Computing and Informatics, Vol. 42, 2023, 457–479, doi: 10.31577/cai_2023_2_457

POINTHUMAN: RECONSTRUCTING CLOTHED HUMAN FROM POINT CLOUD OF PARAMETRIC MODEL

Zongguo Mo, Qicong Wang*

Department of Computer Science and Technology
Xiamen University, Xiamen 361000, China
&
Shenzhen Research Institute, Xiamen University
Shenzhen 518000, China
e-mail: 23020201153790@stu.xmu.edu.cn, qcwang@xmu.edu.cn

Hua Shi

School of Optoelectronic and Communication Engineering
Xiamen University of Technology, Xiamen 361024, Fujian, China
e-mail: shihua@xmut.edu.cn

Baobing Zhang*, Wanxin Sui

Department of Electronic and Electrical Engineering, College of Engineering,
Design and Physical Sciences, Brunel University London, Uxbridge UB8 3PH, UK
e-mail: {Baobing.Zhang, cynthia.sui}@brunel.ac.uk

Abstract. It is very difficult to accomplish the 3D reconstruction of the clothed human body from a single RGB image, because the 2D image lacks the representation information of the 3D human body, especially for the clothed human body. In order to solve this problem, we introduced a priority scheme of different body parts spatial information and proposed PointHuman network. PointHuman combines the spatial feature of the parametric model of the human body with the

* Corresponding author

implicit functions without expressive restrictions. In PointHuman reconstruction framework, we use Point Transformer to extract the semantic spatial feature of the parametric model of the human body to regularize the implicit function of the neural network, which extends the generalization ability of the neural network to complex human poses and various styles of clothing. Moreover, considering the ambiguity of depth information, we estimate the depth of the parameterized model after point cloudization, and obtain an offset depth value. The offset depth value improves the consistency between the parameterized model and the neural implicit function, and accuracy of human reconstruction models. Finally, we optimize the restoration of the parametric model from a single image, and propose a depth perception method. This method further improves the estimation accuracy of the parametric model and finally improves the effectiveness of human reconstruction. Our method achieves competitive performance on the THuman dataset.

Keywords: 3D reconstruction, clothed human reconstruction, SMPL estimation

1 INTRODUCTION

By using intelligent devices to describe and represent the real world has always been a hot and difficult research direction in computer vision and computer graphics areas. The research field of 3D vision has also fast developed in recent years. Lots of 3D human reconstruction research results have been applied in real life. Such as virtual fitting, AR, VR, film, television and 3D games, etc. Creating value for the society while it also brings economic effects. For computer to understand human behavior, participate in human life, realize interaction with humans, it is very important for us to obtain the 3D pose and shape of the human body.

Deep learning is a branch of machine learning. Many traditional machine learning algorithms have a limited learning capacity, and therefore cannot learn the total amount of knowledge with increasing amounts of data. However, deep learning systems can improve performance by accessing more data, a machine surrogate for "more experience". Once a machine has gained enough experience through deep learning, it can be used for specific tasks such as driving a car, face recognition, diagnosing a disease, detecting machine malfunctions, etc. Deep learning can provide a variety of solutions in computer vision, natural language processing, and many other applications. In the future metaverse era, deep learning can play an important role, for example in 3D reconstruction, where deep learning can perform such functions.

The current 3D human body reconstruction methods can be classified into three categories. The first category is to use the existing parametric human body model, such as human parametric model [1], which can directly restore the three-dimensional human body model from a single RGB image or video. The difficulty of recovering 3D model directly from RGB image or video lies in the com-

plexity of the human body, clarity, occlusion, clothing, lighting and the inherent ambiguity of 2D inferring 3D poses. This method does not need specific depth sensor and has a low dependence on external. It is widely used. However, the accuracy of the currently constructed model is far from enough, especially for detailed feature with a hand and face are obviously missing, and no clothing details.

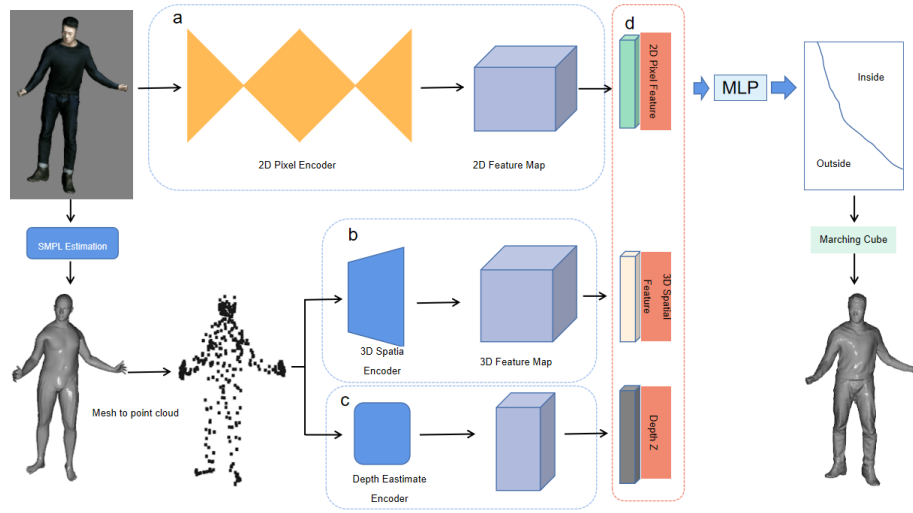


Figure 1. The pipeline of human reconstruction. Given an input image, 2D Pixel Encoder performs pixel feature extraction on the image (a). SMPL estimation is performed on the image to obtain the parametric model, and the parametric model is transformed into a point cloud. 3D Spatial Encoder performs spatial feature extraction on these point clouds (b). Depth Estimation Encoder estimates the offset depth value for these point clouds. The features of a, b and c are fused, and sent to the multi-layer perceptron to predict the distance symbol function value (d), and finally the human body mesh model is obtained.

The second category is the parametric model's deformation. Adding offsets (SMPL + D) to the vertices of the human parametric model to represent a clothed human body is a simple model that is widely used and easy to parameterize. The body geometry of the target pose is obtained by adding the offsets of the vertices under the standard pose of the human parametric model, and then using the skin deformation. There are several previous study [2, 3, 4, 5] to implement. It is difficult to represent SMPL + D for clothes that are not consistent with the SMPL mesh topology, such as open jackets and skirts. Moreover, the binding of clothing to SMPL vertices, especially the binding of mask weights, leads to loose clothing that may be distorted in the mask deformation. And the SMPL + D approach is poorly robust in reconstructing clothing away from the body. It would be better not to adopt a parametric model, such as [6, 7, 8, 9, 10].

