Graph Convolutional Adversarial Network for Human Body Pose and Mesh Estimation

YUANCHENG HUANG AND NANNFENG XIAO

School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China

Corresponding author: Yuancheng Huang (yuancheng.huang@outlook.com)

This work was supported in part the Public Research and Capacity Building of Guangdong Province under Grant 2014B010104001 and in part by the Basic and Applied Basic Research of Guangdong Province under Grant 2015A03030 8018.

ABSTRACT

This paper studies reconstruction of human body shape and pose from a single-view image. While most of current work attempts to regress parameters of human body model such as Skinned Multi-Person Linear Model (SMPL) and Hand Model with Articulated and Non-rigid Deformations (MANO), these parametric approaches underperform compared to non-parametric approaches. Due to the lack of the spatial relationship in the input image, the parametric approaches are hardly used to reconstruct the human body precisely. Besides, the rotation parameter regression is a complex task in parametric approaches. Therefore, we introduce a novel graph convolutional neural network (Graph CNN)-based framework for estimating a non-parametric mesh model. Our key innovation is that the proposed model is trained in a generative adversarial manner. Firstly, Graph CNN utilizes mesh topology to capture integral information of the full 3D human shape and then generate a more smooth and high-quality human mesh model. Secondly, the discriminator in our network acts as a supervisor to specify whether a human shape and pose are real or not. The generator is encouraged to generate human body mesh that is close to the manifold of the real human mesh distribution. Extensive experimental results demonstrate the effectiveness of our proposed framework. In contrast to the state-of-the-art methods, our method can achieve better performance in human shape and pose estimation.

INDEX TERMS

Generative adversarial network, graph convolution network, human shape reconstruction.

I. INTRODUCTION

Reconstructing a 3D space from a 2D image has made a dramatic leap in the human body shape reconstruction. The human body shape reconstruction plays a necessary role in many scenes such as virtual reality games, movie special effects, product customization, augmented reality, etc. Traditionally, a 3D human body or object model can be attained by a binocular camera or a 3D scanner. However, these approaches require expensive devices and it only can be attained when the object is not be occluded. Due to the above constraints, it is difficult to be widely used in real scenarios. Recently, deep learning has made great progress in many applications. A lot of methods for the 3D shape reconstruction have been proposed. By taking advantage of neural networks, these methods present 3D shapes as point clouds [1]–[3], voxels [4], [5], parametric model [6], or mesh model [7].

However, the human body reconstruction is not addressed well in the monocural case because of the wide range of the imaging conditions and lack of the depth information. Most previous approaches use a parametric model for human pose estimation, e.g., the 3D SMPL [6] model or MANO [8]. Given a single image of a person, they try to estimate the parameters of the human body model that conform to the human body pose in the image. Although the recent parameter-based methods have achieved significant improvement, they still have two major shortcomings. Firstly, the particular parametric space has great limitations. Because SMPL just models a single human body, it can not present details such as hair, facial expressions, and clothes. Secondly, the mapping from a high-dimensional space to a particular parametric space omits the spatial relationship of the pixels from the input image. The above shortcomings lead to difficulty of training and as a result, reduce the reconstruction accuracy as addressed in Moon et al. [9] and Choi et al. [10]. Therefore the direct regression of SMPL parameters might not an appropriate regression target. Unlike previous work, we adopt a human body mesh model for reconstructing a human shape and pose. Instead, our regression target is a mesh with n vertices. The 3D coordinates define the location of each vertex.
Our goal is to reconstruct properly the complete 3D geometry of the human body from monocular images. In this paper, we propose a novel end-to-end approach to reconstruct the 3D human body mesh model. The proposed framework contains an encoder, a generator, and a discriminator. The feature vector extracted by the encoder from the input image is attached to template mesh vertices. Then, the graph convolutional neural network (Graph CNN) in the generator is the deformation process of the mesh. In the last layer, the Graph CNN reconstructs the complete 3D geometry of the human body without the SMPL parameters. Both the generated 3D model and the ground truth human body model are sent to the discriminator. Finally, the discriminator is responsible for determining whether a 3D mesh model is from the generator or not. Our key contribution is to generate 3D human body models in a generative adversarial manner. Our framework can not only leverage the traditional heuristic criteria but also leverage the adversarial criteria. In our generative adversarial network, the generator is encouraged to output the mesh which is closer to the manifold of the real human bodies, which helps the generator learn the real data distribution and avoid outputting people with the abnormal body shapes. To improve the performance of the discriminator, we introduce graph convolution in the discriminator. This architecture explicitly encodes the mesh structure and utilizes the spatial information of the human mesh. These components in together generate competitive reconstruction results compared with the state-of-the-art methods.

