




A Robust Multi-View System for High-Fidelity Human Body Shape Reconstruction

Qitong Zhang¹  Lei Wang¹  Linlin Ge¹  Shan Luo¹  Taihao Zhu¹  Feng Jiang¹  Jimmy Ding¹  and Jieqing Feng^{†1} ¹State Key Laboratory of CAD&CG, Zhejiang University, China

Abstract

This paper proposes a passive multi-view system for human body shape reconstruction, namely RHF-Human, to overcome several challenges including accurate calibration and stereo matching in self-occluded and low-texture skin regions. The reconstruction process includes four steps: capture, multi-view camera calibration, dense reconstruction, and meshing. The capture system, which consists of 90 digital single-lens reflex cameras, is single-shot to avoid nonrigid deformation of the human body. Two technical contributions are made: (1) a two-step robust multi-view calibration approach that improves calibration accuracy and saves calibration time for each new human body acquired and (2) an accurate PatchMatch multi-view stereo method for dense reconstruction to perform correct matching in self-occluded and low-texture skin regions and to reduce the noise caused by body hair. Experiments on models of various genders, poses, and skin with different amounts of body hair show the robustness of the proposed system. A high-fidelity human body shape dataset with 227 models is constructed, and the average accuracy is within 1.5 mm. The system provides a new scheme for the accurate reconstruction of nonrigid human models based on passive vision and has good potential in fashion design and health care.

CCS Concepts

- **Computing methodologies** → Computer graphics; Shape modeling; Mesh models;

1. Introduction

Human body shape acquisition and reconstruction are active research topics in many fields ranging from animation to fashion design, health care, and digitized virtual humans. There are two types of human body shape acquisition techniques: (1) dynamic acquisitions techniques [PRMB15, CCS*15, BBLR15, LFB17] that are focused on the natural movements of humans, such as motion capture and surface deformation, and (2) static acquisitions techniques [BBB*10, RZY*20] used for accurate human body shape reconstruction. In recent decades, many static active-vision systems [TZL*12, LCK*21] and model-based techniques [GWBB09, KBJM18] have been proposed for accurate human body shape reconstruction. It is generally believed that passive-vision systems based on multi-view stereo (MVS) are inferior to the above mainstream approaches, but that they have their advantages and tremendous potential.

Early passive-vision methods suffered from 3D body shape capturing challenges, such as calibration accuracy and stereo matching in self-occluded and low-texture skin regions. Even so, a wide range of applications exists for passive-vision techniques based on their unique advantages. Different from active-vision systems

with long scanning times and model-based methods with low-resolution templates, the passive-vision technique possesses the following features: it uses single-shot capture and is template independent. Over the last few years, the passive-vision techniques have developed rapidly with the release of high-resolution datasets [SSG*17, KPZK17] of challenging static scenes. Then, a motivating question is whether it is possible to design a passive system that can overcome the main challenges of body shape capturing and acquire high-quality human body shapes.

To this end, inspired by [BBB*10, RZY*20], a robust passive-vision system (RHF-Human) is proposed to acquire and reconstruct accurate human body shapes robustly. The flowchart of the system is shown in Figure 1. First, a multi-view system consisting of 90 digital single-lens reflex (DSLR) cameras is built. This system can capture a human body shape in a single shot. Second, a two-step robust camera calibration approach is proposed. It can improve the accuracy and robustness of the system calibration process and save calibration time for each new human body acquired. Third, in the dense reconstruction stage, an improved PatchMatch MVS method is adopted to generate a high-quality point cloud. The basic model of joint pixelwise view selection and depth-normal estimation can alleviate the self-occlusion problem and improve the reconstruction accuracy at the subpixel level. To overcome the increasing negative impact of high-frequency details, such as body hair, a local smooth-

