The Visual Computer (2023) 39:3609–3619
https://doi.org/10.1007/s00371-023-02957-0

ORIGINAL ARTICLE

Topology-preserved human reconstruction with details

Lixiang Lin[1] · Jianke Zhu[1]

Accepted: 9 June 2023 / Published online: 5 July 2023

Abstract
Due to the high diversity and complexity of body shapes, it is challenging to directly estimate the human geometry from a single image with the various clothing styles. Most of the model-based approaches are limited to predict the shape and pose of a minimally clothed body with over-smoothing surface. While capturing the fine detailed geometries, the model-free methods are lack of the fixed mesh topology. To address these issues, we propose a novel topology-preserved human reconstruction approach by bridging the gap between model-based and model-free human reconstruction. We present an end-to-end neural network that simultaneously predicts the pixel-aligned implicit surface and an explicit mesh model built by graph convolutional neural network. Experiments on DeepHuman and our collected dataset showed that our approach is effective. The code will be made publicly available at https://github.com/l1346792580123/sdfgcn.

Keywords Topology preserved · Human reconstruction

1 Introduction

Human reconstruction has been studied for decades, which is essential to a large amount of real-world applications, including motion capture, digital entertainments, etc. Generally, it is challenging to directly estimate the geometry of human from a single RGB image due to the high diversity and complexity of body shapes. Moreover, the sophisticated clothing styles often lead to the extra difficulties in human reconstruction.

To address the critical problem of accurately modeling the human body, statistical human models such as SCAPE [1] and SMPL [2] have been proposed. These models use principal component analysis (PCA) and blend skinning to reduce the search space and generate parametric models. Recently, deep neural network-based methods [3,4] have attempted to estimate the model parameters directly from images, avoiding the time-consuming nonlinear optimization process. While these approaches have achieved promising results, they are still limited in their ability to capture the shape and pose of a fully clothed body with fine details, as they tend to produce over-smooth surfaces. In spite of some parametric clothing models [5–7], they may not generalize well in the real-world scenario.

Instead of relying on the parametric models, the model-free approaches [10,11] directly reconstruct the human body from a single image, which enjoy the merits of recovering the fine detailed geometries. To this end, human body is either estimated by the occupancy of small voxels [10,11] or implicitly represented by a function learned by deep neural network [9,12]. The main showstopper for these methods is that there is no commonly shared topology for the reconstructed body geometries. Therefore, it is difficult to find the semantic correspondences between the reconstructed mesh and human body part in contrast to the model-based approaches. This further prevents them from animating the reconstructed body directly.

This paper proposes a new approach to address the limitations of existing methods for human reconstruction. By combining model-based and model-free techniques, we aim to accurately reconstruct the body mesh with the same topology as the SMPLX model. To achieve this, we present an end-to-end neural network that predicts both the pixel-aligned implicit surface and the explicit mesh model using a graph convolutional neural network. The decoder branches of the network all share the same feature encoder, which significantly reduces computational costs during inference. To refine the output of the neural network, we propose an effective implicit registration stage, which is performed in implicit space without the need for the computationally intensive Chamfer distance. Our approach preserves topology to

✉ Jianke Zhu
jkzhu@zju.edu.cn

[1] College of Computer science and technology, Zhejiang University, Hangzhou City, Zhejiang Province, China

ensure a more accurate and efficient reconstruction of human bodies.

In summary, the main contributions of this paper are: (1) an end-to-end neural network that reconstructs the fine detailed body mesh while retaining the fixed topology from a single image; (2) a graph convolutional autoencoder to recover human mesh model with the fixed topology; (3) an efficient implicit registration method to refine the predicted mesh; and (4) empirical evaluations on DeepHuman and our collected dataset showing promising human reconstruction results.

2 Related work

Recovering 3D human body shapes from 2D images or videos is the fundamental problem in computer vision, which has already been extensively studied for decades. [13–15] summarize recent methods for 3D human pose estimation. It is not enough to represent human body using the articulated 3D joint locations. A full mesh is required to represent the human body. Generally, most of the existing approaches can be roughly divided into two categories. The first is dependent on the parametric models, which formulates the human body reconstruction as a regression problem. On the other hand, the model-free methods try to reconstruct detailed human geometry directly.

2.1 Model-based human reconstruction

Due to the high diversity and complexity of poses with various shapes, it is very challenging to build the human body models with the desired generalization capability.

During past fifteen years, a surge of research efforts have been devoted to building the statistical human body models from 3D laser scans [1,2,8,16]. Loper et al. [2] build a skinned vertex-based model with the shape and pose parameters, in which the pose-dependent blend shapes are a linear function of the elements of the pose rotation matrices. This makes it easy to integrate the human body generation process into the deep neural network pipeline for back propagation. Recently, graph convolutional network has became more and more important in dealing with non-rigid shape like face [17], which requires fewer parameters and can achieve higher accuracy compared with the parametric models. Choi et al. [18] propose a graph convolutional network that recovers 3D human mesh from 2D human pose.

With the parametric human models, human reconstruction is reduced to the parameter estimation problem, where the coefficients and joints transformation are directly predicted from the still image. The conventional methods [8,19,20] employ the nonlinear optimization solver to obtain the reasonable solution, which are usually computationally intensive. Kanazawa et al. [3] propose an end-to-end framework to recover the human body shape and pose by estimating SMPL parameters using only 2D joints annotations with an adversarial loss. Kolotouros et al. [4] introduce a self-improving system which combines optimization and prediction methods. [21] propose an alternating successive convex approximation to decouple the relationship between joint positions and SMPL parameters into joint shape and joint pose relationships separately. Most of these approaches only produce a naked human body, where the surfaces of clothing, hair and other accessories are ignored.

To tackle the above problem, clothing is represented as an offset layer from the underlying body in [22–25], which is able to change the pose and shape of the reconstruction using SMPL. Yang et al. [7] train a neural network to regress a PCA-based representation of clothing. Moreover, Lähner et al. [6] learn a garment-specific pose deformation model by regressing the low-frequency PCA components and high-frequency normal maps. Adam model [16] is clothed while the shape is very smooth and not pose-dependent. Recently, Ma et al. [26] have presented a generative 3D mesh model of clothed people, which is conditioned on both pose and clothing type. This enables the capability of drawing clothing samples to dress different body shapes in a variety of styles and poses. Corona et al. [27] propose an implicit model to represent different garment in a unified manner. In contrast to these methods, our proposed approach does not require to build an extra parametric model for the dressing, which is able to handle the cases without clothing as well.