The third category is implicit function without using the parameterized model. The pixel alignment implicit function first introduced by PIFu [6] uses MLP to determine the volume occupancy value for a given 3D location. In order to obtain both global and local feature, PIFu [6] uses a deep network to extract the feature of each pixel, and combine this feature together with the depth information of the corresponding 3D point as the input of the MLP to obtain high-fidelity 3D clothed human body reconstruction. Based on PIFu [6], PIFuHD [11] utilizes higher-precision feature and predictes normal information to obtain clothed human reconstructions with more geometric details. Hong et al. [12] use the stereoscopic sense of binocular camera to introduce voxel features to the human body reconstruction and get better results. Summary, the 3D reconstruction of the clothed human body reconstructed from a single RGB image still has the following problems. First, the complexity of the action pose of the person, the ever-changing and different actions of the same person. Second, the self-occlusion of the person, whether the occluded part or the occluded part will lose the integrity Information. Last, RGB images taken by ordinary cameras lack depth information, resulting in depth ambiguity.

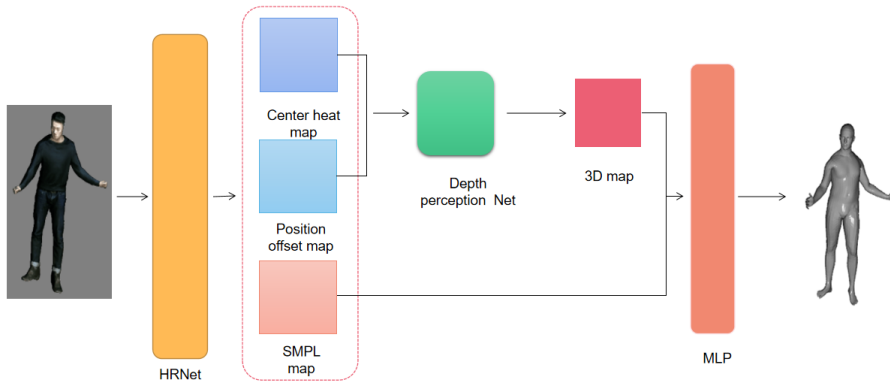


Figure 2. SMPL estimation frame diagram. Inputting an RGB image, HRNet obtains three feature maps: center heat map, position offset map and SMPL map, center heat map and position offset map past depth perception information, and then carry out with the SMPL map Fusion, the SMPL parameters are regressed by the multilayer perceptron

Our three technical contributions are:

- We extract spatial information from the parameterized model, and give the reconstruction network prior knowledge, constrain its spatial expression, meanwhile, impose restrictions on the estimated shape of the human reconstruction. It improves the generalization ability of the neural network to complex human poses and various styles of clothing.
- In order to solve the problem of depth ambiguity, the parametric model contains the relevant coordinate information of each limb of the human body after

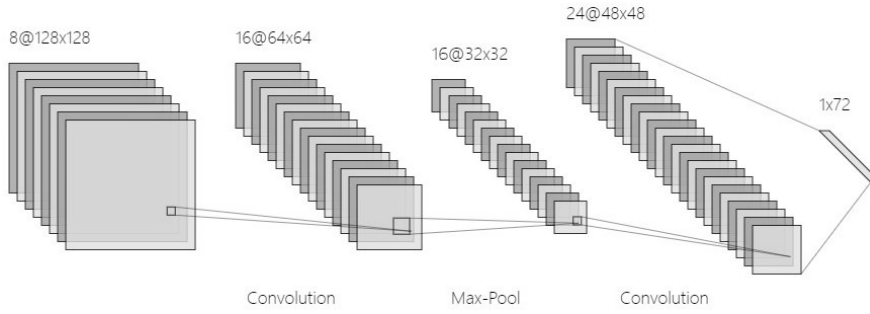


Figure 3. Depth perception Net Frame

point cloud, and the depth can be estimated by the depth network to generate the offset depth value. The offset depth value can use the human body prior information of the predicts depth to guide the occupied space.

- We propose a depth perception method for parameterized model estimation, which reduces the problem of depth ambiguity and restores a more accurate parameterized model.

2 RELATED WORK

2.1 Human Reconstruction Based on Parametric Model

Parametric model capable of changing its parameters to represent the shape of the human body. When human changing its action, the parametric model of the human body will change its parameters to describe the height, short, fat and thin of the human body. Lassner et al. [13] extract 72 joint points of the human body and use random forests to regress SMPL pose and shape parameters. Pavlakos et al. [14] regress SMPL parameters by relying on a smaller number of key points and body contours, further adopt a similar method, then use a segmentation map of human body parts as an intermediate representation. HMR [15] tries to use a weakly supervised method, relying on two-dimensional joint point reprojection penalty and a pre-learned human pose discrimination network, directly using neural network for singe image. Kolotour et al. [5] propose a self-supervised method to solve the same problem. Güler et al. [16] rely on weaker body contour supervision. Rockwell et al. [17] Consider showing only severe occlusion of hand or torso images, to predict the matching SMPL human body. In order to recover more geometric information beyond the body from individual images, such as hand movement and facial expression, Choutas et al. [7] use a body-driven attention technique for extracting high-resolution hand and face from image. A close-up of the part that helps the network predict matching SMPL parameters. Zhang et al. [18] considered how to predict a SMPL human body that matches a 3D scene. For video stream, there are

Figure 4. Our results on a single RGB image. From left to right: the first column is the input image, the second (front), third (side) and fourth (back) columns are the reconstruction results, and the fifth column is the texture inference results, the results show that our method is able to reconstruct high-quality models with robust performance for handling various human poses.

also methods that introduce temporal information to predict SMPL. Among them, Arnab et al. [19] shows that Internet video annotated with SMPLify incorporating temporal continuity can be used to fine-tune HMR results to achieve better results. Kanazaw et al. [20] learn human motion by predicting past and future frames. Sun et al. [15] proposed a temporal model based on a transform network can be used to further improve the effectiveness. VIBE [21] guides action prediction based on priors learned from human sequence motion data. These works focus on using the SMPL parameter space as a homotropic objective. Although the human body reconstruction based on parametric model can capture the movement of the human body and reconstruct the general shape of the human body, it lacks clothing details and is not vivid enough.

