Single RGB-D Fitting: Total Human Modeling with an RGB-D Shot

Xianyong Fang*
Anhui University
Hefei, China
University of Houston
Houston, TX
fangxianyong@ahu.edu.cn

Jikui Yang
Anhui University
Hefei, China
e16301112@stu.ahu.edu.cn

Jie Rao
Anhui University
Hefei, China
e18301094@stu.ahu.edu.cn

Linbo Wang†
Anhui University
Hefei, China
wanglb@ahu.edu.cn

Zhigang Deng
University of Houston
Houston, TX
zdeng4@uh.edu

ABSTRACT

Existing single shot based human modeling methods generally cannot model the complete pose details (e.g., head and hand positions) without non-trivial interactions. We explore the merits of both RGB and depth images and propose a new method called *Single RGB-D Fitting* (SRDF) to generate a realistic 3D human model with a single RGB-D shot from a consumer-grade depth camera. Specifically, the state-of-the-art deep learning techniques for RGB images are incorporated into SRDF, so that: 1) A compound skeleton detection method is introduced to obtain accurate 3D skeletons with refined hands based on the combination of depth and RGB images; and 2) an RGB image segmentation assisted point cloud pre-processing method is presented to obtain smooth foreground point clouds. In addition, several novel constraints are also introduced into the energy minimization model, including the shape continuity constraint, the keypoint-guided head pose prior constraint, and the penalty-enforced point cloud prior constraint. The energy model is optimized in a two-pass way so that a realistic shape can be estimated from coarse to fine. Through extensive experiments and comparisons with the state of the art methods, we demonstrate the effectiveness and efficiency of the proposed method.

CCS CONCEPTS

• **Computing methodologies** → **Virtual reality**; **Reconstruction**.

*Corresponding author
†Corresponding author

VRST '19, November 12–15, 2019, Parramatta, NSW, Australia

ACM ISBN 978-1-4503-7001-1/19/11…$15.00
https://doi.org/10.1145/3359996.3364252

KEYWORDS

Shape modeling, single RGB-D image, 3D reconstruction, depth camera, deep learning

ACM Reference Format:
Xianyong Fang, Jikui Yang, Jie Rao, Linbo Wang, and Zhigang Deng. 2019. Single RGB-D Fitting: Total Human Modeling with an RGB-D Shot. In *25th ACM Symposium on Virtual Reality Software and Technology (VRST '19), November 12–15, 2019, Parramatta, NSW, Australia.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3359996.3364252

1 INTRODUCTION

Recovering the 3D human shape of a loosely clothed human from limited input data has been a long-standing challenge in computer graphics and vision communities. It can be potentially applied for a variety of applications, including virtual try-on, body measurements, and virtual reality. Despite many existing reconstruction efforts that are based on the input of video, multi-view images, or RGB-D images, limited efforts have been attempted to reconstruct the high-quality 3D human models from a *single* shot (e.g., a single RGB-D shot).

Existing single shot based methods may use either one RGB image [Bogo et al. 2016; Chen et al. 2013; Dibra et al. 2016; Guan et al. 2009; Lassner et al. 2017b], one depth image [Oyama et al. 2017; Yi et al. 2015], or one RGB-D image [Li et al. 2019]. Focusing on the body modeling, they often fall short of obtaining the full 3D models with rich details, without manual interventions. Some previous studies [Dibra et al. 2017; Jackson et al. 2018; Omran et al. 2018; Pavlakos et al. 2018; Varol et al. 2018] explored an end-to-end deep modeling scheme with one image, assuming that deep learning techniques can generally obtain better performances than manually-selected features based methods. However, they also cannot model fine details such as hand and head poses. In addition, the performance of such methods largely depends on the training data, besides the required time-consuming training process. Their performances can also be degraded if test images are significantly different from the training ones. Indeed, there is still a clear need for new robust methods to create *total* 3D models for clothed humans. Borrowing the "total" term from [Joo et al. 2018; Xiang et al. 2019],

in this paper we use *Total Human Modeling* (THM) to refer to the simultaneous 3D modeling of the human body, head, and hands.

We are interested in a single RGB-D shot because the depth image encodes the spatially rich 3D information via point clouds and complements with 2D RGB images. Utilizing both the sources has the great potential to improve the reconstruction performance. Fortunately, they can be captured simultaneously by off-the-shelf consumer-grade products such as Microsoft Kinect with one shot.

Therefore, in this paper we propose a new method, called *Single RGB-D Fitting* (SRDF), for a robust and accurate THM of a clothed human based on a single RGB-D shot, by exploiting the information enclosed in both the RGB image and the depth image in the shot. Our method is based on the concept of a standard template of human body widely used in parametric models [Anguelov et al. 2005; Loper et al. 2015; Zhang et al. 2017].

Specifically, deep learning techniques are incorporated into our method in the following ways.

(1) We introduce a compound skeleton detection method to obtain a complete, accurate 3D skeleton from a single RGB-D shot. It adopts an RGB image based deep learning method, LCR-Net [Rogez et al. 2017], to correct the incorrect keypoints. Also, it employs the deep hand pose estimation method [Zimmermann and Brox 2017] for hand skeleton extraction.
(2) A deep RGB image segmentation assisted point cloud preprocessing is proposed to obtain a smooth cloud for a better detail recovery. The deep image segmentation method, Conditional Random Fields (CRFs) and Recurrent Neural Networks (RNNs) combined segmentation method (CRF-RNN) [Zheng et al. 2015], facilitate to obtain a robust hole-free foreground cloud and thus improve the denoising performance.

In our approach, we also introduce a new energy minimization model with three new constraints to ensure an effective recovery, besides the general constraints used in existing methods.

(1) Minor shape variations during the iteration process may make the model converge to incorrect results. Therefore, we propose a *shape continuity constraint* on the neighboring shapes during iterations so that the shapes can be updated progressively without drastic changes.
(2) The topmost joint in the SMPL model [Loper et al. 2015] is the neck joint, *i.e.*, no joints exists in the head. This structure makes the capture of the head pose impossible in the SMPL model. Therefore, we propose a *key point guided head pose prior constraint* to obtain the correct head pose.
(3) A recovered model point should always stay inside the clothes during the optimization process, otherwise incorrect models can be easily generated. Therefore, we propose a *penalty enforced point cloud prior constraint* so that a point would receive a higher penalty if it is outside the cloud.

Also, our energy model is optimized via two passes in a coarse to fine way: the first pass aims for coarse pose and shape estimation; and the second one is designed for details handling, including fine shape and head pose modeling.

2 RELATED WORK

Recovering the geometric appearances of humans under loose clothes has been widely studied. Some of existing approaches [De Aguiar et al. 2008; Venkat et al. 2018; Xu et al. 2018] model the clothed surface for performance capture. We however specifically target 3D human shape modeling and thus hereby will mainly focus on efforts related to it.

A standard body template is an important reference for human modeling [Cheng et al. 2018]. Recently, parametric templates based on the statistical analysis of 3D training data received noticeable successes [Black et al. 2016; Zhang et al. 2017]. Perhaps the most popular models are the Shape Completion and Animation for PEople model (SCAPE) [Anguelov et al. 2005; Pishchulin et al. 2017] and the Skinned Multi-Person Linear model (SMPL) [Loper et al. 2015]. SMPL is more flexible and thus popular [Bogo et al. 2016; Joo et al. 2018; Lassner et al. 2017a; Li et al. 2019; Omran et al. 2018; Romero et al. 2017] than SCAPE for GPU computation because the former's deformation of pose and shape can be linearly applied onto vertices while the latter cannot be used. We also consider SMPL as the reference template. In particular, recently Romero *et al.* [Romero et al. 2017] extended SMPL by integrating a learned hand Model with Articulated and Non-rigid defOrmations (MANO). Termed as SMPL+H, it can obtain models with detailed hand motions. Therefore, we take SMPL+H as the standard human model.

