3D Human Pose and Shape Estimation Based on Parametric Model and Deep Learning

Li, Zhongguo

2021

Document Version:
Publisher's PDF, also known as Version of record

3D Human Pose and Shape Estimation Based on Parametric Model and Deep Learning

3D Human Pose and Shape Estimation Based on Parametric Model and Deep Learning

by Zhongguo Li

LUND
UNIVERSITY

Academic Thesis
Thesis advisors: Anders Heyden
Faculty opponent: Janne Heikkilä, Faculty of Information Technology and Electrical Engineering, University of Oulu

To be presented, with the permission of the Faculty of Engineering of Lund University, for public criticism in the room MH:Hörmander at the Centre for Mathematical Sciences on Friday, the 21st of May 2021 at 10:00.

Organization
LUND UNIVERSITY

Centre for Mathematical Sciences
Box 118
SE–22100 LUND
Sweden

Author(s)
Zhongguo Li

Document name
Doctoral thesis in Mathematical sciences

Date of presentation
2021-05-21

Sponsoring organization China Scholarship Council (CSC)

Title and subtitle
3D Human Pose and Shape Estimation Based on Parametric Model and Deep Learning

Abstract

3D human body reconstruction from monocular images has wide applications in our life, such as movie, animation, Virtual/Augmented Reality, medical research and so on. Due to the high freedom of human body in real scene and the ambiguity of inferring 3D objects from 2D images, it is a challenging task to accurately recover 3D human body models from images. In this thesis, we explore the methods for estimating 3D human body models from images based on parametric model and deep learning.

In the first part, the coarse 3D human body models are estimated automatically from multi-view images based on a parametric human body model called SMPL model. Two routes are exploited for estimating the pose and shape parameters of the SMPL model to obtain the 3D models: (1) Optimization based methods; and (2) Deep learning based methods. For the optimization based methods, we propose the novel energy functions based on some prior information including the 2D joint points and silhouettes. Through minimizing the energy functions, the SMPL model is fitted to the prior information, and then, the coarse 3D human body is obtained. In addition to the traditional optimization based methods, a deep learning based method is also proposed in the following work to regress the pose and shape parameters of the SMPL model. A novel architecture is proposed to put the optimization into a training loop of convolutional neural network (CNN) to form a self-supervision structure based on the multi-view images. The proposed methods are evaluated on both synthetic and real datasets to demonstrate that they can obtain better estimation of the pose and shape of 3D human body than previous approaches.

In the second part, the problem is shifted to the detailed 3D human body reconstruction from multi-view images. Instead of using the SMPL model, implicit function is utilized to represent 3D models because implicit representation can generate continuous surface and has better flexibility for arbitrary topology. Firstly, a multi-scale features based method is proposed to learn the implicit representation for 3D models through multi-stage hourglass networks from multi-view images. Furthermore, a coarse-to-fine method is proposed to refine the 3D models from multi-view images through learning the voxel super-resolution. In this method, the coarse 3D models are estimated firstly by the learned implicit function based on multi-scale features from multi-view images. Afterwards, by voxelizing the coarse 3D models to low resolution voxel grids, voxel super-resolution is learned through a multi-stage 3D CNN for feature extraction from low resolution voxel grids and fully connected neural network for predicting the implicit function. Voxel super-resolution is able to remove the false reconstruction and preserve the surface details. The proposed methods are evaluated on both real and synthetic datasets in which our method can estimate 3D model with higher accuracy and better surface quality than some previous methods.

Key words
3D Human body; Pose and shape estimation; Parametric model; SMPL model; Deep Learning; Implicit function

Classification system and/or index terms (if any)

Supplementary bibliographical information

Language
English

ISSN and key title
1404-0034

ISBN
978-91-7895-787-3 (printed)
978-91-7895-788-0 (electronic)

Recipient's notes

Number of pages
179

Price

Security classification

Signature _____ Date 2021-05-21

3D Human Pose and Shape Estimation Based on Parametric Model and Deep Learning

by Zhongguo Li

LUND
UNIVERSITY

Funding information: The thesis work was financially supported by the China Scholarship Council

Centre for Mathematical Sciences
Lund University
Box 118
SE-22100 Lund
Sweden

www.maths.lth.se

Doctoral Thesis in Mathematical Sciences 2021:5
ISSN:1404-0034

ISBN: 978-91-7895-787-3 (printed)
ISBN: 978-91-7895-788-0 (electronic)
LUTFMA-1073-2021

Printed in Sweden by Media-Tryck, Lund University, Lund 2021

Dedicated to my family ...
谢谢家人的支持与鼓励 ...

Abstract

3D human body reconstruction from monocular images has wide applications in our life, such as movie, animation, Virtual/Augmented Reality, medical research and so on. Due to the high freedom of human body in real scene and the ambiguity of inferring 3D objects from 2D images, it is a challenging task to accurately recover 3D human body models from images. In this thesis, we explore the methods for estimating 3D human body models from images based on parametric model and deep learning.

In the first part, the coarse 3D human body models are estimated automatically from multi-view images based on a parametric human body model called SMPL model. Two routes are exploited for estimating the pose and shape parameters of the SMPL model to obtain the 3D models: (1) Optimization based methods; and (2) Deep learning based methods. For the optimization based methods, we propose the novel energy functions based on some prior information including the 2D joint points and silhouettes. Through minimizing the energy functions, the SMPL model is fitted to the prior information, and then, the coarse 3D human body is obtained. In addition to the traditional optimization based methods, a deep learning based method is also proposed in the following work to regress the pose and shape parameters of the SMPL model. A novel architecture is proposed to put the optimization into a training loop of convolutional neural network (CNN) to form a self-supervision structure based on the multi-view images. The proposed methods are evaluated on both synthetic and real datasets to demonstrate that they can obtain better estimation of the pose and shape of 3D human body than previous approaches.

In the second part, the problem is shifted to the detailed 3D human body reconstruction from multi-view images. Instead of using the SMPL model, implicit function is utilized to represent 3D models because implicit representation can generate continuous surface and has better flexibility for arbitrary topology. Firstly, a multi-scale features based method is proposed to learn the implicit representation for 3D models through multi-stage hourglass networks from multi-view images. Furthermore, a coarse-to-fine method is proposed to refine the 3D models from multi-view images through learning the voxel super-resolution. In this method, the coarse 3D models are estimated firstly by the learned implicit function based on multi-scale features from multi-view images. Afterwards, by voxelizing the coarse 3D models to low resolution voxel grids, voxel super-resolution is learned through a multi-stage 3D CNN for feature extraction from low resolution voxel grids and fully connected neural network for predicting the implicit function. Voxel super-resolution is able to remove the false reconstruction and preserve the surface details. The proposed methods are evaluated on both real and synthetic datasets in which our method can estimate 3D model with higher accuracy and better surface quality than some previous methods.

Popular Scientific Summary

In many scenes like movie, Virtual/Augmented Reality, and video games, 3D human body models play important roles to model the motion and shape of the real human body. Therefore, obtaining the 3D human body model is a practical task for many applications. Although there are many imaging systems for 3D human body reconstruction, people still would like to obtain the 3D human body from single- or multi-view 2D images because 2D images can be captured easily through many mobile devices. However, reconstructing 3D models from 2D images is a quite challenging task in computer vision because this is an ill-posed problem. In addition, human body in real life has very high degree of freedom, which makes it more difficult to reconstruct 3D human body model from 2D images accurately and efficiently.

This dissertation focuses on the methods to reconstruct 3D human body models from 2D images through parametric human body model and deep learning. The dissertation attempts to estimate 3D human body models from images from two aspects: coarse 3D human body model and detailed 3D human body model. The coarse 3D human body model mainly focuses the 3D human pose through estimating the parametric human body, while the detailed 3D human body model contains the details of the appearance of the human body including clothes and shape.

For the coarse 3D human body model reconstruction, the dissertation proposes three methods based on parametric human body models and deep learning. The first two methods mainly reconstruct the coarse 3D human body model through fitting the parametric human body model, SMPL model, to the prior information including joint points and silhouettes. In addition, another method based on deep learning is also proposed. This method estimates the pose and shape parameters of SMPL model through learning collaborating multi-view model-fitting. These methods can reconstruct the coarse 3D human body with good performance for the pose estimation.

For the detailed 3D human body model reconstruction, two methods through multi-view 2D images based on deep learning are proposed in the dissertation. In order to obtain the 3D surface better, the implicit function is learned in the two methods to represent 3D models. The first method learns the implicit function through multi-scale features extracted by multi-stage hourglass network from multi-view images. The second method, which is a coarse-to-fine manner, adds the voxel super-resolution which is also implemented by learning implicit function to the first method. These two methods can reconstruct the 3D human body model with detailed appearance from multi-view images.

This dissertation discusses the problem of 3D human body reconstruction from 2D images and provides some possible solution to the problem. This practical research could be used in many fields like animation and VR/AR.

Publications

Publications concerning the work of this thesis have been made as follows:

I **Zhongguo Li**, Anders Heyden, Magnus Oskarsson, "Parametric Model-Based 3D Human Shape and Pose Estimation from Multiple Views", In: *the 21st Scandinavian Conference on Image Analysis (SCIA 2019)*, p.336-347, Norrköping, Sweden.

Contribution: Zhongguo Li came up with the idea, developed the theory, implemented the algorithm and did all the experiments. Zhongguo Li wrote the most texts of the paper. Anders Heyden and Magnus Oskarsson revised the paper and gave some comments about the experiemts.

II **Zhongguo Li**, Magnus Oskarsson, Anders Heyden, "A Novel Joint Points and Silhouette-based Method to Estimate 3D Human Pose and Shape", In: *the 25th International Conference on Pattern Recognition Workshop (ICPR Workshop 2020), 3D Human Understanding (3DHU)*,p.41-56, selected to submit to *Journal of Imaging*, major revision.

Contribution: Zhongguo Li came up with the idea, developed the theory, implemented the algorithm and did all the experiments. Zhongguo Li wrote the most texts of the paper. Anders Heyden and Magnus Oskarsson revised the paper and gave some comments about the experiments.

III **Zhongguo Li**, Magnus Oskarsson, Anders Heyden, "3D Human Pose and Shape Estimation Through Collaborative Learning and Multi-view Model-fitting", In: *2021 Winter Conference on Applications of Computer Vision (WACV 2021)*, p.1888-1897, Best Paper Award: Applications.

Contribution: Zhongguo Li came up with the idea and developed the theory with Magnus Oskarsson. Zhongguo Li implemented the algorithm, did all the experiments and wrote the most texts of the paper. Anders Heyden and Magnus Oskarsson revised the paper and gave some comments about the experiments.

IV **Zhongguo Li**, Magnus Oskarsson, Anders Heyden, "Learning to Implicitly Represent 3D Human Body From Multi-scale Features and Multi-view Images", In: *the 25th International Conference on Pattern Recognition (ICPR 2020)*, p.8968-8975.

Contribution: Zhongguo Li came up with the idea, developed the theory, implemented the algorithm, did all the experiments and wrote the most texts of the paper. Anders Heyden and Magnus Oskarsson revised the paper, discussed the theory and gave some comments about the experiments.

v **Zhongguo Li**, Magnus Oskarsson, Anders Heyden, "Implicit Function Based 3D Human Body Reconstruction Through Multi-view Images and Voxel Super-Resolution", submitted to *Applied Intelligence*, under review.

Contribution: Zhongguo Li came up with the idea, developed the theory, implemented the algorithm, did all the experiments and wrote the most texts of the paper. Anders Heyden and Magnus Oskarsson revised the paper, discussed the theory and gave some comments about the experiments.

The following paper is not included in the thesis:

vi **Zhongguo Li**, Anders Heyden, Magnus Oskarsson, "Template Based Human Pose and Shape Estimation From a Single RGB-D Image", In:*the 8th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2019)*, p.574-581, Prague, Czech Republic.

Contribution: Zhongguo Li came up with the idea, developed the theory, implemented the algorithm, did all the experiments and wrote the most texts of the paper. Anders Heyden and Magnus Oskarsson revised the paper and gave some comments about the experiments.

Acknowledgements

This thesis marks the end of four and a half years chapter of my life. There are so many people who gave me much support for my PhD career. First of all, I would like to express my deepest appreciation to my supervisors Prof. Anders Heyden and Dr. Magnus Oskarsson for your guidance, patience, fruitful discussion and encouragement during the past four years. I did not know much about the beautiful country when I arrived at Sweden first time four years ago, but you provided me much assistance on the life and work in Sweden so that I could start my study as soon as possible. Besides, I would like to thank my colleagues at the Center for Mathematical Sciences and in particular the Computer Vision Group for your support and help. They are Erik Bylow, Gabrielle Flood, Maria Priisalu, Marcus Valtonen Örnhag, Ida Arvidsson and so on. You gave me many useful information about the course, research and the way of life in Sweden. It is really an inspirational and lovely research environment. Also I would like to thank my friends Jing Yang and Xiaoqing Hou. We had many happy time on sports, traveling and cooking, which enriched my spare time in Sweden. I would also like to thank my Chinese supervisor Prof. Bin Yan and Dr. Jian Chen and other colleagues including Dr. Kai Qiao and Dr. Wenkun Zhang for their support and help during my study in Sweden. As my best friend from undergraduate to PhD, I would also like to thank Dr. Yiwei Pan for his support.

Last but not least I would like to express my gratitude to my family for their tolerance and support when I studied in Sweden which is a very distant country for them. They gave me meticulous care and love. Without their support, I can not study in Sweden with peace of mind. I would also like to thank to my girlfriend Xiaoya Li. Although we met in the last year of my PhD, the support from her is very important and makes me feel love during writing and revising the thesis.

Finally, I would like to thank the China Scholarship Council (CSC) to provide the fund for my study in Sweden.

Contents

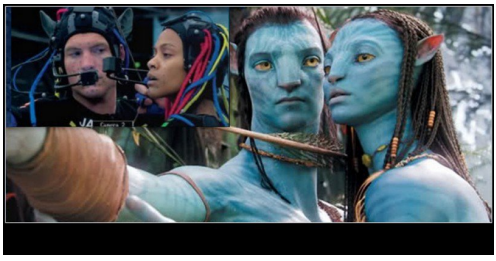
Chapter 1

Introduction

In this chapter, the background of the thesis is introduced to state the significance and importance of our work. In the first part, the motivation and the challenges of the research are presented to show the wide applications of 3D human body and the difficulties of estimating 3D human body from images. Then, the mathematical problems of estimating 3D human body from images are stated according to different methods and ideas. In the following part, from several aspects, we fully summarize the related work on the problem of 3D pose and shape estimation of human body including the advantage and drawbacks of the different approaches. Finally, the contributions and the structure of the thesis are presented.

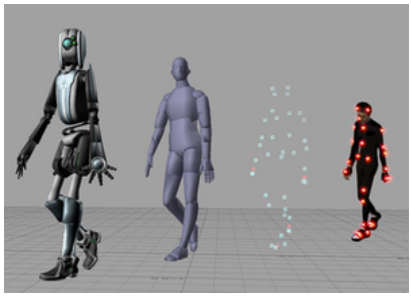
1 Motivation

In the modern life, with the development of the Internet, mobile devices and artificial intelligence, enormous human-centered applications and services are playing more and more important roles, which brings great convenience and is changing the way of our life. Besides, although you might not notice it, as an important way to communicate, body language accounts for at least 60% for the information exchange [146]. Therefore, if the computer can better understand human body motion, it will bring significant progress for the intelligent life and make it possible to build better human-computer interaction system. Image is one of the most important media for computer to understand the world, especially for the human body in the world. It is necessary and important to analyze images containing human bodies in computer vision because this is the basis for the services and communication of the Internet. There are many tasks to analyze humans based on images such as instance and part segmentation [125], pose and shape estimation [155], human

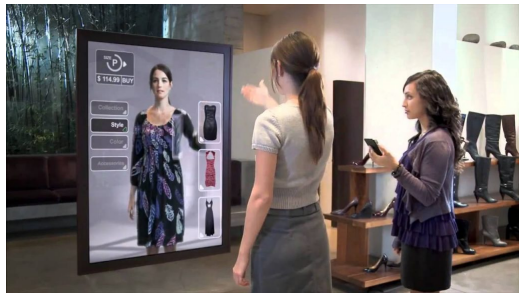


(a) Movie

(b) Game



(c) Animation



(d) Virtual try-on

Figure 1.1: Several examples of 3D human body model: (a) Movie[1]; (b) Game[2]; (c) Animation[3]; (d) Virtual try-on[4].

action recognition [149], 3D reconstruction [158] and so on. Among these tasks, estimating 3D human body including pose and shape from images is one of the most interesting problems. Given some 2D images containing human body, the goal of the problem is to build a 3D model which has the same pose and shape with the human body in the original images. Since the 3D human body has wide application in practical fields and the problem of inferring 3D objects from 3D images is ill-posed, this topic has attracted much attention during past several decades.

The 3D human body model can be widely seen in many practical applications such as the movie industry, the animation, the video games, the e-commerce fashion, biomedical research and so on [123]. It has no doubt that 3D human body model has become an essential role in our daily life, from entertainment to health, because of these applications. In the modern movie industry, the special effects of simulating 3D human body of actors could produce many scenes which may be dangerous for real actors or impossible to achieve in real life such as movies about space and Sic-fi movies. Especially for many commercial blockbusters, the special effect in the movie plays a very important role on the final box

[1]https://www.youtube.com/watch?v=1wK1Ixr-UmM
[2]https://www.pes-patch.com/pes-2020-of-changer-tool-by-nesa24/
[3]https://en.wikipedia.org/wiki/Visual_effects
[4]https://herinkheart.com/fashion-and-technology/

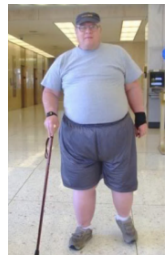
office income because it can bring amazing visual experience for the audience, which will attract more people to watch. In the animation and video games, accurate and realistic human body is a crucial factor to provide us better experience and entertainment. For example, in many Disney animation movies, researchers and animators need to capture 3D annotations of a real human body and transfer them to the cartoon characters, especially for the facial expression and body motion. The carton charactors with realistic facial expression and body motion can attract more audiences. Many video games, especially for sports games like PES, require to create virtual athletic stars so that players can enjoy the games better. The motion and action like running, jumping for the virtual athletic need to be generated in those sport games. Besides, 3D human body model can also be used for Virtual/Augment Reality in the e-commerce fashion so that we can shopping on-line without worrying about the size and style too much, which may be useful to reduce the rate of returning. In the biomedical fields, it is also a new trend to build 3D human body to measure our body index like weight and BMI so that the status of our body can be predicted, which is possible to prevent some diseases. Figure 1.1 shows some examples about the above mentioned applications in our life including movie, animation, video games and e-commerce. Therefore, capturing and building 3D human body model accurately and efficiently has good prospect considering its wide applications in real life, which makes it become one of the most important and active fields in computer vision and graphics during past decades.

2 Challenges

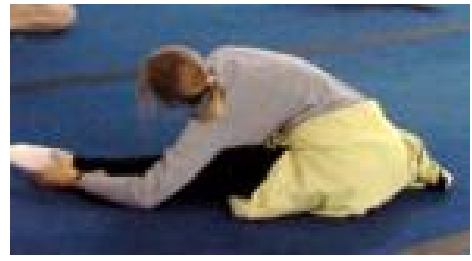
Although 3D human body model plays an important role in many fields, it is a challenging task to build accurate 3D human body model from 2D images. Two factors make the problem difficult: (1) Inferring 3D model from 2D images is an ill-posed problem; (2) Human body in the real scene has very high degree of freedom. As human beings, it is easy to understand and figure out 3D pose and shape only from 2D images because people have seen various and enormous human body motions in the daily life of growing and people even act similar poses in real scene, which accumulates a lot of experience for us. However, this is impossible for a computer to infer accurate 3D objects only from 2D images without any other prior information. It has inherent ambiguity to infer 3D model only from 2D images and the human bodies in real scenes have many complicated motions and various appearance. The examples of human bodies which has challenging properties in real scenes are shown in Figure 1.2. The technical difficulties to estimate 3D human body model from 2D images are summarized in the following paragraphs.

Firstly, the depth information is lost when a 3D model is projected onto 2D image plane by a camera. Recovering 3D objects from 2D images has some degrees of freedom and this is a classic camera calibration problem in computer vision. For example, Structure from

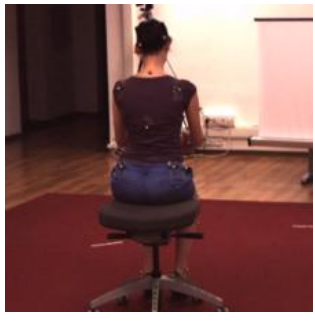
(a) Self-occlusion



(b) Shape



(c) Articulated pose



(d) Covered by chair

(e) Outdoor

(f) Clothes

Figure 1.2: The examples of human body in real scenes. The high uncertainty in human pose, shape, environment and appearance makes the problem extremely challenging.

Motion (SfM) can reconstruct the 3D structure of a scene or object from a series of 2D images which can be used for estimating camera matrix through registering the feature points in the images [110]. SfM can produce sparse or dense point clouds for the rigid objects very well through estimating the camera matrix which can be used for computing depth information. However, human body always has non-rigid motion and the dense 3D model is often needed. Traditional SfM algorithm does not have good performance to recover a moving 3D human body from a limited number of 2D images. In general, recovering 3D model from 2D images faces high ambiguity due to the lack of depth information.

Secondly, human body always shows occlusion or self-occlusion in 2D photograph, which results in that some parts of a human body cannot be seen directly from the 2D photograph. The occlusion means that some parts of the human body are covered by other objects, while the self-occlusion results from the human body itself if the human is performing some actions or is not oriented to the camera. This leads to difficulty of estimating human pose and shape directly from images through some traditional approaches. Estimating the 2D and 3D human body pose is a very important task in computer vision because human pose estimation is an important step for motion tracking and human action recognition. Considering that the pose of human body can be always represented as the locations of the

joint points, there will be high uncertainty and freedom to estimate the accurate locations of joint points only relying on images if some parts cannot be seen in the image. In addition, if some parts cannot be seen in the image, inferring the shape of those covered parts of human body is also a very difficult task.

Thirdly, a human body has internal high-dimension searching space. As we all know, a human body has 206 bones, which is a high dimensional problem to model the motion and deformation of the bones. In practice, different tasks will require different searching space to model the human body motion. More specifically, for the human pose estimation, 20 to 60 joint angles which are defined by a skeleton are the most common situation for the searching space of pose estimation. Although only about 20 joint points are usually used in pose estimation, this is still a quite high dimension. Therefore, the optimization on the problem of pose estimation alway requires highly computational efficiency to obtain the optima.

Besides, human bodies in real life usually have various appearance. Inherently, different people always show different height, weight, hair style, shape of muscles or even different skins for different ages. These variation should be considered when 3D human body is reconstructed from images because these are important to guarantee the realism. Furthermore, since people in ordinary life wear the different clothes both for male and female, it is also crucial to depict the wear in the reconstructed 3D human models. However, there are so many different styles on the clothes like tight clothes, loose clothes, hat, skirts and so on. Obviously, the high degree of freedom of the appearance leads that modeling human body is difficult and it will have great effect on the accuracy of the estimated 3D human body. Therefore, the modeling of clothes is also a key problem for human body reconstruction. On one hand, the very detailed model might be useful for specific problem but might not be general for other scenes. On the other hand, inaccurate body model is not enough to produce realism. The problem on balancing the accuracy and the generalization need to be considered.

Finally, the environment around human body is also an important factor to affect the 3D human body reconstruction from 2D image. There are many different scenarios in our real world, for instance, indoor scene and outdoor scene. For different scenarios, many factors including the lighting, albedo, shadow and the color of background vary significantly, which will affect the quality of 2D images. For many tasks, for example, the segmentation of human body, the background may have great influence on the complexity of segmentation. If the environment results in noise or artifacts in the images, extracting complete human body from the background may be a hard task for many existing algorithms.

Since there are many challenges for the problem, it is an impossible task to tackle all the problems in the thesis. For those human bodies with complicated poses, the thesis mainly focuses on the estimation of the human pose and obtains the coarse human body mod-

els. For those human bodies with simple pose and background, the detailed appearance including the clothes is obtained in the final reconstruction. In the following section, the mathematical problems of the coarse and detail reconstruction are discussed.

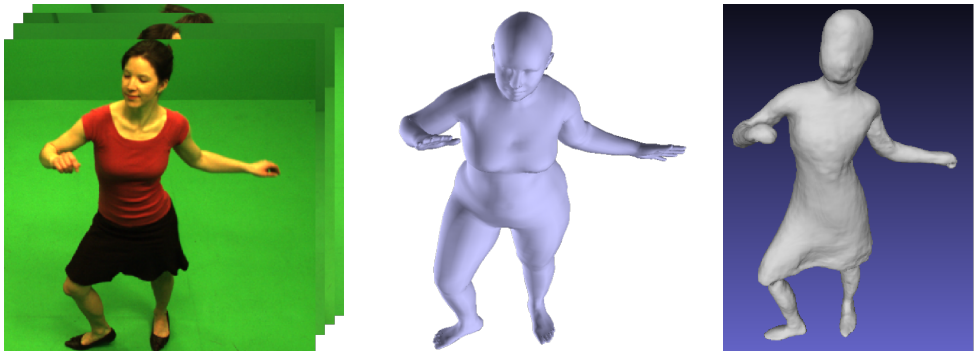
3 Problem statement

The goal of the dissertation is to reconstruct 3D human body model from 2D images, which is an important field of research in computer vision. It involves many different sub-problems, for instance, tracking [16], segmentation[125], 2D/3D skeleton estimation [155], facial expression recognition [176] and hand gesture recognition [150]. Each of the small topics plays important role for human-centered problems and has attracted enormous research. It is impossible to involve and tackle all these problems in one thesis. In this dissertation, two cases are discussed and explored: images based coarse reconstruction and images based detailed reconstruction of 3D human body. Given a series of 2D images containing a human body from multiple views, the task is to estimate a coarse or detailed 3D human body based on the images. For the coarse reconstruction, the details of the clothes and hair are not considered. The most important task for the problem is to estimate 3D joint points and modeling the human body. For the detailed reconstruction, the appearance of human body like clothes and hair should be shown in the final 3D model. This is more complicated than the coarse reconstruction, but the pose for the problem is usually simple. Figure 1.3 gives an example of the goal of this thesis. The coarse reconstruction is obtained based on some parametric human body. It does not have details on the appearance of the human body, but it can handle complicated poses. By contrast, detailed reconstruction displays the clothes on the 3D model with simple pose.

3.1 Coarse reconstruction

Currently, there are two routes for the coarse 3D reconstruction from images: optimization based methods and deep learning based methods. Traditionally, the optimization based methods for the coarse reconstruction obtain the 3D models through fitting a human body template to some prior cues extracted from images. The basic idea for the methods is shown in Figure 1.4. Given 2D images, the prior information like joint points and silhouettes is extracted. Then, defining the parametric human body model, the final reconstruction can be obtained through fitting the parametric model to the extracted prior cues.

There are two problems in the methods: (1) prior information extraction and (2) the parametric human body model. The prior information can be extracted manually by experts or automatically through some methods. The common prior information includes 2D/3D joint points, silhouettes, depth image, video and so on. In addition, the research about



(a) 2D images (b) Corase Reconstruction (c) Detailed Reconstruction

Figure 1.3: Given a series of 2D images (a), the goal of the thesis is to obtain the coarse reconstruction of human body (b) or the detailed reconstruction of human body (c).

building the parametric human body model is also an important problem in computer graphics. There are many classic parametric body models during past decades from simple geometric primitives to linear skinning deformation. Recently, learning from some human body datasets to build the parametric human body model is more popular. In general, the parametric model is a set of 3D vertices and faces and is often defined as a function of pose and shape, i.e., the parametric human body model is defined as $M(\theta, \beta)$ where θ is the pose parameters and β is the shape parameters. After defining the $M(\theta, \beta)$, the prior cues of the parametric model can also be denoted as a function of θ and β, i.e., $P_M(\theta, \beta)$. Then, an energy function can be defined based on the prior information of images and the corresponding information of the parametric human body model P_M as:

$$E(\theta, \beta) = L_P(P_M(\theta, \beta), P(I)) + L_\theta(\theta) + L_\beta(\beta), \tag{1.1}$$

where L_P measures the difference between P_M and P. $P(I)$ is the prior cues extracted from given images I. L_θ and L_β are the regularization terms for the pose and shape, respectively. Through minimizing the energy function, the parametric human body model will fit to the observation of the images because P_M will be close to the observation P. The minimization process is to estimate the pose θ and shape β through optimizing the energy function, i.e., the objective function is

$$\{\hat{\theta}, \hat{\beta}\} = \arg\min_{(\theta,\beta)} \; E(\theta, \beta) \tag{1.2}$$

Some existing optimization library can be adapted to implement the above problem. After obtaining the estimated parameters $\hat{\theta}$ and $\hat{\beta}$, the coarse human body model can be produced by $M(\hat{\theta}, \hat{\beta})$.

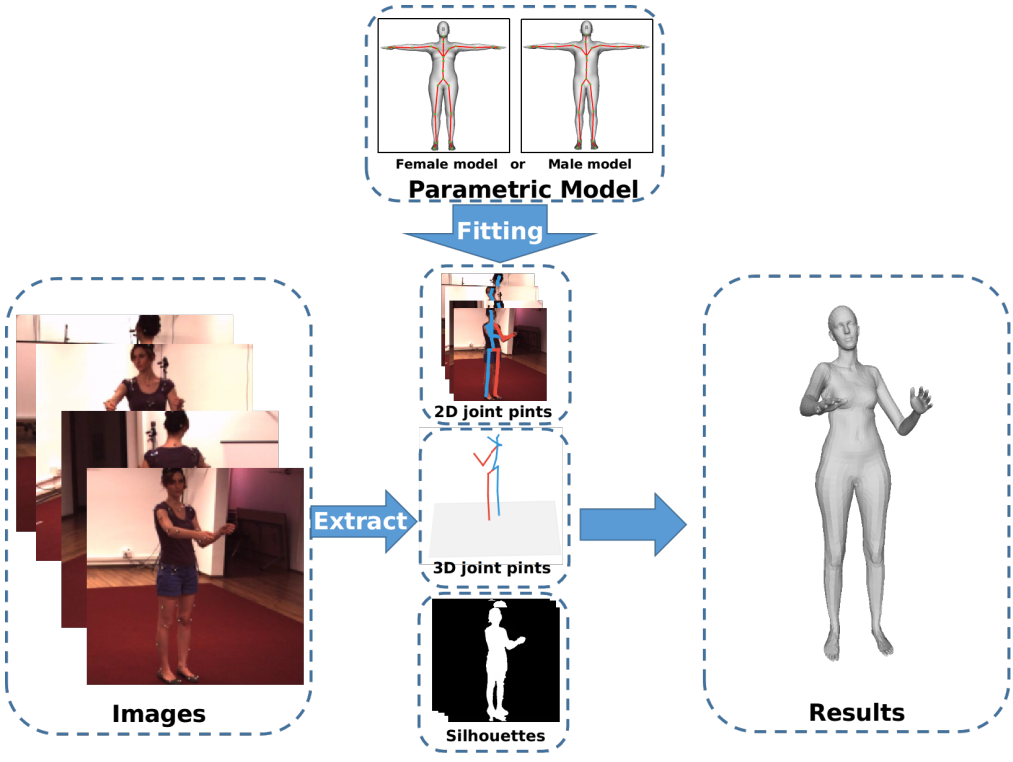


Figure 1.4: Optimization based methods for the coarse reconstruction of 3D human body model from images.

In addition to the optimization based methods, another idea for the coarse reconstruction is to estimate the pose and shape of parametric human body model through Deep Neural Networks (DNN). In pattern recognition and machine learning, neural network is a mathematical structure inspired by the biological neural networks to estimate nonlinear function. In the beginning, neural network only had limited layers due to the restriction of computation and memory of the computer. After 2000, with the development of computer and GPU, researchers started to increase the layers of neural network to depict complex nonlinear function. The classic structure of DNN is shown in Figure 1.5. Currently, DNN, especially for the Convolutional Neural Network (CNN), has become one of the most popular tools in computer vision and has achieved significant success for many tasks. There are many different network structure and many of them have been used for the image classification problem [96, 162, 165, 62]. Here the application of CNN on the coarse human body reconstruction is the main research field. Given a series of 2D images $x_i, i = 1, ..., N$ and its corresponding labels $y_i, i = 1, ..., N$, the goal of a CNN is to fit a function $f(\cdot)$ which takes x_i as input and exports the prediction, i.e.,

$$f(x_i) = Wx_i + b, i = 1, ..., N, \tag{1.3}$$

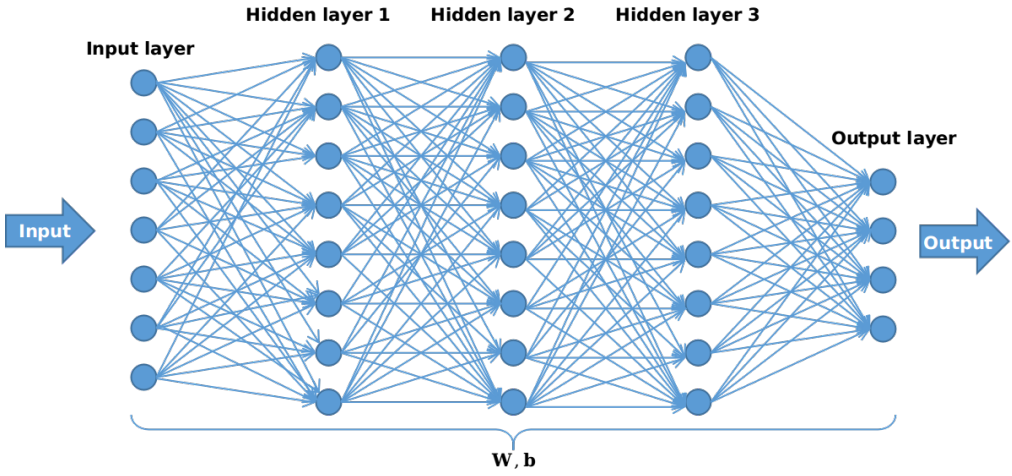


Figure 1.5: An example of deep neural network.

where W and b are the weights and bias of the neurons. For the CNN based methods, the most important task is to estimate the proper W and b. The objective function for the problem is

$$\{\hat{W}, \hat{b}\} = \arg\min_{(W,b)} \; L(Wx_i + b, y_i), \tag{1.4}$$

where $L(\cdot, \cdot)$ is the loss function between the predication and the ground truth. This can be realized through gradient descent algorithm or some improved optimization algorithm. In order to obtain the estimated weights and bias with good generalization, it requires enormous data to train the network sufficiently. In the real implementation of CNN, there are many tricks like nonlinear active function, pooling, normalization and data augmentation to improve the generalization of the learned models. Many open-source machine learning frameworks including Pytorch [37] and Tensorflow [50] have provided good API functions to implement the CNN and to train the CNN, which has promoted the application of machine learning research. As long as proper $\hat{W}$ and $\hat{b}$ are estimated, we can get the estimated results from the input images.

In terms of the coarse human body reconstruction, the goal is to estimate the pose and shape parameters of parametric human body from images. As shown in Figure 1.6, the CNN encodes the input images as the parameters of human body models. Then, the estimated parametric human body and annotations from the training dataset are used for building the loss function. Through minimizing the loss function, the W and b of CNN can be updated to fit a function which exports the parameters of parametric human body according to the input images. Many public datasets for human pose estimation are available now like LSP [76], MPII [9], Human3.6M [72], COCO [107]. These datasets contain various

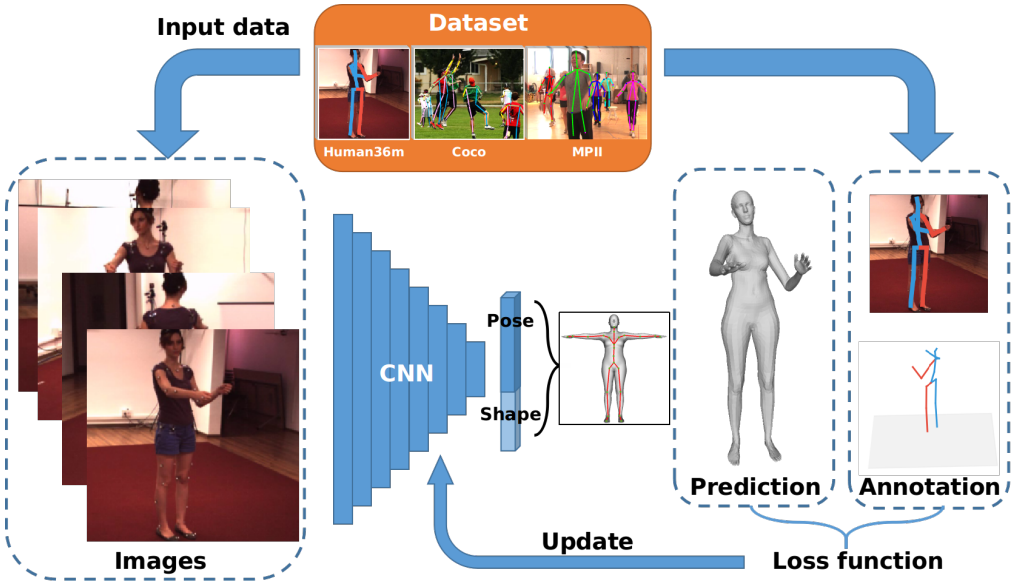
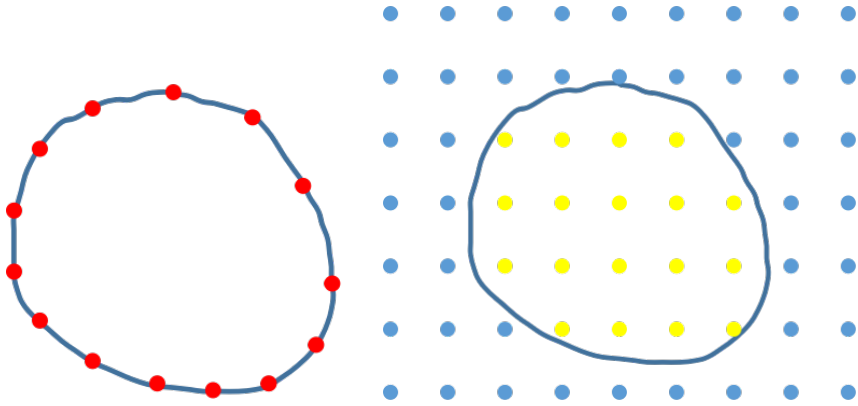


Figure 1.6: Deep learning based methods for coarse human body reconstruction.

poses and scenarios to mainly provide 2D and 3D joint points as annotations for the poses. There are many methods attempting to improve the accuracy of pose estimation through developing novel structure of network or advocating additional cues.

3.2 Detailed reconstruction

For the detailed reconstruction, the goal is to obtain the 3D model with detailed appearance like clothes, wrinkle and so on. This is a more difficult task because both the pose and the appearance of human body need to be considered. Traditionally, methods based on the depth sensor have achieved success on the problem. However, we explore to estimate 3D model from monocular images because monocular images are easier to capture than depth images. As we mentioned above, recovering 3D objects from 2D images is an ill-posed problem, so it is not easy to solve the problem efficiently. Recently, it starts becoming a good idea to estimate 3D objects through learning an implicit function from a deep neural network and the methods based on learning an implicit function are discussed in this section. Implicit function is a function that is defined implicitly by an implicit equation. For example, a surface can be defined by a function $z = f(x, y)$ explicitly. It also can be equally written as an implicit function $F(x, y, z) = z - f(x, y)$, i.e., a surface can be represented by a level set defined by $F(x, y, z) = 0$. Figure 1.7 demonstrates the explicit function and implicit function to represent surface. Considering that the human body has high degree



(a) Explicit representation (b) Implicit representation

Figure 1.7: Two ways to represent surface: Explicit representation VS Implicit representation.

of freedom, implicit function to represent 3D human body has two advantages: (1) easy topology modification; (2) easy point classification on inside or outside.

The implicit function to represent 3D model is often defined as signed distance function or binary occupancy value. Given a 3D point X, the signed distance function is defined as the distance of X to 3D model,i.e.,

$$f(X) = SDF(X) \in \mathbb{R},\ X \in \mathbb{R}^3. \tag{1.5}$$

The sign of $f(X)$ represents that the point lies whether inside of the model $(-)$ or outside of the model $(+)$. For the occupancy value, $f(X)$ is 0 (inside of the model) or 1 (outside of the model). If all the points in a 3D volume are assigned implicit values, the 3D model can be then extracted by the iso-surface through marching cubes algorithm [115]. Considering its good performance to represent 3D model with flexible topology, many researchers attempt to explore 3D shape representation combining implicit function and deep learning recently. The implicit function based on learning has been used for 3D objects reconstruction from images in many research [30, 195]. In order to learn an implicit function from images to represent 3D human body, there are two problems in this process. The first one is feature extraction from the input and the second one is to estimate a proper implicit function to classify the features.

