Parametric Model Estimation for 3D Clothed Humans from Point Clouds

Kangkan Wang* [1,2] Huayu Zheng [1] Guofeng Zhang [3] Jian Yang [1,2]

[1] Key Lab of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education,
[2] Jiangsu Key Lab of Image and Video Understanding for Social Security,
School of Computer Science and Engineering, Nanjing University of Science and Technology, China
[3] State Key Laboratory of CAD&CG, Zhejiang University, China

ABSTRACT

This paper presents a novel framework to estimate parametric models for 3D clothed humans from partial point clouds. It is a challenging problem due to factors such as arbitrary human shape and pose, large variations in clothing details, and significant missing data. Existing methods mainly focus on estimating the parametric model of undressed bodies or reconstructing the non-parametric 3D shapes from point clouds. In this paper, we propose a hierarchical regression framework to learn the parametric model of detailed human shapes from partial point clouds of a single depth frame. Benefiting from the favorable ability of deep neural networks to model nonlinearity, the proposed framework cascades several successive regression networks to estimate the parameters of detailed 3D human body models in a coarse-to-fine manner. Specifically, the first global regression network extracts global deep features of point clouds to obtain an initial estimation of the undressed human model. Based on the initial estimation, the local regression network then refines the undressed human model by using the local features of neighborhood points of human joints. Finally, the clothing details are inferred as an additive displacement on the refined undressed model using the vertex-level regression network. The experimental results demonstrate that the proposed hierarchical regression approach can accurately predict detailed human shapes from partial point clouds and outperform prior works in the recovery accuracy of 3D human models.

Index Terms: Detailed human shape estimation, parametric model, point clouds.

1 INTRODUCTION

In recent years, 3D human shape estimation has received increasing attention in various real-world applications such as virtual reality, augmented reality, modern game, and movie production. The goal of 3D human shape estimation is to accurately recover a full detailed 3D mesh of a human body [7, 8, 13, 22]. Due to large variations of human poses, body shapes, and clothing details, 3D human shape estimation is a challenging task. In addition, self-occlusions of human bodies and camera viewpoint changes also make this problem more difficult. How to estimate detailed 3D human body models accurately and robustly is still a core problem in the 3D computer vision community.

As depth cameras are increasingly accessible in recent years, the capturing of point clouds becomes easier. Various 3D human shape recovery methods [11, 12, 25, 46, 47] have been proposed from point clouds. These traditional methods mainly focus on reconstructing the 3D human models from a sequence of depth frames. Based on deep learning techniques, some works [16, 31, 42] attempt to predict the parameters of a statistical body shape model (such as SMPL [23]) directly from partial point clouds of a single frame. Although they achieve good performance on recovering undressed human shapes, the reconstructed shapes do not contain surface details such as clothes. In practices, it is desirable to recover detailed human shapes beyond the undressed human model. The recent work [10] can recover the non-parametric shapes of clothed bodies from point clouds based on implicit functions [36]. Although the non-parametric representations can recover arbitrary shapes, it may generate unreasonable results like broken arms. Moreover, the lack of semantic information in non-parametric representations [10, 36, 37, 49, 51] limits their applications like body reposing and garment retargeting [6].

Inspired by the parametric models of clothed humans [1, 6, 24], we adopt an additional 3D offset on top of the SMPL model [23] to parameterize the detailed human shapes. Based on the favorable ability of deep neural networks to model the nonlinearity, we aim to estimate detailed parametric human shapes directly from point clouds with deep neural networks. However, there are mainly two challenges to regress detailed parametric models from point clouds. First, it is difficult to accurately estimate detailed 3D body models in a single pass due to complex deformation of human shapes and free-form clothing details. Second, unlike color images that contain sufficient texture information of human bodies, point clouds only provide point positions over the body surface but no color image evidence. Thus, discriminative features need to be exploited to effectively predict the human shapes and cloth geometry.

In this paper, we propose a hierarchical regression approach to address the above problems in predicting parametric models for clothed humans from partial point clouds. In the proposed approach, three regression networks are cascaded to regress 3D model parameters from coarse to fine by exploiting different discriminative features. Global deep features of point clouds are first extracted over the whole point clouds to predict a rough undressed model. Then, local features of joint neighborhoods which characterize the shape and pose of human bodies are utilized to refine the undressed model. Finally, vertex-level features that capture the detailed geometry of the input points are employed to infer 3D displacements of cloth details. The discriminative features of point clouds are carefully extracted to achieve the specific tasks in different networks. The joint neighborhoods on the input points are detected with the initial undressed model, while the detailed geometry can only be captured based on the refined undressed model. The three levels of features function successively, leading to the success of recovering the accurate detailed model. The experimental results on various datasets of dressed humans demonstrate the effectiveness of our proposed hierarchical regression networks.

The main contributions of our method are as follows:

- To the best of our knowledge, this is the first deep learning based work that predicts the parametric models for 3D clothed humans from partial point clouds.
- We propose a novel hierarchical regression framework to estimate detailed 3D human shapes from point clouds in a coarse-

*Corresponding author: wangkangkan@njust.edu.cn

to-fine strategy. Our method can achieve high reconstruction accuracy and outperform the state-of-the-arts.

- We carefully design a scheme of exploiting different discriminative features from point clouds to progressively refine the human shapes.
- Our method is generalized successfully to recover the parametric models of clothed humans on real data with a self-supervised fine-tuning method.

2 Related Work

2.1 3D human shape estimation from color images

Traditional 3D human shape estimation methods attempt to recover 3D model parameters by fitting a parametric body model like SCAPE [3] or SMPL [23] to a set of observations from input color images [13]. As deep learning is successfully applied in many computer vision tasks, some researchers try to directly learn the 3D human pose and shape from a single color image through convolutional networks [8, 18, 21, 22, 28]. These methods can yield good performance on recovering undressed human shapes but fail to estimate clothing details.

In recent years, 3D human shapes with surface details are reconstructed from color images through implicit function [36, 37], hierarchical mesh deformation [51], normal map refinement [49], or UV mapping [2]. Although these methods can recover the detailed human shapes effectively, but the recovered details are not parameterized which limits the applications of shape editing such as garment transfer. For example, Zhu et al. [51] recovers detailed human shapes from a single color image by estimating the mesh deformation with a hierarchical architecture, resulting in nonparametric models. In this work, our goal is to regress the parameters of detailed 3D body models. Some works try to parameterize the human details through an extra offset layer from the body [1,6,24] or by modeling the clothing explicitly [15, 27]. For example, CAPE [24] models the clothing as a displacement layer on the SMPL mesh parameterized by the shape and structure of clothing, body pose and clothing type. Recent works handle the detail recovery of human bodies from videos with photo-consistency among temporal frames in self-supervised learning [40] or through weak supervision [14] under multi-view 2D joint detections and silhouettes.

2.2 3D human shape estimation from depth images

3D human shape can be reconstructed from depth sequences using template-based methods [7, 9, 11, 29, 48], templateless methods [12, 25], or the combination of the two [46, 47]. These methods mainly employ successive information of depth sequences to build dense point correspondences for each frame using the nearest neighboring point searching method. In the case of the 3D human body model estimation from a single depth image, due to large discrepancy between the 3D template model and depth image, it is difficult to build accurate point correspondences with the nearest neighbor searching method. Thus, these methods cannot deal with the recovery of 3D human body model from a single depth image well. Point correspondences can be predicted for a single depth image through metric regression forest [30] or by matching learned descriptors [44]. 3D body models are then recovered by deforming a template model to partial scans with the predicted correspondences.

