JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 38, 963-975 (2022)
DOI: 10.6688/JISE.202209_38(5).0006

Skeletal Joint-based Regressive 3D Human Reconstruction From Partial Point Clouds

JONATHAN THEN SIEN PHANG+, KING HANN LIM AND PO KEN PANG
Department of Electrical and Computer Engineering
Curtin University Malaysia
Miri, 98009 Malaysia
E-mail: jonathanpts@postgrad.curtin.edu.my+; {glkhann; ppoken}@curtin.edu.my

Three dimensional (3D) human model acquisition using single-viewpoint provides flexible setup with a trade-off of partial point clouds. By leveraging the learning capability of neural networks, a complete point clouds can be obtained by sampling a reconstructed 3D human model. A skeletal joints-based regressive 3D human reconstruction is proposed in this paper to infer skeletal joints and variance for reconstructing human shape. A skeletal joints encoder is proposed to learn the latent representation of input partial point clouds using Gaussian maximum likelihood to obtain localized skeletal joints and its variances. The skeletal joints are fed to a regressive human model to reconstruct a synthetic human model to obtain a complete point cloud. Lastly, a two-mode training strategy is proposed to enhance the learning of the proposed method by employing synthetic training in prior and non-synthetic fine-tuning. A real human motion dataset is used in the experiment as the performance evaluation to study the skeletal joints estimation and human shape learning.

Keywords: 3D human model, 3D reconstruction, partial point cloud, skeletal joint, body regressor

1. INTRODUCTION

The reconstruction of three dimension (3D) human model from point clouds [1, 2] is of great interest in computer graphic and computer vision due to its wide applications, such as personalized human model in mixed reality applications [3], human kinematic measurements in sports bio-mechanics [4] and human modeling in video-based motion capture [5]. The common acquisition using multiple sensors can capture accurate 3D human joints from viewpoint occlusion, it is resource expensive in term of equipment, designated space and computation. As a consequence, it is not an ideal solution in applications that require portability and real-time processing. On the other hand, reconstructing a high-fidelity human shape [6, 7] using single viewpoint is challenging because of non-rigid human deformations, low-quality input data, and occlusion. By leveraging the advancement in Deep Learning (DL) techniques [8, 9], these challenges can be overcome using the learning capability of neural networks to achieve human model reconstruction.

Received August 6, 2021; revised January 17, 2022; accepted February 14, 2022.
Communicated by Huo Chong Ling.
+Corresponding author.

Inspired from recent research trend in point cloud reconstruction using autoencoder [10–12] and regressive 3D human model-based reconstruction [13, 14], a skeletal joints-based regressive 3D Human reconstruction model is proposed in this paper with the integration of a skeletal joints encoder and a 3D human model regressor. The proposed method shares a similarity to conventional autoencoder where the skeletal joints encoder is weakly-supervised. This is to maximize the learning of intrinsic property of sparse point cloud. These predicted skeletal joints are then used to guide the human model regressor to produce reconstructed point cloud in full shape. Differentiating from conventional autoencoder, the proposed method utilizes a 3D human model regressor to replace the decoder setup due to decoder-based reconstruction tend to reproduce an output that resemble a training data. Therefore, it is prone to produce noisy reconstructed output especially when trained on non-synthetic data, in which the presence of noises are constant.

2. RELATED WORKS

The use of depth sensors [7] such as structured light, time of flight and stereo visions can provide depth map and point cloud representation. By applying DL approach using 3D structure information can effectively locate and reconstruct human body joints [15,16] in 3D space. Generally, 3D human shape reconstruction can be divided into two categories [16], *i.e*. model-free and model-based 3D reconstruction. The model-free methods do not employ human body models to reconstruct 3D human representation while the model-based methods incorporate parametric body models in the part learning.

2.1 Model-Free 3D Reconstruction

As model-free 3D reconstructions do not take input of human body models with depth information to search for human representation, they predict 3D human pose from 2D images with intermediately estimated 2D pose representation. The common use of deep CNN [17–19] is deployed for 3D human pose estimation. Li and Chan [17] proposed the use of deep CNN for 3D human pose estimation. Their framework was jointly trained with pose regression and body part detectors to achieve pose projection in 3D space. Tekin *et al*. [20] combined the traditional CNNs with autoencoder for structured learning to represents the 3D pose from 2D image. A high-dimensional latent pose representation learned by autoencoder was reprojected to original pose space with a decoding layer to account for joint dependencies.

