

# Northumbria Research Link

Citation: Hu, Pengpeng, Ho, Edmond and Munteanu, Adrian (2022) 3DBodyNet: Fast Reconstruction of 3D Animatable Human Body Shape from a Single Commodity Depth Camera. IEEE Transactions on Multimedia, 24. pp. 2139-2149. ISSN 1520-9210

Published by: IEEE

URL: <https://doi.org/10.1109/tmm.2021.3076340>  
<<https://doi.org/10.1109/tmm.2021.3076340>>

This version was downloaded from Northumbria Research Link:  
<http://nrl.northumbria.ac.uk/id/eprint/46020/>

Northumbria University has developed Northumbria Research Link (NRL) to enable users to access the University's research output. Copyright © and moral rights for items on NRL are retained by the individual author(s) and/or other copyright owners. Single copies of full items can be reproduced, displayed or performed, and given to third parties in any format or medium for personal research or study, educational, or not-for-profit purposes without prior permission or charge, provided the authors, title and full bibliographic details are given, as well as a hyperlink and/or URL to the original metadata page. The content must not be changed in any way. Full items must not be sold commercially in any format or medium without formal permission of the copyright holder. The full policy is available online: <http://nrl.northumbria.ac.uk/policies.html>

This document may differ from the final, published version of the research and has been made available online in accordance with publisher policies. To read and/or cite from the published version of the research, please visit the publisher's website (a subscription may be required.)



**Northumbria  
University**  
NEWCASTLE



**UniversityLibrary**

# 3DBodyNet: Fast Reconstruction of 3D Animatable Human Body Shape from a Single Commodity Depth Camera

Pengpeng Hu, Edmond S. L Ho, and Adrian Munteanu

**Abstract**—Knowledge about individual body shape has numerous applications in various domains such as healthcare, fashion and personalized entertainment. Most of the depth based whole body scanners need multiple cameras surrounding the user and requiring the user to keep a canonical pose strictly during capturing depth images. These scanning devices are expensive and need professional knowledge for operation. In order to make 3D scanning as easy-to-use and fast as possible, there is a great demand to simplify the process and to reduce the hardware requirements. In this paper, we propose a deep learning algorithm, dubbed 3DBodyNet, to rapidly reconstruct the 3D shape of human bodies using a single commodity depth camera. As easy-to-use as taking a photo using a mobile phone, our algorithm only needs two depth images of the front-facing and back-facing bodies. The proposed algorithm has strong operability since it is insensitive to the pose and the pose variations between the two depth images. It can also reconstruct an accurate body shape for users under tight/loose clothing. Another advantage of our method is the ability to generate an animatable human body model. Extensive experimental results show that the proposed method enables robust and easy-to-use animatable human body reconstruction, and outperforms the state-of-the-art methods with respect to running time and accuracy.

**Index Terms**—Human body shape, Body shape under clothing, depth camera, 3D Scanning, deep learning on point clouds

## I. INTRODUCTION

**I**N order to create an accurate 3D shape of the human body in fashion industry or healthcare applications, one either relies on the manual entry of body measurements or on scanning the body using a professional 3D scanner. The measurement-based methods, however, need hundreds of different measures to ensure an accurate body shape reconstruction [1]. Manually extracting such a large number of measurements is a tedious operation that requires professional intervention. Using 3D scanning technologies based on laser or structured light, detailed human models can be created. However, these devices are expensive and require expert knowledge for operation [2]. Multi-view stereo is another solution for human modeling [3], but such methods are very slow due to computational complexity, and often fail due to depth ambiguities or complex occlusions among

different views [4].

Commodity depth cameras, such as the Microsoft Kinect or Intel Realsense, have become increasingly popular 3D devices in recent years and have been recently integrated on mobile phones like iPhone X. The depth image provides range information of the object of interest based on which a 3D reconstruction can be generated. Commodity depth cameras have been used widely in various applications [5][6][7], among which several scanning systems were proposed for body shape reconstruction [8][9][10].

3D body shape reconstruction methods can be classified mainly into three categories: non-parametric, parametric, and template-based methods. Non-parametric methods correspond to the traditional 3D scanning techniques which require point clouds captured by depth cameras from different views in order to reconstruct the overall shape. The point clouds must be registered into a complete shape using rigid [11] and non-rigid [12] registration techniques typically employed for static and dynamic body scans respectively. Parametric methods rely on a parametric body model, e.g. the SCAPE model [13] or the SMPL model [14]. Parametric models factor out the body model using parameters that control the shape and the pose. In practice, in order to obtain the optimal parameters, the parametric body model is fitted to the input data, which usually includes 3D point clouds, 2D RGB images or silhouettes. As opposed to non-parametric methods, parametric modelling can directly generate a watertight, clean body mesh. This enables downstream applications such as virtual try-on, virtual reality and computer animation. Template-based methods deform a specific body template to fit the input body point cloud by non-rigid registration; this enables filling missing areas in the point cloud [15] or finding correspondences [16]. The template can be chosen from the parametric model or handcrafted [16].

In this paper, we propose a novel parametric method for rapidly reconstructing the animatable body shape using a single depth camera. We employ an SMPL parametric model and train an end-to-end network, that (i) takes only two depth images of the front-facing and back-facing human body, captured by a single handheld depth camera, and (ii) regresses the shape parameters of the model. Users do not need to strictly maintain a canonical pose, on the contrary, they are allowed to have an arbitrary pose and even change their pose with large variations when they are captured by the handheld depth camera. The lack of

P. Hu is with the Department of Electronics and Informatics, Vrije Universiteit Brussel, Brussels, Belgium, email: phu@etrovub.be.

E.S.L Ho is with the Department of Computer and Information Sciences, Northumbria University, Newcastle, UK, email: e.ho@northumbria.ac.uk

A. Munteanu is with the Department of Electronics and Informatics, Vrije Universiteit Brussel, Brussels, Belgium, email: acmuntea@etrovub.be (see: <http://www.etrovub.be/acmuntea>).

constraints on freezing the pose makes the 3D scanning closer to a snapshot, becoming more of a user-friendly experience. The main contributions in this paper can be summarized as follows:

- We propose a novel deep learning framework, termed 3DBodyNet, that reconstructs the body model using a single commodity depth camera, by only taking two front- and back-facing depth images as input without the assumption of view alignment. To our best knowledge, it is the first deep learning method that reconstructs the body shape from only two depth images and at the same time it allows for large pose variations between the camera shots. In addition, the reconstructed body model is animatable which facilitates its use in subsequent applications such as garment transfer and virtual fitting.
- We demonstrate that the proposed 3DBodyNet can also work for estimating the body shape under clothing.
- We introduce a large-scale dataset including male/female dressed/undressed bodies and employ it in the task of body shape reconstruction from the front-facing and/or the back-facing depth images.
- We perform comprehensive experiments to validate the proposed method and compare it against the existing state-of-the-art methods, and demonstrate its superiority in objective and subjective terms.

## II. RELATED WORK

### A. Non-parametric body reconstruction

Non-parametric methods include multi-camera and single-camera scanning systems. Multi-camera systems [9][17][18] employ several depth cameras at various positions in order to capture the subject from different viewpoints. Such systems are heavily depending on the quality of extrinsic calibration, and the noise and missing areas in the captured point cloud cannot be avoided. In addition, a multi-camera system needs professional calibration and is expensive to the individual customer. Reconstructing the body shape using a single depth camera is a promising alternative. The pioneering work in this direction is given by KinectFusion [8], which proposes a real-time 3D reconstruction algorithm for complex and arbitrary indoor scenes. Its limitation is that it fails when the objects in the scene are not static or the depth camera moves fast. [19] proposes a non-rigid registration algorithm for modeling human bodies, using a single depth camera. However, non-rigid registration can only tolerate small movements of the subject (e.g. shaking the arms), and it needs to solve a complex optimization problem which is time consuming. In addition, a single-camera scanning system needs a longer capture time compared to a multi-camera scanning system. In this work, we propose a novel method that has the advantages of both categories of scanners, namely, fast data acquisition, fast shape reconstruction and a low-cost device. Moreover, our method can output an animatable body shape which is not possible using existing non-parametric methods.