Recent works attempt to infer the 3D body models from point clouds of a single frame with deep learning methods. Extracting efficient features on point clouds to regress human shapes is much more challenging than on color images. Jiang et al. [16] regresses SMPL parameters from point clouds by extracting skeleton joint features from unordered point features. Prokudin et al. [31] encodes a point cloud as the minimal distances to a fixed basis point set and directly predicts the mesh from the distance vector. Both methods [16, 31] can only recover the undressed SMPL model but cannot handle clothed humans. Wang et al. [42] can predict the sequential 3D body meshes from a sequence of point clouds with a spatial-temporal mesh attention convolution network, which also fails to recover the clothing details. Some recent works [4, 10] can estimate the detailed human shapes, but the shapes are non-parametric or a post-processing step is required to recover the parametric model. Chibane et al. [10] learns 3D point occupancy from the encoded multi-scale features and reconstructs the continuous surface with implicit function-based methods. Bhatnagar et al. [4] first predicts semantic correspondences to a parametric body model and then recovers the parametric models using the optimization-based registration. Bhatnagar et al. [5] proposes a self-supervised learning method to estimate correspondences and parametric models jointly, which is proposed for complete point clouds. TexMesh [50] reconstructs detailed meshes with texture from RGB-D video and self-adapts to real sequences through photometric constraints.

2.3 3D deep learning

Recent research efforts on feature representation of 3D data have been devoted to 3D deep learning. Voxel-based methods [26, 36] transform 3D data into 3D voxels and apply 3D convolution neural network (CNN) over voxels. Multiview-based methods [33, 39] project 3D data into a set of 2D images rendered from different views and use 2D CNN to handle them. Some recent works attempt to extract features directly on the point set [32, 34, 45]. The basic PointNet [32] maps each input point into a feature vector via shared multi-layer perceptron networks and then extracts a global feature over the entire point set using a max pooling operator. Based on the basic PointNet, PointNet++ [34] can capture local structures of the point set by extracting local features in a hierarchical way. Recently, novel features are learned on point clouds to predict the human body shape by computing minimal distances to a basis point set [31] and a 3D grid of multi-scale features [10].

3 Proposed Approach

Our proposed method for detailed human shape estimation takes partial point clouds of a dressed human body as the input and outputs a parametric 3D body model that is consistent with the point clouds in human pose and shape and cloth details. Fig. 1 illustrates the proposed hierarchical regression framework. The framework consists of three successive regression networks and predicts the detailed human body shapes in a coarse-to-fine strategy. The first regression network, i.e., global regression network, takes the global deep features of point clouds as the input to globally estimate a rough unclothed 3D body model. Then, the subsequent regression network, i.e., local regression network, utilizes local deep features of neighborhoods of human joints to refine the unclothed 3D body model. The final regression network, i.e., vertex-level regression network, infers the geometry details of clothes by extracting vertex-level deep features from the neighboring points around the unclothed mesh vertices. The discriminative features in the current stage are produced based on the predicted shapes from the previous stage. In this section, we first briefly introduce the detailed 3D human body model adopted in our paper, and then present our hierarchical regression framework in detail.

3.1 Parametric 3D model for clothed human bodies

We decompose the clothed human body into the undressed body and the clothing detail similar to previous works [1,6,24]. The undressed body is represented using the Skinned Multi-Person Linear model (SMPL) [23]. The SMPL model is parameterized by a set of shape parameters $\boldsymbol{\beta} \in \mathbb{R}^{10}$, and pose parameters $\boldsymbol{\theta} \in \mathbb{R}^{6T}$ ($T = 24$ including 23 joint rotations and 1 global rotation). We employ 3D rotations of a 6-dim vector as [20] and use a translation $\boldsymbol{t} \in \mathbb{R}^3$ to represent global translation of the body. Thus, the parameters of the undressed body mesh can be represented as a 157 dimensional vector $\boldsymbol{\Theta} = (\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{t})$. The output of the SMPL model is a triangulated body mesh $M(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{t}) \in \mathbb{R}^{3\times N}$ with $N = 6890$ vertices. For the mathematical details of SMPL model, please refer to [23].

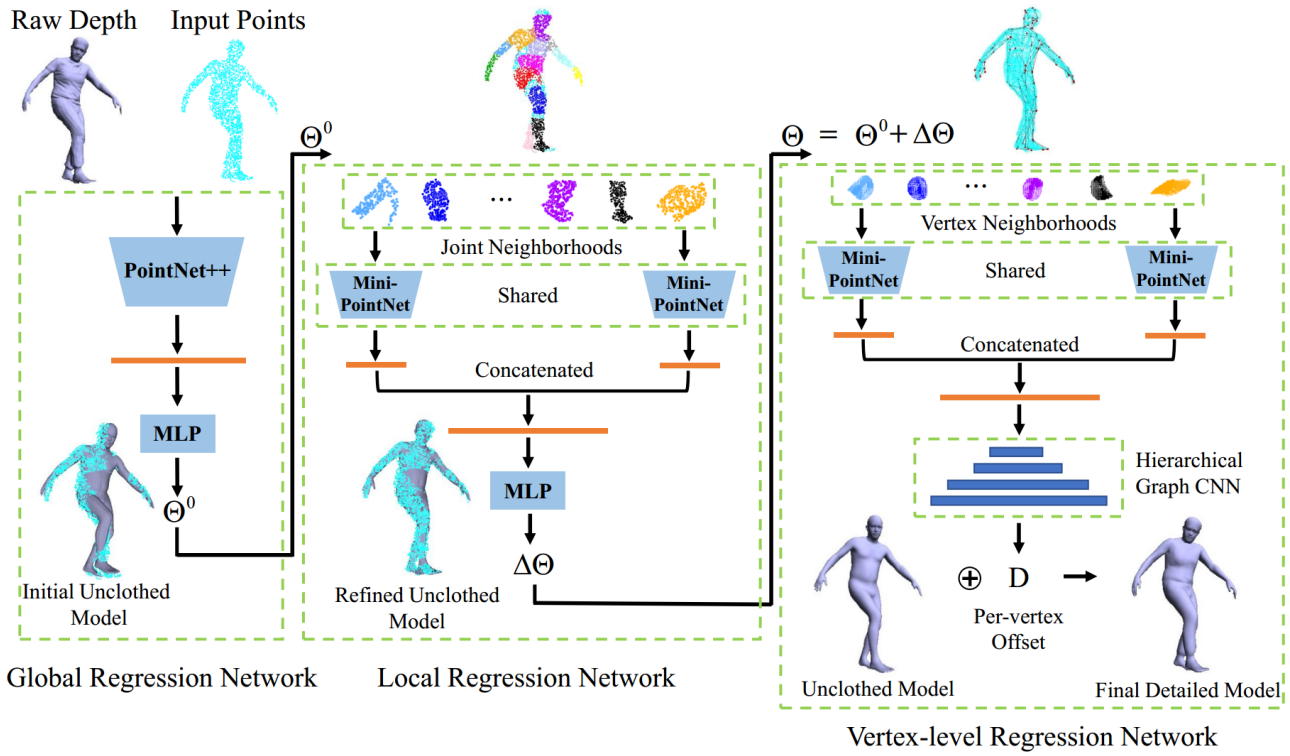


Figure 1: Overview of the proposed hierarchical regression framework. In each stage, by taking an initial parameter and the point clouds as the input, we first extract global/local/vertex-level features on the point clouds and then estimate an update to the model parameter through a regression network. The new parameter is taken as the input to the next stage. Please refer to Sect. 3 for detailed description.

We represent the clothing details [1,6,24] using an extra offset $\boldsymbol{D}$ and add it on top of the SMPL model. Then, we obtain the detailed human shape as follows:

$$M(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{D},\boldsymbol{t}) = W(T(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{D}),J(\boldsymbol{\beta}),\boldsymbol{\theta},\boldsymbol{W})+\boldsymbol{t} \quad (1)$$

$$T(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{D}) = \boldsymbol{T}+B_s(\boldsymbol{\beta})+B_p(\boldsymbol{\theta})+\boldsymbol{D} \quad (2)$$

where linear blend-skinning $W(\cdot)$ with blend weights $\boldsymbol{W}$ deforms the clothed template $T(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{D})$ in the unposed space to the posed space, $M(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{D},\boldsymbol{t})$ is the detailed shape in the posed space after blend-skinning and translation, $J(\boldsymbol{\beta})$ denotes the linear regression to compute 3D joint locations from shape parameters, $B_s(\boldsymbol{\beta})$ and $B_p(\boldsymbol{\theta})$ are shape-dependent deformations and pose dependent deformations, respectively.

