-Real	4.30±1.82	3.51±1.33	3.90±2.33	3.75±3.00	4.76±4.85	7.33±7.50	13.90±15.34	0.620	0.621	0.619	0.628
A2J	4.90±1.60	4.66±1.98	4.40±1.79	4.54±1.97	5.01±3.77	6.87±4.89	11.89±8.32	0.629	0.600	0.579	0.608
HP	4.83±2.61	4.81±2.34	5.89±2.92	5.80±3.72	6.95±6.01	10.42±8.98	16.98±16.32	0.610	0.559	0.532	0.554
F&P-Real	4.99±1.89	4.45±1.56	4.89±1.87	4.34±2.21	5.83±3.64	6.72±5.33	12.99±10.20	0.573	0.548	0.542	0.552
Self-Recon	3.58±1.45	2.57±1.35	2.98±1.45	3.29±2.42	4.52±3.24	4.25±3.26	8.86±7.66	0.674	0.679	0.671	0.687
V2V-Syn	3.44±1.98	2.86±1.89	2.88±2.04	3.44±3.35	4.37±3.85	6.72±5.84	12.43±16.57	0.713	0.708	0.704	0.712
F&P-Syn	3.84±2.44	3.35±1.67	3.97±2.61	4.34±2.78	5.75±4.69	7.32±6.33	11.45±11.00	0.697	0.685	0.684	0.700
Proposed	**2.78**±1.43	**2.54**±1.36	**2.62**±1.49	**3.01**±1.70	**4.32**±3.51	**3.65**±3.32	**7.84**±6.78	**0.767**	**0.745**	**0.750**	**0.759**

Distance and angular errors are expressed under mean±std format, whereas the average values of the classification metrics are reported.

0.0362 to 0.0197 after domain adaptation, to some extent validating the shift mitigation. Larger JSD between $\tilde{\mathbf{o}}^s/\tilde{\mathbf{o}}^r$ and $\mathbf{o}^s$ demonstrates the limitation of only applying Chamfer Distance loss for the completion task.

C. Lower-Limb Pose Estimation

1) Metrics: In order to evaluate the pose estimation methods, we adopted two commonly used metrics: 3D Euclidean distance error (Hip, Knee, Ankle, Toe) and 3D joint angular error (Knee, Ankle, Foot-Progression, see Supplementary Material) [65].

2) Compared Methods: The following methods were implemented for comparison. The experimental settings are based on their online available source codes. The bounding box, if utilized in the preprocessing, are provided by the lower-limb segmentation mask pre-calculated by CDCL [51]. Implementing details are listed in our Supplementary Material.

- V2V [22]: The Voxel-to-Voxel (V2V) pose prediction network performs 3D convolution on voxels and has achieved the state-of-the-art performance for both hand and human pose estimation. For comparison, two versions are considered, V2V-Real and V2V-Syn. V2V-Real is trained on the given labelled real data. On the other hand, the synthetic ground truth 3D skeletal data $\mathbf{y}^s$ is extracted together with $\mathbf{x}^s$ to form synthetic training pairs. V2V-Syn is firstly trained on $\{(\mathbf{x}^s, \mathbf{y}^s)\}$ and then fine-tuned on available labelled real data.
- Hand-PointNet (HP) [36]: This point set based method was initially proposed for the hand pose estimation task. It applies an architecture similar to PointNet++ [32] to learn hand feature in a hierarchical manner.
- A2J [16]: A2J is an algorithm that predicts joint positions in an ensemble way (anchor to joint). It applies 2D-CNN as the backbone network.

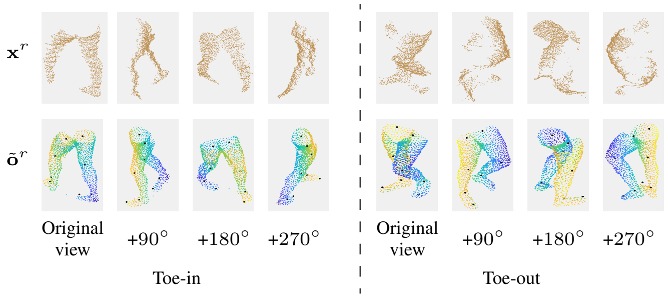


Fig. 7. Visualizations of completed point sets (bottom) from real-scanned data (top). They are rotated around the vertical axis with an increment step of 90° from original-view for visualization. GT skeleton is visualized as black dots in the bottom for reference.