### B. Parametric body reconstruction

Parametric methods became increasingly popular for human body modeling, several methods being recently proposed in

the literature. [20] trained a neural network to reconstruct the body surface by fitting the SMPL body to the input body point cloud. This method requires a complete body point cloud as input, which is not always available. [13] deforms the SCAPE model to fit the scanned data, given a set of markers and specifying the target shape. Given a static body model and markers for motion capture, this method can produce a moving person but still needs to take a complete body as input. Similarly, [10] deforms a SMPL model to a low-quality body scan by manually specifying landmarks. [2] fits their parametric body model to two separate scans of the front-facing and back-facing body and merge them. This method, however, requires the user to keep a canonical standing pose and needs to solve a set of optimization problems. Our work is mainly inspired by the work of [2]. The main functional differences with respect to [2] are: 1) we do not enforce users to freeze in a given pose when acquiring the front- and back-facing scans, 2) we address this inherent pose variation between scans by using a novel deep learning solution, avoiding solving computationally expensive and time-consuming optimizations, 3) our reconstructed body is animatable and 4) our algorithm can reconstruct the body shape under clothing. Another interesting technology is 3D shape reconstruction from a single RGB image [10][21][22][23]; these kinds of methods can only produce visually consistent 3D models rather than accurate 3D shapes due to the lack of depth information and scale ambiguity from 2D to 3D.

### C. Body Shape Under Clothing

The human body is usually covered by layers of clothing. To capture the body geometry via 3D scanning, the subject is asked to wear minimal or very tight and thin clothing for accurate data acquisition [2][24]. However, the procedure is inconvenient and it also raises questions regarding the right to privacy. To address this issue, several studies have been performed in the recent past. Early works fitted the body template into the dressed body model [25][26][27] by solving complex optimization problems. [28] proposed the first deep learning method, Body PointNet, for estimating body shape under clothing from 3D data. However, it requires a complete dressed body scan as input, which is not always available. [29] proposed to reconstruct the body shape and the garments from RGB images. Due to the scale ambiguity from 2D to 3D and lack of depth information, this method assumed that the subjects have the known height which severely affects the estimation accuracy. However, it is a very strong assumption which is not always known. One of the goals in our study is to accurately model the body shape, so our method employs depth images instead of RGB images as input.

### D. Deep Learning on Point Clouds

Inspired by the success of deep learning in 2D application, e.g. image classification [30] and human pose estimation [31], many research works have proposed methods to feed 3D data into neural works for specific tasks. Although other representations, including converting 3D data to regular volumetric data [32] or multi-view 2D images [33] existed, our method takes the point clouds as input since point

clouds can be directly obtained from depth images given the camera intrinsic parameters. The pioneering work, PointNet [34], became a very popular deep method for processing unstructured point clouds. It utilizes a pointwise multi-layer perceptron with a symmetric aggregation function to achieve invariance to permutations, which shows a good performance for extracting features from point clouds. PointNet learns the global features from points while variants of PointNet [35][36] have been proposed to extract the local features. PointNet has been successfully applied to many tasks [16][37][38][39]. To the best of our knowledge, this study is the first work extending the use of PointNet for reconstructing the body shape from only two front- and back-facing depth images obtained with a handheld device.

### III. METHODOLOGY

**Problem Statement** In this section, we formulate the problem in this study. Given a front-facing point cloud of the body  $X = \{x_i \in \mathbb{R}^3, i = 1, \dots, N\}$  and a back-facing point cloud of the body  $Y = \{y_i \in \mathbb{R}^3, i = 1, \dots, M\}$ , the goal is to devise a low-complexity, computationally affordable method that reconstructs an animatable body  $B = \{(v_i \in \mathbb{R}^3, e_j \in \mathbb{Z}^2, j_m \in \mathbb{R}^3, s_i \in \mathbb{R}^D), i = 1, \dots, P, j = 1, \dots, Q, m = 1, \dots, W\}$ , where  $v_i, e_j, j_m$  and  $s_i$  are the vertices, edges, joints and the skinning weights of vertices in  $B$  respectively.  $D$  is the number of attaching joints for each vertex. The existing state-of-the-art method [2] addressed this problem by factoring it into three sub-problems namely, (i) deforming a parametric body template to fit the front-facing scan, (ii) deforming a parametric body template to fit the back-facing scan, and (iii) stitching these two half-body shapes to a fully body. Although this formulation is intuitive, it is prone to template-based fitting errors and it requires the user to keep a canonical pose but tolerating small pose variations between the two partial scans. To avoid these problems, our learning-based approach is (i) to extract features of the front-facing and back-facing point clouds of the body, (ii) to regress the shape parameters of the parametric body model (e.g. the SMPL model) from the features of inputs. We approach it by using supervised learning. Figure 1 illustrates the architecture of the proposed method, dubbed 3DBodyNet. It mainly consists of a Pose-invariant Feature Module (PFM), Parametric Module (PM), Sticking Module (SM) and the SMPL layer (SMPLL).

**Preliminary** In this study, the Skinned Multi-Person Linear (SMPL) model of [14] is used to encode the 3D mesh of a human body due to its good trade-off between high anatomic flexibility and realism. SMPL parameterizes a body mesh with shape and pose parameters. The shape  $\beta \in \mathbb{R}^{10}$  is represented by the first ten coefficients of a PCA shape space. The pose  $\theta \in \mathbb{R}^{3K}$  is parameterized by relative 3D rotation of  $K = 23$  joints in axis-angle representation. Given a  $\beta$  and a  $\theta$ , a triangulated mesh  $M(\beta, \theta)$  with  $N = 6890$  vertices and  $T = 13776$  triangles can be generated. The shape  $B_s(\beta)$  and pose-dependent deformations  $B_p(\theta)$  are first applied to an average body  $M_\mu$  to generate a human body with a specific body shape:

$$T(\beta, \theta) = M_\mu + B_s(\beta) + B_p(\theta) \quad (1)$$

The pose can be controlled by adjusting the joint angles  $J(\beta)$  using the skinning function  $W$ , yielding a human body mesh with a specific shape and pose:

$$M(\beta, \theta) = W(T(\beta, \theta), J(\beta), \theta) \quad (2)$$

SMPL is fully differentiable with respect to  $\beta$  and  $\theta$ . We integrate SMPL as a part of our deep learning model. Thanks to SMPL, the troubles of noisy outputs [39] and the rugged and twisted shapes [16] typically impairing human model generation can be avoided.

#### A. Preprocessing

Given the raw data from the commodity depth camera, the body data is extracted by setting the distance thresholds. We normalized the point clouds in two steps: (i) it is centered to the origin, (ii) and it is scaled by dividing the Z-axis length of its bounding box. Next, we sample fixed number of points to feed them into the deep neural network.

#### B. Pose-invariant Features

Point clouds are unstructured sets, which are not trivial for direct analysis. Traditional features of point sets are handcrafted intrinsically or extrinsically. However, such features are designed for specific tasks. In this study, the features of point sets should be pose invariant. To this end, a learning strategy is adopted. The goal is to find an embedding that is invariant to the posed partial scans of the body. We evaluate two popular choices of learnable embedding modules, namely PointNet [34] and DGCNN [36].

PointNet, the pioneering work of learning on point sets, samples  $M$  points from the raw  $N$  points given as input. Each point is embedded by a nonlinear function from  $\mathbb{R}^3$  into a higher-dimensional space; the output of PointNet is a  $K$ -sized global feature for the whole set of  $M$  points obtained by using a symmetric channel-wise aggregation function (e.g.,  $\max$  or  $\sum$ ). Let  $x_i^l$  be the embedding of point  $i$  in the  $l$ -th layer and let  $h_\theta^l$  be the nonlinear function in the  $l$ -th layer. The forward mechanism can be denoted by  $x_i^l = h_\theta^l(x_i^{l-1})$ .

PointNet performs a per-point embedding and does not capture local structures. In contrast, DGCNN [36] captures local geometric structures by constructing a local neighborhood graph and applying convolution-like operations on the edges connecting neighboring pairs of points. The forward mechanism of DGCNN can be represented as:

$$x_i^l = f(\{h_\theta^l(x_i^{l-1}, x_j^{l-1})\}), j \in N_i \quad (3)$$

where  $N_i$  represents the set of neighbors of vertex  $i$  in the graph.