Multiple shots in RGB images and video sequences have been utilized for 3D shape modeling [Alldieck et al. 2017; Stoll et al. 2010]. Manual interactions to correct the errors in shapes and poses may also be required in such methods [Rogge et al. 2014]. Recently, robust methods [Xiang et al. 2019] for face and hands included THM are proposed with deep learning training and a new 3D body representation. Multiple depth sequences have also been exploited. Most of existing methods require special configurations, *e.g.*, initializing the pose [Neophytou and Hilton 2014] and model [Ye and Yang 2014], dressing skintight clothes [Bogo et al. 2015], or using special photographing poses [Alldieck et al. 2018]. Joo *et al.* [Joo et al. 2018] fulfilled the *total capture* of the 3D shape with both face and hands.

Multi-shots based methods may be inconvenient for certain practical applications. Users generally need to be very careful in preparing consecutive captures to ensure enough overlaps or relatively smaller non-rigid deformations between neighboring views. On the contrary, a single shot based method would be much easier and more robust to use even for novice users.

Some existing methods use a single RGB image to estimate the human model. 3D skeleton often acts as an important cue [Bogo et al. 2016; Guan et al. 2009; Lassner et al. 2017b]. For example, Guan *et al.* [Guan et al. 2009] estimated 3D joints through manual labeling, while Bogo *et al.* [Bogo et al. 2016] fit a 3D skeleton through 2D pose estimation. Lassner *et al.* [Lassner et al. 2017b] additionally utilized the silhouette constraint. The object silhouette prior constraint, which can be accurately obtained through manual labeling, has also been used by researchers [Chen et al. 2013; Dibra et al. 2016; Guan et al. 2009].

A single depth image has also been utilized for shape modeling because depth can provide 3D information and naturally fits with the task of 3D modeling. Yi *et al.* [Yi et al. 2015] first registered

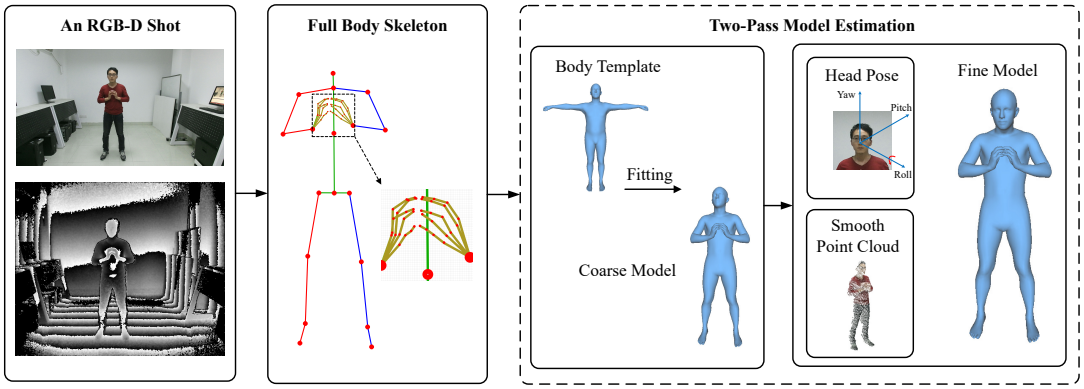


Figure 1: The pipeline of our SRDF approach. Note: (1) Red circles in the skeleton represent the keypoints; (2) different type of bones that connect two neighboring keypoints are shown in different colors for clarity; and (3) the hand skeletons are zoomed in for a detailed view (see Section 4.2 for more details).

joints in a standard human body with the corresponding joints observed in a consumer-grade depth camera and then optimized the human model by propagating the observation. Oyama *et al.* [Oyama et al. 2017] proposed a two-stages model fitting scheme, where the first stage is to recover the coarse surface geometry through the combination of skeleton fitting and Laplacian surface editing, and the second stage recovers fine details by fitting with a fine stitched puppet model.

Recently, Li *et al.* [Li et al. 2019] presented a single RGB-D shot based work. It directly takes both 2D CNN based keypoints and 3D Kinect Keypoints into the energy model without considering refining the 3D keypoints as ours. It also uses the traditional SMPL model while optimizing with traditional shape and pose constraints. They did not consider the fine modelings of the hand and head orientation.

Existing single shot based approaches mainly focus on the body modeling without considering the fine details of the head and hand poses and therefore may not correctly capture the 3D shape as THM expected. Some existing studies did not consider the data source but focus on full 3D scans [Hasler et al. 2009; Pons-Moll et al. 2017]. However, obtaining an accurate 3D full human cloud is a non-trivial task and may require special and expensive scanning devices. Therefore, they may not be applicable for ordinary users. Zuffi *et al.* [Zuffi et al. 2018] proposed an interesting method to model animal shapes based on multiple animal images. Animals are articulated and deformable like humans. However, their furry appearances and physical properties are usually quite different from clothed humans and thus pose different technical challenges.

Recently deep learning techniques [LeCun et al. 2015] have been applied for single image based end-to-end estimation [Bogo et al. 2016; Kanazawa et al. 2018; Omran et al. 2018; Pavlakos et al. 2018; Varol et al. 2018]. Multiple views have also been used to train an end-to-end process [Dibra et al. 2017; Ji et al. 2018]. While most of the above methods rely on a shape model, such as SCAPE or SMPL, Jackson *et al.* [Jackson et al. 2018] use a Volumetric Regression Network (VRN) for a template-free deep reconstruction. The training process of deep learning based methods can be tedious, and its performances highly depends on the similarities between test images and the training data. In addition, they do not consider certain details such as the fine details of the head or hand poses.

3 OVERVIEW OF OUR SRDF APPROACH

Our SRDF approach aims at the THM of humans under loose clothes with a single RGB-D shot from a consumer-level depth camera. We take the Microsoft Kinect as the experimental camera, since its accompanying SDK can conveniently obtain 3D skeletons and head poses as the references for the compound skeleton estimation and the head pose constrained shape modeling. It is noteworthy that, other types of depth cameras and other independent algorithms [Marchand et al. 2016; Murphy-Chutorian and Trivedi 2009; Zhu and Ramanan 2012] can also be used to obtain the poses.

Our method works as follows (refer to Figure 1 for its pipeline). First, a single RGB-D frame of a clothed human subject, captured through one-shot, is used as the input to the system. Then, a compound skeleton detection method is employed so that a fine full body skeleton with key points of the hands are obtained. Subsequently, the resulting 3D human model is computed through a two-passes optimization: The first coarse pass is used to obtain a 3D model by fitting with the general pose and shape of the subject, while the second refinement pass is designed to obtain rich shape and pose details, including the head pose.

The SMPL+H [Romero et al. 2017] is taken as the reference template, which extends SMPL by integrating a learned hand model, MANO, that can capture more detailed hand motions with 14 joints on each hand. Consequently, SMPL+H includes 6890 vertices and 50 keypoints represented joints (22 for body and 28 for hands by excluding the shared keypoint between each hand and its arm). Each hand associates with a 6-elements pose obtained by PCA. To this end, a 78-elements pose $\boldsymbol{\theta}$ is used, including the 72 pose parameters of the original SMPL, *i.e.*, a three-parameters local rotation for each keypoint, and a three-parameters global rotation for the whole model. Additionally, there is a 10-elements shape $\boldsymbol{\beta}$ for the model.