As shown in Figure 1.8, input images are encoded as features by the CNN and a DNN is used for classification of the features. Those features can be corresponding to the pixels which are projected from the 3D points around the ground truth 3D model. For a spatial 3D point X, the corresponding 2D pixel x can be obtained through projected onto image

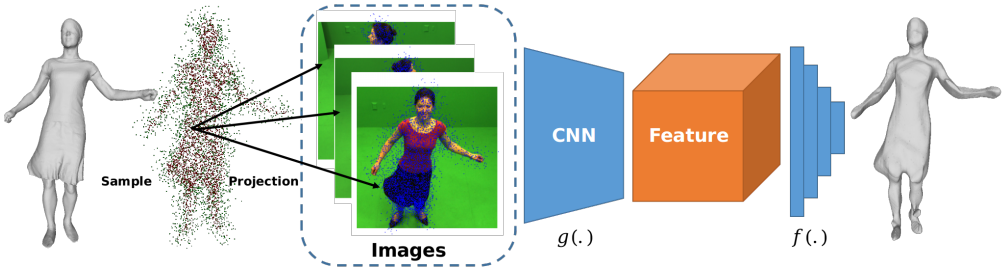


Figure 1.8: The route of learning to implicitly represent 3D human body.

planes. Using fully convolutional network $g(\cdot)$, we can extract a spatially aligned feature map from images. Through training a classifier $f(\cdot)$ to querying the feature, the point can be decided inside or outside of the 3D model. Therefore, the problem is to train a good feature extraction network $g(\cdot)$ and a classifier network $f(\cdot)$, i.e.,

$$\{\hat{g},\hat{f}\} = \arg\min_{(f,g)} L(f(g(x)), o(X)), \tag{1.6}$$

where $o(X)$ is the occupancy value or signed distance of X to the ground truth 3D model. $L(\cdot,\cdot)$ is loss function of prediction $f(g(x))$ and the ground truth $o(X)$. As long as there are enough images and the corresponding ground truth 3D models, it would be possible to train the network to estimate an implicit function with good performance to present 3D human body model. However, the datasets of accurate 3D human bodies are not common. Even though there are some such datasets, many of them are not free and similar datasets are often difficult to capture, which is a challenge for the research.

4 Related work

Image based 3D human body reconstruction is a classic but challenging task in compute vision. Considering 3D human body model is strongly demanded in many applications, there are enormous previous work to tackle the problem during past several decades. In the beginning, the parametric human body model was one of the most important topic in this field because it could provide strong prior information for motion tracking and reconstruction. From simple geometric primitives to learning based models, the parametric human body model are becoming better to represent realistic human body. Since depth sensor became affordable, many researches started to use depth sensor to reconstruct 3D objects because the depth information can provide additional assistance for the reconstruction. With the emergence of deep learning and its significant success in many computer vision tasks, more researchers turned to advocate this tool to explore 3D vision and they made some progress on the problem in the recent years.

In the following, some related work on the topic of 3D human body reconstruction from images is summarized. The work can be divided into three parts: the parametric human body model, the model-based methods for human body reconstruction including optimization based methods and regression based methods, and the model-free methods for human body reconstruction.

4.1 Parametric human body model

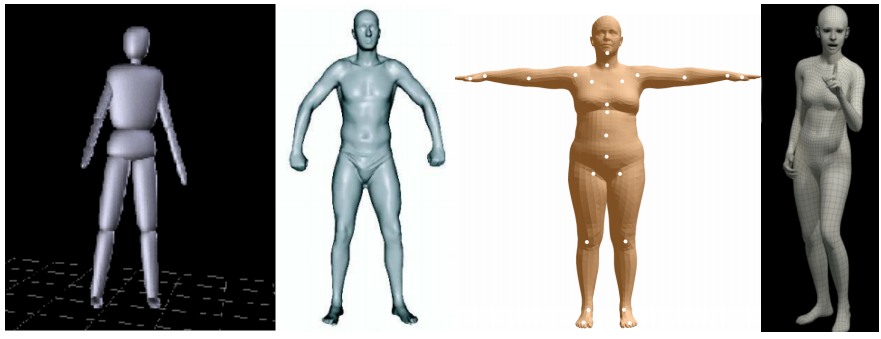
As the above description, human body in real scene demonstrates high freedom of poses and appearance. Therefore, defining a parametric human body model as template is useful for many tasks like motion tracking, 2D/3D pose estimation as well as 3D reconstruction. Early work presented human body model mainly through simple geometric primitives, for instance, cylinder for limbs. Hanavan did a pioneering work in which a mathematical human body model was proposed thorough using 15 simple 3D polygonal shapes [59]. Some 2D human bodies were also proposed based on several 2D ellipses and they had been used for human gait analysis [66, 45, 56]. Besides, 3D human body model based on its kinematic chain received much attention for 3D pose tracking [121, 153, 49]. For more complicated case, some other 3D human body models were formed by using simple sphere [133] or cylinder [121, 153]. Due to the limitation of geometric primitives on describing the shape of human body, some complex human body models were introduced at the same time. For example, Barr *et al.* proposed a synthetic 3D human body through more complex geometric shape called superquadrics [14]. Through introducing local or global deformation on superquadrics, the synthetic model was able to simulate simply flexible motions and there were some applications based on the superquardric model [141, 43, 80, 168, 145].

Driven by the quest of real applications like animations, researchers sought more realism for modeling the human body instead of only using geometric primitives. Artist-driven models [21] were developed and the models could represent simple bones, muscles, and skin of human body. However, obtaining the models was time-consuming and computation-consuming, which limited the application in the early time when the performance of computers was insufficient. Then, Linear Blend Skinning (LBS) method was introduced to build parametric human body model with the better realism [98]. The techniques assigned transformations of skeleton of a human body model to the vertices on the skin. The parametric human body model produced by LBS could depict the deformation and blends of the skin, which produced better realism. Since than, some improved skinning methods were proposed to produce better human body models including dual quaternion blend skinning, spherical skinning, etc. [124, 84, 85, 183]. Although LBS methods and its improvements were used widely, the deformation at joints of the model was still suffering from some unrealistic effects.

Another route to produce 3D human body model through learning from large datasets of

real 3D human scans received more attention after 2000. The basic idea of the method was to deform a template shape to match as many as human bodies with various pose and shape from a dataset of 3D scans through learning algorithm like PCA. The pioneering approach was proposed by Kakadiaris *et al.* [79] who built a body model through identifying human body parts from a list of deformable models based on three orthogonal views. Another important work was proposed in [6] where range data of a human body with 250 different poses was used for learning to model the upper body motion including the torso and arms. They also extended their work to the whole human body [7] for shape variation. The output of the above two models was the global positions of vertices in space, which cannot represent the surface deformation. In [10], a method call SCAPE was proposed to model human pose and shape from many dense 3D scans of different persons with different poses. The pose and shape deformation models were learned from the set of 3D scans, which was able to synthesize muscle deformation automatically to obtain more realism. The learned models can represent pose and shape based on mesh triangle deformation rather than vertices position, which attracted more attention. Then, some improved methods based on learning from enormous 3D scans were proposed to overcome some visual artifacts of SCAPE [184, 61, 40, 65, 23, 143]. In [61], 550 3D scans of full body were utilized to learn a unified model that described both human pose and body shape. Hirshberg *et al.* [65] proposed a BlendSCAPE in which each mesh triangle face in the model was a linear combination of different body parts. Chen *et al.* [23] used tensor decomposition to build a model preserving the dependency between body shape and pose. Aiming at the motion tracking of the human bodies with soft-tissue deformations, a model called DYNA was built based on SCAPE [147] from more than 40, 000 scans of 10 subjects. DYNA predicted the soft-tissue deformation through learning a second-order human body model. In order to make the model consistent with existing animation software, Loper *et al.* [113] proposed a popular statistical human body model called Skinned Multi-person Linear (SMPL) model. This model was learned from CAESAR dataset [152] which had more than 2000 3D scans for each gender. The novelty was that the pose-dependent blend shapes of this model were linear function of pose rotation, which enabled to learn the model from a large number of 3D scans. They also extended SMPL to model the soft-tissue deformation and formed a model called DMPL. For the real-time application of human body with soft-tissue, Bogo *et al.* [19] proposed a method called Dynamic FAUST. Considering its advantage, SMPL model had been widely used for motion tracking, 3D pose estimation, shape recovery etc. Considering SMPL model only focused on the body shape and pose motion, Romero *et al.* [154] proposed SMPL-H which integrated hands motion into SMPL to model the hand gesture. Furthermore, face expressions were considered in SMPL-X [139] to produce better realism. Since the above models were learned from dressing-less 3D scans, they cannot model a person with various clothes. In [148], the authors segmented part clothes and added the to SMPL, but the clothes had no deformation with the variation of pose, which affected its realism in real scene. It had been a popular method to integrate the learned clothes model by DNN and human body to simulate human body with dressing [91, 54, 136]. Be-

sides, generative models on 3D meshes started attracting attention. For example, a novel method called CAPE was proposed to learn the clothing deformation according to the pose of SMPL [116]. In Figure 1.9, several parametric human body models are shown including superquadric model [43], SCAPE [10], SMPL [113] and SMPL-X [139].



(**a**) Superquadric model (**b**) SCAPE (**c**) SMPL (**d**) SMPL-X

Figure 1.9: Some parametric human body models: (a) Superquadric model; (b) SCAPE; (c) SMPL and (d) SMPL-X.

4.2 Model-based methods for human body reconstruction

The parametric human body model could provide strong prior information for the estimation of 3D human body. Therefore, there are many approaches based on the parametric human body to tackle human-related problems including human motion tracking and 3D human body reconstruction. Since the problem of 3D human body reconstruction from images is the main topic of the thesis, the model-based methods for human body reconstruction are summarized in the section. In general, two routes for the problem can be summarized for the 3D human body reconstruction: optimization-based methods and regression-based methods. Traditionally, it is popular to fit the parametric human body to some cues like joint points through minimizing a cost function, which belongs to the optimization based methods. With the successful application of deep learning in many computer vision tasks, more researchers attempt to estimate the 3D human body from images through regressing the parameters of parametric human body, which belongs to the regression based methods. The survey on the two routes are presented in the following sections.

Optimization-based methods

Early work for 3D human body reconstruction was implemented through fitting the parametric human body model to prior information like silhouettes and joint points because of the advantage of the parametric model. In the beginning, many methods based on the parametric human body and optimization were solving the problem of motion tracking [32, 144, 67]. In terms of the shape estimation, early work mainly solved it for the human body with fixed pose. They estimated the human body shape through fitting the parametric human body to the silhouettes of single or multi-view images. The reason was that the parametric human body models at early time were often too simple to recover the complicated pose deformation[64, 51], which also resulted in that the estimated 3D human body models were still very coarse. After the SCAPE model which could represent human body with better realism was proposed, Balan *et al.* [13] proposed to utilize the SCAPE as human body template to track human motion in 3D space. They firstly extracted skeleton from the silhouettes and initialized the SCAPE with the skeleton. Then, they fitted the initialized SCAPE model to the silhouettes extracted from images. This method could handle rich pose and achieved a 3D model with better realism. Similar idea was used for estimating naked 3D human body from images with clothed human [12]. Besides, Guan *et al.* [53] introduced manually labeled 2D joint points as well as silhouettes to build energy function and fitted the SCAPE model to these cues through minimizing the energy function. With the emergence of consumer depth cameras, for example, Kinect, depth images were also used as a prior cue to build energy function for fitting the SCAPE model [186, 17, 108]. Bogo *et al.* [17] proposed an improved SCAPE model called Delta and used this model to fit joints and silhouettes and refined the appearance and remove displacements of depth image through optimization. In addition, many methods also used pre-scanned human body model from laser sensors as a template to track the human body motion [31, 42]. For instance, the authors used segmented image and estimated skeleton to track motion through fitting an articulated template in [42]. Aiming at the problem of optimization based on pre-scanned model, the novel non-rigid optimization algorithm was proposed [101, 100] to achieve better fits. In these methods, the template was non-rigidly registered to the observed 3D scans to track the human body model or some other nonrigid objects. Since the paper [90] was published, deep learning started to play an important role in many computer vision tasks. Human pose estimation and semantic segmentation based on deep learning achieved the state-of-the-art performance [171, 142, 109, 22]. An automatic method for 3D human pose and shape estimation was proposed in [18]. The 2D joint points were autometically estimated by Deepcut [142], and then, the SMPL model was fitted to the 2D joint points to get the 3D human body model. The output of this method was deformed SMPL model and it was a flexible method. Based on this method, silhouettes,3D joint points from depth image, and multi-view images were added to improve the results in some methods [68, 104, 103]. Alldieck *et al.* [3] proposed a novel method to unprojected the SMPL model to improve the shape estimation. They built an energy function between

multi-view silhouettes and unprojected SMPL model to fit the shape. This method can estimate some simple clothes of human body because the unprojected SMPL model relaxed the constraints of the original SMPL model. In [78], the key points of face, hand and body were used and an unified deformation model called "Frank" was fitted to build a 3D models including facial expression, hand gesture and body motion. Besides, there were some other methods in which pre-scanned template was used instead of SMPL model [196, 58]. Xu *et al.* [196] automatically obtained the 2D and 3D joint points and silhouettes though deep neural networks firstly. The skeleton of the pre-scanned model was extracted and was fitted to both the estimated 2D/3D joint points and silhouettes. The method called Livecap [58] tracked the accurate 3D human body in real-time based on single-view RGB video. In general, optimization-based methods rely on the parametric human body model which is used for fitting to the limited information extracted from images such as joint points and silhouettes to reconstruct 3D human body. It is a classic and useful route for human body tracking and reconstruction during past decades. Although the above methods have achieved some success, they are heavily dependent on the accuracy of prior information, which limits its application in practice.

Regression-based methods

Deep learning techniques have achieved great progress to tackle many tasks in computer vision and image processing such as classification [90], detection [47], segmentation [109] etc. In the fields of human-related tasks, 2D or 3D human pose estimation is one of the most impressive application of deep learning. There have been enormous work on 2D and 3D human pose estimation based on DNN [171, 185, 117, 119, 20, 193]. The work in [171] was the pioneer idea to estimate 2D joint points based on CNN. Now these methods have achieved great progress on many public datasets for human pose estimation. Besides, previous work on human body segmentation including full and partial body also had impressive results [130, 190, 63] based on deep learning. Therefore, the impressive performance of the deep learning on human pose estimation and segmentation can automatically provide prior information like joint points and silhouettes. The deep learning can be used for predicting prior cues for the optimization-based methods [18, 3, 78, 196]. However, compared to the whole images, these cues are still quite sparse. Therefore, many researchers have started to use the DNN, more specifically, CNN to regress the human body pose and shape parameters directly from the whole images. At early stages after the emergence of deep neural network, silhouettes were often used for defining the loss function to learn the human body shape through CNN. In [33], a CNN model regressed the silhouettes to learn human body shape based on SCAPE. There was improved method in which the augmented silhouettes with different scales and multiple views were used for cross correlating the model shape [35]. However, the methods in [33] and [35] only handled human body with very simple pose like "A" pose, and thus, they cannot be used for more complicated cases. In [166], the

authors proposed to infer parameters of SMPL model through silhouettes re-projection loss. This work increased the flexibility of the pose. Lassner *et al*. [92] used 91 2D joints obtained by SMPLify and they built loss function based on this to regression the pose and shape parameters of SMPL model. Kanazawa *et al*. [81] proposed an end-to-end 3D human body recovering method. They used CNN to regress the pose and shape parameters of SMPL model from single image. The loss function for training was based on the joint points of image and the joint points of the regressed SMPL model. They achieved impressive performance for 3D pose estimation and 3D human body reconstruction only from one single image. Inspired by this method, some novel methods for 3D human pose and shape estimation based on SMPL model and CNN were proposed in the following years. Image sequence was used in [175] to train the CNN for motion capture. In [131], the idea was similar, but the semantic segmentation of a human body was also used as input of the CNN. Pavlakos *et al*. [138] integrated silhouettes and predicted mesh to improve the shape performance during building the loss function for training. Kolotouros *et al*. [88] put optimzation-based SMPLify into a training loop to form a self-supervised framework. There were still many literatures based on CNN and SMPL to obtain the 3D model through regressing the pose and shape parameters [82, 201, 140, 89, 87]. Since the parametric human body model do not represent appearance like clothes and wrinkles, the above methods can only built coarse 3D human body model without detailed appearance. In order to model the appearance, Varol *et al* [178] proposed a coarse-to-fine method in which the authors firstly regressed SMPL model, and then, used a volumetric representation to fine-tune the details of the appearance of the SMPL model. This coarse-to-fine framework inspired many other approaches [4, 2, 203, 15] . These methods added the appearance to SMPL models to represent clothes. Although the model-based methods have achieved impressive results, they had poor ability to present the details of the appearance because many parametric human body models do not consider the appearance like clothes or hair.

4.3 Model-free methods for human body reconstruction

In contrast to the model-based methods, model-free methods do not depend on the parametric human body model and attempt to directly reconstruct 3D human body model from image based on some other information. Since the human body in real scene always wears various clothes and has different hair style, model-based methods fail to parse these appearance on the parametric model. In order to tackle the problem, model-free methods could be a possible option to recover the detailed 3D human body including the dressing, hair, facial expression, etc. Due to the high freedom, detailed 3D human body reconstruction is quite challenging. Multi-view images based reconstruction, for example, visual hull, was proposed very early because it was an intuitive idea. In addition, depth sensor like Kinect can be used for detailed reconstruction and using depth sensor to reconstruct accurate 3D objects has achieved impressive performance. Recently, benefiting from the deep learning

in 3D vision, there are some approaches which have achieved some success to reconstruct 3D objects from RGB images. The route attracts much attention due to its impressive results. In the following sections, the previous methods based on the above three routes are summarized.

It is an intuitive way to reconstruct 3D human body based on multi-view images. Visual-Hull was a classic concept to estimate 3D objects from multi-view silhouettes [93, 39, 111, 34]. These methods estimated a volume occupancy through fusing silhouettes from multi views and refined the volume by stereo matching or some other optimizations. The 3D model estimated by these methods often had very coarse appearance.

Another route to reconstruct detailed 3D human body is to use depth camera which provides depth information for common RGB images. Although commercial depth scanner [86] had been produced very early and it had been used for building 3D objects, it was hard to be deployed due to its reliance on environments and was expensive for common users. In 2000, Microsoft released a consumer-grade depth sensor called Kinect which is affordable and convenient for many researchers [122]. Since than, many approaches based on RGB-D images captured by Kinect for reconstructing 3D shapes have been proposed. KinectFusion was a classic method to build rigid 3D objects based on Kinect [73, 127]. This method took multi-view RGB-D images and produced the point cloud by Kinect. Then, the Iterative Closet Point (ICP) was utilized to register these multi-view point clouds. Through fusing the point clouds, the full 3D objects or 3D scene can be obtained. Although KinectFusion achieved impressive results, it only focused on the rigid object reconstruction, which was not suitable for nonrigid human body. Aiming at nonrigid reconstruction, many improved fusion algorithms were developed [205, 128, 71, 163]. These methods mainly explored to reconstruct general non-rigid objects. Since human body was an important subject in computer vision, there were also literature only for 3D human body reconstruction [28, 170, 102, 198, 199, 99]. Recent work [105] even can generate detailed 3D human body with extremely loose clothes in seconds based on the Kinect.

Although depth camera is a useful tool to reconstruct 3D objects, it is still inaccessible for common people compared to the RGB images. During past years, learning based 3D reconstruction gained popularity because of the successes of deep learning in 3D vision. In the early stage, the 3D model was represented as explicit volumetric grids. These methods used 3D CNN to operate on voxel grids based on single image [189, 27, 174, 188] or multi-view images [83, 52, 135]. However, the volume resolution of these methods was quite low (e.g. 32^3 or 64^3) because explicit volumetric representation took too much memory. Some improved methods to store the grids as octree structure and they increased the resolution to 512^3 [151, 167, 60]. In addition to the explicit volume representation, generating point cloud from images was also an application of deep learning in 3D vision [38, 1, 57]. These methods mainly focus on general objects and they generated discrete 3D model, which may miss some details on the surface. In order to generate 3D human body with high resolution,

some work proposed a coarse-to-fine manner to solve the problem. They firstly generated discrete voxel grids, and then refined it to a mesh [182, 187, 48]. In contrast to explicit representation or point cloud, implicit function based 3D reconstruction based on deep learning shows some advantages. Implicit function for 3D model is a memory efficient way for the training of CNN and it can produce continuous 3D topology. Instead of storing the x-y-z positions of voxels, 3D model can be represented by the level set which is defined by an implicit function. Based on the idea, some researchers had proposed to reconstruct 3D shape from images through learning an implicit function [120, 134, 24, 26]. In [120], the implicit function was defined by occupancy value, while signed distance function was used in [134]. These methods showed good efficiency on the memory and demonstrated competitive performance on the accuracy of the 3D reconstruction.

Although the above work focus on general simple objects, it still inspired to build 3D human body from images based on deep learning. In [46], the authors fed a coarse 3D model estimated by visual hull into an end-to-end deep neural network to obtain the refined 3D model. Silhouettes were also used in [164] to obtain accurate 3D human body shape. Gabeur *et al.* [41] estimated hidden depth and visible depth from given images through deep neural networks and combined the two depths to generate 3D model. Alldieck *et al.* [5] extracted texture UV maps from images through DeepPose and translate the UV maps to a 3D model. Their results had detailed appearance of human body including the wrinkles of clothes, hair, and facial expression. The approach in [95] had similar idea, but it introduced garment segmentation in the method. In [126], synthetic multi-view silhouettes generated by 2D human pose were used for visual hull to generate 3D model. In [69], a CNN extracted multi-scale features from multi-view images and the features were used for classifying the occupancy value. Satio *et al.* [156] proposed a high-resolution 3D human body reconstruction method from images. They used hourglass network [129] to extract spatial aligned feature grids. They also proposed a refined method by combing the normal of images to improve the details of the appearance [157]. In [132], a novel tetrahedral signed distance function was learned to reconstruct 3D human body from single-view image. Implicit function based 3D human body reconstruction in deep learning is demonstrating competitive detailed 3D human body models from images and it will attract much attention in next years.

5 Contribution

In this thesis both model-based methods and model-free methods are proposed for 3D human body reconstruction from multi-view images. The goal of these methods is to get coarse 3D human body based on parametric human body and to get detailed 3D human body based on learned implicit function. For all of these methods, the inputs of our methods are multi-view images and the output is the estimated coarse or detailed 3D human

body model. In general, the contributions of the thesis are summarized below: (1) Some novel energy functions based on joint points and silhouettes are defined to estimate coarse 3D human body models; (2) An improved architecture collaborating the training of CNN and the multi-view fitting is proposed to use multi-view images to estimate the coarse 3D models; (3) Multi-scale features are extracted to learning the implicit function for detailed 3D model reconstruction; (4) A coarse-to-fine method combining multi-view images based reconstruction and voxel super-resolution is proposed based on learning implicit function; (5) The results of the proposed methods on some datasets achieve competitive performance. The proposed methods are shown from the Chapter 2 to Chapter 6 and the main contributions of each chapter are briefly summarized in the following sections.

Parametric Model-based 3D Human Shape and Pose Estimation from Multiple Views

In this paper an optimization based method for human body reconstruction is proposed. In this method we propose to establish a novel energy function based on the joint points of multi-view images and the parametric human body model, SMPL [113]. During the optimization, the energy function is minimized over through all the multi-view images simultaneously. Not only the pose and shape parameters of SMPL, but also the rotations of multi-view cameras are updated during the optimization. Since the energy function is build based on the multi-view images, the pose and shape estimation of SMPL model can be more accurate and robust. The output of this method is a coarse 3D human body based on the SMPL model. The results of this method achieve better pose estimation on the final 3D human body model than traditional method based on the single-view image.

Joint Points and Silhouette-based Method to Estimate 3D Human Pose and Shape

In this paper a novel method is proposed to estimate the pose and shape parameters of he SMPL model. Both joint points and silhouettes of multi-view images are used to build the energy function for the optimization. Firstly, the energy function for pose fitting is built based on images from sparse views. Afterwards, the silhouettes and the SMPL model after pose fitting are built correspondence in 2D and 3D space, which can be used for defining the energy function of shape fitting. Through minimizing the energy functions of pose and shape fitting, the SMPL model can be fitted to pose and silhouettes and the final 3D model has better pose and shape estimation. Comparing to previous work based on the image sequence, this method can estimate a 3D human body model only from limited number of images and the final model has better shape estimation. The experiments demonstrate that our method can better recover the 3D model from multi-view images than some previous methods.

3D Human Pose and Shape Estimation Through Learning Collaborating Multi-view Model-fitting

In this paper we put optimization based method into a training loop of a CNN model based on multi-view images to estimate 3D human body model. The novelty of the paper is that multi-view images are used during the training instead of only relying on single image. This training strategy can be implemented because some public datasets for human pose estimation are capture from multi-view cameras system. The CNN takes the multi-view images to regress the pose and shape parameters of SMPL model. Then, the regressed SMPL model is used as the initialization for the optimization. During the optimization, the pose, shape as well as the body orientation are optimized simultaneously based on the multi-view images so that the SMPL model achieves the best fit on all the multi-view image planes. Finally, the optimized SMPL model is used for constructing loss functions to supervise the training of CNN. The regression and optimization form a collaborative learning process based on multi-view images, which better uses the relation between the multi-view images. Experiments on several public datasets demonstrate that our method achieves higher accuracy on the 3D pose and the final 3D models also have good shape.

Learning to Implicitly Represent 3D Human Body From Multi-scale Features and Multi-view Images

In this paper a model-free method for 3D human body reconstruction is proposed. This method reconstructs 3D human body model through learning implicit function from multi-view images. In general, the method consists of feature extraction and feature querying. Feature extraction is implemented by multi-stage hourglass networks which can encode the multi-view images as multi-scale features, which is the key novelty of the method. The feature querying is implemented by a fully connected network which classifies the features to decide the corresponding pixel if inside or outside of the ground truth 3D mesh. The multi-scale features encode both local and global spatial information, and thus, can better reconstruct the 3D human body model with details. The experiments on public dataset demonstrate that the proposed method can reconstruct detailed 3D human body model from images and surpasses the previous work.

Detailed 3D Human Body Reconstruction From Multi-view images Combining Voxel Super-Resolution Through Learning Implicit Representation

In this paper we propose a coarse-to-fine method for the detailed 3D human body model reconstruction based on learning the implicit function for 3D representation and voxel super-resolution. The novelty of the method is that the 3D reconstruction and voxel super-

resolution are combined. We firstly use the method in the above paper to estimate 3D human body from multi-view images. Then, the estimated 3D model is voxelized to low resolution voxel grid. Using the low resolution voxel grids as input, we design a multi-stage 3D CNN to extract multi-scale features. Again, a fully neural network is used for classifying the features. The whole 3D model is also implicitly represented during the voxel super-resolution. The voxel super-resolution based on multi-scale features not only preserves the detailed of the appearance, but also removes those incorrect reconstruction of the first step. Therefore, this two steps compose a coarse-to-fine process to reconstruct detailed 3D human body from multi-view images. Experiments on the pubic dataset valid the promising performance of our method. The quantitative and qualitative results demonstrate that our method achieves higher accuracy on the 3D reconstruction than some previous approaches.

6 Thesis outline

The thesis proposes several methods to reconstruct the 3D human body model from images. In Chapter 2, 3 and 4, the SMPL model is used as the template, while optimization-based and regression-based methods are proposed in the three chaptes, respectively. The three chapters mainly focus on the coarse 3D human body reconstruction and the methods achieve good performance on the human pose estimation. In Chapter 5 and 6, learning implicit function is a main tool to reconstruct detailed 3D human body model from images. The experiments on two datasets demonstrates that the proposed methods have better performance.

Overall, this thesis is organized as follows: In the first chapter, the introduction of the thesis is presented including the motivation, challenges, problem statement and the related work summery. Chapter 2 to 6 are the corresponding published papers I to V and we also list the contributions of each paper in the above section. Finally, the conclusion, the discussion about the achieved results and the possible directions of future work are summarized in Chapter 7. The structure of this thesis can be summarized as Figure 1.10.

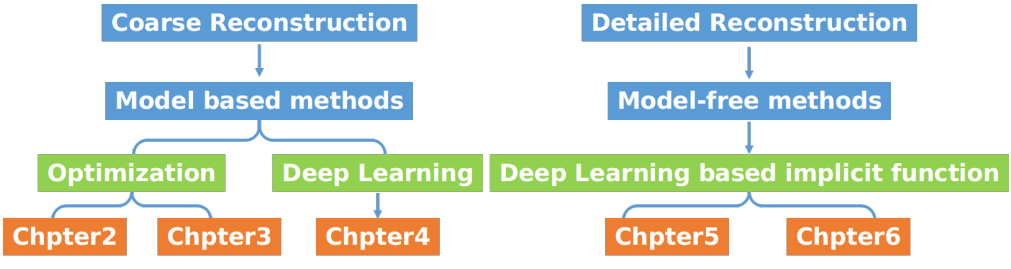


Figure 1.10: The structure of our thesis. The Chapter 2 to Chapter 6 are the main proposed methods corresponding coarse reconstruction and detailed reconstruction.

Paper I

Chapter 2

Parametric Model-Based 3D Human Shape and Pose Estimation from Multiple Views

Abstract

Human body pose and shape estimation is an important and challenging task in computer vision. This paper presents a novel method for estimating 3D human body pose and shape from several RGB images, using detected joint positions in the images and based on a parametric human body model. Firstly, the 2D joint points of the RGB images are estimated using a deep neural network, which provides a strong prior on the pose. Then, an energy function is constructed based on the 2D joint points in the RGB images and a parametric human body model. By minimizing the energy function, the pose, shape and camera parameters are obtained. The main contribution of the method over previous work, is that the optimization is based on several images simultaneously using only estimated joint positions in the images. We have performed experiments on both synthetic and real image datasets, which demonstrates that our method can reconstruct 3D human bodies with better accuracy than previous single-view methods.

1 Introduction

A 3D model of the human body is required in many applications, such as video games, e-commerce, virtual reality, biomedical research, etc. It is, therefore, important to have

robust and accurate methods for recovering models of humans from one or several RGB-images. This is, however, a difficult problem, due to non-rigid motion, different clothing and complex articulation. This makes 3D body reconstruction a very challenging and interesting task in computer vision.

Aiming at effectively acquiring a realistic and personalized 3D human body, many methods have been proposed during the past decades, many using expensive active reconstruction equipment or improving the performance of reconstruction algorithms based on structure from motion methods. Using 3D scanners or multiple calibrated cameras in a controlled environment can obtain 3D models with very high accuracy [77]. The disadvantage of such methods is that these systems are very expensive and relatively complicated to build.

Besides these scanning systems, another line of research is to obtain 3D models from images acquired by ordinary cameras or depth sensors by stereo reconstruction algorithms or fusion algorithms [73, 159, 28]. These methods do not require expensive equipment or complicated set-ups, but are instead based on computationally expensive computer vision algorithms. Structure from motion (SfM) can reconstruct the 3D model of a person with static pose from a moving camera. Using depth sensors, for instance the Kinect, one can also obtain a 3D model through fusion of the geometries obtained from different view-points. These methods do not require any prior information, such as human shape.

Although these ideas have achieved a lot of progress on effectively obtaining 3D reconstructions, there is still a need for simpler methods to reconstruct 3D human body model. With the remarkable progress of human pose estimation based on deep learning, 2D/3D joint points have been the useful information for the reconstruction. Therefore, other methods based on strong prior information are proposed to reconstruct 3D models and have shown good performance. These methods can estimate a 3D human body model from one monocular RGB image by fitting the statistical human body model to the human pose predicted by a DNN [18, 68]. However, only one image is not sufficient to accurately reconstruct 3D models in many cases, due to self-occlusion and complicated articulated motion.

In this chapter a method is proposed to use several (e.g. a sequence of) RGB images which are acquired from different viewpoints to reconstruct the 3D human body based on a skinned multi-person linear shape model (SMPL) [113]. An energy function is defined tp measure the difference between the 2D joint points of the RGB images and the 2D joint points of the projected SMPL model. The 2D joint points of the RGB images are predicted by OpenPose [20]. The difference between our method and SfM-based methods is that we only use the estimated joint positions to reconstruct the 3D model. At the same time, the camera orientations are also regarded as parameters when the energy function is minimized. The advantage is that several images from different viewpoints can provide more accurate 3D information and the number of the images used in our method is in general fewer compared with SfM-based methods. Experiments on synthetic data and Human3.6M [72]

show that our method obtains more accurate pose estimation and 3D shape, than similar methods based on a single image.

2 Related work

As shown in [3], related work is basically divided into two categories: methods that do not use parametric models and methods based on parametric models.

Non-parametric model based methods typically reconstruct 3D models from images acquired by a camera from different viewpoints or from the fusion of depth sensors. The results of the methods can be obtained accurately without using any strong prior information. However, the person should stand still to capture the data and the computation is quite complex and time-consuming. The most well-known algorithm is KinectFusion [73] which creates 3D models in real time by incrementally fusing the partial scans from a moving RGB-D sensor. It has good performance for rigid objects, but is not designed for articulated motion. Therefore, for the 3D reconstruction of a static person, some approaches [159, 28] inspired by KinectFusion are proposed. These methods cannot achieve satisfying result for the dynamic person since the human body typically is moving non-rigidly between different views. DynamicFusion [128], which is the pioneering work for the reconstruction of non-rigid objects, can reconstruct the 3D geometry in real time for a slowly moving person. Other methods such as KillingFusion [163] and BodyFusion [198] are proposed to improve the results based on DynamicFusion. However, these approaches are only suitable for slow motion and have high computational complexity. In order to obtain more accurate results, multiple Kinect sensors or several calibrated cameras can be utilized to create 3D human body models. In [36], the authors propose to use eight Kinects to obtain the 3D model with high accuracy. Multiple cameras are also used in [77, 97] to reconstruct the 3D human body. However, there are technical challenges and it is expensive to build a system with eight Kinects or to build the indoor environment like [77] for many practical applications.

Parametric model-based methods often rely on a template which provides strong prior information during the reconstruction. The template can be reconstructed from depth data or using a pre-computed human body model. In [101, 202, 55], a novel non-rigid registration algorithm is proposed to register a pre-scanned model to other partial depth data acquired by Kinect. In [202], a template is obtained through registering several high quality partial scans and then a personalized 3D model is reconstructed by fitting to the template. Some other algorithms [196, 205] have similar ideas but they use more complicated information or hardware to improve the accuracy and efficiency. Besides pre-scanning the template, a number of statistical human body models have been proposed based on training of a human body set, such as SCAPE [10], SMPL [113] and so on [147]. In [186] the authors use

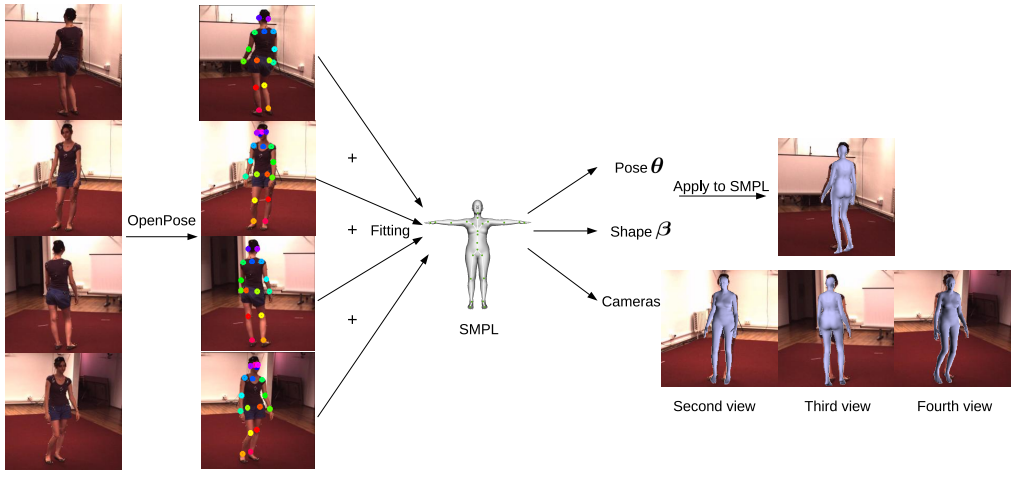


Figure 2.1: The overview of our method.

the SCAPE model to fit the depth image to obtain a 3D model. The improved SCAPE model, Delta, is proposed in [17] and a detailed body reconstruction algorithm is presented in this paper. In [18], the authors propose to fit the SMPL model by using 2D joint points predicted by a DNN-based method. Huang *et al.* [68] use a similar idea but they focus on the video problem, using temporal information. In [81], an end-to-end adversarial learning method is used to estimate the human pose and shape parameters by fitting the SMPL model. Alldieck *et al.* [3] propose an algorithm to obtain the consensus shape and then use both pose and consensus shape to fit the SMPL model in order to obtain better result.

3 Method

The aim of this chapter is to obtain the 3D model of a human body from several RGB images taken from different view-points. Our approach is inspired by the the work in [18] where the 3D human body model is estimated from only one RGB image. Although the method in [18] has achieved some accuracy, the error is still noticeable in many cases since one RGB image cannot supply enough information. As an improvement, we propose to use several RGB images taken from different view-points to reconstruct the 3D model. This leads to a more challenging optimization problem, since the motion of the cameras is unknown, and we need to introduce the parameters of the cameras as variables to estimate. Firstly, we estimate the positions of the 2D joint points of the person in the images through OpenPose. Then, the SMPL model is fitted to the pose of the person in different views by optimizing an energy function in which the camera parameters are included. Finally, the pose, shape and the camera parameters are estimated to obtain the 3D model of the human. The pipeline of our method is summarized in Figure 2.1. In the following, we

firstly introduce SMPL model, then the energy function and finally the optimization that gives the estimation of the camera parameters as well as the pose and shape parameters of the 3D human model.

3.1 SMPL model

The SMPL model encodes both pose and shape parameters [18]. The pose is defined from the parameter θ, which represents the relative rotations of the 23 joint points with respect to the root joint. The shape is represented by the parameter β, which describes the strength of each mode in a shape space obtained from a principal component analysis (PCA) from a registered training set. The pose parameters are represented as a vector $\theta \in \mathbb{R}^{72}$ and the shape parameters as a vector $\beta \in \mathbb{R}^{10}$.

The output of the SMPL model after introducing pose and shape is a mesh with $N = 6890$ vertices and $F = 13776$ faces, $M(\theta, \beta) \in \mathbb{R}^{N\times 3}$. In this model, the 3D joints are obtained by linear regression from the surface mesh vertices, i.e., a function of the pose and shape coefficients. Therefore, the pose and shape parameters can be estimated by optimizing an energy function based on the joint points.

3.2 Energy function

The approach in [18] is called SMPLify, in which the projection of the 3D joints of the SMPL model is fitted to the 2D joint points predicted by a CNN-based method. The advantage of this method is that only one image is utilized to obtain the 3D model. However, one disadvantage of SMPLify is that in some situations one image does not contain enough information for obtaining an accurate 3D reconstruction (due to self-occlusion, articulated motion and ambiguous pose). Other methods based on traditional SfM pipelines, require a lot of images from different views and are computationally intensive. Therefore, we propose to use several images from different views into SMPLify because more images will provide more regularization and it is convenient to not use too many images. The problem of this idea is that the parameters of the cameras from different views are unknown, which makes the projection of the joint points of the SMPL model difficult. The solution to this problem is to use the parameters of the cameras together with the pose and shape of the SMPL model as the variables of an energy function during the optimization. The advantage of this method is that we can obtain not only the estimation of the pose and shape but also an estimate of the cameras parameters (position and orientation).

The energy function contains three parts: the pose-fitting term, the shape parameter regularization term and the pose parameter regularization term. The energy function is defined

as:

$$E(\boldsymbol{\theta}, \boldsymbol{\beta}, R_i) = E_J(\boldsymbol{\theta}, \boldsymbol{\beta}, R_i) + \boldsymbol{\lambda_\theta} E_{\boldsymbol{\theta}}(\boldsymbol{\theta}) + \boldsymbol{\lambda_\beta} E_{\boldsymbol{\beta}}(\boldsymbol{\beta}), \tag{2.1}$$

where $E_J(\theta, \beta, R_i)$ is the pose fitting term, $E_\theta(\theta)$ is the pose parameters regularization term, $E_\beta(\beta)$ is the shape parameters regularization term and λ_θ and λ_β are weights. In the energy function, the pose θ, the shape β and the rotation of the camera R_i can be estimated through

$$\{\hat{\theta}, \hat{\beta}, \hat{R}_i\} = \arg\min E(\theta, \beta, R_i). \tag{2.2}$$

The most important term is E_J in our method and it is defined as

$$E_J(\boldsymbol{\theta}, \boldsymbol{\beta}, R_i) = \sum_{i=1}^{N} \sum_{k=1}^{K} \boldsymbol{\rho}(\boldsymbol{\Pi}_i(J_{S,k}) - J_{2d,k}^{(i)}), \tag{2.3}$$

where N is the number of images, K is the number of joint points, $J_{S,k}$ is the k-th 3D joint points of the SMPL model, Π_i is the i-th camera, $J_{2d,k}^{(i)}$ is the k-th 2D joint point estimated by OpenPose for the i-th image and R_i is the rotation for i-th camera. The error ρ is measured by the Geman-McClure function [44] which gives robustness to large noise and outliers. This function is defined as

$$\rho(x) = \frac{x^2}{\sigma^2 + x^2}, \tag{2.4}$$

where x is the absolute errors of 2D joint points and σ is a constant. The projection of the 3D joint points of the SMPL model in the i-th camera is

$$\Pi_i(J_S) = R_i J_S + t_i,$$

where t_i is the translation of the i-th camera. The translation is calculated separately using the shoulders and hips, which implies that we can assume that the person is standing parallel to the image plane. Because the projection is linear, the derivatives of the error function can be computed easily during the optimization.

The pose regularization is needed for avoiding the knees and elbows bending unnaturally and it is defined as

$$E_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \boldsymbol{\alpha} \sum_i \mathbf{exp}(\boldsymbol{\theta}_i), \tag{2.5}$$

where θ_i denotes the pose of the joint points of elbows and knees and α is a constant that controls the penalization. The shape regularization term is defined as

$$E_{\boldsymbol{\beta}}(\boldsymbol{\beta}) = \sum \boldsymbol{\beta_i}, \tag{2.6}$$

i.e. as the sum of the elements of β.