As shown in Figure 1, our method takes two scans as input and outputs a complete body shape. There are no restrictions on the two inputs scans. In our approach we propose two feature extractors, termed FPFM and BPFM, to extract the pose-invariant features from the front- and back-facing scans respectively. FPFM and BPFM have the same architecture but the weights are different. In addition, the human subjects

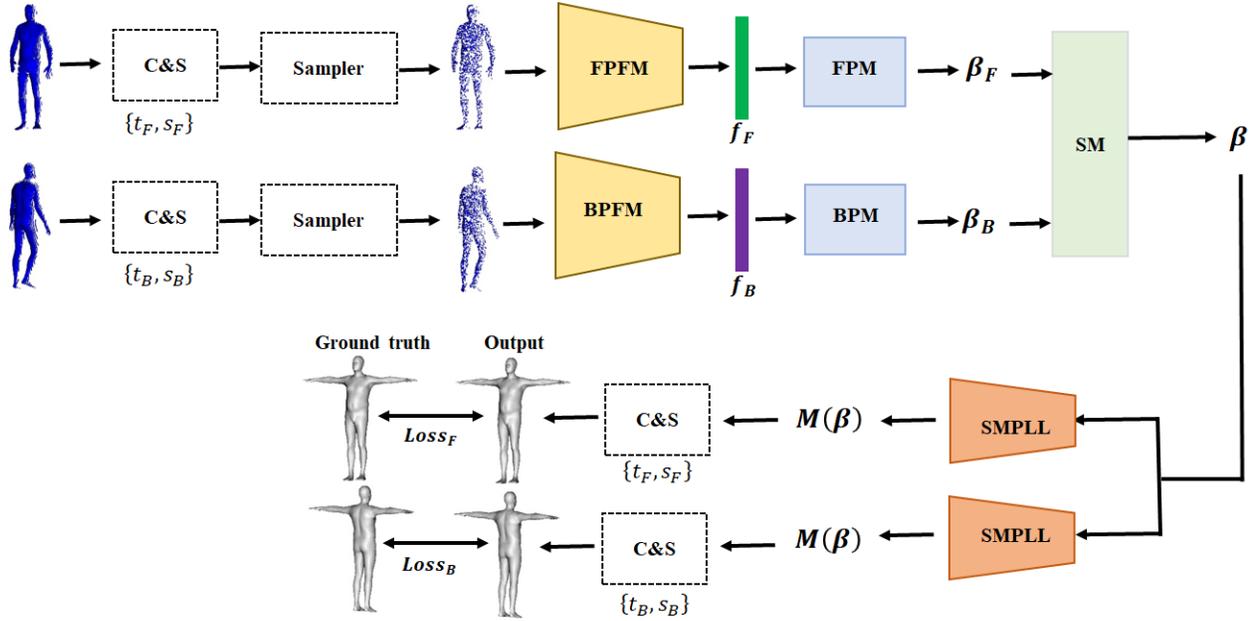


Fig. 1. Architecture of the proposed 3DBodyNet. C&S: the normalization operation that consists of centering to the centroid of the point cloud and scaling it into a unit sphere;  $t$  and  $s$ : the translation vector and scale value for the input data; FPFM and BPFM: pose-invariant feature modules for the front-facing and back-facing body point clouds respectively;  $f$ : the feature encoded in a 1024-dimensional vector; FPM and BPM: parametric modules for converting the features to the  $\beta$  values of SMPL; SM: the stitching module that stitches the  $\beta$ s from the two partial point clouds of the body into a single  $\beta$  value; SMPLL: Skinned Multi-Person Linear model layer.

do not need to maintain a certain pose. More specifically, the subjects can perform arbitrary poses during scanning, we do not assume any prior/ physical relationship between the two poses. As shown in the experimental section, although the input two scans have different poses, the reconstructed body shape is morphologically correct and robust against strong pose variations in the input. In this study, we sample  $M = 2048$  points as input of the proposed PFM and set  $K = 1024$ .

### C. Parametric Module

Given the embedding of the inputs, alternative methods directly regress points of the whole body from an decoder. These methods however, can only output a point cloud [39] or noisy mesh [16], which is also not animatable. To cope with these issues, we design the front-facing parametric module (FPM) and back-facing parametric (BPM) to map the features from the FPFM and BPFM to the shape space of SMPL. More specifically, we directly regress  $\beta$  from the features of point clouds. In this study, the PM consists of multilayer perceptrons (MLPs) with 1024, 1024, 1024 and 10 neurons. FPM and BPM share the same architecture but with different learned weights.

### D. Stitching Module

As shown in Figure 1, two  $\beta$  values ( $\beta_F$  and  $\beta_B$ ), are obtained from the FPM and BFM respectively. However, our goal is to output a single accurate  $\beta$  value. In order to stitch the  $\beta_F$  and  $\beta_B$  together, we design the stitching module (SM) that takes the  $\beta_F$  and  $\beta_B$  as input and outputs an accurate  $\beta$ . To design

the SM, we present three alternative approaches (i) We use a multilayer perceptron to predict  $\beta$  from the concatenation of  $\beta_F$  and  $\beta_B$ ; (ii) We use a multilayer perceptron to estimate  $\beta$  from the concatenation of the  $f_F$  and  $f_B$  that are the learned features of two inputs; (iii) We propose a  $\beta$  based mean pooling operation defined as:

$$\beta = \frac{\beta_F + \beta_B}{2} \quad (4)$$

We will show that the proposed  $\beta$ -based mean pooling operation outperforms these MLP-based method of stitching  $\beta$  in the following ablation study.

### E. SMPL Layer

With the concatenation of introduced PFMs, PMs and the SM, the network can be trained using  $L_2^\beta = \|\beta - \beta_{GT}\|^2$  or  $L_1 = \|\beta - \beta_{GT}\|$  as loss. However, the performance of directly regressing  $\beta$  from the point clouds is not acceptable due to the high non-linearity of SMPL. Our insight is to propose a powerful constraint to guide the learning process. To this end, we integrate SMPL as our final module in our architecture. SMPL is a fully differentiable function that can back-propagate gradients through the network. As we set  $\theta$  to zero, Equation 2 can be rewritten:

$$M(\beta) = W(T(\beta), J(\beta)) \quad (5)$$

### F. Losses

As shown in Figure 1, the module denoted by the dotted lines are not trainable, while the rest of modules are trainable

and are parameterized by a set of neural network weights learned during training. The SMPL layer acts as a pre-trained high-quality mesh decoder. Although the SMPL layer is not parameterized by the neural network weights, its output depends on the results of the stitching module. We train the network in a supervised manner, and propose the following loss function to measure the reconstruction error.

**Vertex Loss.** The ground truth body shape is pre-aligned with the input point clouds of the front-facing and back-facing body respectively using the camera extrinsic parameters obtained during the rendering procedure. We defined the front-facing based reconstruction error as:

$$L_{vert}^F = \frac{1}{|P|} \sum_{x \in P} \min_{y \in P_{GT}} \|x - y\|^2 \quad (6)$$

where  $P = \{p_i \in \mathbb{R}^3, i = 1, \dots, 6890\}$  is the set of vertices of the mesh from the SMPL layer while  $P_{GT}$  is the set of ground truth vertices. The back-facing based reconstruction error  $L_{vert}^B$  is defined the same as  $L_{vert}^F$ .

**Joint Loss.** The model from SMPL has the skinning information, such as the skeleton, consisting of joints and skinning weights. Similar to the vertex loss, we define the front-facing based joint error as:

$$L_{joint}^F = \frac{1}{N} \sum_{i=1}^N \|j_i - j_i^{GT}\|^2 \quad (7)$$

where  $j_i$  denotes the position of the  $i_{th}$  joint. The back-facing based joint error  $L_{joint}^B$  is the same as  $L_{joint}^F$ .

**$\beta$  Loss.** Besides the 3D vertex and joint losses, the ground truth  $\beta$  is also included for supervision of the training.

$$L_{\beta} = \|\beta - \beta_{GT}\|^2 \quad (8)$$

As the two input point clouds are captured from the same body with different postures, the predicted  $\beta_F$  and  $\beta_B$  are supposed to be the same. We, thus, define an additional regularization loss:

$$L_{reg} = \|\beta_F - \beta_B\|^2 \quad (9)$$

**Complete Loss.** Our complete loss is defined as:

$$Loss = L_{vert}^F + L_{vert}^B + \lambda_{joint}^F * L_{joint}^F + \lambda_{joint}^B * L_{joint}^B + L_{\beta} + L_{reg} \quad (10)$$

where  $\lambda_{joint}^F$  and  $\lambda_{joint}^B$  are the scalar weights.