3.2 Hierarchical regression networks

3.2.1 Global regression network

The first regression network in our hierarchical framework, i.e., global regression network, approximately estimates the undressed 3D model based on global deep features of point clouds. We adopt PointNet++ [34] to capture 3D geometry information and extract global deep features from point clouds. By taking the global features as the input, we employ fully-connected networks to learn the parameter $\boldsymbol{\Theta}$ of the undressed mesh with the following objective function:

$$L = \lambda_1 L_{3D} + \lambda_2 L_{para} + \omega L_{adv}, \quad (3)$$

where L_{3D} is the 3D correspondence loss, L_{para} is the parameter loss, L_{adv} is the adversarial loss, λ_1, λ_2, and ω are the regularization parameters. The 3D correspondence loss L_{3D} forces the vertices of the predicted undressed mesh to align to the corresponding points on the ground truth undressed mesh:

$$L_{3D} = \sum_i \|M(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{t})_i - \hat{\boldsymbol{v}}_i\|_2^2, \quad (4)$$

where $M(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{t})_i$ and $\hat{\boldsymbol{v}}_i$ are the i-th vertex of the predicted and the ground truth undressed mesh, respectively. The parameter loss L_{para} encourages the estimated parameters $\boldsymbol{\Theta}$ to be close to the ground truth parameters:

$$L_{para} = \|\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\|_2^2 + \|R(\boldsymbol{\theta})-R(\hat{\boldsymbol{\theta}})\|_2^2 + \|\boldsymbol{t}-\hat{\boldsymbol{t}}\|_2^2, \quad (5)$$

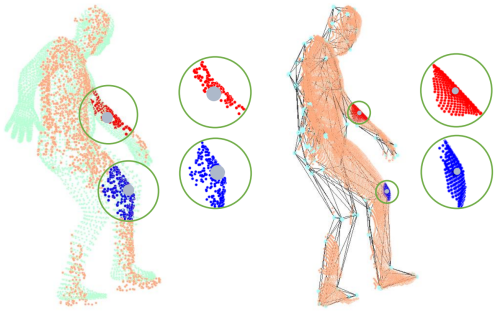
where $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\theta}}$, and $\hat{\boldsymbol{t}}$ are the ground truth parameters. $R(\cdot)$ transforms the pose parameters to vectorized rotation matrices $R(\boldsymbol{\theta}) \in \mathbb{R}^{24\times 9}$. L_{adv} is a least-square adversarial loss [17] used in the GAN to discriminate whether the parameter $\boldsymbol{\Theta}$ is from a real human shape. The adversarial loss can prevent the network from generating 3D models with implausible poses and shapes. With Eq. (3), we can obtain an initial parameter $\boldsymbol{\Theta}^0$ of the unclothed 3D body model from the global features. The predicted 3D body model is a rough but robust approximation of the ground truth.

3.2.2 Local regression network

Due to large discrepancy between the point clouds and the template model, the initial estimation $\boldsymbol{\Theta}^0$ from the global regression network is not close enough to the ground truth parameter $\hat{\boldsymbol{\Theta}}$. In order to obtain the more accurate parameter, the local regression network is proposed to predict the increments $\Delta\boldsymbol{\Theta}$ based on the initial parameter $\boldsymbol{\Theta}^0$ with the local deep features of human joints of the input points.

In the global regression network, the global features of point clouds with PointNet++ ignore local geometric structures of the human body. Thus, the local structures of the 3D body model cannot be fully recovered from the global features, resulting in large reconstruction error. It is expected that the reconstructed 3D model can fit point clouds of the human body locally. Since the joints can reflect the body pose and shape well, we concatenate deep features of neighborhoods of joints as the inputs to the regression network to refine the 3D model. Based on the initial 3D model $M(\boldsymbol{\Theta}^0)$, we can obtain the joint locations of the human body. For each joint

$\boldsymbol{p}_j (j = 1, ..., J)$, we search the neighborhood $\boldsymbol{X}_j$ by finding the neighboring points from the input points within a radius r around the joint. An illustration of the joint neighborhoods is shown in Fig. 2. Then, we extract local features of the neighborhood $\boldsymbol{X}_j$ with a shared mini-PointNet for all neighborhoods. The extracted local features can be denoted by $\boldsymbol{L}(\boldsymbol{X}_j)$, where $\boldsymbol{L}$ is a mapping function. The local features of J joints are then concatenated to form a single feature vector $\boldsymbol{\Phi} = [\boldsymbol{L}(\boldsymbol{X}_1); \boldsymbol{L}(\boldsymbol{X}_2); \cdots; \boldsymbol{L}(\boldsymbol{X}_J)]$ in the order of joint index.



(a) (b)

Figure 2: (a) An illustration of joint neighborhoods detected with the initial SMPL model. (b) An illustration of neighboring patches of mesh vertices extracted based on the refined SMPL model.

Similar to the global regression network, we also use fully-connected networks to regress the update $\Delta\boldsymbol{\Theta}$ from the local feature vector $\boldsymbol{\Phi}$. We use the same objective function as in Eq. (3) to obtain $\Delta\boldsymbol{\Theta}$. Once $\boldsymbol{\Theta}$ is updated, with $\boldsymbol{\Theta} = \boldsymbol{\Theta}^0 + \Delta\boldsymbol{\Theta}$, we can obtain the refined unclothed body mesh $M(\boldsymbol{\Theta})$. It is noted that the hierarchical feature extracting in PointNet++ [34] is different from that in our local regression network. PointNet++ ignores much information around the joints since it uses the iterative farthest point sampling (FPS) and set abstraction to extract local features. Thus, PointNet++ cannot characterize the geometric structure of the human body well. On the contrary, our method localizes the joints with the recovered 3D body model in the previous stage. Therefore, our local regression network can yield more accurate 3D pose and shape estimation by effectively exploiting local features of human joints.

3.2.3 Vertex-level regression network

After the local regression network, we obtain the accurate unclothed body model. Although human poses and shapes are well reconstructed, clothing details are not provided in the unclothed SMPL model. In the vertex-level regression network, we aim to recover the clothing details by inferring per-vertex displacements $\boldsymbol{D}$ added on the undressed model. Then, the estimated 3D body model can fit the input point clouds accurately in the regions of clothes.

We exploit vertex-level features by capturing detailed geometry information on the raw point clouds. Based on the undressed model, the neighboring patch $\boldsymbol{P}_i$ of vertex i can be accurately extracted on the point clouds by searching the neighboring points within a radius. The points in the patch $\boldsymbol{P}_i$ are then encoded with a mini-PointNet to generate vertex-level features $\boldsymbol{V}(\boldsymbol{P}_i)$, where $\boldsymbol{V}$ is a mapping function. The vertex-level features of N vertices are concatenated to form a feature vector $\boldsymbol{\Psi} = [\boldsymbol{V}(\boldsymbol{P}_1); \boldsymbol{V}(\boldsymbol{P}_2); \cdots; \boldsymbol{V}(\boldsymbol{P}_N)]$ in the order of vertex index.