The major contributions of our work are summarized as follows:

- We propose an end-to-end framework based on Graph CNN, which can predict the 3D locations of each mesh vertex directly. This avoids calculating the keypoints loss indirectly and also avoids the need for the SMPL parameters.
- We integrate the graph convolution into the discriminator, which can utilize the graph topology better. The extensive experimental results demonstrate that the proposed approach is effective and necessary for 3D human reconstruction.
- The experimental results show that our approach outperforms the previous state-of-the-art results among human pose and shape estimation on the various publicly available datasets.

II. RELATED WORK

In recent years, the 3D human reconstruction has always been a hot field of artificial intelligence. There are many methods to reconstruct human body, including point cloud [11], simplistic body skeleton [12], [13], posture parameters [14], [15], mesh model [16], [17] etc. These methods have different regression goals. In order to understand the related work, we divide the previous human reconstruction work into the two categories: direct parametric regression and non-parametric regression. Our proposed method belongs to non-parametric regression.

A. DIRECT PARAMETRIC REGRESSION

SMPL model is proposed by Loper et al. [6], which uses posture and shape parameters to represent different human body shapes and pose. After that, the vast majority of the 3D human body estimation works reconstruct the human shape through the regression of the SMPL model parameters. Bogo et al. [14] minimize the loss between the predicted keypoint locations and the ground truth keypoint locations to infer joint keypoint locations and the SMPL model parameters. Tan et al. [18] design an encoder-decoder network. The network uses the corresponding images and ground truth silhouettes pairs instead of the corresponding shape and pose parameter pairs. Madadi et al. [19] take a deep neural network as an encoder and take the SMPL model as a decoder. The denoising autoencoder model can recover a 3D pose from the structured error. Although they used different methods or models, the above work focused on taking SMPL model parameters as the regression parameters. In the case of the SMPL model, a human body model is expressed by the 3D rotations parameters. It is difficult to precisely regress the rotation parameters [20], because of the discontinuities or the periodicity of parameters.

B. NON-PARAMETRIC REGRESSION AND GRAPH CNN

Varol et al. [21] attempt to use two networks to predict the 2D pose and the 2D body part segmentation respectively. They combine two subnetworks to a final network to infer a voxel 3D body. The convolutional layers are widely used in the image processing and have achieved the excellent results. However, the convolution operations can not be directly applied to the mesh vertices. Kipf et al. [22] proposed an efficient variant of convolutional neural networks that can apply directly to graphs. Then, Verma et al. [23] use a novel graph-convolution operator to connect filter weights and graph neighborhoods. Litany et al. [24] design a variational autoencoder with graph convolutional operations, which firstly encodes the latent vector of the partial shape and then reconstructs the complete shape according to the latent vector. Unlike the previous Graph CNN, Wang et al. [25] use a graph that is not fixed. In other words, the connection between each layer is being changed. They use it to better predict the local geometric features of the point clouds. Wang et al. [7] propose an architecture based on the Graph CNN that generates 3D shapes from an initial ellipsoid. However, for the details of some complex objects, the model can not reconstruct well. The object mesh reconstruction is still an unsolved problem.

Recently, the non-parametric regression methods based on graph convolution have also been proposed for the human shape reconstruction. Kolotouros et al. [26] proposed to regress 3D mesh vertices to recover human body shape and posture using Graph CNN. Xie et al. [16] extract the anthropometric measurements from the original image, which regresses these parameters to reconstruct a more accurate human body model. Since these parameters can only be extracted from the standard A-pose, its applications still.
FIGURE 1. Overview of the proposed model. An input image is encoded into a feature vector by the encoder. The generator infers a human body mesh model from the feature vector. The generated mesh model and the real human body model are sent to the discriminator, which aims to tell if these models come from a real human shape and pose.

have limitations. Choi et al. [10] propose Pose2Mesh, which reconstructs 3D human body mesh from 2D human body pose. The network adds vertices through upsampling and recovers a high-resolution mesh model. The Graph CNN allows the explicit encoding of the graph structure, which avoids the dependence of the SMPL parameters.