#### IV. DATASET

In order to train our algorithm, we require a large set of body shapes. More specifically, the dataset requires (i) the front-facing and back-facing depth paired images having the same, similar and totally different postures; (ii) it contains both dressed and undressed bodies; (iii) it should be large-scale. None of the existing datasets can meet these requirements, thus, we propose a new synthetic dataset, dubbed FBB (the **F**ront-facing and **B**ack-facing depth images of human **B**odies), for training our model.

**Posed Body Shapes.** To make use of realistic human body

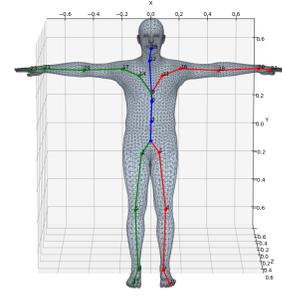


Fig. 2. Layout of 23 joints in the SMPL models.

shapes, we collect a set of 1700  $\beta$  values for the SMPL model from the SURREAL dataset [40]. To mimic the poses of humans, we collect 2667  $\theta$  values for SMPL model from the SURREAL dataset. Considering existing scanning system usually requires that the user to keep "A" pose during capture, we propose a simple yet efficient generative algorithm to mimic a set of "A" poses. As shown in Figure 2, SMPL has 23 joints. It is observed that these joints may have different DOF (degrees of freedom). For example, the knee and elbow joints are hinge joints with 1 DOF while the wrist joint is a universal joint with 2 DOF. Taking into consideration all these factors, we select 18 key joints and assign different probability density functions of uniform distribution for them in order to simulate possible joint angles (see Table I). The rest of joint angles are set to  $\{0\}$ . Each sample of the dataset is built by randomly combining one SMPL  $\beta$  value and two  $\theta$  values. The female and male data are produced separately. Our final dataset has  $3 \cdot 10^5$  samples. By setting all the joint angles to be  $\{0\}$ , we have the human meshes and joints in "T" pose, which are our ground truth data.

**Dressed Bodies.** This paper proposes a method for reconstructing the body shape using a single depth camera, which can be used for subjects with/without clothes. To show its effectiveness for estimating the human body shape under clothing, we put one type of popular clothes (shoes+long-sleeved shirt and long pants) onto the body dataset used for training. Experiments are then performed to validate the idea of body shape estimation under clothing. Other types of clothing can be easily prepared using the same method. This strategy is also applied in [28], which is the state-of-the-art in estimating body shape under clothing from a 3D scan. One notes that the average time of putting clothes on one SMPL body is only 0.6 seconds. This enables generating large training datasets with a broad range of body morphologies for various categories of clothing; for more details, the interested reader is referred to [28].

**Rendering.** The open-source Blender Sensor Simulation plugin Bensor [41] is used for rendering the front-facing and back-facing depth images of bodies. These depth images are further converted to partial point clouds given the intrinsic parameters of the camera. We set the camera as Microsoft Kinect V2 and the max distance as 3.6 meters. To increase realism and generate more variability in point density, the position of camera is randomly selected at the intervals from 1.5 to 2.5 meters; the orientation of camera is set by random

rotation angles at the intervals from  $-10^\circ$  to  $10^\circ$  in all  $x, y, z$  directions. As introduced above, each sample of the body dataset has one body instance with two poses. Therefore, our final dataset has  $1.2 \cdot 10^6$  pair of front-facing and back-facing scans of the body (the 50% for the dressed body and the other 50% for the undressed body ) and  $3 \cdot 10^5$  ground truth body shapes.

TABLE I  
GENERATIVE ALGORITHM OF "A" POSES OF HUMAN.

Joint ID	X angle	Y angle	Z angle
#1	$(-\frac{\pi}{36}, \frac{\pi}{36})$	$(-\frac{\pi}{36}, \frac{\pi}{36})$	$(-\frac{\pi}{36}, \frac{\pi}{9})$
#2	$(-\frac{\pi}{36}, \frac{\pi}{36})$	$(-\frac{\pi}{36}, \frac{\pi}{36})$	$(-\frac{\pi}{9}, \frac{\pi}{36})$
#3	$(-\frac{\pi}{36}, \frac{\pi}{36})$	$(-\frac{\pi}{36}, \frac{\pi}{36})$	$(-\frac{\pi}{36}, \frac{\pi}{36})$
#4	$(-\frac{\pi}{18}, \frac{\pi}{18})$	0	0
#5	$(-\frac{\pi}{18}, \frac{\pi}{18})$	0	0
#7	0	$(-\frac{\pi}{18}, \frac{\pi}{6})$	0
#8	0	$(-\frac{\pi}{6}, \frac{\pi}{18})$	0
#12	$(-\frac{\pi}{18}, \frac{\pi}{18})$	$(-\frac{\pi}{36}, \frac{\pi}{36})$	$(-\frac{\pi}{36}, \frac{\pi}{9})$
#13	0	0	$(-\frac{\pi}{6}, \frac{\pi}{6})$
#14	0	0	$(-\frac{\pi}{6}, \frac{\pi}{6})$
#16	$(-\frac{\pi}{18}, \frac{\pi}{18})$	$(-\frac{\pi}{36}, \frac{\pi}{36})$	$(-\frac{\pi}{3}, \frac{\pi}{6})$
#17	$(-\frac{\pi}{18}, \frac{\pi}{18})$	$(-\frac{\pi}{36}, \frac{\pi}{36})$	$(-\frac{\pi}{6}, \frac{\pi}{3})$
#18	$(-\frac{\pi}{18}, \frac{\pi}{18})$	0	$(-\frac{\pi}{18}, 0)$
#19	$(-\frac{\pi}{18}, \frac{\pi}{18})$	0	$(0, \frac{\pi}{18})$
#20	$(-\frac{\pi}{4}, \frac{\pi}{4})$	0	$(-\frac{\pi}{6}, \frac{\pi}{6})$
#21	$(-\frac{\pi}{4}, \frac{\pi}{4})$	0	$(-\frac{\pi}{6}, \frac{\pi}{6})$
#22	$(-\frac{\pi}{6}, \frac{\pi}{6})$	0	$(-\frac{\pi}{6}, \frac{\pi}{6})$
#23	$(-\frac{\pi}{6}, \frac{\pi}{6})$	0	$(-\frac{\pi}{6}, \frac{\pi}{6})$

## V. EXPERIMENTS

### A. Training setup

We split the dataset into training, validation, and testing using 99% , 0.7% and 0.3% of the samples in the dataset respectively. We train the model for male, female, dressed and undressed data separately. The training is carried out using the Adam optimizer [42] with an initial learning rate of 0.0001 for 50 epochs and a batch size of 16. The training is performed on a desktop PC (Intel(R) Xeon(R) Silver 4112 CPU @2.60GHz 64GB RAM GPU GeForce GTX 1080Ti) based on TensorFlow [43]. We set  $\lambda_{joint}^F = 0.001$  and  $\lambda_{joint}^B = 0.001$  in the loss.

### B. Evaluation metrics

To evaluate the performance of our algorithm, we employ the widely-used reconstruction evaluation metric: vertex-to-vertex error (v2v). This metric is also used in the work of [2] that we mainly used for comparison with our method. The v2v error measures the average Euclidean distance from a vertex of the reconstructed body shape  $V_{recon}$  to its closest point of the ground truth body shape  $V_{gt}$ . The v2v error is defined as:

$$E(V_{recon}, V_{gt}) = \frac{1}{|V_{recon}|} \sum_{x \in V_{recon}} \min_{y \in V_{gt}} \|x - y\|^2 \quad (11)$$

In our experiments, we calculate the average value  $\mu$  and average standard derivation  $\sigma$  of the V2V error.

### C. Qualitative Results and Comparisons

**Human Body Shape Reconstruction.** In this experiment, we test our algorithm based on two public datasets: the PDT13 [44] and FAUST dataset [45]. In the PDT13 dataset, the real human subjects are scanned from the front-facing and back-facing view using the Kinect sensor. But the subject is asked to keep the same posture during scanning, and no accurate ground truth meshes can be used for quantitative comparison. In the FAUST dataset, the subjects wear minimal clothing in different postures. However, no T-pose ground truth meshes are available.

In Figure 3, we compare our results with the results of [44] based on the PDT13 data. Note that our method is robust to the outlier that is observed in the region of inputs highlighted by the red bounding box. It is noticeable that the head shape using the method of [44] is not well reconstructed, while our method can generate high-quality head shapes. Figure 4 depicts our results on the FAUST dataset. It can be seen that the input front-facing and back-facing point clouds have a very large posture variation. However, our method is robust to these large posture variations, and the reconstructed body shapes are visually consistent with the inputs.