SMPL+H cancels the global translation and thus it is difficult to match the template strictly with the captured 3D data. Therefore, if we denote the SMPL and its blend weights as T and W, respectively, and assume the global translation is $\boldsymbol{t}$, we take the following

formulation of the SMPL+H:

$$M(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{t}; T, W) = \mathcal{W}(T(\boldsymbol{\beta}, \boldsymbol{\theta}), J(\boldsymbol{\beta}), \boldsymbol{\theta}, W) + \boldsymbol{t}, \quad (1)$$

where $\mathcal{W}$ is a linear skinning function generating the new shape and $J = \{j_1, j_2, \ldots, j_N\}$ represents the N joints corresponding to $\boldsymbol{\beta}$ under the standard pose. The joint positions under the current pose $\boldsymbol{\theta}$ are,

$$J_{\boldsymbol{\beta}, \boldsymbol{\theta}} = \mathcal{G}(J(\boldsymbol{\beta}), \boldsymbol{\theta}), \quad (2)$$

where $\mathcal{G}$ represents the transformation functions for all the joints under the Rodrigues's formula [Koks 2006] according to $\boldsymbol{\theta}$.

In the following, a compound method is first introduced to estimate an accurate 3D skeleton (Section 4). Then, an energy model is proposed to robustly recover the human model (Equation 1) and the joint positions (Equation 2) in Section 5. The details of the proposed novel constraints and deep segmentation assisted cloud pre-processing are also described in Section 5.

4 COMPOUND SKELETON ESTIMATION

Existing methods to obtain accurate 3D skeletons from point clouds, including those used by the Kinect SDK, generally only estimate keypoints in the main body without detailed hand joints. For example, the Kinect SDK provides 25 keypoints, but does not have enough keypoints on the hands. Even, they cannot robustly detect all assumed joints due to occlusions-by-clothes, self-occlusions, or interference by other body parts, especially for the joints away from the torso. Therefore, we propose a compound method to integrate two RGB-image based deep methods to recover more accurate and complete full body skeletons.

In our skeleton structure, the keypoints of each hand are replaced with our proposed 19 hand keypoints (see Section 4.2) and the head joint is also removed because of no correspondence in the SMPL+H pose. Therefore, only 18 keypoints from the Kinect SDK are kept in our pose (Figure 3(a))), besides the two 19-keypoints hands.

In the following, we will describe how to refine the main skeleton (including torso, arms, and neck) and the hand joints subsequently.

4.1 Refinement of the Main Skeleton

We adopt one of the recent deep methods, called LCR-Net [Rogez et al. 2017], to predict the 3D main skeleton. It can robustly detect joints using a single image in an end-to-end way even when the body is partially occluded and truncated, while other recent robust methods may only aims for videos [Mehta et al. 2017] and thus do not fit for our needs. LCR-Net adopts a pose proposal network to obtain a list of pose proposals, each of which is scored by a classifier and regressed independently to obtain the final pose. However, it can only predict 13 keypoints (Figure 3(b)), which is very limited for realistic body modeling. The Kinect SDK obtains more joints in the neck, hip, and feet than LCR-Net. Apparently, the Kinect SDK is more flexible than LCR-Net for our purpose. In addition, compared to 2D RGB-image based skeleton extraction [Cao et al. 2018], the main skeleton predicted from point clouds can be more robust due to its embedded 3D spatial information.

Therefore, our method takes the pose estimated by Kinect SDK as the basis whose incorrect keypoints are then corrected with the keypoints computed from LCR-Net. Here we assume the complete body needs to appear in the single shot to obtain the complete skeleton. The update is done by comparing the local node structures of the keypoints and those of their counterparts in LCR-Net. The spatial relationship between the keypoints and their parents from LCR-Net acts as the reference to update the location of the incorrect keypoints from the Kinect SDK.

Figure 2 shows the above update process. For the incorrect keypoint j'_b, the local spatial relationship between the corresponding j_b and its parent keypoint j_a in LCR-Net can be depicted as a vector (called the peer vector), $\overrightarrow{j_a j_b}$, which is expected to be similar to $\overrightarrow{j'_a j'_b}$. That is, the orientations of $\overrightarrow{j_a j_b}$ and $\overrightarrow{j'_a j'_b}$ are expected to be similar if not the same. Consequently, the true keypoint position of j'_b, j^*_b, is updated according to the orientation of $\overrightarrow{j_a j_b}$.

This update process can be formulated as follows. Denote $\boldsymbol{u} = (u_x, u_y, u_z)^T$ as a unit vector pointing to the axis z and K is the following skew symmetric matrix:

$$K = \begin{bmatrix} 0 & -u_z & u_y \\ u_z & 0 & -u_x \\ -u_y & u_x & 0 \end{bmatrix}. \quad (3)$$

We also assume α is the angle between the peer vector $\overrightarrow{j_a j_b}$ and $\boldsymbol{u}$. The orientation of the peer vector can be defined in the Rodrigues's form,

$$R = I \cos\alpha + (1 - \cos\alpha)\boldsymbol{u}^T\boldsymbol{u} + \sin\alpha K, \quad (4)$$

where I is the identity matrix.

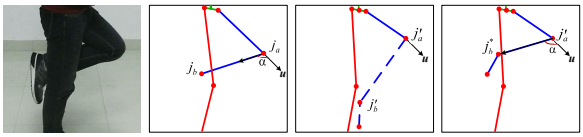


Figure 2: Illustration of the main skeleton refinement process. Note that, j'_b connected by dashed line in (c) represents the incorrect joint whose correct position j^*_b in (d) is estimated by our method.

Then, the correct position j^*_b of the incorrect keypoint j'_b can be computed by rotating $\boldsymbol{u}$ by R and then scaling it to the same length as the peer vector $\overrightarrow{j_a j_b}$,

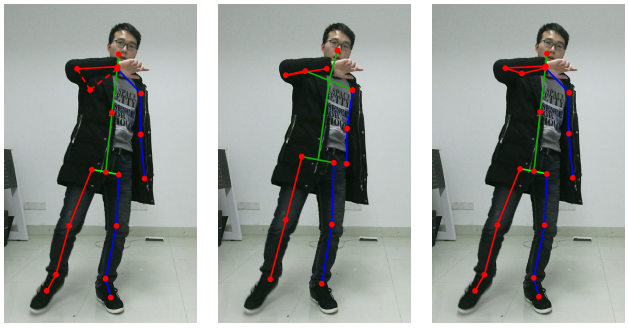
$$j^*_b = j'_a + \|\overrightarrow{j'_a j'_b}\|_2 R\boldsymbol{u}. \quad (5)$$

Figure 3 shows an example of the main skeleton refinement, where the incorrectly estimated keypoints by Kinect SDK (Figure 3(a)) are corrected by LCR-Net (Figure 3(b)) for the final correct pose (Figure 3(c)).

4.2 Refinement of the Hand Skeleton

Accurate hand tracking is important to a realistic reconstruction. Existing 3D point cloud based pose detectors such as Kinect SDK often cannot provide enough hand details. Therefore, we propose a deep RGB image based method to obtain detailed hand poses.

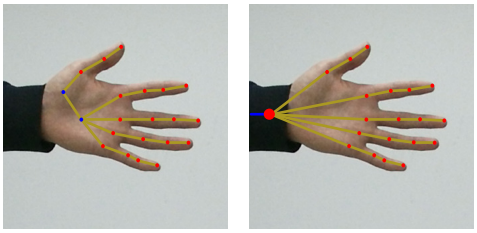
Our work is inspired by the recent work of [Zimmermann and Brox 2017] that can automatically detect an accurate 21-keypoints hand pose from a single RGB image. In the work of [Zimmermann and Brox 2017], three deep networks are introduced to segment the



(a) (b) (c)

Figure 3: An example of the main skeleton refinement. (a): Kinect SDK; (b) LCR-Net; and (c) the refined result. Note: Dashed skeleton lines represent error estimations.

hand, localize the keypoints, and derive the skeleton, respectively. Fine hand details can be obtained from the captured RGB image by this method. Consequently, we will obtain 19 keypoints for each hand based on the hand skeleton in SMPL+H. Figure 4 shows an example of an obtained hand skeleton.