C. GENERATIVE ADVERSARIAL NETWORK

Goodfellow et al. [27] proposed Generative Adversarial Network including a generator and a discriminator. The generated data tends to the distribution of the real dataset during the confrontation between the generator and the discriminator. Brock et al. [28] build a GAN model based on scaling. The architecture is deeper than the other models but only needs about half the parameters of other models, which can generate a more realistic high-resolution image that the human eye can not distinguish. Shaham et al. [29] first proposed an unconditional generative model training on a single natural image, which uses a pyramid of full convolutional GAN to generate images of arbitrary size and aspect ratio. The model maintains the overall structure and the texture of the original image, and it also has obvious variability.

For the human shape reconstruction, Kanazawa et al. [15] proposed an end-to-end recovery model that also contains a discriminator, which acts as weak supervision. The model minimizes the joint reprojection loss and the adversarial loss to regress the SMPL model parameters. Alldieck et al. [30] use a PatchGAN discriminator to enhance the reality. Without the paired accurate 3D pose, it can even recover details including hair and clothes. We propose a capable generative model for human body estimation. The proposed system can benefit from the topology of mesh representation.

III. APPROACH

To reconstruct a full geometric human body model from the input image, we introduce our non-parametric proposed approach for human body estimation. Firstly, we briefly show the model architecture. Then, we introduce the graph convolution layer, which is a vital part in the generator and the discriminator. Finally, loss function is described in detail.

A. MODEL

Fig.1 shows the overview of the proposed framework. The first part of our pipeline is an encoder which extracts features from input RGB images. In our approach, the ResNet-50 convolution network pretrained on the Image-Net classification task [31] is adopted as the encoder. By taking advantage of the transfer learning technology, the 2048-D image-based feature vector is extracted as the input of the generator. Template human mesh attached to extracted features is sent to the Graph CNN. The Graph CNN infers the 3D coordinates of the mesh vertices in the generator. The output mesh model of the generator is our estimation result, which is labeled by fake. Ground truth SMPL parameters are used to recover the real human body mesh, which is labeled by real. The last part of our framework is a discriminator. The real and fake mesh are both sent to the discriminator. The discriminator contains a graph convolutional network and a fully connected layer to estimate the probability that a mesh came from the real human dataset. Since the discriminator encourages the generator to output a smooth surface like the SMPL model, we do not need an additional smooth loss function. The generator is responsible for generating the human shape from the image, and the discriminator is responsible for determining whether the input human body mesh is from the generator or the real dataset. In addition, a Multi-Layer Perceptron (MLP) regresses shape and pose parameters of SMPL from regressed human mesh, so that our network can be applied to applications that require SMPL parameters.

There are two main functions of the discriminator. Firstly, the discriminator avoids outputting human mesh with the abnormal body shape because the discriminator can
distinguish human mesh via body shape. Secondly, for the situation with heavy occlusions in body parts, the discriminator encourages the generator to output reasonable human shape and limb positions.

Kanazawa et al. [15] take the SMPL parameters as the input of discriminator. Instead, the input of ours is a 3D mesh model. This allows the discriminator to use the topology of the graph to capture the complete information of the human body instead of the simple SMPL parameters. We believe that the human mesh is a better representation. The experimental results in part C of section IV also proved it.

B. GRAPH CNN

The 3D coordinates of the mesh vertices are predicted by the Graph CNN. A 3D mesh can be defined as an undirected graph \( G = (\mathcal{V}, \mathcal{E}, \mathcal{X}) \), where \( \mathcal{V} \) is a set of mesh vertices with 3D coordinate, \( \mathcal{E} \) is a set of edges and \( \mathcal{X} \) is the feature vector of each vertex.

The graph convolution layer is proposed by Kipf et al. [22], which is defined as:

\[
Y = \tilde{A} \mathcal{X} W
\]

where, \( \mathcal{A} \in \mathbb{R}^{N \times N} \) is the adjacency matrix of the undirected graph, and \( \tilde{A} \) is the row-normalized matrix of \( \mathcal{A} \). \( \mathcal{X} \in \mathbb{R}^{N \times k} \) denotes the input features matrix and \( W \in \mathbb{R}^{k \times l} \) is the trainable weight matrix. In order to speed up the training of the neural network, we add residual connections and Group Normalization [32] on the graph convolution layer. They are combined into a Residual Graph Convolution Block, which is based on the residual block [31]. The detail of a residual block is shown in Fig. 2. The generator and the discriminator both consist of a serial of residual blocks.