2.2 Human Reconstruction Based on Parametric Model Deformation

Adding offsets (SMPL + D) to the vertices of a parametric model of the human body to represent a clothed human body is a simple approach that is widely used and easy to parameterize. By adding the offset of the vertices under the standard pose of the parametric model of the human body, and then using the skin deformation, we obtain the clothing body geometry of the target pose. ClothCap [8] use this representation to separate and reconstruct human clothing for 4D high-quality scan sequences. Zhang et al. [22] use this representation to optimize the shape of naked body that best fit the scan sequences of people. Loop Reg [23] create a self-supervised loop, through end-to-end training, register the scan data of the clothed human body on the SMPL + D representation. Alldieck et al. [2] extract the contour of a rotation sequence of a person roughly in the A pose, and optimize the clothing based on this The SMPL + D representation of the human body. They propose a neural network that uses a few color images and some semantic information to directly return the target SMPL + D representation, greatly increasing the computational speed [24]. Move the texture map space defined in SMP to achieve a higher-resolution SMPL + D representation, which can represent small clothing wrinkles. MGN [4] segmentes the SMPL vertex for different clothing types, so that the reconstructed SMPL + D representation can better express the boundary of the clothing. Bhatnag et al. [25] parameterize the clothing vertex offset as SMPL parameters with the graph convolution representation of clothing parameters, and a generative model of SMPL+D is learned, which supports a small number of clothing types. Inspired by the SMPL + D representation, Sun et al. [28] use hierarchical free-form 3D deformation techniques to improve the predicted body geometry and capture image-compliant details. Weng et al. [26] deform the SMPL model from the normal estimated from a single image to obtain a drivable clothed human body. SMPL + D is simple and compact, but has some limitions. First, there are limited types of clothing that can be expressed.

For clothing that is inconsistent with the SMPL mesh topology, such as open coats, skirts, etc, SMPL + D is difficult to represent. Secondly, due to the binding of the garment to the top of the SMPL, especially the binding of skin weight, resulting in loose clothing and possible skin deformation distortion. SMPL + D method is less robust to garment reconstruction away from the body.

2.3 Human Reconstruction Based on Non-Parametric Model

In order to get rid of the constraints of parametric representation on the complex geometry of the clothed human body, some implicit representations are used for geometric reconstruction. By implicit representation, we mean that a continuous three-dimensional spatial scalar-valued function is defined, and some of its equivalent surfaces are defined as geometric surfaces. The most common implicit representations are the occupancy field (OF) and the signed distance field (SDF). The scalar value of OF is usually a binary value of whether the spatial point is inside

the represented object, while the scalar value of SDF represents the signed distance of the spatial point relative to the represented surface. In the computer, in order to regularize the representation, the spatial implicit function is often discretized with three-dimensional lattice points. More recently, more compact neural representations, capable of efficiently modeling continuous functions, have also become popular in geometric reconstruction. The discrete occupancy field is a lattice discrete representation of a spatially continuous occupancy field. And BodyNet [16] is one of the early works that introduced this representation to human reconstruction. Voxel regression network (VPN) [27] uses an end-to-end convolutional neural network to directly perform voxel regression on 3D human geometry based on various inputs. DeepHuman [9] integrates multi-scale image features into 3D voxel features, solving the problem of poor voxel regression details. Based on the voxel field representation, since voxels reflect occupancy information unlike SDF fields, which have richer geometric information.

The triple memory consumption limits the improvement of resolution and the results are often coarse. The truncated signed distance field is a discretized representation of the SDF field based on three-dimensional lattice points, and at the same time, truncation is performed for larger distance values. This representation is widely used in fusion-based methods using RGB-D inputs.

The pixel-aligned implicit function first introduced by PIFu [6] uses MLP to determine the volume occupancy value for a given 3D location. In order to obtain global and local feature at the same time, PIFu uses a deep network to extract the feature of each pixel, and uses the feature together with the depth information of the corresponding 3D point as the input of MLP, thus obtaining a high-fidelity 3D clothed human body reconstruction. Stereo-PIFu [12] adds voxel-aligned features to pixel-aligned PIFu features to binocular images. And using the predicted voxel for guiding MLP predictions to high accuracy depths that can effectively combat depth blurring, with the recovered geometry details has richer information. Based on PIFu [6], PIFuHD [11] obtained a clothed human body reconstruction with more geometric details by utilizing higher-precision features and predicted normal information. Huang et al. [28] propose a novel multi-scale surface localization algorithm and a direct rendering method without explicit extraction of surface meshes, and for the first time demonstrated real-time reconstruction of the occupancy field of a clothed human body from monocular video and rendering a new perspective. ARCH [28] and ARCH++ [29] try to solve the problem by converting the problem from the pose space to the normative space, but this conversion depends firstly on the pose estimation (HPS) accuracy. Moreover, since the conversion depends on the mask weight attached to SMPL, this weight is hard-coded and defined on the bare body. And forced application it to a clothed person, driven by the action less natural details of the clothes. ICON [30] Predicts SMPL body from image, rendering front and back body normal, and merging it with the original image. Through a normal prediction network, get the positive and negative through normals, apply the normal map to the SMPL. For particularly complex poses, ICON rebuilds as well, but can't do much with looser clothes.

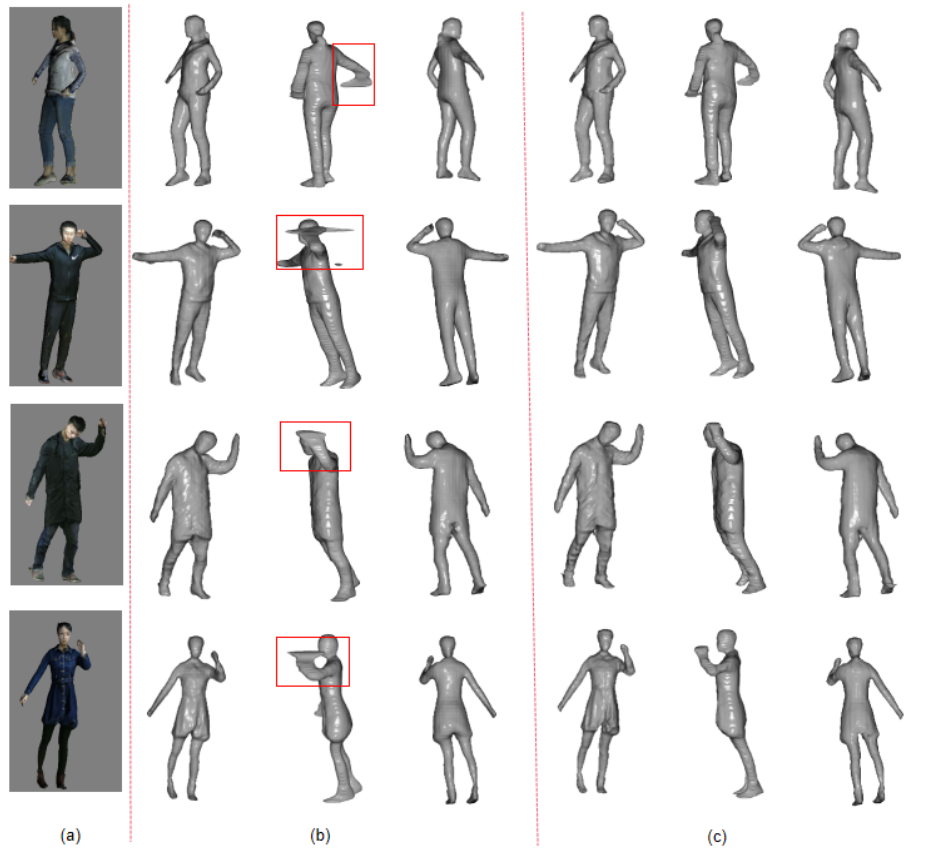


Figure 5. Qualitative comparison against current methods for single-image human model reconstruction: (a) input images, (b) results by PIFu and (c) results by ours