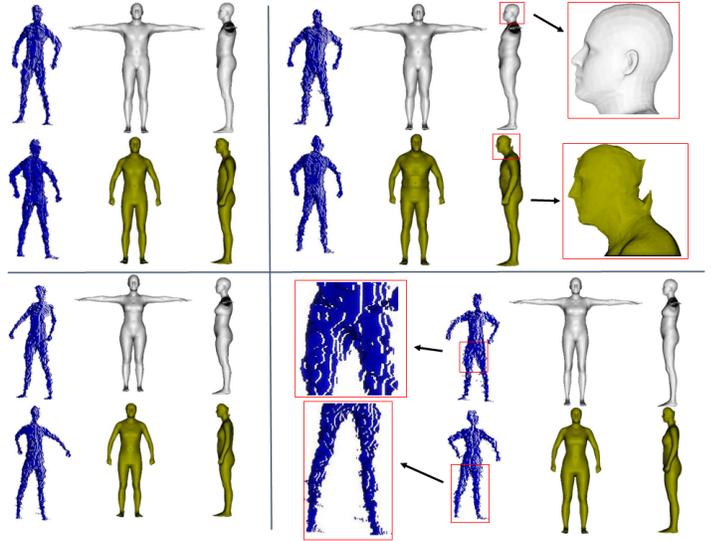


Fig. 3. Comparison our method with the method of [44] on the PDT13 data. The two input point clouds are colored in blue, our results are the white meshes, while the yellow meshes are the results from [44].

**Body Shape Under Clothing.** In this experiment, we test our algorithm based on the BUFF dataset [27] for the task of estimating body shape under clothing. BUFF is a scanned dressed body dataset obtained by capturing 6 real subjects wearing 2 clothing styles (T-shirt and long pants, and soccer outfit) in two motion sequences. Figure 5 shows the results obtained with our method on the BUFF dataset. It is noticeable that the T-shirt is not included in our training dataset. But our algorithm shows a good generalization performance.

**Animation of Reconstructed Body.** Compared to the existing methods, one of the advantages of using our

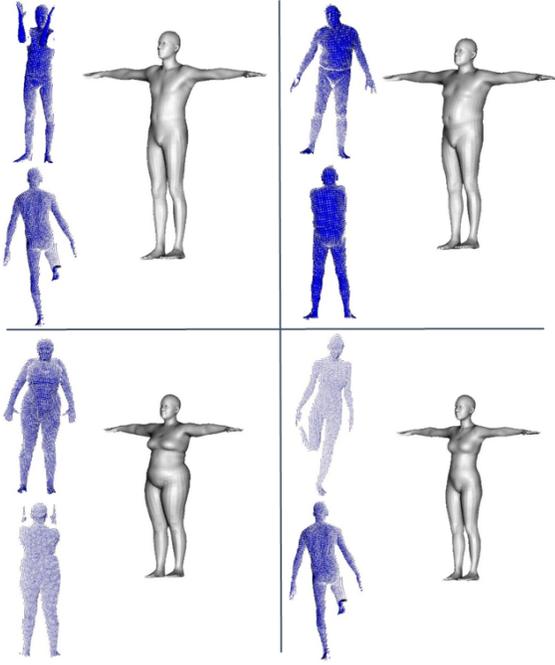


Fig. 4. Results obtained on the FAUST data using our method.

method is to output an animatable body model. Unlike traditional animation methods that have to be configured with skeleton and skinning weights, our reconstructed body model can be animated easily with different pose parameters. As shown in Figure 6, the top is the example of animated results using our method, and the bottom is the dressed results.

#### D. Quantitative Comparison

To quantitatively compare our algorithm with existing state-of-the-art methods, we perform experiments using the BUG dataset [28] as it has the ground-truth body shapes and dressed bodies. We compare our method against various state-of-the-art methods including that of [2] that uses two depth-images as input, the single depth-image based method of [38], and the single RGB image based method of [47]. In contrast to these methods, our approach also works for estimating the body shape under clothing. Thus, we compare our method against Body PointNet [28], which is the state-of-the-art deep learning method taking a complete dressed human body scan as input and outputting the body shape.

**Body Shape Reconstruction.** We first compare our results with related works on the human body shape reconstruction based on the undressed bodies from BUG. For a fair comparison, we apply the same posture with the pose of the input front-facing point cloud to our reconstructed body model. Table II illustrates the comparisons of our method and related works. Figure 7 shows the error maps of the reconstructed shape. It can be seen that the majority of vertices of the reconstructed body using our method is less than 20 millimeters, which outperforms other methods.

**Body Shape Estimation Under Clothing.** Our method can also work for body shape estimation under clothing. Figure 8 and Table III compare our method and the related works for the

TABLE II  
COMPARISON OF RECONSTRUCTION ERROR WITH RELATED WORKS  
(UNIT:  $mm$ ).

input	methods	[2]	[38]	[47]	ours
	$\mu$	5.1	10.4	51.0	1.3
	$\sigma$	6.9	7.4	52.9	4.0
	$max$	27.0	44.9	215.6	19.1

task of estimating body shape under clothing. It is noticeable that the methods of [2], [38] and [47] do not perform well for predicting body shape under clothing. Compared to the results from Body PointNet [28], our results are better. As shown in Table II and Table III, it is important to observe that method [47] is a single RGB image-based method while the other approaches ([2], [38], [28] and ours) are depth-based methods. The experiments reveal that directly reconstructing the 3D human body shape from a single RGB image requires the body height as input due to the scale ambiguity from 2D to 3D. In contrast, depth images offer 3D information, which is exploited by [2], [38], [28] and our proposed method; this proves to us more accurate for 3D human body shape reconstruction compared to the use of a single 2D picture as in [47]. We also note that the training dataset used by [47] lacks RGB images of the human body in tight clothing which explains why [47] performs better when estimating the body shape under clothing than when reconstructing the body shape for tight clothing.

TABLE III  
COMPARISON OF RECONSTRUCTION ERROR FOR BODY UNDER CLOTHING  
WITH RELATED WORKS (UNIT:  $mm$ ).

input	methods	[2]	[38]	[47]	[28]	ours
	$\mu$	8.2	34.7	32.8	9.3	2.2
	$\sigma$	9.0	35.9	28.5	8.5	5.0
	$max$	56.5	233.1	137.7	45.0	18.9

Table IV compares our method with existing state-of-the-art methods in terms of input data type, human body reconstruction, estimation of body shape under clothing, being animatable the speed and being deep learning. The method of [2], which takes two depth images as input, obtains better result than the methods taking a single depth image or a single RGB image as input; this is to be expected as two depth images offer more reliable geometry information. We also take two depth images as input. Compared to [2], our method is about 16 times faster and our result is also better.

#### E. Ablation study

We conduct ablation experiments based on the 450 testing undressed male samples that have no overlap with the data used in the training in order to understand the value of our design and the influence of the different terms in our loss function.

**PointNet or DGCNN.** Firstly, we evaluate which types of features yield better performance in this study: local features or global features? To this end, we used the popular DGCNN [36] and PointNet [34] to learn local and global features

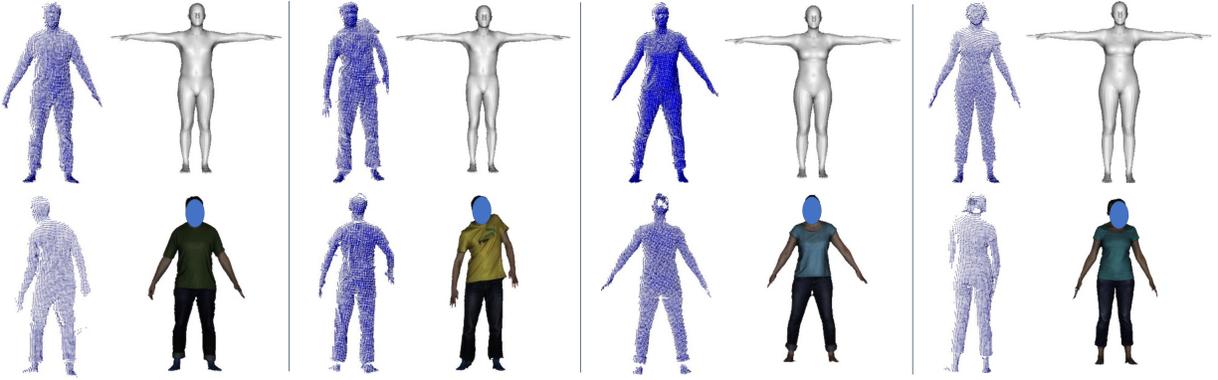


Fig. 5. Results obtained on the BUFF data using our method. Our results are the white meshes. The textured meshes are the BUFF samples.