Bhatnagar et al. [28] recover human mesh from the incomplete point cloud by an implicit neural network to jointly predict the outer 3D surface of the dressed person, the inner body surface and the semantic correspondences to the SMPL model. Saito et al. [29] propose an end-to-end trainable framework that takes raw 3D scans of a clothed human and turns them into an animatable avatar. Ma et al. [30] predict the articulated surface elements to dress the bodies with the realistic clothing that moves and deforms naturally even in the presence of topological changes. Although the above methods get the detailed human mesh with the fixed topology, they requires point cloud as the input comparing to the RGB images used in this work (Fig. 1).

2.2 Model-free human reconstruction

Model-free approaches try to directly estimate human body geometry like voxels or implicit surface from the still image without resorting to a prior model, which have much larger solution space to represent the fine details.

Varol et al. [10] suggest to learn a voxel representation of human body through the deep neural network, where the fine-scale details are often missing due to the high memory requirements of voxel representations. Zheng et al. [11] introduce a discretized volumetric representation to reconstruct

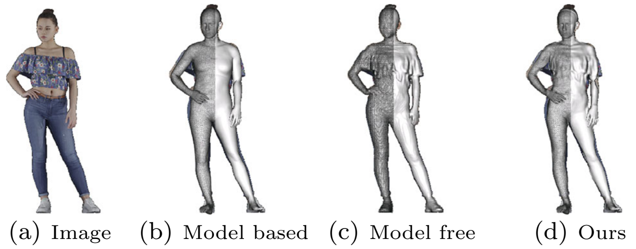


Fig. 1 **a** is the input image. **b** Model-based approach [8] tries to estimate SMPLX parameters, which mainly captures the shape and pose without the details like clothing. **c** Model-free method [9] recovers the fine detailed body geometry while the reconstructed mesh does not have the predefined topology. **d** Our approach can directly estimate the accurate body mesh with the fixed topology

the detailed human, which fuses the different scales of image features in order to recover the accurate surface geometry. In spite of impressive results, the cubic memory requirement imposed by the discrete voxel representation prevents these methods from obtaining the high-resolution reconstruction results. Instead of using the voxels, some approaches [31,32] try to predict the depth maps of human as output. Natsume et al. [33] present a multi-view inference method by synthesizing silhouette views from a single image. Although multi-view silhouettes are more memory efficient, the concave regions are difficult to infer as well as the consistently generated views.

Saito et al. [12] propose a memory efficient approach that represents the detailed human by a pixel-aligned implicit function. Instead of explicitly discretizing the output space into voxels, it learns a function that determines the occupancy for any locations. With such implicit representation, the occupancy for the sampled 3D point can be computed on the fly, which greatly saves the memory. Later, Saito et al. [9] introduce a multi-level architecture for high-resolution 3D human reconstruction, where the coarse level focuses on the holistic reasoning and the fine level estimates the highly detailed geometry.

3 Methods

In this section, we present our proposed approach for human reconstruction from a single image. Firstly, we propose the end-to-end neural framework that reconstructs the fine detailed body mesh while retaining the fixed topology. Secondly, we describe the model-free reconstruction using implicit surface loss. Thirdly, we introduce the graph convolutional network approach to recover human mesh model. Finally, we propose an effective implicit registration stage to fill the gap between the pixel-aligned implicit surface and the recovered mesh.

3.1 Overview

The model-based human reconstruction method [3] enjoys the merit of the predefined mesh topology, which is able to preserve the body shape through the statistical models. On the other hand, the model-free method can recover the fine detailed geometry like wrinkles on the clothing. The key idea of our proposed approach is to take advantage of both representations. To this end, we aim to reconstruct the triangulated body mesh accurately while preserving the same topology as SMPLX model. As illustrated in Fig. 2, we present an end-to-end deep neural network with a typical encoder–decoder structure.

Our overall framework consists of four parts. Firstly, we train a Pix2PixHD network [34] with nine residual blocks

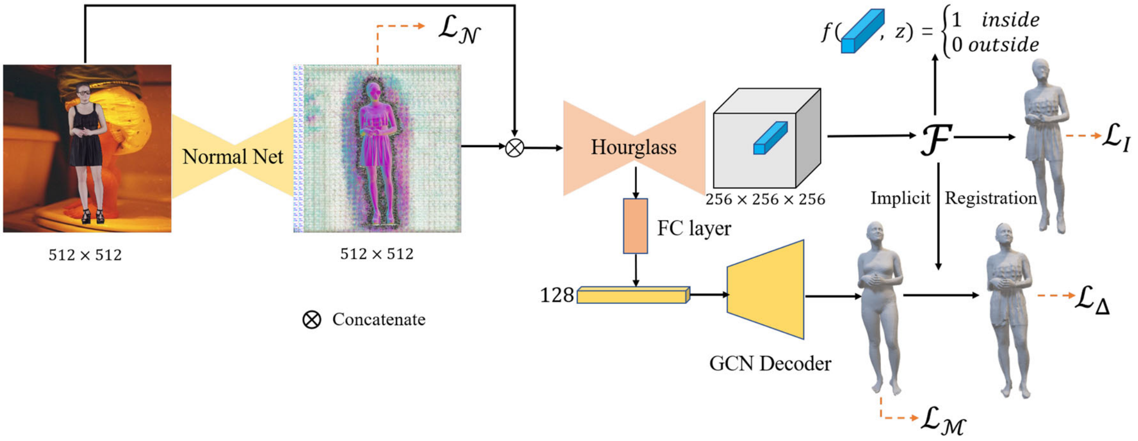


Fig. 2 Overview of our proposed human reconstruction approach. We firstly concatenate the input image with the estimated normal map to feed the hourglass encoder. Then, a pixel-aligned implicit function predicts the occupancy, and a GCN (graph convolutional network) decoder estimates the mesh model. Finally, the mesh model is refined through an effective implicit registration stage

and four downsampling layers to obtain the frontal normal map. Secondly, we concatenate the predicted normal map with the original image as the input for the stacked hourglass network [35] with four stacks to extract deep features. Hourglass network produces a feature map, where we perform average pooling to get a feature vector as the input of our GCN (graph convolutional network) decoder. Thirdly, a fully connected layer with the number of neurons of (1024, 1024, 1024, 128) is used to adapt the number of features. Finally, a fully connected layer with the number of neurons (257, 1024, 512, 256, 128, 1) and the skip connections at (3, 4, 5) layers are employed to predict the binary occupancy value for any given positions. A pretrained GCN decoder is used to get the human mesh with the fixed topology.