From vertex-level features $\boldsymbol{\Psi}$, we adopt a hierarchical graph CNN to infer detail offsets $\boldsymbol{D}$. In the hierarchical graph CNN, graph CNN [21] and mesh upsampling [35] are applied iteratively to regress the detail displacements of all template vertices. The detail loss L_{detail} is defined as follows:

$$L_{detail} = \sum_i \|M(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{D}, \boldsymbol{t})_i - \hat{\boldsymbol{p}}_i\|_1, \tag{6}$$

where $\hat{\boldsymbol{p}}_i$ is the coordinate of the i-th vertex on the groundtruth detailed models. We use $L1$ loss [21] because it leads to better performance than $L2$ loss. The vertex-level features encode correspondence information in the neighboring points relative to each vertex of the undressed models. By minimizing the detail loss L_{detail}, discriminative correspondences can be extracted in the vertex-level features and the vertex offsets are then estimated effectively.

3.3 Self-supervised fine-tuning on real data

We pre-trained the hierarchical regression networks on a synthetic dataset with supervision. However, the pre-trained model does not perform well on real data (e.g. captured by Kinect) due to the domain gap. As the raw point clouds contain shape information and cloth details of the body, we can use them to fine-tune the networks on unseen real data in a self-supervised manner. By testing with the pre-trained model, an initial mesh $\mathscr{V}$ is generated that is consistent with the input data roughly. Then, the correspondences are built between the input points and vertices of $\mathscr{V}$, and these corresponding points serve as pseudo-ground truth correspondences to fine-tune our networks. We adopt the closest point method in [43] to find the correspondences iteratively that are conditioned on point distance and normal compability. We define the self-supervised 3D correspondence loss of regressing SMPL model as follows:

$$L'_{3D} = \sum_i \|(\boldsymbol{x}_i - M(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{t})_{\boldsymbol{\varepsilon}_i}) \cdot \boldsymbol{\sigma}_i\|_2^2, \tag{7}$$

where $\boldsymbol{\varepsilon}_i$ is the correspondence index of point $\boldsymbol{x}_i$ in the model vertices, $\boldsymbol{\sigma}_i \in \{0, 1\}$ is a mask (if the correspondence is correct, $\boldsymbol{\sigma}_i = 1$; otherwise, $\boldsymbol{\sigma}_i = 0$). The mask $\boldsymbol{\sigma}_i$ is obtained for each correspondence based on whether the conditions of point distance and normal compability are satisfied or not as in [43]. This loss measures the correspondence distance which is similar to the Chamfer distance in the unsupervised loss of [16]. As demonstrated in the experiment, the correspondence finding method of [43] is more effective than Chamfer distance under partial point clouds. Similarly, we define the self-supervised detail loss as follows:

$$L'_{detail} = \sum_i \|(\boldsymbol{x}_i - M(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{D}, \boldsymbol{t})_{\boldsymbol{\varepsilon}_i}) \cdot \boldsymbol{\sigma}_i\|_1 + L_{Laplacian}. \tag{8}$$

We introduce a Laplacian loss [38] to prevent unnatural details:

$$L_{Laplacian} = \sum_i \| (\boldsymbol{\delta}_i - \hat{\boldsymbol{\delta}}_i) \cdot u_i \|_2^2, \tag{9}$$

where u_i is the weight of vertex i, $\boldsymbol{\delta}_i$ and $\hat{\boldsymbol{\delta}}_i$ are the Laplacian coordinate of vertex i on the estimated mesh and $\mathscr{V}$, respectively. We set $u_i = 100$ for vertices of head, hand, and foot, and 1 for other vertices. Then, we form the fine-tuning objectives $L' = L'_{3D} + \omega L_{adv}$ to regress the undressed model and L'_{detail} to infer the clothing details, respectively.

4 Experiments

In this section, we first illustrate our implementation details and perform ablative analysis on our method. Then, our method is evaluated by comparing to the state-of-the-art methods. Finally, we show two applications and test our method on the real data.

4.1 Implementation details

4.1.1 Datasets

We conduct experiments on the SURREAL [41], DFAUST [9], BUFF [48], CAPE [24], and real data. These datasets contain large pose and shape variations and a wide variety of clothing details. The 3D models are rendered to depth images from 10 different viewpoints. The depths are converted to point clouds for both training and test, and the synthetic point clouds are partial from a single view. We

train the model for male and female separately. In the training data of SURREAL, we uniformly sample $10,000$ models and render $100,000$ depth images, which mainly cover the space of human pose and shape. From CAPE dataset, we uniformly sample $20,000$ models and create $200,000$ depth images, which contain a wide range of common garments. The training data contains $300,000$ depth images totally for both male and female. The ground truth undressed SMPL models and detail offsets are provided by the raw training dataset. The test data does not include the same subjects in the training data. From each dataset, we generate $10,000$ frames for both male and female as our test data.

4.1.2 Architecture and experimental settings

In the global and local regression stages, the raw point clouds are uniformly downsampled to 2500 points. The complete point clouds are used in the vertex-level regression network. For local feature extracting, we search neighboring points within a radius $r = 0.3m$ around all the $J = 23$ joints. The parameters of human pose, shape, and global translation are regressed separately with three different fully-connected networks. We adopt one fully-connected layer for each network and use 3 iterations for each fully-connected layer as [17]. The discriminator network setting is the same as that in [17]. For vertex-level feature extracting, we downsample the full mesh to 108 vertices and search neighboring patches within a radius $r = 0.1m$ around each sampled vertex. The sampled vertices scatter evenly over the model and other vertex sampling also works. We use $mlp\{64, 64, 128\}$ for the mini-PointNet to encode both local features and vertex-level features. In the hierarchical Graph CNN, the input feature vector is transformed to a (1034×32)-dim vector with a full-connected layer and reshaped to 1034 vertices with 32-dim feature vector. Four successive graph residual blocks [24] are then applied with upsampling factors of $\{2, 2, 2, 2\}$ and feature channels of $\{32, 32, 16, 8\}$, finally outputting the 3D detail displacements of a template mesh with 16535 vertices. We obtain the high-resolution template by sampling on the original SMPL model.

The learning rate is set to 1×10^{-4}. We use the Adam optimizer [19] with batch size of 4. At the beginning of the training in the global regression stage, to avoid getting stuck in the local minima, we set $\lambda_1 = \lambda_2 = 1$ and $\omega = 1000$ with a large weight for L_{adv}. After this training process converges, we then set $\lambda_1 = \lambda_2 = 60$ and $\omega = 1$ with large weights for L_{3D} and L_{para}. In the local regression stage, we only train once using the weights $\lambda_1 = \lambda_2 = 60$ and $\omega = 1$. We train the three regression networks separately and successively to predict the shape updates. While training a new stage, we keep the networks of previous stages fixed. The total training time of our hierarchical networks is about 87.4h, and the inference time is about 53.4ms tested on a sample. We quantify reconstruction error as Mean Average Vertex Error (MAVE) [42] in millimeter.

4.2 Ablative analysis

The number of sampled points. We first investigate the influence of the number of sampled points on undressed SMPL estimation. The test is performed on the SURREAL dataset. Table 1 lists the reconstruction errors using different number of sampled points. As shown in Table 1, we observe that there is a slight improvement on the performance with an increasing number of points. But, the training process also takes much longer. By balancing between reconstruction accuracy and running efficiency, we choose 2500 points in the following experiments. This experiment verifies that our approach is robust to a small number of sampled points since the accuracy does not decrease a lot with 2500 points.

Reconstruction error in different stages. We also evaluate the effectiveness of our hierarchical regression network by comparing the reconstruction error in different stages. The reconstruction errors at the begin and end of different stages are listed in Table 2. Two reconstructed models in different stages are shown in Fig. 3, including CAPE data and DFAUST data. It can be seen that there is an

Different stages	Point number			
	1000	2500	5000	7500
Global R.	43.7	35.6	35.1	34.9
Local R.	31.9	23.4	23.1	23.0

Table 1: Reconstruction errors (mm) in the global and local regression stage with different number of sampled points.