Besides that, implicitly learning of the pose structure from 2D data has recently drawn into attention to infer 3D human pose with two separate sequential training steps. They first perform 2D joint prediction and then reconstruct the 3D pose via optimization or search. Pavlakos *et al*. [18] applied ConvNet for 2D joint location and subsequently perform optimization step to recover 3D pose. They used volumetric representation for 3D human pose and employed a coarse-to-fine prediction scheme to do refinement in 3D pose estimation. To improve the 3D ground truth accuracy, they used the ordinal depths of human joints as the supporting signal to perform weakly supervision in the 3D human pose learning [21]. Sun *et al*. [22] proposed the use of bones instead of joints as pose representation. Subsequently, it exploited the joint connection structure to define a

compositional loss function that encodes long range interactions between the bones.

Zhao *et al*. [23] pre-trained a 2D pose estimation network to predict 2D joint locations. Subsequently, a semantic graph convolutional network is trained to predict 3D pose from 2D joints features. Cheng *et al*. [24] proposed an occlusion-aware DL framework to estimate a 2D confidence heatmaps of keypoints. With the optical flow consistency constraint, unreliable estimations of occluded keypoints were filtered and subsequently fed into 3D temporal convolutional networks to produce a complete 3D pose. Xu *et al*. [25] performed deep kinematics analysis using 2D noisy pose inputs to obtain 3D pose estimation concurrently by considering the static and dynamic body structures. The limitation of model-free 3D reconstruction is restricted by its accuracy of depth estimation from sensors because the captured data contains numerous artifacts such as occlusion, outliers and non-uniform surface.

2.2 Model-based 3D Reconstruction

While Model-free approaches are relying on estimation of depth from 2D images, hence the predicted 3D pose can be unreliable due to estimation process. Moreover, Model-free approaches do not provide expressive visual output of 3D human such as parametric body shapes. The Model-based 3D reconstruction [**?**,26–29] on the other hand incorporates parametric body shapes such as body pose and body volume to perform 3D human reconstruction. Widely used volumetric models for synthetic human body construction are Fine Alignment Using Scan Texture (FAUST) model [26], Skinned Multi-Person Linear (SMPL) model [27]. Litany *et al*. [30] proposed the use of variational autoencoder incorporated with graph convolutional operations for the completion of human shape reconstruction. It learns a latent space for complete realistic shape with vertex-wise correspondence using FAUST synthetic dataset. In recent works, SMPL model is commonly used to perform 3D parameter estimation to construct a full human body. Several extended SMPL-based models such as SMPLify [**?**] and Vposer [28] are devised to reconstruct 3D human model through skeletal joints regression. The latter, Sparse Trained Articulated Human Body Regressor (STAR) [29] is introduced with improvement over SMPL by training with additional 14,000 human subjects and a learning set of sparse local pose corrective blend shapes. In addition, the number of parameters in STAR is reduced to 20% of that in SMPL model.

Kinematic model in 3D space has gained increasing attention in 3D human pose estimation recently because it is a realistic and accurate articulated body representation. By using single depth sensor, Zhou *et al*. [5] infers 3D joint positions from partial point cloud with a 3D pose regression network without 3D human reconstruction. Phang *et al*. [12] proposed an end-to-end training mechanism to learn skeletal joints inference and reconstructs complete human point clouds from partial point clouds. It is a generative reconstruction method comprises a skeletal joints-based autoencoder. In recent trend, kinematic and synthetic 3D human models are used to enhance 3D human reconstruction. For example, Jiang *et al*. [13] proposed to incorporate skeleton joints into a DL network for 3D human shape reconstruction. The basic structure of this model uses PointNet++ to extract point features and then map point features to skeleton joint features and finally SMPL parameters for the 3D human reconstruction. In general, SMPL offers a simple integration and compact representation for 3D human reconstruction. Overall, Model-based can reconstruct an expressive output of 3D human, however most existing methods

that reconstruct a 3D human are inferring on complete point clouds rather than partial point clouds.

3. SKELETAL JOINTS-BASED REGRESSIVE 3D HUMAN RECONSTRUCTION

Point clouds acquisition from widely accessible depth sensors often possess attributes of un-ordered, non-uniform and sparse data distribution in the 3D space resulting low quality input. Further, due to single viewpoint acquisition, acquired human point cloud may suffer from occlusion of human parts and non-rigid deformation that can greatly degrades the performance of human model reconstruction and skeletal joints estimation. To overcome this, a DL architecture skeletal joints-based regressive 3D human point cloud reconstruction is proposed to reconstruct 3D human model using input partial point cloud on two-stage operations as illustrated in Fig. 1. By leveraging the learning capability of DL, point clouds learning can be efficiently achieved by training process of a DL network. For an instance, by training a DL network to learn latent representation of input partial point cloud, a 3D pose can be inferred from the latent representation [12]. Consequently, this also provides generalization towards variation in input point clouds such as low quality input and non-rigid deformation by fine tuning the network. The proposed method operates in two stages, comprises a skeletal joints encoder to determine localized skeletal joint and variance in 3D space, and a regressive 3D human reconstruction model to obtain complete point clouds from synthetic 3D human. The proposed method is optimized using two-mode training mechanism comprises synthetic human model training and real-world dataset fine tuning.