TABLE IV  
COMPARISON WITH DIFFERENT HUMAN BODY SHAPE RECONSTRUCTION METHODS.

methods	input	human body shape reconstruction	estimation of human body under clothing	being animatable	run speed	being deep learning
[2]	two depth images	✓	×	×	120 seconds	×
[38]	one depth image	✓	×	×	3.7 seconds	×
[47]	one RGB image	✓	×	✓	4 seconds	✓
[28]	one complete 3D scan	×	✓	×	5.8 seconds	✓
ours	two depth images	✓	✓	✓	7.5 seconds	✓

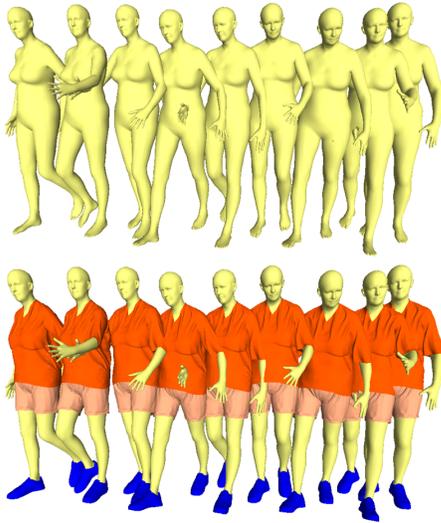


Fig. 6. The example of animated results using our method.

respectively. As shown in Table V, global features outperform local features for this task, so we adopted PointNet as the feature extractor in this work.

TABLE V  
ABLATION STUDY ON THE FEATURES (UNIT:  $mm$ ).

Feature	Local	Global
$\mu$	0.090	0.059
$\sigma$	0.121	0.094
$max$	1.153	0.896

**$\beta$ -based Mean Pooling.** In order to simultaneously capture the information from the two input point clouds, it is necessary to obtain a single accurate prediction. We propose three alternative stitching modules and then compare their performances.

As shown in Figure 9,  $\beta 2\beta$  denotes the stitching module that takes the concatenation of  $\beta_F$  and  $\beta_B$  as input and output the  $\beta$  value;  $f2\beta$  represents the stitching module that regresses the  $\beta$  value from the concatenation of latent features  $f_F$  and  $f_B$ . Table VI gives the comparison for the three methods, it is noticeable that the proposed  $\beta$  mean pooling outperforms the other two methods although they have more complicated architectures.

TABLE VI  
ABLATION STUDY ON THE STITCHING MODULE (UNIT:  $mm$ ).

Stitching Module	$\beta 2\beta$	$f2\beta$	$\beta$ mean pooling
$\mu$	0.070	0.066	0.059
$\sigma$	0.111	0.156	0.094
$max$	1.405	1.587	0.896

**With or without SMPL layer.** The SMPL layer is integrated in our neural network design in the training phase. In the testing phase, the trained networks directly output the SMPL  $\beta$  values. The SMPL layer, thus, does not contribute to the prediction. We attempt to explore directly regressing the SMPL  $\beta$  values by minimizing the loss based on them. Therefore, we remove the SMPL layer and train the networks only based on the  $\beta$  loss. As shown in Table VII, the SMPL layer can significantly reduce the reconstruction error. It guides the deep neural network to search for the optimal SMPL  $\beta$  in the SMPL model space.

TABLE VII  
ABLATION STUDY ON THE SMPL LAYER (UNIT:  $mm$ ).

SMPL layer	With	Without
$\mu$	0.059	0.114
$\sigma$	0.094	0.136
$max$	0.896	0.896

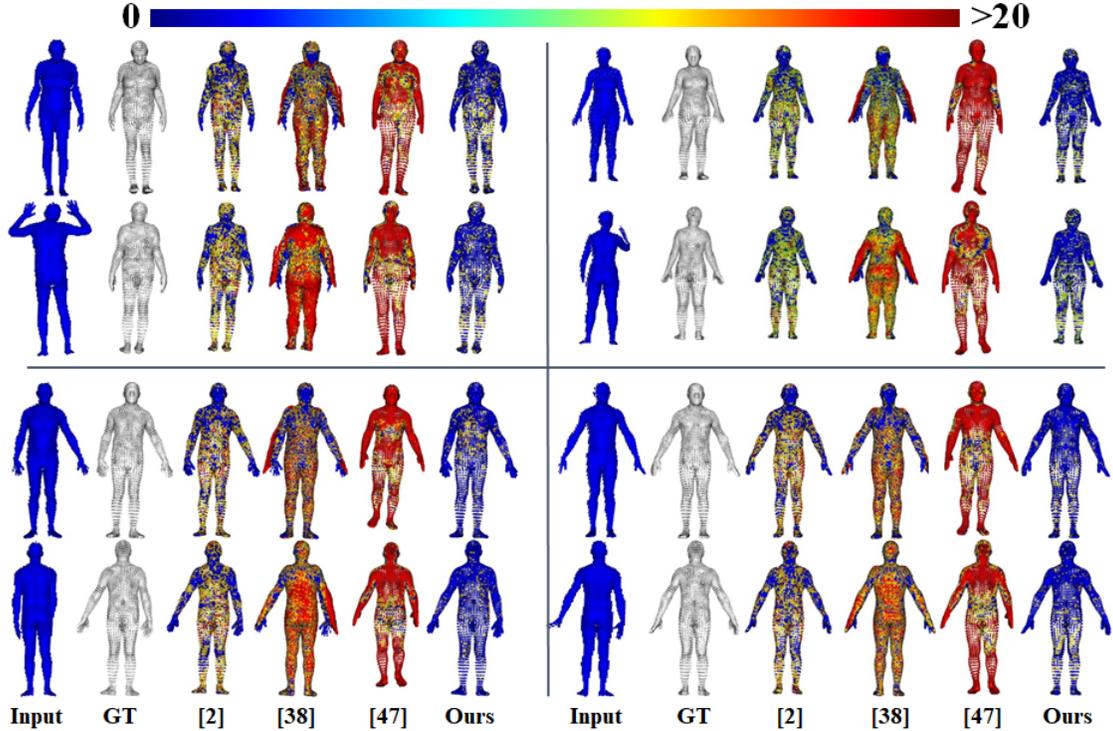


Fig. 7. Comparison of human body reconstruction error with start-of-the-art methods. From left to right: the two input point clouds (blue point sets), the ground truth body (white point set), the result of [2], the result of [38], the result of [47] and our result. For each pair, the top is the front-facing view, and the bottom is the back-facing view. The color of each point is colored by per-vertex error in millimeters.

TABLE VIII  
ABLATION STUDY: LOSS SELECTION (UNIT:  $mm$ ).

Loss	$L_{vert}$	$L_{vert+joint}$	$L_{vert+joint+beta}$	$L_{vert+joint+beta+reg}$
$\mu$	0.109	0.101	0.092	0.059
$\sigma$	0.152	0.142	0.145	0.094
$max$	1.349	1.389	1.40	0.896

TABLE IX  
COMPARISON OF THE RESULTS FROM SINGLE DEPTH IMAGES AND TWO DEPTH IMAGES. (UNIT:  $mm$ ).

Input type	single front-facing depth image	single back-facing depth image	two depth images
$\mu$	0.060	0.085	0.059
$\sigma$	0.124	0.183	0.094
$max$	1.605	1.919	0.896

**Loss Selection.** Our loss mainly consists of four type of terms:  $L_{vert}$ ,  $L_{joint}$ ,  $L_{\beta}$  and  $L_{reg}$ .  $L_{vert}$  and  $L_{joint}$  have definitions for both the front-facing and back-facing data of the body. To validate the contribution of these terms, we compare  $L_{vert} = L_{vert}^F + L_{vert}^B$ ,  $L_{vert+joint} = L_{vert}^F + L_{vert}^B + 0.001 * L_{joint}^F + 0.001 * L_{joint}^B$ ,  $L_{vert+joint+beta} = L_{vert}^F + L_{vert}^B + 0.001 * L_{joint}^F + 0.001 * L_{joint}^B + L_{beta}$  and our complete loss  $L_{vert+joint+beta+reg} = L_{vert}^F + L_{vert}^B + 0.001 * L_{joint}^F + 0.001 * L_{joint}^B + L_{beta} + L_{reg}$ . Table VIII shows the reconstruction comparison. The results show that our full loss obtained the best accuracy for the shape reconstruction.

