Applied Intelligence (2022) 52:6739–6759
https://doi.org/10.1007/s10489-021-02783-8

Detailed 3D human body reconstruction from multi-view images combining voxel super-resolution and learned implicit representation

Zhongguo Li[1] · Magnus Oskarsson[1] · Anders Heyden[1]

Accepted: 20 August 2021 / Published online: 15 September 2021

Abstract

The task of reconstructing detailed 3D human body models from images is interesting but challenging in computer vision due to the high freedom of human bodies. This work proposes a coarse-to-fine method to reconstruct detailed 3D human body from multi-view images combining Voxel Super-Resolution (VSR) based on learning the implicit representation. Firstly, the coarse 3D models are estimated by learning an Pixel-aligned Implicit Function based on Multi-scale Features (MF-PIFu) which are extracted by multi-stage hourglass networks from the multi-view images. Then, taking the low resolution voxel grids which are generated by the coarse 3D models as input, the VSR is implemented by learning an implicit function through a multi-stage 3D convolutional neural network. Finally, the refined detailed 3D human body models can be produced by VSR which can preserve the details and reduce the false reconstruction of the coarse 3D models. Benefiting from the implicit representation, the training process in our method is memory efficient and the detailed 3D human body produced by our method from multi-view images is the continuous decision boundary with high-resolution geometry. In addition, the coarse-to-fine method based on MF-PIFu and VSR can remove false reconstructions and preserve the appearance details in the final reconstruction, simultaneously. In the experiments, our method quantitatively and qualitatively achieves the competitive 3D human body models from images with various poses and shapes on both the real and synthetic datasets.

Keywords Detailed 3D human body · Implicit representation · Multi-scale features · Multi-view images · Voxel super-resolution

1 Introduction

Recovering detailed 3D human body models from images attracts much attention because of its wide applications in movie industry, animations, and Virtual/Augmented Reality. Although professional capture systems [25, 59] are now able to reconstruct accurate 3D human bodies, these systems remain inconvenient for common users because they are often expensive and difficult to deploy. With the developing of deep learning in 3D vsion, estimating 3D human bodies from common 2D images attracts much attention and has achieved some progress because it is much easier to obtain 2D images for the community. However, current approaches cannot get the 3D models from 2D images with sufficient accuracy and the task is still far from being finished. The goal of the work is to achieve better 3D human body models from multi-view 2D images.

Traditionally, reconstructing 3D human body from RGB images mainly depends on the pre-defined parametric human body models. From simple geometric primitives [51] to data-driven models [5, 36], parametric human body models play important roles in human related research. The main idea of the route is to fit the parametric human body model to some prior information including the body skeleton, 2D joint points and the silhouettes [2, 6, 8]. The route has been used for human motion tracking and 3D pose estimation successfully. However, the 3D human body models estimated by these methods cannot satisfy the requirements of the realism in many applications because the parametric models often do not encode the detailed appearance.

Benefiting from the great success of deep learning in 3D vision, it has achieved some progress to learn to reconstruct 3D human body from images recently. During the past several years, convolutional neural networks (CNN) have

✉ Zhongguo Li
zhongguo.li@math.lth.se

[1] Lund, Sweden

shown impressive performance on 2D/3D human pose estimation [4, 41, 46] and human body segmentation [20, 58]. Therefore, some methods automatically estimated 3D human body model from images by fitting the parametric human body to prior cues like the 2D/3D poses and silhouettes which can be estimated by the CNN [2, 8, 21, 60]. Since the poses and silhouettes comprise sparse information, directly inferring the pose and shape of a parametric human body model from the full image through the CNN becomes another useful route and has achieved impressive performance [26, 30, 31, 44, 45]. However, the 3D human body models obtained by these methods have poor appearance. Recently, many approaches came up with a refining process on the parametric human body to add clothes on the naked 3D human body model [3, 49, 61]. Through refining the parametric human body model, these methods can obtain some details including the clothes and hair on the final 3D model. However, these methods require the parametric human body model has high accuracy on the pose estimation.

Recently, learning to reconstruct 3D models has gained popularity. Explicit volumetric representations are straightforward for learning to infer 3D objects from RGB images [12, 15, 28, 55]. Due to the limitation of memory, these methods can only produce low-resolution 3D objects (e.g. 32^3 or 64^3 number of voxels). Even though some methods reduce the memory footprint through octrees, the final resolutions are sill relatively small (e.g. 256^3) [47]. In addition, these results are always discrete, which results in the missing of many details on the surface. In contrast to explicit representations, implicit function for 3D model representation in deep learning shows impressive performance [10, 11, 39, 43]. Compared to learning the explicit volumetric representation, learning an implicit function to represent 3D shape can be implemented in a memory efficient way, especially for the training process. Another advantage of implicit representation is that the 3D model can be decided by the continuous decision boundary, which is able to produce a high-resolution 3D model. Considering the advantages, there are some methods based on learning implicit function to reconstruct detailed 3D human body from images [22, 48, 49]. However, these methods may still produce some false reconstruction on the final 3D model.

In this paper we propose a novel method to estimate a detailed 3D human body model from multi-view images, through learning an implicit representation. Our method works in a coarse-to-fine manner, and thus, consists of two parts: (1) 3D human body reconstruction from multi-view images through learning pixel-aligned implicit function based on multi-scale features (MF-PIFu), and (2) voxel super-resolution (VSR) from low-resolution voxel grids obtained by MF-PIFu. In both of the two parts, we attempt to learn an implicit function to represent the 3D models. For the MF-PIFu, the structure of multi-stage hourglass networks is designed to produce the multi-scale features and a fully connected neural network predicts the occupancy values of the features to implicitly represent 3D models. Through training the above model, the coarse 3D models can be estimated from multi-view images. Then, low-resolution grids can be generated by voxelizing the coarse models. Taking the low-resolution grids as input, a multi-stage 3D CNN is built to produce multi-scale features and a fully connected neural network is also utilized to predict the occupancy values of the features. The final 3D model is generated by the implicit representation through refining the coarse model by VSR. Our method is summarized in Fig. 1.

Our method differs from previous work in three aspects. Firstly, it is a coarse-to-fine method combining 3D reconstruction from multi-view images by MF-PIFu and VSR into one route to infer 3D human body models. MF-PIFu produces a coarse 3D human body from multi-view images and VSR refines the coarse result to generate a final detailed 3D model. Secondly, the implicit representation for the 3D model is used both in MF-PIFu and VSR, which is memory efficient for training and can produce high resolution geometry through extracting a continuous decision boundary. Finally, the multi-scale features are extracted from multi-view images and low-resolution voxel-grids for coarse reconstruction and refining the models, respectively. The multi-scale features are able to fully encode the local and global spatial information of the pixels in the images and the voxels in the low resolution voxel grids. In order to better represent the method, a list of acronyms and corresponding denotations used in the paper is shown in Table 1.

The paper is organized as follows. The introduction and related work of our method are presented in Sections 1 and 2, respectively. The following Section 3 describes the detailed coarse-to-fine structure of our method and the implementation details including the MF-PIFu and VSR. In Section 4, some quantitative and qualitative experiments are illustrated to evaluate the performance of our method. Finally, the conclusion and future work are stated in Section 5.

2 Related work

The related work on 3D human body reconstruction from images is summarized in this section. There are three parts in the section: (1) Optimization based methods; (2) Parametric human body model based regression, and (3) Non-parametric human body model based regression.

Optimization based methods The classic route to recover 3D human body models from an image is to fit a template

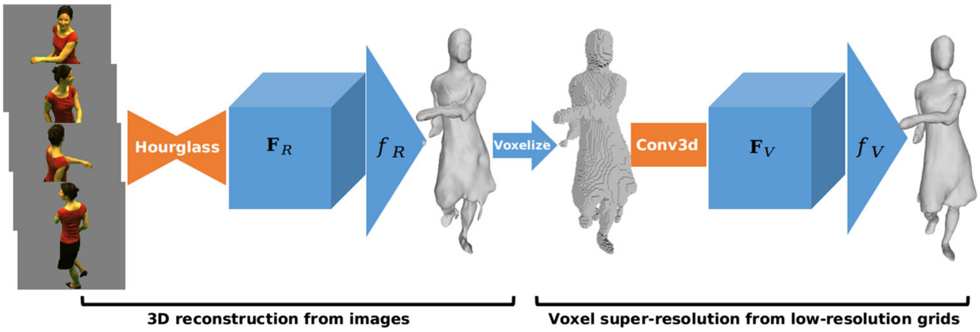

Hourglass
$\mathbf{F}_R$
f_R
Voxelize
Conv3d
$\mathbf{F}_V$
f_V
3D reconstruction from images
Voxel super-resolution from low-resolution grids

Fig. 1 The pipeline of our method. It consists of 3D reconstruction from images and voxel super-resolution from low-resolution grids. The 3D reconstruction from multi-view images is implemented by MF-PIFu and estimates a coarse 3D human body model. After voxelizing coarse model to a low-resolution grid, the voxel super-resolution refines the low-resolution grid to obtain detailed model

such as SCAPE [5] or SMPL [36] to prior cues. SCAPE, which was a data-driven parametric human body model to represent human pose and shape, was learned from 3D human body scans [5]. Some methods fitted SCAPE to the silhouettes and joint points from the images to recover human pose and shape [6, 18, 50]. With the emergence of Kinect, the depth images were also used for fitting the SCAPE [7, 35, 56]. With the success of deep learning on human pose estimation [4, 9, 38, 41], the joint points can be obtained automatically with high accuracy. In [8], an automatic method for 3D human body estimation was proposed through fitting a novel parametric human body model called SMPL [36] to the 2D joint points predicted by DeepCut [46]. Then, more methods used SMPL or prescanning models for human body reconstruction based on 3D joint points, multi-view images, video and silhouettes [2, 19, 21, 33, 60]. These methods tried to build better energy function based on various prior cues and the 3D human body was estimated by optimizing the energy function. Although the optimization based methods were classic, the estimated 3D human body had poor realism.

Parametric human body model based regression Since deep learning has achieved impressive performance on 3D vision tasks [57, 63], it also attracts much attention on 3D human body estimation through regressing the parametric human body model. In the beginning, the shape parameters of SCAPE were regressed from silhouettes to estimate 3D human body model in [13, 14]. In [52], the shape and pose of the SMPL model were regressed through the images and the corresponding SMPL silhouettes. Instead of using silhouettes, the authors proposed to take the whole image as the input of the CNN to regress the pose and shape parameters of the SMPL model through building the loss function about the joint points [26]. Since then, many improved methods were proposed through designing novel network structure or using more constraints on the loss function [27, 29–31, 34, 44, 45]. Pavlakos et al. [45] combined joint points and silhouettes in the loss function to better estimate the shape. There were some other approaches in which various cues were used for building sufficient loss function to train the network including the mesh [31], the texture [44], the multi-view images [34], the optimized SMPL model [30] and the video [27, 29]. In order to model the detailed appearance, some method attempt to refine the regressed SMPL model to obtain the detailed 3D model [1, 3, 23, 32, 42, 53, 61, 62]. In [1, 3, 32], after estimating the pose and shape of SMPL model, the authors used shape from shading and texture translation to add the details to SMPL like face, hairstyle, and clothes with garment wrinkles. In addition, the explicit representation of 3D human body model were also used in detailed reconstruction. BodyNet [53] added the volume loss function to better estimate the pose and shape of SMPL. DeepHuman [61]

Table 1 A list of acronyms and corresponding denotations used in the paper

Acronyms	Denotation
MF-PIFu	Pixel-aligned Implicit Function based on Multi-scale Features
VSR	Voxel Super Resolution
SPIN	SMPL optimization in the loop
SMPL	Skinned Multi-person linear model
P2S	Point-to-surface error
IoU	Intersection over Union

refined the appearance of volumetric SMPL model through transferring the image normal to the volumetric SMPL. In [42], a novel tetrahedral representation for SMPL model was used and the detailed model was obtained by learning the sign distance function of tetrahedral representation. Another recent work also refined the normal and color of image to the estimated SMPL model [23] from single image.