(a) (b)

Figure 4: An example of the hand skeleton refinement. (a): The 21-keypoints hand pose in the work of [Zimmermann and Brox 2017]; and (b): the refined 19-keypoints hand pose by our method. Note that the two blue key points in (a) are removed in (b).

Our experiments show that this deep learning based approach are generally robust to capture the hand poses. Some extremely challenging hand gesture, for example, when the figures seriously curved and stuck, cannot be accurately estimated. In this case, simple drag-and-zooming style of manual interactions are additionally applied to obtain an accurate pose.

5 ENERGY MINIMIZATION MODEL IN SRDF

We now describe the energy minimization model in our SRDF approach. The aim of our work is to reconstruct an accurate human model from a single RGB-D shot. Therefore, besides typical constraints, additional constraints are required to safeguard the target, including shape penalty, body pose, and point cloud priors.

First, drastic shape variations during the optimization process can lead to significant model deformations and thus affect the optimization result. Therefore, shape continuity is also considered in our approach as a constraint. Second, head pose is important to show the true face orientation, however, there is only a key point on the neck to show the head position in the SMPL+H model. Consequently, the SMPL+H model cannot directly show the head pose. Therefore, we introduce the head pose estimated from the cloud as a prior constraint. In addition, the smoothness of the foreground point cloud as a 3D measurement of the subject can also be important for detailed recovery, especially for a single RGB-D shot that only has a single 3D view of the subject. Therefore, we also introduce a new foreground cloud pre-processing step to obtain a smooth foreground. We further present a novel point cloud prior constraint such that the points outside of the cloud are eliminated.

To this end, the following energy minimization model can be obtained:

$$E(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{t}) = w_{data}E_{data} + w_pE_p + w_{sp}E_{sp} + w_{sc}E_s + w_hE_h + w_cE_c, \tag{6}$$

where: 1) E_{data} is the data term; 2) E_p, E_h and E_c are the constraints for body pose, head pose, and point cloud priors, respectively; 3) E_{sp} and E_{sc} are the shape penalty and shape continuity constraints, respectively; and 4) w_ψ, $\psi \in \{data, p, h, c, sp, sc\}$, are their corresponding weights.

5.1 Data Term

The data term aims to minimize the joint distance between the estimated model and the SMPL+H template. Let $\boldsymbol{j}_{k,i}$ be the estimated i-th joint position and $\boldsymbol{j}_{m,i}$ be the corresponding position in the SMPL+H template.

$$E_{data} = \sum_i \lambda_i \rho(\|\boldsymbol{j}_{k,i} - \boldsymbol{j}_{m,i}\|_2^2), \tag{7}$$

where $\rho(e) = \frac{e^2}{\sigma^2+e^2}$ is the Geman-McClure penalty function with bandwidth σ [Pons-Moll et al. 2017] and λ_i represents the weight for the joint i. The weights can be different for different key points considering their different estimation robustness. For example, generally the key points on the torso can be more robustly estimated than those on the limbs; therefore, their associated weights are larger than those on the limbs.

5.2 Body Pose Prior Constraint

The pose prior is adopted from the work of [Zhang et al. 2017], which aims to penalize unnatural pose estimation. It can be formulated as a Gaussian prior [Loper et al. 2015] by the Mahalanobis distance:

$$E_p = (\boldsymbol{\theta} - \mu_{\boldsymbol{\theta}}) \sum\nolimits_{\boldsymbol{\theta}}^{-1} (\boldsymbol{\theta} - \mu_{\boldsymbol{\theta}}), \tag{8}$$

where $\mu_{\boldsymbol{\theta}}$ is the mean and $\sum_{\boldsymbol{\theta}}$ is the covariance matrix, both computed from a pose training set.

5.3 Pose Penalty Constraint

The interpenetration between different human parts can affect the accuracy of pose estimation and thus lead to a wrong model and high computational load. We adopt the idea of capsule [Bogo et al. 2016] to solve the collision problem efficiently.

A capsule has a radius and an axis length. It can be simplified as a sphere centered at $\pi(\boldsymbol{\theta}, \boldsymbol{\beta})$ with a radius $r(\boldsymbol{\beta})$ for the computational convenience. An isotropic Gaussian is defined for each sphere with

$\sigma(\boldsymbol{\beta}) = \frac{\gamma(\boldsymbol{\beta})}{3}$ and, consequently, the penalty of the incompatible parts can be defined as the integral of the product of Gaussians,

$$E_{sp} = \sum_i \sum_{j \in \mathcal{I}(i)} \exp\left(\frac{\|(\pi_i(\boldsymbol{\theta}, \boldsymbol{\beta})) - \pi_j(\boldsymbol{\theta}, \boldsymbol{\beta})\|}{\sigma_i^2(\boldsymbol{\beta}) + \sigma_j^2(\boldsymbol{\beta})}\right), \tag{9}$$

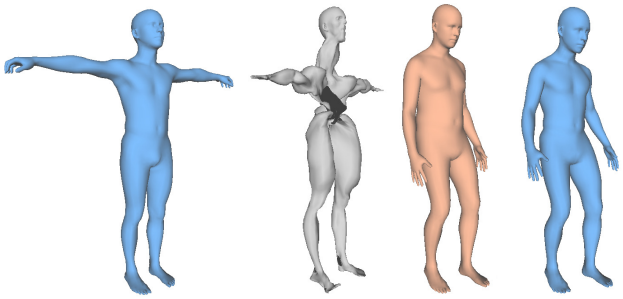
where $\mathcal{I}(i)$, $\pi_i(\boldsymbol{\theta}, \boldsymbol{\beta})$, and $\sigma_i^2(\boldsymbol{\beta})$ are the set of incompatible spheres, the center and variance of the sphere i, respectively.

5.4 Shape Continuity Constraint

The shape continuity constraint aims to avoid large variations of the model during the fitting due to local minimization. Such variations may generate a deformed shape that is quite different from the real shape (Figure 5(b)). Existing shape constraints (*e. g.*, [Bogo et al. 2016]) focus on the continuity of the current shape,

$$E_{sc} = \|\boldsymbol{\beta}\|_2^2, \tag{10}$$

which can easily lead to a distorted figure (Figure 5(c)), compared with the ground truth (Figure 5(a)).



(a) (b) (c) (d)

Figure 5: Demonstration of the shape continuity constraint. (a): The ground truth shape in T pose; (b): the estimated shape without the constraint; (c) the estimated shape using the traditional constraint (Equation 10); and (d) the estimated shape using our proposed constraint.

The obtained shape can be close to the ground truth in a few iterations according to our experiments. Therefore, iterative shape fitting can be adopted to avoid the possibly large shape variations due to local minimization. Based on this idea, we take the progressive similarity of shapes between the consecutive iterations as a continuity constraint:

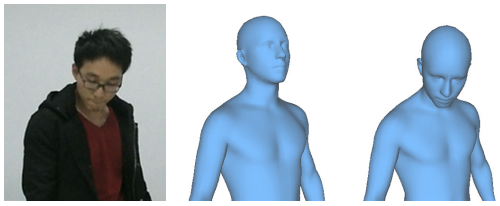
$$E_{sc} = \|\boldsymbol{\beta} - \boldsymbol{\beta}'\|_2^2, \tag{11}$$

where $\boldsymbol{\beta}'$ is the shape computed in the previous iteration. Figure 5(d) shows that the proposed shape continuity constraint can facilitate to recover the correct shape effectively.

5.5 Key-point Guided Head Pose Prior Constraint

Head pose is important for the accurate modeling of head. However, the SMPL+H model only contains the neck joint to represent the head position. Therefore, the obtained head pose is inaccurate and consequently the head modeling can be incorrect (Figure 6(b)).