3 METHOD

Our reconstruction of a human body with clothing from a single RGB image is shown in Figure 1. Given a 2D RGB image containing a person, we first estimate its parametric model. The hourglass network performs feature extraction on the image to obtain pixel feature. The parametric model is a grid structure, which consists of vertices and faces. The parametric model is converted into point cloud from mesh. Every point cloud has x, y, and z coordinates. Point Transformer performs 3D spatial feature extraction on every point cloud. After obtaining the spatial information, ResNet extracts the depth information from the point cloud to obtain the depth offset value. The three features are fused as input of the multi-layer perceptron to predict the SDF value. In Figure 1, PointHuman takes a color

Algorithm 1 Training for PointHuman

Input: set of data D. number of optimization steps K and batch size B.
Initialization: randomly initialize g, h, z and fv.

$x \leftarrow 1$
while $x \leq K$ **do**
  $\mathcal{B} \leftarrow \{s_i \in \mathcal{D}\}_{i=1}^{N}$
  for $x_i \in \mathcal{B}$ **do**
    $F(x) = g(I(x))$
    $S(x) = h(I(x))$
    $z(X) = d(I(x))$
    $fv = f((F(x), S(x), z(X))$
    $\mathcal{L}_V = \frac{1}{n}\sum_{i=1}^{n} \left| f_v\left(F_V\left(x_i\right), S\left(x_i\right), z\left(X_i\right)\right) - f_v^*\left(X_i\right)\right|^2$
  end for
  update g, h, z and fv by back-propagation
end while

Output: $s \in \{0, 1\}$

image:

$$f(F(x), S(x), z(X)) = s : s \in \mathbb{R}, \tag{1}$$

where for a 3D point X, $x = \varpi(X)$ is its 2D projection, $S(x) = h(I(x))$ is spatial feature of its parametric model at x. $z(X) = d(I(x))$ is the offset depth value in the camera coordinate space, $F(x) = g(I(x))$ is the image feature at x. For surface reconstruction, we represent the ground truth surface as a 0.5 level-set of a continuous 3D occupancy field:

$$f_v^*(X) = \begin{cases} 1, & \text{if } X \text{ is inside mesh surface,} \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

The total loss function of our network can be formulated as:

$$\mathcal{L}_V = \frac{1}{n}\sum_{i=1}^{n} \left| f_v\left(F_V\left(x_i\right), S\left(x_i\right), z\left(X_i\right)\right) - f_v^*\left(X_i\right)\right|^2, \tag{3}$$

where $X_i \in \mathbb{R}^3$, $F_V(x) = g(I(x))$ is the image feature from the image encoder g at $x = \varpi(X)$ and n is the number of sampled points. Given the input image, the corresponding parameterized model and the corresponding mesh, the parameters of the image encoder, 3D spatial encoder, depth estimation encoder and fv are updated jointly by minimization so that they are consistent with the input image. The parameters of the image encoder, 3D spatial encoder, depth estimation encoder and fv are updated jointly by minimizing Equation (3). The Algorithm 1 provides the training procedure of our proposed framework.

3.1 Spatial Information Extraction

Spatial shape information is one of the characteristics of a 3D object, which contains the representation information of the object and it is an important input for 3D reconstruction. The mesh structure is one of the manifestations of a 3D object, which reflects the size and shape of the object itself. The mesh consists of multiple triangles on one side and multiple discrete points are used to represent continuous faces in the real world. The point cloud of the mesh is ignore the lines between the vertices, take only the vertices, and use all points on the grid, preserving their spatial shape information. Combining them together is the point cloud of the mesh. The point cloud is a collection of three-dimensional data. The point cloud of the grid still retains its size and shape structure, i.e. spatial information. In order to obtain the spatial geometric information of the parametric model, we perform feature extraction on its point cloud. Transformer has achieved impressive results in the NLP domain and 2D image analysis. Compared with language or image processing, transformer may be more suitable for point cloud processing, because the point cloud is essentially a collection of embedded metric spaces, and the core self-attention of Transformer is a collection operator. In addition to this conceptual matching, Transformer has actually achieved good results in the field of point cloud data processing. Therefore, in this paper, we use Point Transformer [31] to extract geometric information from the point cloud of the parametrized model. Point Transformer adopts a network structure similar to U-net [32]. The first half is down-sampling, The second half has the application of trilinear interpolation to obtain the surface information. The first half and the second half are connected to the information, and the network can then extract the deep spatial information of the parameterized model. Point Transformer uses the subtraction relation and add a position encoding δ to both the attention vector γ and the transformed features α:

$$\mathbf{y}_i = \sum_{\mathbf{x}_j \in \mathcal{X}(i)} \rho\left(\gamma\left(\varphi\left(\mathbf{x}_i\right) - \psi\left(\mathbf{x}_j\right) + \delta\right)\right) \odot \left(\alpha\left(\mathbf{x}_j\right) + \delta\right), \tag{4}$$

where $\mathcal{X}(i) \subseteq \mathcal{X}$ is a set of points in a local neighborhood (specifically, k nearest neighbors) of $\mathbf{x}_i$.

3.2 Estimation of Depth Information

Point cloud of the parametric model has 6 890 vertices, i.e. 6 890 3D coordinates, which contain the relevant depth information of the body. To solve the depth ambiguity problem, point cloud of the parametric model contains relative coordinate information of each human limb, which can be used by the network to estimate depth and generate offset depth values. In addition, the offset depth value can be used to guide occupancy prediction using priority information of the predicted depth. Specifically, the offset depth value makes the network easier to train and allows us to produce good surface detail, reducing the occurrence of limb breakage and breakage. Thus, the offset depth value actually acts

as a bridge between predicted depth and occupancy prediction. For some cases, such as the hand in front of the torso, there will be some discontinuous areas in the predicted depth map. In these cases, the back side of the obscured query point will change discontinuously, leading to unnaturally distorted reconstruction results.