Non-parametric human body model based regression Recently, deep learning also achieved some success on reconstruction of 3D objects from images without relying on any parametric models. Some methods tried to extract coarse 3D information from 2D images and attempted to refine the 3D information through deep neural network such as volume, visual hull and depth images [16, 17, 22, 24, 40]. Jackson et al. [24] reconstructed 3D geometry of humans through training an end-to-end CNN to regress the volumes which were provided in the training dataset. In [17], a coarse model was obtained though Visual Hull from sparse view images and the coarse model was refined by a deep neural network. Natsume et al. [40] generated multi-view silhouettes through deep learning from single image and proposed a deep visual hull to infer the detailed 3D models based on the estimated silhouettes. Huang et al. [22] estimated detailed models by deciding if a spatial point inside or outside of 3D mesh through classifying the features extracted by the CNN. Gabeur et al. [16] estimated the visible and invisible point clouds of the human body from image through deep learning and the full detailed body can be formed by the point clouds. Instead of inferring 3D information from images, some other methods gained popularity to reconstruct general 3D models directly from images with explicit representation such as voxels and point cloud [12, 15, 28, 55]. Due to the limitation of resolution of an explicit representation, implicit representation of 3D models based on deep learning have been used for reconstruction of general objects [10, 11, 28, 39]. Inspired by the idea, some methods only for detailed 3D human body reconstruction also proposed based on learning implicit representation. Saito et al. [48] extracted the pixel-aligned features from images through end-to-end networks. Associating the depth of pixel, the implicit representation can be learned from the features. The method can produce the high-resolution detailed 3D human body including the facial expression, clothes and hair can be estimated from by the above methods. However, there existed many errors on the estimation because only 2D images were used. An improved method called PIFuHD [49] was proposed to reconstruct high-resolution detailed 3D human body from images through introducing image normal to PIFu. The coarse-to-fine methods could obtain more accurate 3D model because more cues were used for the reconstruction.

3 Method

In this section the details of our method are described. The background of implicit function to represent the 3D shape is firstly introduced. Then, we present the 3D human body reconstruction from multi-view images through learning the MF-PIFu. Afterwards, an implicit representation based network for VSR is presented to refine the 3D human body model obtained from the multi-view images. Finally, the implementation details of our method are introduced.

3.1 Learning an implicit function for 3D models

For 3D reconstruction based on deep learning, implicit functions to represent 3D shape is memory efficient for training. Instead of storing all voxels of the volume in an explicit volumetric representation, an implicit function for 3D representation assigns the signed distance or occupancy probability to a spatial point to decide if the point lies inside or outside of the 3D mesh. The estimated 3D mesh can be extracted by a level set surface. In our method, we use occupancy probability as the output of the implicit function. Given a spatial point and a water-tight mesh, the occupancy function is defined as:

$$f(X) := x, X \in \mathbb{R}^3, x \in \{0, 1\}, \tag{1}$$

where X is the 3D point and x is the value of occupancy function for X. The value of x indicates if X lies inside (0) or outside (1) of the mesh. The 3D mesh can be implicitly represented and extracted by the level set of $f(X) = 0.5$.

For 3D reconstruction based on learning implicit representation, the key problem is to learn the occupancy function $f(\cdot)$. More specifically, a deep neural network encodes 3D shape as a vector $\mathbf{v} \in \mathcal{V} \subset \mathbb{R}^m$, and then, the occupancy function takes the vector as input to decide the value of the 3D point, i.e.,

$$f(\mathbf{v}, X) : \mathcal{V} \times \mathbb{R}^3 \mapsto [0, 1]. \tag{2}$$

As long as $f(\cdot)$ can be learned, the continuous occupancy probability field of a 3D model can be predicted and the 3D model can be extracted by the iso-surface of the field through the classic Marching Cubes algorithm.

In PIFu [48], the authors presented a pixel-aligned implicit function for high-resolution 3D human body reconstruction. It is defined as:

$$f(F(\pi(X)), z(X)) : \mathcal{V} \times \mathbb{R} \mapsto [0, 1], \tag{3}$$

where $F(\cdot)$ is the feature grids of CNN, $\pi(X)$ is the projection of X on the image plane by π and $z(X)$ is the depth of X. PIFu showed impressive performance on detailed reconstruction of human bodies for fashion poses, for instance, walking and standing. However, the features extracted by multi-stage networks from input images have the same scale, which may result in the missing of some details. In addition, for some complicated poses, only using 2D images may result in false reconstructions. Aiming at the above two drawbacks, we propose two improvements. On one hand, the multi-scale features are extracted in both 3D reconstruction from images and voxel super-resolution. On the other hand, the voxel super-resolution refines the coarse 3D models to reduce false reconstructions.

The outline of our method is shown as Fig. 1. It has two parts: (1) 3D reconstruction from images by MF-PIFu; and (2) refining 3D models by VSR. The details of the two parts are presented in the following sections.

3.2 MF-PIFu

The method for 3D reconstruction from multi-view images is inspired by PIFu [48]. The difference is that multi-scale features are extracted from multi-view images through multi-stage hourglass networks. Therefore, we call our method as MF-PIFu and the architecture of MF-PIFu is shown in Fig. 2.

Given images with N views $I_i, i = 1, ..., N$, multi-stage hourglass networks which are denoted as $g_R(\cdot)$ encode the images as feature grids $\mathbf{F}_R^{(j)}$, $j = 1, ..., M$ where M is the number of hourglass networks. Then, for the i-th image I_i, its multi-scale feature grids are defined as:

$$g_R(I_i) := \mathbf{F}_R^{(i,1)}, ..., \mathbf{F}_R^{(i,M)}, \tag{4}$$

where the feature grids $\mathbf{F}_R^{(i,1)}, ..., \mathbf{F}_R^{(i,M)}$ have different scales and the j-th grid $\mathbf{F}_R^{(i,j)}$ belongs to feature space $\mathcal{F}_j^{C\times K\times K}$. C is the depth of feature grid and K is the width and height of the feature grid. In our method, C is kept constant (e.g. 256) and K deceases as 2^{j-1} for the j-th hourglass network. Before the $\mathbf{F}_R^{(i,j-1)}$ is fed into the j-th hourglass network, we use a max-pooling layer to downsample $\mathbf{F}_R^{(i,j-1)}$. Through this max-pooling layer, the multi-scale feature grids can be generated by the multi-stage hourglass networks. For the pixel x in the image I_i, the feature vector in $\mathbf{F}_R^{(i,j)}$ can be obtained at the corresponding location through interpolation, which is denoted as $\mathbf{F}_R^{(i,j)}(x) \in \mathcal{F}_j^C$.

After getting the multi-scale features, the multi-scale features need to be queried to predict the occupancy value. The prediction is implemented by a fully connected neural network which is defined as $f_R(\cdot)$. Similar to PIFu, not only the features are used for prediction, but also the depth of the corresponding pixel is also used. The multi-scale features and the depth form new feature vector for prediction. For the pixel x in the image I_i, we define the new feature vector as $\mathbf{F}_R^{(i)}(x) = \{\mathbf{F}_R^{(i,1)}(x), ..., \mathbf{F}_R^{(i,M)}(x), z(x)\} \in \mathcal{F}_1^C \times ... \times \mathcal{F}_M^C \times \mathbb{R}$. The fully connected neural network takes into the feature vector to predict the occupancy value of x:

$$f_R(\mathbf{F}_R^{(i)}(x)) : \mathcal{F}_1^C \times ... \times \mathcal{F}_M^C \times \mathbb{R} \mapsto [0, 1]. \tag{5}$$

In contrast to PIFu, we form the features of each stage and the depth as a new feature vector. This new feature

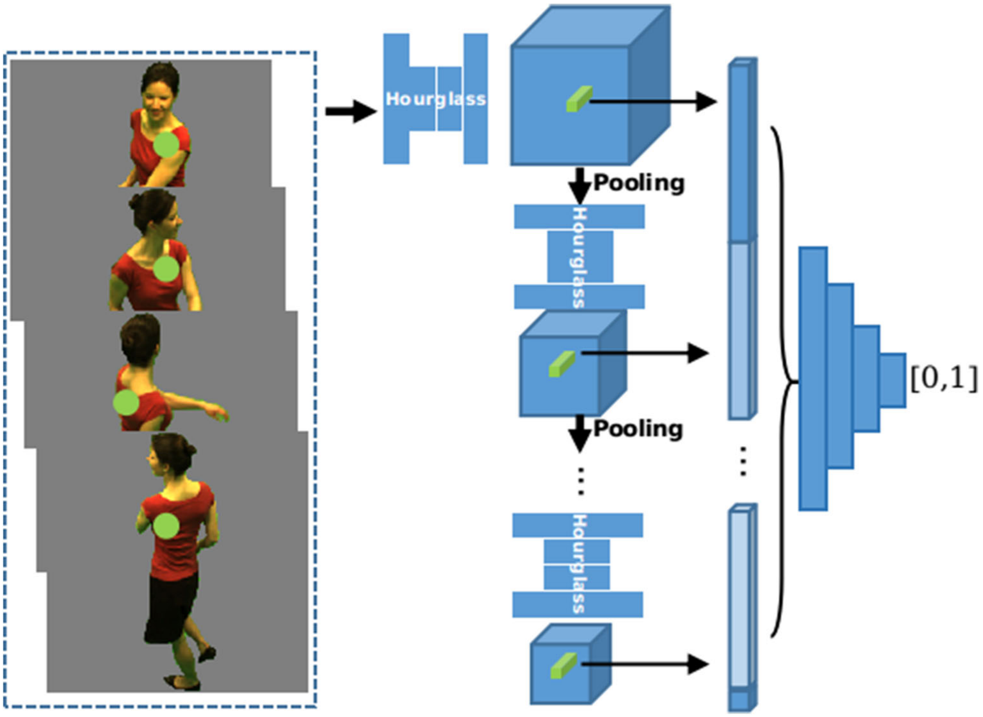


Fig. 2 The structure of MF-PIFu to learn the implicit representation of 3D human body model. Multi-stage hourglass networks are used for multi-scale feature extraction and a fully connected neural network predicts the occupancy value of the feature

encodes both the local and global information of the pixels. The feature grids at the early stage encode more local information, while the feature grids at the last stage represent the global information. Associating the depth information, the new features encode more information than the features used in PIFu, and thus, it is more reliable to predict the occupancy value.

To train $g_R(\cdot)$ and $f_R(\cdot)$ from multi-view images $I_i, i = 1, ..., N$, the pairs $\{I_i, \mathcal{S}\}$ are required in which $\mathcal{S}$ is the corresponding ground truth of 3D model for the multi-view images I_i. As shown in Fig. 3, 3D spatial points $X_i, i = 1, ..., K$ are sampled from the 3D model $\mathcal{S}$ and random displacements with normal distribution $\mathcal{N}(0, \sigma)$ on the points are added. This means that the points to be queried are $\hat{X}_i = X_i + n_i$ where $n_i \sim \mathcal{N}(0, \sigma)$. The binary occupancy values of the points $o(\hat{X}_i)$ can be obtained according to the location of $\hat{X}_i$. If $\hat{X}_i$ lies in $\mathcal{S}$, $o(\hat{X}_i) = 0$ (the red points in Fig. 3). Otherwise, $o(\hat{X}_i)$ is 1 (the green points in Fig. 3). The points $\hat{X}_i$ are projected onto the multi-view images through the given camera parameters. The corresponding pixel of point $\hat{X}_j$ on the i-th image is $x_{ij} = \pi_i(\hat{X}_j)$. Then, the loss function for the pair $\{I_i, \mathcal{S}\}$ can be defined as:

$$L_R = \sum_{i=1}^{N} \sum_{j=1}^{K} \| f_R(\mathbf{F}_R^{(i)}(x_{ij})) - o(X_j) \|. \tag{6}$$

In the above loss function, $\mathbf{F}_R^{(i)}(x_{ij})$ is the multi-scale features of pixel x_{ij} which is the projection of 3D point $\hat{X}_j$ on the i-th view image. This loss function is defined based on the multi-view images jointly, which can predict the occupancy values more accurately. Through minimizing the loss function, $g_R(\cdot)$ and $f_R(\cdot)$ can be trained end-to-end.

3.3 Voxel super-resolution

The 3D models recovered by MF-PIFu are still coarse because MF-PIFu only relies on 2D images. We observe two problems in the estimated 3D models by MF-PIFu. The first one is that the surface of the 3D model is not smooth and the second one is that some extra unnecessary parts are reconstructed on the models due to the false classification of some voxels. In order to overcome the problems, voxel super-resolution (VSR) is learned to refine the coarse 3D models of MF-PIFu. As shown in Fig. 4, our VSR method also uses a multi-scale structure for feature extraction and implicit representation for the 3D model. In contrast to MF-PIFu which uses images as input, the input of VSR is a low resolution voxel grid which is produced by the voxelization of the 3D model of MF-PIFu.