- Self-Recon: We replace the completion C with the reconstruction module R for reconstruction, both of which share the same architecture. It applies self-reconstruction loss in the real-world to learn the 3D geometrical information, which is similar to [37].
- F&P: The feature extractor F together with pose estimation net P are utilized to constitute F&P model. F&P-Real it is trained on available $\{(\mathbf{x}^r, \mathbf{y}^r)\}$ pairs, while F&P-Syn is firstly trained on $\{(\mathbf{x}^s, \mathbf{y}^s)\}$ and then finetuned on available $\{(\mathbf{x}^r, \mathbf{y}^r)\}$ pairs.

We followed the experiment settings for simulating real-world scenarios as illustrated in Section V-A. In this manner, for those train-on-real test-on-real supervised-training methods (V2V-Real, A2J, HP, F&P-Real), only the real data whose labels are available (1%) were used for direct supervised training. For those train-on-syn test-on-real supervised-training ones (V2V-Syn, F&P-Syn), as the GT skeleton can be easily acquired from the synthetic model, we used all the synthetic data (with label) for supervised training and then fine-tuned the model with the real data whose labels are available (1%). For Self-Recon and Proposed, only GT of 1% real was used in the pose estimation task, while all the real training data (w/o label) was used for adversarial training (Proposed) or autoencoder (Self-Recon).

3) Quantitative Results:

a) ***Train-on-real** Versus **Train-on-syn**:* In Table II, we provide the quantitative results of lower-limb pose estimation. As noticed, the train-on-syn test-on-real supervised methods achieve better performance compared to those train-on-real test-on-real counterparts. This validates our motivation of utilizing synthetic data for training when the ground truth data in the real world is difficult to obtain. With access to larger amounts of ground truth labels from the synthetic human model, train-on-syn test-on-real methods can avoid over-fitting to some degree and achieve performance gains.

b) ***Self-Supervised Learning** Versus **Supervised Learning**:* As per our experimental settings, we set available the ground truth skeletons of only 1% of the real-world training data. The remaining 99% unlabelled data were not used in supervised learning strategies. By contrast, our proposed learning strategy can make the most of the rest unlabelled by self-reconstruction loss. Such strategy outperforms those supervised training ones, showing its effectiveness. Especially, compared with F&P-Syn, it can effectively learn geometric

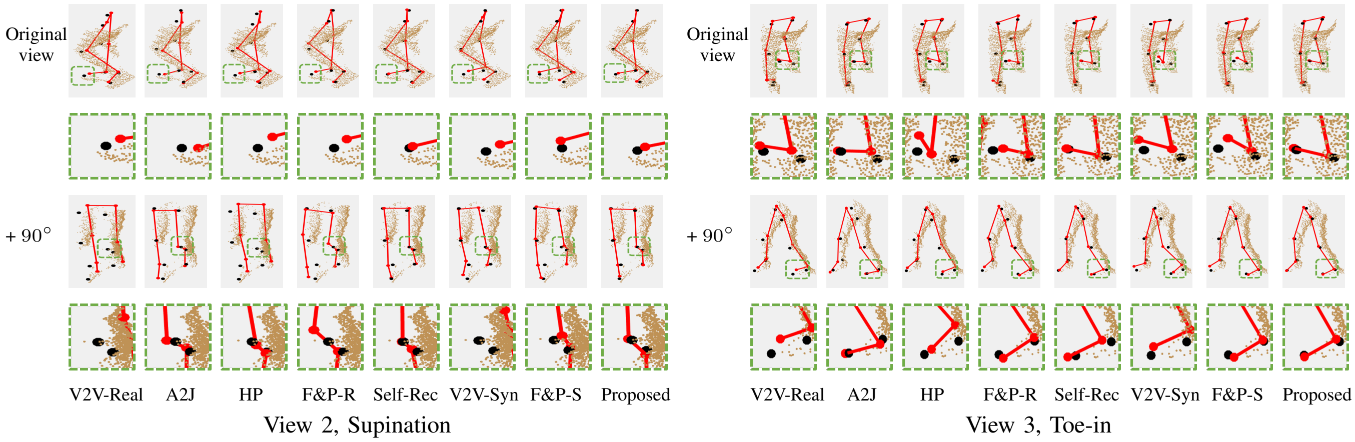
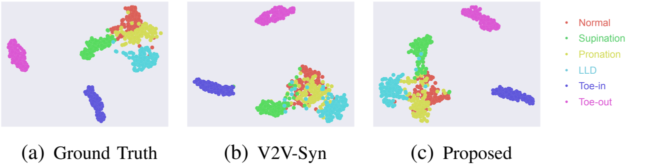