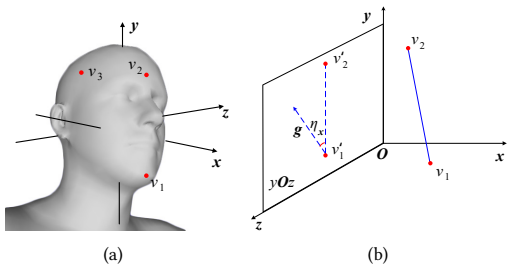
Head pose estimation of a human point cloud has been well studied in the literature [Guan et al. 2009; Shotton et al. 2011]. In our work, we directly use the pose information provided by the Kinect SDK. Specifying several key points on the head template for the pose estimation of the model, we consequently propose a keypoint guided head pose prior constraint.



(a) (b) (c)

Figure 6: Demonstration of the keypoint guided head pose prior constraint. (a): The source shot; (b): the estimated shape without the constraint; and (c) the estimated shape with the constraint.

In our method, the head pose is represented by three rotation angles around the three axes: roll, pitch, and yaw, respectively, as $\boldsymbol{\eta}_h = (\eta_x^h, \eta_y^h, \eta_z^h)^T$. Then, three key points (v_1, v_2 and v_3) on the head template can be manually specified so that three vectors can be generated to compute the head pose (Figure 7(a)).



(a) (b)

Figure 7: Principle of the head pose estimation. (a): The three manually-specified key points; and (b) the computation of η_x^h.

Each rotation angle around an axis is defined to be the angle between the two vectors on the coordinate plane perpendicular to the axis: one vector is the projection of the line connecting two specified points and the other is the projection of any vector in the initial head template. Figure 7(b) shows η_x^h is computed with the projection of the vector $\overrightarrow{v_1v_2}$, $\overrightarrow{v_1'v_2'}$, to the coordinate plane yOz as

$$\eta_x = \arccos \frac{\overrightarrow{v_1'v_2'} \cdot \boldsymbol{g}}{\|\overrightarrow{v_1'v_2'}\|\|\boldsymbol{g}\|}, \tag{12}$$

where $\boldsymbol{g}$ is the projection of any vector in the initial head template to yOz.

The other two angles, η_y and η_z, can be computed similarly. To this end, the head pose prior constraint can be formulated as:

$$E_h = \|\boldsymbol{\eta}_h - \boldsymbol{\phi}_h\|_2^2, \tag{13}$$

where $\boldsymbol{\phi}_h$ is the head pose estimated from the point cloud. Figure 6(c) demonstrates the usefulness of this constraint.

5.6 Point Cloud Prior Constraint

The foreground point cloud need to be extracted and pre-processed (e.g., holes-free) before imposing various prior constraints. Therefore, we propose a deep segmentation assisted foreground cloud pre-processing method. Then, we further introduce a penalty oriented cloud prior constraint.

5.6.1 Deep segmentation assisted foreground cloud pre-processing. We adopt one of the latest deep segmentation methods, CRF-RNN [Zheng et al. 2015], to obtain the segmentation of the input RGB image, which guides to eliminate the missing depths. Assuming the segmented foreground by CRF-RNN is $\mathbb{S}_f$, the missing area, $\mathbb{M}$, can be defined as

$$\mathbb{M} = \mathbb{S}_f \ (\mathcal{B}(\mathbb{D}) \cap \mathbb{S}_f), \tag{14}$$

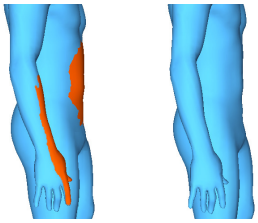
where $\mathcal{B}(\cdot)$ is the binary segmentation of $\mathbb{D}$ for removing the points with null depth. The depth of a point in $\mathbb{M}$ is updated with the average depth of its nearest neighbors along its four coordinates. In addition, noise depths are often significantly different from the true depths of a human; therefore, they can be removed by thresholding.

5.6.2 Penalty Enforced Cloud Prior Constraint. The foreground point cloud represents the general shape of the model and thus can be taken as a prior constraint for a better fitting. Previous methods typically take the point-to-plane distance between a model point to its correspondence in the point cloud as this constraint [Joo et al. 2018; Newcombe et al. 2015; Yu et al. 2018]. However, as shown in Figure 8(a), some model points may be outside of the cloud (see the orange parts on the arm, hand, and belly), which could lead to a distorted fatter shape than its ground truth. Therefore, one additional property needs to be considered to model a clothed human: all estimated model points should not be outside of the cloud, *i. e.*, always staying inside the clothed surface. Consequently, we propose a new point cloud prior constraint with a heavy penalty for the points outside of the cloud.

The constraint is formulated with the correspondences between the model points and the cloud points. However, the model point cannot correspond to the cloud point strictly one by one, considering the loosely clothed subject. Therefore, for a 3D model point, $\boldsymbol{v}_m$, its closest point in the cloud, $\boldsymbol{v}_c$, is taken as the correspondence. Assuming the normal of $\boldsymbol{v}_m$ is $\boldsymbol{n}_m$ and the correspondences set is $\mathbb{S}$, this penalty enforced constraint can be described as follows.

$$\begin{aligned} E_c = \sum_{(\boldsymbol{v}_m, \boldsymbol{v}_c) \in \mathbb{S}} & (\xi(\boldsymbol{v}_m)\lambda_0 \rho(\boldsymbol{n}_c^T \cdot (\boldsymbol{v}_c - \boldsymbol{v}_m) \\ & + (1 - \xi(\boldsymbol{v}_m))\lambda_1 \rho(\boldsymbol{n}_c^T \cdot (\boldsymbol{v}_c - \boldsymbol{v}_m))), \end{aligned} \tag{15}$$

where $\xi(\boldsymbol{v}_m)$ is an indicator of whether the model point $\boldsymbol{v}_m$ is inside the cloud (set as one) or not (set as zero). The zero indicator leads to a negative output in the second addition item in Equation 15. Therefore, this item can act as a penalty by setting λ_1 bigger than λ_0, and then ensures a strong penalty for the model points outside the cloud. Experimentally, λ_0 and λ_1 are set to 10 and 100, respectively.



(a) (b)

Figure 8: Demonstration of the cloud prior constraint without/with the penalty for model points outside of the cloud for the same subject shown in Figure 5. (a): The estimated shape without the penalty; and (b) the estimated shape with the penalty. Note that the orange surfaces in (a) represent the model points outside of the cloud.

$\boldsymbol{n}_c$ in Equation 15 is estimated by a least-squares plane fitting method among the k nearest neighbors of $\boldsymbol{v}_c$, This method relies on the covariance matrix among $\boldsymbol{v}_c$ and its k-nearest neighbors. Assuming one of the k nearest neighbor is $\boldsymbol{v}_i$, the covariance matrix Q for $\boldsymbol{v}_c$ can be computed as:

$$Q = \frac{1}{k} \sum_{i=1}^{k} (\boldsymbol{v}_i - \overline{\boldsymbol{v}}) \cdot (\boldsymbol{v}_i - \overline{\boldsymbol{v}})^T, \tag{16}$$

where $\overline{\boldsymbol{v}}$ is the centroid of the k nearest neighbors. $\boldsymbol{n}_c$ in Equation 15 is the eigenvector of Q corresponding to its minimal eigenvalue. Figure 8(b) shows the shape without outside-model points after applying the additional penalty term and thus is closer to the true shape (Figure 5(a)).

6 OPTIMIZATION DETAILS

A total of six terms are included in our energy minimization model, which may be difficult to converge due to the possible interference during the optimization. Therefore, we take a progressively coarse-to-fine strategy. First, a coarse model is computed by fitting the skeleton so that the general shape and pose are recovered. Then, an accurate model is computed by constraining the head pose and point cloud so that the fine THM can be obtained.