Suppose the 3D model estimated by MF-PIFu is $\hat{\mathcal{S}}$ which is stored as the voxel positions. The voxelization of $\hat{\mathcal{S}}$ can produce a low resolution grid as $\mathcal{V} \in \mathbb{R}^{N \times N \times N}$ (e.g. N=128). Then, as shown in Fig. 4, 3D convolution kernels are utilized to extract 3D feature grids from $\mathcal{V}$. 3D CNN with n convolutional layers is used to generate the multi-scale feature grids $\mathbf{F}_V^{(1)}, ..., \mathbf{F}_V^{(n)}$. The resolution of the k-th feature grid is $N/(2^{k-1})$, i.e., $\mathbf{F}_V^{(k)} \in \mathcal{F}_k^{K \times K \times K}$ where $K = N/(2^{k-1})$. The resolution of the feature grids decreases with the depth of the network. We denote the 3D CNN for VSR as $g_V(\cdot)$ and the multi-scale features can be generated as:

$$g_V(\mathcal{V}) := \mathbf{F}_V^{(1)}, ..., \mathbf{F}_V^{(n)}. \tag{7}$$

The feature grid at the early stage encodes more local information such as the shape details, while the feature grid at the late stage captures the global information of the voxel grid because of the large receptive fields at the late stage.

For a voxel $\mathbf{v} \in \mathcal{V}$, its corresponding multi-scale feature is formed by the features from $\mathbf{F}_V^{(1)}, ..., \mathbf{F}_V^{(n)}$. Since the feature grid is discrete, the feature of voxel $\mathbf{v}$ in $\mathbf{F}_V^{(k)}$ is extracted by trilinear interpolation and is denoted as $\mathbf{F}_V^{(k)}(\mathbf{v})$. The multi-scale feature for the voxel $\mathbf{v}$ is

$$\mathbf{F}_V(\mathbf{v}) = \{\mathbf{F}_V^{(1)}(\mathbf{v}), ..., \mathbf{F}_V^{(n)}(\mathbf{v})\}, \tag{8}$$

where $\mathbf{F}_V(\mathbf{v}) \in \mathcal{F}_1 \times ... \times \mathcal{F}_n$. After obtaining the multi-scale feature for a voxel $\mathbf{v}$, we also use a fully connected network to classify the multi-scale feature and and we denote it

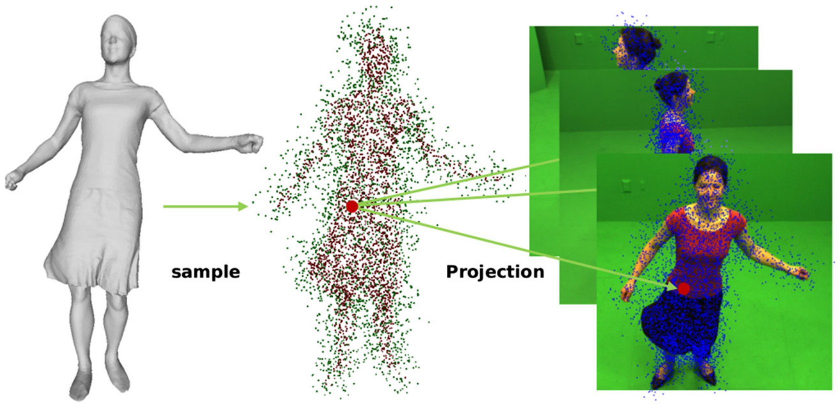


Fig. 3 Sampling 3D points from 3D model and projecting the points to multi-view images

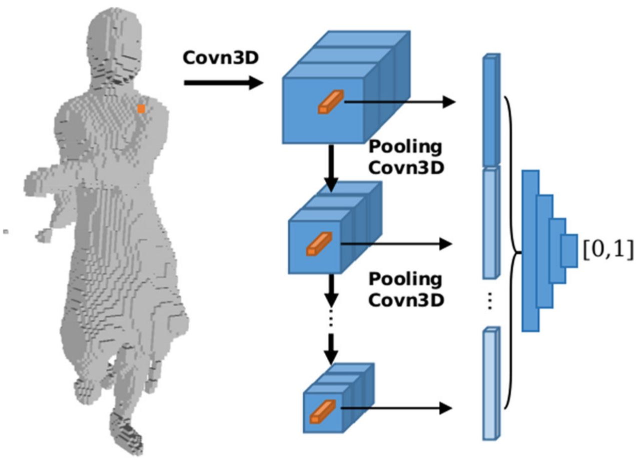


Fig. 4 The structure of VSR based on learning implicit representation. 3D CNN is used for extracting the multi-scale features from low-resolution grid. A fully connected neural network is used for predicting occupancy value of features

$f_V(\cdot)$. The fully connected network predicts the occupancy value of the multi-scale feature of $\mathbf{F}_V(\mathbf{v})$:,

$$f_V(\mathbf{F}_v(\mathbf{v})) : \mathcal{F}_1 ... \times \mathcal{F}_n \mapsto \in [0, 1] \tag{9}$$

This fully connected neural network classifies the voxel based on the multi-scale feature if the corresponding point lies inside or outside of 3D mesh. The implicit representation enables to produce a continuous surface. Besides, since multi-scale feature encodes both the local and global information, the 3D model after super-resolution can keep the global shape and preserve details of the shape.

In order to train the $g_V(\cdot)$ and $f_V(\cdot)$ from low-resolution voxel grids $\mathcal{V}$, the 3D model $\hat{\mathcal{S}}$ estimated by MF-PIFu and its ground truth $\mathcal{S}$ are given as a pair $\{\hat{\mathcal{S}}, \mathcal{S}\}$. The input low-resolution voxel grids are generated by voxelizing $\hat{\mathcal{S}}$. Instead of sampling points from $\mathcal{S}$, we sample N points $\mathbf{v}_i, i, ..., N$ on the surface of $\hat{\mathcal{S}}$ and add random displacements with normal distribution $n_i \sim N(0, \sigma)$ to these points, i.e., $\hat{\mathbf{v}}_i = \mathbf{v}_i + n_i$. Here the same strategy for generating the 3D points and labels are used as [11], i.e., 50% points $\mathbf{v}_i$ are added random displacements with small $\sigma_{\min}$ and the other 50% points $\mathbf{v}_i$ are added random displacements with large $\sigma_{\max}$. During the voxelization, the grid coordinates of the points $\hat{\mathbf{v}}_i$ in the low-resolution voxel grids $\mathcal{V}$ can be indexed and it is denoted as $\rho(\hat{\mathbf{v}}_i)$. One example of sampling points and voxelization to a 128^3 grid is shown in Fig. 5. According to whether the point lies inside or outside of the ground truth 3D model $\mathcal{S}$, the binary occupancy value of the points $\hat{\mathbf{v}}_i$ can also be obtained as $o(\hat{\mathbf{v}}_i)$. This is possible because the estimated 3D model by MF-PIFu has been close to the ground truth. Through sampling the points on the estimated 3D model, the occupancy values of the points are reliable to do the VSR.

After getting the occupancy value of the points, the loss function for training the model of VSR can be defined as:

$$\begin{aligned} L_{VSR} &= \sum_{i=1}^{N} \mathcal{L}\left(f_V(g_V(\rho(\hat{\mathbf{v}}_i))), o(\hat{\mathbf{v}}_i)\right) \\ &= \sum_{i=1}^{N} \mathcal{L}(f_V(\mathbf{F}_V(\rho(\hat{\mathbf{v}}_i))), o(\hat{\mathbf{v}}_i)). \end{aligned} \tag{10}$$

In the loss function, multi-scale features are used, and thus, the local and global information of the low-resolution voxel gird are encoded, which can preserve the details and the global shape simultaneously. We use standard cross-entropy loss function $\mathcal{L}(\cdot, \cdot)$ to measure the loss between the prediction and ground truth. Through minimizing the loss function L_{VSR}, the multi-stage 3D convolutional neural networks and the fully connected network are trained.

3.4 Implementation details

As shown in Fig. 1, our model is a coarse-to-fine architecture in which MF-PIFu reconstructs coarse 3D models from multi-view image and VSR refines the coarse models to produce models with high accuracy. In this section the implementation details about the network structure, training and testing of our method are presented.

Network structure of MF-PIFu We use four stages of hourglass networks to generate multi-scale features and four layers in the fully connected neural network for prediction of occupancy value. For the extraction of multi-scale features, the input of the networks is the multi-view images (e.g. four views in the most of our experiments) which have removed backgrounds and are cropped to 256 × 256. The hourglass network consists of two convolutional layers and two deconvolutional layers to generate pixel-aligned feature maps. Max pooling is used for downsampling the feature maps. The output feature grids of each hourglass network has the size of 256 × 128 × 128, 256 × 64 × 64, 256 × 32 × 32, and 256 × 16 × 16. The fully connected network has four convolutional layers and the number of neurons in each layer is (1024, 512, 128, 1). The input feature of the fully connected layer has size 1025 because the multi-scale features also consider the depth of queried pixel.

Training for MF-PIFu During the training, the batch size of input images is 4 and the model is trained for 12 epochs. In addition, 10,000 points are sampled from the ground truth of 3D mesh and they are added normally random noise with $\sigma = 5\ cm$. These points are used for prediction of the occupancy value to build the loss function. The Mean Square Error (MSE) is used for building the loss function. The RMSProp algorithm with initial learning rate 0.001

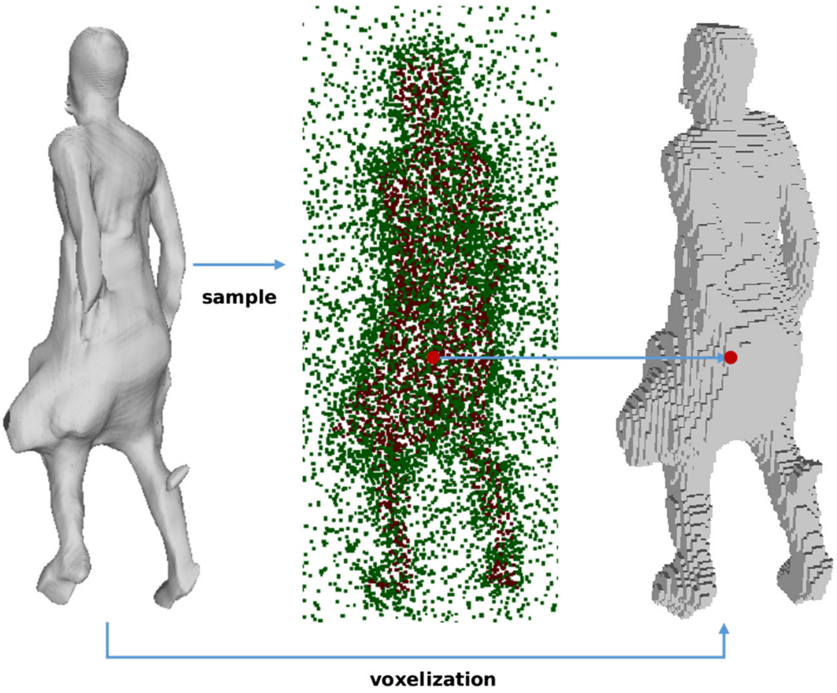


Fig. 5 Sampling 3D points from 3D model estimated by MF-PIFu and the voxelization of the 3D model estimated by MF-PIFu (The resolution is 128^3). The 3D points can be indexed by the grid coordinates in the low-resolution grids

is used for updating the weights of the networks and the learning rate decreases by a factor of 0.1 after 10 epochs. It takes about 7 hours for training on our dataset.

Network structure of VSR The architecture for VSR has the multi-stage 3D convolutional layers for generating multi-scale features from low resolution voxel grids and the fully connected neural network to predict the occupancy value of the multi-scale features. The input of the 3D CNN is the low resolution voxel grids which have the size 128^3. The 3D CNN has 5 convolutional layers and the max pooling is used for downsampling the feature maps. The output feature grid of each convolution block has size of $16 \times (128 \times 128 \times 128)$, $32 \times (64 \times 64 \times 64)$, $64 \times (32 \times 32 \times 32)$, $128 \times (16 \times 16 \times 16)$, $128 \times (8 \times 8 \times 8)$. Therefore, the input feature vector of the fully connected nerual network has 368 elements. The fully connected neural network for predicting the occupancy value consists of four convolutional layers and the number of neurons in each layer is (256,256,256,1).