Fig. 8. Visualizations of pose estimation results based on different methods (selective results from {View 2, Supination} and {View 3, Toe-in}). The black points are the ground truth skeletons whereas the red points and lines are the estimation. The green dashed rectangle annotates a region of interest for comparison, the zoomed-in plot of which is displayed underneath. They are rotated around the vertical axis with 90° from original-view for better visualization.



(a) Ground Truth (b) V2V-Syn (c) Proposed

Fig. 9. t-SNE plot of the feature distribution extracted by the classifier in one session of leave-one-subject-out cross validation.

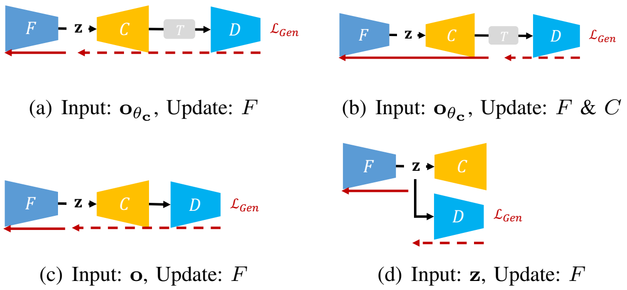
representations by utilizing the unlabelled real data, therefore performing better in terms of pose estimation.

*c) **Reconstruction** Versus **Completion**:* The method Self-Recon is adopted from [37], which utilizes point cloud reconstruction from latent space as auxiliary tasks. It shows preferable results compared to all other methods except for the proposed method. Our proposed method applied point cloud completion task, and its better pose estimation result demonstrates its higher capability of capturing complete geometric representations.

4) Qualitative Results: In Fig. 8, qualitative results of the lower-limb pose estimation are displayed. The region of interest is annotated in green dashed rectangles and zoomed in views are provided for better visualization. It can be observed that our method presents superior performance compared to the others in selected views and gait types. Although most compared methods can achieve reasonably good visual results, they cannot achieve the same precision level as our method.

D. Abnormal Gait Recognition

1) Quantitative Results: Based on the extracted lower-limb pose sequence from different methods, we compared their capabilities of discriminating different gait abnormalities. Based on the generated heel-strike event from Mocap data, a pose sequence was segmented into each cycle and then normalized to the same length and orientation. A network based on 1D-CNN adopted from [6] was trained on all the ground truth (100%) trajectories of training subjects for classification training, details of which can be found in the Supplementary Material. We present the results of Macro Precision, Recall, F1, and Accuracy. The result based on ground truth data is {Prec:0.840; Rec:0.843; F1:0.844; Acc:0.832}. This performance is limited by the individual difference across subjects [6], which is out of the scope of this paper. The classification results using the estimated skeleton of the testing subject is reported in Table II, which reflect the precision as well as discriminative capability of different methods. Over 75% accuracy can be achieved by our method, which demonstrates better discriminativeness across gait patterns based on the estimated skeletons, compared to other methods.



