2022 IEEE International Conference on Robotics and Automation (ICRA)
May 23-27, 2022. Philadelphia, PA, USA

PoseSDF: Simultaneous 3D Human Shape Reconstruction and Gait Pose Estimation Using Signed Distance Functions

Jianxin Yang, Yuxuan Liu, Xiao Gu, Guang-Zhong Yang, *Fellow*, *IEEE*, Yao Guo, *Member*, *IEEE*

***Abstract*— Vision-based 3D human pose estimation and shape reconstruction play important roles in robot-assisted healthcare monitoring and personal assistance. However, 3D data captured from a single viewpoint always encounter occlusions and exhibit substantial heterogeneity across different views, resulting in significant challenges for both tasks. Extensive approaches have been proposed to perform each task separately, but few of them present a unified solution. In this paper, we propose a novel network based on signed distance functions, namely PoseSDF, to simultaneously reconstruct 3D lower limb shape and estimate gait pose by two dedicated branches. To promote multi-task learning, several strategies are developed to ensure that these two branches leverage the same latent shape code while exchanging information between them. More importantly, an auxiliary RotNet is incorporated into the inference phase, overcoming the inherent limitations of implicit neural functions under cross-view scenarios. Experimental results demonstrate that our proposed PoseSDF can achieve both high-quality shape reconstruction and precise pose estimation, generalizing well on the data from novel views, gait patterns, as well as real-world.**

I. INTRODUCTION

Recent advances in artificial intelligence and robotics are transforming healthcare, where robots can provide personal assistance, long-term ambient monitoring, as well as remote disease diagnosis at home environments [1]. Particularly, visual information endows the robot with sufficient perception capabilities, especially perceiving 3D human poses and performing subsequent analysis seamlessly [2], [3]. For instance, the embedded pose tracking and quantitative analysis modules can recognize pathological gait patterns and reveal the potential influencing factors, enabling ambient vision-based gait analysis at the end of robots [4], [5].

Research effort has been devoted to vision-based 3D human pose estimation recently. As an ill-posed problem, inferring 3D human poses from a single RGB image may lead to inaccurate 3D pose estimation by leveraging only 2D information [6]. Existing 3D sensing devices provide a more convenient solution by directly estimating 3D poses from direct access to 3D data [7], [8]. This line of pose estimation research utilized various representations of 3D data, such as depth maps, point clouds, and voxels, as the input to dedicated computational models [9], [10], [11]. Although these methods achieve reasonably good performance through either dedicated model design or training strategies, they are highly data-dependent approaches. Besides, the characteristics of observed 3D data are typical with the occlusions induced by partial scans and are of large variations across different viewpoints, yet the viewpoints of robots keep changing. Furthermore, the representations derived from data-driven models may not well depict unseen poses, leading to failure when inferring unseen walking patterns. More importantly, they focus only on the human pose, whereas the indispensable human shape information is overlooked.

The implicit function is a powerful tool for characterizing 3D shapes, which has gained great popularity as such a representation is not limited by fixed topology structure and resolution [12]. With the advent of deep learning models, implicit neural representations parameterized by neural networks have enabled great progress in shape reconstruction, 3D data compression, novel view synthesis, model animation [12], [13], [14], [15]. The commonly-used implicit functions can be categorized into two groups: classification-based occupancy [13] and regression-based signed distance function (SDF) [12]. Recently, extensive works built upon SDF have been developed for human shape reconstruction [16], [17]. However, the inputs of these approaches are pre-aligned in a canonical coordinate system and few of them can deal with the real data. Moreover, they cannot directly predict the key joints of human pose, impeding the real-world applications.

To address the aforementioned challenges, we propose a novel network based on implicit neural function, namely PoseSDF, for reconstructing the completed shape of a partial point cloud and simultaneously estimating 3D gait pose. Our PoseSDF is partially inspired by DeepSDF [12], but introducing several strategies to enable multi-task learning and improve performance. By regarding each joint as the center of a small virtual sphere, we first propose a pose decoder to learn the latent surfaces of nine virtual spheres, which can leverage the sufficient shape information represented in the latent code. Meanwhile, we introduce an encoder as well as the feature exchanging mechanism, enabling the information flow between two decoders, thus stabilizing the training process. To overcome the inherent challenge of SDF on novel views, a RotNet is designed for cross-view inference, which predicts the rotation between the input partial point cloud and its representation in a canonical view. To demonstrate the effectiveness of PoseSDF, extensive experiments were conducted on the dataset build upon ICL-Gait [18]. Comparison results demonstrate that our proposed PoseSDF achieves superior shape reconstruction and comparable pose

This work was 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: Yao Guo*)

J. Yang, Y. Liu, G.-Z. Yang, and Y. Guo are with the Institute of Medical Robotics, Shanghai Jiao Tong University, China. (jianxinyang@sjtu.edu.cn, 20000905lyx@sjtu.edu.cn, gzyang@sjtu.edu.cn, yao.guo@sjtu.edu.cn)

X. Gu is with Hamlyn Centre, Imperial College London, United Kingdom. (xiao.gu17@imperial.ac.uk)

2022 IEEE International Conference on Robotics and Automation (ICRA) | 978-1-7281-9681-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICRA46639.2022.9812051