Due to the nature of non-uniform surface, sparse and irregular density data distribution of acquired partial point cloud [31], farthest point sampling (FPS) algorithm [32] is implemented in the pre-processing step to evenly distribute input 3D partial point cloud $S_1 = \{p_i\}_{i=1}^{N_1}$ in a grid-less space. Subsequently, 3D space skeletal joints components $C = (M, \Sigma)$, where $M = \{\mu_k\}_{k=1}^{K}$ is skeletal joints and the variance $\Sigma = \{\sigma_k\}_{k=1}^{K}$, is in-

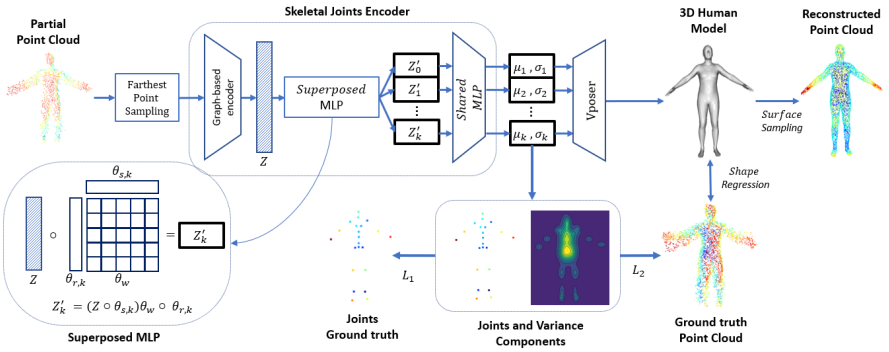


Fig. 1. The proposed skeletal joints-based regressive 3D human reconstruction network is trained using two-mode strategy, *i.e.* synthetic training and fine-tuning on real-world dataset.

ferred from the input partial point cloud using complete point cloud $S_2 = \{p_i\}_{i=1}^{N_2}$ as ground truth. N_1 and N_2 are the number of points in respective point cloud and each k-th local human part is governed by each skeletal joint μ_k and a variance σ_k. In order to infer the skeletal joints components, latent representation of input partial point cloud is obtained using a skeletal joint encoder denoted by f_1 as follows,

$$Z = f_1(\theta_g, \mathscr{G}), \tag{1}$$

where θ_g is network parameter of graph-based point cloud encoder. The skeletal joints encoder implemented a graph-based encoder [33] to encode latent features $Z \in \mathbb{R}^d$ of input point cloud, where d is the dimension of the latent features. The latent feature encoding is achieved by constructing directed k-nearest neighbour (k-NN) graphs $\mathscr{G} = (V,E)$. $V = \{v_i\}_{i=1}^{N_1}$ and $V = S_1$ is vertices from the input partial point cloud and $E = \{e_{ij_k}\}_{k=1}^{K}$ are the edges of k-nearest neighbouring vertices v_j respective to v_i. The edges are computed based on k-smallest pair-wise distance, such that:

$$E = \{e_{ij_k} : e_{ij_k} > e_{ij} \mid \quad \forall j \neq i, \forall j_k \neq j, \{i,j\} \in N\}_{k=\{1,\ldots,K\}}, \tag{2}$$

where e_{ij} are the edge features for each point v_i as follows,

$$e_{ij} = ||V||^2 - 2V^T V + ||V^T||^2. \tag{3}$$

By constructing directed graph of input point cloud, it allow aggregation of global information of input partial point cloud with notion of local neighbourhood geometrical details into a condensed latent representation which provides latent inference in the latter stage of the proposed method. Following that, a vectorized stacked MLPs denoted by f_2 is proposed inspired from [12] to infer skeletal joints components from the latent features as follows,

$$Z' = f_2(\theta_s, \theta_w, \theta_r, Z). \tag{4}$$

where $Z' = \{Z_k^{'} \in \mathbb{R}^{d'}\}_{k=1}^{K}$ is latent representation of skeletal joints component and $\theta_s, \theta_w, \theta_r$ are network parameters of vectorized stacked MLPs. For comparison, latent inference implemented in the concurrent work [12] is performed in iterative forward propagation of independent MLPs, hence huge forward propagation time can incur in the latent inference due to iterations. In contrast, the proposed latent inference vectorized independent MLPs into matrices to eliminate the iterations operation resulting in largely reduced forward propagation time by taking advantage of matrix multiplication operation. Furthermore, the vectorization of network parameters enabled implementation

Table 1. Comparison of parameters and operations of stack MLPs and superposed MLP.