(a) Input: $\mathbf{o}_{\theta_c}$, Update: F (b) Input: $\mathbf{o}_{\theta_c}$, Update: F & C

(c) Input: $\mathbf{o}$, Update: F (d) Input: $\mathbf{z}$, Update: F

Fig. 10. Illustrations of different adversarial training strategies for domain adaptation. (Red solid refers to the update of weights while red dash refers to freezing/no update.)

TABLE III
COMPARISON OF DIFFERENT DOMAIN ADAPTATION STRATEGIES AS SHOWN IN FIG. 10 AND W/O DOMAIN ADAPTATION. (1% ANNOTATED)

Strategies	Dist (cm)↓	Angle (degree)↓	Acc↑
a (proposed)	2.74±1.48	5.48±4.89	0.759
b	3.54±1.60	7.81±7.22	0.660
c	5.79±3.29	8.76±9.34	0.585
d	5.45±3.80	8.25±8.81	0.556
w/o DA	3.32±1.92	6.89±6.29	0.642

Mean±Std for Dist and Angle; Mean for Acc.

2) Qualitative Results: The distributions of the features extracted from the last but one layer of our classifier, is presented in Fig. 9 by t-SNE (t-Distributed Stochastic Neighbor Embedding) plot. We compared the features from GT, estimation of V2V-Syn, and estimation of our proposed method. As shown in Fig. 9(a), the features of GT are reasonably discriminative across patterns. Slight marginal ambiguity between normal and pronation might be caused by the overfitting due to the heterogeneity between different subjects, since our classifier is trained on the other seven subjects. Comparing Fig. 9(b) and Fig. 9(c), we can observe that our proposed method can better keep discriminative categorical distributions, which is enabled by its superior performance on skeleton estimation.

TABLE IV
PERFORMANCE CHANGES WITH THE CHANGE OF NUMBER OF SYNTHETIC TRAINING/VALIDATION DATA.

Ratio	Dist (cm)↓	Angle (degree)↓	Acc↑
10%	4.99±3.01	9.89±7.35	0.574
50%	3.19±2.30	6.32±5.90	0.709
100%	2.74±1.48	5.48±4.89	0.759
200%	2.71±1.40	5.49±4.90	0.759

Mean±Std for Dist and Angle; Mean for Acc.

TABLE V
COMPARISON BETWEEN OUR PROPOSED METHOD AND V2V-REAL WHEN 1% OR 100% GROUND TRUTH SKELETON DATA IS AVAILABLE.

Methods	Dist (cm)↓	Angle (degree)↓	Acc↑
Proposed-1%	2.74±1.48	5.48±4.89	0.759
Proposed-100%	2.25± 1.44	4.63±3.24	0.805
V2V-1%	3.92±2.24	9.23±12.05	0.628
V2V-100%	2.22±1.22	4.62±3.50	0.809

Mean±Std for Dist and Angle; Mean for Acc.

This is consistent with the quantitative classification results reported in Table II. To address the overfitting problem, in our future work, the distribution shift between subjects, as well as between ground truth and estimation, is considered to be mitigated by domain adaptation as well.

E. Other Results

1) Ablation study - variants of adversarial training: To evaluate the effectiveness of our proposed architecture, especially the adversarial architecture, we explore and compare various variants with the proposed one, as shown in Fig. 10. The results are shown in Table III. The superior performance of our method demonstrates the effectiveness of our architecture. We observe that w/o DA can achieve relatively good results, which to some extend shows the relative similarity between $\mathbf{x}^r$ and $\mathbf{x}^s$ as well as the relative robustness to noise of our multi-resolution encoder architecture.

2) Performance with the change of synthetic data volume: Extended from the synthetic training/validation data (4k×7, 100%), experiments based on 10%, 50%, 200% were done. It is observed from Table IV that less data would decrease the data diversity, thus decreasing the overall performance, whereas the result starts to converge with more data added.

3) Performance changes with the whole ground truth: We compare our proposed method with the state-of-the-art V2V-Real [22] when all the ground truth of real data is available. Results in Table V show that V2V outperforms our method slightly when the full ground truth is available. This shows the strength of voxelized heatmap prediction proposed in V2V; however, it is computationally expensive (Running time with Pytorch Titan Xp: V2V 26.1 ms; Proposed 3.5 ms).