† Corresponding author: jqfeng@cad.zju.edu.cn

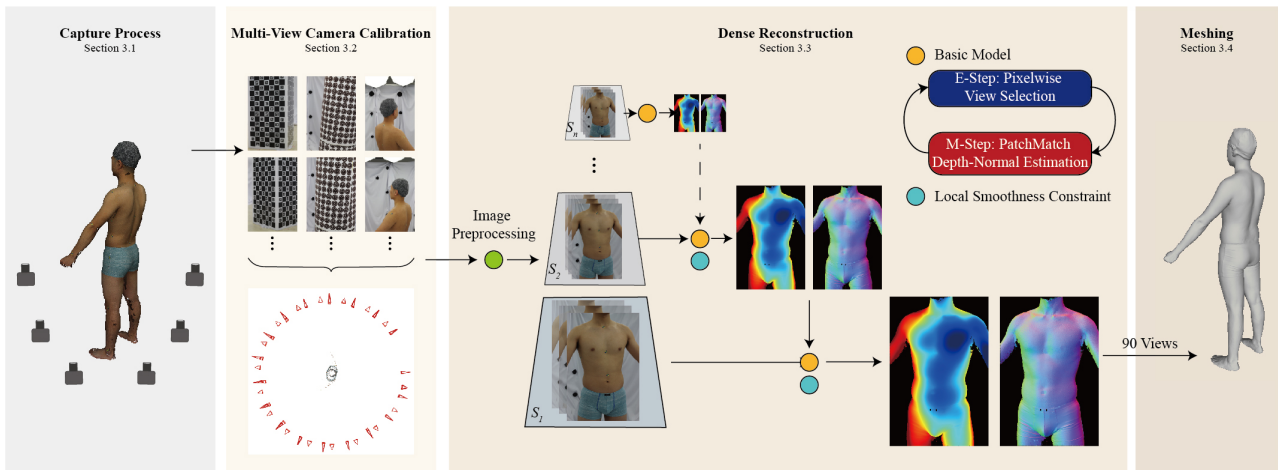


Figure 1: Overview of the proposed human body shape acquisition and reconstruction pipeline.

ness constraint is introduced to optimize the reconstruction of skin occluded by the body hair. In addition, the proposed hierarchical framework can further decrease the impact of body hair and alleviate matching ambiguity in low-texture skin regions. Finally, some digital geometry processing techniques are performed to generate a watertight mesh of the acquired human body shapes; these techniques include outlier removal, hole filling, and surface reconstruction. A total of 227 watertight meshes of 57 subjects with different genders, poses, and skin with different amounts of body hair are tested, and an average reconstruction accuracy of 1.5 mm is obtained. The reconstructed human body shape dataset demonstrates the robustness and accuracy of the proposed system.

In conclusion, this paper provides a systematic contribution, which is a novel passive-vision system for accurate static human body acquisition, and two special contributions, including a robust system calibration process and an accurate and self-optimizing dense reconstruction method. In the professional area, we develop a high-quality passive-vision system for nonrigid human body shape capture in a single-shot way, which overcomes the weakness of mainstream active-vision systems and the main challenges of passive-vision methods, and generates body shape models with high accuracy. For practical purposes, we show that the proposed system offers a universal solution for passive human body shape reconstruction with various genders, poses, and skin with different amounts of body hair, which has great potential applications.

2. Related Work

According to the type of vision system and whether a parametric model is used, human body shape reconstruction methods can be roughly divided into three types: (1) model-free, active-vision methods; (2) model-based methods; and (3) model-free, passive-vision methods. Many dynamic acquisition systems focus on the capture of motion and surface deformation. In this section, we review the most relevant static methods that reconstruct accurate human body shapes.

Model-Free, Active-Vision Reconstruction. Active-vision

methods utilize depth sensors to obtain raw multi-view 3D data and then fuse them via point cloud registration. In recent decades, high-end scanning systems have been used, such as laser [ACP03, YWK20] or structured light [BRLB14, PRMB15] systems, to robustly scan high-accuracy human body shapes. In contrast, some cheap, consumer-level systems [LVG*13, TZL*12] using RGB-D (color and depth) scanners, such as Kinect, have drawn much attention from the community but suffer from noise disturbances, which lead to low-precision results. Due to this concern, recent learning-based work [LCK*21] focused on how to reconstruct implicit surface representations from noisy and incomplete depth maps. However, the performance of such a method is highly dependent on the training datasets and limited by the GPU memory size. In particular, the static active-vision acquisition systems require the scanned person to either stand still, be rotated on a turntable to be scanned from different views, or rotate in front of the sensor while trying to roughly maintain the same pose. Thus, these methods are not ideal for static nonrigid human body shape reconstruction. In contrast, this paper designs a robust passive-vision system that is one-shot and easy-to-deploy for reconstructing high-quality human body shape models.