Training for VSR The low-resolution voxel grids for training the VSR is generated by the coarse 3D models estimated by MF-PIFu through voxelization. The input low-resolution voxel grids have resolution 128^3. We sample 10,000 points from the coarse 3D models, in which 50% of the points are added normal distribution displacements with $\sigma_{\max} = 15\ cm$ and the other 50% of the points are added normal distribution displacements with $\sigma_{\min} = 5\ cm$. The standard cross-entropy loss is used as the loss function. The batch size of input voxel grids is 4 and the network is trained for 30 epochs. The Adam optimizer with learning rate 0.0001 is used for updating the weights of the networks. This will take about 12 hours for training on our dataset.

Testing During the testing process, multi-view images are fed into the trained model of MF-PIFu to generate occupancy predictions for a volume. Then, the predicted 3D human bodies are extracted by an iso-surface through marching cubes from the volume. After voxelizing the predicted 3D model to low-resolution with 128^3, the low-resolution voxel grid is fed into the trained model of VSR to refine the occupancy predictions of the volume. Through use of the march cubes again, the final 3D human body model is extracted from the iso-surface of the volume. Therefore, this process is an image-based coarse-to-fine 3D human body reconstruction method. The MF-PIFu produces the coarse 3D models and the VSR can refine the coarse results through learning another implicit function. After the VSR, the false reconstructed parts can be removed and the details of the appearance can be preserved.

4 Experimental results

In this section some experiments are presented to evaluate our method. We firstly introduce the datasets and metrics for training and testing. Then, several previous methods are used for comparison on the quantitative and qualitative results. Finally, we discuss several factors which may affect the performance of our methods.

4.1 Datasets and metrics

Datasets To train and test our method, two datasets are used in the experiments: Articulated dataset [54] and CAPE dataset [37]. Articulated dataset is captured by 8 cameras and it contains 10 indoor scenarios. Two male subjects have four scenarios, respectively, and one female subject performs two scenarios. For each scenario, RGB images, silhouettes, camera parameters as well as 3D meshes are given. Totally, there are 2000 frames with eight-view images and 3D meshes. We split the dataset as 80% frames (1600) for training and 20% frames (400) for testing. The CAPE dataset is a 3D dynamic dataset of clothed humans generated by learning the clothing deformation from the SMPL body model. There are 15 generative clothed SMPL models with various poses. Since it has a large number of frames, we extract a small dataset from the original CAPE dataset. For each action of each subject, we take the 80-th, 85-th, 90-th, 95-th, and 100-th frames if the action has more than 100 frames. Totally, the small CAPE dataset has 2910 frames with 3D meshes. Since the dataset only provides 3D meshes, we render each mesh to four-view images from front, left, back and right side. The small CPAE dataset is split as 80% for training and 20% for testing in our experiments. Figure 6 gives an example of four-view images and 3D mesh from the small CAPE dataset.

Metrics In order to quantitatively evaluate our method, three metrices are used to measure the estimated 3D models: Euclidean distance from points on the estimated 3D models to surface of ground truth 3D mesh (P2S), Chamfer-L_2 and intersection over union between estimated 3D model and ground truth 3D model (IoU). For P2S and Chamfer-L_2, the lower value means the estimated 3D model is more accurate and complete. For IoU, the higher value means the estimated 3D model better match the ground truth. The detailed definition can be referred to [11].

4.2 The results of the two steps

In order to demonstrate the performance of MF-PIFu and VSR, the results of the two steps are evaluated on the two datasets. Figure 6 gives the examples of the CAPE and Articulated dataset in the first and second rows, respectively. From left to right columns, the figure shows (a) original multi-view images, (b) the ground truth of 3D mesh from two views, (c) the corresponding estimated 3D meshes by the MF-PIFu and (d) the final results of VSR. We can see that the estimated 3D models by MF-PIFu are almost the same as the ground truth. However, there are still some false reconstruction and the details of appearance are not fully recovered, which can be seen from the two examples in Fig. 6c. For instance, the arms of the 3D model from the CAPE dataset are not fully reconstructed by MF-PIFu and there are some extra reconstructed parts around the legs of the 3D models for the Articulated dataset. Figure 6d shows the refined results by VSR, which illustrates that those extra reconstruction in the estimated 3D models of MF-PIFu are removed and the details of the appearance are preserved, especially for arms of the 3D model for the CAPE example and the neck of the 3D model for the Articulated. This figure demonstrates that MF-PIFu can produce the coarse 3D models from multi-view images and VSR can generate better results through refining the coarse 3D models.

The quantitative results of the two steps on the two datasets are also shown in Table 2. The results of P2S, Chamfer-L_2 and IoU of the coarse 3D models by MF-PIFu and the refined 3D models by VSR are given in this table. The bold numbers in the table show that the P2S and Chamfer-L_2 of the VSR are smaller and the corresponding IoU is higher on both the two datasets. For the CAPE dataset, the P2S and Chamfer-L_2 after VSR decrease from 0.9428 cm to 0.4954 cm and from 0.0196 cm to 0.0062 cm, respectively. The IoU after VSR increases from 78.29% to 84.40%. For the Articulated dataset, the P2S and Chamfer-L_2 after VSR reduce from 0.7332 cm to 0.3754 cm and from 0.0194 cm to 0.0032 cm, respectively. The IoU after VSR increases from 84.29% to 90.51%. Therefore, the refined 3D models on the two datasets are more accurate and complete than the coarse 3D models. The VSR is useful to refine the models and obtains better 3D models. The conclusion of this table is consistent with Fig. 6.

4.3 Qualitative results

Our method is qualitatively compared with several previous approaches for 3D human body reconstruction from images including PIFuHD [49], SPIN [30], DeepHuman [61] and PIFu [48]. For the PIFuHD, SPIN and DeepHuman, the trained models provided by the authors are used to estimate 3D models for our test datasets. For PIFu, we trained and tested it on the same training dataset as our method from four-view images. SPIN estimated the pose and shape parameters of SMPL model through collaborating regression and optimization to get the naked 3D models. DeepHuman used encoder-decoder on the volume of deformed SMPL model and used normal image to refine the SMPL model to get detailed appearance. For PIFuHD, the estimated normal images of the front and back side were used in the training to learn pixel-aligned implicit function, which is able to produce 3D models with high resolution. Due to lacking the training code of PIFuHD, the testing code were used to produce the 3D models based single image from our datasets. In Fig. 7, the original images and the ground truth of 3D models are given in the first and second row. The results of PIFuHD [49],

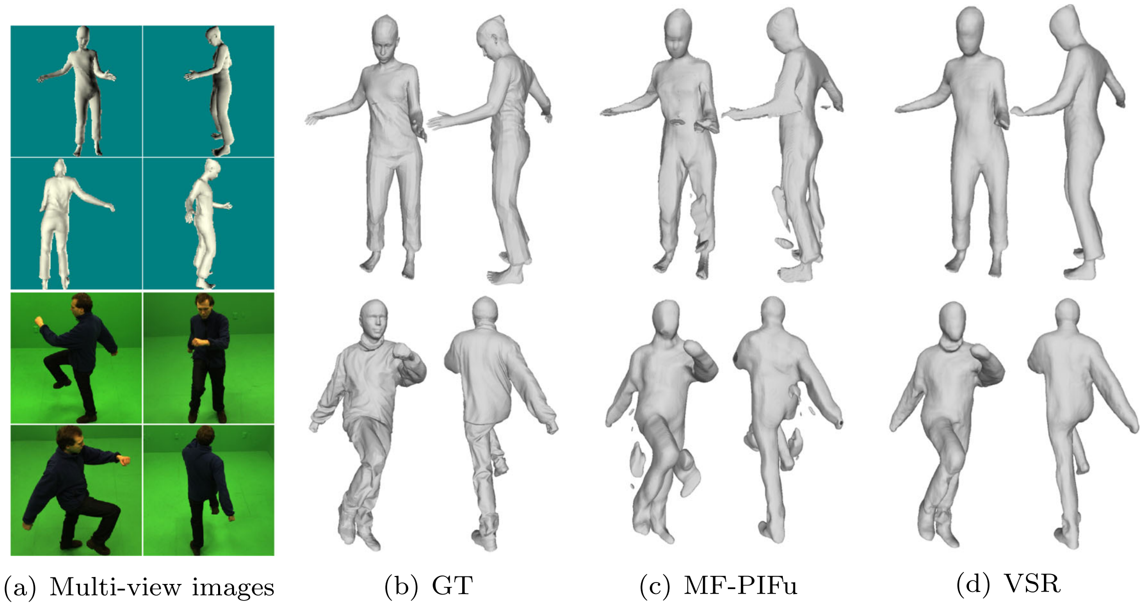


Fig. 6 The 3D models estimated by MF-PIFu and VSR. From the left to right column: The original images (**a**), the ground truth of 3D model from two views (b), the estimated 3D models of MF-PIFu (**c**), and the final 3D models after VSR (**d**)

SPIN [30], DeepHuman [61], PIFu [48] and our method are demonstrated from the third row to the last row. It shows the 3D models estimated by our method have better shape details and less false reconstruction. Since SPIN and DeepHuman rely on the SMPL model, they cannot handle the detailed appearance like clothes and wrinkles on the 3D models. Although DeepHuman attempts to recover the clothes on the 3D model, the results are not satisfying because the trained model of DeepHuman is based on a different dataset. Note that the results of PIFuHD are not so good as the original paper because the normal images of these images are not well estimated and training code is not given. The results of PIFu are better than SPIN and DeepHuman because of learning an implicit representation, while PIFu is better than PIFuHD because we use our dataset to train it. Overall, our method can recover the 3D human body models from multi-view images with plausible pose and surface quality.

In Fig. 8, the P2S between the estimated 3D models by different methods and the ground truth in Fig. 7 are visualized by Meshlab. In Meshlab, the P2S is computed through the Hausdorff Distance and the distances are shown by the heatmaps. The range is from 0 to 10 *cm* to map the color from blue to red for all of the models. The red parts stand for high errors and the blue parts mean small distance. The figure clearly shows that the estimated 3D human bodies of our method have higher accuracy than the other four previous methods.

Table 2 The quantitative results of the CAPE and Articulated datasets by the two steps of our method

		P2S ↓	Chamfer-L_2 ↓	IoU ↑
CAPE	MF-PIFu	0.9482	0.0196	0.7829
	VSR	**0.4954**	**0.0062**	**0.8440**
Articulated	MF-PIFu	0.7332	0.0194	0.8484
	VSR	**0.3754**	**0.0032**	**0.9051**

4.4 Quantitative results

In addition to the qualitative comparison, quantitative results are compared through computing the P2S, Chamfer-L_2 and IoU of different methods on the testing datasets of CAPE and Articulated in Table 3 and Table 4. Note that these metrics are computed after normalizing the 3D models estimated by different methods. It shows that the results of PIFuHD are the worst and our method has the best performance. Although PIFuHD, SPIN and DeepHuman use the trained model, PIFuHD does not estimate camera parameters, which leads to the estimated 3D models have different coordinates with the ground truth and the results of the metrics are bad. SPIN [30] and DeepHuman [61] have similar performance, but the results are still not good. Comparing to PIFuHD, SPIN and DeepHuman, the results of PIFu are better because PIFu are retrained and represents the 3D model through learning implicit function. Our method achieves the best performance among these methods as shown by the bold numbers in the tables because VSR can refine the coarse results of MF-PIFu. The P2S and Chamfer-L_2 are the smallest in our method, which means that the results of our method are more accurate. The IoU of