On the other hand, for our proposed method, only a small reduction is observed when only 1% ground truth is available compared to the full training. This demonstrates the robustness of our method against the number of available ground truth, and also indicates that the proposed method can effectively reduce the efforts of acquiring ground truth. Our future work would consider applying more sophisticated and advanced models for point cloud analysis, to better capture movement subtle changes.

VI. CONCLUSION

Acquiring Mocap data in complex laboratory settings for gait analysis is expensive and laborious. In this paper, we have proposed a novel cross-domain self-supervised learning framework. It not only enables gait analysis in home environments, but also minimizes the need for ground truth Mocap data for the pose estimation model training. Detailed comparative experiments based on leave-one-subject-out cross validation were conducted with the state-of-the-art approaches. The results show that our approach achieved superior performance in terms of pose estimation with minimal Mocap for training. Furthermore, with accurate pose estimation of lower limbs, our proposed method can better capture subtle yet important abnormal gait deviations for improved gait pattern recognition.

Future work should consider the occlusion caused by arm swinging under daily self-paced walking settings as well as the generalization capability of our proposed method to novel arbitrary viewpoints in the real world.

REFERENCES

[1] Ł. Kidziński, B. Yang, J. L. Hicks, A. Rajagopal, S. L. Delp, and M. H. Schwartz, "Deep neural networks enable quantitative movement analysis using single-camera videos," *Nature Communications*, vol. 11, no. 1, pp. 1–10, 2020.

[2] F. Temporiti, G. Zanotti, R. Furone, S. Molinari, M. Zago, M. Loppini, M. Galli, G. Grappiolo, and R. Gatti, "Gait analysis in patients after bilateral versus unilateral total hip arthroplasty," *Gait & Posture*, vol. 72, pp. 46–50, 2019.

[3] S. Chen, J. Lach, B. Lo, and G.-Z. Yang, "Toward pervasive gait analysis with wearable sensors: A systematic review," *IEEE J Biomed Health Inform*, vol. 20, no. 6, pp. 1521–1537, 2016.

[4] G.-Z. Yang, *Body Sensor Networks*. Springer, 2014.

[5] Y. Adesida, E. Papi, and A. H. McGregor, "Exploring the role of wearable technology in sport kinematics and kinetics: A systematic review," *Sensors*, vol. 19, no. 7, p. 1597, 2019.

[6] X. Gu, Y. Guo, F. Deligianni, B. Lo, and G.-Z. Yang, "Cross-subject and cross-modal transfer for generalized abnormal gait pattern recognition," *IEEE Trans Neural Netw Learn Syst*, 2020.

[7] M. Kocabas, N. Athanasiou, and M. J. Black, "Vibe: Video inference for human body pose and shape estimation," in *Comput Vis ECCV*, 2020, pp. 5253–5263.

[8] Z. Cao, G. H. Martinez, T. Simon, S.-E. Wei, and Y. A. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *IEEE Trans Pattern Anal Mach Intell*, 2019.

[9] Y. Xiu, J. Li, H. Wang, Y. Fang, and C. Lu, "Pose Flow: Efficient online pose tracking," in *BMVC*, 2018.

[10] S. L. Colyer, M. Evans, D. P. Cosker, and A. I. Salo, "A review of the evolution of vision-based motion analysis and the integration of advanced computer vision methods towards developing a markerless system," *Sports medicine-open*, vol. 4, no. 1, p. 24, 2018.

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

[12] N. Seethapathi, S. Wang, R. Saluja, G. Blohm, and K. P. Kording, "Movement science needs different pose tracking algorithms," *arXiv preprint arXiv:1907.10226*, 2019.

[13] A. Pfister, A. M. West, S. Bronner, and J. A. Noah, "Comparative abilities of microsoft kinect and vicon 3d motion capture for gait analysis," *J Med Eng Technol*, vol. 38, no. 5, pp. 274–280, 2014.

[14] S. Springer and G. Yogev Seligmann, "Validity of the kinect for gait assessment: A focused review," *Sensors*, vol. 16, no. 2, p. 194, 2016.