Model-Based Reconstruction. Model-based methods fit shape and pose parameters to incomplete inputs (i.e., 3D point clouds, images, and silhouettes) utilizing parametric models [ASK*05, LMR*15] to obtain complete naked [WHB11, ZLNW19] or clothed [YFHW16, ZPBP17] human body shapes. For inputs of 3D point clouds, Weiss et al. [WHB11] and Zhao et al. [ZLNW19] captured several point clouds from different views, estimated per-view optimal pose and shape parameters, and finally generated consistent shape and pose parameters to reconstruct 3D human body shapes. Achenbach et al. [AWLB17] computed dense point clouds through MVS and then fitted a template model to the scanner data to generate models that are ready to be animated. Regarding 2D information inputs, early works [BSB*07, GWBB09] estimated the parameters of the SCAPE model [ASK*05] utilizing silhouettes and 2D joints with manual intervention. Recent works have focused on automatic methods incorporating cues (e.g. 2D joints [BKL*16], sil-

houettes [LRK*17], and multi-view images [Hua17]) and learning-based methods [KBJM18, AMB*19] without any 2D detections. Although model-based methods can handle abundant poses, the low-resolution template is a limitation, which tends to eliminate high-frequency details. In contrast, the proposed method is template independent when reconstructing high-precision human body shapes.

Model-Free, Passive-Vision Reconstruction. Passive-vision methods acquire 3D information from 2D images that are usually captured from multi-view viewpoints simultaneously in a studio environment. There are many passive multi-view systems [LDX10, Rem04, FRS17] focused on human body shape reconstruction. For the acquisition systems, sparse camera setups [SH07, VPB*09, TNM09] with extremely wide baselines or full-body capture systems [VPB*09, LDX10, JLT*15] with low percentages of body pixels limit the quality of reconstructed human body shapes. The calibration process is easily damaged by inaccurate matching and even small camera movements caused by studio staff or camera gravity. For the reconstruction methods, early attempts were based on visual hulls [MBR*00, VBMP08, FP09] but could not handle concavities or generate fine-scale details. More accurate geometries can be acquired by utilizing multi-view stereo constraints [SCD*06, FH15]. However, the performances of these methods are restricted by several challenges, such as stereo matching in strongly occluded [ES04, SH07] and low-texture regions [FP10, LQ05], and the limited estimation accuracy based on the front-parallel assumption [VETC07, CVHC08]. Thus, in the past, the reliability and accuracy of passive systems were generally considered to be inferior to those of active methods.

In the last decade, some more advanced techniques have focused on calibration [LMS16, SF16], visibility estimation [SZFP16, XT19], low-texture regions [RZY*20, XLS*20], and mesh refinement integrated with shading cues [WVT12, WWMT11, WLDW11], to make the resulting 3D models as accurate as possible. Recently, PatchMatch MVS methods [GLS15, SZFP16], which discard the front-parallel assumption and adopt the core idea of the PatchMatch algorithm [BSFG09], have shown great power in solving dense matching problems with high accuracy and efficiency. Based on PatchMatch MVS, pixelwise view selection strategies [SZFP16, XT19] and some cues (e.g. planar priors [RM19, XT20]) and multi-scale frameworks [XT19, LFYX19]) used for low-texture regions were proposed to further improve the accuracy and completeness of reconstruction. In addition, learning-based methods have also achieved excellent performance on single-view [TTC*19, NSH*19] and multi-view [YLL*18, LFB18] reconstruction tasks based on voxels [GVCH18, HLC*18], depth maps [LFB18, TTC*19], and implicit functions [SHN*19, SSSJ20]. These learning-based methods offer a new direction for passive-vision approaches but are limited by the bottlenecks of the training datasets and GPU memory size.