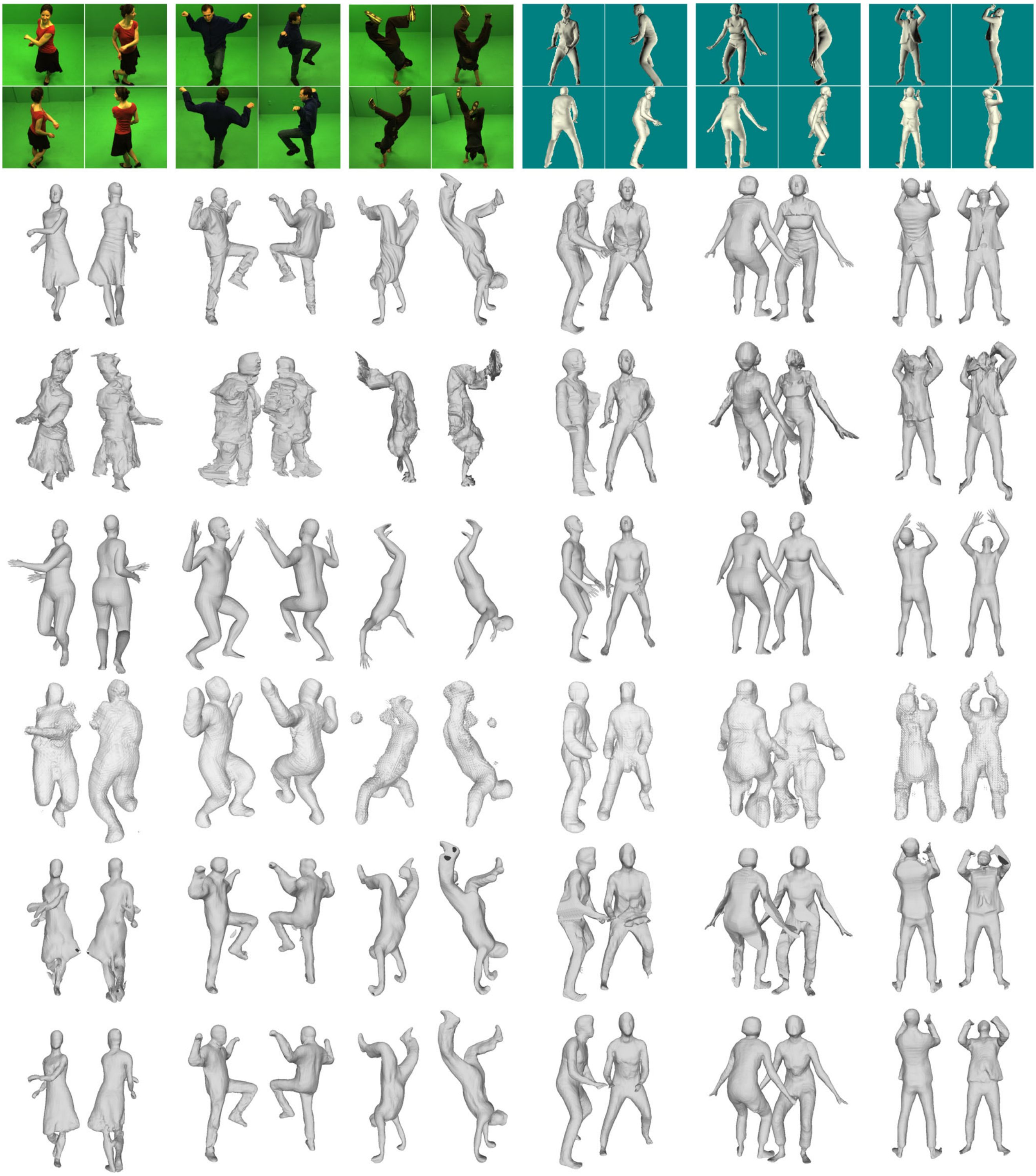


Fig. 7 The comparison between our method and several previous methods on the two datasets. From top to down rows: the original images, the ground truth of 3D models from two views, the estimated 3D models of PIFuHD [49], SPIN [30], DeepHuman [61], PIFu [48] and our method

our method is the highest, which means that the estimated 3D models are more complete. The two tables demonstrate that our method had good performance on both synthetic and real datasets.

In order to clearly show the metric on the testing datasets, the P2S of each sample in the two testing data of the CAPE and Articulated dataset is shown in Fig. 9. There are 582 samples in the testing dataset of CAPE

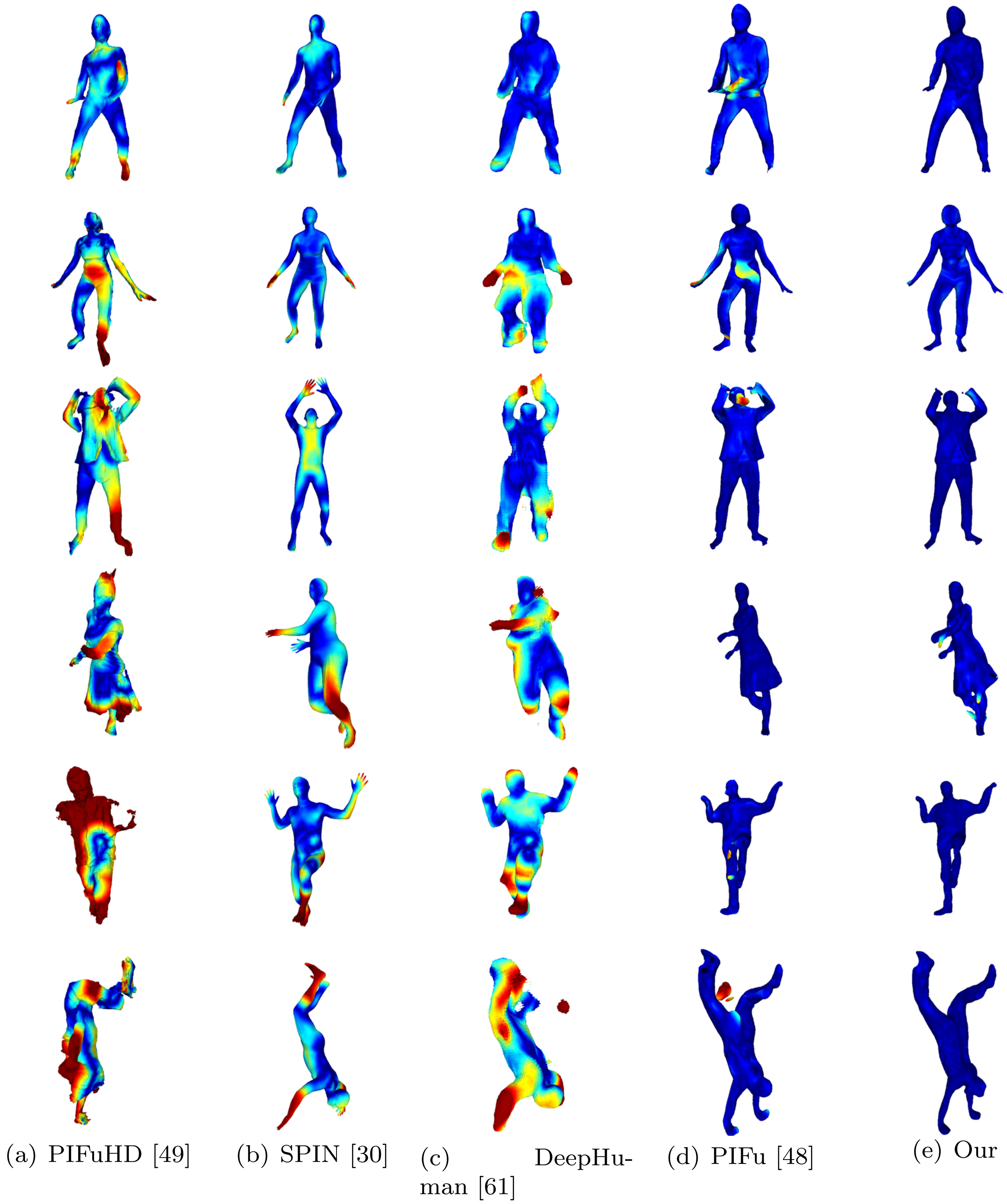


Fig. 8 Visualization of the P2S between the estimated 3D models and the ground truth for different methods in Fig. 7. The distance are represented by the heatmaps in Meshlab and mapped to the estimated 3D models

and 400 samples in the testing dataset of Articulated, respectively. Our method (the red line) has the lowest errors on the two datasets comparing to the other methods. Besides, for the testing samples, our method is more stable and robust because the red lines do not have serious fluctuation.

4.5 Discussion on the PIFu

As shown above, PIFu [48] is a similar approach which also learns an implicit representation for 3D model from images. Therefore, we discuss more about the performance of PIFu in this section. The results of PIFu, MF-PIFu,

Table 3 The quantitative results of PIFuHD [49], SPIN [30], DeepHuman [61], PIFu [48] and our method on the testing dataset of the CAPE

Method	P2S ↓	Chamfer-L_2 ↓	IoU ↑
PIFuHD [49]	5.7136	0.5650	0.2469
SPIN [30]	2.2134	0.1271	0.4044
DeepHuman [61]	3.4028	0.1850	0.3861
PIFu [48]	1.0330	0.0212	0.7571
Ours	**0.4954**	**0.0062**	**0.8440**

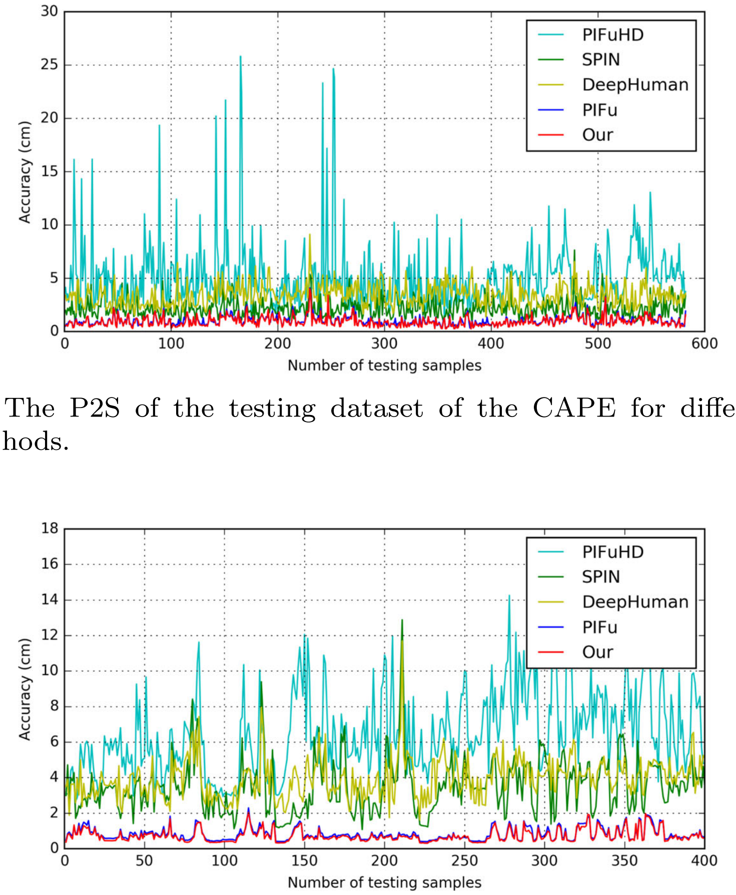
Our method achieves better performance

Table 4 The quantitative results of PIFuHD [49], SPIN [30], DeepHuman [61], PIFu [48] and our method on the testing dataset of the Articulated

Method	P2S ↓	Chamfer-L_2 ↓	IoU ↑
PIFuHD [49]	6.5592	0.6497	0.2056
SPIN [30]	3.5206	0.2679	0.3506
DeepHuman [61]	3.9448	0.2675	0.3742
PIFu [48]	0.8194	0.0210	0.8255
Ours	**0.3754**	**0.0032**	**0.9051**

Our method achieves better performance

Fig. 9 The P2S of each sample in the testing data of the two datasets for different methods. The y axis stands for the accuracy of P2S. The x axis is the number of samples in the testing data



(a) The P2S of the testing dataset of the CAPE for different methods.

(b) The P2S of the testing dataset of the Articulated for different methods.

Table 5 The qualitative results of PIFu, MF-PIFu, PIFu+VSR and our method

View	P2S ↓	Chamfer-L_2 ↓	IoU ↑
PIFu	0.8194	0.0210	0.8255
MF-PIFu	0.7332	0.0194	0.8484
PIFu+VSR	0.4322	0.0041	0.8865
Our	**0.3754**	**0.0032**	**0.9051**

PIFu+VSR and our method are evaluated to demonstrate the advantage of MF-PIFu and our method on the Articulated dataset. Table 5 gives the quantitative results of PIFu, MF-PIFu, PIFu+VSR and our method on the testing dataset of the Articulated. PIFu+VSR means that PIFu is trained by the same Articulated dataset as MF-PIFu, and the testing results of PIFu is refined by the VSR which is trained by the low-resolution voxel grids obtained by MF-PIFu. This table shows that MF-PIFu achieves better results than PIFu and the VSR can refine the coarse models obtained by PIFu and MF-PIFu. Our method combines the MF-PIFu and VSR, and thus, our method achieves the best performance on the dataset as shown by the bold numberd in the table. Figure 10 gives the the P2S of the four cases on the testing dataset of the Articulated. It shows that the accuracy of our method on most samples is the highest. For the MF-PIFu, it has smaller P2S on the most samples than the original PIFu, which provides more reliable inputs for the voxel super-resolution. Therefore, our method combining MF-PIFu and VSR achieves the smallest P2S on most samples. This is consistent with Table 5.