MLP	Parameters configuration	Operation
Stack [12]	$(d \times d' \times K)$	$Z' = \{\theta_k \circ Z\}_{k=1}^{K}$
Superposed	$(d \times d') + (d \times K) + (d' \times K)$	$Z' = ((Z \circ \theta_s)\theta_w) \circ \theta_r$

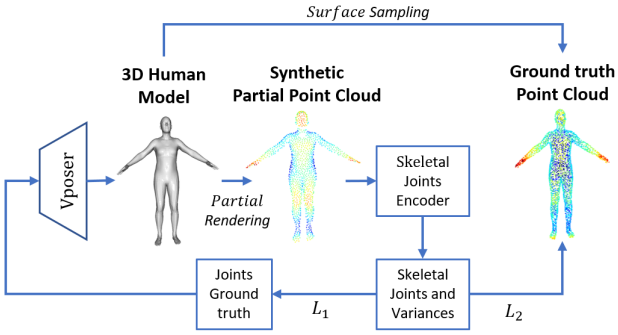


Fig. 2. The training of the proposed skeletal joint encoder on synthetic partial point cloud rendered from Vposer 3D human model regressed on ground truth joints.

of superposed matrix compression, *i.e.* matrix decomposition [34] to reduce the overall network parameters. In Table 1, the network parameters configuration and operations of proposed and concurrent work is tabulated for detailed comparison, where $\theta_k \in \mathbb{R}^{d \times d' \times K}$, $\theta_w \in \mathbb{R}^{d \times d'}$, $\theta_r \in \mathbb{R}^{K \times d'}$, $\theta_s \in \mathbb{R}^{K \times d}$.

$$C = f_3(\theta_p, Z') \tag{5}$$

Next, latent representation of skeletal joints component is inferred and subsequently projected into ambient space by employing a shared MLP denoted with f_3 with network parameter θ_p. After obtaining the estimated skeletal joints and variances, the estimated skeletal joints are fed into Vposer [28] denoted with f_{vposer} , a fully differentiable regressive 3D human that directly reconstruct a 3D human $\mathbb{H} \in \mathbb{R}^{3 \times N_H}$, where $N_H = 6890$ is number of vertices of the 3D human mesh, from human joints as follows,

$$\mathbb{H} = f_{vposer}(M). \tag{6}$$

Finally, surface sampling using Barycentric coordinate interpolation [35] is implemented to obtain reconstructed 3D human point cloud such that $S_3 \sim \mathbb{H}$ and $S_3 = \{p_i\}_{i=1}^{N_3}$, where N_3 is number of point in reconstructed 3D human point cloud.

4. 3D HUMAN RECONSTRUCTION MODEL TRAINING

In order to achieve skeletal joints estimation and 3D human reconstruction from partial point cloud, a large dataset of partial point clouds that covers vast majority of possible human pose is required for network training. Moreover, it is resource expensive to acquire complete point clouds and skeletal joints of the corresponding partial point clouds. Existing works [12, 13] on network training directly used real-world dataset as their training and testing data. However, method in [13] is only applicable for complete point clouds inference. While output of [12] is non-synthetic, it may carry forward noises that are present in training data and less expressive in modeling a 3D human such as body shape.

To overcome the aforementioned challenges, a two-mode training phase is proposed in the model training as illustrated in Fig. 2, where the first training phase learns general human shape and pose from synthetic partial point cloud synthesised from Vposer [28] regressed from ground truth joints. The synthetic partial point cloud is partially rendered using rendering technique from [36] on single-viewpoint. The second training phase fine-tunes the proposed method by using real-world dataset training data as input and guided by ground truth joints. A set of scaling and shape parameters are regressed in the fine-tuning process to capture the human size and shape of real-world dataset. This is to further adapt the proposed method toward non-synthetic nature of real-world dataset captured from a depth sensor.