3.3 Optimization

The optimization is performed in two steps. In the first step the camera translation is estimated. Here the focal length of the camera is assumed to be known. The camera translation can be estimated through fitting the shoulders and hips in the SMPL model and the predicted 2D pose.

In the second step the model is fitted through minimization of (2.1). The parameters λ_θ and λ_β are decreased gradually during the optimization. The minimization method is based on Powell' s dogleg method, which is provided by the python modules, OpenDR [114] and Chumpy [112]. For four-view images with 320×240 size, it takes about 2 minutes for the minimization on a desktop machine.

4 Experiments

In this section some experiments are presented to illustrate the performance of our method. In the first experiment, a small synthetic dataset is generated based on SURREAL [177] in which a large amount of synthetic human bodies with different poses and shapes are created based on the SMPL model. Since the SURREAL only provides videos from one view, we generate three more images from the other views. Then, for the real images, our method is evaluated on the Human3.6M which is for the evaluation of human pose estimation.

In order to quantitatively compare the results, the metric for evaluation is defined as:

$$Error = \frac{1}{N}\sum_{i=1}^{N} ||J_i^{gt} - J_i^{est}||_2, \tag{2.7}$$

where J_i^{gt} is the ground truth of the 3D joint points and J_i^{est} is the estimated 3D joint points. In this part, there are a total of 24 joint points for the SMPL model.

In our experiments, the parameters $(\lambda_\theta, \lambda_\beta)$ decrease as (404,100), (404,50), (58,5), (4.78,1). The σ is set to 100 and α is 10. The maximum number of iterations is 100 for every stage and the stopping criteria is that the error of the energy function is smaller than 10^{-3}. The experiments are implemented in Python and our desktop machine has a 4 core Intel i5-6500 CPU @ 3.20GHz with 8 GB RAM.

Table 2.1: The mean errors of SMPLify compared with our method using respectively 2, 3 and 4 images.

	SMPLify	Ours-2	Ours-3	Ours-4
Mean error	0.0177	0.0113	0.0108	0.00525

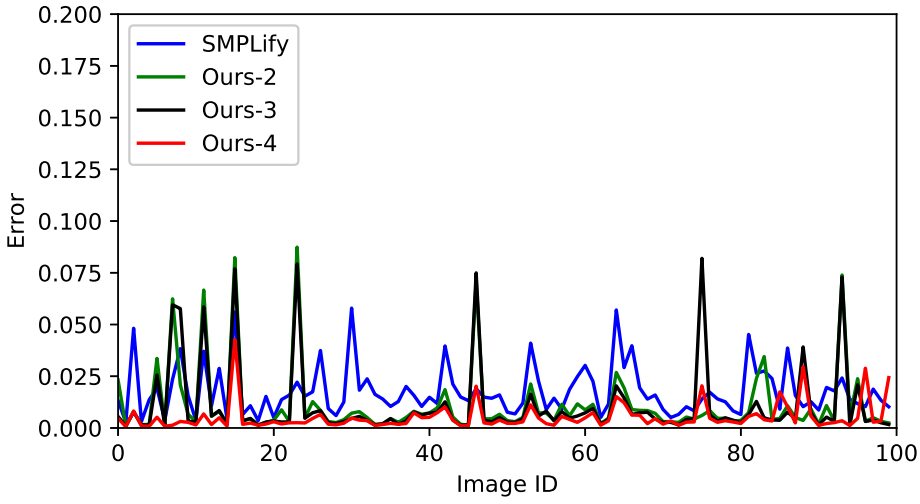


Figure 2.2: The errors of the 100 samples for multi-view images. We compare the results from one image (blue), two images (Green), three images (Black) and four images (Red).

4.1 Results on synthetic data

The synthetic images are generated based on SURREAL. SUREAL dataset is generated by the SMPL model through giving different pose and shape parameters to the SMPL model. It contains more than $10,000$ small videos and each small video consist of 100 frames. In each video, the human body performs a continuous action like walking and jumping. Since this is a very large dataset, it is difficult to evaluate all the image sequences. We decide to extract a small syntheic dataset based on SURREAL. This small dataset consists of the first 100 videos from the training set of SURREAL. Considering that some videoes in SURREAL have the same pose and shape parameters, we only take the first 100 videos who have different poses and shapes. Only the parameters in the first frame of the 100 videos are taken out. We utilize 100 pose and shape parameters from the training data of SURREAL into the SMPL model to generate 100 different 3D human bodies. Then, four images whose sizes are 320×240 are rendered by cameras from different view-points for each human body model. The multi-view images have black backgrounds and white body, which can be seen in Figure 2.4. By doing this, we can also get the ground truth of 2D joint points from all views. In following experiments on the synthetic dataset, we implement our method to get the 3D models from two, three and four views respectively. As comparison, we use the method called SMPLify in which only one single image is used.

The errors for the 100 samples and the mean error using different number of images are given in Figure 2.2 and Figure 2.3, respectively. Figure 2.2 is the errors of the 100 examples

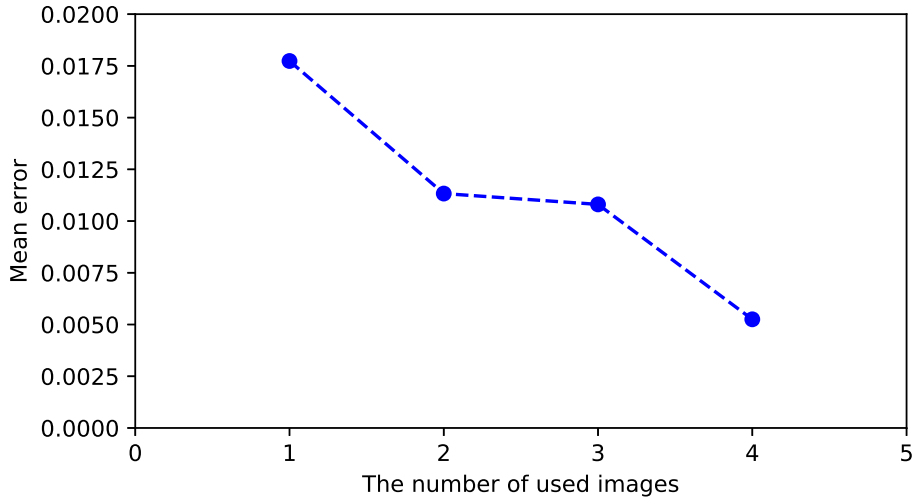


Figure 2.3: The mean errors of the 100 samples for mulltiview images.

when different number of images is utilized, while Figure 2.3 shows the mean errors of the 100 examples when the number of used images is different. It is shown from Figure 2.2 that the error is smaller when more images are used in the method for most samples. In some cases, the error of our method with two or three images is greater. This is because images from two different views may influence each others since the camera at the side position cannot capture all of the joint points. Besides, we can see from this figure that the results of our method are more stable. For the Figure 2.3, it clearly illustrates that the mean error decreases when more images are used. The images from more views can provide more information for the estimation and this will reduce the ambiguity of the 2D joint points in the image planes. In general, the estimation will be more accurate if multi-view images can be acquired.

The mean error of the 100 samples is also given in Table 2.1. We can see that the mean error decreases when more images are utilized and that the performance of our method surmounts that of SMPLify, which shows that more images indeed can provide more useful information. The mean error of SMPLify on the small synthetic dataset is 0.0177, while the mean errors of our method from 2, 3, and 4 images are 0.0113, 0.0108 and 0.00525, respectively. This table is consistent with Figure 2.2 and Figure 2.3.

In order to better demonstrate the results, Figure 2.4 gives qualitative results from the synthetic dataset. In this figure, we only give the results of SMPLify and the results obtained by our method from four-view images. We give four examples in this figure and the white bodies in each row are the multi-view images used in our method. Since SMPLify only uses

Figure 2.4: The figure shows results on synthetic data. Each row corresponds to one person with some unknown shape and pose. For each row the left hand image in each column is the input image for frame one to four. The middle image in column one is the result from SMPLify using only the input image from frame one. The right hand image in each column is the result of our method using all four input frames.

one single image, only the first view image is used for comparison. The middle image in column one is the result from SMPLify using only the input image from the first view. The right hand image in each column is the result of our method using all four input frames. We can see from the first image that our method has better performance, especially the last one. The images from other views show that the estimation of the cameras by our method is very correct, which demonstrates the effectiveness of our method.

4.2 Results on Human3.6M

Table 2.2: The mean errors of SMPLify and our method based on four views for the eight actions of S1.

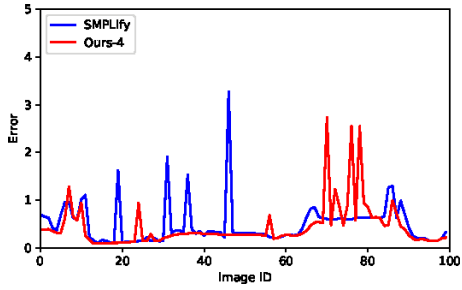
	Directions	Discussion	Eating	Greeting
SMPLify [18]	0.4866	0.4205	0.7401	0.6624
Ours-4	**0.4281**	**0.3048**	**0.6952**	**0.3357**
	Phoning	Posing	Purchasing	Sitting
SMPLify	0.7270	0.6746	0.5410	0.4784
Ours-4	**0.4144**	**0.3938**	0.5471	**0.3470**

Table 2.3: The mean errors of SMPLify and our method based on four views for the eight actions of S6.

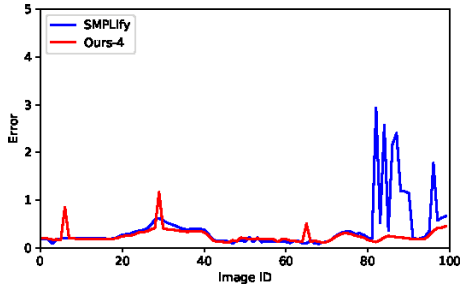
	Directions	Discussion	Eating	Greeting
SMPLify [18]	0.4880	0.3870	0.3597	0.4895
Ours-4	**0.4280**	**0.2416**	**0.3114**	**0.4395**
	Phoning	Posing	Purchasing	Sitting
SMPLify	0.4543	0.5815	0.5530	0.4291
Ours-4	**0.2711**	**0.4203**	**0.3372**	**0.2891**

There are total of 11 subjects (6 males, 5 females) in Human3.6M and every person has 15 actions. In order to test our method sufficiently, we choose S1 which is a female and S6 which is a male to evaluate our method on 8 actions: Directions, Discussion, Eating, Greeting, Phoning, Posing, Purchasing and Sitting. For each action, we sample the video every five frames and take total of 100 frames. The results of SMPLify and our method with four images are compared. The metric for the comparison is also computed according to (2.7). In this part there are 16 joint points because the number of joints in Human3.6M is 16. Similarly, the errors of every frame in the different actions for S1 and S6 are shown in Figure 2.5 and Figure 2.6. The mean errors of the 100 frames in each action for S1 and S2 are shown in Table 2.2 and Table 2.3. It is shown in these results that our method can obtain more accurate estimation in most cases.

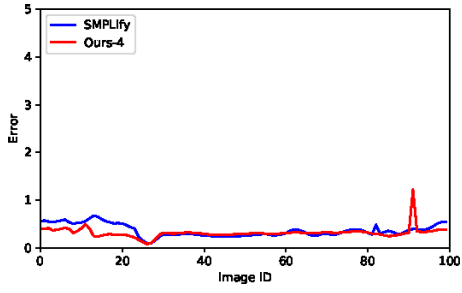
In addition, some images from the dataset are shown in Figure 2.7 and Figure 2.8. We can see from Figure 2.7 and Figure 2.8 that SMPLify has obvious errors such as the occlusion of the arms and bodies. Our method sometimes also have unexpected errors because of having a side-view such as the first sample in Figure 2.8. The reason is that our method relies on all of the observed images. Therefore, if one camera translation is not estimated correctly, it will affect the images from other views, and then, the final results may be incorrect after the optimization. However, SMPLify only uses one image and if this image is not captured from the side view, the result is sometimes better than ours. In general, our method can achieve better estimation than SMPLify.



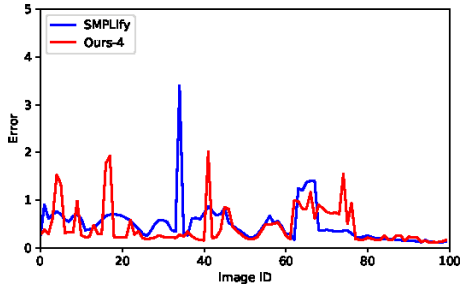
(a) Directions



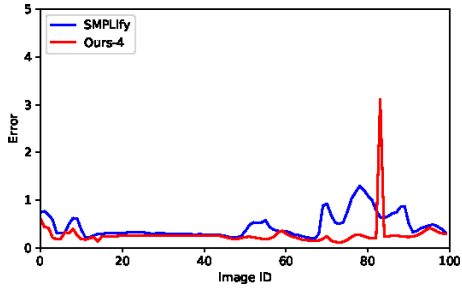
(b) Discussion



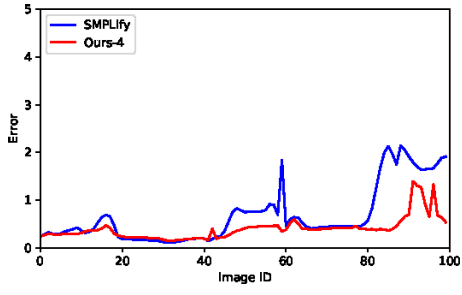
(c) Eating



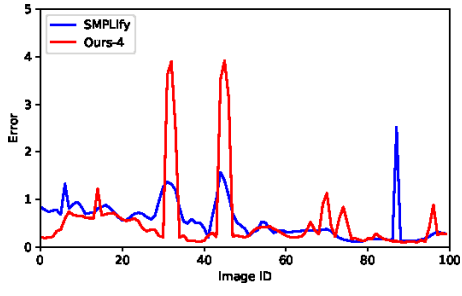
(d) Greeting



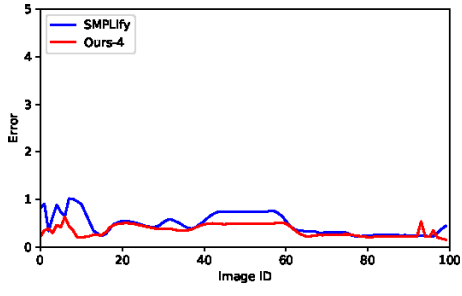
(e) Phoning



(f) Posing

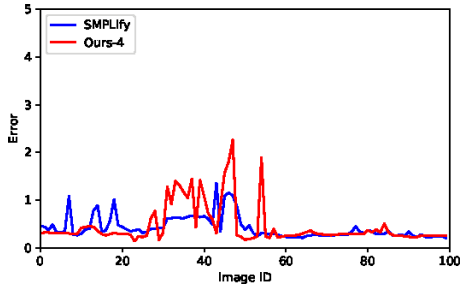


(g) Purchasing

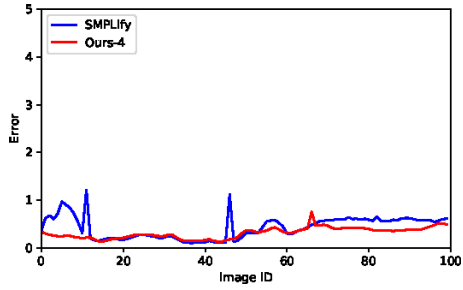


(h) Sitting

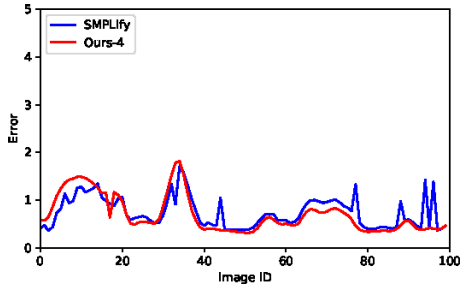
Figure 2.5: The errors of every frame of the eight actions for S1.



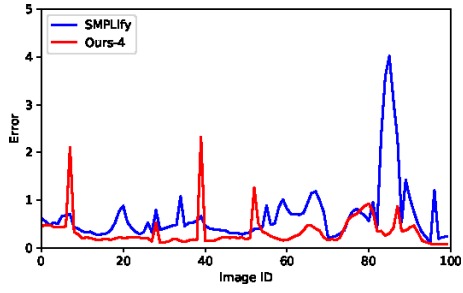
(a) Directions



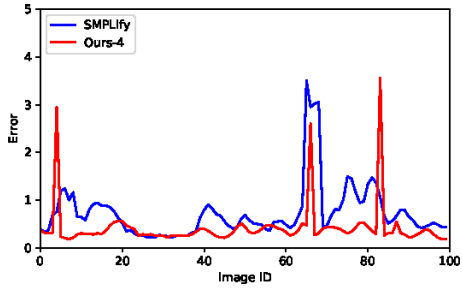
(b) Discussion



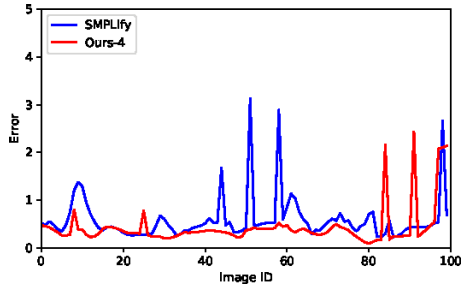
(c) Eating



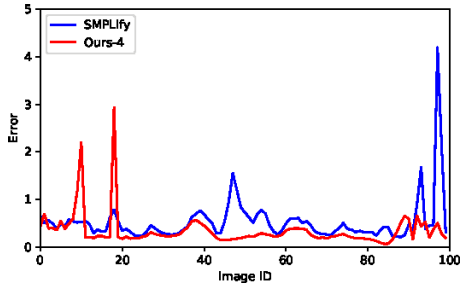
(d) Greeting



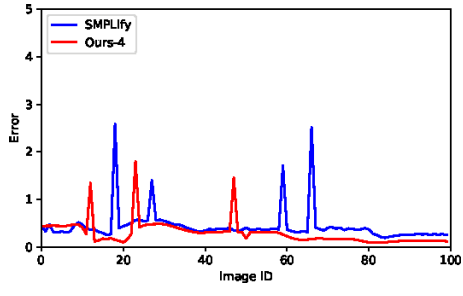
(e) Phoning



(f) Posing



(g) Purchasing



(h) Sitting

Figure 2.6: The errors of every frame of the eight actions for S6.

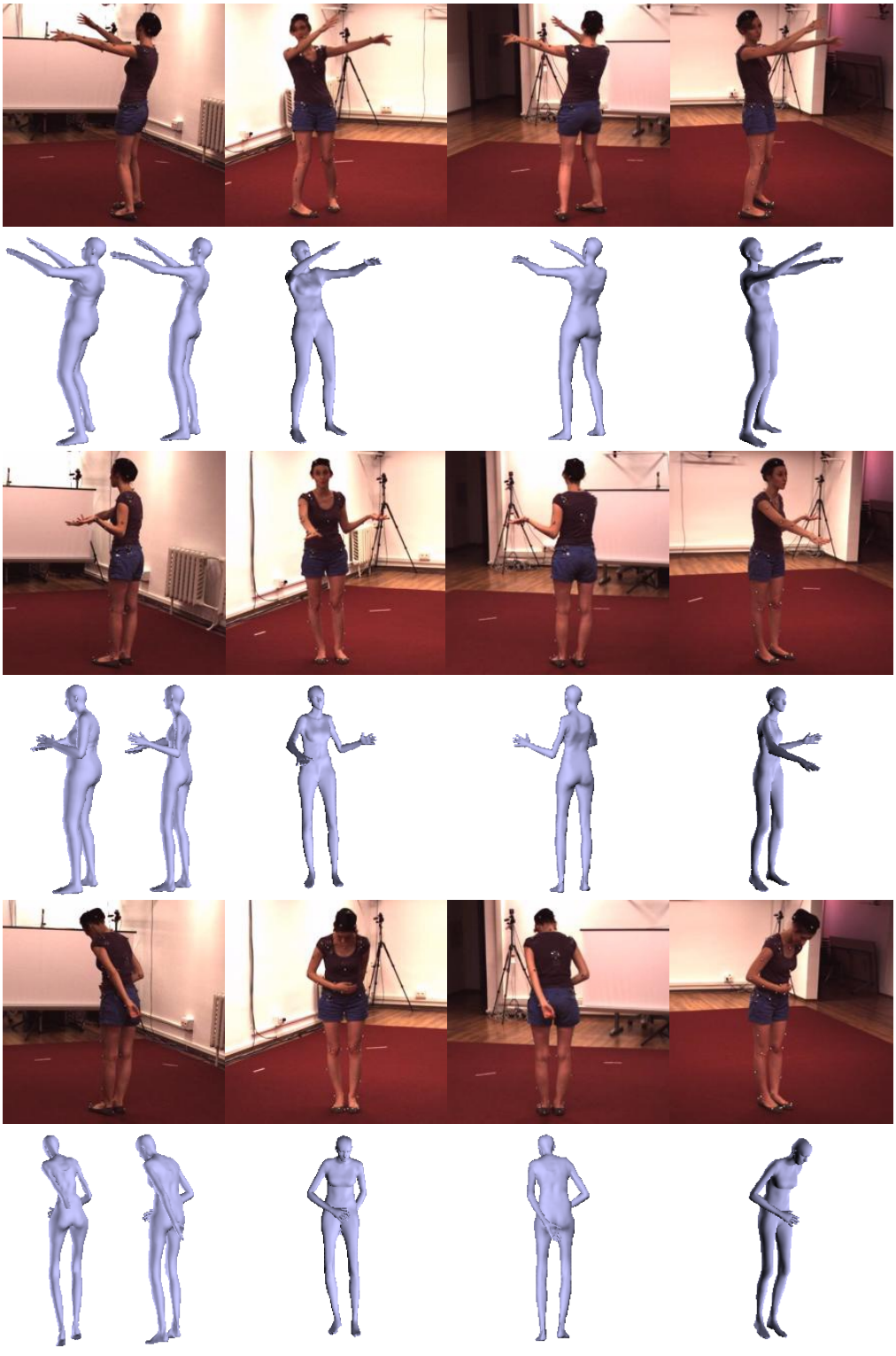


Figure 2.7: Some samples from S1. The first images are given the results of SMPLify and our method from left to right.

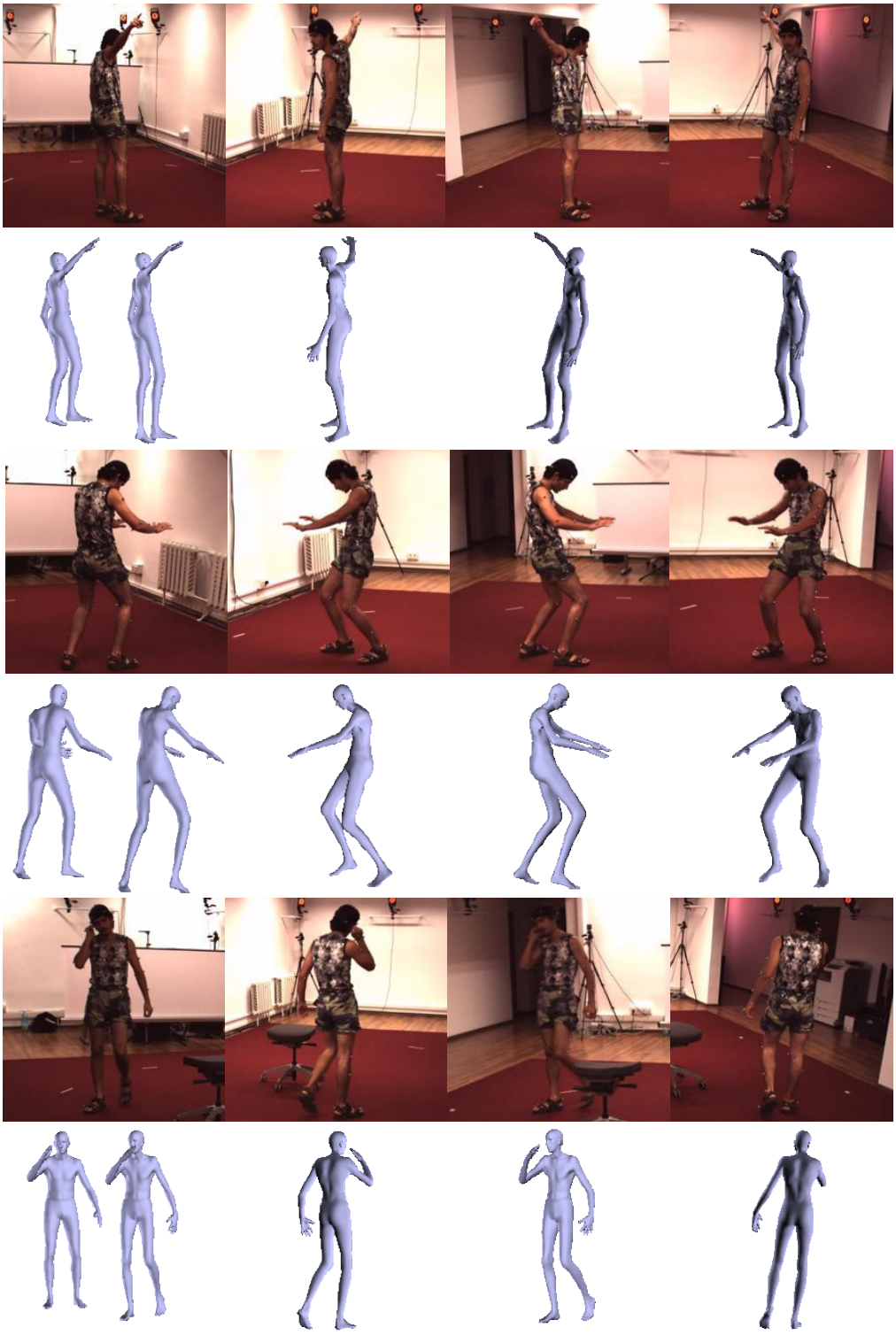


Figure 2.8: Some samples from S6. The first images are given the results of SMPLify and our method from left to right.

5 Conclusion

We have proposed a method to reconstruct a 3D human body model from several RGB images taken from different view-points. Our approach starts by estimating the 2D joint points of the images by using a DNN-based method called OpenPose. Then, a statistical human body model, SMPL, is utilized to fit the predicted 2D joint points from the images by minimizing an energy function over all images simultaneously. Finally, our method estimates both the pose and shape parameters of the human body as well as the camera parameters. Experiments on synthetic and real data quantitatively and qualitatively demonstrate that the results of our method are better regarding the pose error compared to the previous method based on only one image .

Our method also has some limitation. If the images are captured from the side view, the joint points will be very close to each other or even at the same position, which makes our method unstable. Also, we mainly focus on the estimation of the pose and this implies that the shape of the reconstruction is less accurate. However, this is a fundamental limitation of all methods that only use the joint positions and disregard the contours of the body.

Paper II

Chapter 3

A Novel Joint Points and Silhouette-based Method to Estimate 3D Human Pose and Shape

Abstract

This paper presents a novel method for 3D human pose and shape estimation from images with sparse views, using joint points and silhouettes, based on a parametric model. Firstly, the parametric model is fitted to the joint points estimated by deep learning-based human pose estimation. Then, we extract the correspondence between the parametric model of pose fitting and the ground-truth silhouettes in 2D and 3D space. A novel energy function based on the correspondence is built and minimized to fit the posed parametric model to the silhouettes. Our approach uses comprehensive shape information because the energy function of silhouettes is built from both 2D and 3D space. This also means that our method only needs images from sparse views, which balances data used and the required prior information. Results on synthetic data and real data demonstrate the competitive performance of our approach on 3D pose and shape estimation of the human body with medium detailed appearance.

1 Introduction

Estimation of 3D human body models from images is an important but challenging task in computer vision. In many practical fields, for instance, video games, VR/AR, E-commerce

and biomedical research, 3D human body models are needed and play vital roles. However, the human body in real scenes naturally exhibits many challenging properties, such as non-rigid motion, clothes and occlusion. These factors make it difficult to accurately and efficiently estimate the 3D human body model from images, and many approaches have been proposed to obtain 3D human body models during the past decades.

Time-of-flight cameras, and other types of hardware solutions, can provide depth information and have been one of the solutions to the reconstruction of 3D human bodies [73, 128, 163, 199, 194]. More specifically, depth cameras are utilized to capture RGB images and the corresponding depth images of the scenes. The 3D meshes of each view can be computed from the RGB-D images and the complete 3D model can be estimated by fusing the 3D meshes of each view. The process of fusion is often implemented using the Iterated Closest Point (ICP) algorithm [73] or other similar improved algorithm, which are often computation-consuming. Since these methods only can handle rigid scenes well, research on dynamic scenes has been explored [128, 163, 199, 194]. However, compared to ordinary cameras, the cameras with depth sensors are still expensive and calibration of the depth camera can also be complicated.

With the development of deep learning architectures, 3D human body models can be estimated by optimization- [18, 3] or regression-based methods [81, 88]. For these methods based on optimization, prior information, for example, human poses and silhouettes can be estimated by deep neural networks. The 3D model can, then, be obtained by fitting the parametric human body model to the prior information. The regression-based methods use deep neural networks, for example, convolutional neural network (CNN), to directly estimate the parameters of the given parametric human body model from images, by training the deep neural networks[81, 178, 88]. Both approaches have been explored extensively and have achieved good performance in 3D human body reconstruction. However, regression-based methods require a large amount of data to train the neural network. This often requires much work and it is difficult, and sometimes expensive, to generate the dataset. Compared to regression-based methods, the human pose estimation and semantic segmentation based on deep neural networks have been well developed and many pre-trained models can be utilized directly. This means that prior information in optimization-based methods can be more easily estimated through deep neural networks. For these reasons, our choice of method, proposed in this paper, is also optimization-based.

In this paper, the goal is to estimate the 3D human body from images. Since this is a very complex problem and one single image can only provide limited prior information, a number of images taken from different view-points are used in our paper. The human pose estimation based on deep neural network [192] is adopted to estimate the joint points of the human body in the multiple-view images. The Skinned multi-person linear model (SMPL) [113], which is widely used in the methods based on optimization, is the parametric human body model also used in our paper. Then, an energy function is established based

on the predicted joint points and the joint points of the SMPL model. By minimizing the energy function, we can achieve an estimated 3D human body model which has a pose consistent to the observed images. In addition, the silhouettes are exploited to improve the shape of the estimated human body model. Through building the correspondence edges between the estimated human body model and the given silhouettes from 2D and 3D space, the energy function for the silhouettes is constructed. The shape parameters of the human body model are obtained by optimizing the energy function. Therefore, the final 3D human body model is generated by the estimated pose and shape parameters after pose fitting and shape fitting. The experiments on synthetic data and a public real dataset validate the performance of our method. The 3D models obtained by our method have medium detailed appearance.

In summary, the contribution of our method consists of two parts. Firstly, an improved energy function for silhouettes is constructed from 2D and 3D perspectives to estimate the parameters of shape. Secondly, a small number of images (four in our experiments) from different views are applied in our method, which balances the number of images and the prior information.

The article is organized as follows: Section 2 contains a review of related work. Section 3 presents the idea of our method. Then, the experiments on synthetic and real data are described in Section 4. Finally, Section 5 concludes the whole work and discusses possible future work of this paper.

2 Related work

In order to obtain 3D human body model from images, researchers have explored a lot of methods from hardware and software during the past decades. These work can be basically categorized according to whether a parametric human body model is adopted in the methods. For the approaches which do not depend on any parametric human body model, the 3D reconstruction of human body is mainly implemented from RGB-D images captured by depth cameras. In contrast to the above methods, the approaches based on a parametric human body model often attempt to estimate the pose and shape from common RGB images. We call the two categories *parametric model-free* and *parametric model based* methods, respectively.

Parametric model-free methods often reconstruct 3D human body models from RGB-D images, which means that these methods often require depth cameras. KinectFusion [73] was the typical work which used a Kinect depth camera to reconstruct the 3D meshes of an indoor scene with static objects. However, KinectFusion was mainly aiming at reconstructing rigid objects rather than dynamic scene like a moving human body. In order to tackle

non-rigid reconstruction, DynamicFusion [128], VolumeDeform [71], KillingFusion [163] were proposed over the next several years. These methods can handle reconstruction of non-rigid and moving objects, but they typically only obtain good performance for partial body or small slow moving objects. Yu *et al.* proposed BodyFusion [198] and DoubleFusion [199] to reconstruct the whole 3D human body model for moving persons with high accuracy. One common thing in all of the above work is that they utilize one single Kinect to recover the 3D human body model. In order to improve the accuracy more, some methods based on multiple Kinects [197, 36] were proposed to reconstruct 3D geometry, which was more complicated to set than single a Kinect. Besides, commercial depth cameras [194] was also a tool to reconstruct 3D models. The core idea of the parametric-model free methods is that they utilized depth cameras to capture RGB-D images and fused the meshes of each view to obtain the final 3D model. Although the work has achieved good performance for 3D reconstruction of human body, cameras with depth sensor are still inconvenient in many applications.

Parametric model based methods often tackle the problem through fitting a parametric human body model to prior information of the given images. The parametric human body model is often trained by a dataset and is defined as a function of variables which can represent prior information like pose and shape. The parametric models such as SCAPE [10] and SMPL [113] have been used in many methods. Recently, an improved model called SMPL-X was proposed by considering the motion of face and hands [139]. With the development of deep learning, some methods exploited deep neural networks to regress the parameters of the parametric model, hence we call them regression based methods. In [81], the pose and shape parameters of the SMPL model were estimated by training a deep encoder network. In [178], the loss function of mesh was added to further finetune the mesh of the 3D model. In [138], the pose and shape parameters were separately trained in two pipelines to make the result better. Kolotouros *et al.* [88] used the output of deep neural network to initialize the SMPL model and then supervise the training process of deep neural networks through the SMPL model. In [140], texture was utilized to capitalizes on the appearance constancy of images from different viewpoints. Although these methods have achieved competitive results, collecting datasets for training is still cumbersome work and training the network is also time-consuming. Another way to solve the problem is to fit the parametric human body model to prior information through optimizing an error function (optimization based methods). Early work [161] used SCPAE to estimate the articulated pose and non-rigid shape. In [53], silhouettes and joint points were manually obtained and the SCAPE model was fitted to the priori clues to estimate the parameters of SCAPE. In [186, 17], RGB-D images were utilized to estimate the parameters of SCAPE model. Xu *et al.*[196] scanned a template as the parametric model and used it to fit the prior information through optimizing an energy function. Bogo *et al.* [18] proposed a method called SMPLify in which the joint points were predicted by human pose estimation based on a deep neural network, and then SMPL was fitted to the estimated joint points. Moreover, silhouettes

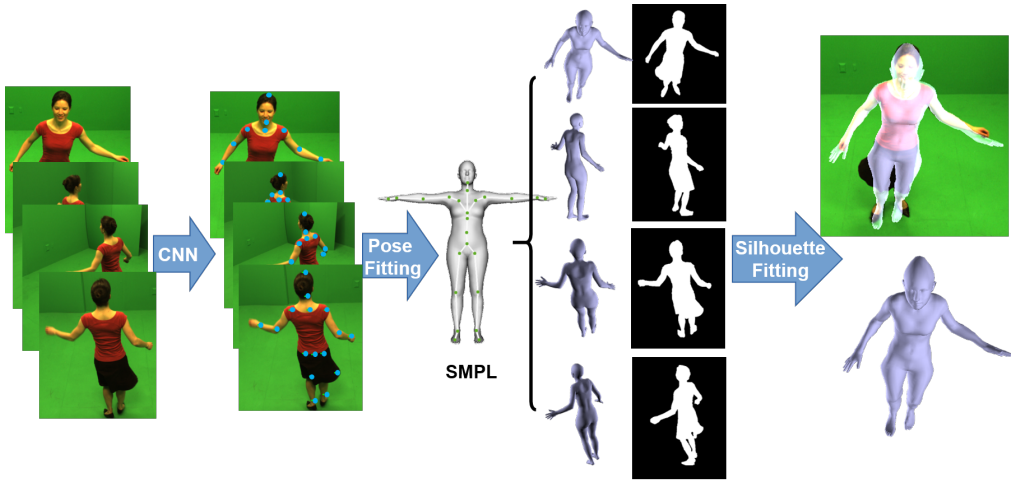


Figure 3.1: The overview of our method. 2D joint points of multi-view images are extracted by CNN. Then, pose fitting are proposed to fit the SMPL model based on the 2D joint points. Finally, the deformed SMPL model are fitted to the silhouettes to obtain the final results.

[3] and multiple images with different views [68, 103] were introduced as prior information for the SMPL model. Overall, optimization based methods are often easier to implement, since it is unnecessary to create datasets and to do training.

3 Method

In this section we present the method to obtain a 3D human body model from a small number of images taken from different view-points, using the joint points and silhouettes based on the SMPL model. The overview of the proposed method is shown in Figure 3.1. Here we use four images as the example. The pipeline of our method consists of two stages: pose fitting based on joint points and shape fitting based on silhouettes. For the pose fitting, the joint points of images from multiple viewpoints are estimated firstly using a CNN-based human pose estimation. The parametric human body model SMPL is fitted to the predicted joints through minimizing an energy function. Furthermore, we need to establish the correspondence between the pose-fitted SMPL model obtained by the first step and the silhouettes of the multiple-view images. The energy function can be built based on the correspondence from 2D and 3D space. Through minimizing the energy function, SMPL is fitted to the silhouettes of the four images to estimate the shape of the human body.

3.1 Parametric human body model

The parametric human body model in our method is called SMPL which is learned from an aligned human body dataset [113]. SMPL is defined as a function of pose $\vec{\theta} \in \mathbb{R}^{3\times 24}$ and shape $\vec{\beta} \in \mathbb{R}^{1\times 10}$ of the human body. The output of the function is a mesh with $V = 6890$ vertices and $F = 13776$ faces. This means that we can generate different 3D human bodies as long as we can get proper parameters of $\vec{\theta}$ and $\vec{\beta}$. There are 24 joint points in SMPL and each of them is represented as the rotation vector in terms of the root point, i.e., the *i-th* joint point is represented as $\theta_i \in \mathbb{R}^3$. The shape parameters $\vec{\beta}$ are the first 10 coefficients of the principle components of the training dataset.

3.2 Pose fitting

In the following, we explain the pose fitting in our method between SMPL and estimated joint points. For given multiple-view RGB images, the joint points are predicted by a CNN-based human pose estimation method [192]. In order to ensure the accuracy of human pose estimation, we firstly use CornerNet [94] to detect the bounding box of the person, and then use the image with bounding box into [192] to predict the joint points. Note that the order of the output of [192] is different from the order of joints of SMPL. For given N images from different views, the joint points are defined as $J_{2d}^{(i)}, i = 0, ..., N-1$. For the SMPL model, the joint points J_S are in 3D space and J_S is a function of pose $\vec{\theta}$ and shape $\vec{\beta}$. Suppose that the camera transformation matrix is $\Pi_i = (R_i, t_i)$ for the *i-th* camera. The projected 2D joint points of the SMPL model on the image plane can be represented as $\Pi_i(J_S(\vec{\theta}, \vec{\beta}))$. Therefore, the energy function to fit the SMPL model using joint points is defined as

$$E(\vec{\theta}, \vec{\beta}, \mathbf{R}, \mathbf{t}) = E_{jt}(\vec{\theta}, \vec{\beta}, R, t) + \omega_\theta E_\theta(\vec{\theta}) + \omega_\beta E_\beta(\vec{\beta}) \quad , \tag{3.1}$$

where E_{jt} is the joint points term and $E_\theta(\vec{\theta}), E_\beta(\vec{\beta})$ are the regularization term for $\vec{\theta}, \vec{\beta}$. ω_θ and ω_β are the weights of the regularization terms. $\mathbf{R}$ is $\{R_1, R_2, R_3\}$ and $\mathbf{t}$ is $\{t_1, t_2, t_3\}$. The joint points term E_{jt} measures the difference between all of the joint points $J_{2d}^{(i)}$ and $\Pi_i(J_S(\vec{\theta}, \vec{\beta}))$

$$E_{jt}(\vec{\theta}, \vec{\beta}, \mathbf{R}, \mathbf{t}) = \sum_{i=0}^{N-1} \rho\left(J_{2d}^{(i)} - \Pi_i(J_S(\vec{\theta}, \vec{\beta}))\right) \quad , \tag{3.2}$$

where ρ is the Geman-McClure function [44] and is defined as $\rho(x) = x^2/(\sigma^2 + x^2)$. σ is a constant and it is set as 100. Geman-McClure function can better deal with large noise and outliers. The regularization term for $\vec{\theta}$ is defined as

$$E_\theta(\vec{\theta}) = \alpha \sum_{i=55,58,15,12} exp(\theta_i) \quad , \tag{3.3}$$

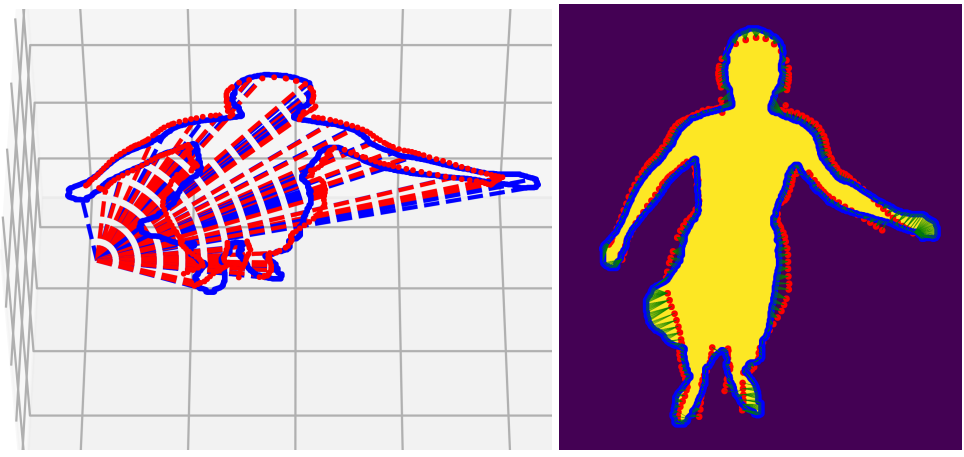


Figure 3.2: An example of correspondence between silhouette and SMPL model in 2D and 3D space. The left is the 3D correspondence and the right is the 2D correspondance between SMPL and silhouettes. The red points are SMPL vertices and the blue points are the corresponding points on silhouettes.

where α is a constant which is set as 10 and the 55-th, 58-th, 15-th, and 12-th elements in $\vec{\theta}$ are the joint points on the left and right elbows and keens. This can avoid the arms and legs to exhibit strange bending. The regularization term of $\vec{\beta}$ is defined as

$$E_\beta(\vec{\beta}) = \sum_{i=0}^{9} \beta_i \ . \tag{3.4}$$

The advantage of our method is that the camera parameters are also regarded as variables. After the optimization, the rotation and translation of the cameras will also be estimated. Therefore, through the minimization of the energy function, the pose, shape parameters of the SMPL model and the camera parameters can be obtained.