Different stages	SURREAL		DFAUST		CAPE	
	Begin	End	Begin	End	Begin	End
Global R.	132.9	35.6	136.4	35.8	137.3	35.9
Local R.	35.4	23.4	35.7	23.8	35.6	24.1
Vertex-level R.	23.2	16.4	23.6	16.7	23.7	16.8

Table 2: Reconstruction errors (mm) at the begin and end of each stage on the SURREAL, DFAUST, and CAPE dataset.

improvement of reconstruction accuracy after each stage, and the recovered 3D model is registered more closely to the input point cloud. The results imply that our hierarchical structure is very effective to progressively estimate the detailed human shapes. Particularly, we apply another local regression after the first one, and find that the errors decrease about 3mm on the first one. Since the reduced error is not obvious, we only adopt one local regression.

Our hierarchical regression networks progressively refine the parametric model of 3D clothed humans with different discriminative features from point clouds. Owing to efficient deep features extracted in each stage, the task of each stage can be finished successfully. The global regression network regresses a rough SMPL model with the global features. The local regression network estimates a more accurate SMPL model since local joint features can reflect the human pose and shape well. The vertex-level regression network further recovers the cloth details that cannot be represented by the SMPL model, so the recovered detailed model can fit the input points more accurately.

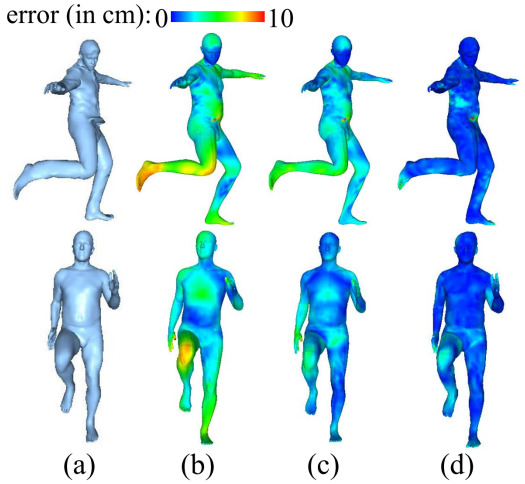


Figure 3: The visualization of reconstruction accuracies in different stages of our method on the CAPE data (top row) and DFAUST data (bottom row). (a) Input depth. (b) The global regression. (c) The local regression. (d) The vertex-level regression.

Detail regression. We experiment with different features to regress the detail displacements: global features, vertex-level features, and cascade of the two. The global features are extracted by PointNet++ [34]. In extracting vertex-level features, we use 108 and 431 vertices obtained by downsampling the full mesh [35]. The cascaded regression first predicts initial details using global features and then updates the details using vertex-level features of 108 vertices. The reconstruction errors using different features are listed in Table 3. We can see that the recovery accuracy using vertex-level features is higher than using global features since detailed geometry

Different methods	DFAUST	BUFF	CAPE
Global features	20.7	19.8	21.1
108 vertices	16.7	16.0	16.8
431 vertices	16.2	15.6	16.4
Cascaded regression	15.9	15.3	16.1
Fully connected layers	21.9	21.6	22.3

Table 3: Reconstruction errors (mm) in the detail regression using global features of PointNet++, vertex-level features of 108 and 431 vertices, and the cascaded regression. We also report the errors by replacing the hierarchical graph CNN with fully connected layers.

information of the point clouds can be exploited more effectively in vertex-level features. The cascaded regression achieves a slight higher accuracy, but it takes much more training time. The accuracy is increased little using more vertices since 108 vertices can capture sufficient detailed information. Thus, we adopt 108 vertex features to regress details in our experiments.

Hierarchical graph CNN. We demonstrate the effectiveness of the hierarchical graph CNN on regressing detail offsets by comparing it with fully connected layers. In the comparison method, we replace the hierarchical graph CNN with three fully connected layers and keep other settings the same as the graph CNN. The three fully connected layers are fulfilled with $mlp\{1024, 1024, 16535 \times 3\}$ which map the vertex-level features to the offsets of 16535 vertices. The reconstruction errors of fully connected layers are listed in the last row of Table 3, which are about 5mm larger than that of our method. The reconstruction examples using two methods are shown in Fig. 4. The comparison results show the hierarchical graph CNN is more effective to regress the pre-vertex offset than the fully connected layers. The hierarchical graph CNN sufficiently exploits the human shape priors contained in the graph and adopts a coarse-to-fine strategy to progressively recover the full-resolution vertex offsets, resulting in higher accuracy of detail recovery.

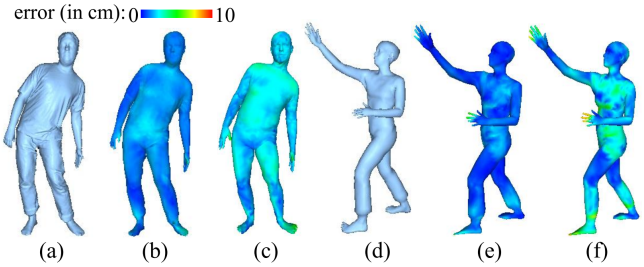


Figure 4: Comparison between two detail regression methods on the BUFF data (left) and CAPE data (right). (a,d) The input depth. (b,e) Hierarchical graph CNN. (c,f) Fully connected layers.

Self-supervised fine-tuning on real data. In this experiment, the effectiveness of the self-supervised fine-tuning is validated on real data. We first compare the prediction results before and after the fine-tuning. Then, we explore the influence of correspondence finding on the prediction results by comparing two methods. One method is Chamfer distance employed in [16], and the other one is the closest point method [43] adopted in our paper. A comparison example is shown in Fig. 5. Before the fine-tuning, there is large discrepancy between the predicted undressed model and the input point clouds. After the fine-tuning, the undressed model becomes more consistent with the point clouds in both human pose and shape, and the recovered detailed model fits to the point clouds accurately due to the success of detail recovery. In addition, both the undressed and detailed models fine-tuned using Chamfer distance cannot fit to the point clouds well. The correspondences implicitly found by minimizing the Chamfer distance are prone to be incorrect for partial point clouds because back-facing surfaces of the SMPL model may be matched to front-facing surfaces of the point clouds (e.g., the right leg in this example). In contrast, we search correct correspondences

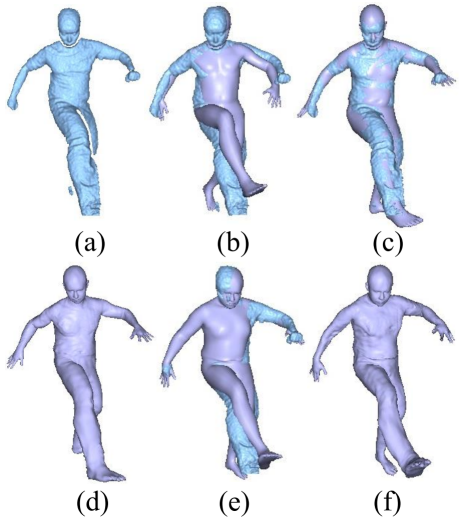


Figure 5: An example of self-supervised fine-tuning on "Kungfu" data from [11]. (a) The input depth. (b,c) The undressed SMPL model before and after our fine-tuning method, respectively. (d) The detailed model after our fine-tuning method. (e,f) The undressed SMPL model and detailed model after fine-tuning with Chamfer distance, respectively. The overlay with alignment is shown between the undressed model and the raw depth.

Methods	DFAUST	BUFF	CAPE
Prokudin et al. [31]	32.2	33.7	31.3
Wang et al. [42]	29.4	30.5	28.3
Bhatnagar et al. [4]	26.0	26.8	25.6
Chibane et al. [10]	22.5	24.5	21.1
Our method	**16.7**	**16.0**	**16.8**

Table 4: Reconstruction errors (mm) using different methods of human shape estimation from point clouds.

using the strategy in [43] based on the fact that correspondences should be spatially close and have similar normal orientation, making our method more effective for partial point clouds.