[15] A. Haque, B. Peng, Z. Luo, A. Alahi, S. Yeung, and L. Fei-Fei, "Towards viewpoint invariant 3d human pose estimation," in *Comput Vis ECCV*. Springer, 2016, pp. 160–177.

[16] F. Xiong, B. Zhang, Y. Xiao, Z. Cao, T. Yu, J. T. Zhou, and J. Yuan, "A2j: Anchor-to-joint regression network for 3d articulated pose estimation from a single depth image," in *Proc IEEE Int Conf Comput Vis*, 2019, pp. 793–802.

[17] Z. Zhang, L. Hu, X. Deng, and S. Xia, "Weakly supervised adversarial learning for 3d human pose estimation from point clouds," *IEEE Trans Vis Comput Graph*, vol. 26, no. 5, pp. 1851–1859, 2020.

[18] T. N. Nguyen, H. H. Huynh, and J. Meunier, “3d reconstruction with time-of-flight depth camera and multiple mirrors,” *IEEE Access*, vol. 6, pp. 38 106–38 114, 2018.

[19] S. Yuan, G. Garcia-Hernando, B. Stenger, G. Moon, J. Yong Chang, K. Mu Lee, P. Molchanov, J. Kautz, S. Honari, L. Ge *et al.*, “Depth-based 3d hand pose estimation: From current achievements to future goals,” in *Proc IEEE Conf Comput Vis Pattern Recognit*, 2018, pp. 2636–2645.

[20] H. Kainz, D. Graham, J. Edwards, H. P. Walsh, S. Maine, R. N. Boyd, D. G. Lloyd, L. Modenese, and C. P. Carty, “Reliability of four models for clinical gait analysis,” *Gait & posture*, vol. 54, pp. 325–331, 2017.

[21] M. Wang and W. Deng, “Deep visual domain adaptation: A survey,” *Neurocomputing*, vol. 312, pp. 135–153, 2018.

[22] G. Moon, J. Yong Chang, and K. Mu Lee, “V2v-posenet: Voxel-to-voxel prediction network for accurate 3d hand and human pose estimation from a single depth map,” in *Proc IEEE Conf Comput Vis Pattern Recognit*, 2018, pp. 5079–5088.

[23] V. Ganapathi, C. Plagemann, D. Koller, and S. Thrun, “Real-time human pose tracking from range data,” in *Comput Vis ECCV*. Springer, 2012, pp. 738–751.

[24] M. Ye and R. Yang, “Real-time simultaneous pose and shape estimation for articulated objects using a single depth camera,” in *Proc IEEE Conf Comput Vis Pattern Recognit*, 2014, pp. 2345–2352.

[25] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, “Real-time human pose recognition in parts from single depth images,” in *CVPR 2011*. IEEE, 2011, pp. 1297–1304.

[26] J. Shotton, R. Girshick, A. Fitzgibbon, T. Sharp, M. Cook, M. Finocchio, R. Moore, P. Kohli, A. Criminisi, A. Kipman *et al.*, “Efficient human pose estimation from single depth images,” *EEE Trans Pattern Anal Mach Intell*, vol. 35, no. 12, pp. 2821–2840, 2012.

[27] H. Yub Jung, S. Lee, Y. Seok Heo, and I. Dong Yun, “Random tree walk toward instantaneous 3d human pose estimation,” in *Proc IEEE Conf Comput Vis Pattern Recognit*, 2015, pp. 2467–2474.

[28] L. Ge, H. Liang, J. Yuan, and D. Thalmann, “Robust 3d hand pose estimation in single depth images: from single-view cnn to multi-view cnns,” in *Proc IEEE Conf Comput Vis Pattern Recognit*, 2016, pp. 3593–3601.

[29] ——, “3d convolutional neural networks for efficient and robust hand pose estimation from single depth images,” in *Proc IEEE Conf Comput Vis Pattern Recognit*, 2017, pp. 1991–2000.

[30] M. Vasileiadis, C.-S. Bouganis, G. Stavropoulos, and D. Tzovaras, “Optimising 3d-cnn design towards human pose estimation on low power devices.” in *BMVC*, 2019, p. 42.