Supported by the great potential of passive methods, some works on human faces [BBB*10] and human body shapes [RZY*20] were developed using pairwise stereo reconstruction and obtained excellent results. To further propel the development of passive methods in the field of human body shape reconstruction, we propose a passive multi-view system utilizing a robust camera calibration ap-

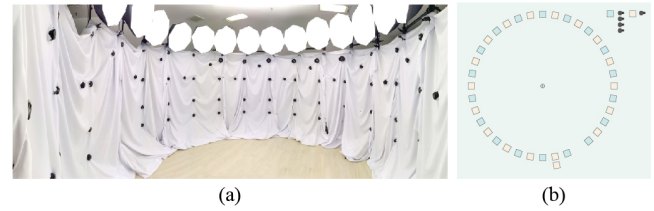


Figure 2: The camera setup of the multi-view system. (a) Real capturing system. (b) Diagram of the camera setup.

proach and an advanced PatchMatch MVS method to acquire accurate results. Although our body scanning pipeline is similar to the previous work [FRS17], they focused on the efficiency of model generation while the proposed system focuses on the accuracy.

3. Multi-View Shape Acquisition and Reconstruction

This section describes the proposed robust passive multi-view system, as shown in Figure 1, including capture process, multi-view camera calibration, dense reconstruction, and meshing.

3.1. Capture Process

Human body shapes are captured using the proposed multi-view camera system as shown in Figure 2. The multi-view setup consists of 90 DSLR cameras arranged around a circular capture space with a radius of 2 m. We set four layers of cameras, focusing on the main torso from top to bottom (four cameras form a group) to capture various human body shapes and heights (up to 2 m). These cameras constitute 18 groups arranged approximately every 20 degrees. Six Canon 5D Mark IV cameras with 70 mm prime lenses are placed on the front of the body placed at the top layer for the human face capture, and 66 Canon 600D and 700D cameras with 50 mm prime lenses are placed for other views of the main torso. The rest of the 18 cameras (Canon 750D) with 28 mm wide-angle lenses are inserted among every neighboring group for wider views of the various human body shapes to complement the information and ensure a flexible pose space.

The cameras are synchronizable to approximately 0.5 ms using the remote shutter, and this is sufficient for static subjects. The highest resolution setting ($> 5000 * 3000$) of every type of camera is utilized with more than 50% of the pixels accounted for the human body. Diffuse environmental lighting is fixed and directed to human skin to prevent specular highlights. Significantly, the camera setup can be modified for different requirements. Thus, the system is easy to adapt for practical use.

3.2. Robust Multi-View Camera Calibration

Based on the theoretical foundation of camera calibration, we concentrate on the practical problem of designing a reliable and efficient calibration system for capturing human body shapes. First, we estimate the initial camera parameters using calibration objects to lay the foundation for high-accuracy results. A rigid right parallelepiped (Figure 3(a)) equipped with ChAruco patterns

[GMMM14] (Figure 3(b)) is placed at different positions and rotation angles, captured 40-50 times in total, and used to acquire all intrinsic camera parameters K and distortion parameters via [Zha00]. The exact corners of the ChAruco patterns with distinguishable tags provide correspondences between cameras, and they give a known metric distance D between two corners for setting a scale factor. An incremental structure-from-motion (SfM) method [SF16] is applied to obtain extrinsic camera parameters $\{R, t\}$ utilizing these discriminable pattern corners. Furthermore, points on a calibration cylinder with encoded patterns [RZY*20] (Figure 3(c)-(d)), which are captured 20-30 times in total, are integrated into the SfM pipeline to improve the camera pose accuracy. We move, rotate, and incline the cylinder to acquire more points and cover the space missed by the encoded points of the right parallelepiped. Second, a refinement process is applied to avoid the slight camera motions caused by studio staff or camera gravity during capturing, and to increase the robustness of the hardware. We directly match the DSP-SIFT [DS15] features in human images with a strict threshold, triangulate these points, and then optimize the initial camera poses and intrinsic parameters via bundle adjustment (BA) [TMHF99]. Exact camera parameters can be acquired to guide the subsequent dense reconstruction process. The whole calibration pipeline is shown in Figure 4.