4.3 Comparison to state-of-the-art methods

We compare our method with state-of-the-art deep learning methods [31], [10], and [42] for human shape estimation from point clouds. We also compare our method to the recent optimization-based registration method [4] of recovering the parametric model for clothed humans. We fulfill these methods using the codes released or provided by the authors. These methods are trained on our dataset. The reconstruction errors with different methods are listed in Table 4, and two reconstructed examples using these methods are shown in Fig. 6. The comparison results demonstrate that our method outperforms state-of-the-art methods in recovering detailed 3D body models from point clouds. The compared methods can recover the 3D shapes with human details, but the recovered shapes are not parameterized for [31], [10], and [42]. The minimal distances of [31] to a fixed basis point set are sensitive to rotations of the body, so the performance degrades dramatically on human bodies with large poses or rotations. The reconstructed shapes using Chibane et al. [10] can preserve body details in the input, but it may produce unreasonable results due to data missing and the error of the occluded parts is very high. Bhatnagar et al. [4] first adopts the implicit function based method [10] to predict a nonparametric shape, so the recovery accuracy is limited to that of [10]. Also, the error in the predicted body part labels for each point can significantly affect the final registration accuracy of [4].

Wang et al. [42] estimates the 3D coordinates of model vertices but not the parametric model (i.e., pose and shape parameters in undressed SMPL model). A post-preprocessing model fitting is

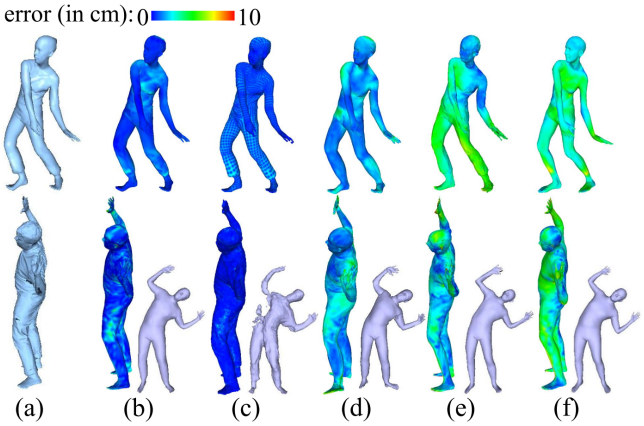


Figure 6: The visualization of reconstruction accuracies using different methods on the CAPE data (top row) and BUFF data (bottom row). (a) Input depth. (b) Our method. (c) Chibane et al. [10]. (d) Bhatnagar et al. [4]. (e) Wang et al. [42]. (f) Prokudin et al. [31]. The recovered models on BUFF data are shown in another view to visualize the reconstruction results of occluded parts.

applied in [42] on the vertex coordinates to obtain the SMPL model. In addition, Wang et al. [42] mainly focuses on predicting the human pose and shape but not the cloth details, which limits its applications like garment retargeting. In contrast, the proposed method in this paper not only directly regresses the parametric model but also recovers the 3D clothes humans, which is much more challenging and also promising. The method of [42] fails to recover geometry details only with global features of PointNet++, while our method exploits different discriminative features on the point clouds to progressively regress the detailed human shapes. The error of [42] on DFAUST in the original paper is much lower because partial DFAUST data are contained in the training dataset which alleviates the problem of domain gap. In this experiment, we train [42] directly on our training dataset without any DFAUST data. So the error of [42] is higher when tested on DFAUST data due to the domain gap. The errors of [42] on BUFF and CAPE are large because the cloth details are not recovered in the 3D models and there is large fitting error between the recovered models and the input points. The comparison results demonstrate our method can achieve higher reconstruction accuracy than the state-of-the-arts.

4.4 Comparisons with model fitting methods

We compare our method to three kinds of model fitting methods using purely searched correspondences [23], detected joints [8], and learned correspondences [44], respectively. These methods are implemented with their released codes. We test all methods on the CAPE dataset. The reconstruction error of our method is 16.8*mm* while those of the compared methods are 147.5*mm*, 60.5*mm* and 62.3*mm* for the pure model fitting, [8] and [44], respectively. An example of the recovered models using different methods is shown in Fig. 7. The model fitting methods can gain good accuracy for point cloud sequences since correspondences are built reliably with temporal information. However, for a single frame or rather fast motion, performance of these methods degrades dramatically as the correspondence searching easily fails. Thus, markers [9], key points [8], or learnt correspondences [44] are used to improve performance for these cases. These model fitting methods can work well for the data with relatively small poses. Nonetheless, the reconstruction error is still large when tested on samples with strong occlusions or large poses. The model fitting methods highly rely on the detection of 2D joints or correspondences with point clouds. In the cases of strong occlusions or large poses, human joints might be detected inaccurately and searched correspondences are prone to be unreliable which cause large reconstruction errors using the model fitting methods. The high errors under strong occlusions or large poses enlarge the average error over all test samples.

In contrast, our deep learning method has the stronger capacity of representing the complicated mapping function, and modern techniques allow the deep networks to be successfully trained. Thus, by learning from sufficient training samples, our method can achieve much higher reconstruction accuracy even under strong occlusions or large poses. Particulary, point correspondences are not required to be constructed explicitly and the cloth details are successfully recovered using our method. The comparison results with model fitting methods demonstrate the superiority of our deep learning based method in 3D human body modeling from a single depth frame. In addition, the inference time of our method (53.4ms per frame) is hundreds of times faster than model fitting methods.

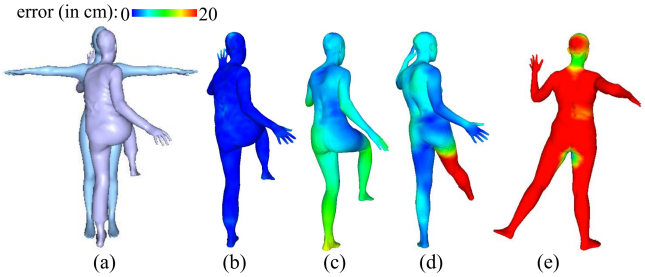


Figure 7: Comparison with model fitting methods on the CAPE data. (a) The template and input depth. (b) Our method. (c) Wei et al. [44]. (d) Bogo et al. [8]. (e) Pure model fitting [23].

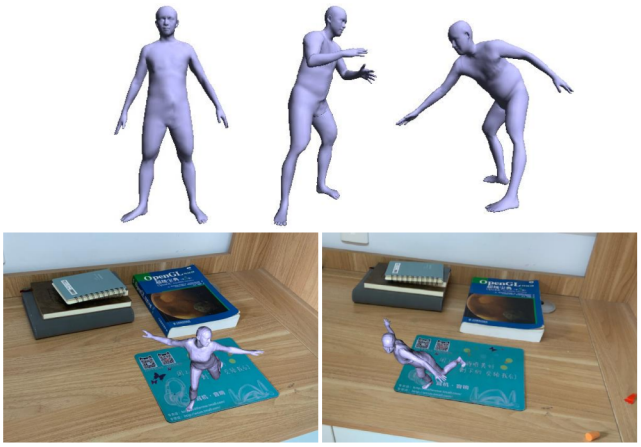


Figure 8: Applications of our method. (Top row) Body re-posing. From the estimated parametric model, we can generate the detailed human shapes with new poses. (Bottom row) Augmented reality. We first estimate the camera parameters and then render the clothed humans on the desk.