The qualitative examples from the Articulated dataset are shown in Fig. 11. From the figure, it is clearly shown that the results of PIFu, MF-PIFu and PIFu+VSR have some false reconstruction, especially for the first example. The 3D models estimated by our method are the best because the false reconstruction is removed and the surface quality is improved by VSR, which can be demonstrated by the areas indicated by the red circles. The visualization of the errors on the 3D models is also given in the figure, which clearly shows that the 3D models of our method have the smallest distance to the ground truth among the four cases.

4.6 Spatial sampling

Spatial sampling is used in both MF-PIFu and VSR to generate the ground truth of the occupancy value of spatial 3D points. It is an important factor in the sharpness of the final 3D model. In the two parts of our method, we use the same sampling strategy. Firstly, the points are uniformly sampled from the surface of the 3D model. Then, the random displacements with normal distribution $\mathcal{N}(0, \sigma)$ are added to the points. The σ defines the distance of the points to the surface. The larger σ makes the points further from the 3D mesh. For the MF-PIFu, we choose $\sigma = 5\ cm$ for the random displacements because the paper of PIFu [48] has demonstrated that $\sigma = 5\ cm$ can achieve the best performance for the 3D reconstruction from images. Here we evaluate the effects of σ on the VSR on the Articulated dataset. As shown in the implementation details, the 3D points are added random displacements with large $\sigma_{\max}$ and small $\sigma_{\min}$ during training the VSR. In order to discuss the effect of $\sigma_{\max}$ and $\sigma_{\min}$, five pairs of $(\sigma_{\max}, \sigma_{\min})$ are chosen and the corresponding performance under the five cases is compared. Table 6 shows the quantitative values of the P2S, Chamfer-L_2 and IoU for different $(\sigma_{\max}, \sigma_{\min})$ on the testing dataset of the Articulated and the best results are shown as bold numbers. Figure 12 shows the mean P2S of different $\sigma_{\max}$ for the testing dataset of the Articulated. The table and the figure demonstrate that the performance is almost the same for

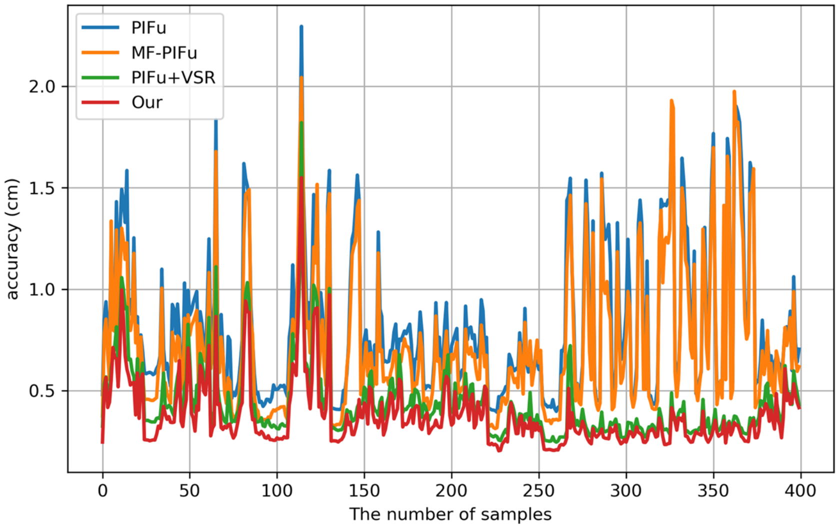


Fig. 10 The P2S of each sample in the testing data of the Articulated for PIFu, MF-PIFu, PIFu+VSR, and our method. The y axis stands for the accuracy of P2S. The x axis is the number of samples in the testing data

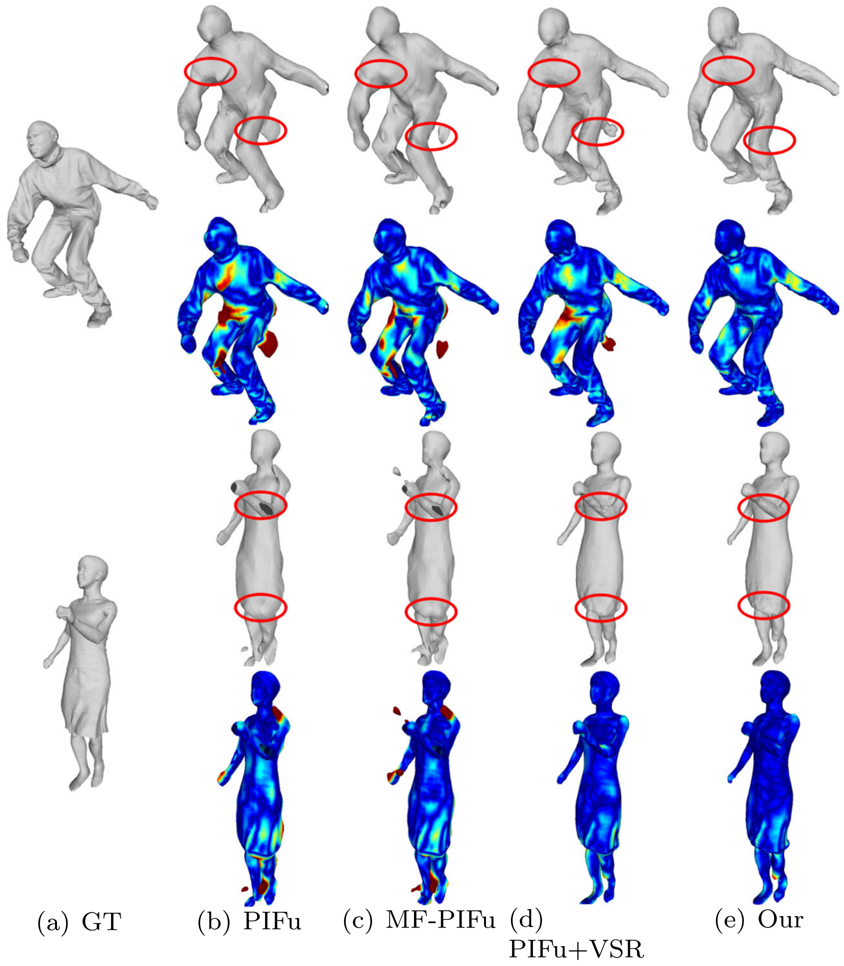


Fig. 11 The qualitative results of PIFu, MF-PIFu, PIFu+VSR, and our method on the Articulated dataset

$(\sigma_{\max}, \sigma_{\min}) = (15, 1.5), (25, 2.5), (35, 3.5)$. The P2S and IoU of the results for $(\sigma_{\max}, \sigma_{\min}) = (15, 1.5)$ are the best, but it does not have too much difference with (25, 2.5) and (35, 3.5). This is the reason that $(\sigma_{\max}, \sigma_{\min}) = (15, 1.5)$ is used in our method.

Figure 13 shows two examples for different σ from the Articulated dataset. We also give the visualization of the errors for the 3D models. From the figure, it can be seen that the estimated models of $\sigma_{\max} = 5$ have extra unnecessary parts. The errors of $\sigma_{\max} = 10$ are also relatively high from the visualization map, while the results of $\sigma_{\max} = 15, 25, 35$ are almost the same level. However, as shown in the areas indicated by the red circles, the surface details of the estimated 3D models of $\sigma_{\max} = 15$ are better preserved, especially for the neck part of the first example. Therefore, according to the above observation, the best choice for $(\sigma_{\max}, \sigma_{\min})$ is (15, 1.5) for the Articulated dataset. It is also acceptable to use larger $(\sigma_{\max}, \sigma_{\min})$, for instance, (25, 2.5) and (35, 3.5). However, this does not mean that $\sigma_{\max}$ can be too large because the results may not be good if $\sigma_{\min}$ is larger than 5 cm. The reasonable range for $(\sigma_{\max}, \sigma_{\min})$ is (15, 1.5) ~ (35, 3.5) according to the experiments.

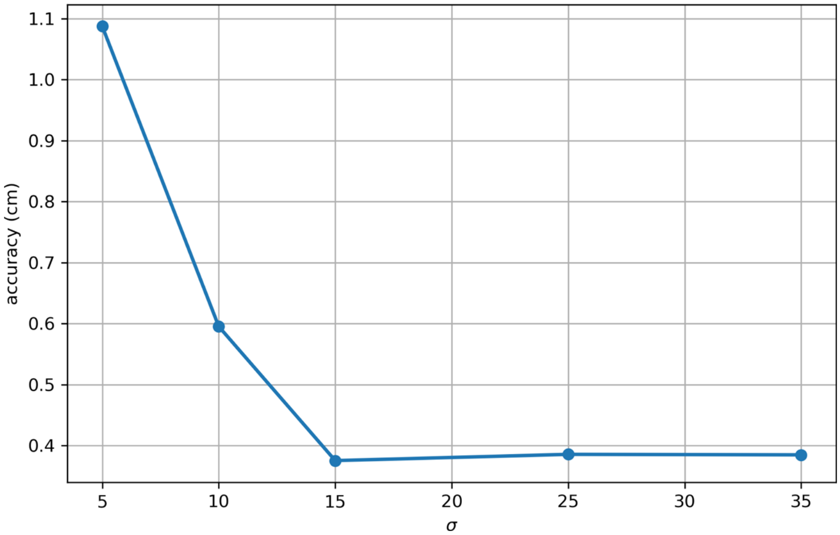
Table 6 Quantitative results of different $(\sigma_{\max}, \sigma_{\min})$ on the Articulate dataset

$(\sigma_{\max}, \sigma_{\min})$ (cm)	P2S ↓	Chamfer-L_2 ↓	IoU ↑
(5,0.5)	1.0874	0.1151	0.9006
(10,1.0)	0.5953	0.0110	0.8466
(15,1.5)	**0.3754**	0.0032	**0.9051**
(25,2.5)	0.3856	0.0030	0.8986
(35,3.5)	0.3848	**0.0029**	0.8984

4.7 Voxel grid resolution

The resolution of input voxel grids for VSR will also affects the performance of VSR to refine 3D models. In order to demonstrate the effects, the results of VSR with the input resolution of 32^3 and 128^3 for the Articulated dataset are compared. The voxel grids with different resolutions are generated from the estimated 3D models of MF-PIFu. Using the VSR which is trained by voxel grids with 128^3, the final results are generated from voxel grids with 32^3 and 128^3, respectively. Table 7 shows the P2S, Chamfer-L_2

Fig. 12 The mean P2S on the testing dataset of the Articulated for different $\sigma_{\max}$. The y axis stands for the mean P2S. The x axis is the $\sigma_{\max}$



and IoU of the results on the testing dataset of the Articulated for the input low-resolution voxel grids with 32^3 and 128^3 resolution which are the bold numbers in the table. We can see that the quantitative values of results for 128^3 resolution are better than 32^3. It is reasonable because higher resolution can provide more details for the voxel super-resolution. Figure 14 shows some examples of the 32^3 and 128^3 resolution. The 3D models after voxel super-resolution and the corresponding visualization of errors are shown in the figure. It also demonstrates that the results of VSR with 128^3 resolution voxel grids has better details on the shape, especially for those areas indicated by the red circles. Therefore, the resolution of input voxel grid for voxel super-resolution should be as high as possible. In our observation, the resolution 128^3 is reasonable to obtain good 3D model estimation considering the limitation of memory footprint.