In [9], the frontal normal map is predicted as a proxy for 3D geometry. Features extracted from the frontal normal map can generate the sharper reconstructed results. Therefore, we firstly employ a Pix2PixHD network to obtain the frontal normal map, and then concatenate it with the original image as the input for the feature encoder. The Pix2PixHD network is trained with the following loss function:

$$\mathcal{L}_{\mathcal{N}} = \lambda_N \sum_{\{i,j\}\subset P} |n_{i,j} - n^*_{i,j}|, \tag{1}$$

where $\mathcal{L}_{\mathcal{N}}$ is the regular L_1 loss. $\{i, j\} \subset P$ represents the valid set of pixels in image, and $n_{i,j}$ and $n^*_{i,j}$ are the corresponding ground-truth normal vector and predicted normal vector, respectively. We try to predict the frontal normal map of the person in the image. The weight is $\lambda_N = 5$. We use Adam optimizer with the learning rate of 2×10^{-4} until the convergence.

It is worth mentioning that we suggest to share the same feature encoder for all the decoder branches. This greatly reduces the computational cost during the inference. Moreover, a decoder branch is employed to predict the implicit surface function for the model-free human reconstruction, and another decoder branch is used to extract the explicit mesh surface using graph convolutional neural network trained on a large corpus. More importantly, we propose an extra implicit registration stage to fill the gap between the other two branches, which intends to reduce the registration error between the triangulated mesh and implicit surface.

From the above all, the proposed deep neural network minimizes the following loss function:

$$\mathcal{L} = \mathcal{L}_{\mathcal{I}} + \mathcal{L}_{\mathcal{M}} + \mathcal{L}_{\Delta}, \tag{2}$$

where $\mathcal{L}_{\mathcal{I}}$ is the loss for implicit surface function estimation, $\mathcal{L}_{\mathcal{M}}$ is the loss to recover the human mesh through GCN decoder and $\mathcal{L}_{\Delta}$ is the loss for the implicit registration stage, which bridges the gap between the model-free reconstruction and parametric mesh model.

3.2 Implicit reconstruction loss

Motivated by the previous model-free human reconstruction approach [12], we try to estimate the body surface through an implicit function $f(\cdot)$ that approximates the signed distance of zero level set. The implicit surface shares the same coordinate space as SMPLX mesh model. Specifically, a fully connected layer is employed to predict the binary occupancy value for any given positions $X_i = (x_i, y_i, z_i) \in \mathbb{R}^3$ in the continuous 3D space:

$$f(\mathcal{F}_{\mathbf{x}_i}, Z_i) = \begin{cases} 1, & \text{if } X_i \text{ is inside mesh} \\ 0, & \text{otherwise} \end{cases}, \tag{3}$$

where $\mathcal{F}_{\mathbf{x}_i}$ denotes the deep features sampled at the location $\mathbf{x}_i = (u_i, v_i) = \pi(X_i)$ in image space $\Omega \subset R^2$. The projection function $\pi : R^3 \rightarrow \Omega$ can be either orthogonal projection or perspective projection. $Z_i = (MX_i)^z$ is the depth value in camera coordinate space, and M is the camera extrinsic matrix.

Given the ground-truth occupancy $y(X_i)$ at point X_i, we employ the extended binary cross-entropy (BCE) loss [11] to supervise our proposed implicit surface representation layer. Therefore, the implicit reconstruction loss $\mathcal{L}_I$ can be derived as follows:

$$\begin{aligned} \mathcal{L}_I = & \sum_{X_i \in \mathcal{S}} \eta y(X_i) \log f(X_i) \\ & + (1-\eta)(1-y(X_i)) \log(1 - f(X_i)), \end{aligned} \tag{4}$$

where $\mathcal{S}$ is the set of the sampled points. η represents the ratio of points outside surface in $\mathcal{S}$, which is computed from the sampling results. A mixture sampling strategy is used to select the points for the implicit reconstruction loss computation. In our experiment, the sampled points are composed of the uniform sampling and importance sampling with the standard deviations of 0.04 and the ratio of 8 : 1. For the ground-truth points and their occupancy, we make use of the DeepHuman dataset [11] and our collected high-resolution human scans.

3.3 Mesh recovery loss

The model-based human reconstruction has the merit of the watertight mesh representation with the data-driven priors, where the generative SMPLX model [8] is recently proposed. It has a mesh with 10,475 vertices and 54 body joints. Moreover, an extra joint is used to control the global rotation, which is parameterized by the PCA shape coefficients and poses. Although formulating the pose blend shapes as a linear function of the rotation matrices, the whole procedure of mesh generation is still highly nonlinear. Therefore, it is challenging to regress them from a single image directly.

To deal with this problem, we suggest to make use of a graph convolutional network-based autoencoder to capture human body shapes, which shares the same mesh topology as SMPLX model without the blend skinning. In this paper, we employ the same loss function described in [18] to train our proposed GCN autoencoder and regress the latent coefficients from the extracted feature. For each vertex V_i in $\mathcal{M}$ with the target V_i^*, the mesh recovery loss $\mathcal{L}_{\mathcal{M}}$ is defined as below:

$$\mathcal{L}_{\mathcal{M}} = \lambda_v \sum_{V_i \in \mathcal{M}} ||V_i - V_i^*||_1 + \lambda_e \mathcal{L}_{edge} + \lambda_n \mathcal{L}_{normal}, \quad (5)$$

where the first term denotes the per vertices L_1 fitting loss, and the last two terms regularize the mesh deformations on edges and normals, respectively. Let $\mathcal{T}$ represent a facet in $\mathcal{M}$, and (i, j) are the vertex indices in $\mathcal{T}$. The edge length loss is derived as follows:

$$\mathcal{L}_{edge} = \sum_{\mathcal{T} \in \mathcal{M}} \sum_{\{i,j\} \in \mathcal{T}} |||V_i - V_j||_2 - ||V_i^* - V_j^*||_2|. \quad (6)$$

Given the target normal n_f^* for each facet $\mathcal{T}$, the normal consistency loss is defined as in [18]:

$$\mathcal{L}_{normal} = \sum_{\mathcal{T} \in \mathcal{M}} \sum_{\{i,j\} \in \mathcal{T}} \left| \left\langle \frac{V_i - V_j}{||V_i - V_j||_2}, n_f^* \right\rangle \right|. \quad (7)$$

The weights are $\lambda_v = 10$, $\lambda_e = 40$ and $\lambda_n = 0.5$, respectively.

As in [26], our proposed autoencoder consists of an encoder–decoder pair built by graph convolutional network. To embed the input data into the latent space, the encoder is made of eight Chebyshev residual blocks with Chebyshev polynomial of order two, a Chebyshev convolution with order one and a fully connected layer. Each graph convolution layer is followed by a Leaky ReLU [36]. The architecture of decoder is similar to the encoder. For the detailed network structure, please refer to the supplementary materials. To effectively capture the various body shapes and poses, we train this autoencoder on AMASS datasets [37]. In contrast to COMA [17] reconstructing the smoothing facial meshes, our proposed method has to tackle the critical challenges of body articulations and blend skinning.