3.3 Shape fitting

The following section will describe the progress of shape fitting in our method. Since joint points mainly provide information about human pose in the first step, the silhouettes are used in this part to improve the estimation of shape. Here we assume that the silhouettes have been given. Now let us revisit the SMPL model about the vertex position. As shown in [113], the vertex of SMPL is transformed as

$$\mathbf{t}_i = \sum_{k=1}^{K} \omega_{k,i} G'_k(\vec{\theta}, J(\vec{\beta})) \left(\bar{\mathbf{t}} + B_S(\vec{\beta}) + B_P(\vec{\theta}) \right) \ . \tag{3.5}$$

In addition, since the rotation R and translation t of the camera are estimated after pose fitting, the positions of the cameras can be computed as $c = -R^T t$. Thus, we can define

a ray from the camera c to the vertex $\mathbf{t}_i$ of the transformed SMPL model as $\vec{r}$, as shown in Figure 3.2. Then, for the untransformed SMPL model, the corresponding ray is

$$\vec{r}' = \left[\sum_{k=1}^{K} \omega_{k,i} G_k'(\vec{\theta}, J(\vec{\beta}))\right]^{-1} \vec{r} - B_P(\vec{\theta}) \ . \tag{3.6}$$

We would like to find the correspondence between $\vec{r}'$ and the boundary points of the observed silhouette. This ray can be decomposed using Plücker coordinates $(\vec{r}'_m, \vec{r}'_n)$. Given the silhouette of the image, we can find the boundary points $\mathbf{v}$ of the silhouette and then backproject $\mathbf{v}$ to $\mathbf{V}$ in the camera coordinates since we have estimated the camera parameters. Then, the distance from the points to the ray can be computed as $d = \mathbf{V} \times \vec{r}'_n - \vec{r}'_m$. Those points and rays whose distance is smaller then a threshold are regarded as corresponding pairs. These pairs are defined as a set P which is the correspondence in 3D space. On the other hand, the vertices of SMPL model intersected by ray $\vec{r}'$ can be projected to the image plan as $\mathbf{v}'$ using camera parameters. The point set $\mathbf{v}$ and $\mathbf{v}'$ are defined as Q, which is the correspondence in 2D space. Figure 3.2 shows one example of the correspondence on the SMPL vertices and the silhouettes in 2D and 3D space. We can see that the correspondence in this case seems to be correct and can provide additional information for the shape fitting. Overall, the energy function using silhouettes is defined as

$$E(\vec{\beta}) = E_{silh}(\vec{\beta}) + E_{reg}(\vec{\beta}) \ . \tag{3.7}$$

The silhouette term $E_{silh}(\vec{\beta})$ is constructed by using the set P and Q and is defined as

$$E_{silh}(\vec{\beta}) = \sum_{(\mathbf{V},\vec{r}) \in P} \rho(\mathbf{V} \times \vec{r}'_n - \vec{r}'_m) + \sum_{(\mathbf{v},\mathbf{v}') \in Q} \rho(\mathbf{v} - \mathbf{v}') \ , \tag{3.8}$$

where $\mathbf{V} \times \vec{r}'_n$ is the cross product of $\mathbf{V}$ and $\vec{r}'_n$, ρ is the Geman-McClure function as Eq.(2). The first part of E_{silh} measures the difference of 3D points of backprojected silhouette boundary and rays, while the second part shows the difference of 2D silhouette points and projected SMPL vertices. Therefore, the silhouette term considers the silhouette information from both 3D and 2D perspective in contrast to the paper [3].

The regularization term is defined based on the SMPL model with zero pose, i.e., $\vec{\theta} = \vec{0}$. This is because this part only focuses on the shape estimation. Then, the SMPL model is computed as $\mathbf{t}(\vec{\beta}, D) = \bar{t} + B_S\vec{\beta} + D$, where D is the offset given by the SMPL model. The regularization term contains the Laplacian term E_L as well as the body model term E_B and it is represented as in [3]

$$E_{reg}(\vec{\beta}) = \omega_L E_L + \omega_B E_B, \tag{3.9}$$

where ω_L and ω_B are the weights. The Laplacian term E_L is defined as

$$E_L = \sum_{i=1}^{N} ||L(\mathbf{t}_i) - \delta_i||^2 \ , \tag{3.10}$$

where L is the Laplace operator and $\delta_i = L(\mathbf{t}_i(\vec{\beta}, 0))$. This term enforces smooth deformation. The body model term E_B is represented as

$$E_B = \sum_{i=1}^{N} ||\mathbf{t}_i(\vec{\beta}, D) - \mathbf{t}_i(\vec{\beta}, 0)||^2 \quad . \tag{3.11}$$

Through minimizing (3.7), the shape parameters can be estimated and the final results are obtained.

3.4 Optimization

After building the energy functions based on joint points and silhouettes, we need to optimize the energy functions. We have used Python to implement our optimization method. The energy functions in (3.1) and (3.7) are minimized by Powell's dogleg method which is provided in the Python modules called OpenDR [114] and Chumpy [112]. For four images with different views, it takes about 2 minutes to obtain the final estimation of the 3D human body.

Table 3.1: The values of relating parameters for the optimization.

	Synthetic dataset						Real dataset					
	Pose fitting			Shape fitting			Pose fitting			Shape fitting		
k	ω_θ	ω_β	σ	ω_L	ω_B	σ	ω_θ	ω_β	σ	ω_L	ω_B	σ
1	91.0	100	100	6.5	0.9	0.05	91.0	100	100	6.5	0.9	0.08
2	91.0	50	100	5.25	0.75	0.03	91.0	50	100	5.25	0.75	0.04
3	47.4	10	100	4	0.6	0.01	47.4	10	100	4	0.6	0.03
4	4.78	5	100				4.78	5	100			

The parameters used during the optimization are shown in Table 3.1. In the following experiments, we mainly used a synthetic dataset and a real dataset to evaluate our approach. Table 3.1 gives the parameters that we used in the experiments for the two datasets. For pose fitting, we assume that the focal length of the camera is known, but the translation and rotation of the camera are unknown. We initialize the rotation matrix as the identity matrix. The translation vector is initialized according to the torso length of SMPL model and the torso length of human body in the images. The weights ω_θ and ω_β in the energy function are decreased gradually after some iterations or when the value of the energy function is smaller than a threshold. For the silhouettes based energy function, we assume that the ground truth of the silhouettes are given. The weights ω_L, ω_B and the parameter σ in Geman-McClure function are decreased gradually after some iterations or when the value of the function is smaller than a threshold. The pseudo-code for the procedure of optimization

is demonstrated in the Algorithm 1.

Algorithm 1: Procedure of optimization

Pose fitting:
Input: Images from different views, 2D joint points, and SMPL model
Output: $\vec{\theta}, \vec{\beta}, \mathbf{R}, \mathbf{t}$
for *k=0:3* **do**
    update: ω_θ, ω_β and σ according to Table 3.1
    iter = 0
    while $E > 1e\text{-}3$ *in (3.1) or iter*> 100 **do**
        iter += 1;
        compute derivative of E in (3.1);
        update $\{\vec{\theta}, \vec{\beta}, \mathbf{R}, \mathbf{t}\}$

Shape fitting:
Data: Silhouettes, $\vec{\theta}$ and $\vec{\beta}$ after pose fitting
Output: $\vec{\beta}$
for *k=0:2* **do**
    update: ω_L, ω_B and σ according to Table 3.1
    iter = 0
    while $E > 1e\text{-}3$ *in (3.7) or iter*> 100 **do**
        iter += 1;
        compute derivative of E in (3.7);
        update $\vec{\beta}$

4 Experiments

In this section, the experiments to evaluate our proposed method are presented. We firstly introduced the datasets which were used in the experiments. Then, we discussed the effect of joint points on pose fitting and the influence of silhouettes on shape fitting, respectively. Besides, we also evaluated the pose fitting and shape fitting on the final estimation. Finally, we compared our method to several previous approaches on the datasets to validate the advantage of our approach.

4.1 Datasets

To evaluate our approach for a variety of poses and shapes, we generated a synthetic dataset and also used a public real dataset. The synthetic dataset consisted of 50 male and 50 female 3D human bodies which were created by the SMPL model. We set all the human bodies as "A" pose through giving the same pose parameters of the SMPL model, while the shape of

each human bodies was different by varying the shape parameters of the SMPL model. For each 3D human body, we used four cameras from different views to project the 3D model into four 2D images. Since the 3D joint points of SMPL model relied on the pose and shape parameters, the ground truth of 2D joint points and silhouettes can also be obtained when we projected the SMPL model. The samples of male and female human body models are shown in Figure 3.3. For the experiments based on the synthetic dataset, the values of parameters in pose fitting and shape fitting for optimization are set according to Table 3.1. The procedure of optimization is shown in Algorithm 1.

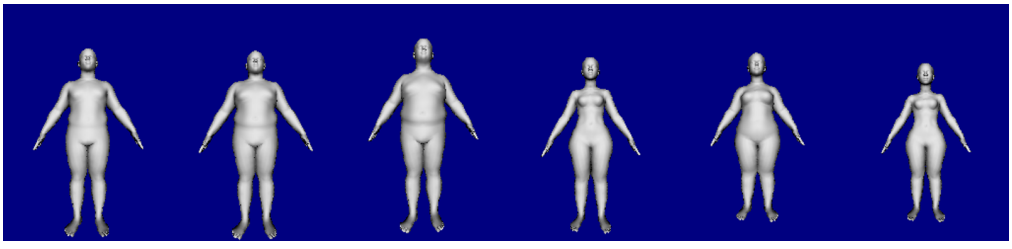
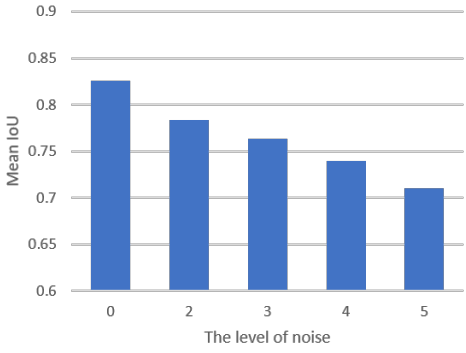


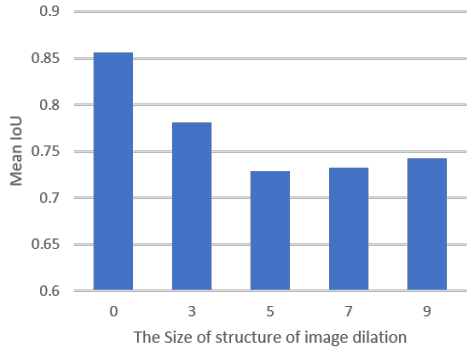
Figure 3.3: Some samples in the synthetic dataset.

Figure 3.4: Example of joint points and silhouettes from the real dataset.

In terms of the real dataset, we used the public data from [179] which consisted of ten image sequences. Each sequence was captured from eight different views by eight cameras in an indoor scene. Four images which are taken by the 2-nd, 4-th, 6-th and 8-th cameras are adopted in our experiments. Note that there are two marches and squats in the dataset, so we evaluate the results of march1 and squat2, i.e, the experiments are implemented based on 8 image sequences. For the joint points, we predicted the bounding box the person through CornerNet [94], and then estimated the joint points of the dataset through CNN-based human pose estimation in [192] using the cropped images by the bounding boxes. In terms of the silhouettes, the ground truth was given in this dataset. However, the silhouettes



(a) Joint points on pose fitting



(b) Silhouettes on shape fitting

Figure 3.5: The effect of joint points and silhouettes on pose fitting and shape fitting for the synthetic dataset.

can be extracted through threshold and filter since the background can be easily removed. In practice, semantic segmentation can be used for silhouettes extraction. Silhouette segmentation is not the key problem in our method, so we directly use the ground truth of silhouettes like [3]. One example of joint points and silhouette is shown in Figure 3.4. Similarly, the values of parameters for optimization are shown in Table 3.1 and the procedure of optimization is shown in Algorithm 1. If the final estimation was not good under the weights, these parameters can be adjusted to make the results better.

The metric for quantitatively comparing to other methods is the intersection over union (IoU) between the ground truth of silhouettes and the estimated silhouettes. Since the experiments are based on images from four views, the IoU is also computed using four silhouettes from four views. The higher of IoU means the better of the final estimation on the shape.

4.2 Evaluation of pose fitting and shape fitting

Figure 3.5(a) shows the mean IoU over 100 samples in the synthetic dataset when the joint points of the human body are added different levels of noise. The noise is the standard normal distribution $N(0, 1)$ and it can be added to the joint points by $J_{2d} + N(0, 1) \times k$, where k is the level of noise and is set as 2, 3, 4 and 5. After pose fitting for each human body based on the joint points with the k-th noise, we can obtain the pose and shape parameters of SMPL model as well as the parameters of four cameras. Then, four estimated images can be generated by projecting the estimated SMPL model to 2D images using four camera parameters. The IoU of the four images can be computed by using the ground truth of silhouettes. Finally, the mean IoU over the 100 samples in the synthetic dataset can be

computed. Note that the first data in Figure 3.5(a) is the result when we use the ground truth of joint points in pose fitting. Figure 3.5(a) shows that the accuracy of pose fitting decreases when the joint points are added larger noise. This means that the joint points should be as accurate as possible if we want to obtain better results.

Figure 3.5(b) shows the mean IoU over 100 samples in the synthetic dataset when we use image dilation with different size of structure to process the silhouettes. We varied the size of structure of image dilation as 3, 5, 7 and 9 to change the silhouettes. For each human body, we firstly used the ground truth of joint points to fit the pose, and then used the silhouettes after image dilation with different size of structure to implement shape fitting. The IoU of the human body can be computed and the mean IoU in the synthetic data can be obtained under the size of structure of image dilation. We can see from the figure that the size of structure will affect the accuracy of the final results. Generally, the larger the size of structure is, the worse the final results are. When the size of structure is larger than 5, the accuracy will keep stable because distance between the points on SMPL model and points on the edge of silhouettes is too large to find the correspondence.

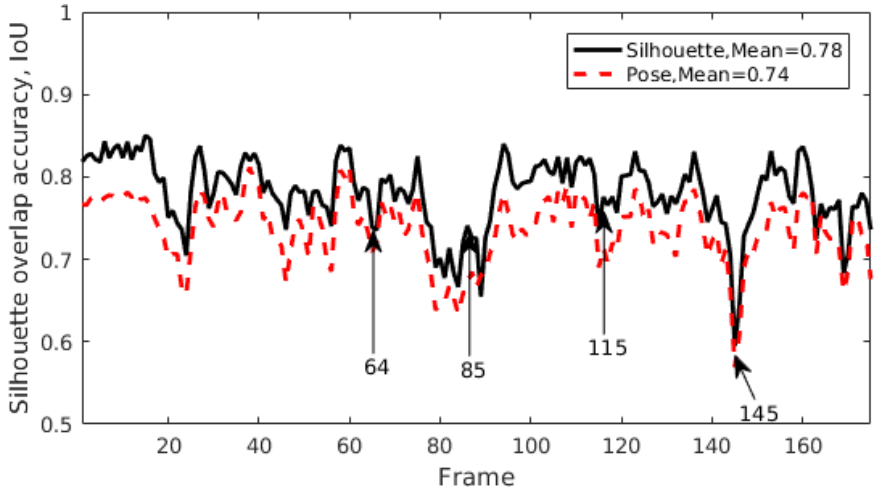


Figure 3.6: Comparison of pose fitting and shape fitting of our method for the real dataset of *Bouncing*.

We use the real data to evaluate the performance of pose fitting and shape fitting on the final results. The results after pose fitting and shape fitting from *Bouncing* are shown in Figure. 3.7. We can see from the figure that pose fitting can recover the pose of human body correctly and shape fitting can improve the 3D model on the shape details. In order to quantitatively demonstrate the performance of pose fitting and shape fitting, the IoU of silhouettes from four views of the image sequence *Bouncing* are calculated and shown in Figure 3.6. It is shown from the figure that the IoU after silhouette fitting is higher than the IoU only using pose fitting for most frames in the sequence. The mean IoU after shape

Figure 3.7: Comparison of pose fitting and shape fitting of our method for the examples from *Bouncing*. From top to down: Original images, results after pose fitting and results after shape fitting.

fitting is 0.78, while the mean IoU only using pose fitting is 0.74, which shows that shape fitting is a step to improve the accuracy of human body reconstruction.

4.3 Comparison to previous approaches

In this section we evaluate our method on both the synthetic and real dataset. To show the performance of our method, we compared to three previous approaches: SMPLify [18], SMPLify4 [103] and VideoAvatar [3]. Figure 3.8 qualitatively shows the comparison of two examples from the synthetic data. We can see from this figure that our method can better recover the shape of human body model than the other three methods. Since the pose of the human body is quite simple, these methods can estimate the pose well. However, our method can better recover the waist part of the human body. The estimated 3D models obtained by the other previous methods were not good on the waist parts. Especially for VideoAvatar, since our method established the energy function of silhouettes in 2D and

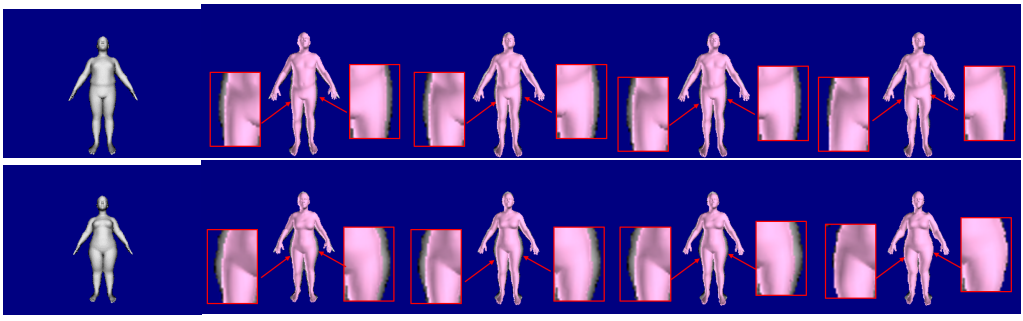


Figure 3.8: Two results on the synthetic data of one view. From left to right: the original images, SMPLify [18], SMPLify4 [103], VideoAvatar [3] and our method.

3D space, the results using four images are better than VideoAvatar which used 120 images from different view-points to obtain the 3D human body model. Figure 3.9 shows the results of our method from the other three view-points of the male and female model in Figure 3.8. The figure demonstrates that our method can recover the shape of human body model not only for the single view-point but also for the other view-points, which means that the 3D model estimated by our method is satisfying.

Figure 3.10 shows the IoU of silhouette overlap of our methods compared to other three methods on the synthetic dataset. The IoU of silhouette overlap is computed based on the silhouette of the projected 3D human body model and the corresponding ground truth of silhouette. The IoU for SMPLify is calculated based on one view-point, while the IoU of the other methods is based on four view-points. Although the results of some samples for SMPLify are better, the accuracy of our method is still higher than the results of the other three methods for the most samples in the dataset. Since SMPLify only adopts one image, the optimization is not sensitive to the initialization, which is the reason that the results of SMPLify on some samples are better. Compared to the SMPLify4 [103], our method introduced silhouette after the pose optimization, and thus, the results of our method are better on the synthetic dataset. In addition, for VideoAvatar [3] which also uses silhouettes, our results are still better. The reason could be that the energy function of silhouette in VideoAvatar was only built from 3D space. In VideoAvatar, the 3D model was acquired from a video stream containing 120 frames from multiple viewpoints. By contrast, the improved energy function of silhouettes in our method considers both 2D and 3D, which ensures that our method has good performance only using four images.

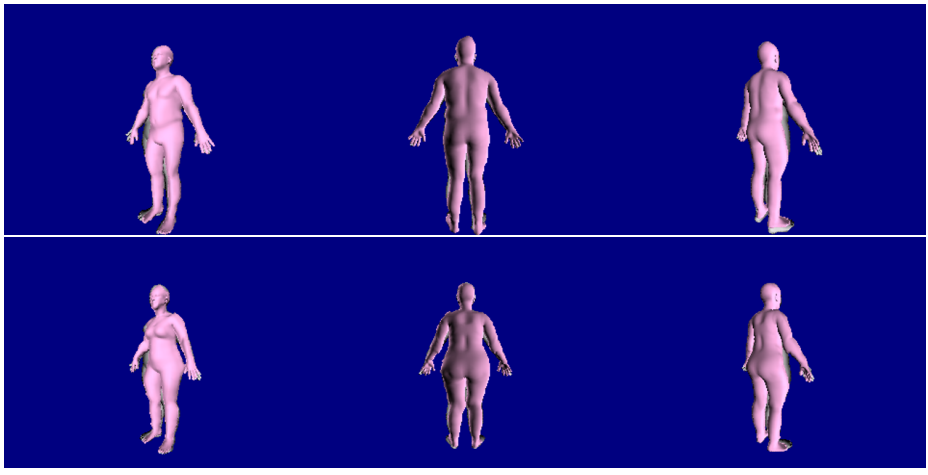


Figure 3.9: The results of the other three views obtained by our method.

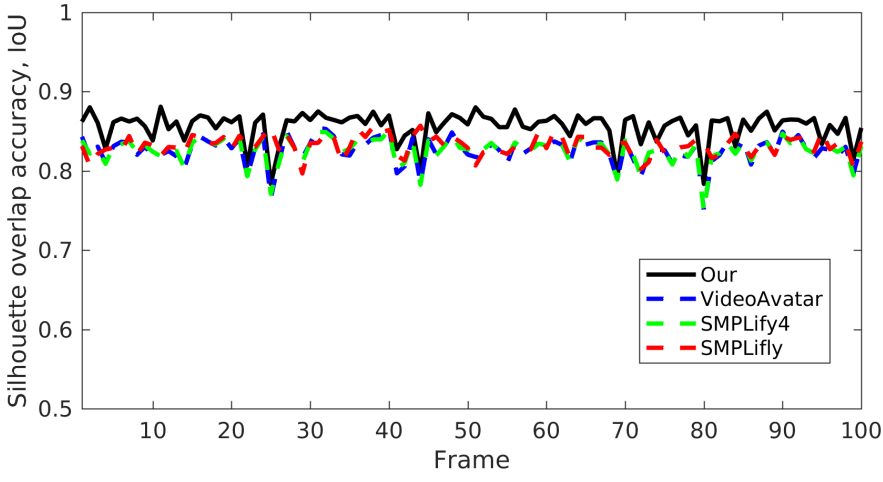


Figure 3.10: The comparison of IoU of silhouette overlap between our method and other methods on the synthetic dataset.

In the following, we evaluate our method on the real dataset in [179]. Firstly, we show the IoU of silhouette overlap for image sequence *Crane* in Figure 3.11. It shows the IoU of the results of SMPLify [18], SMPLify4 [103], VideoAvatar [3] and our method. Note that the IoU of SMPLify is also computed based on four images because the pose in the real data is much more complicated than the human body in synthetic dataset. The four images are the projection of SMPL model generated by SMPLify using the cameras estimated by our method. This can better reflect the accuracy of 3D model. We can see that our methods obtains higher accuracy than the other three methods for the most samples in this image

sequence. The results of SMPLify are the worst because only joint points from one single image are used as prior information. The results of SMPLify4 and VideoAvarta are almost the same because VideoAvatar requires enough number of images from different views. Furthermore, we also give the average of IoU of silhouettes overlap for the 8 different actions in the real dataset in Table 2. It is shown from this table that our method achieves the best performance comparing to other three previous methods because the IoU of our method is the highest. The results of SMPLify are worst, while the SMPLify4 and VideoAvatar have almost the same performance and they are better than SMPLify. The results of *Handstand* are not good because the pose estimation for the images in the sequence is not good. The pretrained model in human pose estimation of [192] cannot achieve good estimation for human body with handstand. Even in this case our results are sill the best comparing to other methods. Overall, our method has competitive performance among these approaches according to Table 3.2.

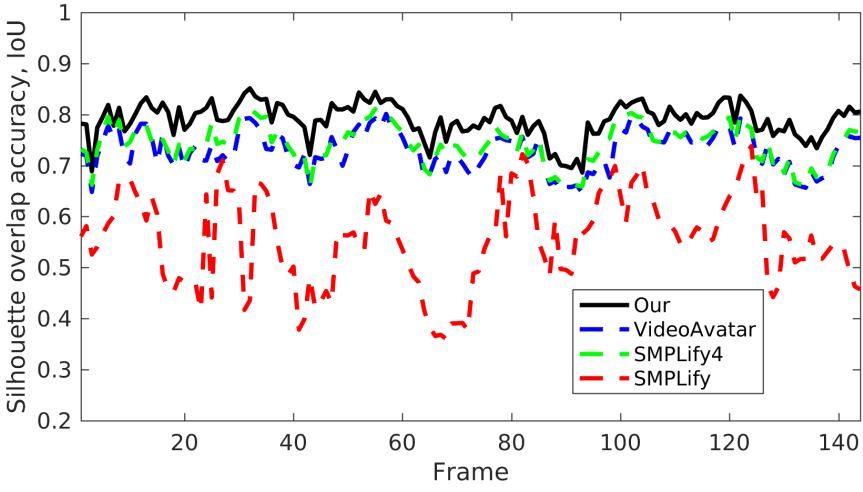


Figure 3.11: The comparison of IoU of silhouette overlap between our method and other methods on the *Crane* image sequence in the real dataset.

Table 3.2: The mean IoU of silhouette overlap of the 8 image sequence for different methods in the real dataset.

	Frames	SMPLify [18]	SMPLify4 [103]	VideoAvatar [3]	Our
Swing	150	0.5649	0.7570	0.7573	**0.7748**
Crane	175	0.5558	0.7425	0.7296	**0.7900**
Bouncing	175	0.5660	0.7367	0.7337	**0.7811**
Jump	150	0.5664	0.7078	0.7035	**0.7590**
Samba	175	0.5255	0.7544	0.7559	**0.7734**
Handstand	175	0.5384	0.6131	0.6118	**0.6504**
March1	250	0.5224	0.6930	0.6887	**0.7227**
Squat1	250	0.5256	0.7316	0.7304	**0.7726**

Several images from *Swing*, *Crane*, *Samba* and *Bouncing* are shown in Figure 3.12. In the figure, we illustrate one frame from each above sequence and give the qualitative results of the three previous methods and our method from one view. We can see from the figure that the shape of our method gives a better fit to the original images, which can be seen from the parts that are zoomed in. More specifically, the *Bouncing* results of SMPLify are not correct. This is because using only a single RGB image gives a too high uncertainty concerning the spatial information. Compared to SMPLify4 and VideoAvatar, which are shown in the second and third columns in Figure 3.12, the shapes of our method provides a better fit to the original images. This demonstrates the effectiveness to use the energy function based on silhouettes from 2D and 3D space. Therefore, our method achieves a good estimation not only for the pose but also for the shape of the human body. We also provide the results obtained by our method from the other three views in Figure 3.13. It demonstrates that the results from other views are correct, which means that the 3D model estimated by our method has better accuracy.

5 Conclusion

We have proposed a novel method for human pose and shape estimation using joint points and silhouettes based on SMPL model from multi-view images. SMPL model provides better representation about the 3D human body and the prior information including joint points and silhouettes gives strong cues for human body estimation. Our method consists of two steps: joint points based fitting and silhouettes based fitting. The joint points of the images were firstly predicted by deep learning-based human pose estimation. Then, the pose and shape parameters of SMPL model were estimated by fitting the SMPL model to the joint points of the four images simultaneously. Furthermore, we identified the corresponding points on the edge of silhouettes and SMPL model to build a novel energy function from 2D and 3D space. The shape parameters of SMPL were improved by minimizing the novel energy function. Our method not only estimated the pose of the human body, but also obtained better shape appearance of the human body. The experiments on synthetic dataset and real dataset indicated that our approach can obtain better human body shape comparing to the previous methods. The limitation of our method is that we strongly depends on the estimated joint points and silhouettes, which may result in that the estimation of pose and shape is not correct when the joint points or silhouettes are not predicted correctly. Besides, the texture of the images is not mapped to the 3D model, which makes the appearance is not realistic enough. Overall, our method can be used in many practical fields such as VR video games or biomedical research.

(a) Original (b) SMPLify [18] (c) SMPLify4 [103] (d) VideoAvatar [3] (e) Proposed

Figure 3.12: The results of *Swing, Crane, Samba and Bouncing* from top to down. The original images and the results of SMPLify [18], SMPLify4 [103], VideoAvatar [3] and proposed method are shown in (a), (b), (c), (d) and (e).

Figure 3.13: The results of *Swing, Crane, Samba and Bouncing* from other three views obtained by our method.

Paper III

Chapter 4

3D Human Pose and Shape Estimation Through Learning Collaborating Multi-view Model-fitting

Abstract

3D human pose and shape estimation plays a vital role in many computer vision applications. There are many deep learning based methods attempting to solve the problem only relying on single RGB image for training the network. However, since some public datasets are captured from multi-view cameras system, we propose a novel method to tackle the problem by putting optimization-based multi-view model-fitting into a regression-based learning loop from multi-view images. Firstly, a convolutional neural network (CNN) regresses the pose and shape of a parametric human body model (SMPL) from multi-view images. Then, utilizing the regressed pose and shape as initialization, we propose an improved multi-view optimization method based on the SMPLify method (MV-SMPLify) to fit the SMPL model to the multi-view images simultaneously. Subsequently, the optimized parameters can be adopted to supervise the training of the CNN model. This whole process forms a self-supervising framework which can combine the advantages of the CNN approach and the optimization-based approach through a collaborative process. Besides, the multi-view images can provide more sufficient supervision for the training. Experiments on public datasets qualitatively and quantitatively demonstrate that our method outperforms previous approaches in a number of ways.

1 Introduction

Human pose and shape estimation has many applications in virtual/augmented reality and computer games. However, this is a challenging problem since human bodies typically exhibit various motions and shapes in real scenes. Aiming at the problem, there are usually two routes to estimate 3D human pose and shape: optimization-based methods and regression-based methods [88]. Both of the approaches have achieved some success for the problem recently.

Traditionally, through defining a parametric human body model [10, 53, 113, 18, 17] or prescanning a 3D model as template [101, 55, 198, 196, 199], optimization-based approaches use some prior information including joint points [18], skeleton[53], silhouettes [17] and RGB-D images [186] to build an energy function. Some work adopt more than one cues in order to achieve better results [68, 3, 196] or propose novel optimization algorithms [101, 55]. By minimizing the energy function, the pre-defined human body model will fit to the prior information, and then, the estimated human pose and shape can be obtained. Although optimization based methods can be used to estimate 3D human body models in many different situations, it is often difficult to automatically extract accurate prior information due to the complexity of real human bodies. In addition, the optimization is often time-consuming.

On the other hand, regression-based methods for human pose and shape estimation have attracted much research with the significant achievements of deep neural networks in many image processing problems [171, 185, 119, 172, 137, 173, 169]. Regression-based methods [81, 178, 138, 131, 11] use deep neural networks that take all or subsets of pixels in the images and regress the human body and shape parameters based on training on large datasets. Many novel frameworks have been proposed to improve on accuracy of 3D human body estimation [69, 106, 140]. A dataset containing a large number of images and corresponding annotations is required for the methods to train the networks. Both the development of datasets and the time for training are serious drawbacks of regression-based methods. Recently, Kolotouros *et al.* put an optimization-based method into the loop of the regression-based framework and achieved good performance [88]. However, they only used one single-view image during the training.

Considering that some public datasets are captured from multi-view cameras, we propose a novel method for 3D human pose and shape estimation through a collaboration between learning and multi-view model fitting based on multi-view images in this paper. Firstly, a convolutional neural network (CNN) is advocated to regress the pose and shape parameters of a skinned multi-person linear model (SMPL) from multi-view images. Then, we fit the regressed SMPL model to all the multi-view images simultaneously through optimizing an energy function which is defined according to the joint points of the SMPL model

and the ground-truth joint points of the human body in the multi-view images. During the optimization, unlike the single view case in which only pose and shape parameters are optimized, we also optimize the orientation (i.e. the camera view) of the SMPL model for different views to reflect the relation of multi-view images. Finally, in addition to the typical 2D joint points supervision for training, the optimized pose and shape parameters as well as the optimized SMPL model are also adopted to supervise the training of the CNN. Therefore, the CNN can provide initialization of the SMPL model for optimization, while the optimized results can supervise the training process of the CNN, which builds a tight collaboration between the two parts. In addition to this, the multi-view optimization considers the inner relations of the given multi-view images, which can supply more accurate and complete information for the estimation. An overview of our method is shown in Figure 4.1. The code of our method is public at `https://github.com/leezhongguo/MVSPIN`.

The main contributions of our work have three parts. Firstly, a novel multi-view images based training strategy is used for the training of network, which better explores the information of the multi-view datasets. Besides, we propose a multi-view model-fitting, merged into a multi-view learning loop to form a novel framework for 3D human pose and shape estimation. Since multi-view model-fitting has better performance than single-view fitting, this provides reliable supervision for training the CNN and the results of our method surpass several recent methods. Finally, our framework can be used for 3D human pose and shape estimation from both single-view image and multi-view images after training the network with multi-view images. The experiments on some public datasets show that our method can better estimate 3D human pose and shape than some previous methods.

2 Related work

There are a large number of previous studies on the problem of human pose and shape estimation aiming at different tasks like joint points estimation, silhouette segmentation, part segmentation and so on. Here we mainly describe those relevant approaches for 3D human pose and shape estimation.

Parametric human body models have been widely used in the estimation of pose and shape. Anguelov *et al.* proposed a data-driven method called SCAPE to generate a deformable human body model [10]. It contained two models which were functions of pose and shape, respectively. They could be combined to create a 3D mesh with realistic muscle deformation. Some improvements based on SCAPE were proposed over the next several years [186, 184]. A new parametric human body was proposed by Loper *et al.* and it was skinned multi-person linear model (SMPL) [113]. It can model various body shapes with natural human poses by defining a function of pose and shape parameters, which made the model be widely used in human pose and shape estimation tasks. Pavlakos *et al.* extended SMPL

to SMPL-X by adding more key points on the face, hands and feet [139]. In [147], a dynamic human body model was proposed for modeling human body motion. The above human body models were all learned from a large human body dataset.

Optimization-based methods have traditionally been used to estimate human pose and shape parameters. In [53], a 3D human body model was estimated by fitting SCAPE to the manually acquired joint points and silhouettes. With the development of depth sensors, range data acquired by Kinect was used as prior information and an improved SCAPE model was fitted to the range data in [186, 17]. In addition to the use of prior cues, novel optimization methods were also explored by many researchers [101, 55, 198, 199]. It was also popular to use several different cues to estimate 3D human body [196, 58]. With the success of human pose estimation by deep neural networks, an automatic approach called SMPLify was proposed to estimate the parameters of the SMPL by using 2D joint points predicted by deep neural networks [18]. Inspired by the method, some approaches based on multi-view images [68, 103] and video [3] were proposed to improve the estimation.

Regression-based methods have also been developed and achieved significant success on the 2D [171, 185, 20] and 3D [137, 172, 8, 173, 169] human pose estimation. Most regression-based work used deep neural networks as encoders to estimate the pose and shape parameters directly from images. The training of the networks often relied on the annotation of 2D/3D joint points [88, 81], dense pose [95], multi-view images [106], silhouettes [33, 138], texture [140, 5] and part segmentation [131]. In [33], silhouettes were used to train a network to estimate the shape of a human body in a simple pose. For human bodies with complicated poses, Kanazawa *et al.* proposed an end-to-end framework using 2D joint locations [81]. In this method, the pose and shape parameters of the SMPL model were learned by the deep neural networks, using the reprojection loss which was defined by ground truth of 2D joint points and the projection of skeleton joints from the SMPL model. Inspired by this framework, many approaches were proposed by designing new routes to acquire various information to better supervise the network. Even for multiple people in images, Zanfir *et al.* [200] proposed a regression-based method to solve the problem. In addition to 2D CNN, some papers use 3D CNN to regress a volume and use a signed distance function to represent a detailed 3D model [69, 156]. In the above methods, Kolotouros *et al.* incorporated SMPLify into the training loop of the CNN, which was the first attempt to combine optimized-based method and regression-based method [88]. This made the training of the CNN self-supervised and achieved competitive performance.

3 Method

The details of our method are presented in this section. We will first introduce the learning-based parametric human body model used in our method. Then, the regression part and

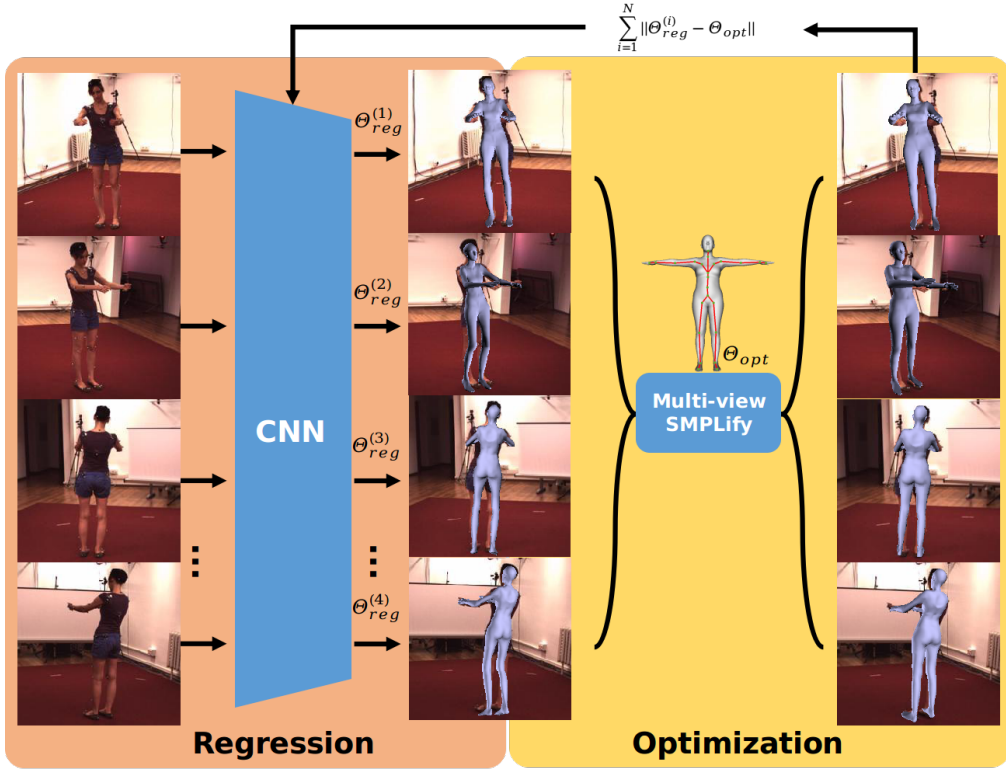


Figure 4.1: Overview of the proposed method. The CNN regresses the parameters Θ_{reg} from multi-view images. Then, using Θ_{reg} as initialization, multi-view SMPLify optimizes the parameters to obtain Θ_{opt}. The optimized parameters Θ_{opt} of the multi-view images are used to supervise the training of CNN.

the optimization part of our approach are presented, respectively. Based on these two parts, we define the collaboration of them to complete our whole method. Finally, we present the implementation details of our method.

3.1 The SMPL model

The SMPL model is a parametric human body model learned from a very large number of aligned human body shapes. It is a triangulated mesh with $N = 6890$ vertices and the position of each vertex is a linear function $M(\theta, \beta)$ of the pose parameters $\theta \in \mathbb{R}^{72}$ and the shape parameters $\beta \in \mathbb{R}^{10}$. The pose θ encodes the rotation angle of each skeleton joint point in terms of the root point. The shape β contains the coefficients of the ten most significant PCA vectors of the human body models extracted from the human body shape space. In addition, the skeleton joint points $\mathcal{J}$ of the SMPL model are also a linear function of pose θ and shape β. Since it is a linear model, a CNN is expected to perform well, when estimating a regression function to infer the pose and shape parameters. The

skeleton joint points of the SMPL model can also be used for the optimization on joint points in order to estimate the pose and shape parameters. Therefore, the SMPL model can be used for both regression and optimization.