4.8 The number of images

Since we estimate 3D human body from multi-view images, the effect of the number of views on the final estimation also needs to be discussed. The performance of our method for four images and eight images on the Articulated dataset is evaluated. Note that the MF-PIFu is trained by the four-view images and eight-view images, respectively. For the VSR, it is only trained by the voxel grids with 128^3 resolution generated by the four-view images. Table 8 shows the quantitative results on the Articulated dataset when the four-view and eight-view images are used. Figure 15 is the P2S of each sample in the testing dataset of Articulated for the four-view and eight-view cases. We can see that the results of eight-view case are a little better than the four-view case as shown by the bold numbers in the table. Since eight-view images could provide more information for the MF-PIFu than the four-view images, the coarse 3D models obtained by MF-PIFu are more accurate, which ensures the coarse 3D models can provide more information for VSR to obtain better refined 3D models. During the VSR, the training on the 3D space can help to reduce the ambiguity of four-view and eight-view cases. The final estimation does not have too much difference in the two cases.

Two examples from the Articulated dataset are shown in Fig. 16 for the four-view and eight-view images. The figure

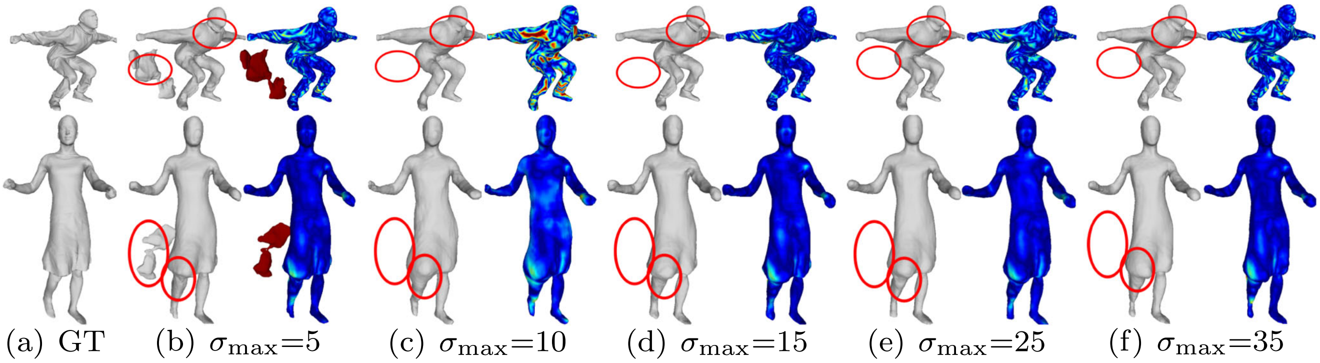


Fig. 13 The comparison for different $\sigma_{\max}$ on the Articulated dataset. From (**a**) to (**f**), two examples from the testing dataset are shown for $\sigma_{\max} = 5, 10, 15, 25, 35$. For each $\sigma_{\max}$, the visualization of the error between the estimated result and the ground truth is given

Table 7 Quantitative results of 32^3 and 128^3 resolutions on the Articulate dataset

voxel res.	P2S ↓	Chamfer-L_2 ↓	IoU ↑
Ours(32^3)	1.9322	0.1626	0.6902
Ours(128^3)	**0.3754**	**0.0032**	**0.9051**

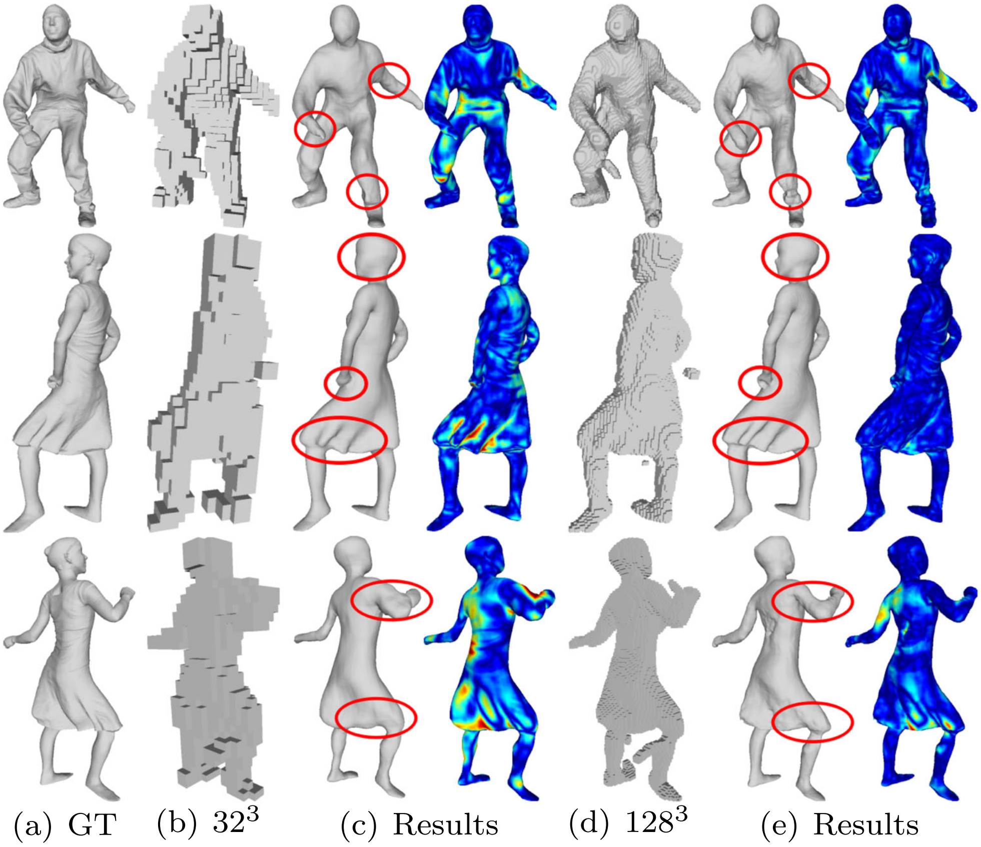


Fig. 14 The comparison between 32^3 and 128^3 resolution on the Articulated dataset. **a** is the ground truth of 3D models; (b) is the voxel grids with 32^3; **c** is the results of super resolution trained by 32^3 voxel grids; **d** is the voxel grids with 128^3; **e** is the results of super resolution trained by 128^3 voxel grids

Table 8 Quantitative results for the four-view and eight-view images on the Articulated dataset

View	P2S ↓	Chamfer-L_2 ↓	IoU ↑
Ours(four views)	0.3754	0.0032	0.9051
Ours(Eight views)	**0.3606**	**0.0021**	**0.9042**

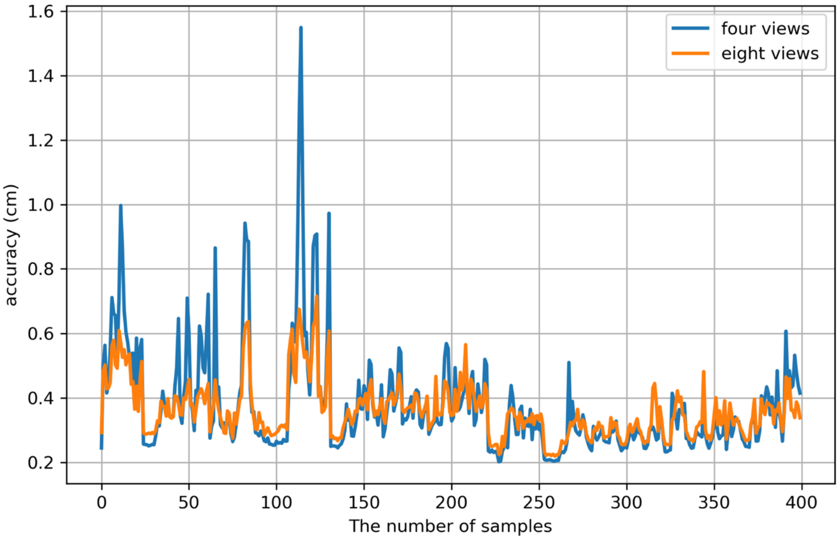


Fig. 15 The P2S of each sample in the testing data of the Articulated for four-view and eight-view images. The *y* axis stands for the accuracy of P2S. The *x* axis is the number of samples in the testing data

gives the results of MF-PIFu (b), the results of VSR (c) for the four-view images and the results of MF-PIFu (d), the results of VSR (e) for the eight-view images. We can see that there exists some error reconstruction on the 3D models of MF-PIFu for the four views, especially for the areas indicated by the red circles. The results of MF-PIFu of eight-view images looks better than four-view images. After the VSR, the coarse 3D models are refined to more accurate models, but the errors are not removed completely for the four-images. By contrast, the results of eight-view images look more smooth and accurate. Therefore, it is useful to obtain better estimation if there are more views. In this paper, it has been enough to obtain satisfying 3D models by four-view images.

5 Conclusion

Detailed 3D human body reconstruction from 2D images is a challenging task because of the high freedom of human body and the ambiguity of inferring 3D objects from 2D images. In this paper we propose a coarse-to-fine method for detailed 3D human body reconstruction from multi-view images through learning an implicit representation. The coarse 3D models are estimated from multi-view images through learning pixel-aligned implicit function based on multi-scale features which encode both local and global information. Then, generating the low-resolution voxel grids through voxelizing the coarse 3D models, VSR is learned to refine the coarse 3D models. For learning

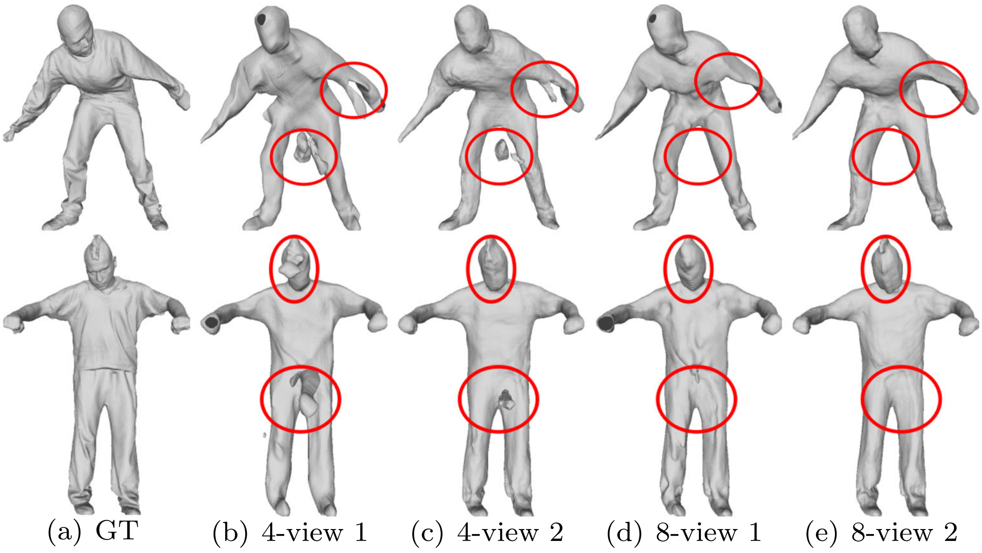


Fig. 16 The results of four-view and eight-view images on the Articulated dataset. From left to right columns: ground truth, the results of MF-PIFu of four-view images, the final results of four-view images, the results of MF-PIFu of eight-view images, and the final results of eight-view images

VSR, multi-stage 3D convolutional layers are used to extract multi-scale features from low-resolution voxel grids. The implicit representation is also learned based on the multi-scale features for VSR. Benefiting from the voxel super-resolution, the coarse 3D models can be refined to have higher accuracy and better surface quality because the false reconstruction on the coarse 3D models can be removed and the details on the shape can be preserved. The experiments on the public datasets demonstrate that our method can recover detailed 3D human body models from multi-view images with higher accuracy and completeness than previous approaches.

Some work needs to be done in the future. Firstly, the variety of the training dataset need to be added. The models in the two datasets of our paper mostly have the same color clothes. If there is a new model with colourful clothes, our method will fail to obtain good results. However, the high-quality 3D human body models are not easy to be acquired and many datasets are not free, which increases the difficulty for the research. Besides, the texture of the detailed model is not considered in our method which should be done in the future. Finally, single-view image based reconstruction is needed in the future to increase the convenience of our method.

Funding Open access funding provided by Lund University.