**Single depth image or two depth images.** The proposed method can be directly applied to single depth image-based body shape reconstruction. In this study, we take a single front-facing depth image, a single back-facing depth image and two depth images as input and compare the performance when using our model. As shown in Table IX, the results obtained by using two depth images is systematically the best while the result obtained a single back-facing depth image is the worst.

## VI. CONCLUSIONS

In this work, we propose a novel learning-based framework for reconstructing the human body shape using a single commodity depth camera. Compared to the existing methods, our method has the following advantages: 1) we only need two depth images from the front-facing and back-facing views of the subject as input; 2) the posture variation of two depth images is large; 3) our method is able to estimate the body shape under clothing; 4) our reconstructed body model is animatable; 5) the proposed method is of low complexity and fast. We also present a novel dataset dubbed FBB consisting of dressed and undressed human body shapes. Extensive experimental results on PDT13, FAUST, BUFF and BUG datasets show that our method outperforms the existing methods. Since it is a parametric method, the main limitation of our approach is that it cannot preserve fine details from the raw partial scans. In future work, we will focus on non-rigidly aligning two partial scans of subjects with large pose variations based on deep learning.

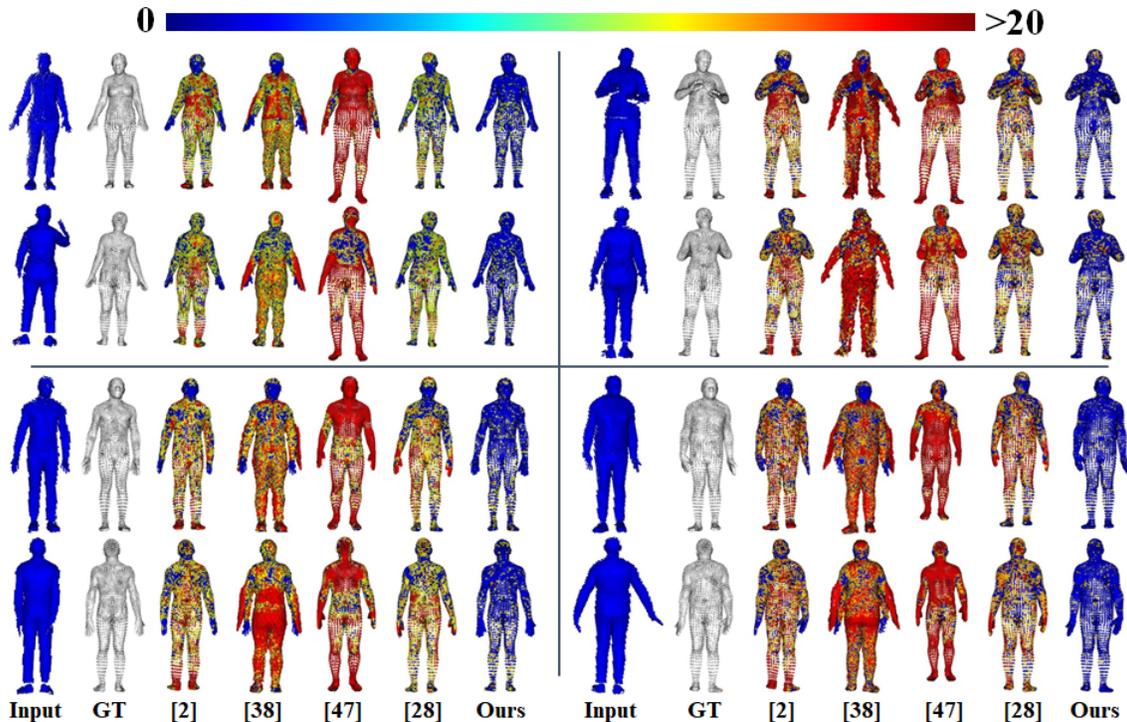


Fig. 8. Comparison of error of estimated body shape under clothing with start-of-the-art methods. From left to right: the two input point clouds (blue point sets), the ground truth body (white point set), the result of [2], the result of [38], the result of [47], the result of [28] and our result. For each pair, the top is the front-facing view, and the bottom is the back-facing view. The color of each point is colored by per-vertex error in millimeters.

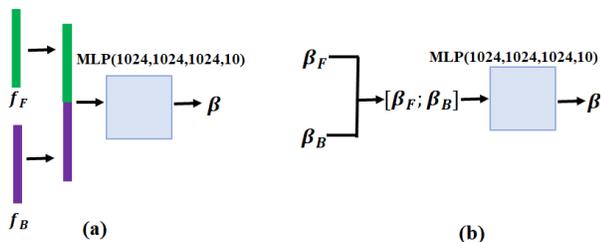


Fig. 9. Our other alternative stitching modules. (a)  $f2\beta$ : regressing the  $\beta$  value from the concatenation of latent features  $f_F$  and  $f_B$ ; (b)  $\beta2\beta$ : regressing the  $\beta$  value from the concatenation of  $\beta_F$  and  $\beta_B$ .

## VII. ACKNOWLEDGMENTS

The authors would like to acknowledge the support of Innoviris (project eTailor), the close collaboration with Treedy's in the framework of this project.

## REFERENCES

- [1] P. Hu, D. Li, G. Wu, T. Komura, D. Zhang, and Y. Zhong, "Personalized 3d mannequin reconstruction based on 3d scanning," *International Journal of Clothing Science and Technology*, 2018.
- [2] T. Zhao, S. Li, K. N. Ngan, and F. Wu, "3-d reconstruction of human body shape from a single commodity depth camera," *IEEE Transactions on Multimedia*, vol. 21, no. 1, pp. 114–123, 2018.
- [3] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski, "A comparison and evaluation of multi-view stereo reconstruction algorithms," in *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, vol. 1. IEEE, 2006, pp. 519–528.
- [4] Y. M. Kim, C. Theobalt, J. Diebel, J. Kosecka, B. Matusik, and S. Thrun, "Multi-view image and tof sensor fusion for dense 3d reconstruction," in *2009 IEEE 12th international conference on computer vision workshops, ICCV workshops*. IEEE, 2009, pp. 1542–1549.
- [5] K. Zhu, R. Wang, Q. Zhao, J. Cheng, and D. Tao, "A cuboid cnn model with an attention mechanism for skeleton-based action recognition," *IEEE Transactions on Multimedia*, 2019.
- [6] G. Zhou, Y. Yan, D. Wang, and Q. Chen, "A novel depth and color feature fusion framework for 6d object pose estimation," *IEEE Transactions on Multimedia*, 2020.
- [7] Y. Lu, T. Stathopoulou, M. F. Vasiloglou, S. Christodoulidis, Z. Stanga, and S. Mouggiakakou, "An artificial intelligence-based system to assess nutrient intake for hospitalised patients," *IEEE transactions on multimedia*, 2020.
- [8] 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*. IEEE, 2011, pp. 127–136.
- [9] 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*, vol. 18, no. 4, pp. 643–650, 2012.
- [10] Z. Xu, W. Chang, Y. Zhu, D. Le, H. Zhou, and Q. Zhang, "Building high-fidelity human body models from user-generated data," *IEEE Transactions on Multimedia*, 2020.
- [11] P. J. Besl and N. D. McKay, "Method for registration of 3-d shapes," in *Sensor fusion IV: control paradigms and data structures*, vol. 1611. International Society for Optics and Photonics, 1992, pp. 586–606.
- [12] H. Chui and A. Rangarajan, "A new point matching algorithm for non-rigid registration," *Computer Vision and Image Understanding*, vol. 89, no. 2-3, pp. 114–141, 2003.
- [13] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis, "Scape: shape completion and animation of people," in *ACM SIGGRAPH 2005 Papers*, 2005, pp. 408–416.
- [14] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "Smpl: A skinned multi-person linear model," *ACM transactions on graphics (TOG)*, vol. 34, no. 6, pp. 1–16, 2015.
- [15] Z. Liu, J. Huang, S. Bu, J. Han, X. Tang, and X. Li, "Template deformation-based 3-d reconstruction of full human body scans from low-cost depth cameras," *IEEE transactions on cybernetics*, vol. 47, no. 3, pp. 695–708, 2016.
- [16] T. Groueix, M. Fisher, V. G. Kim, B. C. Russell, and M. Aubry, "3d-