4.5 Applications: Re-posing and augmented reality

Our approach has many meaningful applications, and we show two applications here. Our detailed human shape representation decouples the underlying body and the clothing details, and parameterizes them separately. This allows us to edit the detailed human shape conveniently, for example generating clothed bodies with new poses. Our method is first applied on the CAPE data to recover the 3D parametric models. The examples of body re-posing using our method are shown in the top row of Fig. 8. The results are impressive and realistic which show the advantages of our method over the non-parametric methods. In addition, the detailed human models can be integrated into a real scene for VR/AR applications. Given the camera parameters of each frame in a video, we can render the

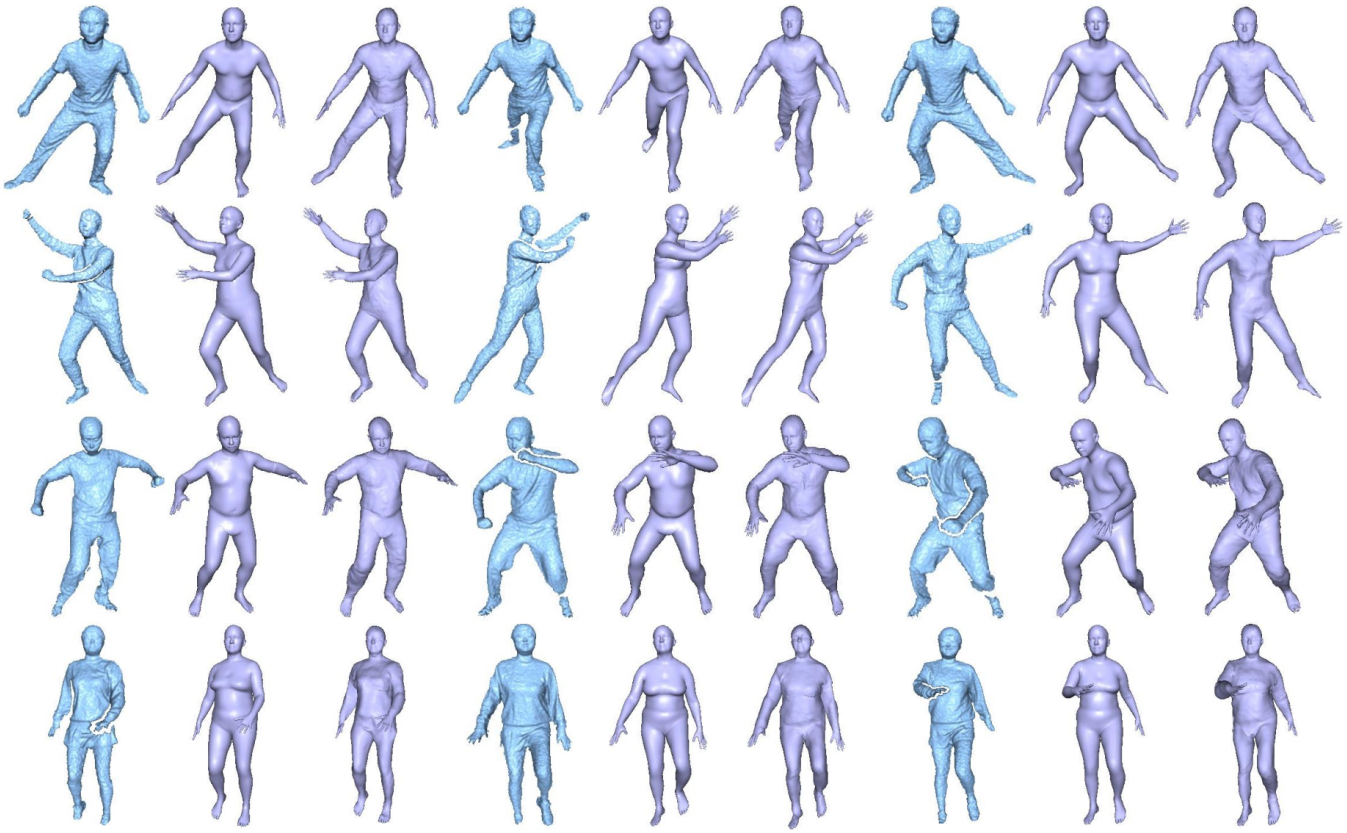


Figure 9: Results of our method tested on real data. For each result, we show the raw depth scan, the predicted undressed model, and the estimated detailed model. From top to bottom: "Crouching" data, "Girl" data from [42], "Kungfu" data from [11], and "Walking" data. The detailed models can fit the raw scans better than the undressed models by recovering the surface details.

recovered human models in a particular scene and observe 3D humans from different viewpoints. The example of augmented reality is shown in the bottom row of Fig. 8. This AR example can give rise to many practical applications in certain circumstances, e.g., AR showing in the museum, entertainment effect in online games, avatar navigation under a SLAM environment, etc. Also, recovering 3D humans has promising applications in telepresence and virtual dressing. Realistic 3D humans recovery plays a footstone role in the immersive telepresence (i.e., Google Starline). In virtual dressing, personalized 3D humans can make online trying on more natural and flexible. Applying our method in these VR/AR applications is a valuable work in the near future.

4.6 Test on real data

We test our method on real data captured by a Kinect V2 sensor. The data of real bodies have a variety of shapes, poses, and clothes. The recovered results using our method are shown in Fig. 9. Although there are strong occlusions and arbitrary deformations on the real data, our method can robustly and accurately estimate 3D shape and pose and clothing details from real point clouds of human bodies. The impressive reconstruction results demonstrate that our method can generalize effectively to real point clouds of clothed bodies with the proposed self-supervised fine-tuning method. The predicted models using our method lack of high-frequency details such as cloth wrinkles which may be due to the smooth features among neighboring graph nodes generated during graph upsampling. We can resort to normal map refinement to further enhance the finer details on the recovered models. Also, our method may fail to recover the details of loose clothes such as skirts since they cannot be represented by the 3D offsets. This problem can be alleviated by using an explicit clothing layer independent from the SMPL model like [15]. We also plan to reconstruct the parametric and detailed 3D human motions by applying our method on sequential point clouds.

5 Conclusion

In this paper, we proposed a hierarchical regression framework for estimating parametric models of 3D dressed humans from partial point clouds of a single depth image. We carefully exploit different levels of discriminative features on point clouds to progressively infer the parametric human model with cascaded regression networks. The global regression network estimates a rough undressed SMPL model using global features, and the local regression network updates the SMPL model with local features of human joints. With the updated SMPL model, the vertex-level regression network extracts vertex-level features from the raw point clouds and regresses the offsets of cloth details. The experimental results on diverse datasets of clothed humans demonstrate the effectiveness of the proposed hierarchical regression networks.

Acknowledgments

The authors would like to thank Xianghui Chen for his kind help in generating the AR example. This work was partially supported by the Natural Science Foundation of China under Grant Nos.61906089, U1713208, and Program for Changjiang Scholars. This work was also supported in part by the Fundamental Research Funds for the Central Universities (NJ2020023), in part by the Open Project Program of the State Key Laboratory of Virtual Reality Technology and Systems of Beihang University (No.VRLAB2021C03), in part by the Open Project Program of the State Key Laboratory of CAD&CG of Zhejiang University (Grant No.A2106), and in part by the Open Project Program of the State Key Laboratory of Novel Software Technology of Nanjing University (No.KFKT2021B19).

REFERENCES

[1] T. Alldieck, M. Magnor, B. L. Bhatnagar, C. Theobalt, and G. Pons-Moll. Learning to reconstruct people in clothing from a single RGB camera. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, jun 2019.

[2] T. Alldieck, G. Pons-Moll, C. Theobalt, and M. Magnor. Tex2Shape: Detailed full human body geometry from a single image. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019.

[3] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis. SCAPE: shape completion and animation of people. *ACM Trans. on Graphics*, 24:408–416, July 2005.

[4] B. L. Bhatnagar, C. Sminchisescu, C. Theobalt, and G. Pons-Moll. Combining implicit function learning and parametric models for 3D human reconstruction. In *European Conference on Computer Vision*, 2020.