[31] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, “Pointnet: Deep learning on point sets for 3d classification and segmentation,” in *Proc IEEE Conf Comput Vis Pattern Recognit*, 2017, pp. 652–660.

[32] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, “Pointnet++: Deep hierarchical feature learning on point sets in a metric space,” in *Adv Neural Inf Process Syst*, 2017, pp. 5099–5108.

[33] Z. Liu, H. Tang, Y. Lin, and S. Han, “Point-voxel cnn for efficient 3d deep learning,” in *Adv Neural Inf Process Syst*, 2019, pp. 965–975.

[34] S. Shi, C. Guo, L. Jiang, Z. Wang, J. Shi, X. Wang, and H. Li, “Pv-rcnn: Point-voxel feature set abstraction for 3d object detection,” in *Proc IEEE/CVF Conf Comput Vis Pattern Recognit*, 2020, pp. 10 529–10 538.

[35] Z. Huang, Y. Yu, J. Xu, F. Ni, and X. Le, “Pf-net: Point fractal network for 3d point cloud completion,” in *Proc IEEE/CVF Conf Comput Vis Pattern Recognit*, 2020, pp. 7662–7670.

[36] L. Ge, Y. Cai, J. Weng, and J. Yuan, “Hand pointnet: 3d hand pose estimation using point sets,” in *Proc IEEE Conf Comput Vis Pattern Recognit*, 2018, pp. 8417–8426.

[37] Y. Chen, Z. Tu, L. Ge, D. Zhang, R. Chen, and J. Yuan, “So-handnet: Self-organizing network for 3d hand pose estimation with semi-supervised learning,” in *Proc IEEE Int Conf Comput Vis*, 2019, pp. 6961–6970.

[38] S. Li and D. Lee, “Point-to-pose voting based hand pose estimation using residual permutation equivariant layer,” in *Proc IEEE Conf Comput Vis Pattern Recognit*, 2019, pp. 11 927–11 936.

[39] Z. Ren and Y. Jae Lee, “Cross-domain self-supervised multi-task feature learning using synthetic imagery,” in *Proc IEEE Conf Comput Vis Pattern Recognit*, 2018, pp. 762–771.

[40] X. Gu, Y. Guo, F. Deligianni, and G.-Z. Yang, “Coupled real-synthetic domain adaptation for real-world deep depth enhancement,” *IEEE Trans Image Process*, vol. 29, pp. 6343–6356, 2020.

[41] X. Chen, B. Chen, and N. J. Mitra, “Unpaired point cloud completion on real scans using adversarial training,” in *International Conference on Learning Representations*, 2019.

[42] L. Jing and Y. Tian, “Self-supervised visual feature learning with deep neural networks: A survey,” *IEEE Trans Pattern Anal Mach Intell*, 2020.

[43] M. Kocabas, S. Karagoz, and E. Akbas, “Self-supervised learning of 3d human pose using multi-view geometry,” in *Proc IEEE Conf Comput Vis Pattern Recognit*, 2019, pp. 1077–1086.

[44] J. Sauder and B. Sievers, “Self-supervised deep learning on point clouds by reconstructing space,” in *Adv Neural Inf Process Syst*, 2019, pp. 12 962–12 972.

[45] I. Achituve, H. Maron, and G. Chechik, “Self-supervised learning for domain adaptation on point-clouds,” *arXiv preprint arXiv:2003.12641*, 2020.

[46] I. Mahmood, U. Martinez-Hernandez, and A. A. Dehghani-Sanij, “Evaluation of gait transitional phases using neuromechanical outputs and somatosensory inputs in an overground walk,” *Human Movement Science*, vol. 69, p. 102558, 2020.

[47] A. Beeck, V. Quack, B. Rath, M. Wild, R. Michalik, H. Schenker, and M. Betsch, “Dynamic evaluation of simulated leg length inequalities and their effects on the musculoskeletal apparatus,” *Gait & posture*, vol. 67, pp. 71–76, 2019.