To compute the coarse model, the weights of both the head pose and the point cloud, w_h and w_c, are set to zero, *i.e.*, the consistencies of both head pose and point cloud are not considered. w_{data} and w_{sc} are progressively increased while decreasing w_p and fixing w_{sp}. Such a setup makes the shapes progressively obtained through the decreasing of the contribution of E_p, avoiding the over-fitting of the body pose during the optimization.

To compute the fine model, the weights of the terms used for the coarse model are fixed while the weights for both the head pose and the point cloud, w_h and w_c, are set to non-zero values so that the details of the head pose and other details can be obtained for an accurate modeling.

The optimization is fulfilled using the same method in the works of [Bogo et al. 2016; Zhang et al. 2017], *i.e.*, solving Equation 6 with

the "dogleg" gradient-based descent minimization method [Nocedal and Wright 2006], which is implemented in the Python based auto-differentiation tool, Chumpy [Loper 2018]. In the coarse pass of our experiments, w_{data}, and w_{sc} of Equation 6 are set to 1 and 5 initially and increased by 1 and 0.5, respectively, during the iterations. w_p is set to 1 initially and decreases 0.1 in each iteration while w_{sp} is fixed to 0.003. In the fine pass, w_h and w_c are set to 1 and 0.5, respectively. σ in Equation 7 is set to 100 to decrease the possibility of an incorrect estimation due to wrong key points.

7 EXPERIMENTAL RESULTS

We conducted many experiments both qualitatively and quantitatively on loosely clothed subjects[1] . Also, we compared our method with three state-of-the-art, single image-based shaped modeling approaches, using the codes provided by their authors: SMPLify [Bogo et al. 2016], Neural Body Fitting (NBF) [Omran et al. 2018], and Human Mesh Recovery (HMR) [Kanazawa et al. 2018]. SMPLify computes the optimized model iteratively with the 2D skeleton estimated by deep learning techniques. Both NBF and HMR directly apply different deep learning techniques for end-to-end modeling: The former uses a standard semantic segmentation CNN to obtain 12 semantic parts before estimating them with CNN, while the latter estimates the parameters using an iterative 3D regression module.

7.1 Qualitative Results

Figure 9 shows some intermediate results from SRDF for the shot shown in Figure 1. The final model on the right is obtained gradually from the initial model (standard template) on the left.

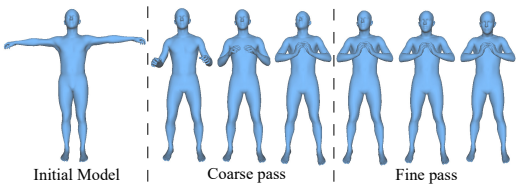


Figure 9: Intermediate models computed by SRDF for the shot shown in Figure 1.

Figure 10 shows some experimental results of SRDF for seven shots in different poses. We can see that our method can effectively model the whole human shape including the hands and head.

Comparative performances of SRDF and the other three selected approaches (i.e., SMPLify, NBF, and HMR) are shown in Figure 11. In this comparison, five subjects in six challenging poses are tested and compared. Our method can model the head, hands, and the overall shape more accurately than all the other three methods, especially for the head and hand parts.

We also visualize the per-vertex errors to show the modeling accuracies (Figure 12). The error of each vertex is computed as its distance between the estimated model and the corresponding ground truth and visualized via a heat map. The ground truth was obtained by laser scanning. Among all the approaches in this comparison, SRDF achieved the smallest errors almost everywhere, thanks to the detail-preserving property of our method. In addition, the second subject can be better modeled by SRDF with smaller errors in almost everywhere than the first subject. This is because the second subject is fatter than the first and therefore can be more easily fitted with less parameter variations.

[1]Codes available at https://sites.google.com/view/xianyong/home/publications.

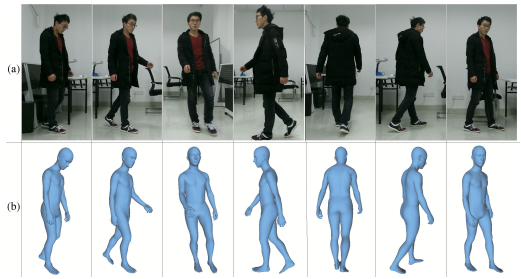


Figure 10: Experimental results of SRDF with seven shots. (a): A single source shot; and (b) the estimated human model.

7.2 Quantitative Results

We performed quantitative evaluations on 200 shots with different poses (Figure 13). The statistical performances of skeleton estimation and reconstruction accuracy are experimented and quantified.

First, the accuracy of skeleton detection is experimented by comparing the detection performance of the key points in the main skeleton by the Kinect SDK and SRDF. Figure 14(a) shows the ratio of each successfully detected key point by Kinect SDK. The key points on the torso (i.e., those on the spine, neck, and hip) can be accurately estimated while the remaining ones are more or less inaccurate. We also compared the skeleton estimation accuracies between the two methods. The length of each skeletal part connecting two neighboring key points in the estimated skeleton is first accumulated, and then its average error in comparison with the corresponding ground truth length is computed, as shown in Figure 14(b). Compared to the Kinect SDK, the length of each skeletal part by SRDF is closer to the ground-truth after applying our LCR-Net incorporated skeleton refinement step.

The reconstruction performance was experimented with the average per-vertex error of each shot between its estimated model and the ground truth (Figure 15). Here, the per-vertex error is computed in the same ways as the experiment in Figure 12. Our SRDF approach returned the lowest errors for almost all the shots than the other methods.

8 LIMITATIONS

Some limitations exist in our current work, which can be summarized as follows.

- SRDF adopts the Kinect as the experimental camera which is easy to deploy and experiment. However, the Kinect SDK may not correctly estimate the body and head poses, especially when the subject wears thick clothes or his/her face is partially visible. Consequently, in such cases the main body skeleton refinement and head pose prior estimation could fail. In addition, the Kinect can only capture images in 1080p,

Figure 11: Performances comparison of different methods under challenging poses. The two small images in each sub-window show the enlarged views of the head and hands, respectively. (a): the source shots; (b) SMPLify[Bogo et al. 2016]; (c) NBF[Omran et al. 2018]; (d) HMR[Kanazawa et al. 2018]; and (e) SRDF.

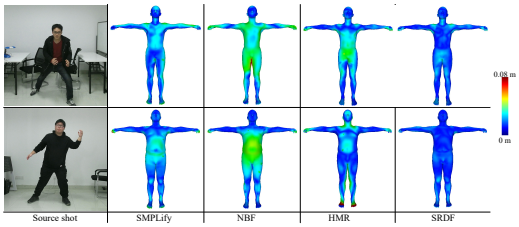


Figure 12: Comparison of different methods (SMPLify[Bogo et al. 2016], NBF[Omran et al. 2018], HMR[Kanazawa et al. 2018] and SRDF) using the per-vertex error.

whose frame rate can be automatically decreased to 15 fps. Therefore, it may easily capture blurred silhouette or hands, especially when the subject stands 3 meters or further away. Then, the foreground segmentation of CRF-RNN or the deep hand pose estimator may not obtain a correct foreground point cloud or the keypoints on the hands respectively, which could make the fine THM failed.