[5] 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. In *Advances in Neural Information Processing Systems (NeurIPS)*, December 2020.

[6] B. L. Bhatnagar, G. Tiwari, C. Theobalt, and G. Pons-Moll. Multi-Garment Net: Learning to dress 3D people from images. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019.

[7] F. Bogo, M. J. Black, M. Loper, and J. Romero. Detailed full-body reconstructions of moving people from monocular RGB-D sequences. In *Proceedings of the IEEE International Conference on Computer Vision*, 2015.

[8] 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 *European Conference on Computer Vision*, pp. 561–578, 2016.

[9] F. Bogo, J. Romero, G. Pons-Moll, and M. J. Black. Dynamic FAUST: Registering human bodies in motion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, July 2017.

[10] J. Chibane, T. Alldieck, and G. Pons-Moll. Implicit functions in feature space for 3D shape reconstruction and completion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[11] K. Guo, F. Xu, Y. Wang, Y. Liu, and Q. Dai. Robust non-rigid motion tracking and surface reconstruction using L0 regularization. In *Proceedings of the IEEE International Conference on Computer Vision*, 2015.

[12] K. Guo, F. Xu, T. Yu, X. Liu, Q. Dai, and Y. Liu. Real-time geometry, albedo and motion reconstruction using a single rgbd camera. *ACM Trans. on Graphics*, 2017.

[13] M. Habermann, W. Xu, M. Zollhoefer, G. Pons-Moll, and C. Theobalt. LiveCap: Real-time human performance capture from monocular video. *ACM Trans. on Graphics*, 2019.

[14] M. Habermann, W. Xu, M. Zollhoefer, G. Pons-Moll, and C. Theobalt. DeepCap: Monocular human performance capture using weak supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[15] B. Jiang, J. Zhang, Y. Hong, J. Luo, L. Liu, and H. Bao. BCNet: Learning body and cloth shape from a single image. In *European Conference on Computer Vision*, 2020.

[16] H. Jiang, J. Cai, and J. Zheng. Skeleton-aware 3D human shape reconstruction from point clouds. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019.

[17] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7122–7131, 2018.

[18] A. Kanazawa, J. Y. Zhang, P. Felsen, and J. Malik. Learning 3D human dynamics from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[19] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *International Conference for Learning Representations*, 2015.

[20] N. Kolotouros, G. Pavlakos, M. J. Black, and K. Daniilidis. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In *ICCV*, 2019.

[21] N. Kolotouros, G. Pavlakos, and K. Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[22] C. Lassner, J. Romero, M. Kiefel, F. Bogo, M. J. Black, and P. V. Gehler. Unite the people: Closing the loop between 3D and 2D human representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6050–6059, 2017.

[23] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. SMPL: A skinned multi-person linear model. *ACM transactions on graphics*, 34(6):248, 2015.

[24] Q. Ma, J. Yang, A. Ranjan, S. Pujades, G. Pons-Moll, S. Tang, and M. J. Black. Learning to dress 3D people in generative clothing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[25] R. Newcombe, D. Fox, and S. Seitz. DynamicFusion: Reconstruction and tracking of non-rigid scenes in real-time. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[26] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove. DeepSDF: Learning continuous signed distance functions for shape representation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[27] C. Patel, Z. Liao, and G. Pons-Moll. TailorNet: Predicting clothing in 3D as a function of human pose, shape and garment style. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[28] 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 the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 459–468, 2018.

[29] G. Pons-Moll, S. Pujades, S. Hu, and M. J. Black. ClothCap: seamless 4D clothing capture and retargeting. *ACM transactions on graphics (TOG)*, 36(4), July 2017.

[30] G. Pons-Moll, J. Taylor, J. Shotton, A. Hertzmann, and A. Fitzgibbon. Metric regression forests for correspondence estimation. *International Journal of Computer Vision*, 113:163–175, July 2015.

[31] S. Prokudin, C. Lassner, and J. Romero. Efficient learning on point clouds with basis point sets. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019.

[32] C. R. Qi, H. Su, K. Mo, and L. J. Guibas. PointNet: Deep learning on point sets for 3D classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 652–660, 2017.

[33] C. R. Qi, H. Su, M. Nießner, A. Dai, M. Yan, and L. J. Guibas. Volumetric and multi-view cnns for object classification on 3D data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5648–5656, 2016.

[34] C. R. Qi, L. Yi, H. Su, and L. J. Guibas. PointNet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in Neural Information Processing Systems*, pp. 5099–5108, 2017.

[35] A. Ranjan, T. Bolkart, S. Sanyal, and M. J. Black. Generating 3D faces using convolutional mesh autoencoders. In *European Conference on Computer Vision*, pp. 725–741, 2018.

[36] S. Saito, Z. Huang, R. Natsume, S. Morishima, A. Kanazawa, and H. Li. PIFu: Pixel-aligned implicit function for high-resolution clothed human digitization. *Proceedings of the IEEE International Conference on Computer Vision*, 2019.

[37] S. Saito, T. Simon, J. Saragih, and H. Joo. PIFuHD: Multi-level pixel-aligned implicit function for high-resolution 3D human digitization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[38] O. Sorkine, D. Cohen-Or, Y. Lipman, M. Alexa, C. Rössl, and H.-P. Seidel. Laplacian surface editing. In *Proceedings of the EUROGRAPHICS/ACM SIGGRAPH Symposium on Geometry Processing*, pp. 179–188, 2004.

[39] H. Su, S. Maji, E. Kalogerakis, and E. Learned-Miller. Multi-view convolutional neural networks for 3D shape recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*,

pp. 945–953, 2015.

[40] F. Tan, H. Zhu, Z. Cui, S. Zhu, M. Pollefeys, and P. Tan1. Self-supervised human depth estimation from monocular videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[41] G. Varol, J. Romero, X. Martin, N. Mahmood, M. J. Black, I. Laptev, and C. Schmid. Learning from synthetic humans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 109–117, 2017.

[42] K. Wang, J. Xie, G. Zhang, L. Liu, and J. Yang. Sequential 3D human pose and shape estimation from point clouds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7275–7284, 2020.

[43] K. Wang, G. Zhang, J. Yang, and H. Bao. Dynamic human body reconstruction and motion tracking with low-cost depth cameras. *The Visual Computer*, pp. 1–16, 2020.

[44] L. Wei, Q. Huang, D. Ceylan, E. Vouga, and H. Li. Dense human body correspondences using convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[45] W. Wu, Z. Qi, and F. Li. PointConv: Deep convolutional networks on 3D point clouds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[46] T. Yu, K. Guo, F. Xu, Y. Dong, Z. Su, J. Zhao, J. Li, Q. Dai, and Y. Liu. BodyFusion: Real-time capture of human motion and surface geometry using a single depth camera. In *Proceedings of the IEEE International Conference on Computer Vision*, October 2017.

[47] T. Yu, Z. Zheng, K. Guo, J. Zhao, Q. Dai, H. Li, G. Pons-Moll, and Y. Liu. DoubleFusion: Real-time capture of human performances with inner body shapes from a single depth sensor. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2018.

[48] C. Zhang, S. Pujades, M. Black, and G. Pons-Moll. Detailed, accurate, human shape estimation from clothed 3D scan sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[49] Z. Zheng, T. Yu, Y. Wei, Q. Dai, and Y. Liu. DeepHuman: 3D human reconstruction from a single image. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019.

[50] T. Zhi, C. Lassner, T. Tung, C. Stoll, S. G. Narasimhan, and M. Vo. TexMesh: Reconstructing detailed human texture and geometry from RGB-D video. In *European Conference on Computer Vision*, 2020.

[51] H. Zhu, X. Zuo, S. Wang, X. Cao, and R. Yang. Detailed human shape estimation from a single image by hierarchical mesh deformation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.