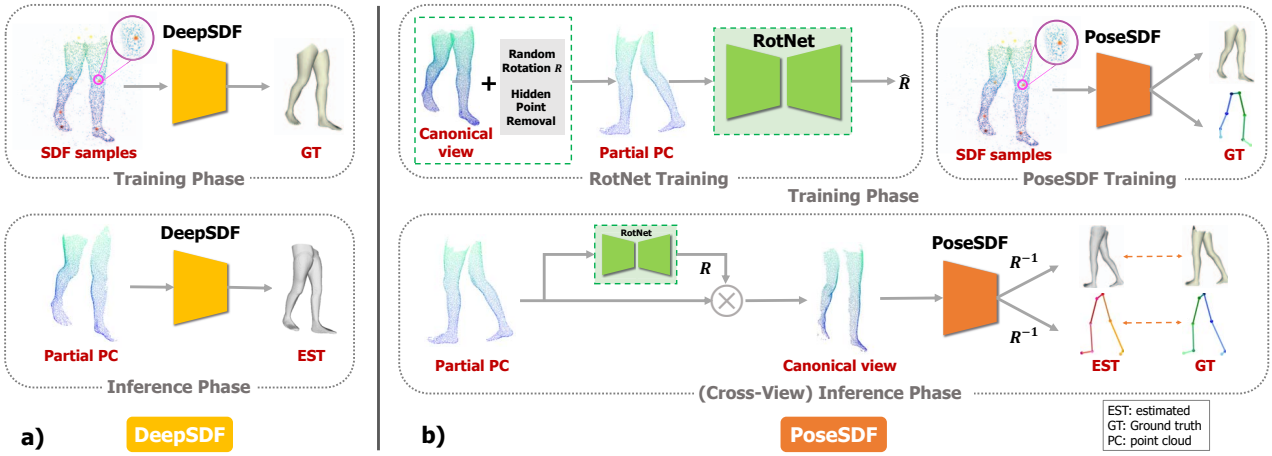

Fig. 1. Illustration of the training and inference phases of DeepSDF and our PoseSDF. a) Given the partial point clouds represented in a canonical coordinate system, DeepSDF can reconstruct the completed shapes; b) In order to perform both shape reconstruction and pose estimation with one network, we introduce an additional pose decoder following the similar structure as DeepSDF auto-decoder. Instead of directly estimating the joint positions, we treat this task as the reconstruction of nine spheres that centered on key joints. Although PoseSDF is trained with the data captured from a single view, a dedicated RotNet is incorporated with PoseSDF to first transform the input point cloud from an unseen view to the canonical representation, thus addressing the challenges on cross-view reconstruction.

estimation with a small amount of training data captured from a canonical view. It also shows good generalization capability in terms of the cross-view scenarios and real-world data. In summary, the contribution of this paper is three-fold:

- We propose a unified network for 3D lower limb reconstruction and gait pose estimation. To the best of our knowledge, this is the first work to perform these two tasks simultaneously using signed distance functions.
- Different from data-driven models, our PoseSDF can achieve high-quality shape reconstruction and precise pose estimation with a small amount of training data.
- Our PoseSDF can generalize well to unseen views and gait patterns, and moreover can be directly applied to real data.

II. Related Work

A. Depth Based 3D Human Body Modeling

1) Pose estimation: One line of research for body modeling aims to estimate skeletons from observed depth images. Existing deep learning based approaches are grounded on different representations of depth data [10], [11], [19], among which most are discriminative models optimized in a data-driven manner. They are limited in considering the 3D surface information encoded in depth data, and may fail in pose estimation with varied shapes. Recent work incorporating shape recovery/completion as auxiliary tasks has demonstrated that making use of shape information can provide coherent improvement for pose estimation [20].

2) Shape reconstruction: Existing shape reconstruction works are based on parametric models, explicit representations or implicit representations. One of the main advantages of parametric models is that they allow easy control of the pose and shapes, such as the sophisticated 3D human models like Skinned Multi-Person Linear Model (SMPL) [21]. However, the gap between real-world human bodies and these parametric models may lead to oversmoothing and cannot deal well with shape deformations. On the other hand, existing explicit representation based methods generate the complete point sets, meshes, or voxels directly [22], which mostly are limited to a single topology and cannot produce high-quality results. Till now, reconstruction based on implicit functions has received increasing attention, as it allows prediction in a continuous space implicitly and thus facilitates high-quality reconstruction and complex pose information capture [12], [23], which is discussed next.

B. Implicit Neural Representations

Recent advances in multi-layer-perceptron (MLPs) based implicit neural representations allow characterizing the 3D surface data in a continuous and differentiable manner. It has gained increasing popularity in modeling 3D representations of scenes and objects [12], [13], [23], [24]. One representative pioneer work DeepSDF [12] proposed an auto-decoder structure to map 3D coordinates to the SDF values of corresponding positions, implicitly modeling the shape boundary as the zero-level-set of learned representations. Based on DeepSDF, Mihenhall *et al.* [24] added a positional encoding of the input coordinates to facilitate the high-frequency representations parameterized by MLPs.