The network training uses weakly supervised joint loss $L = L_1 + L_2$ with mean squared error (MSE) shown in Eq. (7) and Gaussian maximum likelihood shown in Eq. (8) as follows,

$$L_1(\mu, \mu_{gt}) = \frac{1}{K} \sum_{k=1}^{K} ||\mu_k - \mu_{gt,k}||_2^2, \tag{7}$$

$$L_2(S_2|\mu, \sigma) = \sum_{i=1}^{N} \prod_{k=1}^{K} \mathcal{N}(p_i|\mu_k, \sigma_k), \tag{8}$$

where,

$$\mathcal{N}(p_i|\mu_k, \sigma_k) = \frac{1}{(2\pi)^{\frac{3}{2}} |\sigma_k|^{\frac{1}{2}}} e^{-\frac{1}{2}(p_i-\mu_k)^T \sigma_k^{-1} (p_i-\mu_k)}. \tag{9}$$

The MSE of skeletal joints is computed between the estimated skeletal joints with the ground truth to optimize encoder's parameter learning. Gaussian maximum likelihood function is implemented to improve the estimation by maximizing the likelihood of estimated skeletal joints and its variances respective to complete point clouds.

5. EXPERIMENT SETUP

The experimental evaluation of the proposed method is developed based on Berkeley Multimodal Human Action Database (MHAD) [37] dataset. MHAD is a non-synthetic human point cloud dataset that contains 11 actions performed by 12 human subjects. All subjects performed five repetitions of each action, yielding 660 action sequences which correspond to averagely 200 frames per action sequence. The dataset is split into four repetitions for training set and one repetition for testing set for each subject. As for performance bench-marking [13, 14], Action #1 (jumping) and Action #2 (jumping jack) are used to perform evaluation to demonstrate the working principle of the proposed method. Each sample is normalized into $[-0.5, 0.5]^3$ using the hip skeletal joints as the centroid of a bounding box with volume of approximately 2.0 cubic meter.

In this experiment, the number of skeletal joints is set $K = 22$ as dictated by Vposer. The number of points of point clouds are set $N_1 = 1024$, $N_2 = 6890$ and $N_3 = 6890$. The latent feature dimension is set $d = 1024$ and $d' = 512$. For the evaluation metrics, the average point-to-vertex distance d_{p2v} in Eq. (10) and average vertex-to-point distance d_{v2p}

in Eq. (11) are used to evaluate the precision and recall performance of the reconstructed human vertices and ground truth point cloud. In order to compute an overall reconstruction quality ground truth and reconstructed point clouds, an average Chamfer distance $d_{CD} = d_{p2v} + d_{v2p}$ is implemented due to its efficient computation and permutation invariant properties.

$$d_{p2v}(S_2, S_3) = \frac{1}{|S_2|} \sum_{x \in S_2} \min_{y \in S_3} ||x - y||_2^2 \tag{10}$$

$$d_{v2p}(S_2, S_3) = \frac{1}{|S_3|} \sum_{y \in S_3} \min_{x \in S_2} ||x - y||_2^2 \tag{11}$$

The training of the proposed method is set for 100 epochs with scheduling decay rate of 0.5 per 20 epoch and starting learning rate is set 1×10^{-4}. Parameters are initialized using Xavier normal and batch size 64 is used. ADAM optimizer is implemented as optimization method for both parts of training. Meanwhile, the fine-tuning is set for 5 epoch with scheduling decay rate of 0.5 per 20 epoch and starting learning rate is set 1×10^{-5}. The results of concurrent work [12] is recreated and trained using non-synthetic real-world dataset generated in the proposed method. All networks are built and executed on Pytorch 1.8.0 with batch size of 128. The specifications of test bench for the experiments are Intel-i7-4790K with 32GB RAM and Quadro P6000 GPU with 24GB VRAM.

6. PERFORMANCE EVALUATION

The qualitative results of the proposed method in complete 3D human point cloud reconstruction from real-world non-synthetic partial point clouds is demonstrated in Fig. 3. Fig. 3 (a) illustrates a sequence of partial point clouds acquired from single-viewpoint depth sensor. Subsequently, by feeding the estimated skeletal joints illustrated in Fig. 3 (b) into Vposer, 3D human models are reconstructed illustrated in Fig. 3 (c). Through surface sampling on the reconstructed 3D human model, the complete human point cloud is obtained as illustrated in Fig. 3 (d). Due to the surface sampling process, the surface quality of complete human point cloud is significantly superior compared to [12]. Moreover, noises such as outliers that can occur during reconstruction from a decoder is avoided.