The proposed calibration approach has the following advantages. The calibration process with 3D objects is suitable for human body shapes. Both the parallelepiped and the cylinder are sized to match the body of the subject, moved, and rotated many times around the position at which the subject stands. Thus, the calibration process is well estimated with sufficient calibration data in the subject-occupied regions, which are the same as the workspaces of the calibration objects. Unlike the required complete views of the checkerboard calibration and inexact features of LED-based calibration, the ChAruco patterns of the parallelepiped provide unique and accurate subpixel features and allow for partial views. The accurate initial camera parameters lay a foundation of high accuracy and restrict the possible subsequent optimization errors caused by the incorrect matching of human features. Finally, the calibration does not need to be repeated for each new subject due to the introduced refinement process. Whenever slight camera movements caused by external disturbances occur during capture, the next group of captured human images can be used to rectify the disturbed camera parameters via strict BA. Thus, the refinement process enhances the robustness of the calibration and further improves the quality of the reconstructed human body shapes.

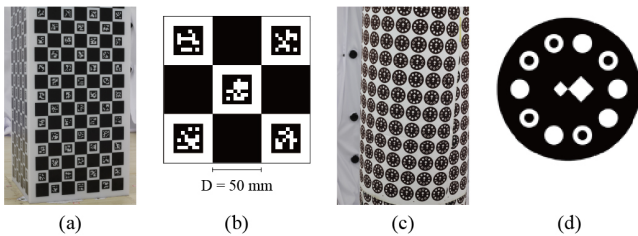


Figure 3: Calibration objects. (a) The rigid calibration right parallelepiped. (b) Examples of ChAruco patterns. (c) The calibration cylinder. (d) One example of an encoding pattern.

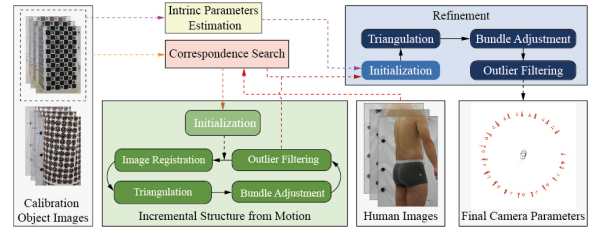


Figure 4: Overview of the robust multi-view calibration pipeline.

3.3. Accurate Dense Reconstruction

In this section, we introduce the proposed MVS algorithm for universal human body shape reconstruction. In Section 3.3.1, the basic graphical model is described and used in the presented algorithm to overcome the matching problem in self-occluded regions and the inexact depth estimation relied on the front-parallel assumption, and to improve the estimated accuracy of the algorithm. In Section 3.3.2, the detailed MVS algorithm is described to solve the challenges encountered in reconstruction, including the increased negative impact of body hair caused by the improved subpixel accuracy, and the matching ambiguity in the low-texture skin regions.

3.3.1. Basic Model

This section describes the adopted basic model [ZDJF14, SZFP16] for joint pixelwise view selection and depth-normal estimation to alleviate the matching issue in self-occluded regions and the limited accuracy caused by the front-parallel assumption. Since each row/column is processed independently and alternatively in parallel computing, we describe this framework as being limited to a single line.