[48] W. Cui, C. Wang, W. Chen, Y. Guo, Y. Jia, W. Du, and C. Wang, “Effects of toe-out and toe-in gaits on lower-extremity kinematics, dynamics, and electromyography,” *Applied Sciences*, vol. 9, no. 23, p. 5245, 2019.

[49] X.-S. Gao, X.-R. Hou, J. Tang, and H.-F. Cheng, “Complete solution classification for the perspective-three-point problem,” *IEEE Trans Pattern Anal Mach Intell*, vol. 25, no. 8, pp. 930–943, 2003.

[50] S. Ghorbani, K. Mahdaviani, A. Thaler, K. Kording, D. J. Cook, G. Blohm, and N. F. Troje, “Movi: A large multipurpose motion and video dataset,” *arXiv preprint arXiv:2003.01888*, 2020.

[51] K. Lin, L. Wang, K. Luo, Y. Chen, Z. Liu, and M.-T. Sun, “Cross-domain complementary learning with synthetic data for multi-person part segmentation,” *arXiv preprint arXiv:1907.05193*, 2019.

[52] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, “SMPL: A skinned multi-person linear model,” *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, vol. 34, no. 6, pp. 248:1–248:16, Oct. 2015.

[53] G. Varol, J. Romero, X. Martin, N. Mahmood, M. J. Black, I. Laptev, and C. Schmid, “Learning from synthetic humans,” in *Proc IEEE Conf Comput Vis Pattern Recognit*, 2017, pp. 109–117.

[54] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black, “Keep it smpl: Automatic estimation of 3d human pose and shape from a single image,” in *Comput Vis ECCV*. Springer, 2016, pp. 561–578.

[55] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik, “End-to-end recovery of human shape and pose,” in *Proc IEEE Conf Comput Vis Pattern Recognit*, 2018, pp. 7122–7131.

[56] K. M. Robinette, S. Blackwell, H. Daanen, M. Boehmer, and S. Fleming, “Civilian american and european surface anthropometry resource (caesar), final report. volume 1. summary,” SYTRONICS INC DAYTON OH, Tech. Rep., 2002.

[57] S. Katz, A. Tal, and R. Basri, “Direct visibility of point sets,” in *ACM SIGGRAPH 2007 papers*, 2007, pp. 24–es.

[58] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, “Context encoders: Feature learning by inpainting,” in *Proc IEEE Conf Comput Vis Pattern Recognit*, 2016, pp. 2536–2544.

[59] S. Izadi, D. Kim, O. Hilliges, D. Molyneaux, R. Newcombe, P. Kohli, J. Shotton, S. Hodges, D. Freeman, A. Davison *et al.*, “Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera,” in *Proceedings of the 24th annual ACM symposium on User interface software and technology*, 2011, pp. 559–568.

[60] W. Yuan, T. Khot, D. Held, C. Mertz, and M. Hebert, “Pcn: Point completion network,” in *2018 International Conference on 3D Vision (3DV)*. IEEE, 2018, pp. 728–737.

[61] F. Kuhnke and J. Ostermann, “Deep head pose estimation using synthetic images and partial adversarial domain adaption for continuous label spaces,” in *Proc IEEE Int Conf Comput Vis*, 2019, pp. 10 164–10 173.

[62] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, “Improved training of wasserstein gans,” in *Adv Neural Inf Process Syst*, 2017, pp. 5767–5777.

[63] D. W. Shu, S. W. Park, and J. Kwon, “3d point cloud generative adversarial network based on tree structured graph convolutions,” in *Proc IEEE Int Conf Comput Vis*, 2019, pp. 3859–3868.

[64] P. Achlioptas, O. Diamanti, I. Mitliagkas, and L. Guibas, “Learning representations and generative models for 3d point clouds,” in *International conference on machine learning*. PMLR, 2018, pp. 40–49.

[65] N. Sarafianos, B. Boteanu, B. Ionescu, and I. A. Kakadiaris, “3d human pose estimation: A review of the literature and analysis of covariates,” *Comput Vis Image Underst*, vol. 152, pp. 1–20, 2016.