Table 2 shows the evaluation of average joints deviation and average reconstruction loss comparing the proposed method and adopted work [12]. From the evaluation, the proposed method with fine tuning outperformed the existing work with 24.83*mm* in joints deviation compared to 45.29*mm*. Moreover, a significant reconstruction quality improvement is shown with lower Chamfer distance of 0.84×10^{-3} as compared to 1.0×10^{-3}. Notably, the improvement in average joints deviation is contributed by learning on synthetic data in prior which contain more concise information of partial point clouds. Furthermore, the improvement of reconstruction quality is due to uniform surface sampling of complete point cloud. As oppose to reconstruction from scratch in the adopted work, the technique can capture the artifacts such as non-uniform surface and outliers in training data and reproduce the artifacts in the output. In the case of the proposed method without fine tuning, the joint deviation and reconstruction loss are significantly higher. This is

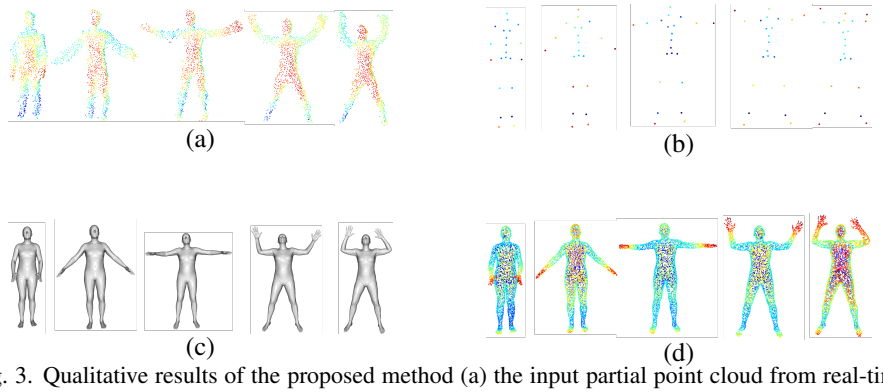


Fig. 3. Qualitative results of the proposed method (a) the input partial point cloud from real-time, (b) the estimated skeletal joints using input partial point cloud, (c) the 3D human mesh, and (d) full reconstructed human point cloud.

because the network is not adapted to non-synthetic data modality when solely trained on synthetic data.

Table 3 shows the quantitative results of mean and max distance measurements on Action #1 and Action #2 in Berkeley MHAD dataset [37] comparing to existing works. The evaluation is based on the point-to-vertex d_{p2v} and vertex-to-point d_{v2p} metric comparing ground truth and reconstructed point clouds. Essentially, the evaluation measures the quality of human shape reconstruction. The proposed method with fine-tuning outperformed the adopted work in overall d_{p2v} and d_{v2p}. This indicates superior shape reconstruction as compared to decoder-based reconstruction. Without fine-tuning, the proposed method has significantly larger shape deviation compared to ground truth point clouds as

Table 2. The average estimated joints deviation against ground truth joints in millimeter (mm) and average reconstruction loss d_{CD} with (w/) and without (w/o) fine tuning (FT).

Methods	Joints Deviation (mm)	d_{CD} ($\times 10^{-3}$)
Phang *et al*. [12]	45.29	1.0
Ours (w/o FT)	71.27	1.29
Ours (w/ FT)	**24.83**	**0.84**

Table 3. Quantitative results of mean and max distance in millimeter (mm) measurements on two action sequences in Berkeley MHAD dataset [37]. Note that in each cell, the first and second numbers denote the distances d_{p2v}/d_{v2p} respectively.

Methods	Action #1		Action #2	
	mean	max	mean	max
SMPLify [14]	31.1/41.1	43.4/58.8	31.3/39.7	48.6/58.4
Jiang *et al*. [13]	21.4/23.5	28.6/34.7	16.9/18.2	21.5/21.2
Phang *et al*. [12]	40.99/46.75	72.91/78.54	45.92/46.09	97.21/63.80
Ours (w/o FT)	37.33/43.70	60.84/85.74	33.81/43.76	41.64/65.78
Ours (w/ FT)	34.32/30.53	47.07/44.13	34.78/35.29	48.21/55.46

indicated by large max value of d_{p2v} and d_{v2p}. The large shape deviation is mainly because the model does not estimate the human shape. On the other hand, other existing work such as [**?**] first estimates the skeletal joints using third party model to regress a 3D human model, while [13] uses complete point clouds to regress a 3D human model. Although the existing methods [**?**, 13] do not infer partially occluded data, they are able to reconstruct more precise human shape.

7. CONCLUDING REMARKS

A skeletal joint-based regressive 3D human model reconstruction from partial point cloud is proposed in this paper. The proposed method first estimate the skeletal joints components of human partial point cloud and a regressive 3D human reconstruct a 3D human model based on the estimated skeletal joints. A superposed MLP comprises matrix vectorization and parameters superposition in the skeletal joints encoder is proposed to improve both the memory footprint and processing efficiency compared to adopted method. In addition, a two-mode model optimization strategy consisting network training on synthesized data and fine tuning on real-world dataset is proposed. This allows generation of simulated partial point cloud that would be resource expensive to acquire in real world. The proposed method is weakly supervised, where it is guided by ground truth joints and maximum likelihood against ground truth complete point cloud. The proposed method achieved an average of 24.83*mm* joint distance deviation against ground truth joints and 0.84×10^{-3} average Chamfer distance on reconstruction fidelity. In the future work, a pose-based encoder can be developed to directly drive the regression of 3D human model. This is due to Vposer is an inverse kinematic based 3D human regression method, hence the regressed 3D human model may present a slightly distorted pose and therefore affect the accuracy of reconstructed human shape.