References

1. Alldieck T, Magnor M, Xu W, Theobalt C, Pons-Moll G (2018) Detailed human avatars from monocular video. In: International conference on 3d vision (3DV), pp 98–109
2. Alldieck T, Magnor M, Xu W, Theobalt C, Pons-Moll G (2018) Video based reconstruction of 3D people models. In: IEEE Conference on computer vision and pattern recognition (CVPR), pp 8387–8397
3. Alldieck T, Pons-Moll G, Theobalt C, Magnor M (2019) Tex2Shape: Detailed full human body geometry from a single image. In: International conference on computer vision (ICCV), pp 2293–2303
4. Alp Güler R, Neverova N, Kokkinos I (2018) DensePose: Dense human pose estimation in the wild. In: IEEE Conference on computer vision and pattern recognition (CVPR), pp 7297–7306
5. Anguelov D, Srinivasan P, Koller D, Thrun S, Rodgers J, Davis J (2005) SCAPE:Shape completion and animation of people. ACM Trans Graph 24(3):408–416
6. Balan A, Sigal L, Black MJ, Davis JE, Haussecker HW (2007) Detailed human shape and pose from images. In: IEEE Conference on computer vision and pattern recognition (CVPR), pp 1–8
7. Bogo F, Black MJ, Loper M, Romero J (2015) Detailed full-body reconstructions of moving people from monocular RGB-d sequences. In: International conference on computer vision (ICCV), pp 2300–2308
8. Bogo F, Kanazawa A, Lassner C, Gehler P, Romero J, Black MJ (2016) Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In: European conference on computer vision (ECCV), pp 561–578
9. Cao Z, Hidalgo Martinez G, Simon T, Wei S, Sheikh YA (2019) OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. IEEE Trans Pattern Anal Mach Intell:1–1
10. Chen Z, Zhang H (2019) Learning implicit fields for generative shape modeling. In: IEEE Conference on computer vision and pattern recognition (CVPR), pp 5939–5948
11. Chibane J, Alldieck T, Pons-Moll G (2020) Implicit functions in feature space for 3D shape reconstruction and completion. In: IEEE Conference on computer vision and pattern recognition (CVPR)
12. Choy CB, Xu D, Gwak J, Chen K, Savarese S (2016) 3D-R2N2: A unified approach for single and multi-view 3D object reconstruction. In: European conference on computer vision (ECCV), pp 628–644
13. Dibra E, Jain H, Öztireli C, Ziegler R, Gross M (2016) Hs-nets: Estimating human body shape from silhouettes with convolutional neural networks. In: International conference on 3d vision (3DV), pp 108–117
14. Dibra E, Jain H, Oztireli C, Ziegler R, Gross M (2017) Human shape from silhouettes using generative hks descriptors and cross-modal neural networks. In: IEEE Conference on computer vision and pattern recognition (CVPR), pp 4826–4836
15. Fan H, Su H, Guibas LJ (2017) A point set generation network for 3D object reconstruction from a single image. In: IEEE Conference on computer vision and pattern recognition (CVPR), pp 605–613
16. Gabeur V, Franco JS, Martin X, Schmid C, Rogez G (2019) Moulding Humans: Non-parametric 3D human shape estimation from single images. In: International conference on computer vision (ICCV), pp 2232–2241
17. Gilbert A, Volino M, Collomosse J, Hilton A (2018) Volumetric performance capture from minimal camera viewpoints. In: European conference on computer vision (ECCV), pp 566–581
18. Guan P, Weiss A, Balan A, Black MJ (2009) Estimating human shape and pose from a single image. In: International conference on computer vision (ICCV), pp 1381–1388
19. Habermann M, Xu W, Zollhoefer M, Pons-Moll G, Theobalt C (2019) LiveCap: Real-time human performance capture from monocular video. ACM Trans Graph 38(2):1–17
20. He K, Gkioxari G, Dollár P, Girshick R (2017) Mask r-CNN. In: International conference on computer vision (ICCV), pp 2961–2969
21. Huang Y, Bogo F, Lassner C, Kanazawa A, Gehler PV, Romero J, Akhter I, Black MJ (2017) Towards accurate marker-less human shape and pose estimation over time. In: International conference on 3d vision (3DV), pp 421–430

22. Huang Z, Li T, Chen W, Zhao Y, Xing J, LeGendre C, Luo L, Ma C, Li H (2018) Deep volumetric video from very sparse multi-view performance capture. In: European conference on computer vision (ECCV), pp 336–354
23. Huang Z, Xu Y, Lassner C, Li H, Tung T (2020) ARCH: Animatable reconstruction of clothed humans. In: IEEE Conference on computer vision and pattern recognition (CVPR), pp 3093–3102
24. Jackson AS, Manafas C, Tzimiropoulos G (2018) 3D human body reconstruction from a single image via volumetric regression. In: Computer vision – ECCV 2018 workshops, pp 64–77
25. Joo H, Simon T, Li X, Liu H, Tan L, Gui L, Banerjee S, Godisart T, Nabbe B, Matthews I et al (2017) Panoptic Studio: A massively multiview system for social interaction capture. IEEE Trans Pattern Anal Mach Intell 41(1):190–204
26. Kanazawa A, Black MJ, Jacobs DW, Malik J (2018) End-to-end recovery of human shape and pose. In: IEEE Conference on computer vision and pattern recognition (CVPR), pp 7122–7131
27. Kanazawa A, Zhang JY, Felsen P, Malik J (2019) Learning 3D human dynamics from video. In: IEEE Conference on computer vision and pattern recognition (CVPR), pp 5614–5623
28. Kar A, Häne C, Malik J (2017) Learning a multi-view stereo machine. In: Advances in neural information processing systems, pp 365–376
29. Kocabas M, Athanasiou N, Black MJ (2020) VIBE: Video inference for human body pose and shape estimation. In: IEEE Conference on computer vision and pattern recognition (CVPR), pp 5253–5263
30. Kolotouros N, Pavlakos G, Black MJ, Daniilidis K (2019) Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In: International conference on computer vision (ICCV), pp 2252–2261
31. Kolotouros N, Pavlakos G, Daniilidis K (2019) Convolutional mesh regression for single-image human shape reconstruction. In: IEEE Conference on computer vision and pattern recognition (CVPR), pp 4501–4510
32. Lazova V, Insafutdinov E, Pons-moll G (2019) 360-degree textures of people in clothing from a single image. In: International conference on 3d vision (3DV), pp 643–653
33. Li Z, Heyden A, Oskarsson M (2019) Parametric model-based 3D human shape and pose estimation from multiple views. In: Scandinavian conference on image analysis (SCIA), pp 336–347
34. Liang J, Lin MC (2019) Shape-aware human pose and shape reconstruction using multi-view images. In: International conference on computer vision (ICCV), pp 4352–4362
35. Liu Z, Huang J, Bu S, Han J, Tang X, Li X (2016) Template deformation-based 3D reconstruction of full human body scans from low-cost depth cameras. IEEE Trans Cybern 47(3):695–708
36. Loper M, Mahmood N, Romero J, Pons-Moll G, Black MJ (2015) SMPL: A skinned multi-person linear model. ACM Trans Graph 34(6):1–16
37. Ma Q, Yang J, Ranjan A, Pujades S, Pons-Moll G, Tang S, Black MJ (2020) Learning to dress 3D people in generative clothing. In: IEEE Conference on computer vision and pattern recognition (CVPR), pp 6469–6478
38. Martinez J, Hossain R, Romero J, Little JJ (2017) A simple yet effective baseline for 3D human pose estimation. In: International conference on computer vision (ICCV), pp 2640–2649
39. Mescheder L, Oechsle M, Niemeyer M, Nowozin S, Geiger A (2019) Occupancy Networks: Learning 3D reconstruction in function space. In: IEEE Conference on computer vision and pattern recognition (CVPR), pp 4460–4470
40. Natsume R, Saito S, Huang Z, Chen W, Ma C, Li H, Morishima S (2019) SiCloPe: Silhouette-based clothed people. In: IEEE Conference on computer vision and pattern recognition (CVPR), pp 4480–4490
41. Newell A, Yang K, Deng J (2016) Stacked hourglass networks for human pose estimation. In: European conference on computer vision (ECCV), pp 483–499
42. Onizuka H, Hayirci Z, Thomas D, Sugimoto A, Uchiyama H, Taniguchi RI (2020) TetraTSDF: 3D human reconstruction from a single image with a tetrahedral outer shell. In: IEEE Conference on computer vision and pattern recognition (CVPR), pp 6011–6020
43. Park JJ, Florence P, Straub J, Newcombe R, Lovegrove S (2019) DeepSDF: Learning continuous signed distance functions for shape representation. In: IEEE Conference on computer vision and pattern recognition (CVPR), pp 165–174
44. Pavlakos G, Kolotouros N, Daniilidis K (2019) TexturePose: Supervising human mesh estimation with texture consistency. In: International conference on computer vision (ICCV), pp 803–812
45. Pavlakos G, Zhu L, Zhou X, Daniilidis K (2018) Learning to estimate 3D human pose and shape from a single color image. In: IEEE Conference on computer vision and pattern recognition (CVPR), pp 459–468
46. Pishchulin L, Insafutdinov E, Tang S, Andres B, Andriluka M, Gehler PV, Schiele B (2016) DeepCut: Joint subset partition and labeling for multi person pose estimation. In: IEEE Conference on computer vision and pattern recognition (CVPR), pp 4929–4937
47. Riegler G, Osman Ulusoy A, Geiger A (2017) OctNet: Learning deep 3D representations at high resolutions. In: IEEE Conference on computer vision and pattern recognition (CVPR), pp 3577–3586
48. Saito S, Huang Z, Natsume R, Morishima S, Kanazawa A, Li H (2019) PIFu: Pixel-Aligned implicit function for high-resolution clothed human digitization. In: International conference on computer vision (ICCV), pp 2304–2314
49. Saito S, Simon T, Saragih J, Joo H (2020) PIFuHD: Multi-level pixel-aligned implicit function for high-resolution 3D human digitization. In: IEEE Conference on computer vision and pattern recognition (CVPR), pp 84–93
50. Sigal L, Balan A, Black MJ (2008) Combined discriminative and generative articulated pose and non-rigid shape estimation. In: Advances in neural information processing systems, pp 1337–1344
51. Sigal L, Bhatia S, Roth S, Black MJ, Isard M (2004) Tracking loose-limbed people. In: IEEE Conference on computer vision and pattern recognition (CVPR), vol 1, pp i–i
52. Tan JKV, Budvytis I, Cipolla R (2017) Indirect deep structured learning for 3d human body shape and pose prediction
53. Varol G, Ceylan D, Russell B, Yang J, Yumer E, Laptev I, Schmid C (2018) BodyNet: Volumetric inference of 3d human body shapes. In: European conference on computer vision (ECCV), pp 20–36
54. Vlasic D, Baran I, Matusik W, Popović J (2008) Articulated mesh animation from multi-view silhouettes. ACM Trans Graph 27(3):1–9
55. Wang N, Zhang Y, Li Z, Fu Y, Liu W, Jiang YG (2018) Pixel2Mesh: Generating 3D mesh models from single RGB images. In: European conference on computer vision (ECCV), pp 52–67
56. Weiss A, Hirshberg D, Black MJ (2011) Home 3D body scans from noisy image and range data. In: International conference on computer vision (ICCV), pp 1951–1958
57. Wu Y, Jiang X, Fang Z, Gao Y, Hamido F (2021) Multi-modal 3D object detection by 2D-guided precision anchor proposal and multi-layer fusion. Appl Soft Comput J 108:107405
58. Xia F, Wang P, Chen X, Yuille AL (2017) Joint multi-person pose estimation and semantic part segmentation. In: IEEE Conference on computer vision and pattern recognition (CVPR), pp 6769–6778

59. Xu L, Su Z, Han L, Yu T, Liu Y, Lu F (2019) UnstructuredFusion: Realtime 4D geometry and texture reconstruction using commercial RGBD cameras. IEEE Transactions on Pattern Analysis and Machine Intelligence
60. Xu W, Chatterjee A, Zollhöfer M, Rhodin H, Mehta D, Seidel HP, Theobalt C (2018) MonoPerfCap: Human performance capture from monocular video. ACM Trans Graph 37(2):1–15
61. Zheng Z, Yu T, Wei Y, Dai Q, Liu Y (2019) DeepHuman: 3D human reconstruction from a single image. In: International conference on computer vision (ICCV), pp 7739–7749
62. Zhu H, Zuo X, Wang S, Cao X, Yang R (2019) 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
63. Zhu K, Jiang X, Fang Z, Gao Y, Fujita H, Hwang JN (2021) Photometric transfer for direct visual odometry. Knowl-Based Syst 213:106671

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