Recent work has successfully applied implicit neural representations on reconstruction, view-synthesis, and registration of humans [16], [25], [26], [27]. This line of research inspires our work of adapting DeepSDF for simultaneously modeling lower-limb shapes and poses. However, they mostly assumed the same or already available viewpoint information, and took complete point sets as input in both training and testing. These assumptions hardly hold true in the real world, as the real-scanned point clouds are only partially observable and are prone to rotation variations [18]. In our work, we proposed series of novel solutions to adapt the original DeepSDF to simultaneously perform pose estimation as well, meanwhile addressing the data shift caused by view variations and complete-partial discrepancy.

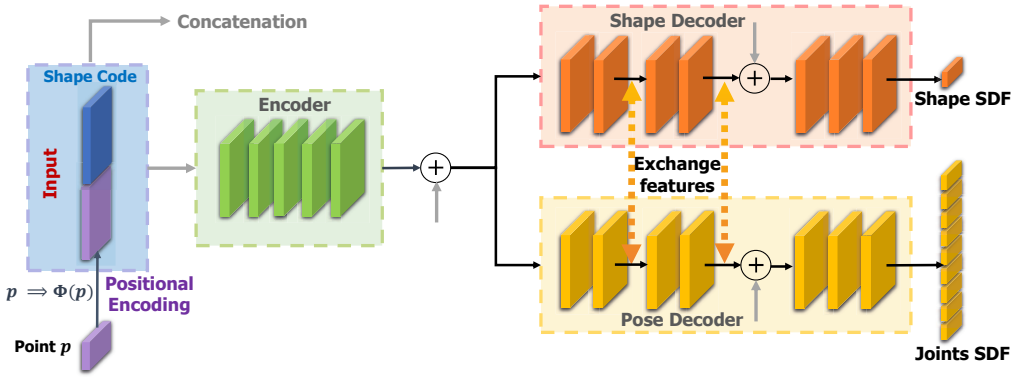

Fig. 2. The network architecture of our proposed PoseSDF. The gray arrow indicates the concatenation with the input vector (i.e., the latent code and positional encoding of the sampled point). The output from the Encoder is feed into the Shape decoder and Pose decoder for predicting shape sdf and joint sdf, respectively. Within these two decoders, the features are exchanged with each other to stabilize the training.

III. Methodology

A. Signed Distance Function for Shape Reconstruction

Before going deep into our proposed PoseSDF, we first give an overview of the signed distance function (SDF) and its pipeline for shape reconstruction. Given a point $p = [x, y, z]^T$ in 3D, SDF is a continuous function that outputs the distance $s = SDF(p; \mathcal{S}) \in \mathbb{R}^1$ to the closest surface of a closed 3D shape $\mathcal{S}$, where the sign of s indicates whether this point p is inside (negative) or outside (positive). The latent surface of this 3D shape is the zero-level-set of SDF, i.e., $SDF(\cdot) = 0$.

DeepSDF [12] is the pioneer work that learns the mapping from point samples to the continuous signed distance function using a multi-layer fully-connected network, which is denoted as auto-decoder. In order to learn a shared representation space of K different shapes $\{\mathcal{S}_k\}_{k=1}^K$, a set of N points $\{p_i\}_{i=1}^N$ and their corresponding SDF values $\{s_i\}_{i=1}^N$ are first sampled from each shape.

The network $f_\theta(z_k, p)$ is trained to learn an approximation of the given point-SDF pairs $\{p_i, s_i\}_{i=1}^N$, where θ denotes the network parameters and z_k is a latent vector of shape $\mathcal{S}_k$. During the training process, the latent code z_k is randomly initialized to obey $\mathcal{N}\left(0, 0.01^2\right)$ and SDF values are clamped to $[-\delta, \delta]$ by $\text{clamp}_\delta(s) := \min(\delta, \max(-\delta, s))$, where δ is a truncated distance. Then the network parameters θ and the latent code z_k are jointly updated by solving the following objective function:

$$\arg\min_{\theta, z_k} \sum_{k=1}^{K} \left(\sum_{i=1}^{N} \mathcal{L}\left(\hat{s}_i, s_i\right) + 1/\sigma^2 \left\| z_k \right\|_2^2 \right) \tag{1}$$

$$\mathcal{L}(\hat{s}_i, s_i) = \left| \text{clamp}_\delta(\hat{s}_i) - \text{clamp}_\delta(s) \right| \tag{2}$$

where $\hat{s}_i = f_\theta(z_k, p_i)$ is the predicted SDF value, and $1/\sigma^2 \left\| z_k \right\|_2^2$ is a regularization term that prevents overfitting.

Different from the conventional encoder-decoder structure, this auto-decoder requires an optimization step of the latent code z_u in the inference phase. Concretely, given a partial scan $X_u = \{p_u, s_u\}$ of an unseen shape $\mathcal{S}_u$, the optimal latent code z_u is first estimated by Maximum-a-Posterior (MAP) estimation with the trained model parameters f_θ. It is noteworthy that the SDF values s_u of the input point cloud are inherently set as zero. Afterwards, with the optimized latent code $\hat{z}_u$, the shape $\mathcal{S}_u$ can be reconstructed by extracting the zero-level-set on uniformly sampled grid through Marching Cubes [28].