Once GCN autoencoder is trained, we freeze the model parameters of decoder and integrate it into our proposed human reconstruction framework to facilitate the mesh model generation. Moreover, we formulate the model-based reconstruction as the GCN latent embedding estimation problem. We employ $\mathcal{L}_{\mathcal{M}}$ to supervise the training process.

3.4 Implicit registration loss

In order to take advantage of both implicit function and topology reserved human model, we propose a novel implicit registration loss to capture the detailed clothing information from implicit function. $\mathcal{L}_{\Delta}$ is defined as follows:

$$\mathcal{L}_{\Delta} = \lambda_{sdf} \mathcal{L}_{sdf} + \mathcal{L}_{reg}, \quad (8)$$

where $\lambda_{sdf} = 10$. $\mathcal{L}_{sdf}$ is defined as follows:

$$\mathcal{L}_{sdf} = \sum_{V_i \in \mathcal{M}} ||f(\mathcal{F}_{\pi(V_i + M^{-1}\Delta_i)}, (MV_i)^z + \Delta_i) - \sigma||_1, \quad (9)$$

where $f(\cdot)$ is the pixel-aligned implicit function defined in Eq. (3) and $\Delta = (0, 0, \Delta_z)$ is an optimizable variable initialized to 0. Since the learned implicit function is fed with the depth along the ray defined by the 2D projection, we only optimize it along z-axis. M is the camera extrinsic matrix, and σ is set to 0.5.

The regularization term $\mathcal{L}_{reg}$ is proposed to enforce the surface smoothing through minimizing the mesh Laplacian differences and the L_2 norm of offset Δ as below:

$$\mathcal{L}_{reg} = \lambda_{lap} ||L(V + M^{-1}\Delta) - L(V)||_2^2 + \lambda_{norm} ||\Delta||_2^2, \quad (10)$$

where L denotes the Laplacian matrix that retains the mesh regularity. The L_2 norm over mesh offset Δ prevents the vertices from shifting too large. The regularization coefficient λ_{lap} is set to 10^4, and λ_{norm} is 50.

Since the neural network prediction is close enough to the implicit surface of model-free reconstruction, $\mathcal{L}_{sdf}$ is able to guarantee the convergence. Our proposed implicit registration method does not calculate the point-to-surface Chamfer distance which is very computationally intensive. Thus, the registration is performed very efficiently in implicit space without extracting the explicit mesh by the marching cube algorithm [38].

4 Experiments

In this section, we give the details of our experimental implementation and discuss the results on human reconstruction. We examine the representation capability of our proposed GCN autoencoder for human body. Moreover, we evaluate our results on DeepHuman and our collected human dataset.

4.1 Experimental setup

AMASS [37] is used to train our GCN autoencoder. AMASS is a large database of human motion datasets with a common framework and parameterization. AMASS contains a large variety of SMPL and SMPLX parameters for human motion. Due to the flexibility of the face and hands, AMASS dataset does not provide the ground-truth SMPLX parameters for face and hands. With the SMPLX topology, the face and hands can be fitted with other algorithm [8]. Since DeepHuman dataset only provides the ground-truth SMPL parameters, we train a SMPL GCN autoencoder. To effectively optimize the GCN parameters, we use Adam optimizer with the learning rate of 10^{-4} and a weight decay of 10^{-4} for 10 epochs.

To facilitate the effective experimental evaluation, we conduct the experiments on DeepHuman dataset [11] and our collected scans. DeepHuman dataset contains the total number of 6,795 items, including RGB image, SMPL parameters and meshes reconstructed by a multi-view fusion algorithm. We randomly split the samples to form training and testing sets with a ratio of 9 : 1, and obtain 6,115 items for training and 680 samples for testing. We crop the images according to their height and place the human at the center. Then, the cropped images are resized into the resolution of 512×512. Due to the privacy issue, the facial regions in image are blurred. Being difficult to recover the thin structures like fingers, the hand geometry of the mesh in the dataset are presented in the form of fists.

We collected 260 high-resolution photogrammetry scans from this website [39], which are collected and uploaded by its users. The whole dataset is spitted into a training set of 234 subjects and a testing set of 26 subjects. We render these meshes using the off-the-shelf software Blender. Each subject is rendered from every 18 degree in yaw axis with an elevation fixed with 0°. As in [9], we randomly augment the background images using COCO dataset [40]. In our experiment, we render the images in the resolution of 1024×1024 and then scale it into 512×512 as the input of our network. After rendering, we employ the conventional optimization-based method [8] to fit the SMPLX model with respect to each scan.

We implemented the proposed approach by PyTorch. The normal estimation network is trained using Adam optimizer with the learning rate of 2×10^{-4} until convergence. We train the pixel-aligned implicit function and deep feature for the GCN latent space encoder with 60 epochs. We use RMSprop optimizer with the learning rate 5×10^{-4} that is decayed by the factor of 0.1 at 40th epoch. We employ same sampling strategy as PIFu [12] to sample 12000 points to train the implicit function. In the implicit registration stage, Adam optimizer with learning rate 0.002 is employed to optimize the offset of mesh vertices with 500 iterations.

Table 1 Evaluation on autoencoder

Dataset	MPVPE (mm)
AMASS (SMPLX)	4.628
Human3.6 M (SMPL)	4.913

Table 2 Performance evaluation on DeepHuman dataset. * indicates this output mesh has the same topology as parametric human model. Units for point-to-surface and Chamfer distance are in cm

Methods	Normal	P2S	Chamfer
PIFu [12]	0.020	2.718	2.327
Ours w/o normal	0.018	2.413	2.229
Ours	**0.010**	**1.317**	**1.152**
*Our SMPL results	0.030	1.434	1.278
*SMPL refined by Chamfer loss	0.020	1.324	1.170
*SMPL refined by implicit loss	0.026	1.335	1.175

The bold in tables means the best performance.

4.2 Evaluation on autoencoder

We evaluate the performance of our GCN autoencoder on AMASS dataset and Human3.6 M dataset [41]. Human3.6 M consists of 3.6 million 3D Human poses acquired by recording the performance of 5 female and 6 male subjects.

The mean per vertex position error (MPVPE) is similar to MPJPE [3] while we make use of all vertices to evaluate the representation capability of our GCN autoencoder. As shown in Table 1, our proposed GCN autoencoder achieves almost the same reconstruction results as the conventional parametric SMPLX and SMPL model with the fewer latent parameters.

4.3 Evaluation on deep human

As in PIFu [12], we adopt three reconstruction performance metrics including the mean point-to-surface Euclidean distance (P2S), Chamfer distance and normal projection error. P2S and Chamfer distance measure the reconstruction accuracy comparing to the ground-truth mesh. Additionally, the normal projection error is used to evaluate the fineness of reconstructed local details as well as the projection consistency.