3.2 The architecture of our regression CNN

In this section the architecture of the CNN to regress the human body parameters from images is introduced. The design of the network is based on the structure in [88]. Instead of using single view image for one training loop as in [88], we propose to form the multi-view images as a small batch and fed the small batch into the network for one training loop. Given the multi-view images, the network encodes the body in each single view image as a $\mathbb{R}^{85}$ vector containing the pose θ, shape β of the SMPL model and the camera Π as shown in Figure 4.1. The camera Π is a weak perspective model and is represented by a 3×1 vector (s, t_x, t_y) where s denotes the scale parameter and it can be converted to camera translation. This can be done because the rotation of the camera is assumed to be the identity. Then, the relative rotation between the human body and the camera is encoded in the root orientation of the body model. Suppose we have several images from different view-points, denoted $I_i, i = 1, ..., N$ along with the corresponding camera parameters $\Pi_i \in \mathbb{R}^{3\times1}$. Since the multi-view images are from the same human body (pose and shape) from different view-points, the multi-view images have the same ground truth for the pose and shape parameters $\Theta = \{\theta, \beta\}$. For the i-th image I_i passing through the networks, the regressed parameters are defined as $\Theta^{(i)}_{reg} = \{\theta^{(i)}_{reg}, \beta^{(i)}_{reg}\}$ and $\Pi^{(i)}_{reg}$. Then, the predicted 2D joint points can be obtained by projecting the skeleton joint points of the SMPL model through the estimated cameras, i.e., $J^{(i)}_{reg} = \Pi^{(i)}_{reg}(\mathcal{J}(\Theta^{(i)}_{reg}))$, where $\mathcal{J}(\Theta^{(i)}_{reg})$ are the skeleton joint points of the regressed SMPL model. In addition, the predicted mesh of the SMPL model can also be generated by $M^{(i)}_{reg}(\Theta^{(i)}_{reg})$. Therefore, the loss function of the 2D joint points on the multi-view images can be defined as:

$$L_{2D} = \sum_{i=1}^{N} ||J^{(i)}_{reg} - J^{(i)}_{gt}|| \ , \tag{4.1}$$

where $J^{(i)}_{gt}$ denotes the ground truth of 2D joint points of the i-th input image I_i. Compared to [88], this loss function considers the 2D joint points from all of the views, which can reduce the ambiguity of 2D joint points from a single-view image and provide stronger supervision of the CNN model. In addition to the loss function on 2D joint points, loss function for pose and shape will be discussed in the following sections.

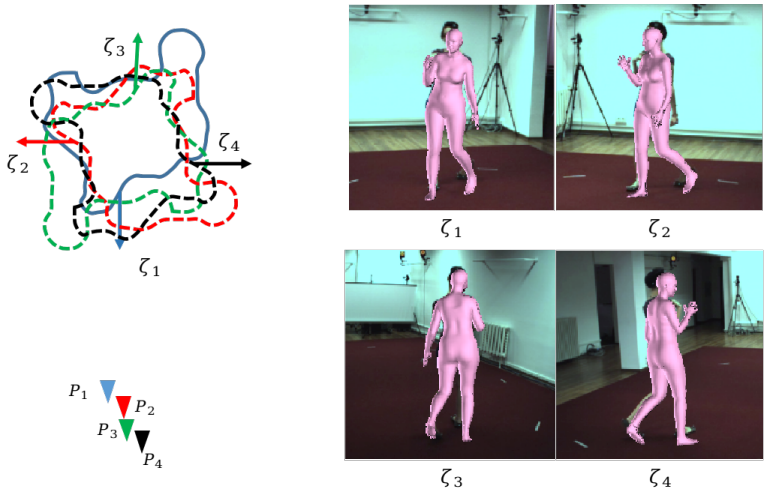


Figure 4.2: Illustration of the cameras, body orientations and projected SMPL models on the image planes. The four images share the same pose and shape parameters, while the camera translations and body orientations are different.

3.3 Multi-view SMPLify

In this section we apply an improved SMPLify method based on multi-view images in order to perform the optimization. SMPLify was proposed in [18] and it fitted the SMPL model to a set of 2D joint points predicted by a deep neural networks from a single image. In order to extend SMPLify from single-view image to multi-view images, an improved method was described in [103]. However, the results of [103] are often not robust enough since they initialize the camera rotation as the identity matrix, which may result in the optimization process ending in local optima. In our method, we optimize the body orientation instead of camera rotation because we have assumed that the camera is oriented to human body. According to the definition of the pose θ of the SMPL, the first three elements represent the body orientation denoted by $\zeta \in \mathbb{R}^3$. Then, we define $\tilde{\theta} = \theta \setminus \zeta$ as the pose of the rest joint points. Since the multi-view images share the same pose and shape, we initialize $\tilde{\theta}$ as the mean of $\tilde{\theta}_{reg}^{(i)}$ and β as the mean of $\beta_{reg}^{(i)}$, over all images $i = 1, \ldots, N$. The body orientations $\zeta^{(i)}$ for different views are initialized as $\zeta_{reg}^{(i)}$. We convert the weakly perspective camera $\Pi_{reg}^{(i)}$ to the camera translation $T_{reg}^{(i)}$ and define the camera rotation as the identity matrix. Then, the camera matrix for the projection can be represented as $P_{reg}^{(i)} = \{I, T_{reg}^{(i)}\}$. Using this camera matrix, the reprojected 2D joint points of the regressed SMPL model can be obtained as $P^{(i)}(\mathcal{J}^{(i)})$. Figure 4.2 illustrates an example of the cameras, body orientations and the corresponding projected regressed SMPL models on the image planes. Based on the above definition, the energy function of the multi-view SMPLify is defined as:

$$E(\tilde{\theta}, \beta, \zeta^{(i)}) = E_J(J_{gt}^{(i)}, P^{(i)}(\mathcal{J}^{(i)})) + \lambda_\theta E_{\tilde{\theta}}(\tilde{\theta}) + \lambda_\beta E_\beta(\beta) \ , \tag{4.2}$$

where E_J measures the errors between J_{gt} and $P^{(i)}(\mathcal{J}^{(i)})$ on all views. $E_\theta(\theta)$ and $E_\beta(\beta)$ are the regularization terms for pose and shape parameters, respectively. For a detailed description of these regularization terms, see [103]. For the energy function above, the minimization is an important step to get the optimized parameters. Similar to [88], fixing the pose and shape parameters, the camera translations of all the images and the orientation of the SMPL model were estimated first. This is implemented by using similar triangles defined by the torso length of regressed SMPL and the ground truth. The initialization of the camera translation T and the body model orientation ζ were obtained from the output of the CNN model. Then, fixing the camera translation, we minimize (4.2) to obtain the optimized pose $\tilde{\theta}_{opt}$, shape β_{opt} and multi-view body orientation $\zeta^{(i)}_{opt}$. Adam with 0.01 learning rate is used for the optimization and the maximum number of iterations is 100 in our experiments. Therefore, the complete optimized pose for the i-th image is $\theta^{(i)}_{opt} = \{\zeta^{(i)}_{opt}, \tilde{\theta}_{opt}\}$.

3.4 Collaborative learning

In this section we combine the CNN and the multi-view SMPLify into one route in a new training loop. As shown in Figure 4.1, the regressed pose and shape parameters $\Theta^{(i)}_{reg}$ and camera translation $T^{(i)}_{reg}$ are obtained after the images have passed through the networks. The loss function based on the 2D joint points is defined as in (4.1), and we use the regressed parameters to initialize the multi-view SMPLify. Through minimizing (4.2), the optimized parameters can be obtained as $\Theta^{(i)}_{opt} = \{\theta^{(i)}_{opt}, \beta_{opt}\}$. Then, using the optimized $\Theta^{(i)}_{opt}$, the optimized SMPL models and the corresponding skeleton joint points with different body orientations can be generated as $M^{(i)}_{opt}$ and $\mathcal{J}^{(i)}_{opt}$.

Now we can define additional contributing losses to train the CNN by using the above results. The loss for the pose and shape parameters is defined as

$$L_\Theta = \sum_{i=1}^{N} ||\Theta^{(i)}_{reg} - \Theta_{opt}|| \ . \tag{4.3}$$

Further, the loss function for the mesh of the SMPL model is defined as

$$L_M = \sum_{i=1}^{N} ||M^{(i)}_{reg} - M_{opt}|| \ . \tag{4.4}$$

If the ground truth of 3D joint points are provided in the training dataset, we can also define the loss function of the 3D joint points as

$$L_{3D} = \sum_{i=1}^{N} ||J^{(i)}_{3D} - \mathcal{J}^{(i)}_{opt}|| \ , \tag{4.5}$$

where $\mathcal{J}_{opt}^{(i)}$ denote the skeleton joint points of the i-*th* optimized SMPL model. Therefore, the complete loss function for training the network is defined as:

$$L = \omega_1 L_{2D} + \omega_2 L_{3D} + \omega_3 L_{\Theta} + \omega_4 L_M \ , \tag{4.6}$$

where $(\omega_1, \ldots, \omega_4)$ is the weighting of the terms. Intuitively, our proposed approach has some advantages compared to other methods. Firstly, multi-view images reduce the ambiguity of inferring 3D human pose from 2D joint points. Both in the regression and optimization process, multi-view images can obtain better results than a single view image. Besides, the CNN and the multi-view SMPLify form a tight collaboration during the training loop. The output of the CNN model can initialize the optimization problem, while the optimized results could supervise the training of the CNN model through the loss function defined by optimized parameters.

3.5 Implementation details

Training. In terms of the number of view points, we use four views in our experiments because the training public datasets that we used were acquired from four or eight views. For each training batch, the real number of images is $4 \times N$ where N is the batch-size used in the code. The CNN in our model is trained by Adam with 3×10^{-5} learning rate for 20 epochs. In the total loss function of (4.6), the weights of each sub-loss $(\omega_1, \omega_2, \omega_3, \omega_4)$ are $(5.0, 5.0, 1.0, 0.001)$. We train our model on two datasets: Human3.6M [72] and MPI-INF-3DHP [118]. In each batch, we use 90% images from Human3.6M and 10% images from MPI-INF-3DHP. All of the images are cropped to 224×224. The network is trained on an NVIDIA TITAN X (Pascal) GPU with 12 GB. The batch-size is set to 16 and each batch takes about 5.5 seconds for one iteration. The total number of iteration is 2441 for one epoch and the whole training takes about 3 days.

Inference. For the inference, we use single-view image to evaluate our method. Note that the optimization part is not used in the inference because 2D joint points should be unknown for the inference in practice. More specifically, Three datasets are used for inference including the S1 and S9 of Human3.6M, the validation dataset of MPI-INF-3DHP and the test set of 3DPW [180]. These testing images contains various poses and shapes under both indoor and outdoor scenarios.

4 Experiments

In this section some experiments are described to evaluate the performance of our method. We will briefly introduce the datasets used in the experiments for training and evaluation. Then, quantitative and qualitative results are demonstrated to compare the previous

methods based on both single-view image and multi-view image, respectively. Finally, an ablation study is given to show the advantage of our method comparing to method only relying on deep learning.

The metric for quantitative comparison in our experiments contains the reconstruction error, Mean Per Joint Position Error (MPJPE), Percentage of Correct Keypoints with threshold 150 mm (PCK@150 mm) and Area Under Curve (AUC) of 3D joint points. The lower of the first two metrics means better results, while the higher of the last two metric means better results. The reconstruction error is the MPJPE after Procrustes post-processing to remove scale ambiguity. PCK and AUC have the same definition as [119].

4.1 Dataset

Human3.6M. The first dataset in our experiments is the Human3.6M [72]. It contains 11 different subjects and each subject performs 15 different actions indoors. All of the data is acquired from four views and the corresponding 2D/3D joint points and part segmentation are also captured. Similar to previous work [81] which used the protocol 1, the video of the S1, S5, S6, S7 and S8 are used as training dataset, while the video S9 and S11 are used for evaluation. For the training set, we extract images from the video every ten frames, while evaluation images are extracted from S9 and S11 every five frames as in [88]. The training set contains 39066×4 images and the evaluation set has 109867 images.

MPI-INF-3DHP. The second dataset is the MPI-INF-3DHP [118]. It contains eight subjects for training and two subjects for testing. For each subject, eight videos from different views are captured and we choose *video_0*, *video_2*, *video_7* and *video_8* as training data. Only those images with a complete human body in all views are extracted from the videos every ten frames. The testing dataset can be used directly. Totally, the training set has 9452×4 images and the testing set has 2929 images.

3DPW. Since the above datasets are indoor scenario, we use the test set of 3DPW to evaluate our method on the outdoor scenario case. 3DPW is captured mostly in outdoor conditions using IMU which can provide ground truth 3D pose in the wild. There are 25 test image sequences in 3DPW. After removing some invalid frames, we can obtain totally 35515 images which are used for evaluation.

4.2 Comparison to single-view methods

We compare to some previous approaches which train the network using single-view image to estimate 3D pose and shape of human body. Table 4.1, Table 4.2 and Table 4.3 show the quantitative results of some previous work on the Human3.6M, 3DPW and MPI-INF-

3DHP, respectively. Note that we use the same testing dataset as the previous methods so that they are comparable. The results of the SPIN [88] are obtained through performing the SPIN using the trained model in the original paper by ourselves, while the results of other methods come from the corresponding paper. We can see from the two tables that our method outperforms most previous approaches on the three datasets. For the SPIN which trains the network using single-view image, our method achieved almost the same performance on Human3.6M. This is because SPIN use four different datasets to train the network, which makes their network more generalization. However, since our method trains the network based on multi-view images, the results of our method outperform SPIN on 3DPW and MPI-INF-3DHP even though we only use Human3.6M and MPI-INF-3DHP to train the network. Therefore, the two tables demonstrate that our method achieves better performance than those approaches trained from single-view image.

Table 4.1: Quantitative comparison to previous work trained by single-view image on Human3.6M.

Methods	Rec.Err. ↓	MPJPE ↓
Pavlakos *et al.* [138]	75.9	-
Omran *et al.* [131]	59.9	-
HMR [81]	56.8	87.97
Kolotouros *et al.* [89]	51.9	74.7
SPIN [88]	44.2	**64.5**
Ours	**43.8**	64.8

Table 4.2: Quantitative comparison to previous work trained by single-view image on 3DPW.

Methods	Rec.Err. ↓	MPJPE ↓
HMR [81]	76.7	130.0
Kanazawa *et al.* [82]	72.6	116.5
Arnab *et al.* [11]	72.2	-
Kolotouros *et al.* [89]	70.2	-
SPIN [88]	59.2	96.5
Ours	**58.6**	**93.4**

Table 4.3: Quantitative comparison to previous work trained by single-view image of MPI-INF-3DHP.

Methods	PCK↑/AUC↑/Rec.Err.↓	PCK↑/AUC↑/MPJPE↓
VNect [119]	83.9/47.3/98.0	76.6/40.4/124.7
HMR [81]	86.3/47.8/89.8	72.9/36.5/124.2
SPIN [88]	92.1/55.0/68.4	75.3/35.3/109.4
Ours	**92.9/56.1/65.6**	**79.2/39.3/98.7**

Table 4.4: Quantitative comparison to previous work based on multi-view images on S9 and S11 of Human3.6M.

Methods	Rec. Error ↓	MPJPE ↓	Known Camera?	Parametric Model?
PVH-TSP [172]	-	87.3	Yes	No
Trumble *et al.* [173]	-	62.5	Yes	No
Pavlakos *et al.* [137]	-	56.89	Yes	No
Tome *et al.* [169]	-	52.8	Yes	No
Liang *et al.* [106]	45.13	79.85	No	Yes
Ours	**43.8**	64.8	No	Yes

4.3 Comparison to multi-view methods

There are some approaches which also used multi-view images to train the network to regress human pose and shape. Table 4.4 gives the results of some previous methods based on multi-view images on the test data of Human3.6M. Note that the first four methods did not rely on parametric model to estimate 3D human pose. They assumed that the cameras were known so that the 2D joint points can be reprojected to 3D space. Therefore, the MPJPE of the three methods was calculated without any ambiguity with the ground truth on the scale or rotation. However, for Liang *et al.* and our method, the 3D poses are the deformed SMPL model and they have different scale with the ground truth due to the unknown cameras, so the MPJPE of the two methods are worse. After Procrustes Alignment on the 3D pose of the deformed SMPL model, the effects of ambiguity can be removed and the reconstruction error is more suitable to compare with the MPJPE of the other methods. We can see from the Table 4.4 that our method achieves the smallest reconstruction error, which demonstrates that our method outperforms the previous methods based on multi-view images on the Human3.6M. Since both Liang *et al.* and our method rely on SMPL model, we also compare to Liang *et al.* on the 3DPW and MPI-INF-3DHP which contain the images in the outdoor scene in Table 4.5 and Table 4.6. Although the method in [106] also use multi-view image to regress the pose and shape parameters of SMPL, our method still outperforms the method because the MV-SMPLify fully explores the relations between the multi-view images and provides better supervision on the training of the CNN. Therefore, our method achieves the satisfying performance on the three datasets even comparing to methods based on multi-view images for training.

Table 4.5: Quantitative comparison to previous work based on multi-view images on 3DPW.

Methods	Rec.Err. ↓	MPJPE ↓
Liang *et al.* [106]	-	96.86
Ours	**58.6**	**93.4**

Table 4.6: Quantitative comparison to previous work based on multi-view images on MPI-INF-3DHP.

Methods	PCK↑/AUC↑/Rec.Err.↓	PCK↑/AUC↑/MPJPE↓
Liang *et al.* [106]	86.0/49.0/89.0	66.0/29.0/137.0
Ours	**92.9/56.1/65.6**	**79.2/39.3/98.7**

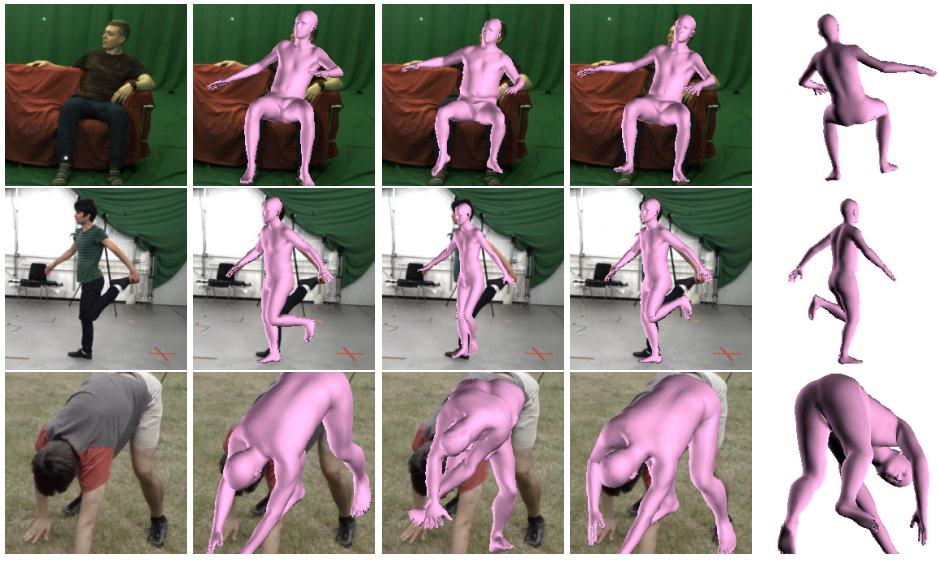
4.4 Qualitative results

In this section, we give some qualitative results of SPIN [88], Liang *et al.* [106] and our method on the datasets of Human3.6M, MPI-INF-3DHP and 3DPW. SPIN is the method based on single-view image, while Liang *et al.* [106] is the method based on multi-view images. Figure 4.3, Figure 4.4 and Figure 4.5 demonstrate the several examples from Human3.6M, MPI-INF-3DHP and 3DPW, respectively. In each figure, the results of SPIN [88], Liang *et al.* [106] and our method are shown from the second column to fourth column. The examples shown in the three figures contain various human poses and are captured both in indoor and outdoor scenes.

We can see that the human bodies in the images shown in the Figure 4.3∼ 4.5 have complicated poses with different backgrounds. The figures demonstrate that our method can recover the 3D human body model with better pose and shape estimation than the other two methods. For the examples from the three datasets, our method has better performance on the pose and shape estimation. The results of SPIN [88] are also better than the results of Liang *et al.* [106], which shows that putting optimization on training loop is more useful. For the images with indoor condition, our method achieved almost the same performance as the SPIN [88] on most images, especially for the Human3.6M. However, for the images with outdoor condition, our method clearly outperforms the SPIN. For example, the last column in Figure 4.5, SPIN [88] has the error on the left and right of the body estimation and the results of [106] are also false. For some complicated scene and pose in 3DPW, for example, the third row in Figure 4.5, our method also has errors but it still looks better than the two other methods. Since our method uses multi-view images and optimization in the training loop, the results on the fencing of our method are correct. The figures are also consistent with the quantitative results.

(a) Images (b) SPIN [88] (c) Liang *et al.* [106] (d) Ours (e) Novel view

Figure 4.3: The qualitative results from Human3.6M. From left to right: The original images, the results of SPIN [88], Liang *et al.* [106] and the results of our method from two views.



(a) Images (b) SPIN [88] (c) Liang *et al.* [106] (d) Ours (e) Novel view

Figure 4.4: The qualitative results from MPI-INF-3DHP. From left to right: The original images, the results of SPIN [88], Liang *et al.* [106] and the results of our method from two views.

(a) Images (b) SPIN [88] (c) Liang *et al.* [106] (d) Ours (e) Novel view

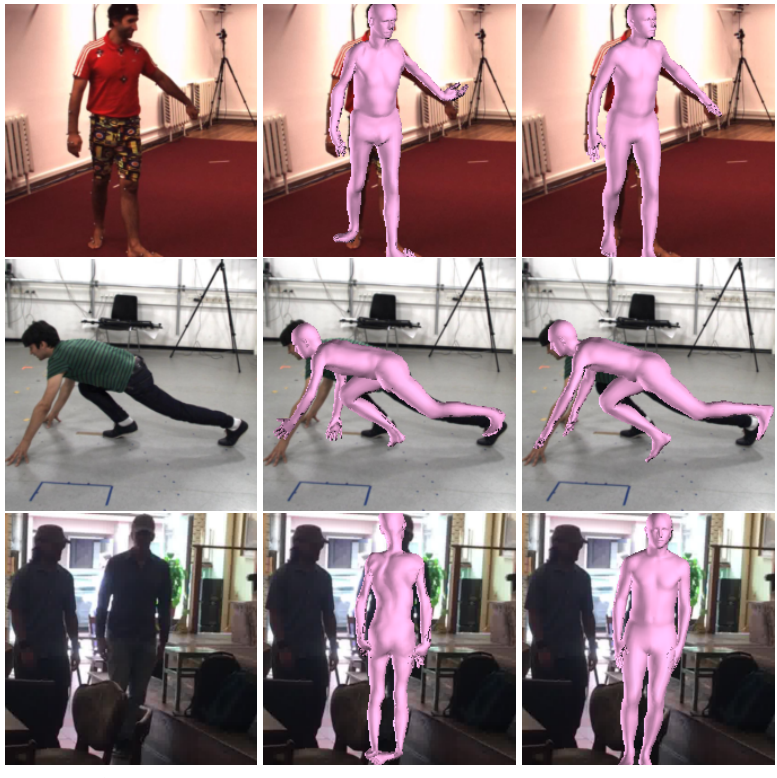
Figure 4.5: The qualitative results from 3DPW. From left to right: The original images, the results of SPIN [88], Liang *et al.* [106] and the results of our method from two views.

4.5 Comparison to training without optimization

We discuss the effect of multi-view SMPLify on the final estimation on the three datasets. The network was trained with multi-view SMPLify and without multi-view SMPLify, respectively. Table 4.7 shows the reconstruction error and MPJPE of the two cases on the three datasets. Our method with $L_{2D}+L_{3D}$ in the table stands for the results withour multi-view SMPLify. We can see that the accuracy has been improved after multi-view SMPLify is used in our training loop. Since the training datasets in our method are Human3.6M and MPI-INF-3DHP, the improvements are not significant. By contrast, the results on the 3DPW shows that our full method achieves more clear improvements. Figure 4.6 shows the qualitative results of our method without and with multi-view SMPLify from the three datasets, respectively. We can see that the results without multi-view SMPLify are worse, especially for the example from 3DPW (the last row in Figure 4.6). From the results of Human3.6M (the first row in Figure 4.6), we can see that the final 3D human body is not natural even though the pose is accurate. The wrist and the arm of the 3D model have unnatural blend and rotation. Therefore, the supervision of 2D and 3D joint points cannot guarantee the correct 3D model. After adding the supervision of multi-view SMPLify, our method can achieve the good estimation on the poses and the natural 3D bodies.

Table 4.7: The evaluation of the effect of multi-view SMPLify on our method for the three datasets.

	Human3.6M		MPI-INF-3DHP		3DPW	
	Rec.Err.↓	MPJPE↓	Rec.Err.↓	MPJPE↓	Rec.Err.↓	MPJPE ↓
Ours($L_{2D} + L_{3D}$)	46.4	65.8	66.8	100.8	61.7	99.0
Ours(Full)	**43.8**	**64.8**	**65.1**	**97.6**	**58.6**	**93.4**



(a) Original images (b) Ours($L_{2D} + L_{3D}$) (c) Ours(Full)

Figure 4.6: Qualitative results of our method without and with multi-view SMPLify in training loop from the three datasets.

4.6 The results of multi-view SMPLify

We first compare the performance of multi-view SMPLify and SMPLify in [18] to demonstrate the advantage of using multi-view images. Taking 100×4 images from S1 in Human3.6M as an example, these images were fed into the CNN with the pre-trained parameters in [88]. Using the output of the CNN as initialization, we optimized the energy functions of the multi-view SMPLify and SMPLify to get the optimized pose and shape,

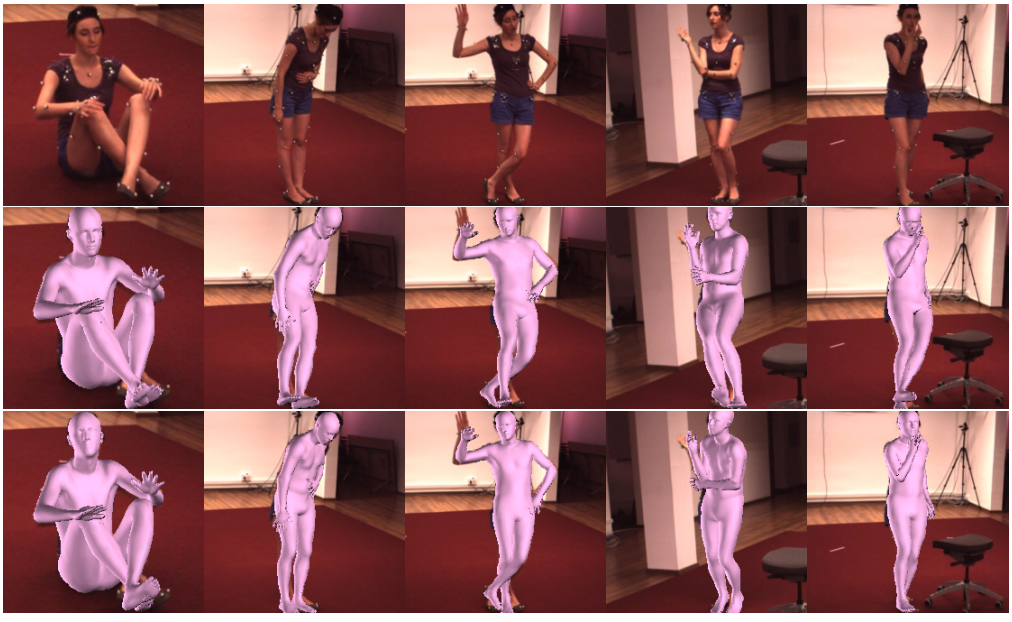


Figure 4.7: The results from SMPLify [18] and multi-view SMPLify. From top to bottom: original image, SMPLify [18] and multi-view SMPLify.

respectively. Some examples from the 100×4 images are shown in Figure 4.7. The second row in Figure 4.7 shows the results of SMPLify, while the third row shows the result from the multi-view SMPLify. We can see that the results from the multi-view SMPLify better fit the ground truth and reduce the ambiguity of limbs in 3D space. Especially for the feet and body orientations, multi-view SMPLify has more robust performance than SMPLify which only relies on a single image. We also compute the reconstruction error, PCK and AUC of 3D joint points of the 100×4 images. The results are shown in Table 4.8 and Figure 4.8. We can see from Table 4.8 that multi-view SMPLify can achieve higher PCK and AUC, while the reconstruction error is lower than when using a single image. Figure 4.8 gives the curve of PCK with different thresholds and it also shows that multi-view SMPLify had higher AUC and PCK with 150 mm as threshold. Therefore, the multi-view SMPLify is more stable and reliable for our method and could provide better supervision for training CNN.

Table 4.8: Comparison of the results from using single image and multi-view images.

	PCK↑	AUC↑	Rec.Err.↓
SMPLify [18]	93.9	54.9	70.0
MV SMPLify	**97.4**	**60.7**	**59.2**

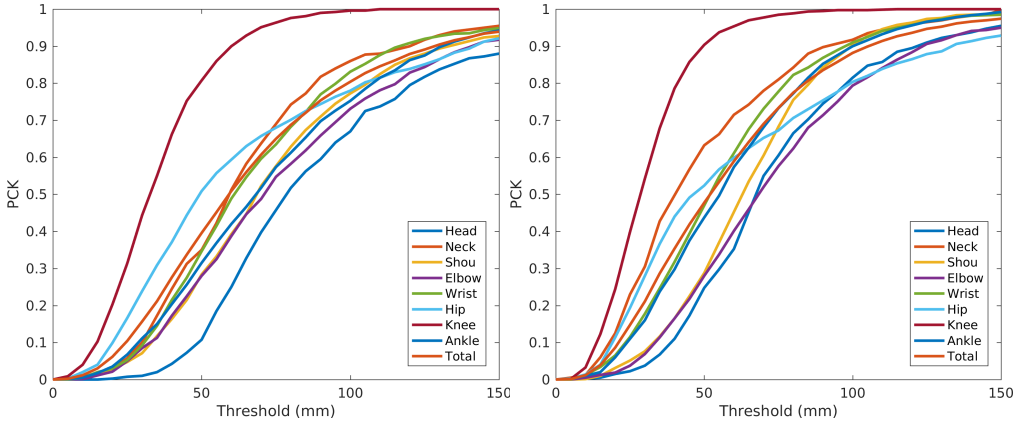


Figure 4.8: The AUC of SMPLify and multi-view SMPLify for different joints. To the left SMPLify and to the right multi-view SMPLify.

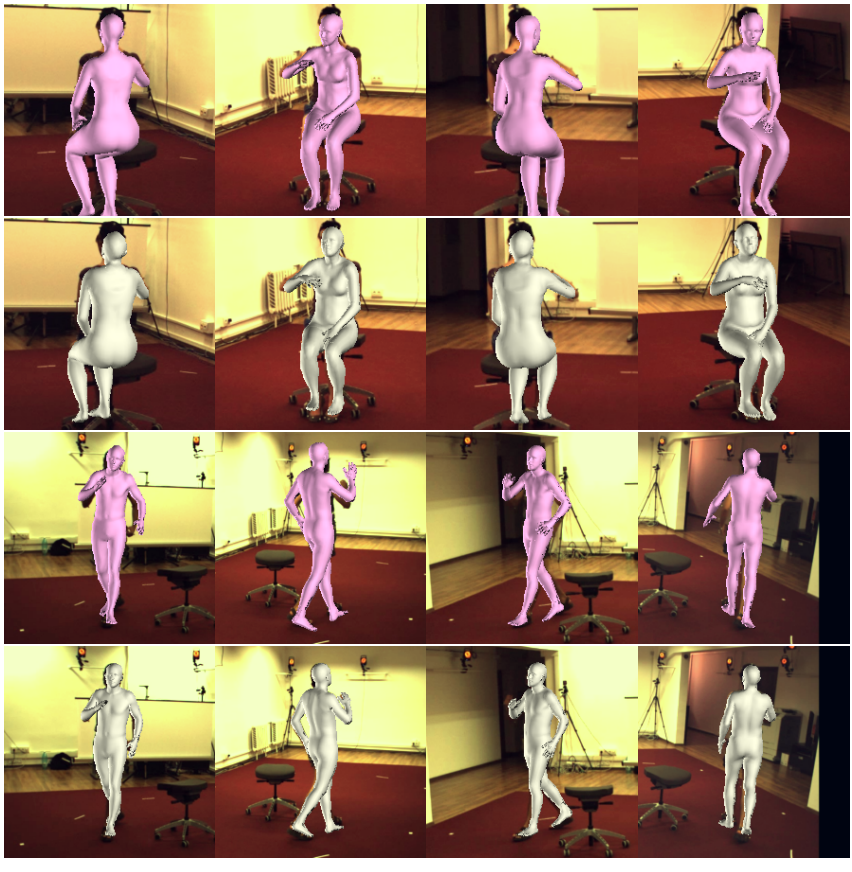


Figure 4.9: The comparison between regressed and optimized SMPL model. The pink models are the results after regression. The white models are the the results after multi-view SMPLify.

In addition, Figure 4.9 shows the comparison of the regressed SMPL model and optimized SMPL model obtained by multi-view SMPLify. In the figure, the pink models are the results of the CNN and the white models are the results after multi-view SMPLify for the multi-view images. We can see that the results after multi-view SMPLify are better because the limbs of the optimized SMPL model are closer to the ground truth, especially for the results of the 3-rd and 4-th rows. This also demonstrates that it is advantageous to use the results of multi-view SMPLify to supervise the network.

4.7 More results

Extra results of our method from the Human3.6M, MPI-INF-3DHP and 3DPW are shown in Figure 4.10. These images show various poses and are captured under both indoor and our scenarios. The first three rows are from Human3.6M, the middle three rows are from MPI-INF-3DHP and the last three rows are from 3DPW. The original image and the 3D model of our method from different view are given for each image. We can see that our method achieve promising 3D pose and shape estimation on the these images. Even for the 3DPW which is only used for testing, the estimated 3D models of our method are also satisfying. This figure demonstrates the effectiveness of our method.

5 Conclusion

In this paper we propose a method to estimate 3D human pose and shape from multi-view images by collaboration between a regression model, a CNN, and an optimization model, multi-view SMPLify. Instead of training the network only relying on single-view image, multi-view images provided by some public datasets are utilized for training. The multi-view images are firstly processed by a CNN to regress the pose and shape parameters of the SMPL model as well as the camera parameters. Then, the multi-view SMPLify takes the output of the CNN as initialization to fit the SMPL model to the multi-view images based on 2D joint points. Multi-view SMPLify achieves better optimized results than single SMPLify, which provides stronger supervision of the training than single-view image. On one hand, our approach sufficiently explores the relations of multi-view images in some public datasets like Human3.6M for the network training. On the other hand, the CNN and multi-view SMPLify form a tight self-supervised framework. We validate our method on three public datasets through comparing to previous work and the results of our method indicate the advantage of using multiple views throughout the training process.

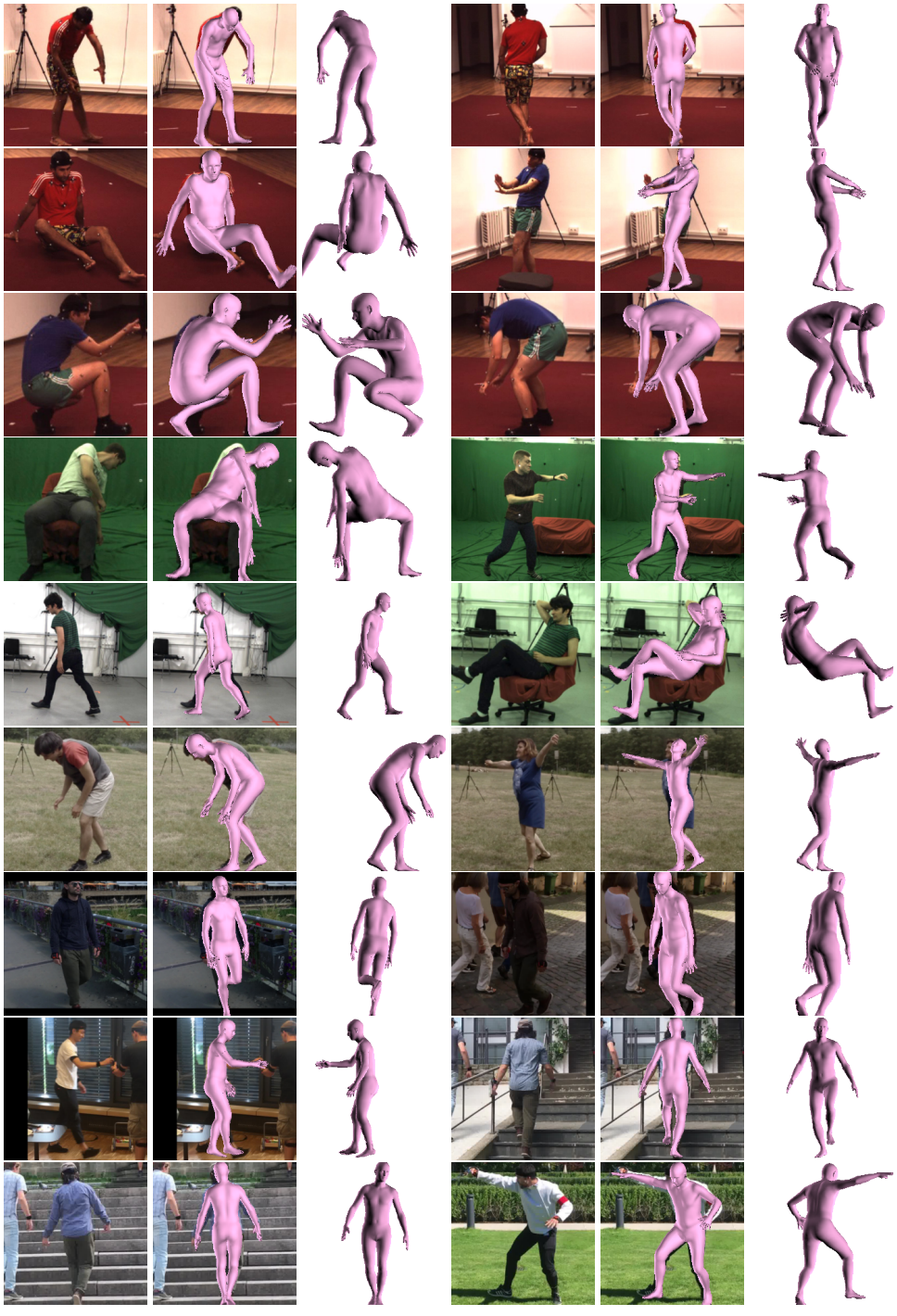


Figure 4.10: The results of our method on the three datasets. The first three rows are from Human3.6M, the middle three rows are from MPI-INF-3DHP and the last three rows are from 3DPW. For each example, the original image, the 3D model and the 3D model from anther view are given.

Paper IV

Chapter 5

Learning to Implicitly Represent 3D Human Body From Multi-scale Features and Multi-view Images

Abstract

Reconstruction of 3D human bodies, from images, faces many challenges, due to it generally being an ill-posed problem. In this paper we present a method to reconstruct 3D human bodies from multi-view images, through learning an implicit function to represent 3D shape, based on multi-scale features extracted by multi-stage end-to-end neural networks. Our model consists of several end-to-end hourglass networks for extracting multi-scale features from multi-view images, and a fully connected network for implicit function classification from these features. Given a 3D point, it is projected to multi-view images and these images are fed into our model to extract multi-scale features. The scales of features extracted by the hourglass networks decrease with the depth of our model, which represents the information from local to global scale. Then, the multi-scale features as well as the depth of the 3D point are combined to a new feature vector and the fully connected network classifies the feature vector, in order to predict if the point lies inside or outside of the 3D mesh. The advantage of our method is that we use both local and global features in the fully connected network and represent the 3D mesh by an implicit function, which is more memory-efficient. Experiments on public datasets demonstrate that our method surpasses previous approaches in terms of the accuracy of 3D reconstruction of human bodies from images.

1 Introduction

In various fields like animation, games and medical research, 3D human body models are needed, and hence the task of acquiring such 3D human body models is a crucial task. However, capturing and reconstructing detailed 3D human body models from monocular images is a quite challenging task in computer vision and graphics, due to the diversity and complexity of the human body in real scenes. In order to conveniently build accurate 3D human body models from images or other 3D input, a large number of approaches have been proposed during the past several decades.

Consumer depth camera based methods have achieved progress in building 3D human body models [186, 17]. The problem is that the depth image is often sensitive to noise with the changing of environment, while some high quality depth camera systems are expensive and difficult to deploy. Therefore, many attempts to estimate 3D human body from monocular images have been done in research. With the success of deep learning in computer vision, learning to estimate 3D human body models from monocular images has achieved remarkable progress recently. In general, two different approaches can be summarized: (1) parametric model based [10, 113] and (2) model-free based [156, 26]. In the first category, a parametric human body is adopted to provide strong prior information and is fitted to e.g. joint points through optimization to estimate the 3D model. This can result in good pose estimation even for bodies partially covered by objects. However, the detailed appearance of human bodies like clothes and hair are typically not be preserved [18, 68, 3]. Recently, model-free based methods try to reconstruct 3D objects from monocular images or 3D point clouds directly, which has shown promising performance on preserving the shape details of the human body as well as the pose [5, 156, 26].

One way to learn to reconstruct 3D human bodies from images is based on fully-convolutional networks that can extract feature maps spatially aligned with the monocular images [178]. The 3D mesh, which is represented as its x-y-z locations directly, can be applied in the fully convolutional manner. The disadvantage is that it requires large amounts of memory to store the features of each voxel in the 3D mesh. By contrast, implicit function representation of 3D models [29] has shown advantages. It is a memory efficient way and it has been used in 3D deep learning for shape completion and reconstruction from incomplete 3D objects and so does the 3D human body reconstruction [120, 24, 134, 26]. Such methods based on learning implicit functions for shape reconstruction often use deep neural networks to extract features from 3D points and define a classifier network to infer if a given 3D point locates in or out of the surface. Recently, a method for voxel super-resolution and shape completion from various 3D inputs proposed a multi-scale feature extraction structure [26] based on an implicit representation 3D model.