Compared to the direct shape reconstruction from point clouds [29] and voxels [20], $f_\theta(z_k, x)$ encodes the continuous surface of a 3D shape, which can achieve hyper-resolution reconstruction. It is also advantageous in processing the input data with an arbitrary number of points by taking a single point at a time in both training and inferring, rather than resample the input data to a fixed number of points as for point clouds [11] and voxels [10].

B. PoseSDF Network

Inspired by DeepSDF, the proposed PoseSDF aims to leverage its advantages in shape reconstruction, meanwhile extending the functionality of implicit functions to pose estimation by performing multiple tasks. Since the latent code z_k is a high-dimensional feature that encodes sufficient information of the input data, therefore, how to make full use of this latent embedding to simultaneously estimate 3D human pose is the main focus of this paper. To address this, as shown in Fig. 2, we propose a unified deep neural network called **Pose Signed Distance Functions (PoseSDF)**, which can simultaneously reconstruct the human body shape and estimate human pose from a partial point cloud. Besides, we introduce several novel solutions to enhance the generalization capability of the model to unseen human shapes, gait patterns, viewpoints, and the challenging real-world data.

1) Positional encoding of sampled points: Given a set of completed human shapes $\{\mathcal{S}_k\}_{k=1}^K$ in the training set, we first follow the same criterion as DeepSDF [12] to sample a set of points $\{x_i\}$ and calculate the corresponding SDF $\{s_i\}$.

Unlike DeepSDF directly concatenating the latent code z_k and point p, we adopt a Fourier Positional Encoding (FPE) [24] as in Eq. (3) to map point p to a high-dimensional space $\Phi(p)$, which can help to recover 3D shape with fine-scale geometry details from unorganized 3D point clouds.

$$\begin{aligned}\Phi(p) = &[\sin(2\pi\omega_1^T p), \cos(2\pi\omega_1^T p), \cdots, \\ &\sin(2\pi\omega_M^T p), \cos(2\pi\omega_M^T p)]\end{aligned} \tag{3}$$

Then, the input to our network is the concatenation of three items, i.e., $[z_k, p, \Phi(p)]$.

2) Encoder module: An intuitive way to adapt DeepSDF for performing multiple tasks is to duplicate the multi-layer fully-connected network with multiple branches. However, our preliminary attempt reveals that the gradients of two branches are directly back-propagated to the latent code under this circumstance, leading to inferior performance due to the competition between the two branches. Alternatively, we design an Encoder module as shown in Fig. 2 to map the input vector to a latent space, which can balance and stabilize the joint training of shape reconstruction and pose estimation. Moreover, to avoid information loss, the input vector is concatenated again with the output of the Encoder.

3) Shape decoder: The shape decoder $f_{\theta_{rec}}(z_k, p_i)$ aims to reconstruct the human body shape from the input vector, which adopts a similar structure as DeepSDF [12] yet with several improvements.

On the one hand, it has been proven in [30] that truncated SDF $\mathrm{clamp}_\delta(s_i)$ used in DeepSDF could lead to unstable training, while incorporating both regression and classification of occupancy into the loss function, i.e., estimate both the SDF value and predict the sign of SDF, will make a faster convergence of the network. Thus, inspired by [31], we utilize a weighted loss function as follow:

$$\mathcal{L}_{rec}(\hat{s}_i, s_i) = \omega \max\{|\hat{s}_i - s_i| - \epsilon, 0\} \tag{4}$$

$$\omega = (1 + \lambda \,\mathrm{sign}(s_i)\,\mathrm{sign}(\hat{s}_i - s_i)) \tag{5}$$

where ϵ is a tolerance distance that treat SDF values within $[-\epsilon, \epsilon]$ as the zero-level-set surface, and λ is a weight coefficient that punishes the wrong estimation of signs and enables the implicit training of the occupancy classification. In specific, we gradually increase λ and decrease ϵ in the training phase to facilitate the network learning coarse surfaces first and then concentrating on the fine details.

On the other hand, DeepSDF [12] uses `tanh` activation layer to regress SDF values, which not only limits the output within $[-1, 1]$, but also makes the regression in a more complicated non-linear manner. To solve this, we replace the activation layer with `LeakyReLU` in our PoseSDF.

4) Pose decoder: Although the latent code is a global representation for each shape, the network's output only depends on local information since it takes a single point p_i as input in each query, which makes the pose estimation challenging. To address this, we adopt a similar idea of DeepSDF to achieve 3D human pose estimation in a unified architecture. Intuitively speaking, the distance from a point p_i to a joint j_i can be expressed as a signed distance as well. However, in this manner, only the SDF of the joint position equals zero, while others are larger than zero. This imbalance scenario would lead to the network fail to convergence. To solve this, we represent each joint as a small sphere with radius r. Along this line, the pose estimation task is transformed to a reconstruction problem of n spheres, wherein this paper we use nine key joints {*pelvis*, *left hip*, *left knee*, *left ankle*, *left foot*, *right hip*, *right knee*, *right ankle*, *right foot*} to represent human lower limb pose.