ACKNOWLEDGMENT

This study was funded by Sarawak Multimedia Authority with the project ID SMA-1077. We would like to gratefully acknowledge the support of NVIDIA Corporation with the donation of the the Quadro P6000 GPU used for this research.

REFERENCES

1. W. Liu, J. Sun, W. Li, T. Hu, and P. Wang, "Deep learning on point clouds and its application: A survey," *Sensors*, Vol. 19, 2019, p. 4188.
2. Y. Jin, D. Jiang, and M. Cai, "3d reconstruction using deep learning: A survey," *Communications in Information and Systems*, Vol. 20, 2020, pp. 389-413.
3. M. K. Bekele, R. Pierdicca, E. Frontoni, E. S. Malinverni, and J. Gain, "A survey of augmented, virtual, and mixed reality for cultural heritage," *Journal on Computing and Cultural Heritag*, Vol. 11, 2018, pp. 1-36.

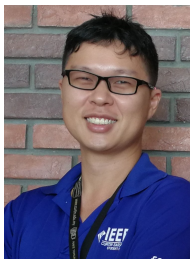
4. J. Wang, K. Qiu, H. Peng, J. Fu, and J. Zhu, "AI coach: Deep human pose estimation and analysis for personalized athletic training assistance," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 374-382.
5. Y. Zhou, H. Dong, and A. El Saddik, "Learning to estimate 3d human pose from point cloud," *IEEE Sensors Journal*, Vol. 20, 2020, pp. 12334-12342.
6. Z.-Q. Cheng, Y. Chen, R. R. Martin, T. Wu, and Z. Song, "Parametric modeling of 3d human body shape - a survey," *Computers & Graphics*, Vol. 71, 2018, pp. 88-100.
7. T. Xu, D. An, Y. Jia, and Y. Yue, "A review: Point cloud-based 3d human joints estimation," *Sensors*, Vol. 21, 2021, p. 1684.
8. G. Pavlakos, L. Zhu, X. Zhou, and K. Daniilidis, "Learning to estimate 3d human pose and shape from a single color image," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 459-468.
9. I. Ramirez, A. Cuesta-Infante, E. Schiavi, and J. J. Pantrigo, "Bayesian capsule networks for 3d human pose estimation from single 2d images," *Neurocomputing*, Vol. 379, 2020, pp. 64-73.
10. P. Achlioptas, O. Diamanti, I. Mitliagkas, and L. Guibas, "Learning representations and generative models for 3d point clouds," in *Proceedings of International Conference on Machine Learning*, 2018, pp. 40-49.
11. M. Liu, L. Sheng, S. Yang, J. Shao, and S.-M. Hu, "Morphing and sampling network for dense point cloud completion," in *Proceedings of AAAI Conference on Artificial Intelligence*, Vol. 34, 2020, pp. 11596-11603.
12. J. T. S. Phang, K. H. Lim, and P. K. Pang, "Generative skeletal joint-based autoencoder for 3d human point clouds reconstruction," in *Proceedings of IEEE International Conference on Green Energy, Computing and Sustainable Technology*, 2021, pp. 1-6.
13. H. Jiang, J. Cai, and J. Zheng, "Skeleton-aware 3d human shape reconstruction from point clouds," in *Proceedings of IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5431-5441.
14. 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 *Proceedings of International Conference on Computer Vision*, LNCS Vol. 9909, 2016, pp. 561-578.
15. 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, 2020, p. 102897.
16. C. Zheng, W. Wu, T. Yang, S. Zhu, C. Chen, R. Liu, J. Shen, N. Kehtarnavaz, and M. Shah, "Deep learning-based human pose estimation: A survey," *arXiv Preprint*, 2020, No. arXiv:2012.13392.
17. S. Li and A. B. Chan, "3d human pose estimation from monocular images with deep convolutional neural network," in *Proceedings of Asian Conference on Computer Vision*, 2014, pp. 332-347.
18. G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis, "Coarse-to-fine volumetric prediction for single-image 3d human pose," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7025-7034.