Inspired by the above research, we propose a multi-scale features based method to learn

an implicit function for 3D human body reconstruction from multi-view images in this paper. After sampling some 3D points near the boundary of the ground-truth mesh, we firstly project them to the multi-view images. Then, the multi-view images are fed into our model to extract multi-scale features and query the occupancy values. The model for feature extraction contains several stages of end-to-end fully convolutional neural networks (hourglass networks). Instead of keeping the size of feature maps of each hourglass network the same, we downsample the feature map of each hourglass network, which produces multi-scale feature maps that encode both global and local information of the given multi-view images. Since we use hourglass networks, the feature map can be spatially aligned to the original images, which can choose the features of those 3D points from the mesh. The multi-scale features as well as depth information are combined as a new feature. Finally, we feed the new feature into a fully connected classifier network which gives the value of an implicit function to the corresponding point. Therefore, through querying enough points of a 3D grid, the 3D model can be inferred by a threshold from the 3D grid. The overview of our method is shown in Figure 5.1.

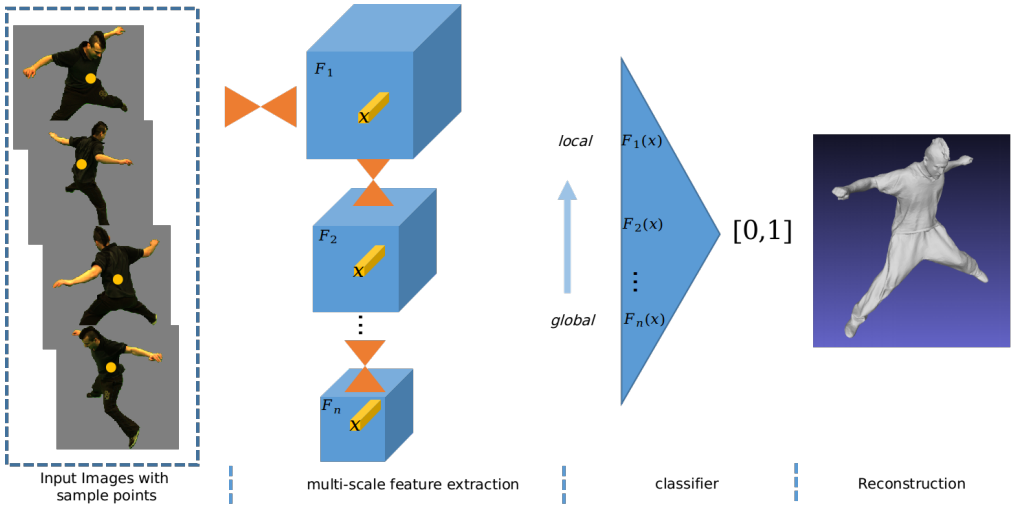


Figure 5.1: The overview of our method. Multi-view images are fed into our model. F_i is the feature grids extracted by the hourglass network shown as the orange "►◄". For the point in the images (yellow "●"), the corresponding features can be extracted from multi-scale features. The features are passed to a classifier to decide the value of the implicit function representation. After training the model, we can reconstruct the 3D mesh from the implicit function.

Our method has two main contributions. Firstly, it is a model-free implicit function based method which can better estimate the shape details in a memory efficient way. Secondly, the novel multi-scale features encode local and global information of the query points, which leads to better performance than approaches based on single scale features. In order to demonstrate the effectiveness of our method, we compare to previous work on synthetic and real clothed human body datasets quantitatively and qualitatively.

2 Related work

Approaches for 3D human body reconstruction can be divided into two categories according to whether a template is used for reconstruction. We call them model-based and model-free methods.

Model-based methods require the use of a template to fit some prior information. The template is often defined by a parametric human body model like SCAPE [10] or SMPL [113] that can provide strong prior information on the pose and shape. Then, various prior cues mainly including silhouettes [13], 2D/3D joint points [53, 25], and depth images [186] were utilized and the parametric human body model was fitted to the cues through optimization. In order to improve the accuracy of prediction, it was common to use more than one prior cue, see e.g. [53, 17]. For the fitting problem, novel optimization algorithms were also discussed and explored [101, 55]. However, the prior information was always manually obtained to ensure the accuracy. With the emergence and success of deep learning on computer vision tasks, 2D/3D joint points can be automatically predicted by training deep neural networks [142, 185, 129], which has become one of the most popular prior cues for fitting. Many automatic methods based on 2D/3D joint points obtained by deep neural networks were proposed [18, 68, 196, 3] since the success on human pose estimation based on learning. Additional cues like silhouettes [68, 196, 3], video [68, 3], and depth information [199] were also added to improve the results. Deep learning can not only predict prior information from images, but also regress the parameters of parametric human bodies directly from images. In [81], an end-to-end structure was proposed to regress the pose and shape parameters of SMPL, obtaining more robust results than fitting-based methods. Recently, many novel approaches based on deep neural networks have been proposed through improving the architecture of networks or defining new loss functions for training [138, 89, 140, 88]. In [88], a weakly self-supervised method combining fitting and regression was proposed and achieved competitive performance. In [5], detailed 3D human body geometry was estimated by translating texture maps obtained by DensePose [8] to augmented SMPL models from a single image. Although these methods have achieved extraordinary performance on human pose and shape estimation, they fail in detailed estimation of human bodies with loose clothes.

Model-free methods for 3D shape reconstruction from images or 3D input like point clouds and low-resolution voxel grids have also made impressive progress recently. Instead of relying on a template, the methods mainly regress 3D volumes explicitly or implicitly, giving better ability to represent detailed shape including clothes and hair. In the beginning, some research [74, 178, 203, 126] proposed to regress the 3D volume containing the complete human body model through explicitly representing 3D shape. In [74], a method for face reconstruction from a single image was proposed by regressing the vertices of a volume using a CNN. Then, Varol *et al.* [178] proposed to extend to the full body, and a

3D volume was fed into the network directly. These methods require large amounts of memory for representing the 3D volume during training. By contrast, learning an implicit function for 3D shape reconstruction and completion is memory efficient and has also obtained some progress recently [69, 120, 24, 134, 156, 26, 132]. The methods only took into those points sampled around the 3D mesh to regress the value of implicit function, and thus, can reduce the memory footprint during training. In [134], the Truncated Signed Distance Function (TSDF) [29] was used for implicitly representing the 3D mesh and the authors proposed a CNN-based method to learn the continuous TSDF for shape completion and interpolation. Instead of using TSDF, binary occupancy was used in [120, 24] and they can reconstruct 3D shape from single image. These methods have inspired 3D human body reconstruction without the use of templates. In [69], the authors trained a CNN for feature extraction and a classifier for implicit function from multi-view images. Satio *et al.* proposed to use an end-to-end method to extract features and depth to learn the implicit function and they can estimate a 3D mesh from both single and multi-view images [156]. Chibane *et al.* proposed an implicit network which extracted multi-scale features to reconstruct 3D human body from low-resolution volume and partial point cloud [26]. In [132], the authors proposed a novel tetrahedral TSDF method for 3D human body reconstruction which built the TSDF between an inflated SMPL model and a deformed ground truth 3D mesh. Model-free methods based on implicit function do not rely on a template, which can better estimate the clothed human body. Therefore, the problem has attracted much attention since the success of the implementation of deep learning. Our method also belongs to the model-free methods.

3 Proposed Method

This section presents the details of our method to reconstruct 3D human bodies from multi-view images, which is shown in Figure 5.1. Firstly, the background on the implicit function representation of surfaces is introduced in section 3.1. Then, we describe the the multi-scale feature extraction and querying in section 3.2. Finally, the loss function for training is defined in section 3.3.

3.1 3D Model Using an Implicit Function

Implicit representation is a memory efficient way to represent 3D models because we do not need to store all voxels of a 3D volume. Given a watertight mesh $\mathcal{S}$ and its corresponding dense grid points X, the implicit function gives a value of those point around the surface instead of storing the x-y-z locations of the points in the volume. The value is usually represented in two forms: the signed distance from points to mesh or binary occupancy. The surface can be decided by a threshold in the grid as long as all the points in the gird

are given values to decide whether points are outside or inside of the mesh. The two forms have similar performance to represent 3D surfaces implicitly. In our method we use the occupancy values to define the implicit function. In general, the implicit function can be defined as:

$$f(X) : \mathbb{R}^3 \rightarrow [0, 1], \tag{5.1}$$

where X is the 3D point around the given 3D mesh $\mathcal{S}$. If the point is inside the surface, the value of f is 0. Otherwise, the value of f for the point is 1. As a result, the surface can be decided by a decision boundary with a threshold 0.5.

Based on the definition of the implicit function, the mesh can be reconstructed as long as the implicit function $f(\cdot)$ can be learned correctly. The advantage is that the implicit function can represent the 3D mesh with any resolution since any points in a 3D volume containing the mesh could be given its corresponding occupancy value through the estimated $\hat{f}(\cdot)$. It is also a memory-efficient representation because we do not use all voxels in the 3D volume. Then, the mesh can be reconstructed through algorithms such as marching cubes on the signed occupancy grid, which enables detailed 3D shape reconstruction. Therefore, the key problem is to learn the implicit function $f(\cdot)$ for our method.

3.2 Multi-scale Features Extraction and Querying

Figure 5.2 shows the projection from sampled 3D points of mesh to 2D multi-view image planes, feature extraction from multi-view images and query of the features. In the following, we describe the details of each step.

The first problem is how the 3D mesh and 2D multi-view images build the correspondence. Similar to previous work [156, 26], we uniformly sample N 3D points $X_i, i = 1, ..., N$ on the ground truth of 3D mesh instead of using all the vertices, which can reduce the consumption of memory. This can be done through an efficient ray tracing algorithm [181]. Note that the mesh should be water-tight to use [181]. Then, random displacements are generated by normal distribution $n_i \sim \mathcal{N}(0, \sigma)$ where σ is the standard deviation and they are added to the 3D points X_i, i.e., $\hat{X}_i = X_i + n_i$. This ensures that those points far from the mesh are not selected, which can reduce unnecessary predictions. The corresponding occupancy value $f(\hat{X}_i)$ of $\hat{X}_i$ can be obtained according to the outside or inside of $\hat{X}_i$. As shown in Figure 5.2, the green points are outside of the mesh, while the red points are inside of the mesh. Given the camera matrices of M views $\Pi_j, j = 1, ..., M$, these 3D points $\hat{X}_i$ can be projected to the M image planes through $\Pi_j(\hat{X}_i)$ and we denote these 2D points as x_{ij} for the i-th point $\hat{X}_i$ and j-th camera Π_j.

Then, we need to get the features of x_{ij} from the multi-view images. We form the multi-view images as a small batch and compute the features of each image. As shown in Figure 5.1, for the j-th view image I_j, n hourglass networks are utilized to create multi-scale feature grids

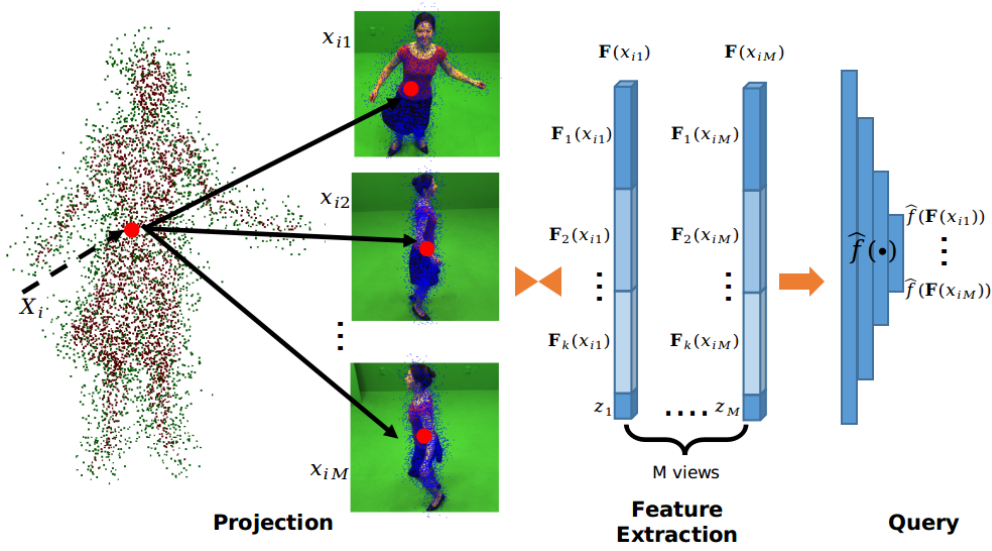


Figure 5.2: The example of projection from 3D points to multi-view images, the multi-scale features extraction and query the multi-scale features.

$\mathbf{F}_1^{(j)}, ..., \mathbf{F}_n^{(j)}$. The hourglass network has two convolution layers and two deconvolution layers and a max pooling layer is followed after each hourglass. Therefore, for the k-*th* feature grid $\mathbf{F}_k^{(j)}$, its resolution is $\mathbb{R}^{H_k \times W_k \times C}$ where H_k and W_k will decrease as 2^{k-1} for the k-th hourglass network. C is the number of convolution kernels and we set it as 256 in our model. This will make the feature grid $\mathbf{F}_k$ at early stage capture the local information, while $\mathbf{F}_k$ at the last stage encodes global information. Then, we can extract features from $\mathbf{F}_k$ at the corresponding position of point x_{ij} in the image plane through interpolation since hourglass network extract network through a fully convolutional way. We denote the feature vector at the k-th stage as $\mathbf{F}_k^{(j)}(x_{ij}) \in \mathbb{R}^{256}$. Totally, the features from all the stages as well as the depth $z^{(j)}$ of x_{ij} at the j-th image are formed as the final feature for the point x_{ij} and we denote it as:

$$\mathbf{F}^{(j)}(x_{ij}) = \{\mathbf{F}_1^{(j)}(x_{ij}), ..., \mathbf{F}_n^{(j)}(x_{ij}), z^{(j)}(x_{ij})\} \tag{5.2}$$

Comparing to PIFu in which features at each stage $\mathbf{F}_f$ are the same scale and are queried separately, we classify the new feature $\mathbf{F}^{(j)}(x_{ij})$ which is formed by multi-scale feature grid and contains local, global and depth information. This feature can better encode spatial information of the image. For the M view images, there will be M feature vectors $\mathbf{F}^{(j)}(x_{ij}), j = 1, ..., M$ for points $x_{ij}, j = 1, ..., M$. These new features are fed into the classifier in Figure 5.1 which is a fully-connected network, respectively. This fully-connected network has four layers as $(1024, 512, 128, 1)$. We denote it $\hat{f}$ and the output of this clas-

sifier is the $[0, 1]$ indicating whether the point is outside or inside of the mesh:

$$\hat{f}(\mathbf{F}^{(j)}(x_{ij})) : \mathbb{R}_1 \times ... \times \mathbb{R}_n \times \mathbb{R}^1 \rightarrow [0, 1] \tag{5.3}$$

Since we have M view images, the feature vectors of each image will be classified separately, leading to better localization of the 3D points. In this classifier, not only local and global information is contained in the input feature vector, but the depth information is also given. This enables the whole learning process to better recover the details.

3.3 Training

According to the above multi-scale extraction and querying process, the loss function for training can be established. For given multi-view images $I_i, i = 1, ..., M$ and its corresponding 3D mesh $\mathcal{S}$, we sample N 3D points from $\mathcal{S}$ and add displacements on these points to get $\hat{X}_i, i = 1, ..., N$. The loss function for learning the implicit function to represent the 3D human body model is defined as:

$$\mathcal{L}_f = \sum_{i=1}^{N} \sum_{j=1}^{M} L(\hat{f}(\mathbf{F}^{(j)}(x_{ij})), o(\hat{X}_i)), \tag{5.4}$$

where x_{ij} is the 2D projection on the j-*th* image for point $\hat{X}_i$. $o(\hat{X}_i)$ is the ground truth of occupancy value for $\hat{X}_i$. $L(\cdot)$ is the standard mean square error loss between $\hat{f}(\mathbf{F}^{(j)}(x_{ij}))$ and $o(\hat{X}_i)$. Through minimizing $\mathcal{L}_f$, the multi-stage hourglass networks and the classifier fully connected network are trained.

After training the multi-stage hourglass network for feature extraction and the fully connected network for classifying, we can query any 3D dense grid. As long as the points in the dense grid is given occupancy values, the 3D model can be extract through classical marching cubes algorithm. Therefore, this method will not be limited by parametric human bodies and is more flexible to process complex human bodies.

4 Experiments

Several experiments are designed to evaluate our method on two dataset: Articulated dataset [179] and clothed auto person encoding (CAPE) dataset [116]. The Articulated dataset contains ten subjects and each subject is captured from eight views. The ground truth of 3D meshes and the camera parameters of the eight views are also given. Totally, there are 2000 frames in this data and we split them 80% for training and 20% for testing. Another dataset is called CAPE which is a synthetic clothed human body dataset. It contains 15

subjects and every subject performs many different actions. In this dataset, 3D meshes of the frames are given, and thus, this dataset is very large. We extract a small subset from the original CAPE. For every action of each subject, the frames at $80, 85, 90, 95, 100$ are taken out if the action has more than 100 frames. Then, we render the mesh into 512×512 images using four cameras: front, right, back and left. The examples from the small CAPE dataset can be seen in the last two rows of Figure 5.5. Totally, there are 2910 frames in the small CAPE and we also split it 80% for training and 20% for testing.

In our experiments we used four-view images to estimate the 3D model and the features were extracted by four stages hourglass networks. All input images were cropped and resized to 256×256 tightly containing the human body in the images. Firstly, a convolution layer and max pooling layer processed the input images and the output features are $128 \times 128 \times 64$. Then, the four-stage hourglass networks were advocated to extract multi-scale features and the size of each feature grids were $128 \times 128 \times 256$, $64 \times 64 \times 256$, $32 \times 32 \times 256$ and $16 \times 16 \times 256$, respectively. In order to train our model, we sampled $10,000$ points from the ground truth mesh and the random displacements generated by $\mathcal{N}(0, 0.05)$ were added to the 3D points. These 3D points with random displacements were projected to the corresponding four-view images and the features of these points were interpolated from the feature grids. The four-view images were formed as a small batch to input our model. The batchsize in our code was set to two due to the limitation of the GPU, i.e., the true number of input images was 2×4. We trained our model on the Articulated dataset and CAPE dataset, respectively. For the two datasets, the model was trained for 12 epochs, which took about 7 hours for each dataset on a NVIDIA TITAN X GPU with 12 GB.

For qualitative and quantitative comparison, we compared to three previous methods: SPIN [88], DeepHuman [203] and PIFu [156]. SPIN [88] reconstructed 3D model through estimating the pose and shape of the SMPL model. DeepHuman [203] regressed the 3D volume based on an estimated SMPL model. PIFu [156] learned an implicit function to reconstruct the 3D human body model. Note that the results of SPIN and DeepHuman in our experiments were obtained by using the trained model in the original paper, while PIFu was trained by our training datasets and was evaluated on our testing dataset because the dataset used in PIFu was not free. The metrics for comparison contain point-to-surface Euclidean distance (P2S) from the vertices on the predicted mesh to the ground truth mesh (Lower is better), volumetic intersection over union (IoU) measuring how well the predicted mesh (Higher is better) and Chamfer-L_2 which shows the accuracy and completeness of the surface (Lower is better). All of them are measured as *cm* and the definition of these metrics can be referred to [26].

4.1 Quantitative results

Table 5.1 and 5.2 showed the P2S, Chamfer-L_2 and IoU of the testing dataset of Articulated and CAPE obtained by SPIN [88], DeepHuman [203], PIFu [156] and our method, respectively. Since SPIN and DeepHuman relied on the parametric human body model SMPL [88] which may have different coordinates with the ground truth of 3D meshes, we firstly normalized the predicted 3D model and true 3D model. Then, the predicted 3D model was registered to the ground truth meshes through the iterative closest point (ICP) algorithm. Through normalization and rigid registration, the estimated 3D mesh was well registered to the ground truth, which ensured that the metrics are calculated correctly. We can see from the two tables that the P2S, Chamfer-L_2 obtained by our method were the lowest on the two datasets, while the IoU of our method was the highest on the two datasets. This demonstrated that our method achieves the best 3D reconstruction on both Articulated dataset and CAPE dataset comparing to the previous methods. In order to clearly show the quantitative results, Figure 5.3 shows the P2S of the four methods on the testing samples of the Articulated and CAPE. We can see that our method achieves the lowest P2S on the most samples of the two datasets. This demonstrates the same conclusion as the Table 5.1 and Table 5.2.

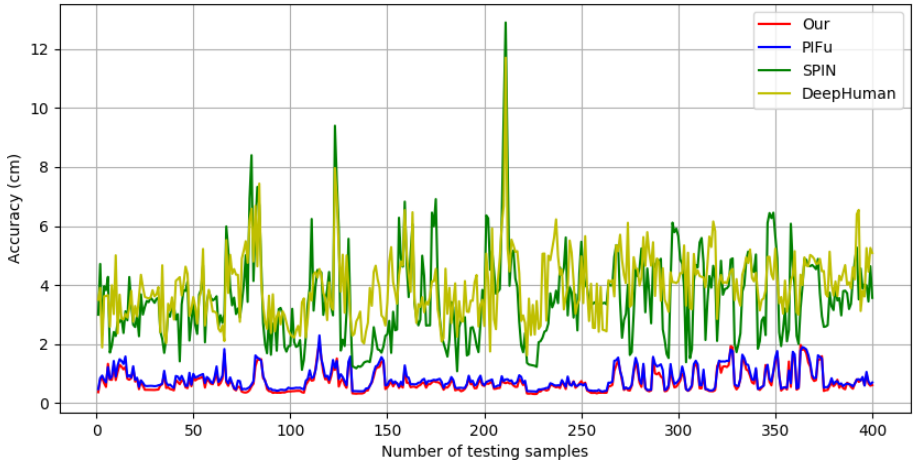
Table 5.1: Quantitative evaluation on Articulated dataset for four-view reconstruction

	P2S ↓	Chamfer-L_2 ↓	IoU ↑
SPIN [88]	3.5206	0.2679	0.3506
DeepHuman [203]	3.9448	0.2675	0.3742
PIFu [156]	0.8194	0.0210	0.8255
Our	**0.7332**	**0.0194**	**0.8484**

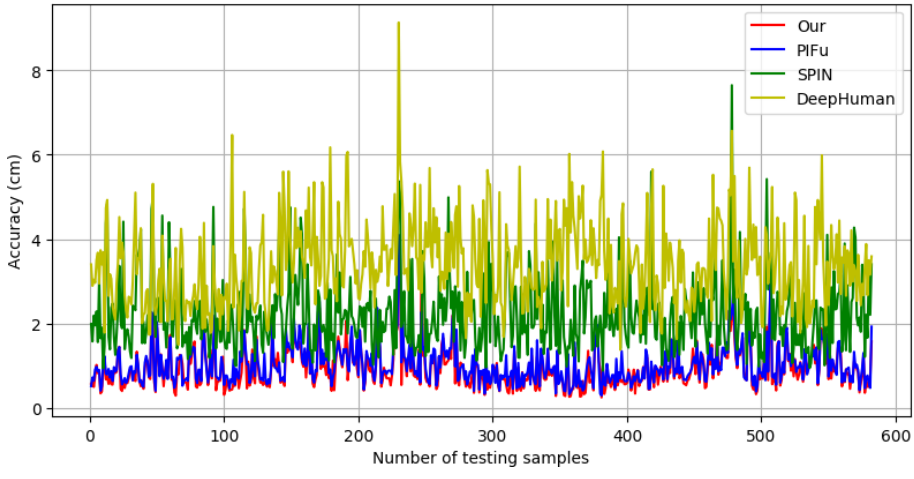
Table 5.2: Quantitative evaluation on CAPE dataset for four-view reconstruction

	P2S ↓	Chamfer-L_2 ↓	IoU ↑
SPIN [88]	2.2134	0.1271	0.4044
DeepHuman [203]	3.4028	0.1850	0.3861
PIFu [156]	1.0330	0.0212	0.7571
Our	**0.9482**	**0.0196**	**0.7829**

For SPIN [88] and DeepHuman [203] which estimated 3D model from single image based on SMPL model, their results were easily affected by the given image. If some parts of the human body can not be seen in the image because of self-occlusion, the two methods may obtain incorrect estimation. PIFu [156] and our method estimated the 3D human body model through an implicit function and could improve the accuracy significantly. However, PIFu [156] used features with the same scale, which cannot fully reflect the spatial



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



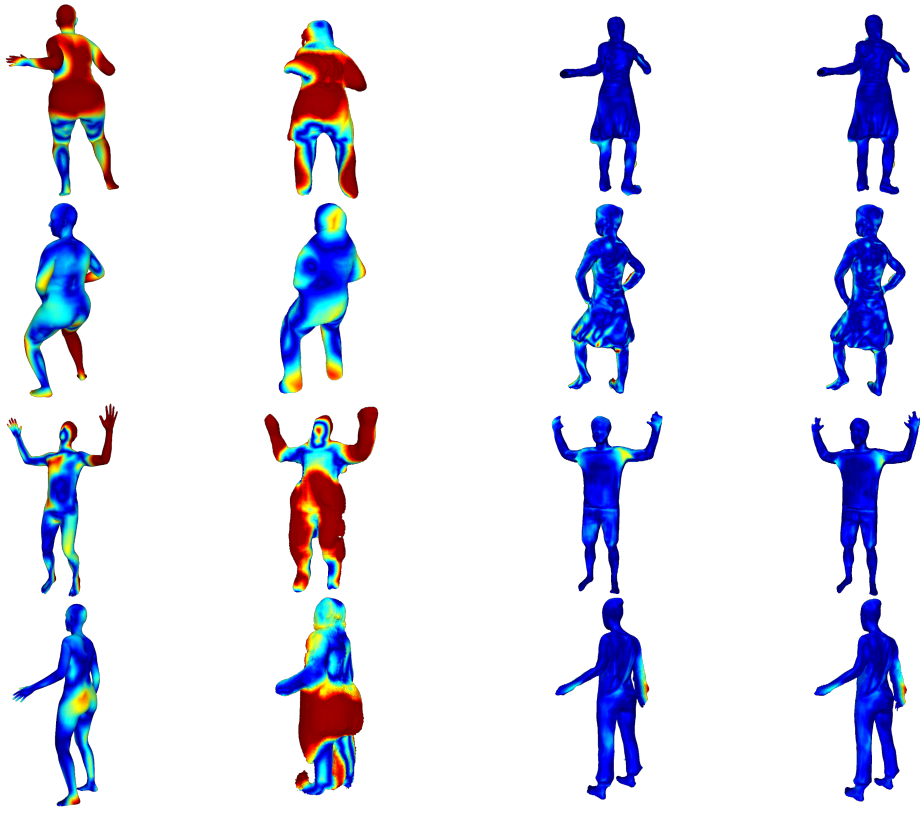
(b) The P2S of the testing dataset of the CAPE dataset for different methods.

Figure 5.3: 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.

information. In our method, we proposed to use multi-scale features to fully use local and global information, and thus, our method can achieve better results.

In Figure 5.4, several examples from Articulated and CAPE are shown to demonstrate the errors from points of predicted mesh to the ground truth mesh. The first two rows are from Articulated and the last two rows are from CAPE. The higher errors are shown as red, while the blue means lower errors. It shows that the results of PIFu [156] and our method have smaller errors. SPIN [88] had better results on CAPE than the results on Articulated because the human pose in CAPE was simple. DeepHuman [203] regressed 3D models based on

the results of SPIN, but the results were not good because this model was not retrained by our dataset, which meant that this method had poor generalization. Our method achieved smaller error than PIFu [156] on the given examples, which was consistent with Table 5.1 and 5.2.



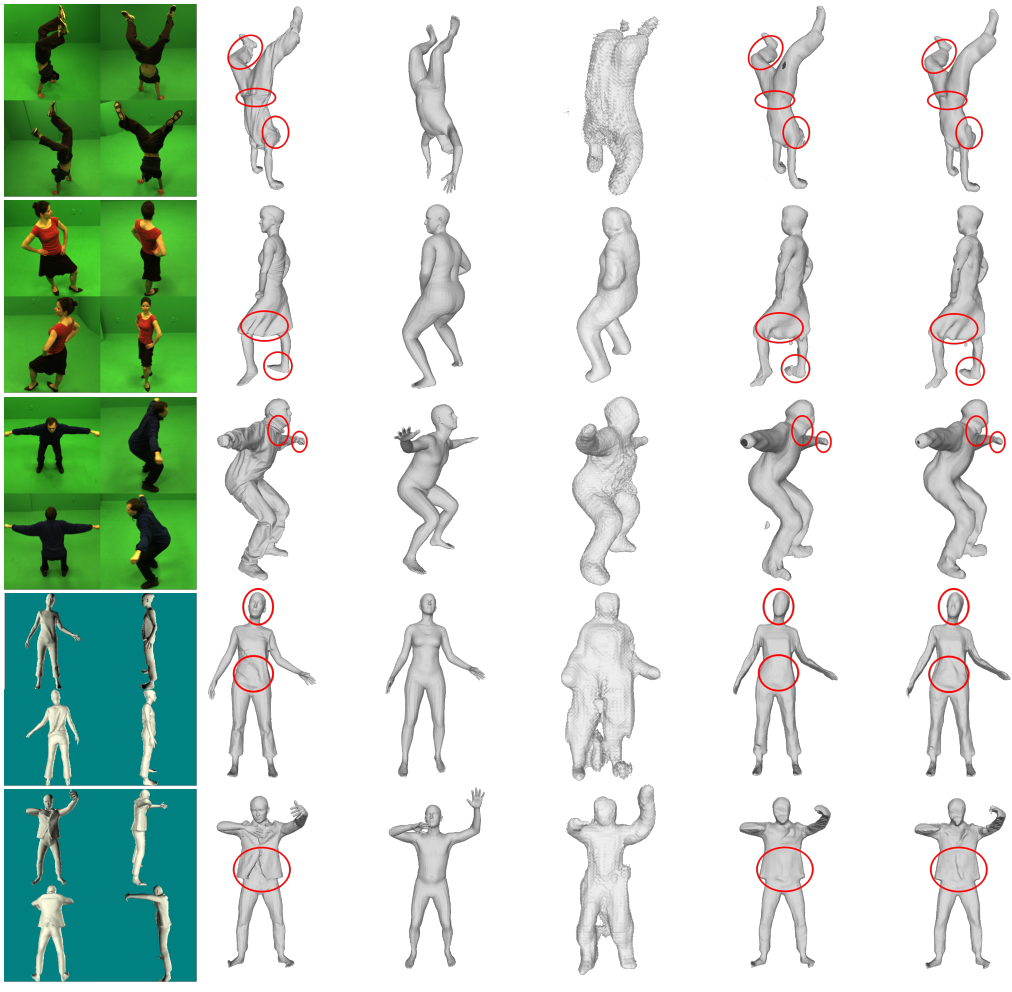
(a) SPIN [88] (b) DeepHuman [203] (c) PIFu [156] (d) Our

Figure 5.4: The errors from points of predicted 3D models to the ground truth 3D meshes. The first two rows are from Articualted [179] and the last two rows are from CAPE [116]. From left to right columns: SPIN [88], DeepHuman [203], PIFu [156], Our Method.

4.2 Qualitative results

In Figure 5.5, some qualitative results of SPIN [88], DeepHuman [203], PIFu [156] and our method from Articulated and CAPE dataset are demonstrated. The first three rows are examples from the Articulated dataset and the last two rows are examples from the CAPE dataset. We can see from the figure that the estimated 3D models of SPIN [88] and DeepHuman [203] have the correct pose compared with the original 3D models. However, the two methods cannot obtain the details about the clothes. Although DeepHuman [203]

tried to infer the clothes, the original trained model of DeepHuman cannot achieve good estimation on our testing dataset due to its poor generalization. For the results of PIFu [156] and our method, the areas indicated by red circles showed that our method achieves better estimation on the clothes because our estimated 3D human body showed more details on the wrinkles of the clothes. Especially for the second row in which the human body wore a skirt, our method achieved more details on the wrinkles of the skirt.



(a) Images (b) GT (c) SPIN [88] (d) DeepHuman [203] (e) PIFu [156] (f) Our

Figure 5.5: The qualitative results of previous methods and our method. The first three rows are from Articulated and the last two rows are from CAPE. From left to right columns: the original four-view images, the ground truth of 3D model, the results of SPIN [88], DeepHuman [203], PIFu [156] and our method.

4.3 Discussion on the Number of Views

In our method the number of views is an important factor to the final reconstruction. In order to demonstrate the effect of the number of views, we also trained our model using eight-view images to compare the results of four views on the Articulated dataset. Table 5.3 lists the quantitative results of four views and eight views on the testing data. The table demonstrates that the accuracy of the estimated 3D model is higher when eight-view images are utilized to train the model. The P2S reduces to 0.43 *cm*, Chamfer-L_2 reduces to 0.052 *cm*, while the IoU increases to 88.75%. This is understandable because eight-view images provide more prior information for the learning. Therefore, it is useful to get better results by using more images in our method.

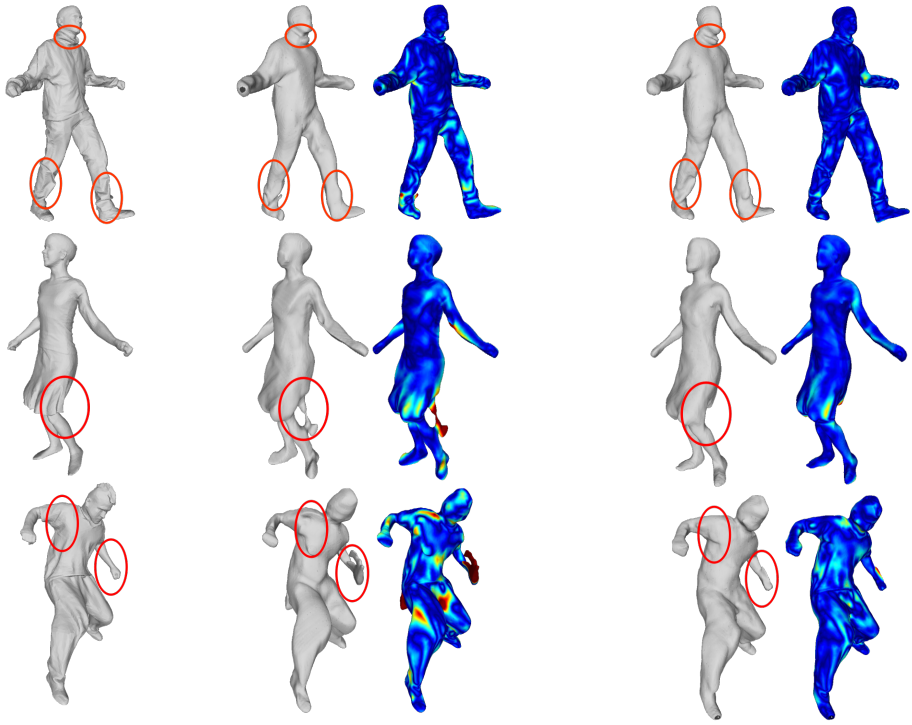
Table 5.3: Quantitative evaluation on Articulated dataset for four-view and eight-view reconstruction

	P2S ↓	Chamfer-L_2 ↓	IoU ↑
four-view	0.7332	0.0194	0.8484
eight-view	**0.4302**	**0.0052**	**0.8875**

In Figure 5.6, several examples from the Articulated dataset on the 3D reconstruction errors are also shown. This figure clearly demonstrates that the estimated 3D meshes from eight-view images have smaller errors. The 3D model from eight-view images can reduce some incorrect parts which existed in the 3D models four-view images, for example, the second example in the figure. Besides, The details of the clothes indicated by the red circle are better recovered from the eight views. The wrinkles of the clothes are better shown in the eight-view results as shown in the first example of Figure 5.6.

5 Conclusion

In this paper we proposed a method to reconstruct detailed 3D human bodies from multi-view images based on an implicit function learned from multi-scale features. Our model consists of multi-stage hourglass networks to extract multi-scale features and a fully connected network to classify the features. The hourglass networks extract the feature grids which are spatially aligned to the original images so that the features of the corresponding points in the images can be obtained. The multi-scale features as well as the depth form a new feature which contains both local and global spatial information. The fully connected network then classifies the new feature as inside or outside of the 3D mesh. Through querying those points around the 3D mesh, we can implicitly represent the 3D models. The experiments demonstrate that our methods can reconstruct detailed 3D models from multi-view images with higher accuracy. The drawback of our method is that the training dataset



(a) GT (b) Four views (c) Eight views

Figure 5.6: The examples of 3D models and errors of four-view images and eight-view images on Articulated dataset.

in our method is limited, and thus, our method does not show good generalization. In the future we will introduce more human bodies to improve our method.

Paper v

Chapter 6

Detailed 3D Human Body Reconstruction From Multi-view Images Combining Voxel Super-Resolution and Learned Implicit Representation

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. In order to tackle the problem, we propose a coarse-to-fine method to reconstruct a detailed 3D human body from multi-view images combining voxel super-resolution based on learning the implicit representation. Firstly, the coarse 3D models are estimated by learning an implicit representation based on multi-scale features 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 voxel super-resolution based on an implicit representation is learned through a multi-stage 3D convolutional neural network. Finally, the refined detailed 3D human body models can be produced by the voxel super-resolution 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 voxel super-resolution 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 reconstructions from images with various poses and shapes on both the real and synthetic datasets.

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. However, inferring 3D objects from 2D images is a challenging task in computer vision due to the ambiguity of reprojection from 2D to 3D space. The high freedom of the human body in real scenes further increases the difficulty of the task. Although multi-view systems [77] and laser scanning systems [194] are now able to reconstruct accurate 3D human bodies, these systems remain inconvenient for common users because they are often hard to deploy and expensive. Thus, estimating 3D human bodies from images is more attractive and many approaches have provided possible directions to tackle the problem from advocating the pre-defined parametric human body as template to recent deep learning based route.

Traditionally, 3D human body reconstruction from RGB images mainly depends on the pre-defined parametric human body models. From simple geometric primitives [160] to data-driven models [10, 113], 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 skeleton, 2D joint points and the silhouettes [13, 18, 3]. Such methods have been used for human motion tracking and 3D pose estimation successfully. However, due to the missing detailed appearance on the most parametric human bodies such as clothes and facial expression, the results of these methods are often unclothed, which cannot satisfy the requirements of the realism in many applications.

Benefiting from the great success of deep learning in many computer vision tasks, 3D human body reconstruction from images based on deep learning has also achieved some progress recently. During the past several years, convolutional neural networks (CNN) have shown impressive performance on 2D/3D human pose estimation [142, 129, 8] and human body segmentation [191, 63]. Therefore, some methods automatically estimated 3D human body model from images by fitting the parametric human body to prior cues like the 2D or 3D joint points of human body and silhouettes which can be estimated by the CNN [18, 68, 3, 196]. 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 [81, 138, 140, 89, 88]. However, the 3D human body models obtained by these

methods are still unclothed. Recently, many approaches came up with a refining process based on CNN on the parametric human body to add clothes on the naked 3D human body model. The refining process includes in the image texture translation [5], inferring the surface normal [157] and volumetric regression [203]. 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 that the parametric human body model has the accurate pose with the observed human body because the final estimation will be affected seriously if the prior information is not predicted correctly.

Recently, deep learning on 3D reconstruction like point clouds or voxels from images for some general objects has gained popularity. Explicit volumetric representations are straightforward for learning to infer 3D objects from RGB images [27, 83, 182, 38]. 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 adaptive representations such as octrees, the final resolutions are sill relatively low (e.g. 256^3) [151]. 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 [134, 120, 24, 26] and is attracting much attention. Compared to learn 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 allows 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 [69, 156, 157]. 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) inferring the 3D human body model from multi-view images, and (2) voxel super-resolution from low-resolution voxel grids obtained by (1). In both of the two parts, we attempt to learn an implicit function to represent the 3D models. For the reconstruction of a 3D human body from multi-view images in (1), the structure of multi-stage hourglass networks is designed to produce 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 voxel super-

resolution. Our method is summarized in Figure 6.1.

Our method differs from previous work in three aspects. Firstly, it is a coarse-to-fine method combining 3D reconstruction from multi-view images and voxel super-resolution into one route to infer 3D human body models. The 3D reconstruction from images produces a coarse result and the voxel super-resolution refines the coarse result to generate a final detailed 3D model. Secondly, the implicit representation for the 3D model is used both in image based 3D reconstruction and voxel super-resolution, 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.

The paper is organized as follows. The introduction and related work of our method are presented in Section 1 and Section 2, respectively. The following Section 3 describes the detailed coarse-to-fine structure of our method and the implementation details including the 3D model reconstruction from multi-view images and voxel super-resolution. 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

We summary the related work on 3D human body reconstruction from images and 3D vision based on deep learning 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 such as SCAPE [10] or SMPL [113] 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 [10]. Some methods fitted SCAPE to the silhouettes and joint points from observed images to recover human pose and shape [13, 161, 53]. With the emergence of Kinect, the depth images were also used for fitting the SCAPE [186, 17, 108]. With the success of deep learning on human pose estimation [129, 117, 8, 20], the joint points can be obtained automatically with high accuracy. In [18], an automatic method for 3D human body estimation was proposed through fitting a novel parametric human body model called SMPL [113] to the 2D joint points predicted by deep learning [142]. Then, more methods turned to use SMPL or pre-scanning models for human body reconstruction based on 3D joint points, multi-view images, video and silhouettes[68, 3, 196, 103, 58].