The output of the pose decoder $f_{\theta_{pose}}(z_k, p_i)$ is a vector consisting of the SDF values to nine virtual spheres, i.e., $\hat{\boldsymbol{J}} = [\hat{j}_1, \cdots, \hat{j}_9]$. Thus, the loss function for pose decoder can be formulated as

$$\mathcal{L}_{pose}\left(\hat{\boldsymbol{J}}, \boldsymbol{J}\right) = \omega/9 \left\| \hat{\boldsymbol{J}} - \boldsymbol{J} \right\|_1, \tag{6}$$

where ω is the same as Eq. (5).

5) Feature exchanging between shape and pose decoders: It is noteworthy that during the inference phases, the input to the network is only the partial point cloud, i.e., the part of the zero-level-set of SDF values. Since there is no specific guidance for the pose decoder, the network would tend to fail in pose estimation. To solve this, we exchange the latent embedding between the first few layers of the shape decoder and pose decoder as shown in Fig. 2. Together with the encoder module, these strategies enable two branches to communicate with each other during both training and inference, rather than purely two independent downstream branches sharing the same input representation and competing with each other.

6) Total loss function: The final loss function for the end-to-end training of our PoseSDF is formulated as follow:

$$\mathcal{L}_{total} = \mathcal{L}_{rec} + \mathcal{L}_{pose} + 1/\sigma^2 \left\| z_k \right\|_2^2 \tag{7}$$

7) Inference phase: As illustrated in Fig. 1, during the inference, PoseSDF takes the partial scan of the human lower limb as input. Firstly, the randomly initialized latent code z_k is optimized through back-propagation by leveraging this zero-level-set of SDF (i.e., partial point cloud) as following:

$$\hat{z}_k = \arg\min_{z_k} \left(\mathcal{L}_{rec} + 1/\sigma^2 \left\| z_k \right\|_2^2 \right) \tag{8}$$

With the optimized latent code, the latent surface of the human shape can be reconstructed via the shape decoder. On the other hand, nine small spheres indicating key joints can be reconstructed by the pose decoder, then we calculate the center of each sphere as the estimated joint position.

An inherent challenge of implicit neural representations (e.g., SDF and occupancy) is that it treats a shape with rotation variances as a totally different representation in the shape space. In this manner, the model trained with the shapes represented in a canonical coordinate system cannot generalize well to other unseen views (see Fig. 3(a)). Besides, even mixing the data generated from multiple viewpoints, the network like DeepSDF cannot reconstruct satisfactory shapes even being tested on the data from trained views, as shown in Fig. 3(b). This would reveal the shortage of such an implicit neural representation for cross-view reconstruction. In practice, we cannot guarantee that the captured depth maps or point clouds are always from a fixed viewpoint. To address this, we introduce an auxiliary RotNet in the inference phase to transform the input point clouds to the canonical viewpoint (i.e., training view). Concretely, we adopt PointNet++ [32] as the backbone of RotNet, predicting the rotation differences between the input point cloud and its canonical representation before being fed into PoseSDF. Such a simple strategy can significantly improve both reconstruction and pose estimation performance in terms of novel views.

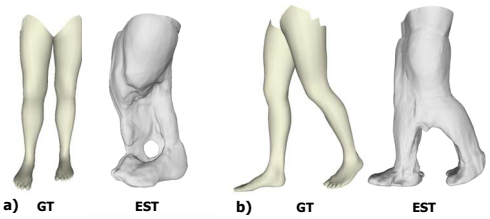

Fig. 3. Illustration of failure cases of DeepSDF for cross-view reconstruction. a) trained on the canonical view and tested on an unseen view; b) trained on the data with multiple views and tested on one seen views.

C. Data Preparation

To train our PoseSDF model, we leveraged a publicly available ICL-Gait dataset [18] for training and testing. The data was recorded from 5 viewpoints and 5 conditions (*Normal*, *Supination*, *Pronation*, *In-toeing*, and *Out-toeing*). We refer readers to `https://xiaogu.site/ICL_gait/` for more details of this dataset. The synthetic data in the training set was generated under *View 1* (canonical coordinates system) by using the SMPL [21], in which the pose parameters were derived from the *Normal* gait kinematics, and the shape parameter for each synthetic sample was randomly chosen from SURREAL dataset [33] (only 300 synthetic shapes in total for training). For the testing set, we added another viewpoint (front view) in addition to original 5 views to improve the diversity of synthetic samples. We generated 100 samples per gait condition and per viewpoint excluding *View 1, Normal* (2900 samples in total). It is noteworthy that different proportions of the test set were utilized in terms of the specific validation protocol.

For each shape, the sampling space was first normalized as a unit sphere that covers the entire shape. Then, 50,000 points were sampled near or on the surface similar to [12] and 8,000 points were sampled for each joint sphere of radius $r = 0.05$ (see SDF samples in Fig. 1). To improve the robustness of the model, we randomly added noise to the 3D shape surface and rotated the shape with a small angle ($< 5^\circ$) along z-axis. It needs to mention that only the partial point clouds of testing samples were used as input to the network during inference, in which they were generated by sampling points from meshes and applying the hidden point removal.