We propose to integrate graph convolution layers in both the generator and the discriminator, which is inspired by Kolotouros et al. [26]. In the generator, the Graph CNN takes mesh vertices and feature vectors as input. The graph convolution layers are equivalent to the deforming of the mesh. It aims to estimate the 3D coordinates for each vertex. In the discriminator, the GraphCNN CNN encourages the generator to output 3D human shapes that are more similar to the real dataset.

C. RECONSTRUCTION AND ADVERSARIAL LOSS

We assume that \( \mathcal{V} \) is the ground truth vertices generated from the SMPL parameters and \( \mathcal{V} \in \mathbb{R}^{N \times 3} \) is the predicted 3D mesh vertices. Because human body mesh model is inferred directly, we can minimize the \( L_1 \) loss between predicted and corresponding ground true mesh vertices, i.e.,

\[
L_{3D} = \sum_i ||\hat{V}_i - V_i||_1.
\]  

The Graph CNN in the generator output 3D mesh vertices directly. Then, we use the MLP to regress the SMPL parameters from inferred mesh model, including shape \( \beta \in \mathbb{R}^{10} \) and \( \theta \in \mathbb{R}^{3K} \). We apply \( L_2 \) loss on this regression task, i.e.,

\[
L_{\text{smpl}} = \sum_i ||\hat{\beta}_i - \beta_i||_2 + \sum_i ||\hat{\theta}_i - \theta_i||_2
\]

where, \( \hat{\beta} \) and \( \hat{\theta} \) are the predicted parameters, \( \beta \) and \( \theta \) are the ground truth shape and pose. In this way, human body model is represented in the parametric space.

Using the predicted 3D mesh vertices, we can obtain the 2D and 3D keypoint location \( \hat{X} \) of the joints. For some datasets, there are paired ground truth keypoint locations. Let \( X \) denote the ground truth keypoints. To align the image with the model, we align the predicted joint locations and the ground truth keypoints. This supervision loss is added to the loss function, which is defined as follows:

\[
L_{\text{joints}} = \sum_i ||\hat{X}_i - X_i||_1.
\]

Let \( L_{\text{recon}} \) denote the reconstruction loss of the generator. It includes 3D mesh vertices loss, SMPL parameters loss, and joints loss.

\[
L_{\text{recon}} = L_{3D} + L_{\text{smpl}} + L_{\text{joints}}
\]

The distance between the real and fake data distribution is measured by the least square formulation [33]. We introduce adversarial loss to improve the reality of the generated human model because our framework is trained on in a generative adversarial manner. The adversarial loss function for the model is:

\[
L_G = \mathbb{E}_{P_G}[D(G(I)) - 1]^2
\]

\[
L_D = \mathbb{E}_{P_D}[D(G(I))^2] + \mathbb{E}_{P_{\text{data}}}[D(V) - 1]^2
\]

\[
L_{\text{adv}} = L_G + L_D.
\]

where, \( I \) is the input image. \( P_G \) and \( P_{\text{data}} \) are the generated human model distribution and real data distribution respectively. The adversarial loss make generator learn the real data distribution.

In the traditional GAN applications, the generator generates clear images from random noise. Training GAN is unstable because it attempts to search for a Nash equilibrium in a high-dimensional space. Besides, the mode collapse is a serious issue during GAN training, as addressed in Bau et al. [34]. These problems make it difficult to train generative adversarial networks. The above problems do not occur in our GAN. Because our GAN also needs to minimize the reconstruction loss while the GAN minimizes the adversarial loss.

Finally, the overall loss function is:

\[
L = L_{\text{res}} + \lambda L_{\text{adv}}
\]

where, \( \lambda \) is a hyper-parameter that assigns a weight on the adversarial loss.
IV. EXPERIMENTS

A. IMPLEMENTATION DETAILS

Datasets: Human3.6M [36] and UP-3D [35] is the training dataset of the proposed approach. Human3.6M is a pose dataset collected indoors, which includes common postures such as walking, exercising, smoking, etc. UP-3D includes 8515 natural images of humans with rich annotations and correct limb rotation, which is selected according to the successful fits. They both provide the SMPL parameters and the 3D joints keypoints. We evaluate our proposed model on Human3.6. Ablation study is run on MPII [37] and LSP [38]. MPII is a 2D pose annotated images dataset, which is provided by Lassner et al. [38]. Each image in LSP has been annotated with the 14 joint locations. We use the test set of this dataset as our validation set. MPII is a human pose dataset, which contains 25K images and 40K people with annotated body joints. We demonstrate the model results on two common protocols (P1 and P2, as defined in [39]).