- The existing RGB-image based deep learning techniques help SRDF in various ways. However, they could fail even when the image is clear with a high resolution. One of the well-known reasons is that their performances largely rely on the training data and can be unstable if the test image is significantly different from them. In addition, the deep hand pose estimator requires the hands to appear clearly. However, hands may be hidden or their fingers can be seriously occluded. Their tiny joints make things even worse, especially when they are a part of the completely captured human shape. Then, obtaining separate key points of the hands is sometimes difficult (Figure 16). This shortcoming generally exists in any other hand pose estimators. Therefore, better deep learning techniques need to be developed.
- Our experiments show the modeling accuracy can be improved further if more effective optimization model can be formulated. New constraints presented in SRDF have already improved the modeling quality. Therefore, possible new constraints could be incorporated to further improve the performance, for example, constraints for recovering the winkles and eyes. Our current work estimates the skeleton as the first step. However, the skeleton alone may not be robustly estimated as discussed above. Therefore, some unified optimizations can be a possible improvement so that the fine skeleton and human shape can be recovered simultaneously and complement with each other.

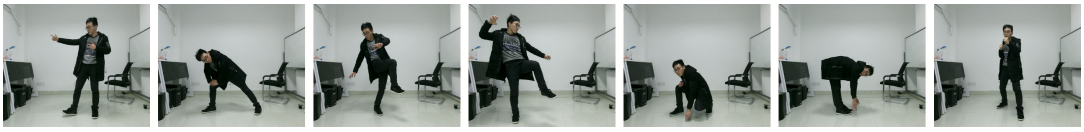


Figure 13: Some example shots in our test dataset

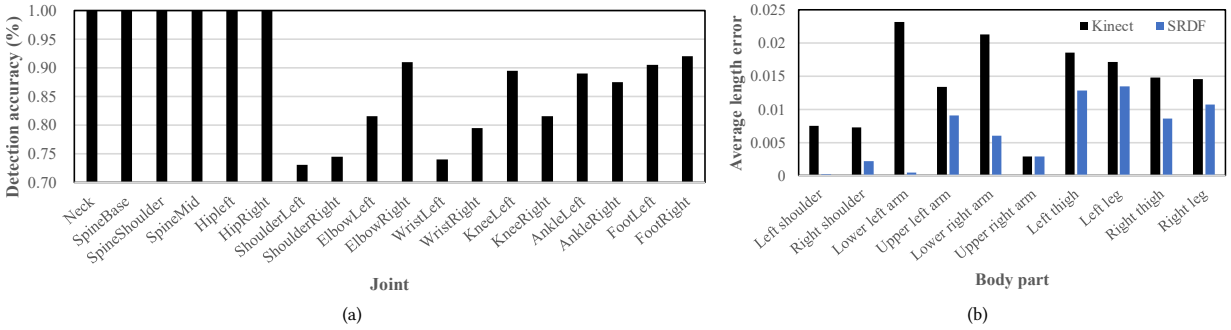


Figure 14: Statistical comparison of the skeleton estimation performance between the Kinect SDK and SRDF. (a): The detection accuracy of each key point by the Kinect SDK; and (b): the average length accuracy of each skeletal part formed by connecting neighboring key points. Note that, (1) the key points in (a) are named according to the Kinect SDK; and (2) the skeletal parts on the torsos are not shown in (b) because of their key points have 100% detection accuracies as shown in (a).

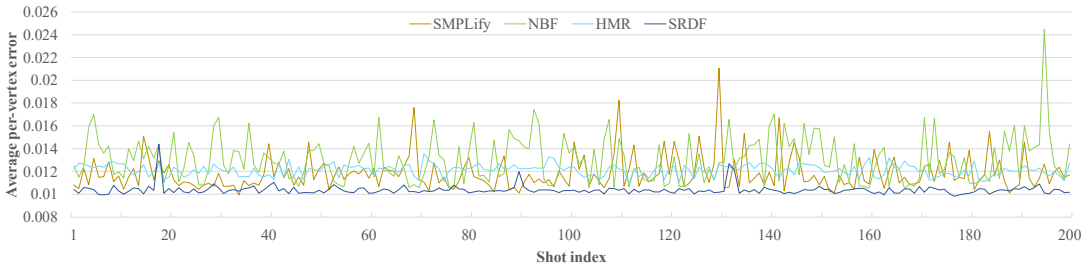


Figure 15: Statistical comparison of the modeling performances of different methods by the average per-vertex error

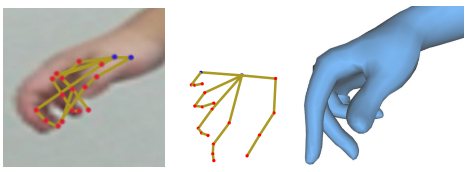


Figure 16: Estimated skeleton of the right hand of the bottom subject in Figure 1 by Zimmermann and Brox [Zimmermann and Brox 2017]. (a) The 2D view; (b) the 3D view; (c) the final model by SRDF.

9 CONCLUSION

In this paper, we present a novel single RGB-D shot based method (called SRDF) to accurately reconstruct 3D human model. It adopts the merits of both 3D point clouds and 2D RGB images in the following aspects, thanks to the advances of deep learning techniques: 1) a full body pose including fine hands is obtained with a compound skeleton detection method, which integrates two single RGB image based deep body and hand pose estimators into a point cloud based body pose detector; and 2) a smoothed foreground point cloud without holes and noises is obtained by a deep RGB image segmentation assisted point cloud pre-processing algorithm. Furthermore, a new energy minimization model is also introduced with three novel constraints, including the shape continuity constraint for the smooth shape update during iterations, the head pose prior constraint for the accurate orientation of the head, and the penalty enforced point cloud prior constraint to punish the points outside the clothed cloud. We also present a two-passes optimization method for a coarse-to-fine human shape recovery. Our experimental results demonstrate the effectiveness and accuracy of our SRDF method.

ACKNOWLEDGMENTS

This work is co-supported by Natural Science Fouccation of China (61502005) and Key Science & Technology Program of Anhui Province (1604d0802004).

REFERENCES

Thiemo Alldieck, Marc Kassubeck, Bastian Wandt, Bodo Rosenhahn, and Marcus Magnor. 2017. Optical flow-based 3D human motion estimation from monocular video. In *German Conference on Pattern Recognition*. 347–360.

Thiemo Alldieck, Marcus A Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. 2018. Video Based Reconstruction of 3D People Models. In *CVPR*. 8387–8397.

Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. 2005. SCAPE: shape completion and animation of people. *ACM Trans. on graphics* 24, 3 (2005), 408–416.

Michael Black, Javier Romero, Gerard Pons-Moll, Naureen Mahmood, and Federica Bogo. 2016. Learning Human Body Shapes in Motion. In *ACM SIGGRAPH 2016 Courses*. 17:1–17:411.

Federica Bogo, Michael J Black, Matthew Loper, and Javier Romero. 2015. Detailed full-body reconstructions of moving people from monocular RGB-D sequences. In *ICCV*. 2300–2308.

Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. 2016. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *ECCV*. 561–578.

Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2018. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. In *arXiv preprint arXiv:1812.08008*.

Xiaowu Chen, Yu Guo, Bin Zhou, and Qinping Zhao. 2013. Deformable model for estimating clothed and naked human shapes from a single image. *The Visual Computer* 29, 11 (2013), 1187–1196.

Zhi-Quan Cheng, Yin Chen, Ralph R Martin, Tong Wu, and Zhan Song. 2018. Parametric modeling of 3D human body shape - A survey. *Computers & Graphics* 71 (2018), 88–100.

Edilson De Aguiar, Carsten Stoll, Christian Theobalt, Naveed Ahmed, Hans-Peter Seidel, and Sebastian Thrun. 2008. Performance capture from sparse multi-view video. *ACM Trans. on Graphics* 27, 3 (2008), 98.

Endri Dibra, Himanshu Jain, A Cengiz Öztireli, Remo Ziegler, and Markus H Gross. 2017. Human Shape from Silhouettes Using Generative HKS Descriptors and Cross-Modal Neural Networks. In *CVPR*, Vol. 2. 5504–5514.

Endri Dibra, Cengiz Öztireli, Remo Ziegler, and Markus Gross. 2016. Shape from selfies: Human body shape estimation using cca regression forests. In *ECCV*. 88–104.