IV. Experiments and Results

A. Implementation Details

Our PoseSDF was mainly built upon multi-layer fully-connected networks, in which the dimension of each layer was 256 except for input and output ones. The dimension of $\Phi(p)$ and latent code z_k was 18 and 64, respectively. We trained the network for 3,000 epochs with an initial learning rate of 0.001 and multiplied it by 0.5 every 600 epochs. The λ is increased from 0 to 0.5 and ϵ is reduced from 0.0025 to 0 during training. To constrain the magnitude of latent codes, $1/\sigma^2$ was 10^{-4} during training and 10^{-2} during inference.

B. Comparison Methods

In the experiment, we compared with DeepSDF [12] for human shape reconstruction, and with PointNet++ [11], Voxel-to-Voxel [10] for 3D pose estimation. We also report the results of the ablated networks of PoseSDF.

- **PointNet++ (PN++)**[11]: a hierarchical pose estimation network with multi-scale grouping that learns local features of point clouds.
- **Voxel-to-Voxel (V2V)**[10]: a state-of-the-art pose estimation model that performs 3D convolution on voxels.
- **DeepSDF**[12]: a baseline model for 3D shape reconstruction using SDF.
- **PoseSDF (w/o Encoder)**: an ablated network directly feed the input to two decoders.
- **PoseSDF (w/o Exchange)**: an ablation study by removing the feature exchanging between two decoders.
- **PoseSDF (w/o RotNet)**: this ablated one aims to evaluate the RotNet under rotation variations. Noted that RotNet is only incorporated during the cross-view inference phase.
- **PoseSDF**: the full version of the proposed model.

C. Experiment Settings and Metrics

To highlight the generalization ability of PoseSDF comprehensively, we used four evaluation protocols as below. It should be pointed out that all models were only trained with the data belonging to the *Normal* gait condition of *View 1*.

- Cross-Condition (CC): We tested the models using the data from the other four abnormal patterns captured from the same viewpoint as training data. This protocol is used to show the robustness of our model in discriminating subtle pose and shape changes.
- Cross-View (CV): This protocol is used to evaluate the performance of the model on the data from the other five views. The rotation augmentation along the z-axis was only used for training the RotNet module. For V2V training, the same rotation augmentation was adopted. We did not apply it to PN++ since it already has a preprocessing module for orientation normalization [11].
- Cross-Condition-Cross-View (CC&CV): Under this protocol, data split was conducted following both CC and CV settings, which is expected to highlight the robustness under a more challenging scenario.
- Real data (Real-CC): To validate the capability of our method in real-world scenarios, the models trained with synthetic data were directly applied on real data of four abnormal gait patterns represented in the canonical view.

In terms of the quantitative analysis on shape reconstruction, Chamfer distance (CD) was used for evaluation [12]. In specific, we sampled 30,000 points from the reconstructed human shape. For pose estimation evaluation, the Mean Per Joint Position error (MPJPE) in millimeters was calculated without any post-process. We report the MPJPE of all 9 joints for synthetic dataset and of 8 joints (w/o *Pelvis*) for real data.

D. Quantitative Results

1) 3D pose estimation: The comparison pose estimation results are presented in Table I. **CC validation:** PoseSDF achieves MPJPE of 20.07 mm, which markedly outperforms PN++ [11] and is slightly better than V2V [10].

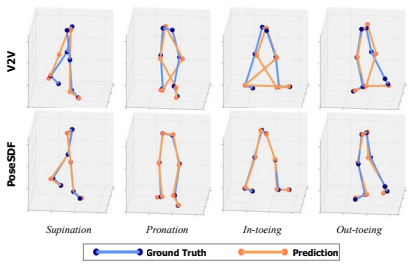

Fig. 4. Visualization of the pose estimation by V2V and PoseSDF on abnormal gait patterns captured from multiple views under **CC&-CV validation** (better view in color).

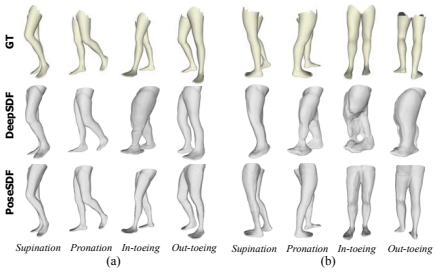

Fig. 5. Qualitative reconstruction results by DeepSDF and our PoseSDF. a) results under **CC validation**; b) results under **CC&CV validation**.

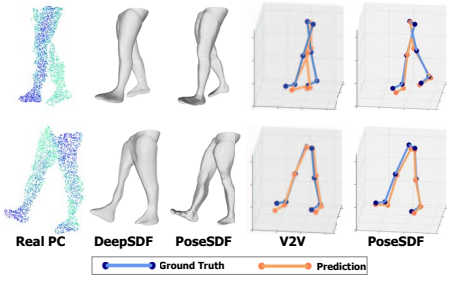

Fig. 6. Visualization of both shape reconstruction (DeepSDF v.s. PoseSDF) and pose estimation (V2V v.s. PoseSDF) results by directly taking **real** partial point cloud as input. (better view in color).