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 was always unclothed due to the limitation of parametric human body, which limited its realism.

Parametric human body model based regression. Since deep learning has achieved impressive performance on many computer vision tasks, 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 [33, 35], which can only handle the standing pose or very simple poses. In [166], 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 [81]. Since then, many improved methods were proposed through designing novel network structure or using more constraints on the loss function [138, 89, 140, 88, 82, 106, 87]. Pavlakos *et al.* [138] 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 [89], the texture [140], the multi-view images [106], the optimized SMPL model [88] and the video [82, 87]. Although these methods can infer the pose and shape of SMPL model very well, they still obtained unclothed human body models. In order to model the detailed appearance, some methods attempted to refine the regressed SMPL model to obtain the detailed 3D model [2, 178, 204, 95, 5, 203, 132, 70]. In [2], 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. They also proposed some improved methods to obtain better results [95, 5]. In addition to the texture, the explicit representation of 3D human body model was also used in detailed reconstruction. BodyNet [178] added the volume loss function to better estimate the pose and shape of SMPL. DeepHuman [203] refined the appearance of volumetric SMPL model through transferring the image normal to the volumetric SMPL. In [132], a novel tetrahedral representation for SMPL model was used and the detailed model was obtained by learning the signed distance function of tetrahedral representation. Another recent work also refined the normal and color of image to the estimated SMPL model from single image [70].

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, depth images [75, 69, 46, 41, 126]. Jackson *et al.* [75] reconstructed 3D geometry of humans through training an end-to-end CNN to regress the volumes which were provided

in the training dataset. In [46], 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.* [126] 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.* [69] 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.* [41] 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 [27, 83, 182, 38]. Due to limitation of resolution of explicit representation, implicit representation for 3D model based on deep learning has been used for reconstruction of general objects [83, 120, 24, 26]. Inspired by the idea, some methods only for detailed 3D human body reconstruction also proposed based on learning implicit representation. Saito *et al.* [156] 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 [157] was proposed to reconstruct high-resolution detailed 3D human body from images through introducing image normal to PIFu. The coarse-to-fine methods could obtained more accurate reconstruction because more cues were used for the reconstruction.

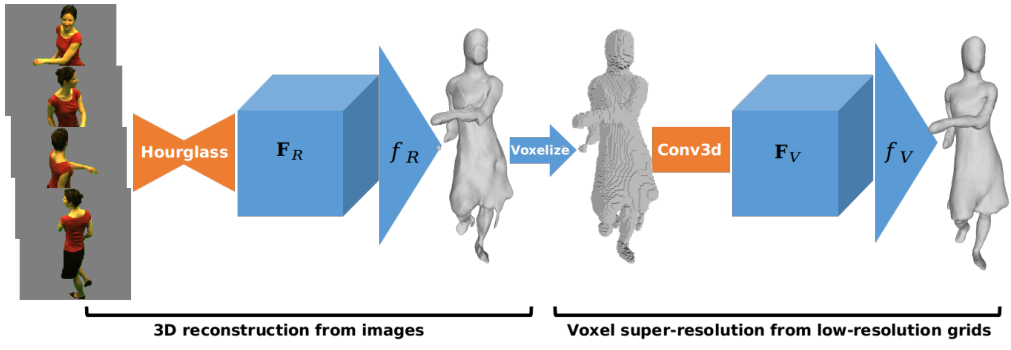


Figure 6.1: The pipeline of our method. It consists of 3D reconstruction from images and voxel super-resolution from low-resolution grids. The 3D reconstruction from images estimates a coarse 3D human body model. After voxelizing to a low-resolution grid, the super-resolution refines the low-resolution grid to obtain detailed model.

3 Method

In this section the details of our method are described. We firstly introduce the background of implicit function to represent the 3D shape. Then, we present the 3D human body reconstruction from multi-view images through learning the implicit representation. Afterwards, an implicit representation based network for voxel super-resolution 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 function 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. Thus, the 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:

$$f(X) := x, X \in \mathbb{R}^3, x \in \{0, 1\}, \tag{6.1}$$

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 generated 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.,

$$f(\mathbf{v}, X) : \mathcal{V} \times \mathbb{R}^3 \mapsto [0, 1]. \tag{6.2}$$

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

$$f(F(\pi(X)), z(X)) : \mathcal{V} \times \mathbb{R} \mapsto [0, 1], \tag{6.3}$$

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 Figure 6.1. It has two parts: (1) 3D reconstruction from images; and (2) Voxel super-resolution from low-resolution grids. 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 [156]. The difference is that we extract Multi-scale Features 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 Figure 6.2.

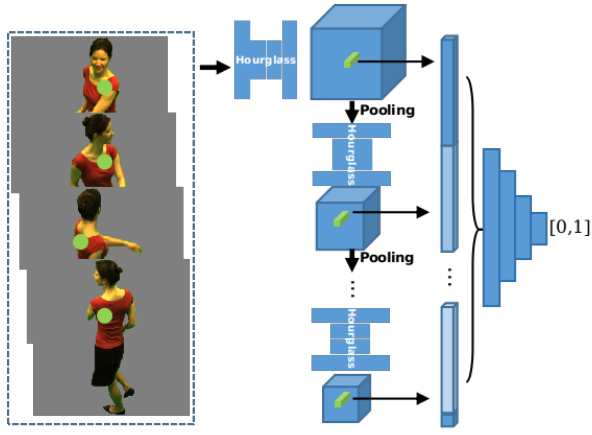


Figure 6.2: The structure of MF-PIFu to learn the implicit representation of 3D human body model. Multi-stage hourglass networks are used for multi-scale feature extraction and a fully connected neural network predicts the occupancy value of the feature.

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

$$g_R(I_i) := \mathbf{F}_R^{(i,1)}, ..., \mathbf{F}_R^{(i,M)}, \tag{6.4}$$

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 nwtwork, 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 $\mathbf{F}_R^{(j,1)}(x) \in \mathcal{F}_j^C$.

After getting the multi-scale features, we need to query the multi-scale features, i.e., predict the occupancy value. The prediction is defined 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:

$$f_R(\mathbf{F}_R^{(i)}(x)) : \mathcal{F}_1^C \times ... \times \mathcal{F}_M^C \times \mathbb{R} \mapsto [0, 1]. \tag{6.5}$$

In contrast to PIFu, we form the features from each stage and the depth as a new feature vector. This new feature encodes both the local and global information of the pixels. The feature grids at an 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 for prediction of occupancy value.

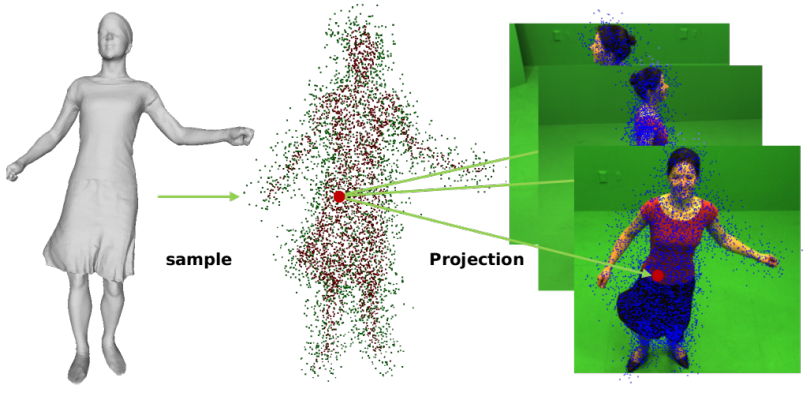


Figure 6.3: Sampling 3D points from 3D model and projecting the points to multi-view images.

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 re-

quired in which $\mathcal{S}$ is the corresponding ground truth of 3D model for the multi-view images I_i. As shown in Figure 6.3, 3D spatial points X_i, $i = 1, ..., K$ are sampled from the 3D model $\mathcal{S}$ and are added random displacements with normal distribution $\mathcal{N}(0, \sigma)$ on the points. 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 $\mathcal{S}$, $o(\hat{X}_i) = 0$. Otherwise, $o(\hat{X}_i)$ is 1. 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:

$$L_R = \sum_{i=1}^{N} \sum_{j=1}^{K} \| f_R(\mathbf{F}_R^{(i)}(x_{ij})) - o(X_j) \|. \tag{6.6}$$

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 due to the multi-view effect. 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, we propose the voxel super-resolution (VSR) to refine the coarse 3D models of MF-PIFu. As shown in Figure 6.4, 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.

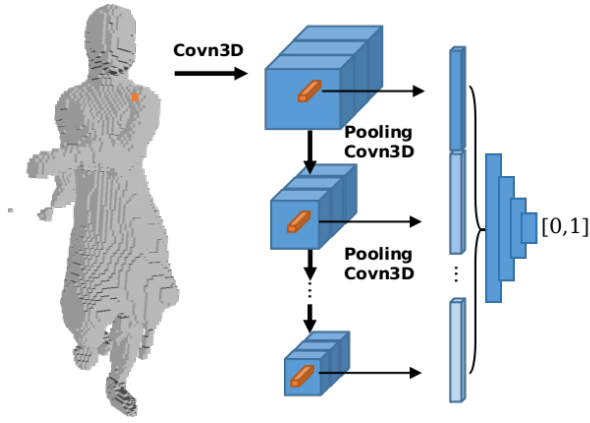


Figure 6.4: The structure of voxel super-resolution based on learning implicit representation. Multi-stage 3D convolutional layers are used for extracting the multi-scale features from low-resolution grid. A fully connected neural network is used for predicting occupancy value of features.

Suppose that 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 Figure 6.4, 3D convolution kernels are utilized to extract 3D feature grids from $\mathcal{V}$. We recursively use n 3D convolution layers 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 convolutional neural network for VSR as $g_V(\cdot)$ and the multi-scale features can be generated as:

$$g_V(\mathcal{V}) := \mathbf{F}_V^{(1)}, ..., \mathbf{F}_V^{(n)}. \tag{6.7}$$

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

$$\mathbf{F}_V(\mathbf{v}) = \{\mathbf{F}_V^{(1)}(\mathbf{v}), ..., \mathbf{F}_V^{(n)}(\mathbf{v})\}, \tag{6.8}$$

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

$$f_V(\mathbf{F}_v(\mathbf{v})) : \mathcal{F}_1 ... \times \mathcal{F}_n \mapsto \in [0, 1] \tag{6.9}$$

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.

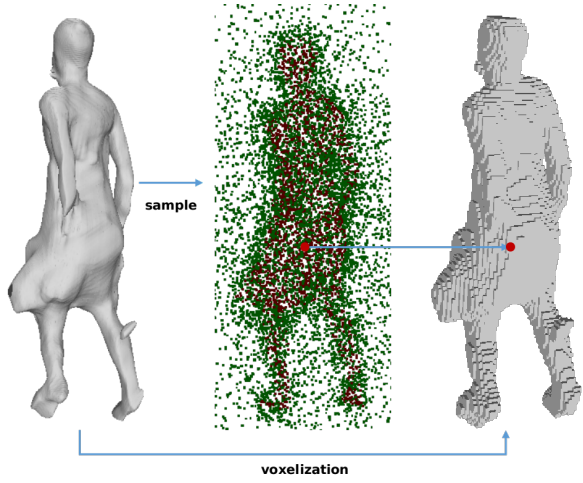


Figure 6.5: Sampling 3D points from 3D model estimated by MF-PIFu and the voxelization of the 3D model estimated by MF-PIFu (The resolution is 128^3). The 3D points can be indexed by the grid coordinates in the low-resolution grid.

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 we take the same strategy as [26] to generate points to be queried, 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 we denote it as $\rho(\hat{\mathbf{v}}_i)$. One example of sampling points and voxelization to a 128^3 grid is shown in Figure 6.5. 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 $o(\hat{\mathbf{v}}_i)$. We can do this 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 voxel super-resolution. After getting the occupancy value of the points, the loss function for training the model of

voxel super-resolution can be defined as:

$$\begin{aligned} L_{VSR} &= \sum_{i=1}^{N} \|f_V(g_V(\rho(\hat{\mathbf{v}}_i))) - o(\hat{\mathbf{v}}_i)\| \\ &= \sum_{i=1}^{N} \|f_V(\mathbf{F}_V(\rho(\hat{\mathbf{v}}_i))) - o(\hat{\mathbf{v}}_i)\|. \end{aligned} \tag{6.10}$$

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 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 Figure 6.1, 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 $256 \times 128 \times 128$, $256 \times 64 \times 64$, $256 \times 32 \times 32$, and $256 \times 16 \times 16$. 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 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 convolution neural network is the low resolution voxel grids which have the size 128^3. We use 5 stage 3D 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 neural 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*. We use standard cross-entropy loss 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 using 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. We firstly obtain a coarse 3D reconstruction from multi-view image through learning the implicit function. Then, based on the coarse 3D prediction, 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 our experiments: Articulated dataset [179] and CAPE dataset [116]. 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 actions 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. Figure 6.6 gives an example of four-view images and 3D mesh from the small CAPE dataset. We also split the dataset as 80% for training and 20% for testing in our experiments.

Metrics. In order to evaluate our method quantitatively, we choose three metrices 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 [26].

4.2 The results of the two steps

In order to demonstrate the performance of MF-PIFu and VSR, we evaluate the results of the two parts on the two datasets. Figure 6.6 gives the examples of the CAPE and Articulated dataset, respectively. The first row is an example from CAPE and the second row is an example from Articulated. The figure from left to right column 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 Figure 6.6 (c). For instance, the arms of the 3D

(a) Multi-view images (b) GT (c) MF-PIFu (d) VSR

Figure 6.6: The 3D models from multi-view images and the 3D model after voxel super-resolution. 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 model after VSR (d).

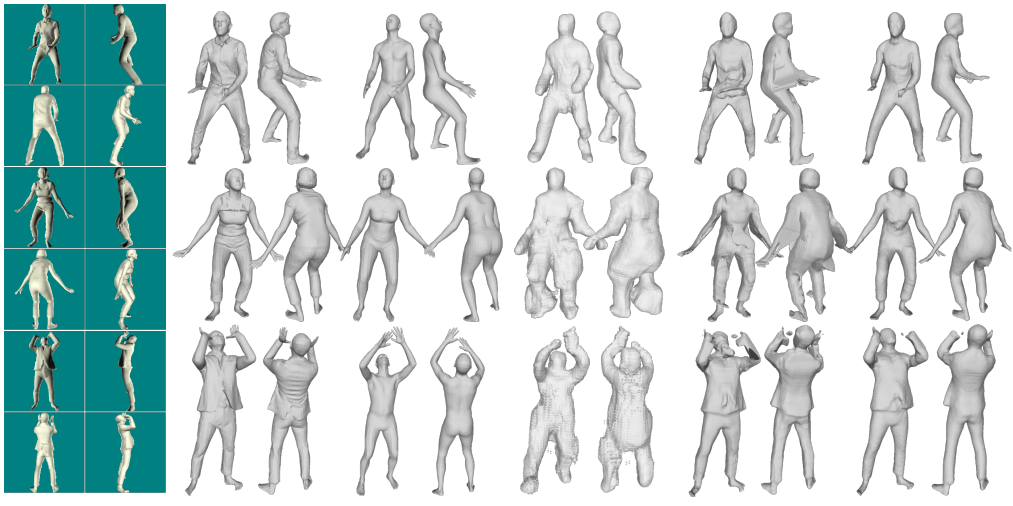
model from the CAPE dataset are not fully reconstructed by MF-PIFu and there are some extra reconstructed parts around the legs of the 3D models from the Articulated dataset. From Figure 6.6 (d), it shows that the results of VSR are refined. Those extra reconstruction in the estimated 3D models of MF-PIFu are removed and the details of the appearance are preserved, especially for the arms of the 3D model for the CAPE example and the neck part of the 3D model for the Articulated example. Therefore, the refined models look more smooth and natural. This figure demonstrates that MF-PIFu can produce the coasrse 3D models from multi-view images and VSR can generate better results through refining the coarse 3D models.

Table 6.1: 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

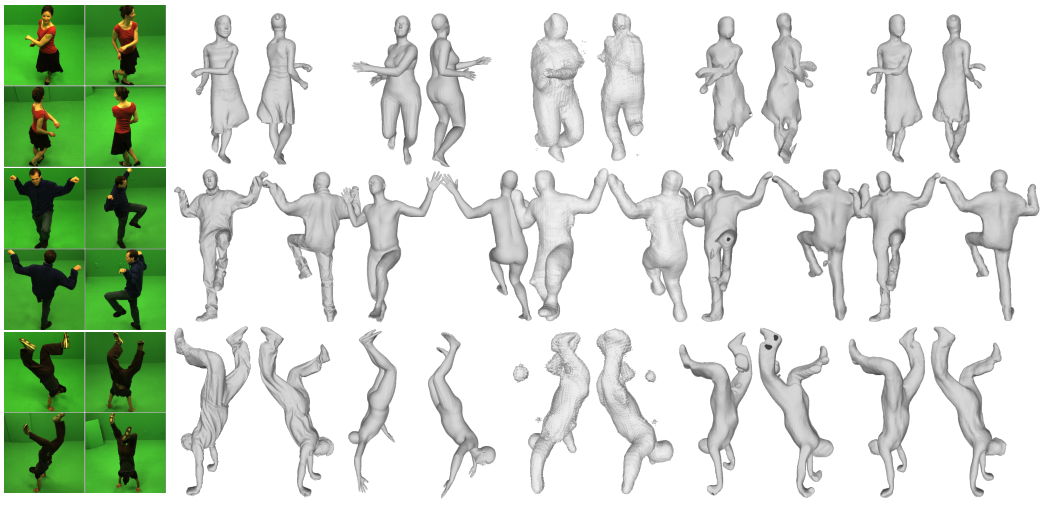
The quantitative results of the two steps on the two datasets are also shown in Table 6.1. The results of P2S, Chamfer-L_2 and IoU of the coarse 3D models by MF-PIFu and the

refined 3D models of VSR are given in this table. We can see from the table 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 can obtain better 3D models. The conclusion of this table is consistent with Figure 6.6.



(a) Image (b) GT (c) SPIN [88] (d) DeepHuman [203] (e) PIFu [156] (f) Our

Figure 6.7: The comparison between our method and several previous methods on the CAPE dataset. Three examples are shown from top to down rows. The multi-view images, the ground truth of 3D models from two views, the estimated 3D models of SPIN [88], DeepHuman [203], PIFu [156] and our method are shown from the left to row column.



(a) Image (b) GT (c) SPIN [88] (d) DeepHuman [203] (e) PIFu [156] (f) Our

Figure 6.8: The comparison between our method and several previous methods on the Articulated dataset. Three examples are shown from top to down rows. The multi-view images, the ground truth of 3D models from two views, the estimated 3D models of SPIN [88], DeepHuman [203], PIFu [156] and our method are shown from the left to row column.

4.3 Qualitative results

We qualitatively compare our method with several previous approaches for 3D human body reconstruction from images including SPIN [88], DeepHuman [203] and PIFu [156]. For the SPIN and DeepHuman, we used the trained model provided by the authors to obtain the results. The two methods rely on the SMPL model [113] to reconstruct 3D human body from single images. For the PIFu, we trained and tested it on the same training dataset of Articulated and CPAE as our method from four-view images. The SPIN estimated the pose and shape parameters of SMPL model through collaborating regression and optimization. The estimated 3D models of SPIN are naked because the results of SPIN are the SMPL models parameterized by the estimated pose and shape parameters. The DeepHuman used encoder-decoder on the volume of deformed SMPL model and used normal image to refine the deformed SMPL model. This method can produce detailed SMPL model because the normal image could refine the appearance of SMPL model. In Figure 6.7 and Figure 6.8, some examples from the two datasets and the results of SPIN [88], DeepHuman [203], PIFu [156] and our method are demonstrated, respectively. For each dataset, we give three examples which cover various poses and clothes to compare the performance of the methods. We can see that the estimated 3D models of SPIN and DeepHuman are not good enough but the results of PIFu and our method are better. Since the 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 in the original paper is based on a different dataset. The results of PIFu are better than SPIN and DeepHuman because of learning an implicit representation, but there are some false parts in the results since the features in PIFu are at the same scales. By contrast, our method uses a coarse-to-fine manner to better reconstruct 3D human body models. The MF-PIFu estimates the coarse 3D models based on multi-scale features and implicit representation, and the VSR refines the coarse models to generate final results also based on multi-scale features and implicit representation. Our method can recover the 3D human body models from multi-view images with plausible pose and surface quality.

In Figure 6.9, we visualize the P2S between the reconstructed 3D models in Fiugre 6.7 and Figure 6.8 by the different methods and the ground truth. We use Meshlab to visualize the P2S to show the accuracy of the estimated 3D models by different methods. In Meshlab, the P2S is computed through the Hausdorff Distance. The distances are shown by the heatmaps and are mapped to the reconstructed 3D models. For every sample, the color range of different methods is based on the value of the P2S of our method. 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 three previous methods.

(a) SPIN [88] (b) DeepHuman [203] (c) PIFu [156] (d) Our

Figure 6.9: Visualization of the P2S between the estimated 3D models in Figure 6.7 and Figure 6.8 and the ground truth for different methods. The distance are represented by the heatmaps in Meshlab and mapped to the estimated 3D models.

4.4 Quantitative results

In addition to the qualitative comparison, we also quantitatively compare to previous methods through computing the P2S, Chamfer-L_2 and IoU of results by different methods on the testing datasets of CAPE and Articulated. Table 6.2 and Table 6.3 demonstrate the mean values of the above metrics of different methods on the testing dataset of CAPE and Articulated, respectively. For the CAPE, the results of DeepHuman [203] are the worst because the CAPE is a synthetic dataset, but the trained model of DeepHuman is based on a real dataset. The SPIN [88] is better than DeepHuman, but it is still worse than PIFu [156] and our method because the estimated 3D models of SPIN are naked and the poses of the estimated 3D models might not be accurate. Comparing to SPIN and DeepHuman, the results of PIFu are better because PIFu uses four-view images and represents the 3D model through learning implicit function. Our method achieves the best performance among these methods because VSR can refine the coarse results of MF-PIFu. Both MF-PIFu and VSR in our method extract multi-scale features and learn the implicit function from multi-view images. The coarse-to-fine manner is an efficient way to obtain better models. 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 our method is the highest, which means that the estimated 3D models are more complete. For the Articulated dataset, Table 6.3 shows similar conclusion. The SPIN and DeepHuman achieve similar level on the real dataset and PIFu is better than the above two methods. However, our method also achieves the smallest P2S and Chamfer-L_2 and the highest IoU on the Articulated dataset. The two tables demonstrate that our method had good performance on both synthetic and real datasets.

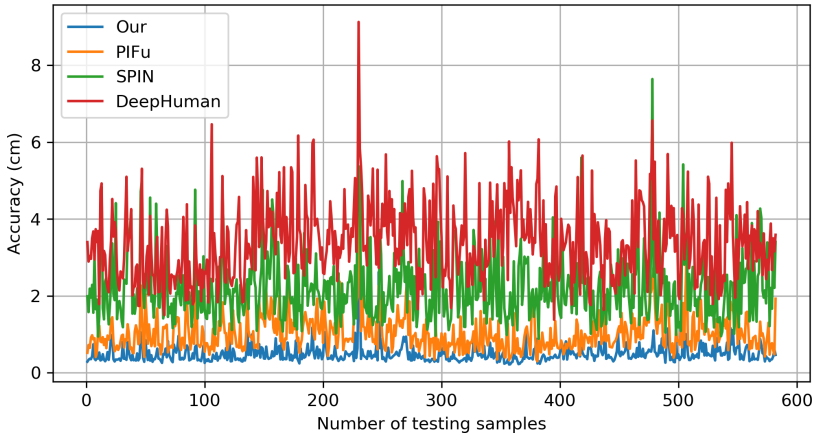
Table 6.2: The quantitative results of SPIN [88], DeepHuman [203], PIFu [156] and our method on the testing dataset of the CAPE. Our method achieves better performance.

Method	P2S ↓	Chamfer-L_2 ↓	IoU ↑
SPIN [88]	2.2134	0.1271	0.4044
DeepHuman [203]	3.4028	0.1850	0.3861
PIFu [156]	1.0330	0.0212	0.7571
Ours	**0.4954**	**0.0062**	**0.8440**

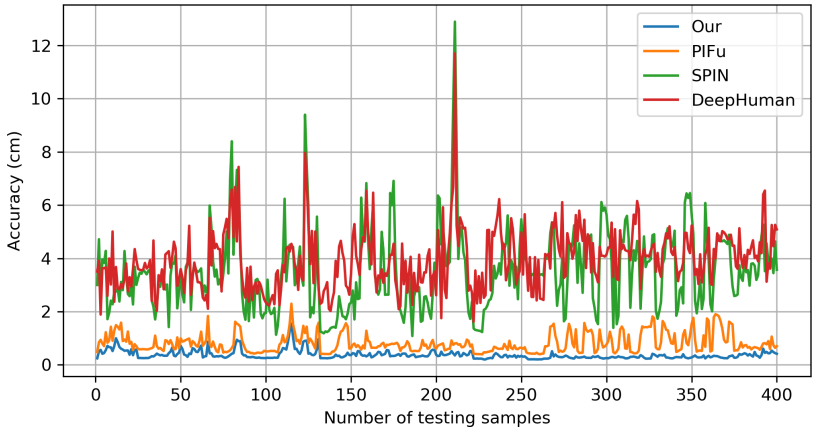
Table 6.3: The quantitative results of SPIN [88], DeepHuman [203], PIFu [156] and our method on the testing dataset of the Articulated. Our method achieves better performance.

Method	P2S ↓	Chamfer-L_2 ↓	IoU ↑
SPIN [88]	3.5206	0.2679	0.3506
DeepHuman [203]	3.9448	0.2675	0.3742
PIFu [156]	0.8194	0.0210	0.8255
Ours	**0.3754**	**0.0032**	**0.9051**

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 Figure 6.10. There are 582 samples in the testing dataset of CAPE and 400 samples in the testing dataset of Articulated, respectively. Our method (the blue 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 blue lines do not have serious fluctuation.



(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.

Figure 6.10: 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.

4.5 Discussion on the PIFu

As shown above, PIFu [156] 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, PIFu+VSR and our method are evaluated to demonstrate the advantage of MF-PIFu and our method on the Articulated dataset. Table 6.4 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 was 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 coasrse 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. Figure 6.11 gives the the P2S of the four cases on the testing dataset of the Articulated. We can see from the figure 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 6.4.

Table 6.4: 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

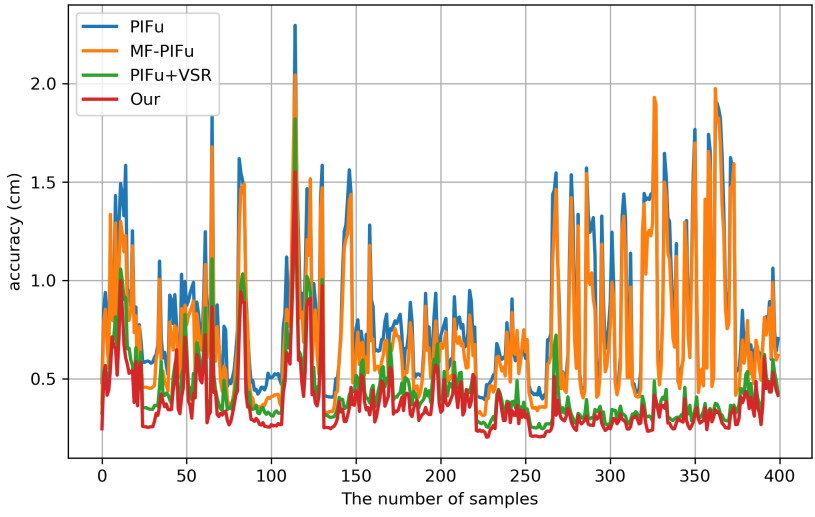
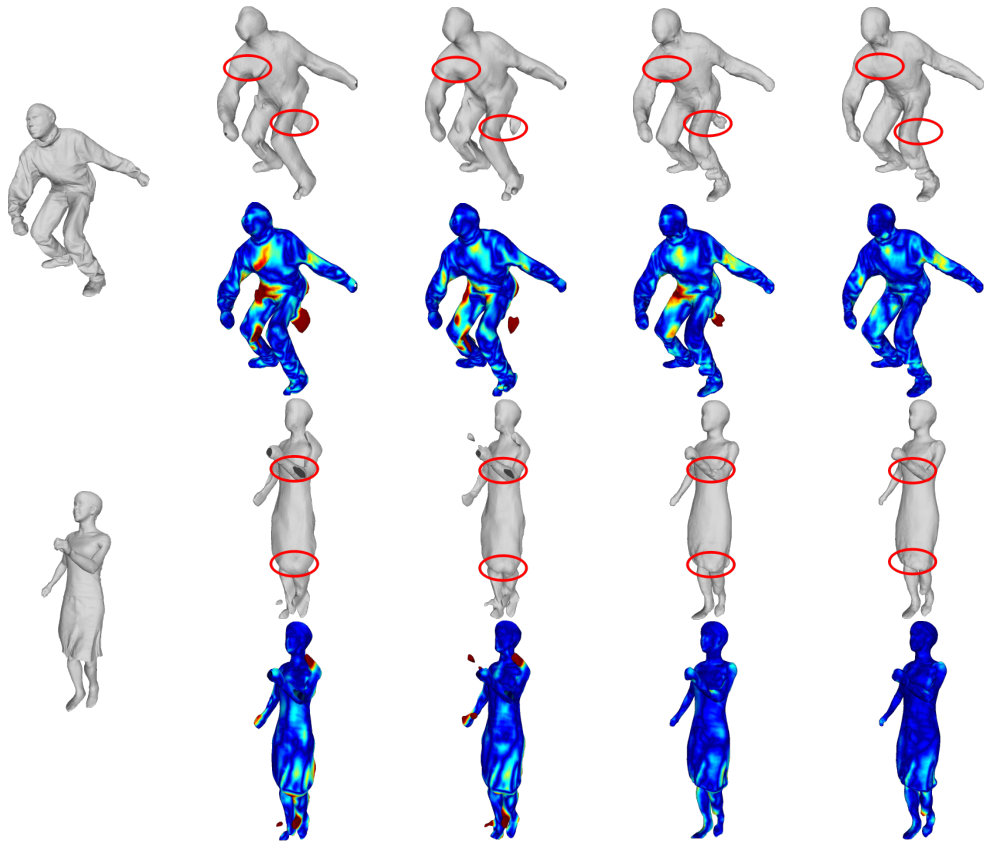


Figure 6.11: The P2S of each sample in the testing data of the Articulated for PIFu, MF-PIFu, PIFu+VSR, and our method. The y axis stands for the accuracy of P2S. The x axis is the number of samples in the testing data.

The qualitative examples from the Articulated dataset are shown in Figure 6.12. 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.



(a) GT (b) PIFu (c) MF-PIFu (d) PIFu+VSR (e) Our

Figure 6.12: The qualitative results of PIFu, MF-PIFu, PIFu+VSR, and our method on the Articulated dataset.

4.6 Spatial sampling

Spatial sampling is used in both MF-PIFu and VSR to generate the ground truth of the implicit 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 [156] 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 voxel super-resolution 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 $\sigma_{\min}$, we choose five pairs of $(\sigma_{\max}, \sigma_{\min})$ and compare the corresponding performance under the five cases. Table 6.5 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. Figure 6.13 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 $(\sigma_{\max}, \sigma_{\min}) = (15, 1.5), (25, 2.5), (35, 3.5)$. 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 we use $(\sigma_{\max}, \sigma_{\min}) = (15, 1.5)$ in the quantitative and qualitative comparison to the previous methods.

Table 6.5: 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

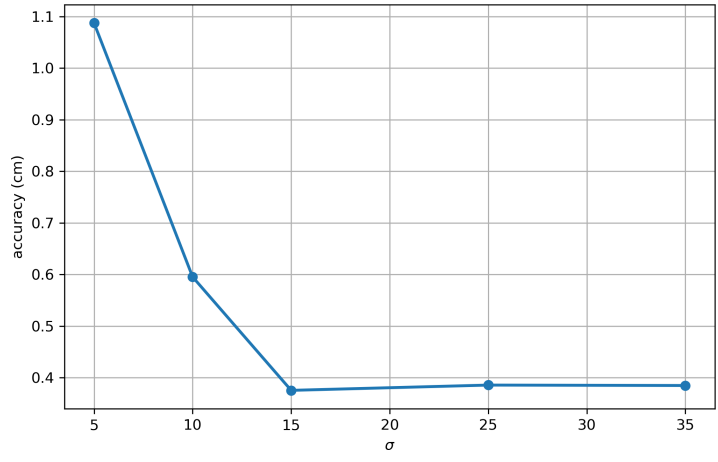
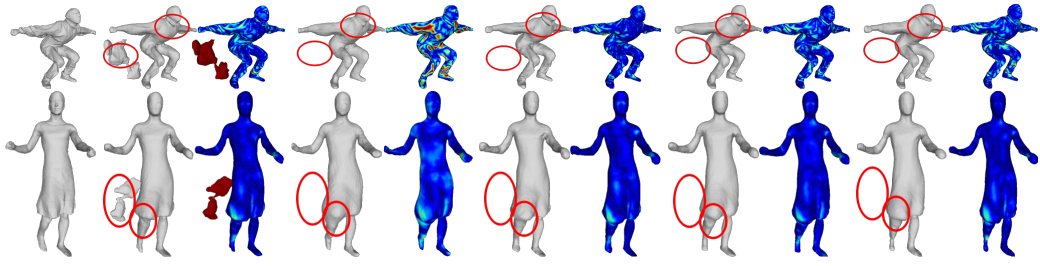


Figure 6.13: The mean P2S on the testing dataset of the Articulated for different $\sigma_{\max}$. The *y* axis stands for the mean P2S. The *x* axis is the $\sigma_{\max}$.

Figure 6.14 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, we can see 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 $(15, 1.5) \sim (35, 3.5)$ according to the experiments.



(a) GT (b) $\sigma_{\max} = 5$ (c) $\sigma_{\max} = 10$ (d) $\sigma_{\max} = 15$ (e) $\sigma_{\max} = 25$ (f) $\sigma_{\max} = 35$

Figure 6.14: The comparison for differen $\sigma_{\max}$ 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 estimated result and the ground truth is given.

4.7 Voxel grid resolution

The resolution of input voxel grids for VSR will also affects the refinement of VSR to generate 3D models. In order to demonstrate the effects, we compare the results of VSR with the input resolution of 32^3 and 128^3 for the Articulated dataset. 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 6.6 shows the P2S, Chamfer-L_2 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. We can see that the quantitative values of the 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 6.15 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 have 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.

Table 6.6: 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**

Figure 6.15: 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.

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. We evaluate the performance of our method for four images and eight images on the Articulated dataset. 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 6.7 shows the quantitative results on the Articulated dataset when the four-view and eight-view images are used. Figure 6.16 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. 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 voxel super-resolution, 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.

Table 6.7: 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

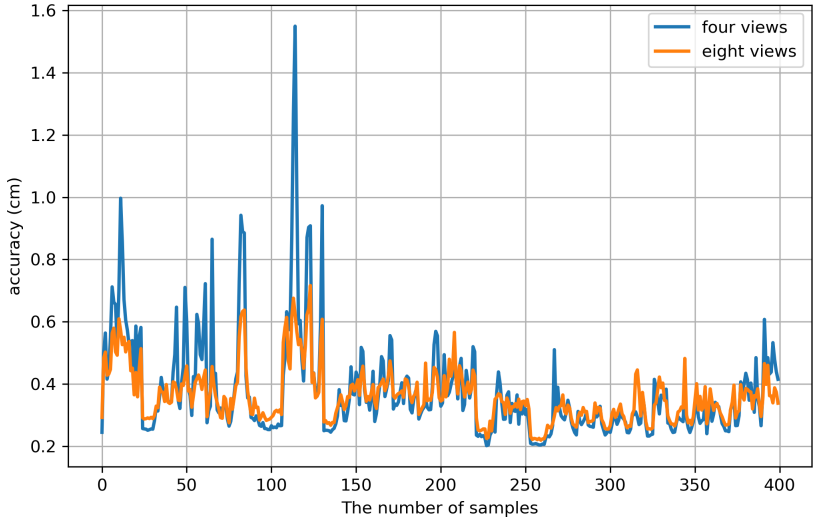
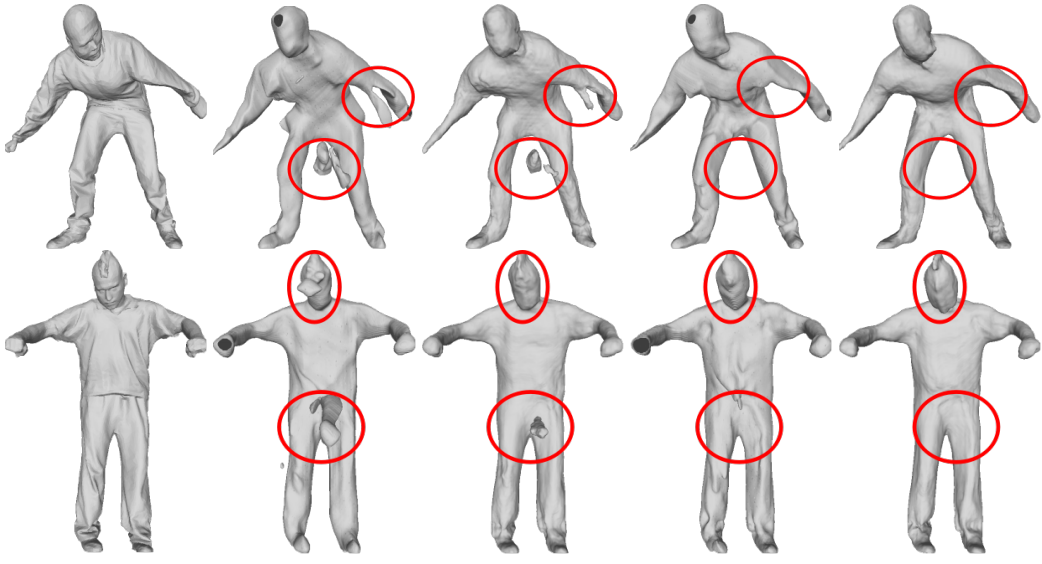


Figure 6.16: The P2S of each sample in the testing data of the Articulated for four-view and eight-view images. The y axis stands for the accuracy of P2S. The x axis is the number of samples in the testing data.

Two examples from the Articulated dataset are shown in Figure 6.17 for the four-view and eight-view images. The figure 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 voxel super-resolution, the coarse 3D models are refined to more accurate models, but the errors are not removed completely for the four-view. 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.



(a) GT (b) 4-view 1 (c) 4-view 2 (d) 8-view 1 (e) 8-view 2

Figure 6.17: The results of four-view and eight-view images on the Articulated dataset. From left ot 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.

5 Conclusion

Detailed 3D human body reconstruction from RGB images is a challenging task because of the high number of degrees of the 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 im-

plicit representation. The coarse 3D models are estimated from multi-view images through learning implicit representations 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, we use voxel super-resolution to refine the coarse 3D models. For the voxel super-resolution, 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 voxel super-resolution. 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, we need to increase the variety of the training dataset. The models in the two datasets of our paper mostly have the same color clothes. If there is a new model with colorful 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.

Conclusion

Chapter 7

Conclusions and Outlook

In the thesis, we present several methods for 3D human body reconstruction from images including the estimation of coarse human body models, especially for the human poses and shapes, and the estimation of detailed human body models, especially for the detailed appearance like clothes. This chapter revisits and summarizes the key idea and contribution of the proposed methods in the thesis. In addition to the conclusion, we also give an outlook for future research direction and possible enhancements.

1 Conclusion

This thesis aims at the problem of 3D human pose and shape estimation from images. Considering the 3D human body model is needed in many applications, five methods are proposed to tackle the problem in different aspects from Chapter 2 to Chapter 6. In the five methods, some of them use parametric models based on optimization, and some of them advocate deep neural network to learn to represent 3D models from images. We obtain both the coarse and detailed 3D human body models from the five methods. Those methods obtaining coarse 3D models mainly focus on the estimation of various poses and shapes, while those methods obtaining detailed 3D models pay more attention on the appearance details.

In Chapter 2, a method for 3D human pose and shape estimation is proposed based on a parametric model from multi-view 2D images. The key idea is to fit a parametric human body called SMPL to the 2D joint points of the multi-view images simultaneously. The SMPL model can be parametrized by its pose and shape parameters, which is convenient to estimate 3D model through optimizing pose and shape parameters. Another important problem is to extract the accurate 2D joint points from image. Benefiting from the advance

of deep learning, we can use OpenPose to estimate the 2D joint points of the multi-view images automatically. During the fitting, the pose, shape as well as the camera rotation are optimized in order to reflect the relations between the multi-view images. We evaluate the method on the subsets of two subjects (S1 and S6) in Human3.6M and the experiments demonstrate that our method outperforms the single-view image based method, SMPLify, on the estimation of 3D human pose.

In Chapter 3, we introduce silhouettes to improve the estimation of the shape of the SMPL model. Firstly, the SMPL model was fitted to the joint points of the multi-view images, which makes the SMPL model have the same pose with the observed human body. Since joint points mainly provide pose information, the shape parameters of the SMPL are not estimated well. Then, in the second part, silhouettes are used to enhance the results. We build the correspondence between the boundary of the silhouettes and the SMPL model after pose fitting over the multi-view images. The energy function for silhouettes fitting is defined on both 2D and 3D space, which gives more constrains on the shape estimation. The synthetic datasets and public real datasets are advocated to evaluate our method and the results demonstrate that our method achieves the better accuracy on the shape of human body than previous methods.