Table 2 gives the experimental results on DeepHuman dataset. It can be seen that our proposed approach performs better than PIFu. Moreover, the normal map can significantly improve the reconstruction accuracy and capture the clothing details, which makes it easier for the implicit function to retain the rich local details.

In the implicit registration stage, we add the clothing details obtained from our trained implicit function to the SMPL mesh. The results show that our proposed implicit

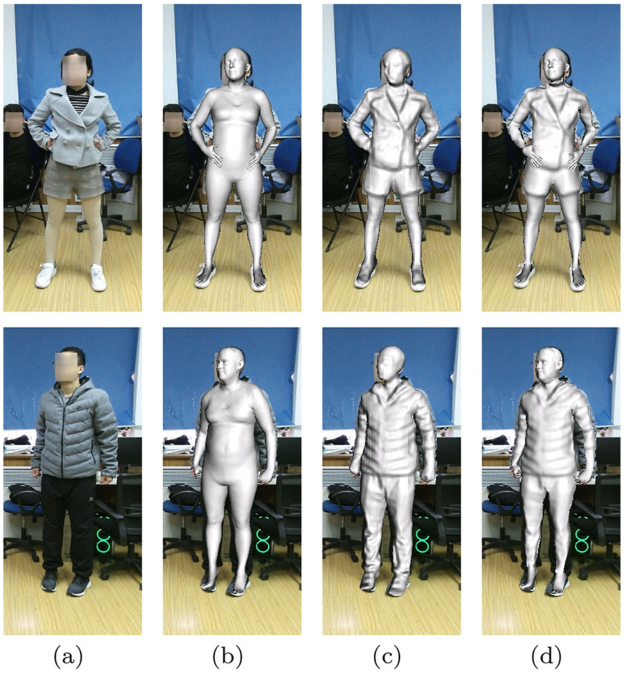


Fig. 3 Reconstruction results on DeepHuman dataset. We show the results of SMPL **b**, our implicit function results **c** and implicit registration results **d** on input image **a**, respectively

Table 3 Performance evaluation on model-based reconstruction

Methods	MPJPE	Reconst. error
Linear model	50.058	42.367
GraphCMR GCN	43.227	39.247
GraphCMR linear	40.276	36.023
Ours GCN decoder	**37.598**	**32.140**

The bold in tables means the best performance.

registration method performs comparable with conventional Chamfer distance-based method. Qualitative results of our method are shown in Fig. 3, and GCN decoder predicts the coarse mesh with the same topology as SMPL. After implicit registration stage, the vertices offsets representing the clothing details are obtained from the implicit function. Due to the flexibility and low reconstruction quality of hands, feet and face, we do not optimize them in the SMPL model.

For model-based reconstruction results, we compare our GCN results with HMR [3] and GraphCMR [42]. We employ the same input and backbone network for all the methods. The mean per joint position error (MPJPE) and reconstruction error are used as the performance metrics. Table 3 shows the experimental results. It can be clearly seen that our proposed GCN decoder obtains the lower estimation error comparing to other methods, which demonstrates the effectiveness of our GCN for body mesh representation. Figure 4 shows the visual results. As we perform graph convolution in spectral domain and have the normal and edge regularization, the

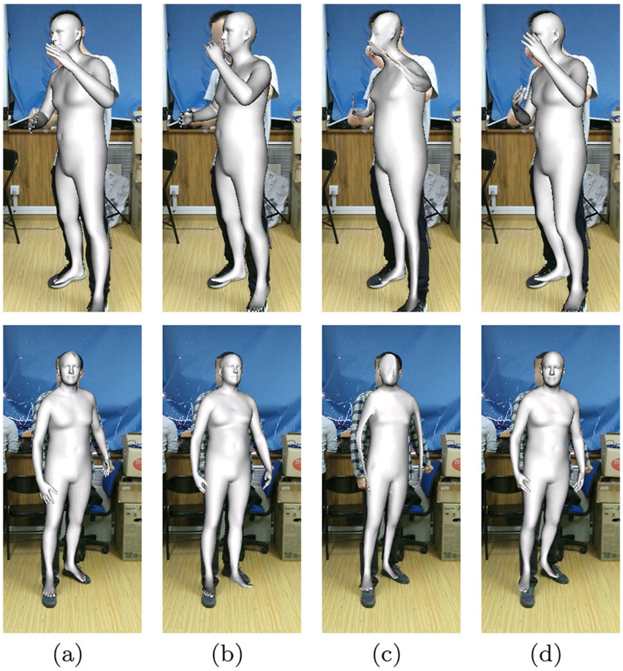


Fig. 4 Comparisons on the model-based reconstruction. We show the results of our GCN decoder **a**, linear model to predict the SMPL parameters **b**, GCN results of GraphCMR **c** and final results of GraphCMR **d**

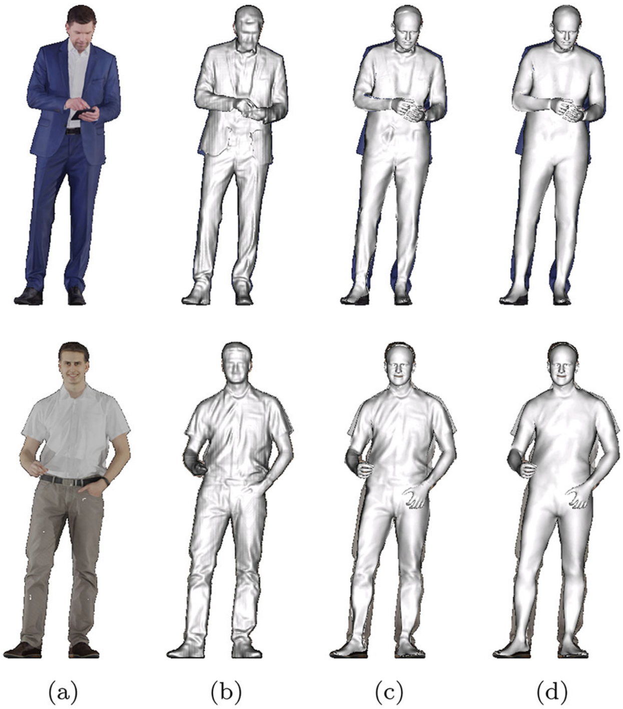


Fig. 5 Comparison on the implicit registration results **b**, results with subdivided SMPLX topology **c** and not **d** on input image **a**

Table 4 Reconstruction Performance evaluation on our collected dataset. Same settings and notations as in Table 2

Methods	Normal	P2S	Chamfer
*SMPLicit [27]	0.029	2.245	2.055
*HMD [43]	0.035	2.495	2.532
PIFuHD (our implementation)	0.007	1.023	1.053
Ours	**0.005**	**0.969**	**0.965**
*Our SMPLX results	0.026	1.101	1.097
*SMPLX refined by Chamfer loss	0.013	0.977	0.975
*SMPLX refined by implicit loss	0.018	1.010	1.016

The bold in tables means the best performance.

mesh generated by our GCN is almost the same as the SMPL mesh. GraphCMR [42] generates over-smoothing mesh and needs another linear model to predict the SMPL parameters from the vertices predicted by GCN.