We alternate training the generator and the discriminator. Optimizing the generator to minimize $L_{\text{recon}}$ and the $L_D$ and then optimizing the discriminator to minimize $L_D$. In our architecture, the generator includes five Residual Graph Convolution Blocks, and the discriminator includes three. We noticed that reducing residual blocks lead to a decrease in reconstruction accuracy. Adding residual blocks will consume the additional resources, while not improving the reconstruction accuracy and the classification accuracy. For our training, the learning rates for two different architecture. The Graph CNN architecture with Multi-Layer Perceptron (MLP) is used as the test set. We compare the discriminator loss of them are used as the training set, and the remaining one is used as the test set. We compare the discriminator loss of Graph CNN architecture with Multi-Layer Perceptron (MLP) with Batch Normalization. For our experiments, we use a batch size of 16 and Adam optimizer. We set appropriate learning rates for two different architecture. The Graph CNN is $3 \times 10^{-4}$ and the MLP is $1 \times 10^{-3}$. The results are shown in Fig.4. Compared with the MLP, ablation results suggest that the graph convolutions significantly increase the training speed of the discriminator. Besides, it can achieve lower loss and stable training.

2) QUALITATIVE EVALUATION

We present the qualitative comparison between our proposed approach and HMR [15] in Fig. 3. We regard HMR as a suitable comparison target because it also trained in a generative adversarial manner. As the figure shows, our proposed approach provides more accurate limb positions than HMR. However, both methods are difficult to reconstruct a precise human body model in complex scenes with occlusions.

C. ABLATION STUDY ON DISCRIMINATOR

Kolotouros et al. [26] demonstrate that graph convolution layers are beneficial compared to fully connected layers in the generator. Our contribution is the insight that the discrimination task of mesh vertices can be simplified when the graph convolution layers are used in discriminator. In this subsection, we focus on explaining the effectiveness of graph convolution on the discriminator. The performance of the discriminator plays a vital role during GAN training. If the discriminator can not distinguish between the real and generated samples, the generator will not be able to learn reliable gradients. A better discriminator allows a generator to obtain a higher quality gradient. We design a series of experiments to prove that the graph convolutions are effective on the discriminator.

We fix the generator parameters, only train the discriminator on MPII dataset. For every ten batches of samples, nine of them are used as the training set, and the remaining one is used as the test set. We compare the discriminator loss of Graph CNN architecture with Multi-Layer Perceptron (MLP) with Batch Normalization. For our experiments, we use a batch size of 16 and Adam optimizer. We set appropriate learning rates for two different architecture. The Graph CNN is $3 \times 10^{-4}$ and the MLP is $1 \times 10^{-3}$. The results are shown in Fig.4. Compared with the MLP, ablation results suggest that the graph convolutions significantly increase the training speed of the discriminator. Besides, it can achieve lower loss and stable training.

### Table 1. The MPIPE (mm) and Reconstruction loss (mm) on Human3.6M, Protocol 2 and Protocol 2. All methods are trained on Human3.6M.

| Method                  | MPIPE P1 | MPIPE P2 | Reconst. Error |
|-------------------------|----------|----------|----------------|
| HMR [15]                | -        | -        | 77.6           |
| Kolotouros et al. [26]  | 105.2    | 102.1    | 69.0           |
| Ours                    | 101.6    | 97.2     | 61.3           |

### Table 2. The Reconstruction loss (mm) comparison between the previous methods and our method on Human3.6M. Protocol 2. All methods are trained on the different combinations of datasets.

| Method                  | Reconst. Error |
|-------------------------|----------------|
| Lassner et al. [35]     | 93.9           |
| Rogenz et al. [6]       | 87.3           |
| SMPLify [14]           | 82.3           |
| IMA [13]               | 36.8           |
| Kolotouros et al. [26]  | 50.1           |
| Ours                    | 48.4           |

We evaluate our method both qualitatively and quantitatively on a variety of datasets.