TABLE I

COMPARISON RESULTS ON 3D GAIT POSE ESTIMATION

Methods	MPJPE(mm)↓			
	CC	CV	CC&CV	Real-CC
PN++[11]	73.51	199.88	195.06	258.09
V2V[10]	20.23	132.53	122.35	79.83
PoseSDF (w/o RotNet)	-	267.81	218.95	-
PoseSDF (w/o Encoder)	22.13	28.08	35.58	57.41
PoseSDF (w/o Exchange)	28.28	34.90	40.98	73.56
PoseSDF	**20.07**	**26.27**	**34.25**	**56.18**

TABLE II

COMPARISON RESULTS OF CD METRIC ON 3D SHAPE RECONSTRUCTION

Methods	CD (1e-3) ↓			
	CC	CV	CC & CV	Real-CC
DeepSDF[12]	2.33	13.38	11.42	18.42
PoseSDF (w/o RotNet)	-	12.64	11.95	-
PoseSDF (w/o Encoder)	2.11	2.22	2.31	16.26
PoseSDF (w/o exchange)	2.40	2.32	2.40	16.99
PoseSDF	**1.97**	**2.12**	**2.15**	**16.04**

The performance of *PoseSDF (w/o encoder)* drops by 10% due to the competition during the joint training of two branches. It is also found that *PoseSDF (w/o exchange)* performs much worse than the full version and the one w/o encoder. This is because that the latent code is optimized completely based on the shape decoder during the inferring phase, leading to large discrepancies over the trained latent embedding. **CV and CC&CV validation:** Under these two challenging cross-view protocols, our PoseSDF incorporating with the RotNet model still performs best in terms of MPJPE, with an acceptable degradation compared to that under CC protocol. Without the RotNet, the ablated model cannot estimate the correct pose, resulting in the MPJPE over 100 mm. **Real-CC validation:** While directly applying the models trained with synthetic data to real-world data, V2V shows kinds of robustness against domain gap between real and synthetic data, whereas PoseSDF performs much better (MPJPE=56.18 mm). This comparison highlights the generalization capability of our PoseSDF to real raw partial scan without any domain adaptation techniques.

2) 3D shape reconstruction: Table II shows the comparison results on shape reconstruction. It can be observed that the proposed PoseSDF outperforms the DeepSDF and its ablations under all protocols. Especially, *w/o Encoder* and *w/o exchange* show similarly inferior results across all protocols. The main reason is that two branches are competed with each other during the training, while in the inference, the latent shape code is optimized by only using shape SDF. Such a situation will cause the distinction in the latent code distributions of the training and test data. Due to the noisy input of real partial scan, the CD metrics of all methods are clearly degrading, but PoseSDF is still better than others.

E. Qualitative Results

Fig. 4 demonstrates several pose estimation results by V2V and PoseSDF under the CC&CV protocol. Thanks to the proposed RotNet module, our PoseSDF is able to generate more accurate joint positions in terms of cross-view scenarios. For the **CC validation** in Fig. 5(a), although DeepSDF can generate reasonable well shapes in the single view setting, our proposed PoseSDF can preserve more fine details, showing good potential in human biomechanical analysis. In terms of the challenging **CC&CV validation**, DeepSDF cannot reconstruct the correct shapes in most cases. On the contrary, the shapes generated by PoseSDF are satisfactory, only with a minor defect on the thickness of thighs and the fine details of feet. Fig. 6 shows the visualization results under **Real-CC validation**. For both shape reconstruction (compared to DeepSDF) and pose estimation (compared to V2V [10]), it can be seen that PoseSDF has better fine details, especially for the feet. This also indicates that simultaneously performing these two body modeling tasks can mutually improve each other.

V. CONCLUSION

This paper proposes PoseSDF for reconstructing the 3D lower-limb shape from a partial point cloud and simultaneously estimate the 3D pose. This is achieved by regarding the estimation of each joint as the reconstruction of a small sphere centered on this joint. In specific, two branches, i.e., Shape decoder and Pose decoder, are designed to provide coherent performance boost. Furthermore, a front-end Encoder and a feature exchanging strategy are proposed to promote the multi-task training of shape and pose estimation. To achieve satisfactory results under cross-view scenarios, a dedicated RotNet is developed to normalize the input to its canonical view. In the future, we will generalize our work to 3D rotation equivariant, as well as ease the optimization of latent shape code during the inference.

REFERENCES

[1] D. Portugal, P. Alvito, E. Christodoulou, G. Samaras, and J. Dias, "A study on the deployment of a service robot in an elderly care center," *International Journal of Social Robotics*, vol. 11, no. 2, pp. 317–341, 2019.

[2] C. Zimmermann, T. Welschehold, C. Dornhege *et al.*, "3d human pose estimation in rgbd images for robotic task learning," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*. IEEE, 2018, pp. 1986–1992.

[3] S. Liu, G. Tian, Y. Zhang, and P. Duan, "Scene recognition mechanism for service robot adapting various families: A cnn-based approach using multi-type cameras," *IEEE Transactions on Multimedia*, 2021.

[4] W. Chi, J. Wang, and M. Q.-H. Meng, "A gait recognition method for human following in service robots," *IEEE Trans. Syst., Man, and Cybern., Syst.*, vol. 48, no. 9, pp. 1429–1440, 2017.

[5] Y. Guo, F. Deligianni, X. Gu, and G.-Z. Yang, "3-d canonical pose estimation and abnormal gait recognition with a single rgb-d camera," *IEEE Robotics and Automation letters*, vol. 4, no. 4, pp. 3617–3624, 2019.