4.4 Evaluation on our collected dataset

Since either SMPL or SMPLX model has too few vertices to capture all clothing information during implicit registration stage, we subdivide its topology to generate more vertices. More specifically, for each face, we add the midpoint of each edge to subdivide every triangle into four facets. The effect of subdivision is shown in Fig. 5. It can be seen that we can get more detailed registration results after subdivision.

We compare our proposed topology-preserved human reconstruction approach against the conventional Chamfer distance-based registration method used in IPNet [28], HMD [43], PIFuHD [9] and SMPLicit [27]. Note that PIFuHD does not make their training code publicly available and the coordinate space of PIFuHD in human reconstruction is inconsistent with our collected scans, we re-implement PIFuHD for evaluation. Due to the limited number of high-resolution scans having collected, the latent feature for GCN decoder cannot generalize well. Therefore, we further refine our GCN output to get the correct pose.

We use the same performance metrics described above to evaluate our approach. As shown in Table 4, our implicit function network performs comparable with PIFuHD. The mesh is predicted by GCN decoder with the same topology as SMPLX model, which can be effectively refined by our proposed implicit registration scheme. It can be seen that the proposed approach performs comparable with conventional Chamfer distance-based method. Figure 6 shows the reconstruction results. Due to the limited training data, SMPLicit [27] cannot represent all kinds of clothing and lack of details. HMD [43] refines the SMPL model according to per-pixel shading information. The reconstruction results lack a lot of details and is inconsistent with the input image. Since there are no ground-truth SMPLX parameters for face and hand in AMASS dataset, we employ the traditional optimization-based approach [8] to capture the hand and face for better visualization. By taking advantage of the fixed topology, we can easily animate the recovered mesh with the arbitrary poses, as shown in Fig. 7. The latent vector for arbitrary poses can be generated by our pretrained GCN encoder.

As depicted in Table 5, the speed of our implicit registration process is about seven times faster than the conventional Chamfer distance-based method. This is because the proposed implicit loss is efficient to compute while the Chamfer loss is computationally intensive requiring to find the nearest neighbors in target mesh for each vertex in query mesh. Moreover, our method does not require to extract the mesh by marching cube [38], which saves the extra computational time. Although the optimization time of HMD [43] is short, the optimization results is not detailed and realistic.

5 Ablation studies

In this section, we conduct ablation studies on the loss weight. There is only $\mathcal{L}_{\mathcal{N}}$ loss in normal net training. We choose a suitable λ_N to match the learning rate. In our framework, λ_v and λ_{sdf} are the weights for data term, while λ_e, λ_n, λ_{lap} and λ_{norm} are the weights for regularization term. We set the appropriate weights so that the regularization term is about a quarter of the data term. We change one of the weights to different values and leave the remaining weights unchanged. The experimental results are shown in Table 6. It can be seen that different loss weights have less effect on the results.

6 Conclusions and limitations

We introduced a new method for human reconstruction that aims to combine the strengths of both model-based and model-free approaches. Our method preserves the topology of the reconstructed human by utilizing an end-to-end neural network that predicts both the pixel-aligned implicit surface and the explicit mesh surface through graph convolutional neural network. Additionally, we propose an efficient implicit registration method to refine the network output in the implicit space. We have conducted the evaluation on DeepHuman and our collected high-resolution human dataset. The encouraging experimental results showed that our proposed approach is able to effectively recover the accurate mesh model while preserving its topology.

While our proposed approach has yielded promising results, it is important to note that it may encounter challenges due to the complexity of the background and the significant gap between the GCN mesh and implicit surface. However, we can overcome these obstacles by making use of the off-the-shelf human segmentation techniques to elim-

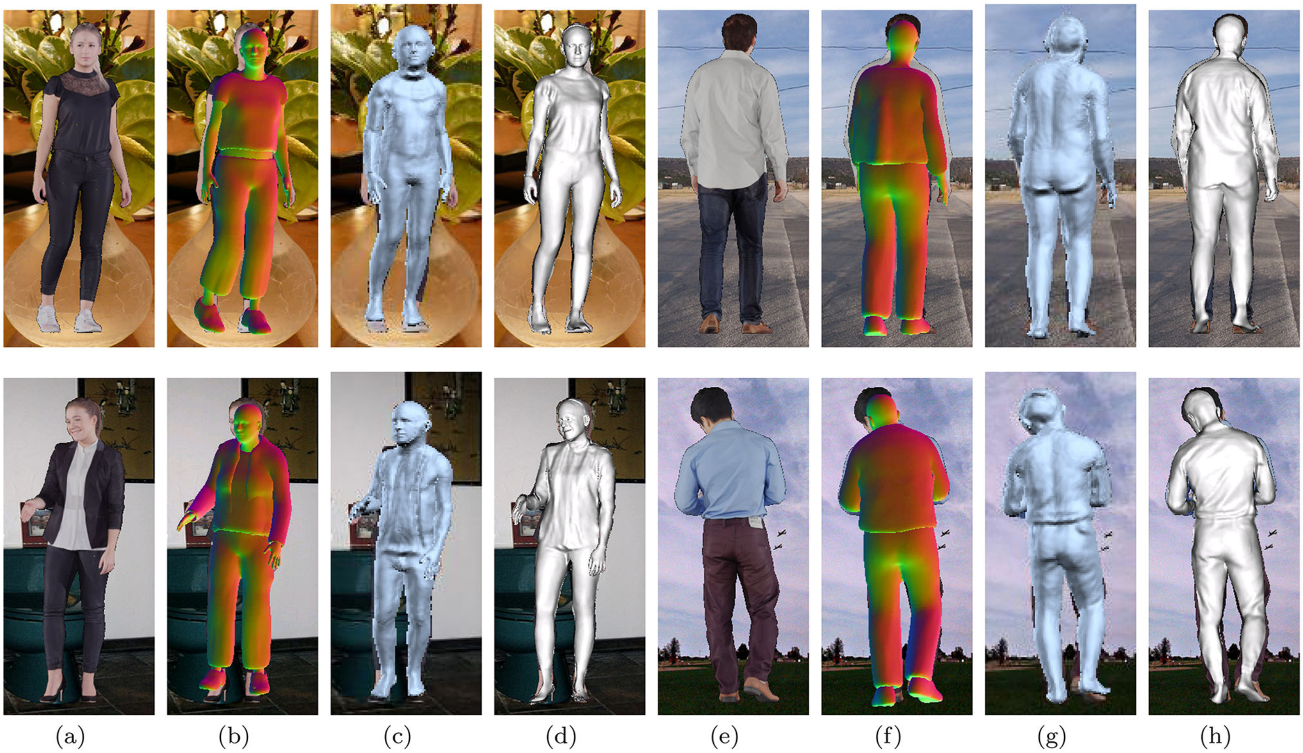


Fig. 6 Results on our collected dataset. We show the results of SMPLicit [27] (**b**, **f**), hmd [43] (**c**, **g**) and our implicit registration results (**d**, **h**) on the input image (**a**, **e**), respectively