1) QUANTITATIVE EVALUATION

To quantify the accuracy of our generated human body mesh, we use two metrics: (i) the mean per joint position error (MPIPE), and (ii) Reconstruction error, as defined in [42]. First, we compare our method to the two well-performing baseline models, namely the HMR method of [15] and the recent approach of Kolotouros et al. [26]. They also reconstruct body shape from a single image. For a fair comparison, all methods are only trained on Human3.6M. The detailed results on Human3.6 Protocol 1 and Protocol 2 are shown in Table 1. We observe that the proposed approach yields performance improvements compared to the baselines. Second, we evaluate all methods on Human3.6M Protocol 2 and report the errors in Table 2. In this experiment, our approach is trained on Human3.6 and UP-3D and other methods are trained on the different combinations of datasets. Here our method also outperforms the other baselines.

We fix the generator parameters, only train the discriminator on MPII dataset. For every ten batches of samples, nine of them are used as the training set, and the remaining one is used as the test set. We compare the discriminator loss of Graph CNN architecture with Multi-Layer Perceptron (MLP) with Batch Normalization. For our experiments, we use a batch size of 16 and Adam optimizer. We set appropriate learning rates for two different architecture. The Graph CNN is $3 \times 10^{-4}$ and the MLP is $1 \times 10^{-3}$. The results are shown in Fig.4. Compared with the MLP, ablation results suggest that the graph convolutions significantly increase the training speed of the discriminator. Besides, it can achieve lower loss and stable training.
To intuitively explain the effectiveness of graph convolution in human body reconstruction, we use an MLP instead of the Graph CNN in the discriminator to test its performance (Mesh + MLP). Another comparative experiment using an MLP to regress SMPL parameters from the predicted 3D mesh, and send the predicted SMPL parameters and ground truth SMPL parameters to the discriminator with the MLP architecture (SMPL + MLP). In Table 3, we show the classification accuracy of three architectures after training. The reconstruction results are shown in Table 4. The architecture using MLP is basically the same as Kolotouros et al. [26] on reconstruction accuracy. Compared with our original architecture, the other architecture is clearly underperforming. The results indicate that it is useful to integrate graph convolutions into the discriminator. The experimental results demonstrate the Graph CNN is beneficial from the mesh structure.

### Table 3. Classification accuracy of the different architecture. F1 metric denotes F1-score.

| Architecture | Classification accuracy(%) | F1 |
|--------------|----------------------------|----|
| HMR [18]     | 98.5                       | 0.985 |
| SMPL + MLP   | 99.4                       | 0.993 |
| Mesh + MLP   | 98.8                       | 0.987 |
| Mesh + Graph CNN | 99.8                     | 0.998 |

### Table 4. The MPJPE (mm) and Reconstruction loss (mm) comparison with or without mesh or GraphCNN on Human3.6M, Protocol 1.

| Method               | MPJPE  | Reconstruction loss |
|----------------------|--------|---------------------|
| Kolotouros et al. [26]| 74.7   | 51.9                |
| SMPL + MLP           | 76.3   | 55.4                |
| Mesh + MLP           | 77.8   | 54.2                |
| Ours (Mesh + Graph CNN) | 70.3   | 48.6                |

### V. CONCLUSION

In this paper, we introduce an end-to-end generative adversarial framework for human pose and shape estimation from a monocular RGB image. Compared with the directly regressing model parameters, using the graph convolution can utilize the topology of the mesh. The extensive experiments demonstrate that our approach outperforms the relevant baselines.

However, there are still several constraints to our method. Recovering 3D mesh models miss the details such as clothes and hair facial expressions. Theoretically, these details can be reconstructed in the non-parametric mesh model. Another problem is that if there are too many vertices in the mesh, the requirement of computing power is unbearable, which will prevent us from recovering a higher resolution model. Future work can focus on the above problems.

### REFERENCES

[1] H. Fan, H. Su, and L. Guibas, “A point set generation network for 3D object reconstruction from a single image,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 605–613.

[2] R. Q. Charles, H. Su, M. Kaichun, and L. J. Guibas, “PointNet: Deep learning on point sets for 3D classification and segmentation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 652–660.

[3] Y. Li, R. Bu, M. Sun, W. Wu, X. Di, and B. Chen, “PointCNN: Convolution on X-transformed points,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 820–830.