We use ResNet [33] for depth estimation, the z-coordinate values of 6 890 vertices are used as input to obtain the offset depth difference of the body torso of the parameterized model. So the input information is vector of size

$$\boldsymbol{R} \in \mathbb{R}^{6\,890\times 1}.$$

That is the coordinates of the point cloud. The last layer of output is vector of size

$$\boldsymbol{R} \in \mathbb{R}^{256\times 5\,000}.$$

The offset depth value and the depth value of the camera are stitched together to get fused depth value. Fused depth value are added together and fed into the multilayer perceptron.

3.3 Estimating Parametric Model

ROMP [34] aims to recover 3D human body from a single image, but due to the lack of depth information, the correct human body cannot be recovered for self-occlusion. Based on this, we propose a depth perception method to solve this problem. We use the HRnet [35] network to process the image, output the center heat map, the position offset map and the SMPL feature map.

The center heat map and the position offset map are fused and fed into the depth perception network to obtain a 3D feature map. 3D feature map and SMPL feature map get parameters of parametric model through multilayer perceptron regression.

The flow chart is shown in Figure 2. The layout of the depth perception network is convolutional layer–pooling layer–convolutional layer–activation function–output layer, the output is the feature vector of size

$$\boldsymbol{R} \in \mathbb{R}^{72\times 1}.$$

Depth perception network structure is shown in Figure 3. And the SMPL feature map is the feature vector of size

$$\boldsymbol{R} \in \mathbb{R}^{82\times 1}.$$

Center heatmap: The front view center heatmap of size

$$\boldsymbol{M_F} \in \mathbb{R}^{1\times H\times W}.$$

It is aligned in pixel space and uses a Gaussian kernel to represent the likelihood of an object being in 2D. We are adding a second 2D heatmap of size

$$\boldsymbol{M_t} \in \mathbb{R}^{1\times D\times W},$$

which represents an unseen top view. This heatmap represents the likelihood that a person is at a certain depth point. However, this map does not represent metric depths. We synthesize and refine these two maps into a 3D center heatmap

$$\boldsymbol{M_o} \in \mathbb{R}^{1\times D\times H\times W},$$

which uses a 3D Gaussian kernel to represent the 3D position of the detected body center.

Position Offset Map: The discretized center heatmap roughly localizes the body, but we expect the network to produce more precise estimates. Likewise, the position offset map includes a front view and a top view. To improve the granularity of 3D localization, we use additional feature map to refine coarse detections by adding estimated offset vectors at each location. Front view offset feature maps of size

$$\boldsymbol{R_f} \in \mathbb{R}^{1\times H\times W}$$

contain 3D offset vectors. The top view offset map of size

$$\boldsymbol{R_t} \in \mathbb{R}^{1\times D\times W}$$

contains a 1-dimensional offset vector for depth correction.

$$\boldsymbol{R_o} \in \mathbb{R}^{1\times D\times H\times W}$$

corresponds to a 3D center map and contains a 3D offset vector.

SMPL map:

$$\boldsymbol{R} \in \mathbb{R}^{128\times H\times W}$$

contains a 128 grid feature vector at each 2D location. These features are aligned with the input 2D image at the pixel level. After feature fusion with 3D feature map, the SMPL parameters are regressed using a multilayer perceptron.

The front view and top view must work together to estimate the position and depth of the person image. We take the concatenation of the front view map and the backbone feature map as input. We unroll and synthesize 2D maps from front view and top view to generate 3D feature map. The 3D feature map and the SMPL feature map are fused, and the parameters of the parameterized model are regressed through the multilayer perceptron.

4 EXPERIMENTS

In this section we evaluate our approach. Details about the implementation are given in Section 4.1. Our ablation experiment in Section 4.3 and Section 4.4. In

Section 4.2 we demonstrate that our method is able to reconstruct human models with challenging poses. We then compare our method to others methods in Section 4.5. The quantitative evaluation results are given in Table 3.

4.1 Implementation Details

Network Architecture. For image feature extraction, we adapt the Hourglass Stack same encoders in PIFu, take an image of 512×512 as input and outputs a 256-channel feature map with size of 128×128. For spatial feature extraction, we use Point Transformer. Its input resolution is $6\,890 \times 3$, and its output is a 64-channel feature volume with a resolution of $64 \times 128 \times 128$.

For depth formation extraction, we make use of ResNet, its input is the depth value of point cloud of Parametric Model resolution is $6\,890 \times 1$, and its output is a 1-channel vector with a resolution of $1 \times 5\,000$.

Training Data. We use THuman dataset, and it contains 6 795 human meshes with various clothes, shape and poses. We split the dataset into a training set of 5 436 meshes and a testing of 1 359 meshes. THuman dataset is more challenging to learn and less likely to cause over-fitting on upstanding human poses and horizontal camera angles than the dataset used in PIFu. The downside of the dataset is that it lacks high quality texture map for photo-realistic rendering, which might hurt model generalization on in-the-wild natural images.

Network Training. We use Adam optimizer for network training with the learning rate of 1×10^{-3}, the batch size is 8, the number of epochs is 45, and the number of sampled points is 5 000 per subject. The learning rate is decayed by the factor of 0.1 at every $10\,000^{\text{th}}$ iteration. It takes 204 hours for a 3 090 graphics card to complete a training session.

Network Infering. A single RGB image as input, and the improved ROMP predicts its corresponding parametrized model. The parametric model is converted into a point cloud through OPEN3D, which is input to the network together with the image, and the final network outputs the parametric model and the textured reconstructed surface.

4.2 Results

We present the results of our method for 3D human reconstruction from a single RGB image in Figure 4. The input image in Figure 4 contains a variety of complex body poses. The results demonstrate that the ability of our method to reconstruct high-quality 3D human models, as well as its strong ability to handle a variety of human poses. Figures 6, 7 and 8 show the training error, IOU and precision for baseline and different fusing methods. In Figure 4, we can see that after we introduce the parametric model, the reconstructed human body achieves good results, the results show that our method is able to reconstruct high-quality models with robust performance for handling various human poses. Compared with PIFu with only

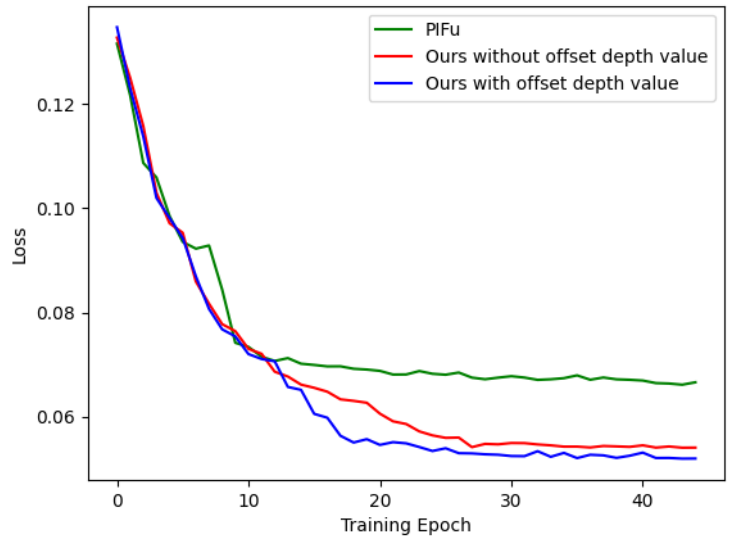


Figure 6. Evaluation of training error. Green line represents PIFu, red line represent our method without offset depth value, and blue line represents our method with offset depth value.