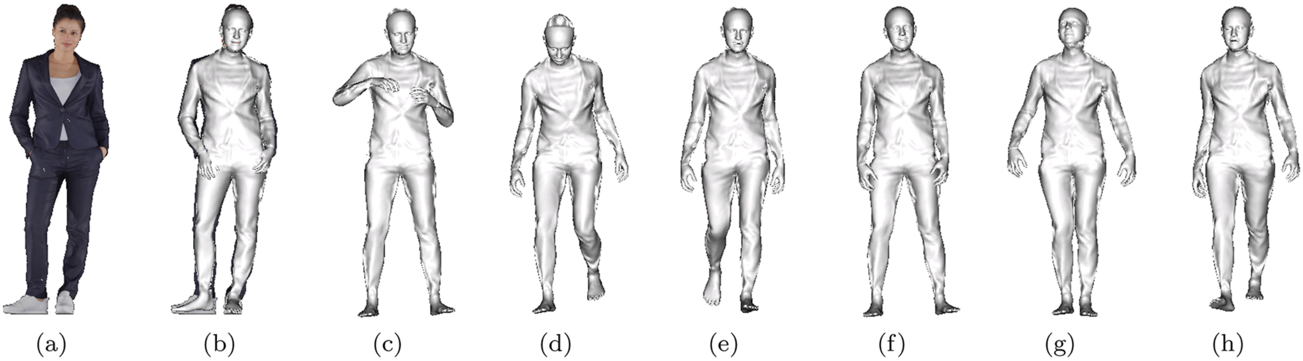


Fig. 7 Visual results of reposing the recovered mesh

Table 5 Comparison on computational time

Methods	Marching cube	Optimization	Total
Chamfer registration	58.2 s	140.3 s	198.5 s
HMD [43]	–	14.0 s	14.0 s
Ours	–	27.0 s	27.0 s

inate background. Additionally, we can align the results of the GCN and implicit models using Chamfer distance-based optimization to address the gap issue. It is worth mentioning that we optimize the latent vector of GCN rather than the offset per vertex. Nevertheless, our method may not be ideal for loose clothing items such as long skirts like other SMPL-based techniques.

7 Changes

Compared to our conference version, we have carefully revised the whole manuscript according to the reviewers' comments. Firstly, we have included the additional reconstruction results from the extra views to demonstrate the generalization capability of our proposed approach. Secondly, we have conducted ablation studies on the parameter weights to fully evaluate the impact on reconstruction results.

Table 6 Ablation studies on the loss weights. Same settings and notations as in Table 4

	Normal	P2S	Chamfer
$*\lambda_v = 1$	0.028	1.127	1.124
$*\lambda_v = 100$	0.027	1.117	1.115
$*\lambda_e = 1$	0.028	1.129	1.127
$*\lambda_e = 100$	0.027	1.121	1.119
$*\lambda_n = 0.1$	0.029	1.127	1.122
$*\lambda_n = 1$	0.028	1.123	1.120
$\lambda_{sdf} = 1$	0.005	1.034	1.021
$\lambda_{sdf} = 100$	0.005	0.983	0.978
$*\lambda_{lap} = 10^2$	0.019	1.027	1.029
$*\lambda_{lap} = 10^6$	0.020	1.025	1.022
$*\lambda_{norm} = 1$	0.019	1.023	1.020
$*\lambda_{norm} = 100$	0.019	1.025	1.023

In the revised manuscript, we also provided the further clarification on our motivation and highlighted the key differences between our proposed approach and the previous methods. Furthermore, we have thoroughly discussed the advantages and limitations of these methods. Finally, we have carefully proofread the entire manuscript and corrected the written errors, including typos and grammar mistakes.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00371-023-02957-0.

Funding This work is supported by the National Natural Science Foundation of China under Grants (61831015).

Declarations

Conflict of interest The authors declare that they have no conflicts of interest.