19. X. Zhou, Q. Huang, X. Sun, X. Xue, and Y. Wei, "Towards 3d human pose estimation in the wild: a weakly-supervised approach," in *Proceedings of IEEE International Conference on Computer Vision*, 2017, pp. 398-407.
20. B. Tekin, I. Katircioglu, M. Salzmann, V. Lepetit, and P. Fua, "Structured prediction of 3d human pose with deep neural networks," *arXiv Preprint*, 2016, No. arXiv:1605.05180.
21. G. Pavlakos, X. Zhou, and K. Daniilidis, "Ordinal depth supervision for 3D human pose estimation," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7307-7316.
22. X. Sun, J. Shang, S. Liang, and Y. Wei, "Compositional human pose regression," in *Proceedings of IEEE International Conference on Computer Vision*, 2017, pp. 2602-2611.
23. L. Zhao, X. Peng, Y. Tian, M. Kapadia, and D. N. Metaxas, "Semantic graph convolutional networks for 3d human pose regression," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3425-3435.
24. Y. Cheng, B. Yang, B. Wang, W. Yan, and R. T. Tan, "Occlusion-aware networks for 3d human pose estimation in video," in *Proceedings of IEEE/CVF International Conference on Computer Vision*, 2019, pp. 723-732.
25. J. Xu, Z. Yu, B. Ni, J. Yang, X. Yang, and W. Zhang, "Deep kinematics analysis for monocular 3d human pose estimation," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 899-908.
26. F. Bogo, J. Romero, M. Loper, and M. J. Black, "Faust: Dataset and evaluation for 3d mesh registration," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 3794-3801.
27. M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A skinned multi-person linear model," *ACM Transactions on Graphics*, Vol. 34, 2015, pp. 1-16.
28. G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. A. Osman, D. Tzionas, and M. J. Black, "Expressive body capture: 3d hands, face, and body from a single image," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10975–10985.
29. A. A. Osman, T. Bolkart, and M. J. Black, "Star: Sparse trained articulated human body regressor," in *Proceedings of the 16th European Conference on Computer Vision*, 2020, pp. 598-613.
30. O. Litany, A. Bronstein, M. Bronstein, and A. Makadia, "Deformable shape completion with graph convolutional autoencoders," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1886-1895.
31. M. Berger, A. Tagliasacchi, L. Seversky, P. Alliez, J. Levine, A. Sharf, and C. Silva, "State of the art in surface reconstruction from point clouds," in *Eurographics State of the Art Reports*, Vol. 1, 2014, pp. 161-185.
32. M.-H. Guo, J.-X. Cai, Z.-N. Liu, T.-J. Mu, R. R. Martin, and S.-M. Hu, "PCT: Point cloud transformer," *arXiv Preprint*, 2020, No. arXiv:2012.09688.
33. Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, "Dynamic graph cnn for learning on point clouds," *ACM Transactions on Graphics*, Vol. 38, 2019, pp. 1-12.

34. Y. Wen, D. Tran, and J. Ba, "Batchensemble: an alternative approach to efficient ensemble and lifelong learning," *arXiv Preprint*, 2020, No. arXiv:2002.06715.
35. M. Xu, W. Dai, Y. Shen, and H. Xiong, "Msgcnn: Multi-scale graph convolutional neural network for point cloud segmentation," in *Proceedings of IEEE 5th International Conference on Multimedia Big Data*, 2019, pp. 118-127.
36. W. Yuan, T. Khot, D. Held, C. Mertz, and M. Hebert, "PCN: Point completion network," in *Proceedings of IEEE International Conference on 3D Vision*, 2018, pp. 728-737.
37. F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy, "Berkeley MHAD: A comprehensive multimodal human action database," in *Proceedings of IEEE Workshop on Applications of Computer Vision*, 2013, pp. 53-60.

Jonathan Then Sien Phang received his bachelor degree in Electronics and Communication from Curtin University Malaysia in 2019. He is currently pursuing his Ph.D. in the Electronics and Computer Department at Curtin University Malaysia. He focuses on the application of DL in human gait rehabilitation. His current research interest involves developing learning-based methods in 3D reconstruction for gait analysis applications.

King Hann Lim received the M.Eng. and Ph.D. degrees in Electrical and Electronic Engineering from the University of Nottingham, in 2007 and 2012, respectively. He is currently an Associate Professor with the Department of Electrical and Computer Engineering in Curtin University Malaysia. His research areas include computer vision, artificial intelligence, and optimization. He has been publishing more than 80 journals and conference papers in the related areas of his research interest.



Po Ken Pang received his bachelor degree of Engineering in Electronics and Computing Engineering from the University Malaysia Sarawak in 2011 and Master of Science in Electronic Systems Design Engineering from University Science Malaysia in 2012. He is currently a staff member with the Department of Electrical and Computer Engineering, Curtin University Malaysia. His research areas are image processing and artificial intelligence.