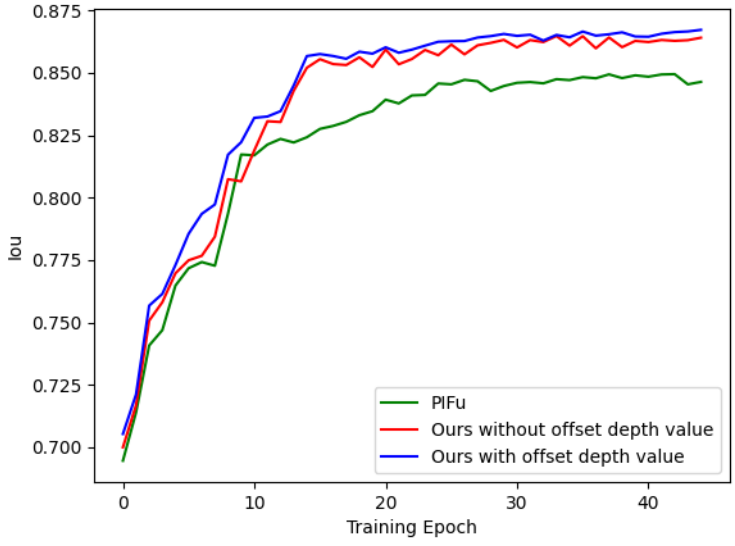


Figure 7. Evaluation of IOU. Green line represents PIFu, red line represents our method without offset depth value, and blue line represents our method with offset depth value.

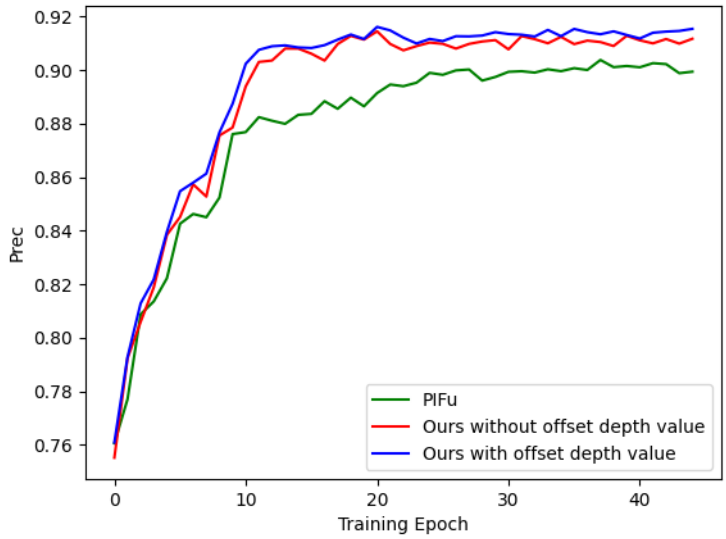


Figure 8. Evaluation of precision. Green line represents PIFu, red line represents our method without offset depth value, and blue line represents our method with offset depth value.

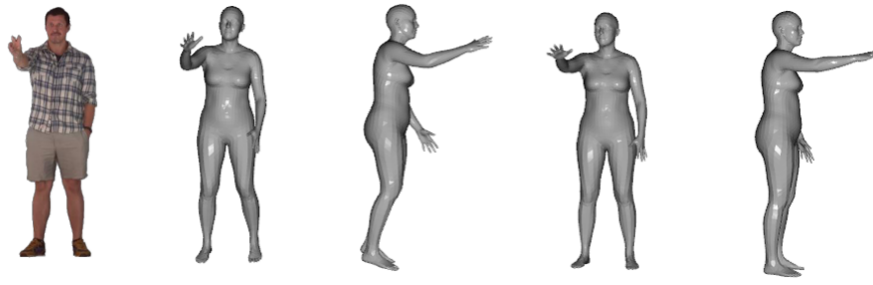


Figure 9. Visualization of the body optimization process. The leftmost column: the input image. The 2nd to 3rd columns: the reconstruction results before reference body optimization. The 4th to 5th columns: the reconstruction results after optimization.

pixel features, after we extract the spatial information of the parameterized model, the Loss, IOU and Prec have made great progress. In addition, after the depth estimation of the parameterized model, the surface details of the reconstructed human body are richer. These indicator is further optimized.

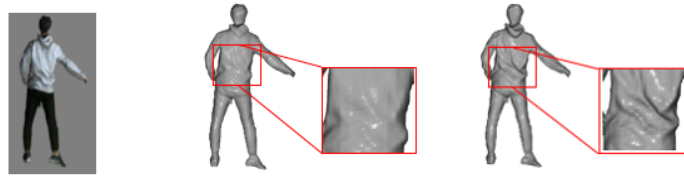


Figure 10. Offset depth values can reconstruct richer details. The leftmost column: the input image. The 2nd to 3rd columns: the reconstruction results without offset depth value and the results with offset depth value.

4.3 Offset Depth Value

We conduct ablation experiments to demonstrate the importance of the inputs in our designed equations for occupancy inference and high-fidelity reconstruction. As shown in Figure 10, PIFu [6] fails to reconstruct reasonable human body geometry using pixel-aligned feature and absolute z-coordinates from a single image due to the complexity and different spatial locations. In contrast, a variant of our PointHuman successfully learns human priors from the same dataset by taking pixel-aligned features, space-aligned features, and the offset depth value as input. Experiments show that our space-aligned feature indeed encode the depth-scale information of query points and further enhance the expressive power of previous work. Table 1 also shows that by replacing absolute z-values with offset depth value, geometric detail can be better recovered. The increase in PointHuman may come from the reconstruction process of the occupancy field, which verifies that our offset depth value indeed effectively utilizes human priors from predicted depth map to guide occupancy inference.

	PSD (cm)	Chamfer (cm)	Normal (cm)
w/o offset depth	2.197	2.312	0.292
w/ offset depth	2.100	2.288	0.281

Table 1. Numerical ablation study of offset depth

4.4 SMPL Estimation

To evaluate the effectiveness of the improved parametric model, we compare the human fitting results before and after improvement using evaluation image. As shown in Figure 9, the optimization step can further fit the SMPL model to the actual human body, resulting in a more accurate body pose estimation. This is also demonstrated in the quantitative evaluation in Table 2, we can also see that the body mesh model reconstruction is also improved after the reference body is optimized.