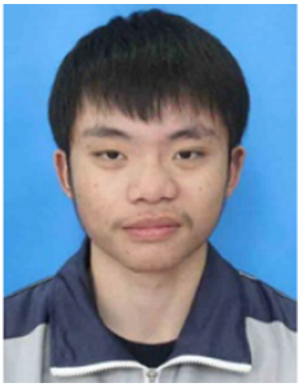
References

1. Anguelov, D., Srinivasan, P., Koller, D., Thrun, S., Rodgers, J., Davis, J.: SCAPE: shape completion and animation of people. ACM Trans. Graph. **24**(3), 408–416 (2005)
2. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: a skinned multi-person linear model. ACM Trans. Graph. **34**(6), 248:1-248:16 (2015)
3. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, pp. 7122–7131 (2018)
4. Kolotouros, N., Pavlakos, G., Black, M.J., Daniilidis, K.: Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In: International Conference on Computer Vision ICCV, pp. 2252–2261 (2019)
5. Neophytou, A., Hilton, A.: A layered model of human body and garment deformation. In: International Conference on 3DV, pp. 171–178 (2014)
6. Lähner, Z., Cremers, D., Tung, T.: Deepwrinkles: Accurate and realistic clothing modeling. In: European Conference on Computer Vision ECCV, vol. 11208, pp. 698–715 (2018)
7. Yang, J., Franco, J., Hétroy-Wheeler, F., Wuhrer, S.: Analyzing clothing layer deformation statistics of 3d human motions. In: European Conference on Computer Vision ECCV, vol. 11211, pp. 245–261 (2018)
8. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3d hands, face, and body from a single image. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, pp. 10,975–10,985 (2019)
9. Saito, S., Simon, T., Saragih, J.M., Joo, H.: Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, pp. 81–90 (2020)
10. Varol, G., Ceylan, D., Russell, B.C., Yang, J., Yumer, E., Laptev, I., Schmid, C.: Bodynet: Volumetric inference of 3d human body shapes. In: European Conference on Computer Vision ECCV, vol. 11211, pp. 20–38 (2018)
11. Zheng, Z., Yu, T., Wei, Y., Dai, Q., Liu, Y.: Deephuman: 3d human reconstruction from a single image. In: International Conference on Computer Vision ICCV, pp. 7738–7748. IEEE (2019)
12. Saito, S., Huang, Z., Natsume, R., Morishima, S., Li, H., Kanazawa, A.: Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In: International Conference on Computer Vision ICCV, pp. 2304–2314 (2019)
13. Chen, Y., Tian, Y., He, M.: Monocular human pose estimation: a survey of deep learning-based methods. Comput. Vis. Image Underst. **192**, 102,897 (2020)
14. Desmarais, Y., Mottet, D., Slangen, P., Montesinos, P.: A review of 3d human pose estimation algorithms for markerless motion capture. Comput. Vis. Image Underst. **212**, 103,275 (2021)
15. Wang, J., Tan, S., Zhen, X., Xu, S., Zheng, F., He, Z., Shao, L.: Deep 3d human pose estimation: a review. Comput. Vis. Image Underst. **210**, 103,225 (2021)
16. Joo, H., Simon, T., Sheikh, Y.: Total capture: A 3d deformation model for tracking faces, hands, and bodies. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, pp. 8320–8329 (2018)
17. Ranjan, A., Bolkart, T., Sanyal, S., Black, M.J.: Generating 3d faces using convolutional mesh autoencoders. In: European Conference on Computer Vision ECCV, vol. 11207, pp. 725–741 (2018)
18. Choi, H., Moon, G., Lee, K.M.: Pose2mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose. In: European Conference on Computer Vision ECCV, vol. 12352, pp. 769–787 (2020)
19. Bogo, F., Kanazawa, A., Lassner, C., Gehler, P.V., Romero, J., Black, M.J.: Keep it SMPL: automatic estimation of 3d human pose and shape from a single image. In: European Conference on Computer Vision ECCV, vol. 9909, pp. 561–578 (2016)
20. Xiang, D., Joo, H., Sheikh, Y.: Monocular total capture: Posing face, body, and hands in the wild. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, pp. 10965–10974 (2019)
21. Sun, W., Wang, L., Ma, S., Ma, Q.: Estimating 3d body mesh without smpl annotations via alternating successive convex approximation. Comput. Vis. Image Underst. **224**, 103539 (2022)
22. Alldieck, T., Magnor, M.A., Xu, W., Theobalt, C., Pons-Moll, G.: Detailed human avatars from monocular video. In: International Conference on 3DV, pp. 98–109 (2018)
23. Alldieck, T., Magnor, M.A., Bhatnagar, B.L., Theobalt, C., Pons-Moll, G.: Learning to reconstruct people in clothing from a single

RGB camera. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, pp. 1175–1186 (2019)

24. Zhu, H., Zuo, X., Wang, S., Cao, X., Yang, R.: Detailed human shape estimation from a single image by hierarchical mesh deformation. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, pp. 4491–4500 (2019)
25. Pons-Moll, G., Pujades, S., Hu, S., Black, M.J.: Clothcap: seamless 4d clothing capture and retargeting. ACM Trans. Graph. **36**(4), 73:1-73:15 (2017)
26. Ma, Q., Yang, J., Ranjan, A., Pujades, S., Pons-Moll, G., Tang, S., Black, M.J.: Learning to dress 3d people in generative clothing. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, pp. 6468–6477 (2020)
27. Corona, E., Pumarola, A., Alenyà, G., Pons-Moll, G., Moreno-Noguer, F.: Smplicit: Topology-aware generative model for clothed people. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, pp. 11875–11885 (2021)
28. Bhatnagar, B.L., Sminchisescu, C., Theobalt, C., Pons-Moll, G.: Combining implicit function learning and parametric models for 3d human reconstruction. In: European Conference on Computer Vision ECCV, vol. 12347, pp. 311–329 (2020)
29. Saito, S., Yang, J., Ma, Q., Black, M.J.: Scanimate: Weakly supervised learning of skinned clothed avatar networks. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, pp. 2886–2897 (2021)
30. Ma, Q., Saito, S., Yang, J., Tang, S., Black, M.J.: SCALE: modeling clothed humans with a surface codec of articulated local elements. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, pp. 16082–16093 (2021)
31. Gabeur, V., Franco, J., Martin, X., Schmid, C., Rogez, G.: Moulding humans: Non-parametric 3d human shape estimation from single images. In: International Conference on Computer Vision ICCV, pp. 2232–2241 (2019)
32. Tang, S., Tan, F., Cheng, K., Li, Z., Zhu, S., Tan, P.: A neural network for detailed human depth estimation from a single image. In: International Conference on Computer Vision ICCV, pp. 7749–7758 (2019)
33. Natsume, R., Saito, S., Huang, Z., Chen, W., Ma, C., Li, H., Morishima, S.: Siclope: Silhouette-based clothed people. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, pp. 4480–4490 (2019)
34. Wang, T., Liu, M., Zhu, J., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, pp. 8798–8807 (2018)
35. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: European Conference on Computer Vision ECCV, vol. 9912, pp. 483–499 (2016)
36. Maas, A., Hannun, A., Ng, A.: Rectifier nonlinearities improve neural network acoustic models. In: Proceedings of the International Conference on Machine Learning vol. 30, p. 3 (2013)
37. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: AMASS: Archive of motion capture as surface shapes. In: International Conference on Computer Vision ICCV, pp. 5442–5451 (2019)
38. Lorensen, W.E., Cline, H.E.: Marching cubes: a high resolution 3d surface construction algorithm. In: SIGGRAPH, pp. 163–169. ACM (1987)
39. renderpeople: https://www.renderpeople.com
40. Lin, T., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: common objects in context. In: European Conference on Computer Vision ECCV, vol. 8693, pp. 740–755 (2014)
41. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human 3.6m: large scale datasets and predictive methods for 3d human sensing in natural environments. IEEE Trans. Pattern Anal. Mach. Intell. **36**(7), 1325–1339 (2014)
42. Kolotouros, N., Pavlakos, G., Daniilidis, K.: Convolutional mesh regression for single-image human shape reconstruction. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, pp. 4501–4510 (2019)
43. Zhu, H., Zuo, X., Wang, S., Cao, X., Yang, R.: Detailed human shape estimation from a single image by hierarchical mesh deformation. In: IEEE Conference on Computer Vision and Pattern Recognition, (CVPR), pp. 4491–4500 (2019)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Lixiang Lin received his bachelor's degree in Computer Science from Nankai University in 2019. He is currently a Ph.D. candidate in the College of Computer Science and Technology, Zhejiang University. His research interests include machine learning and computer vision.

Jianke Zhu received the master's degree from University of Macau in Electrical and Electronics Engineering, and the Ph.D. degree in computer science and engineering from The Chinese University of Hong Kong, Hong Kong in 2008. He held a post-doctoral position at the BIWI Computer Vision Laboratory, ETH Zurich, Switzerland. He is currently a Professor with the College of Computer Science, Zhejiang University His research interests include computer vision and multimedia information retrieval.