[6] D. Drover, R. MV, C.-H. Chen, A. Agrawal, A. Tyagi, and C. Phuoc Huynh, "Can 3d pose be learned from 2d projections alone?" in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, September 2018.

[7] Y. Guo, Y. Li, and Z. Shao, "Rrv: A spatiotemporal descriptor for rigid body motion recognition," *IEEE Transactions on Cybernetics*, vol. 48, no. 5, pp. 1513–1525, 2017.

[8] Y. Guo, Y.-F. Li, and Z. Shao, "Dsrf: A flexible trajectory descriptor for articulated human action recognition," *Pattern Recognition*, vol. 76, pp. 137–148, 2018.

[9] J. Shotton, T. Sharp, A. Kipman *et al.*, "Real-time human pose recognition in parts from single depth images," *Commun. ACM*, vol. 56, no. 1, pp. 116–124, 2013.

[10] G. Moon, J. Y. Chang, and K. M. Lee, "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*, 2018, pp. 5079–5088.

[11] L. Ge, Y. Cai, J. Weng, and J. Yuan, "Hand pointnet: 3d hand pose estimation using point sets," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8417–8426.

[12] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove, "Deepsdf: Learning continuous signed distance functions for shape representation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 165–174.

[13] L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger, "Occupancy networks: Learning 3d reconstruction in function space," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4460–4470.

[14] V. Reijgwart, A. Millane, H. Oleynikova, R. Siegwart, C. Cadena, and J. Nieto, "Voxgraph: Globally consistent, volumetric mapping using signed distance function submaps," *IEEE Robotics and Automation Letters*, vol. 5, no. 1, pp. 227–234, 2019.

[15] G. Li, F. Caponetto, E. Del Bianco, V. Katsageorgiou, I. Sarakoglou, and N. G. Tsagarakis, "A workspace limit approach for teleoperation based on signed distance function," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 5589–5596, 2021.

[16] S. Peng, Y. Zhang, Y. Xu, Q. Wang, Q. Shuai, H. Bao, and X. Zhou, "Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 9054–9063.

[17] M. Atzmon and Y. Lipman, "Sal: Sign agnostic learning of shapes from raw data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2565–2574.

[18] X. Gu, J. Yang, H. Zhang, J. Qiu, F. P. W. Lo, Y. Guo, G.-Z. Yang, and B. Lo, "Occlusion-invariant rotation-equivariant semi-supervised depth based cross-view gait pose estimation," *arXiv preprint arXiv:2109.01397*, 2021.

[19] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 793–802.

[20] J. Malik, I. Abdelaziz, A. Elhayek, S. Shimada, S. A. Ali, V. Golyanik, C. Theobalt, and D. Stricker, "Handvoxnet: Deep voxel-based network for 3d hand shape and pose estimation from a single depth map," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7113–7122.

[21] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "Smpl: A skinned multi-person linear model," *ACM transactions on graphics (TOG)*, vol. 34, no. 6, pp. 1–16, 2015.

[22] X. Wen, T. Li, Z. Han, and Y.-S. Liu, "Point cloud completion by skip-attention network with hierarchical folding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1939–1948.

[23] J. Mu, W. Qiu, A. Kortylewski, A. Yuille, N. Vasconcelos, and X. Wang, "A-sdf: Learning disentangled signed distance functions for articulated shape representation," 2021.

[24] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," in *European conference on computer vision*. Springer, 2020, pp. 405–421.

[25] B. L. Bhatnagar, C. Sminchisescu, C. Theobalt, and G. Pons-Moll, "Loopreg: Self-supervised learning of implicit surface correspondences, pose and shape for 3d human mesh registration," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[26] Z. Yang, S. Wang, S. Manivasagam, Z. Huang, W.-C. Ma, X. Yan, E. Yumer, and R. Urtasun, "S3: Neural shape, skeleton, and skinning fields for 3d human modeling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 13 284–13 293.

[27] S. Wang, A. Geiger, and S. Tang, "Locally aware piecewise transformation fields for 3d human mesh registration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7639–7648.

[28] W. E. Lorensen and H. E. Cline, "Marching cubes: A high resolution 3d surface construction algorithm," *ACM siggraph computer graphics*, vol. 21, no. 4, pp. 163–169, 1987.

[29] Z. Huang, Y. Yu, J. Xu, F. Ni, and X. Le, "Pf-net: Point fractal network for 3d point cloud completion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7662–7670.

[30] V. Sitzmann, E. R. Chan, R. Tucker, N. Snavely, and G. Wetzstein, "Metasdf: Meta-learning signed distance functions," *arXiv preprint arXiv:2006.09662*, 2020.

[31] Y. Duan, H. Zhu, H. Wang, L. Yi, R. Nevatia, and L. J. Guibas, "Curriculum deepsdf," in *European Conference on Computer Vision*. Springer, 2020, pp. 51–67.

[32] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," *arXiv preprint arXiv:1706.02413*, 2017.

[33] G. Varol, J. Romero, X. Martin, N. Mahmood, M. J. Black, I. Laptev, and C. Schmid, "Learning from synthetic humans," in *CVPR*, 2017.