	PSD (cm)	Chamfer (cm)	Normal (cm)
w/o SMPL optimization	2.203	2.367	0.291
w/ SMPL optimization	2.175	2.301	0.285
Ours using ground-truth SMPL	2.100	2.288	0.281

Table 2. Numerical ablation study of SMPL optimization

4.5 Comparison

We compare our method with several current methods, DeepHuman and PIFu. Among them, PIFu uses deep implicit functions as geometric representation, DeepHuman combines volume representation with SMPL model. We compare with PIFu in Figure 5, PIFu struggles to reconstruct model in challenging pose, while also suffering from self-occlusion. Unlike these methods, our method is able to perform in challenging body poses. Our method outperforms these methods in terms of surface quality and pose generalization ability.

The results of the comparison are shown in Table 3, and the quantitative comparison shows that our method outperforms the methods of Deephuman and PIFu in terms of surface reconstruction accuracy. Overall, our method is more general, more robust and more accurate than DeepHuman and PIFu.

	PSD (cm)	Chamfer (cm)	Normal (cm)
Deephuman [9]	11.246	11.928	0.464
PIFu [6]	4.026	2.604	0.300
Ours	2.100	2.288	0.281

Table 3. Numerical comparison results

5 CONCLUSION

Accurately and robustly reconstructing a 3D human body from a single RGB image is a challenging problem due to the diversity of body movements, clothing types, and other factors. We propose PointHuman to fuse feature of pixel feature, spatial feature and offset depth values implements single-view human mesh reconstruction. Our construction method addresses both spatial priors and deep blurring. The key idea behind our approach to overcome these challenges is to decompose the pose estimation from the surface reconstruction. To this end, we provide a deep-learning based framework that combines the point cloud form of a parametric SMPL model with a non-parametric deep implicit function for reconstructing a 3D human body model from a single RGB image. Our method performs well in terms of robustness and surface detail. For very complex poses and very loose clothing, our method cannot generate reasonable human bodies. Therefore, although the proposed method has taken a step forward in terms of generalization ability, it still

fails in the case of extremely challenging poses. Point Transformer has limited ability to extract spatial information from parametric models and cannot extract spatial information from extremely complex poses. For the invisible area, our PointHuman can only predict a plausible result while can not guarantee its accuracy. The network for spatial feature extraction can be improved, or multi-view reconstruction can be used, so that the reconstructed human body is better. An important future direction is to alleviate the reliance on ground truth and save costs by exploring large-scale image dataset and video dataset for unsupervised training. Additionally, we can consider combining semantic segmentation for reconstruction to solve the problem of not being able to reconstruct loose clothes.

Acknowledgements

This work was supported by Shenzhen Science and Technology Projects under Grant JCYJ20200109143035495 and Natural Science Foundation of Fujian Province (2022J011275).

REFERENCES

[1] Loper, M.—Mahmood, N.—Romero, J.—Pons-Moll, G.—Black, M. J.: SMPL: A Skinned Multi-Person Linear Model. ACM Transactions on Graphics (ToG), Vol. 34, 2015, No. 6, pp. 1–16, doi: 10.1145/2816795.2818013.

[2] Alldieck, T.—Magnor, M.—Xu, W.—Theobalt, C.—Pons-Moll, G.: Video Based Reconstruction of 3D People Models. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8387–8397, doi: 10.17863/CAM.85609.

[3] Alldieck, T.—Pons-Moll, G.—Theobalt, C.—Magnor, M.: Tex2shape: Detailed Full Human Body Geometry from a Single Image. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 2293–2303, doi: 10.1109/ICCV.2019.00238.

[4] Feng, Y.—Choutas, V.—Bolkart, T.—Tzionas, D.—Black, M. J.: Collaborative Regression of Expressive Bodies Using Moderation. 2021 International Conference on 3D Vision (3DV), IEEE, 2021, pp. 792–804, doi: 10.1109/3DV53792.2021.00088.

[5] Kolotouros, N.—Pavlakos, G.—Black, M. J.—Daniilidis, K.: Learning to Reconstruct 3D Human Pose and Shape via Model-Fitting in the Loop. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 2252–2261, doi: 10.1109/ICCV.2019.00234.

[6] Saito, S.—Huang, Z.—Natsume, R.—Morishima, S.—Kanazawa, A.—Li, H.: Pifu: Pixel-Aligned Implicit Function for High-Resolution Clothed Human Digitization. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 2304–2314, doi: 10.1109/ICCV.2019.00239.

[7] Choutas, V.—Pavlakos, G.—Bolkart, T.—Tzionas, D.—Black, M. J.: Monocular Expressive Body Regression Through Body-Driven Attention. European Conference on Computer Vision, Springer, 2020, pp. 20–40, doi: 10.1007/978-3-030-58607-2_2.

[8] Pons-Moll, G.—Pujades, S.—Hu, S.—Black, M. J.: Clothcap: Seamless 4d Clothing Capture and Retargeting. ACM Transactions on Graphics (ToG), Vol. 36, 2017, No. 4, pp. 1–15, doi: 10.1145/3072959.3073711.

[9] Zheng, Z.—Yu, T.—Wei, Y.—Dai, Q.—Liu, Y.: Deephuman: 3D Human Reconstruction from a Single Image. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 7739–7749, doi: 10.1109/ICCV.2019.00783.

[10] Sitzmann, V.—Martel, J.—Bergman, A.—Lindell, D.—Wetzstein, G.: Implicit Neural Representations with Periodic Activation Functions. Advances in Neural Information Processing Systems, Vol. 33, 2020, pp. 7462–7473.

[11] Saito, S.—Simon, T.—Saragih, J.—Joo, H.: Pifuhd: Multi-Level Pixel-Aligned Implicit Function for High-Resolution 3D Human Digitization. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 84–93, doi: 10.1109/CVPR42600.2020.00016.

[12] Hong, Y.—Zhang, J.—Jiang, B.—Guo, Y.—Liu, L.—Bao, H.: Stereopifu: Depth Aware Clothed Human Digitization via Stereo Vision. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 535–545, doi: 10.1109/CVPR46437.2021.00060.

[13] Lassner, C.—Romero, J.—Kiefel, M.—Bogo, F.—Black, M. J.—Gehler, P. V.: Unite the People: Closing the Loop Between 3D and 2D Human Representations. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6050–6059, doi: 10.1109/CVPR.2017.500.