- coded: 3d correspondences by deep deformation,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 230–246.
- [17] M. Kowalski, J. Naruniec, and M. Daniluk, “Livescan3d: A fast and inexpensive 3d data acquisition system for multiple kinect v2 sensors,” in *2015 international conference on 3D vision*. IEEE, 2015, pp. 318–325.
- [18] A. Fuster-Guilló, J. Azorín-López, M. Saval-Calvo, J. M. Castillo-Zaragoza, N. García-D’Urso, and R. B. Fisher, “Rgb-d-based framework to acquire, visualize and measure the human body for dietetic treatments,” *Sensors*, vol. 20, no. 13, p. 3690, 2020.
- [19] Y. Cui, W. Chang, T. Nöll, and D. Stricker, “Kinectavatar: fully automatic body capture using a single kinect,” in *Asian Conference on Computer Vision*. Springer, 2012, pp. 133–147.
- [20] H. Jiang, J. Cai, and J. Zheng, “Skeleton-aware 3d human shape reconstruction from point clouds,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 5431–5441.
- [21] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik, “End-to-end recovery of human shape and pose,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7122–7131.
- [22] A. Tuan Tran, T. Hassner, I. Masi, and G. Medioni, “Regressing robust and discriminative 3d morphable models with a very deep neural network,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5163–5172.
- [23] F. Wu, L. Bao, Y. Chen, Y. Ling, Y. Song, S. Li, K. N. Ngan, and W. Liu, “Mvf-net: Multi-view 3d face morphable model regression,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 959–968.
- [24] P. Hu, T. Komura, D. Holden, and Y. Zhong, “Scanning and animating characters dressed in multiple-layer garments,” *The Visual Computer*, vol. 33, no. 6–8, pp. 961–969, 2017.
- [25] N. Hasler, C. Stoll, B. Rosenhahn, T. Thormählen, and H.-P. Seidel, “Estimating body shape of dressed humans,” *Computers & Graphics*, vol. 33, no. 3, pp. 211–216, 2009.
- [26] J. Yang, J.-S. Franco, F. Hétyroy-Wheeler, and S. Wuhner, “Estimation of human body shape in motion with wide clothing,” in *European Conference on Computer Vision*. Springer, 2016, pp. 439–454.
- [27] C. Zhang, S. Pujades, M. J. Black, and G. Pons-Moll, “Detailed, accurate, human shape estimation from clothed 3d scan sequences,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4191–4200.
- [28] P. Hu, N. N. Kaashki, V. Dadarlat, and A. Munteanu, “Learning to estimate the body shape under clothing from a single 3d scan,” *IEEE Transactions on Industrial Informatics*, 2020.
- [29] B. L. Bhatnagar, G. Tiwari, C. Theobalt, and G. Pons-Moll, “Multi-garment net: Learning to dress 3d people from images,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 5420–5430.
- [30] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in neural information processing systems*, 2012, pp. 1097–1105.
- [31] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black, “Keep it smpl: Automatic estimation of 3d human pose and shape from a single image,” in *European Conference on Computer Vision*. Springer, 2016, pp. 561–578.
- [32] M. Tatarchenko, A. Dosovitskiy, and T. Brox, “Octree generating networks: Efficient convolutional architectures for high-resolution 3d outputs,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2088–2096.
- [33] H. Su, S. Maji, E. Kalogerakis, and E. Learned-Miller, “Multi-view convolutional neural networks for 3d shape recognition,” in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 945–953.
- [34] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, “Pointnet: Deep learning on point sets for 3d classification and segmentation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 652–660.
- [35] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, “Pointnet++: Deep hierarchical feature learning on point sets in a metric space,” in *Advances in neural information processing systems*, 2017, pp. 5099–5108.
- [36] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, “Dynamic graph cnn for learning on point clouds,” *Acm Transactions On Graphics (tog)*, vol. 38, no. 5, pp. 1–12, 2019.
- [37] W. Yuan, T. Khot, D. Held, C. Mertz, and M. Hebert, “Pcn: Point completion network,” in *2018 International Conference on 3D Vision (3DV)*. IEEE, 2018, pp. 728–737.
- [38] N. Lunscher and J. Zelek, “Deep learning whole body point cloud scans from a single depth map,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 1095–1102.
- [39] F. P.-W. Lo, Y. Sun, J. Qiu, and B. P. Lo, “Point2volume: A vision-based dietary assessment approach using view synthesis,” *IEEE Transactions on Industrial Informatics*, vol. 16, no. 1, pp. 577–586, 2019.
- [40] G. Varol, J. Romero, X. Martin, N. Mahmood, M. J. Black, I. Laptev, and C. Schmid, “Learning from synthetic humans,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 109–117.
- [41] M. Gschwandtner, R. Kwitt, A. Uhl, and W. Pree, “Blensor: Blender sensor simulation toolbox,” in *International Symposium on Visual Computing*. Springer, 2011, pp. 199–208.
- [42] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.
- [43] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, “Tensorflow: A system for large-scale machine learning,” in *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, 2016, pp. 265–283.
- [44] T. Helten, A. Baak, G. Bharaj, M. Müller, H.-P. Seidel, and C. Theobalt, “Personalization and evaluation of a real-time depth-based full body tracker,” in *2013 International Conference on 3D Vision-3DV 2013*. IEEE, 2013, pp. 279–286.
- [45] F. Bogo, J. Romero, M. Loper, and M. J. Black, “Faust: Dataset and evaluation for 3d mesh registration,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 3794–3801.
- [46] L. Kavan, S. Collins, J. Žára, and C. O’Sullivan, “Skinning with dual quaternions,” in *Proceedings of the 2007 symposium on Interactive 3D graphics and games*, 2007, pp. 39–46.
- [47] W. Zeng, W. Ouyang, P. Luo, W. Liu, and X. Wang, “3d human mesh regression with dense correspondence,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7054–7063.



**Pengpeng Hu** is researcher at the Electronics and Informatics (ETRO) department of the Vrije Universiteit Brussel (VUB), Belgium. He received a Doctorate in Digital Textile Engineering from Donghua University, China, in 2017. He was a visiting scholar at the School of Informatics of the Edinburgh University, the UK in 2016. He was a postdoctoral fellow at the Computer and Information Sciences Department of the Northumbria University, the UK in 2017. Since 2018, he started working at VUB. He is the outstanding paper winner of the Emerald

Literati Award 2019. His research interests include geometric deep learning, 3D human body reconstruction, RGB-D image, and digital fabrication. He serves as guest editor for MDPI SENSORS. He served as the Technical Support Chair of BMVC 2018, and the member of Program Committee in SKIMA 2017, SKIMA 2018, and SKIMA 2019.



**Edmond Shu-lim Ho** is currently the Programme Leader for BSc (Hons) Computer Science and a Senior Lecturer in the Department of Computer and Information Sciences at Northumbria University, Newcastle, UK. Prior to joining Northumbria University in 2016, he was a Research Assistant Professor in the Department of Computer Science at Hong Kong Baptist University. He received the BSc degree in Computer Science from the Hong Kong Baptist University, the MPhil degree from the City University of Hong Kong, and the PhD degree from

the University of Edinburgh. His research interests include Computer Graphics, Computer Vision, Robotics, Motion Analysis, and Machine Learning.



**Adrian Munteanu** is professor at the Electronics and Informatics (ETRO) department of the Vrije Universiteit Brussel (VUB), Belgium. He received the MSc degree in Electronics and Telecommunications from Politehnica University of Bucharest, Romania, in 1994, the MSc degree in Biomedical Engineering from University of Patras, Greece, in 1996, and the Doctorate degree in Applied Sciences from Vrije Universiteit Brussel, Belgium, in 2003. In the period 2004-2010 he was post-doctoral fellow with the Fund for Scientific Research – Flanders (FWO), Belgium, and since 2007, he is professor at VUB. Adrian Munteanu contributed to more than 350 publications and holds 7 patents. He is the recipient of the 2004 BARCO-FWO prize for his PhD work, the (co-)recipient of the Most Cited Paper Award from Elsevier for 2007. Adrian Munteanu served as Associate Editor for IEEE Transactions on Multimedia and currently serves as Associate Editor for IEEE Transactions on Image Processing.