In Chapter 4, we address the 3D human pose and shape estimation through putting optimization based on multi-view images into the training loop of deep neural network. Since some popular public datasets for human pose estimation are captured from multi-view images, we could use this advantage during the training. The multi-view images are fed into the CNN model to regress the pose and shape parameters of SMPL. Then, an energy function is built by the 2D joint points of the multi-view images and the SMPL model. In order to minimize the function, the pose, shape of SMPL model as well as the body orientations of multi-view images are optimized by using the regressed parameters as initialization. Afterwards, The optimized SMPL model can be used to supervise the training of CNN. The experiments on several public datasets demonstrate the 3D pose and shape estimation by our method outperforms the previous methods.

In Chapter 5, we shift to the detailed 3D human body reconstruction from images through learning implicit representation for 3D models. The contribution in this chapter is that multi-stage hourglass networks are used to extract multi-scale features from multi-view images, which encodes both local and global information of images. Then, a fully connected neural network predicts the occupancy value of the multi-scale features to decide the corresponding 3D point lies inside or outside of the 3D mesh. The feature maps extracted by hourglass networks are spatially aligned with the multi-view images, so we can predict the occupancy value of enough number of spatial points. Finally, the 3D models can be represented implicitly and extracted by marching cubes algorithm. Unlike the previous chapters, the method in this chapter does not depend on any parametric models, and thus, the method can estimate 3D human body models with detailed appearance. Results on the

public synthetic and real datasets show that the estimated 3D model by our method has higher accuracy and better details on the appearance.

In Chapter 6, we also focus on the detailed 3D human body reconstruction from multi-view images. Instead of only relying on the images, a refined process based on voxel super-resolution is introduced to better estimate 3D models. In general, the method in this chapter is a coarse-to-fine manner to estimate detailed 3D human body model from multi-view images. The coarse 3D models are firstly estimated from multi-view images through learning implicit 3D representation based on multi-scale features. The refined process is implemented by learning the voxel super-resolution from low-resolution voxel grids generated by the coarse 3D models. The low-resolution voxel grids are fed into multi-stage 3D convolutional layers to extract multi-scale features and the fully connected neural network predicts the occupancy values of the multi-scale features to obtain the implicit representation. The final 3D models are extracted by marching cubes algorithm from the implicit volume. The voxel super-resolution can both remove the false parts of coarse models and preserve the details of appearance. The experiments on synthetic and real dataset demonstrate that our method can recovery 3D human body with higher accuracy and better surface details than previous methods. The method in this chapter is also better than the method in Chapter 5.

2 Future work

Although the works in the thesis estimate the both coarse and detailed 3D human body models from images, there are still many problems of 3D human body reconstruction which are not involved in our works, for example, texture, facial expression, hand gesture and multiple persons. Therefore, this leaves much room for future work, which is summarized below.

Since our methods in Chapter 2, Chapter 3, and Chapter 4 rely on a parametric human body model, SMPL, it is an important factor to obtain reasonable and natural 3D human body models. The SMPL model mainly focuses on the pose and shape representation, but the other human body elements like hands and facial expression are not considered in this model. Recently, some researches started to recover the full human body including hand gesture, feet, and facial expression. However, these works are still far from obtaining accurate results. The topic to parameterize hands, feet and facial expression in the parametric human body model is an interesting problem. In addition, fitting the parametric model to prior information containing face and hand is another direction for the research. Therefore, considering the requirement of realism, adding face and hand on the parametric model is an important direction.

Another direction is to infer the texture on the 3D human body. As shown in results of all the chapters, we do not add the texture on the 3D human body, which affects on the realism of the final 3D model. Recently, some methods learn to infer the color of voxels from images through deep neural networks. Some other methods also infer the UV space and add them on the parametric human body. However, we still face the challenges to infer the texture of voxels accurately, especially for those invisible parts of human body. The first problem is that the high-resolution 3D human body dataset is hard to acquire. For some recent works, the datasets they used are not free and public, which is not easy and convenient for research. Therefore, predicting the color of voxels by deep learning is difficult due to the limitation of dataset. Another problem is that only front side can be seen in the 2D image. Inferring the texture of the invisible part on the 3D human body model is a challenging problem. Therefore, the future work on inferring the texture of 3D human body is an interesting topic.

Computation efficient is also an issue for 3D human body reconstruction from images no matter for optimization and deep learning. The optimization based methods often fit the parametric model to prior information, which is computation-consuming and always falls into local minimization. The methods based on deep learning often require enormous dataset for training. Even for the inference, due to that the network is often very deep, it is also take some time to obtain the final results. For some real application, for example, motion capture of human body, reducing the computation and the complexity of deep neural network is also important.

Recovering 3D human body model with more challenging conditions are also attracting much attention. In real life, there might be multiple people in one scenario, various illumination in the wild, and self-occlusion of human body or occlusion by other objects. The high freedom of environment and human motion in real life leaves many research topics on 3D human body reconstruction from images. These problems are far from achieving satisfying results now.

In this chapter, the conclusion of the thesis is presented and some possible directions are summarized. Since 3D human body estimation is a very practical task and has high demand in many applications, it is very important and has significance to solve the problem. This thesis only contributes several methods for the topic and many work still need to be done in the future.

References

[1] P. Achlioptas, O. Diamanti, I. Mitliagkas, and L. Guibas. Learning representations and generative models for 3D point clouds. In *International Conference on Machine Learning*, pages 40–49, 2018.

[2] T. Alldieck, M. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll. Detailed human avatars from monocular video. In *2018 International Conference on 3D Vision (3DV)*, pages 98–109. IEEE, 2018.

[3] T. Alldieck, M. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll. Video based reconstruction of 3D people models. In *CVPR*, pages 8387–8397. IEEE, 2018.

[4] T. Alldieck, M. Magnor, B. L. Bhatnagar, C. Theobalt, and G. Pons-Moll. Learning to reconstruct people in clothing from a single RGB camera. In *CVPR*, pages 1175–1186. IEEE, 2019.

[5] T. Alldieck, G. Pons-Moll, C. Theobalt, and M. Magnor. Tex2Shape: Detailed full human body geometry from a single image. In *ICCV*, pages 2293–2303. IEEE, 2019.

[6] B. Allen, B. Curless, and Z. Popović. Articulated body deformation from range scan data. *ACM Trans. Graph.*, 21(3):612–619, July 2002.

[7] B. Allen, B. Curless, Z. Popović, and A. Hertzmann. Learning a correlated model of identity and pose-dependent body shape variation for real-time synthesis. In *Proceedings of the 2006 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pages 147–156, 2006.

[8] R. Alp Güler, N. Neverova, and I. Kokkinos. DensePose: Dense human pose estimation in the wild. In *CVPR*, pages 7297–7306. IEEE, 2018.

[9] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2D human pose estimation: New benchmark and state of the art analysis. In *CVPR*, pages 3686–3693. IEEE, June 2014.

[10] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis. SCAPE: shape completion and animation of people. *ACM Trans. Graph.*, 24(3):408–416, July 2005. ISSN 0730-0301.

[11] A. Arnab, C. Doersch, and A. Zisserman. Exploiting temporal context for 3D human pose estimation in the wild. In *CVPR*, pages 3395–3404. IEEE, 2019.

[12] A. O. Bălan and M. J. Black. The naked truth: Estimating body shape under clothing. In *Proceedings of the European Conference on Computer Vision*, pages 15–29. Springer, 2008.

[13] A. O. Balan, L. Sigal, M. J. Black, J. E. Davis, and H. W. Haussecker. Detailed human shape and pose from images. In *CVPR*, pages 1–8. IEEE, 2007.

[14] A. H. Barr. Global and local deformations of solid primitives. *SIGGRAPH Comput. Graph.*, 18(3):21–30, 1984.

[15] B. L. Bhatnagar, G. Tiwari, C. Theobalt, and G. Pons-Moll. Multi-Garment Net: Learning to dress 3D people from images. In *ICCV*, pages 5420–5430. IEEE, 2019.

[16] R. Bodor, B. Jackson, and N. Papanikolopoulos. Vision-based human tracking and activity recognition. In *Proc. of the 11th Mediterranean Conf. on Control and Automation*, volume 1, 2003.

[17] F. Bogo, M. J. Black, M. Loper, and J. Romero. Detailed full-body reconstructions of moving people from monocular RGB-D sequences. In *ICCV*, pages 2300–2308. IEEE, 2015.

[18] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 561–578. Springer, 2016.

[19] F. Bogo, J. Romero, G. Pons-Moll, and M. J. Black. Dynamic FAUST: Registering human bodies in motion. In *CVPR*, pages 6233–6242. IEEE, 2017.

[20] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2019.

[21] J. E. Chadwick, D. R. Haumann, and R. E. Parent. Layered construction for deformable animated characters. *SIGGRAPH Comput. Graph.*, 23(3):243–252, 1989.

[22] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and

fully connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2017.

[23] Y. Chen, Z. Liu, and Z. Zhang. Tensor-based human body modeling. In *CVPR*, pages 105–112. IEEE, 2013.

[24] Z. Chen and H. Zhang. Learning implicit fields for generative shape modeling. In *CVPR*, pages 5939–5948. IEEE, 2019.

[25] K.-L. Cheng, R.-F. Tong, M. Tang, J.-Y. Qian, and M. Sarkis. Parametric human body reconstruction based on sparse key points. *IEEE Transactions on Visualization and Computer Graphics*, 22(11):2467–2479, 2015.

[26] J. Chibane, T. Alldieck, and G. Pons-Moll. Implicit functions in feature space for 3D shape reconstruction and completion. In *CVPR*, pages 6970–6981. IEEE, 2020.

[27] C. B. Choy, D. Xu, J. Gwak, K. Chen, and S. Savarese. 3D-R2N2: A unified approach for single and multi-view 3D object reconstruction. In *Proceedings of the European conference on computer vision (ECCV)*, pages 628–644. Springer, 2016.

[28] Y. Cui, W. Chang, T. Nöll, and D. Stricker. KinectAvatar: fully automatic body capture using a single kinect. In *Asian Conference on Computer Vision*, pages 133–147. Springer, 2012.

[29] B. Curless and M. Levoy. A volumetric method for building complex models from range images. In *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*, pages 303–312, 1996.

[30] A. Dai, C. Ruizhongtai Qi, and M. Nießner. Shape completion using 3D-encoder-predictor cnns and shape synthesis. In *CVPR*, pages 5868–5877. IEEE, 2017.

[31] E. De Aguiar, C. Stoll, C. Theobalt, N. Ahmed, H.-P. Seidel, and S. Thrun. Performance capture from sparse multi-view video. In *ACM SIGGRAPH 2008 papers*, pages 1–10. 2008.

[32] Q. Delamarre and O. Faugeras. 3D articulated models and multi-view tracking with silhouettes. In *ICCV*, volume 2, pages 716–721. IEEE, 1999.

[33] E. Dibra, H. Jain, C. Öztireli, R. Ziegler, and M. Gross. HS-Nets: Estimating human body shape from silhouettes with convolutional neural networks. In *2016 International Conference on 3D Vision (3DV)*, pages 108–117. IEEE, 2016.

[34] E. Dibra, C. Öztireli, R. Ziegler, and M. Gross. Shape from selfies: Human body shape estimation using CCA regression forests. In *Proceedings of the European conference on computer vision (ECCV)*, pages 88–104. Springer, 2016.

[35] E. Dibra, H. Jain, C. Oztireli, R. Ziegler, and M. Gross. Human shape from silhouettes using generative hks descriptors and cross-modal neural networks. In *CVPR*, pages 4826–4836. IEEE, 2017.

[36] M. Dou, S. Khamis, Y. Degtyarev, P. Davidson, S. R. Fanello, A. Kowdle, S. O. Escolano, C. Rhemann, D. Kim, J. Taylor, et al. Fusion4d: Real-time performance capture of challenging scenes. *ACM Trans. Graph.*, 35(4):1–13, 2016.

[37] Facebook. Pytorch. `https://pytorch.org/`.

[38] H. Fan, H. Su, and L. J. Guibas. A point set generation network for 3d object reconstruction from a single image. In *CVPR*, pages 605–613. IEEE, 2017.

[39] J.-S. Franco and E. Boyer. Exact polyhedral visual hulls. In *British Machine Vision Conference*, volume 1, pages 329–338, 2003.

[40] O. Freifeld and M. J. Black. Lie bodies: A manifold representation of 3D human shape. In *Proceedings of European Conference on Computer Vision (ECCV)*, pages 1–14. Springer, 2012.

[41] V. Gabeur, J.-S. Franco, X. Martin, C. Schmid, and G. Rogez. Moulding humans: Non-parametric 3D human shape estimation from single images. In *ICCV*, pages 2232–2241. IEEE, 2019.

[42] J. Gall, C. Stoll, E. De Aguiar, C. Theobalt, B. Rosenhahn, and H.-P. Seidel. Motion capture using joint skeleton tracking and surface estimation. In *CVPR*, pages 1746–1753. IEEE, 2009.

[43] D. M. Gavrila and L. S. Davis. 3D model-based tracking of humans in action: a multi-view approach. In *CVPR*, pages 73–80. IEEE, 1996.

[44] S. Geman. Statistical methods for tomographic image reconstruction. *Bull. Int. Stat. Inst*, 4:5–21, 1987.

[45] A. Geurtz. Model based shape estimation. *Signal Process.*, 37(2):303, 1994.

[46] A. Gilbert, M. Volino, J. Collomosse, and A. Hilton. Volumetric performance capture from minimal camera viewpoints. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 566–581, 2018.

[47] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, pages 580–587. IEEE, 2014.

[48] G. Gkioxari, J. Malik, and J. Johnson. Mesh R-CNN. In *ICCV*, pages 9785–9795. IEEE, 2019.

[49] L. Goncalves, E. Di Bernardo, E. Ursella, and P. Perona. Monocular tracking of the human arm in 3D. In *ICCV*, pages 764–770. IEEE, 1995.

[50] Google. Tensorflow. `https://www.tensorflow.org/`.

[51] D. Grest, D. Herzog, and R. Koch. Human model fitting from monocular posture images. In *Proceedings of the Conference on Vision, Modeling and Visualization*, pages 665–1344. ACM, 2005.

[52] X. Gu, Z. Fan, S. Zhu, Z. Dai, F. Tan, and P. Tan. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *CVPR*, pages 2495–2504. IEEE, 2020.

[53] P. Guan, A. Weiss, A. O. Balan, and M. J. Black. Estimating human shape and pose from a single image. In *ICCV*, pages 1381–1388. IEEE, 2009.

[54] E. Gundogdu, V. Constantin, A. Seifoddini, M. Dang, M. Salzmann, and P. Fua. GarNet: A two-stream network for fast and accurate 3D cloth draping. In *ICCV*, pages 8739–8748. IEEE, 2019.

[55] K. Guo, F. Xu, Y. Wang, Y. Liu, and Q. Dai. Robust non-rigid motion tracking and surface reconstruction using L0 regularization. In *ICCV*, pages 3083–3091. IEEE, 2015.

[56] Y. Guo, G. Xu, and S. Tsuji. Understanding human motion patterns. In *IAPR International Conference on Pattern Recognition (ICPR)*, volume 2, pages 325–329, 1994.

[57] Y. Guo, H. Wang, Q. Hu, H. Liu, L. Liu, and M. Bennamoun. Deep learning for 3D point clouds: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[58] M. Habermann, W. Xu, M. Zollhoefer, G. Pons-Moll, and C. Theobalt. LiveCap: Real-time human performance capture from monocular video. *ACM Trans. Graph.*, 38(2):1–17, 2019.

[59] E. P. Hanavan Jr. A mathematical model of the human body. *AMRL-TR. Aerospace Medical Research Laboratories*, 1964.

[60] C. Häne, S. Tulsiani, and J. Malik. Hierarchical surface prediction for 3d object reconstruction. In *2017 International Conference on 3D Vision (3DV)*, pages 412–420. IEEE, 2017.

[61] N. Hasler, C. Stoll, M. Sunkel, B. Rosenhahn, and H.-P. Seidel. A statistical model of human pose and body shape. In *Computer graphics forum*, volume 28, pages 337–346. Wiley Online Library, 2009.

[62] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778. IEEE, 2016.

[63] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask R-CNN. In *ICCV*, pages 2961–2969. IEEE, 2017.

[64] A. Hilton, D. Beresford, T. Gentils, R. Smith, W. Sun, and J. Illingworth. Whole-body modelling of people from multiview images to populate virtual worlds. *The Visual Computer*, 16(7):411–436, 2000.

[65] D. A. Hirshberg, M. Loper, E. Rachlin, and M. J. Black. Coregistration: Simultaneous alignment and modeling of articulated 3D shape. In *Proceedings of European conference on computer vision (ECCV)*, pages 242–255. Springer, 2012.

[66] D. Hogg. Model-based vision: a program to see a walking person. *Image and Vision Computing*, 1(1):5–20, 1983. ISSN 0262-8856.

[67] C. Hu, Q. Yu, Y. Li, and S. Ma. Extraction of parametric human model for posture recognition using genetic algorithm. In *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition*, pages 518–523. IEEE, 2000.

[68] Y. Huang, F. Bogo, C. Lassner, A. Kanazawa, P. V. Gehler, J. Romero, I. Akhter, and M. J. Black. Towards accurate marker-less human shape and pose estimation over time. In *2017 International Conference on 3D Vision (3DV)*, pages 421–430. IEEE, 2017.

[69] Z. Huang, T. Li, W. Chen, Y. Zhao, J. Xing, C. LeGendre, L. Luo, C. Ma, and H. Li. Deep volumetric video from very sparse multi-view performance capture. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 336–354, 2018.

[70] Z. Huang, Y. Xu, C. Lassner, H. Li, and T. Tung. ARCH: Animatable reconstruction of clothed humans. In *CVPR*, pages 3093–3102. IEEE, 2020.

[71] M. Innmann, M. Zollhöfer, M. Nießner, C. Theobalt, and M. Stamminger. VolumeDeform: Real-time volumetric non-rigid reconstruction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 362–379. Springer, 2016.

[72] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, jul 2014.

[73] S. Izadi, D. Kim, O. Hilliges, D. Molyneaux, R. Newcombe, P. Kohli, J. Shotton, S. Hodges, D. Freeman, A. Davison, et al. KinectFusion: real-time 3D reconstruction and interaction using a moving depth camera. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, pages 559–568, 2011.

[74] A. S. Jackson, A. Bulat, V. Argyriou, and G. Tzimiropoulos. Large pose 3d face reconstruction from a single image via direct volumetric cnn regression. In *ICCV*, pages 1031–1039. IEEE, 2017.

[75] A. S. Jackson, C. Manafas, and G. Tzimiropoulos. 3D human body reconstruction from a single image via volumetric regression. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018.

[76] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *BMVC*, volume 2, page 5. Citeseer, 2010.

[77] H. Joo, T. Simon, X. Li, H. Liu, L. Tan, L. Gui, S. Banerjee, T. Godisart, B. Nabbe, I. Matthews, et al. Panoptic Studio: A massively multiview system for social interaction capture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(1): 190–204, 2017.

[78] H. Joo, T. Simon, and Y. Sheikh. Total Capture: A 3D deformation model for tracking faces, hands, and bodies. In *CVPR*, pages 8320–8329. IEEE, 2018.

[79] I. A. Kakadiaris and D. Metaxas. 3d human body model acquisition from multiple views. In *ICCV*, pages 618–623. IEEE, 1995.

[80] I. A. Kakadiaris and D. Metaxas. Model-based estimation of 3D human motion with occlusion based on active multi-viewpoint selection. In *CVPR*, pages 81–87. IEEE, 1996.

[81] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik. End-to-end recovery of human shape and pose. In *CVPR*, pages 7122–7131. IEEE, 2018.

[82] A. Kanazawa, J. Y. Zhang, P. Felsen, and J. Malik. Learning 3D human dynamics from video. In *CVPR*, pages 5614–5623. IEEE, 2019.

[83] A. Kar, C. Häne, and J. Malik. Learning a multi-view stereo machine. In *Advances in neural information processing systems (NIPS)*, pages 365–376, 2017.

[84] L. Kavan and J. Žára. Spherical blend skinning: A real-time deformation of articulated models. In *Proceedings of the 2005 Symposium on Interactive 3D Graphics and Games*, pages 9–16, 2005.

[85] L. Kavan, S. Collins, J. Žára, and C. O' Sullivan. Geometric skinning with approximate dual quaternion blending. *ACM Trans. Graph.*, 27(4), 2008.

[86] S. Knoop, S. Vacek, and R. Dillmann. Sensor fusion for 3D human body tracking with an articulated 3D body model. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 1686–1691. IEEE, 2006.

[87] M. Kocabas, N. Athanasiou, and M. J. Black. VIBE: video inference for human body pose and shape estimation. In *CVPR*, pages 5253–5263. IEEE, 2020.

[88] N. Kolotouros, G. Pavlakos, M. J. Black, and K. Daniilidis. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In *ICCV*, pages 2252–2261. IEEE, 2019.

[89] N. Kolotouros, G. Pavlakos, and K. Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *CVPR*, pages 4501–4510. IEEE, 2019.

[90] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in neural information processing systems (NIPS)*, pages 1097–1105, 2012.

[91] Z. Lahner, D. Cremers, and T. Tung. DeepWrinkles: Accurate and realistic clothing modeling. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 667–684, 2018.

[92] C. Lassner, J. Romero, M. Kiefel, F. Bogo, M. J. Black, and P. V. Gehler. Unite the people: Closing the loop between 3D and 2D human representations. In *CVPR*, pages 6050–6059. IEEE, 2017.

[93] A. Laurentini. The visual hull concept for silhouette-based image understanding. *IEEE Transactions on pattern analysis and machine intelligence*, 16(2):150–162, 1994.

[94] H. Law and J. Deng. CornerNet: Detecting objects as paired keypoints. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 734–750, 2018.

[95] V. Lazova, E. Insafutdinov, and G. Pons-Moll. 360-degree textures of people in clothing from a single image. In *2019 International Conference on 3D Vision (3DV)*, pages 643–653. IEEE, 2019.

[96] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[97] V. Leroy, J.-S. Franco, and E. Boyer. Multi-view dynamic shape refinement using local temporal integration. In *ICCV*, pages 3094–3103. IEEE, 2017.

[98] J. P. Lewis, M. Cordner, and N. Fong. Pose space deformation: A unified approach to shape interpolation and skeleton-driven deformation. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*, pages 165–172, 2000.

[99] C. Li, Z. Zhao, and X. Guo. ArticulatedFusion: Real-time reconstruction of motion, geometry and segmentation using a single depth camera. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 317–332, 2018.

[100] H. Li, R. W. Sumner, and M. Pauly. Global correspondence optimization for non-rigid registration of depth scans. In *Computer graphics forum*, volume 27, pages 1421–1430. Wiley Online Library, 2008.

[101] H. Li, B. Adams, L. J. Guibas, and M. Pauly. Robust single-view geometry and motion reconstruction. *ACM Trans. Graph.*, 28(5):1–10, 2009.

[102] H. Li, E. Vouga, A. Gudym, L. Luo, J. T. Barron, and G. Gusev. 3d self-portraits. *ACM Trans. Graph.*, 32(6):1–9, 2013.

[103] Z. Li, A. Heyden, and M. Oskarsson. Parametric model-based 3d human shape and pose estimation from multiple views. In *Scandinavian Conference on Image Analysis*, pages 336–347. Springer, 2019.

[104] Z. Li, A. Heyden, and M. Oskarsson. Template based human pose and shape estimation from a single RGB-D image. In *8th International Conference on Pattern Recognition Applications and Methods, ICPRAM 2019*, pages 574–581. SciTePress, 2019.

[105] Z. Li, T. Yu, C. Pan, Z. Zheng, and Y. Liu. Robust 3D self-portraits in seconds. In *CVPR*, pages 1344–1353. IEEE, 2020.

[106] J. Liang and M. C. Lin. Shape-aware human pose and shape reconstruction using multi-view images. In *ICCV*, pages 4352–4362. IEEE, 2019.

[107] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 740–755. Springer, 2014.

[108] Z. Liu, J. Huang, S. Bu, J. Han, X. Tang, and X. Li. Template deformation-based 3D reconstruction of full human body scans from low-cost depth cameras. *IEEE Transactions on Cybernetics*, 47(3):695–708, 2016.

[109] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440. IEEE, 2015.

[110] H. C. Longuet-Higgins. A computer algorithm for reconstructing a scene from two projections. *Nature*, 293(5828):133–135, 1981.

[111] C. Loop, C. Zhang, and Z. Zhang. Real-time high-resolution sparse voxelization with application to image-based modeling. In *Proceedings of the 5th High-Performance Graphics Conference*, pages 73–79, 2013.

[112] M. Loper. Chumpy. `https://github.com/mattloper/chumpy`.

[113] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. SMPL: a skinned multi-person linear model. *ACM Trans. Graph.*, 34(6):1–16, 2015.

[114] M. M. Loper and M. J. Black. OpenDR: An approximate differentiable renderer. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 154–169. Springer, 2014.

[115] W. E. Lorensen and H. E. Cline. Marching cubes: A high resolution 3D surface construction algorithm. *ACM siggraph computer graphics*, 21(4):163–169, 1987.

[116] Q. Ma, J. Yang, A. Ranjan, S. Pujades, G. Pons-Moll, S. Tang, and M. J. Black. Learning to dress 3D people in generative clothing. In *CVPR*, pages 6469–6478. IEEE, 2020.

[117] J. Martinez, R. Hossain, J. Romero, and J. J. Little. A simple yet effective baseline for 3D human pose estimation. In *ICCV*, pages 2640–2649. IEEE, 2017.

[118] D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, and C. Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *2017 international conference on 3D vision (3DV)*, pages 506–516, 2017.

[119] D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei, H.-P. Seidel, W. Xu, D. Casas, and C. Theobalt. VNect: Real-time 3D human pose estimation with a single RGB camera. *ACM Trans. Graph.*, 36(4):1–14, 2017.

[120] L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger. Occupancy Networks: Learning 3D reconstruction in function space. In *CVPR*, pages 4460–4470. IEEE, 2019.

[121] D. Metaxas and D. Terzopoulos. Shape and nonrigid motion estimation through physics-based synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(6):580–591, 1993.

[122] Microsoft. Azure kinect dk. `https://azure.microsoft.com/en-us/services/kinect-dk/`.

[123] T. B. Moeslund, A. Hilton, V. Krüger, and L. Sigal. *Visual analysis of humans*. Springer, 2011.

[124] A. Mohr and M. Gleicher. Building efficient, accurate character skins from examples. *ACM Trans. Graph.*, 22(3):562–568, July 2003.

[125] G. Mori, X. Ren, A. A. Efros, and J. Malik. Recovering human body configurations: Combining segmentation and recognition. In *CVPR*, volume 2, pages II–II. IEEE, 2004.

[126] R. Natsume, S. Saito, Z. Huang, W. Chen, C. Ma, H. Li, and S. Morishima. SiCloPe: Silhouette-based clothed people. In *CVPR*, pages 4480–4490. IEEE, 2019.

[127] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohi, J. Shotton, S. Hodges, and A. Fitzgibbon. KinectFusion: Real-time dense surface mapping and tracking. In *2011 10th IEEE International Symposium on Mixed and Augmented Reality*, pages 127–136. IEEE, 2011.

[128] R. A. Newcombe, D. Fox, and S. M. Seitz. DynamicFusion: Reconstruction and tracking of non-rigid scenes in real-time. In *CVPR*, pages 343–352. IEEE, 2015.

[129] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 483–499. Springer, 2016.

[130] G. L. Oliveira, A. Valada, C. Bollen, W. Burgard, and T. Brox. Deep learning for human part discovery in images. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1634–1641. IEEE, 2016.

[131] M. Omran, C. Lassner, G. Pons-Moll, P. Gehler, and B. Schiele. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *2018 International Conference on 3D Vision (3DV)*, pages 484–494. IEEE, 2018.

[132] H. Onizuka, Z. Hayirci, D. Thomas, A. Sugimoto, H. Uchiyama, and R.-i. Taniguchi. TetraTSDF: 3D human reconstruction from a single image with a tetrahedral outer shell. In *CVPR*, pages 6011–6020. IEEE, 2020.

[133] J. O'Rourke and N. I. Badler. Model-based image analysis of human motion using constraint propagation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-2(6):522–536, 1980.

[134] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove. DeepSDF: Learning continuous signed distance functions for shape representation. In *CVPR*, pages 165–174. IEEE, 2019.

[135] D. Paschalidou, O. Ulusoy, C. Schmitt, L. Van Gool, and A. Geiger. Raynet: Learning volumetric 3d reconstruction with ray potentials. In *CVPR*, pages 3897–3906. IEEE, 2018.

[136] C. Patel, Z. Liao, and G. Pons-Moll. TailorNet: Predicting clothing in 3D as a function of human pose, shape and garment style. In *CVPR*, pages 7365–7375. IEEE, 2020.

[137] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis. Harvesting multiple views for marker-less 3D human pose annotations. In *CVPR*, pages 6988–6997. IEEE, 2017.

[138] G. Pavlakos, L. Zhu, X. Zhou, and K. Daniilidis. Learning to estimate 3D human pose and shape from a single color image. In *CVPR*, pages 459–468. IEEE, 2018.

[139] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. Osman, D. Tzionas, and M. J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *CVPR*, pages 10975–10985. IEEE, 2019.

[140] G. Pavlakos, N. Kolotouros, and K. Daniilidis. TexturePose: Supervising human mesh estimation with texture consistency. In *ICCV*, pages 803–812. IEEE, 2019.

[141] A. P. Pentland. Automatic extraction of deformable part models. *International Journal of Computer Vision*, 4:107–126, 1990.

[142] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. V. Gehler, and B. Schiele. DeepCut: Joint subset partition and labeling for multi person pose estimation. In *CVPR*, pages 4929–4937. IEEE, 2016.

[143] L. Pishchulin, S. Wuhrer, T. Helten, C. Theobalt, and B. Schiele. Building statistical shape spaces for 3D human modeling. *Pattern Recognition*, 67:276–286, 2017.

[144] R. Plaenkers and P. Fua. Model-based silhouette extraction for accurate people tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 325–339. Springer, 2002.

[145] R. Plankers and P. Fua. Articulated soft objects for video-based body modeling. In *ICCV*, volume 1, pages 394–401 vol.1. IEEE, 2001.

[146] G. Pons-Moll. *Human Pose Estimation from Video and Inertial Sensors*. -, 2014.

[147] G. Pons-Moll, J. Romero, N. Mahmood, and M. J. Black. DYNA: A model of dynamic human shape in motion. *ACM Trans. Graph.*, 34(4):1–14, 2015.

[148] G. Pons-Moll, S. Pujades, S. Hu, and M. J. Black. ClothCap: Seamless 4D clothing capture and retargeting. *ACM Trans. Graph.*, 36(4):1–15, 2017.

[149] R. Poppe. A survey on vision-based human action recognition. *Image and vision computing*, 28(6):976–990, 2010.

[150] S. S. Rautaray and A. Agrawal. Vision based hand gesture recognition for human computer interaction: a survey. *Artificial intelligence review*, 43(1):1–54, 2015.

[151] G. Riegler, A. Osman Ulusoy, and A. Geiger. OctNet: Learning deep 3D representations at high resolutions. In *CVPR*, pages 3577–3586. IEEE, 2017.

[152] K. M. Robinette, S. Blackwell, H. Daanen, M. Boehmer, and S. Fleming. Civilian american and european surface anthropometry resource (caesar), final report. volume 1. summary. Technical report, SYTRONICS INC DAYTON OH, 2002.

[153] K. Rohr. Towards model-based recognition of human movements in image sequences. *CVGIP: Image understanding*, 59(1):94–115, 1994.

[154] J. Romero, D. Tzionas, and M. J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Trans. Graph.*, 36(6), 2017.

[155] R. Rosales and S. Sclaroff. Inferring body pose without tracking body parts. In *CVPR*, volume 2, pages 721–727. IEEE, 2000.

[156] S. Saito, Z. Huang, R. Natsume, S. Morishima, A. Kanazawa, and H. Li. PIFu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *ICCV*, pages 2304–2314. IEEE, 2019.

[157] S. Saito, T. Simon, J. Saragih, and H. Joo. PIFuHD: Multi-level pixel-aligned implicit function for high-resolution 3D human digitization. In *CVPR*, pages 84–93. IEEE, 2020.

[158] A. Sappa, N. Aifanti, S. Malassiotis, and M. G. Strintzis. Monocular 3D human body reconstruction towards depth augmentation of television sequences. In *ICIP*, volume 3, pages 325–328. IEEE, 2003.

[159] A. Shapiro, A. Feng, R. Wang, H. Li, M. Bolas, G. Medioni, and E. Suma. Rapid avatar capture and simulation using commodity depth sensors. *Computer Animation and Virtual Worlds*, 25(3-4):201–211, 2014.

[160] L. Sigal, S. Bhatia, S. Roth, M. J. Black, and M. Isard. Tracking loose-limbed people. In *CVPR*, volume 1, pages I–I. IEEE, 2004.

[161] L. Sigal, A. Balan, and M. J. Black. Combined discriminative and generative articulated pose and non-rigid shape estimation. In *Advances in neural information processing systems (NIPS)*, pages 1337–1344, 2008.

[162] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[163] M. Slavcheva, M. Baust, D. Cremers, and S. Ilic. KillingFusion: Non-rigid 3D reconstruction without correspondences. In *CVPR*, pages 1386–1395. IEEE, 2017.

[164] B. M. Smith, V. Chari, A. Agrawal, J. M. Rehg, and R. Sever. Towards accurate 3D human body reconstruction from silhouettes. In *2019 International Conference on 3D Vision (3DV)*, pages 279–288. IEEE, 2019.

[165] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, pages 1–9. IEEE, 2015.

[166] V. Tan, I. Budvytis, and R. Cipolla. Indirect deep structured learning for 3D human body shape and pose prediction. In *BMVC*, 2017.

[167] M. Tatarchenko, A. Dosovitskiy, and T. Brox. Octree generating networks: Efficient convolutional architectures for high-resolution 3D outputs. In *ICCV*, pages 2088–2096. IEEE, 2017.

[168] D. Thalmann, Jianhua Shen, and E. Chauvineau. Fast realistic human body deformations for animation and VR applications. In *Proceedings of CG International*, pages 166–174, 1996.

[169] D. Tome, M. Toso, L. Agapito, and C. Russell. Rethinking pose in 3D: Multi-stage refinement and recovery for markerless motion capture. In *2018 international conference on 3D vision (3DV)*, pages 474–483. IEEE, 2018.

[170] J. Tong, J. Zhou, L. Liu, Z. Pan, and H. Yan. Scanning 3D full human bodies using kinects. *IEEE Transactions on Visualization and Computer Graphics*, 18(4):643–650, 2012.

[171] A. Toshev and C. Szegedy. DeepPose: Human pose estimation via deep neural networks. In *CVPR*, pages 1653–1660. IEEE, 2014.

[172] M. Trumble, A. Gilbert, C. Malleson, A. Hilton, and J. P. Collomosse. Total capture: 3D human pose estimation fusing video and inertial sensors. In *BMVC*, volume 2, pages 1–13, 2017.

[173] M. Trumble, A. Gilbert, A. Hilton, and J. Collomosse. Deep autoencoder for combined human pose estimation and body model upscaling. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 784–800, 2018.

[174] S. Tulsiani, T. Zhou, A. A. Efros, and J. Malik. Multi-view supervision for single-view reconstruction via differentiable ray consistency. In *CVPR*, pages 2626–2634. IEEE, 2017.

[175] H.-Y. Tung, H.-W. Tung, E. Yumer, and K. Fragkiadaki. Self-supervised learning of motion capture. In *Advances in Neural Information Processing Systems (NIPS)*, pages 5236–5246, 2017.

[176] M. A. Turk and A. P. Pentland. Face recognition using eigenfaces. In *CVPR*, pages 586–587. IEEE, 1991.

[177] G. Varol, J. Romero, X. Martin, N. Mahmood, M. J. Black, I. Laptev, and C. Schmid. Learning from synthetic humans. In *CVPR*, pages 109–117. IEEE, 2017.

[178] G. Varol, D. Ceylan, B. Russell, J. Yang, E. Yumer, I. Laptev, and C. Schmid. BodyNet: Volumetric inference of 3D human body shapes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 20–36, 2018.

[179] D. Vlasic, I. Baran, W. Matusik, and J. Popović. Articulated mesh animation from multi-view silhouettes. In *ACM SIGGRAPH 2008 papers*, pages 1–9, 2008.

[180] T. von Marcard, R. Henschel, M. J. Black, B. Rosenhahn, and G. Pons-Moll. Recovering accurate 3D human pose in the wild using IMUs and a moving camera. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 601–617. Springer, 2018.

[181] I. Wald, S. Woop, C. Benthin, G. S. Johnson, and M. Ernst. Embree: a kernel framework for efficient CPU ray tracing. *ACM Trans. Graph.*, 33(4):1–8, 2014.

[182] N. Wang, Y. Zhang, Z. Li, Y. Fu, W. Liu, and Y.-G. Jiang. Pixel2mesh: Generating 3D mesh models from single RGB images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 52–67, 2018.

[183] R. Y. Wang, K. Pulli, and J. Popović. Real-time enveloping with rotational regression. In *ACM SIGGRAPH 2007*, pages 73–82, 2007.

[184] O. Weber, O. Sorkine, Y. Lipman, and C. Gotsman. Context-aware skeletal shape deformation. In *Computer Graphics Forum*, volume 26, pages 265–274. Wiley Online Library, 2007.

[185] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *CVPR*, pages 4724–4732. IEEE, 2016.

[186] A. Weiss, D. Hirshberg, and M. J. Black. Home 3D body scans from noisy image and range data. In *ICCV*, pages 1951–1958. IEEE, 2011.

[187] C. Wen, Y. Zhang, Z. Li, and Y. Fu. Pixel2mesh++: Multi-view 3D mesh generation via deformation. In *CVPR*, pages 1042–1051. IEEE, 2019.

[188] J. Wu, C. Zhang, X. Zhang, Z. Zhang, W. T. Freeman, and J. B. Tenenbaum. Learning shape priors for single-view 3d completion and reconstruction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 646–662, 2018.

[189] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao. 3D shapenets: A deep representation for volumetric shapes. In *CVPR*, pages 1912–1920. IEEE, 2015.

[190] F. Xia, P. Wang, L.-C. Chen, and A. L. Yuille. Zoom better to see clearer: Human part segmentation with auto zoom net. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 648–663. Springer, 2016.

[191] F. Xia, P. Wang, X. Chen, and A. L. Yuille. Joint multi-person pose estimation and semantic part segmentation. In *CVPR*, pages 6769–6778. IEEE, 2017.

[192] B. Xiao, H. Wu, and Y. Wei. Simple baselines for human pose estimation and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 466–481, 2018.

[193] R. Xie, C. Wang, and Y. Wang. MetaFuse: A pre-trained fusion model for human pose estimation. In *CVPR*, pages 13686–13695. IEEE, 2020.

[194] L. Xu, Z. Su, L. Han, T. Yu, Y. Liu, and F. Lu. UnstructuredFusion: Realtime 4D geometry and texture reconstruction using commercial RGBD cameras. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[195] Q. Xu, W. Wang, D. Ceylan, R. Mech, and U. Neumann. DISN: Deep implicit surface network for high-quality single-view 3D reconstruction. In *Advances in Neural Information Processing Systems (NIPS)*, pages 492–502, 2019.

[196] W. Xu, A. Chatterjee, M. Zollhöfer, H. Rhodin, D. Mehta, H.-P. Seidel, and C. Theobalt. MonoPerfCap: Human performance capture from monocular video. *ACM Trans. Graph.*, 37(2):1–15, 2018.

[197] G. Ye, Y. Liu, N. Hasler, X. Ji, Q. Dai, and C. Theobalt. Performance capture of interacting characters with handheld kinects. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 828–841. Springer, 2012.

[198] T. Yu, K. Guo, F. Xu, Y. Dong, Z. Su, J. Zhao, J. Li, Q. Dai, and Y. Liu. BodyFusion: Real-time capture of human motion and surface geometry using a single depth camera. In *ICCV*, pages 910–919. IEEE, 2017.

[199] T. Yu, Z. Zheng, K. Guo, J. Zhao, Q. Dai, H. Li, G. Pons-Moll, and Y. Liu. DoubleFusion: Real-time capture of human performances with inner body shapes from a single depth sensor. In *CVPR*, pages 7287–7296. IEEE, 2018.

[200] A. Zanfir, E. Marinoiu, M. Zanfir, A.-I. Popa, and C. Sminchisescu. Deep network for the integrated 3D sensing of multiple people in natural images. In *Advances in Neural Information Processing Systems (NIPS)*, pages 8410–8419, 2018.

[201] J. Y. Zhang, P. Felsen, A. Kanazawa, and J. Malik. Predicting 3D human dynamics from video. In *ICCV*, pages 7114–7123. IEEE, 2019.

[202] Q. Zhang, B. Fu, M. Ye, and R. Yang. Quality dynamic human body modeling using a single low-cost depth camera. In *CVPR*, pages 676–683. IEEE, 2014.

[203] Z. Zheng, T. Yu, Y. Wei, Q. Dai, and Y. Liu. DeepHuman: 3D human reconstruction from a single image. In *ICCV*, pages 7739–7749. IEEE, 2019.

[204] H. Zhu, X. Zuo, S. Wang, X. Cao, and R. Yang. Detailed human shape estimation from a single image by hierarchical mesh deformation. In *CVPR*, pages 4491–4500. IEEE, 2019.

[205] M. Zollhöfer, M. Nießner, S. Izadi, C. Rehmann, C. Zach, M. Fisher, C. Wu, A. Fitzgibbon, C. Loop, C. Theobalt, et al. Real-time non-rigid reconstruction using an RGB-D camera. *ACM Trans. Graph.*, 33(4):1–12, 2014.