[14] Pavlakos, G.—Zhu, L.—Zhou, X.—Daniilidis, K.: Learning to Estimate 3D Human Pose and Shape from a Single Color Image. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 459–468, doi: 10.1109/CVPR.2018.00055.

[15] Sun, Y.—Ye, Y.—Liu, W.—Gao, W.—Fu, Y.—Mei, T.: Human Mesh Recovery from Monocular Images via a Skeleton-Disentangled Representation. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 5349–5358, doi: 10.1109/ICCV.2019.00545.

[16] Güler, R. A.—Neverova, N.—Kokkinos, I.: Densepose: Dense Human Pose Estimation in the Wild. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7297–7306, doi: 10.1109/CVPR.2018.00762.

[17] Rockwell, C.—Fouhey, D. F.: Full-Body Awareness from Partial Observations. European Conference on Computer Vision, Springer, 2020, pp. 522–539, doi: 10.1007/978-3-030-58520-4_31.

[18] Zhang, Y.—Hassan, M.—Neumann, H.—Black, M. J.—Tang, S.: Generating 3D People in Scenes Without People. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 6194–6204, doi: 10.1109/CVPR42600.2020.00623.

[19] Arnab, A.—Doersch, C.—Zisserman, A.: Exploiting Temporal Context for

3D Human Pose Estimation in the Wild. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 3395–3404, doi: 10.1109/CVPR.2019.00351.

[20] Kanazawa, A.—Zhang, J. Y.—Felsen, P.—Malik, J.: Learning 3D Human Dynamics from Video. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 5614–5623.

[21] Kocabas, M.—Athanasiou, N.—Black, M. J.: Vibe: Video Inference for Human Body Pose and Shape Estimation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 5253–5263, doi: 10.1109/CVPR42600.2020.00530.

[22] Zhang, C.—Pujades, S.—Black, M. J.—Pons-Moll, G.: Detailed, Accurate, Human Shape Estimation from Clothed 3D Scan Sequences. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4191–4200, doi: 10.1109/CVPR.2017.582.

[23] Bhatnagar, B. L.—Sminchisescu, C.—Theobalt, C.—Pons-Moll, G.: Loopreg: Self-Supervised Learning of Implicit Surface Correspondences, Pose and Shape for 3D Human Mesh Registration. Advances in Neural Information Processing Systems, Vol. 33, 2020, pp. 12909–12922.

[24] Alldieck, T.—Magnor, M.—Bhatnagar, B. L.—Theobalt, C.—Pons-Moll, G.: Learning to Reconstruct People in Clothing from a Single RGB Camera. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 1175–1186, doi: 10.1109/CVPR.2019.00127.

[25] Bhatnagar, B. L.—Tiwari, G.—Theobalt, C.—Pons-Moll, G.: Multi-Garment Net: Learning to Dress 3D People from Images. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 5420–5430, doi: 10.1109/ICCV.2019.00552.

[26] Weng, C. Y.—Curless, B.—Kemelmacher-Shlizerman, I.: Photo Wake-Up: 3D Character Animation from a Single Photo. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 5908–5917, doi: 10.1109/CVPR.2019.00606.

[27] Jackson, A. S.—Manafas, C.—Tzimiropoulos, G.: 3D Human Body Reconstruction from a Single Image via Volumetric Regression. Proceedings of the European Conference on Computer Vision (ECCV) Workshops, 2018, pp. 64–77, doi: 10.1007/978-3-030-11018-5_6.

[28] Huang, Z.—Xu, Y.—Lassner, C.—Li, H.—Tung, T.: Arch: Animatable Reconstruction of Clothed Humans. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 3093–3102, doi: 10.1109/CVPR42600.2020.00316.

[29] He, T.—Xu, Y.—Saito, S.—Soatto, S.—Tung, T.: ARCH++: Animation-Ready Clothed Human Reconstruction Revisited. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 11046–11056, doi: 10.1109/ICCV48922.2021.01086.

[30] Xiu, Y.—Yang, J.—Tzionas, D.—Black, M. J.: Icon: Implicit Clothed Humans Obtained from Normals. Proceedings of the IEEE Conference on

Computer Vision and Pattern Recognition (CVPR), Vol. 2, 2022, doi: 10.1109/CVPR52688.2022.01294.

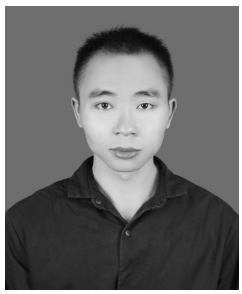
[31] Zhao, H.—Jiang, L.—Jia, J.—Torr, P. H.—Koltun, V.: Point Transformer. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 16259–16268.

[32] Ronneberger, O.—Fischer, P.—Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2015, pp. 234–241.

[33] Targ, S.—Almeida, D.—Lyman, K.: Resnet in Resnet: Generalizing Residual Architectures. Arxiv Preprint Arxiv:1603.08029, 2016.

[34] Sun, Y.—Bao, Q.—Liu, W.—Fu, Y.—Michael J., B.—Mei, T.: Monocular, One-Stage, Regression of Multiple 3D People. ICCV, 2021, doi: 10.1109/ICCV48922.2021.01099.

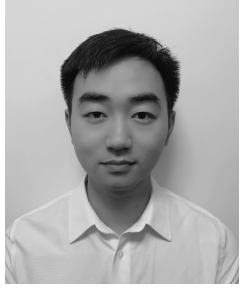
[35] Yu, C.—Xiao, B.—Gao, C.—Yuan, L.—Zhang, L.—Sang, N.—Wang, J.: Lite-Hrnet: A Lightweight High-Resolution Network. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 10440–10450, doi: 10.1109/CVPR46437.2021.01030.



Zongguo Mo is pursuing a graduate degree from the Department of Computer Science and Technology, at Xiamen University, Fujian, China. His research interests include deep learning, human reconstruction and human pose estimation.

Qicong Wang received his Ph.D. in information and communication engineering from Zhejiang University, Hangzhou, China. He is currently Associate Professor at the Department of Computer Science and Technology, Xiamen University, Xiamen, China. His research interests include computer vision, machine learning, and big data analytics.

Hua Shi is a lecturer at the School of Optoelectronic and Communication Engineering, Xiamen University of Technology in China. He received his Ph.D. from the Xiamen University, P.R. China in 2014. His research is in the areas of machine learning, computer vision, and artificial intelligence.



Baobing Zhang received his Ph.D. degree in artificial intelligence from the Brunel University London, UK in 2020. He is currently Post-Doctoral Research Fellow at the Brunel University London. His research interests include deep learning, computer vision, image processing, data privacy, and AI applications.

Wanxin Sui is currently Ph.D. candidate at the Brunel University London, UK. Her research interests are in the areas of data protection and privacy, privacy-preserving AI techniques, and AI applications in higher education.