Peng Guan, Alexander Weiss, Alexandru O Balan, and Michael J Black. 2009. Estimating human shape and pose from a single image. In *ICCV*. 1381–1388.

Nils Hasler, Carsten Stoll, Bodo Rosenhahn, Thorsten Thormählen, and Hans-Peter Seidel. 2009. Estimating body shape of dressed humans. *Computers & Graphics* 33, 3 (2009), 211–216.

Aaron S Jackson, Chris Manafas, and Georgios Tzimiropoulos. 2018. 3D Human Body Reconstruction from a Single Image via Volumetric Regression. In *ECCV 2018 Workshops*. 64–77.

Zhongping Ji, Xiao Qi, Yigang Wang, Gang Xu, Peng Du, and Qing Wu. 2018. Shape-from-Mask: A Deep Learning Based Human Body Shape Reconstruction from Binary Mask Images. *arXiv preprint arXiv:1806.08485* (2018).

Hanbyul Joo, Tomas Simon, and Yaser Sheikh. 2018. Total Capture: A 3D Deformation Model for Tracking Faces, Hands, and Bodies. In *CVPR*. 8320–8329.

Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. 2018. End-to-end recovery of human shape and pose. In *CVPR*. 7122–7131.

Don Koks. 2006. *Explorations in mathematical physics: the concepts behind an elegant language*. Springer-Verlag New York, New York, NY.

Christoph Lassner, Gerard Pons-Moll, and Peter V Gehler. 2017a. A generative model of people in clothing. In *ICCV*, Vol. 6. 853–862.

Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J Black, and Peter V Gehler. 2017b. Unite the people: Closing the loop between 3D and 2D human representations. In *CVPR*, Vol. 2. 4704–4713.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature* 521, 7553 (2015), 436.

Zhongguo Li, Anders Heyden, and Magnus Oskarsson. 2019. Template based human pose and shape estimation from a single RGB-D image. In *8th International Conference on Pattern Recognition Applications and Methods*, Ana Fred, Maria De Marsico, and Gabriella Sanniti di Baja (Eds.). 574–581.

Matthew Loper. 2018. Chumpy. https://pypi.python.org/pypi/chumpy.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. 2015. SMPL: A skinned multi-person linear model. *ACM Trans. on Graphics* 34, 6 (2015), 248.

Eric Marchand, Hideaki Uchiyama, and Fabien Spindler. 2016. Pose estimation for augmented reality: a hands-on survey. *IEEE Trans. on Visualization and Computer Graphics* 22, 12 (2016), 2633–2651.

Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. 2017. VNect: Real-time 3D Human Pose Estimation with a Single RGB Camera. *ACM Trans. on Graphics* 36, 4 (2017), 44:1–44:14.

Erik Murphy-Chutorian and Mohan Manubhai Trivedi. 2009. Head pose estimation in computer vision: A survey. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 31, 4 (2009), 607–626.

Alexandros Neophytou and Adrian Hilton. 2014. A Layered Model of Human Body and Garment Deformation. In *International Conference on 3D Vision (3DV)*. 171–178.

Richard A Newcombe, Dieter Fox, and Steven M Seitz. 2015. DynamicFusion: Reconstruction and tracking of non-rigid scenes in real-time. In *CVPR*. 343–352.

Jorge Nocedal and Stephen Wright. 2006. *Numerical optimization* (2nd ed.). Springer, New York, NY.

Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter Gehler, and Bernt Schiele. 2018. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *International Conference on 3D Vision (3DV)*. 484–494.

Mei Oyama, Naoshi Kaneko Aoyama, Masaki Hayashi, Kazuhiko Sumi, and Takeshi Yoshida. 2017. Two-stage model fitting approach for human body shape estimation from a single depth image. In *IAPR International Conference on Machine Vision Applications (MVA)*. 234–237.

Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. 2018. Learning to Estimate 3D Human Pose and Shape from a Single Color Image. In *CVPR*. 459–468.

Leonid Pishchulin, Stefanie Wuhrer, Thomas Helten, Christian Theobalt, and Bernt Schiele. 2017. Building statistical shape spaces for 3D human modeling. *Pattern Recognition* 67 (2017), 276–286.

Gerard Pons-Moll, Sergi Pujades, Sonny Hu, and Michael J Black. 2017. Clothcap: Seamless 4D clothing capture and retargeting. *ACM Trans. on Graphics* 36, 4 (2017), 73.

Gregory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. 2017. LCR-Net: Localization-classification-regression for human pose. In *CVPR*. 1216–1224.

Lorenz Rogge, Felix Klose, Michael Stengel, Martin Eisemann, and Marcus Magnor. 2014. Garment replacement in monocular video sequences. *ACM Trans. on Graphics* 34, 1 (2014), 6.

Javier Romero, Dimitrios Tzionas, and Michael J Black. 2017. Embodied hands: Modeling and capturing hands and bodies together. *ACM Trans. on Graphics* 36, 6 (2017), 245.

Jamie Shotton, Andrew Fitzgibbon, Mat Cook, Toby Sharp, Mark Finocchio, Richard Moore, Alex Kipman, and Andrew Blake. 2011. Real-time human pose recognition in parts from single depth images. In *CVPR*. 1297–1304.

Carsten Stoll, Juergen Gall, Edilson De Aguiar, Sebastian Thrun, and Christian Theobalt. 2010. Video-based reconstruction of animatable human characters. *ACM Trans. on Graphics* 29, 6 (2010), 139.

Gül Varol, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid. 2018. BodyNet: Volumetric Inference of 3D Human Body Shapes. In *ECCV*. 20–38.

Abhinav Venkat, Sai Sagar, and Avinash Sharma. 2018. Deep Textured 3D Reconstruction of Human Bodies. In *BMVC*. 286.

Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. 2019. Monocular Total Capture: Posing Face, Body, and Hands in the Wild. In *CVPR*.

Weipeng Xu, Avishek Chatterjee, Michael Zollhöfer, Helge Rhodin, Dushyant Mehta, Hans-Peter Seidel, and Christian Theobalt. 2018. Monoperfcap: Human performance capture from monocular video. *ACM Trans. on Graphics* 37, 2 (2018), 27.

Mao Ye and Ruigang Yang. 2014. Real-time simultaneous pose and shape estimation for articulated objects using a single depth camera. In *CVPR*. 2345–2352.

Jaeho Yi, Seungkyu Lee, Sujung Bae, and Moonsik Jeong. 2015. Human Body Volume Recovery from Single Depth Image. In *International Symposium on Visual Computing*. 396–405.

Tao Yu, Zerong Zheng, Kaiwen Guo, Jianhui Zhao, Qionghai Dai, Hao Li, Gerard Pons-Moll, and Yebin Liu. 2018. DoubleFusion: Real-Time Capture of Human Performances With Inner Body Shapes From a Single Depth Sensor. In *CVPR*. 7287–7296.

Chao Zhang, Sergi Pujades, Michael J Black, and Gerard Pons-Moll. 2017. Detailed, accurate, human shape estimation from clothed 3D scan sequences. In *CVPR*, Vol. 2. 5484–5493.

Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip HS Torr. 2015. Conditional random fields as recurrent neural networks. In *ICCV*. 1529–1537.

Xiangxin Zhu and Deva Ramanan. 2012. Face detection, pose estimation, and landmark localization in the wild. In *CVPR*. 2879–2886.

Christian Zimmermann and Thomas Brox. 2017. Learning to estimate 3D hand pose from single RGB images. In *ICCV (1)*. 3.

Silvia Zuffi, Angjoo Kanazawa, and Michael J Black. 2018. Lions and Tigers and Bears: Capturing Non-Rigid, 3D, Articulated Shape From Images. In *CVPR*. 3955–3963.