[4] J. Wu, C. Zhang, T. Xue, B. Freeman, and J. Tenenbaum, “Learning a probabilistic latent space of object shapes via 3D generative-adversarial modeling,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 82–90.
[5] X. Yan, J. Yang, E. Yumer, Y. Guo, and H. Lee, “Perspective transformer nets: Learning single-view 3D object reconstruction without 3D supervision,” in *Adv. Neural Info. Process. Syst.*, 2016, pp. 1696–1704.

[6] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, “SMPL: A skinned multi-person linear model,” *ACM Trans. Graph.*, vol. 34, no. 6, pp. 1–16, Nov. 2015.

[7] N. Wang, Y. Zhang, Z. Li, Y. Fu, W. Liu, and Y.-G. Jiang, “Pixel2Mesh: Generating 3D mesh models from single RGB images,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 52–67.

[8] G. Roegz and C. Schmid, “Mocap-guided data augmentation for 3D pose estimation in the wild,” in *Adv. Neural Info. Process. Syst.*, 2016, pp. 3108–3116.

[9] J. Y. Chang, G. Moon, 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5079–5088.

[10] H. Choi, G. Moon, and K. M. Lee, “Pose2Mesh: Graph convolutional network for 3D human pose and mesh recovery from a 2D human pose,” 2020, arXiv:2008.09047. [Online]. Available: http://arxiv.org/abs/2008.09047

[11] K.-C. Chan, C.-K. Koh, and C. S. George Lee, “A 3-D-point-cloud system for human-pose estimation,” *IEEE Trans. Man, Cybern. Syst.*, vol. 44, no. 11, pp. 1486–1497, Nov. 2014.

[12] D. C. Luvison, D. Picard, and H. Tabia, “2D/3D pose estimation and action recognition using multitask deep learning,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5137–5146.

[13] D. Pavilo, C. Feichtenhofer, D. Grangier, and M. Auli, “3D human pose estimation in video with temporal convolutions and semi-supervised training,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7753–7762.

[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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 7122–7131.

[15] 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. (CVPR)*, Jun. 2018, pp. 561–578.

[16] H. Xie, Y. Zhong, Z. Yu, and A. Hussain, “Non-parametric anthropometric graph convolutional network for virtual mannequin reconstruction,” *IEEE Access*, vol. 8, pp. 3539–3550, 2020.

[17] G. Moon and K. Mu Lee, “2L-MeshNet: Image-to-lixel prediction network for accurate 3D human pose and mesh estimation from a single RGB image,” 2020, arXiv:2008.03713. [Online]. Available: http://arxiv.org/abs/2008.03713

[18] V. Tan, I. Budvytis, and R. Cipolla, “Indirect deep structured learning for 3D human shape and pose prediction,” in *Proc. Brit. Mach. Vis. Conf.*, 2017, pp. 1–11.

[19] M. Madadi, H. Bertiche, and S. Escalera, “SMPLR: Deep SMPL reverse for 3D human pose and shape recovery,” 2018, arXiv:1812.10766. [Online]. Available: http://arxiv.org/abs/1812.10766

[20] J. Martinez, R. Hossain, J. Romero, and I. J. Little, “A simple yet effective baseline for 3D human pose estimation,” in *Proc. IEEE Int. Conf. Comput. Vis.*, (ICCV), Oct. 2017, pp. 2640–2649.

[21] G. Varol, D. Ceylan, B. Russell, J. Yang, E. Yumer, I. Laptev, and C. Schmid, “Bodynet: Volumetric inference of 3D human body shapes,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 20–36.

[22] T. N. Kipf and M. Welling, “Semi-supervised classification with graph convolutional networks,” in *Proc. Int. Conf. Learn. Represent.*, 2016, pp. 1–14.

[23] N. Verma, E. Boyer, and J. Verbeek, “FeaStNet: Feature-steered graph convolutions for 3D shape analysis,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2598–2606.

[24] O. Litany, A. Bronstein, M. Bronstein, and A. Makadia, “Deformable shape completion with graph convolutional autoencoder,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1886–1895.

[25] 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 Trans. Graph.*, vol. 38, no. 5, pp. 1–12, Nov. 2019.

[26] N. Kolotouros, G. Pavlakos, and K. Daniilidis, “Convolutional mesh regression for single-image human shape reconstruction,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4501–4510.