Given a reference image X^{ref} and a set of source images $X^{\text{src}} = \{X^m \mid m = 1, \dots, M\}$ with known camera parameters, this method models the depth θ_l and the normal \mathbf{n}_l as a Markov process, together with the hidden state $Z_l^m \in \{0, 1\}$ which defines whether pixel l is visible in the source image m . Then, an inference is formulated as a maximum-a posteriori (MAP) and optimized iteratively. To solve the posterior $P(Z, \theta, \mathbf{N} \mid X)$, variational inference was used in [ZDJF14] to approximate the real posterior with a function $q(Z, \theta, \mathbf{N})$, which ensures that the Kullback-Leibler divergence between these two posteriors is minimized. Analog to [ZDJF14], Schönberger et al. [SZFP16] factorized the approximation $q(Z, \theta, \mathbf{N}) = q(Z)q(\theta, \mathbf{N})$ and constrained $q(\theta, \mathbf{N})$ to the family of Kronecker delta functions $q(\theta_l, \mathbf{n}_l) = \delta(\theta_l = \theta_l^*, \mathbf{n}_l = \mathbf{n}_l^*)$. A variant of the generalized expectation-maximization (GEM) algorithm [NH98] is utilized to estimate this approximation. In the E step of the GEM algorithm, the forward-backward algorithm is applied to infer Z in the hidden markov chain while keeping (θ, \mathbf{N}) fixed. In the M step of the GEM algorithm, the hidden variable Z is fixed and (θ, \mathbf{N}) is calculated through PatchMatch sampling and sequential propagation. The optimal pair $(\hat{\theta}_l^{\text{opt}}, \hat{\mathbf{n}}_l^{\text{opt}})$ is estimated as follows:

$$(\hat{\theta}_l^{\text{opt}}, \hat{\mathbf{n}}_l^{\text{opt}}) = \arg \min_{\theta_l^*, \mathbf{n}_l^*} \frac{1}{|S|} \sum_{m \in S} \xi_l^m(\theta_l^*, \mathbf{n}_l^*), \quad (1)$$

$$\xi_l^m(\theta_l^*, \mathbf{n}_l^*) = 1 - \rho_l^m(\theta_l, \mathbf{n}_l) + \eta \min(\psi_l^m, \psi_{\max}). \quad (2)$$

where S is a subset of the source images selected from a distribution $P_l(m)$, which encourages the sampling images to have sufficient baselines, similar resolutions, and nonoblique viewing directions. The cost $\xi_l^m(\theta_l^*, \mathbf{n}_l^*)$ includes the photometric cost ρ_l^m using a bilaterally weighted NCC and the geometric consistency cost as the forward-backward reprojection error $\psi_l^m = \|x_l - H_l^m H_l x_l\|$, where H_l^m and H_l denote the homography matrix transformation of the patch from the source to the reference image and from the reference to the source image, respectively. This function uses $\eta = 0.5$ as a constant regularizer and $\psi_{\max} = 3px$ as the maximum reprojection error. In addition, the inference stage is decomposed into two stages due to the memory constraints incurred when computing the geometric consistency. In the first stage, the initial depths and normals for each image are estimated by the photometric consistency using only the photometric cost in Equation (2). In the second stage, the final estimations are acquired by combining the photometric and geometric consistency as in Equation (2).

According to the PatchMatch scheme in [SZFP16], the pair $(\theta_l^*, \mathbf{n}_l^*)$ is selected from the hypotheses set during each sweep:

$$\left\{ (\theta_l, \mathbf{n}_l), (\theta_{l-1}^{\text{prp}}, \mathbf{n}_{l-1}), (\theta_l^{\text{rnd}}, \mathbf{n}_l), (\theta_l, \mathbf{n}_l^{\text{rnd}}), (\theta_l^{\text{rnd}}, \mathbf{n}_l^{\text{rnd}}), (\theta_l^{\text{prt}}, \mathbf{n}_l), (\theta_l, \mathbf{n}_l^{\text{prt}}) \right\}, \quad (3)$$

where $\theta_{l-1}^{\text{prp}}$ and \mathbf{n}_{l-1} are the propagated depth and normal estimations of the previous pixel, respectively, θ_l^{prt} and $\mathbf{n}_l^{\text{prt}}$ denote the perturbed parameters of the current estimation θ_l and \mathbf{n}_l , respectively, and θ_l^{rnd} and $\mathbf{n}_l^{\text{rnd}}$ are the randomly generated samples, respectively.

In this way, the basic PatchMatch MVS algorithm estimates the visibility values of neighboring views and the plane parameters, including the depth and normal for each pixel. The pixel-wise view selection strategy can acquire accurate visibility information for the self-occluded regions and improve the matching accuracy of the algorithm. The introduced normal estimation for each pixel avoids the incorrect estimation caused by the front-parallel assumption. With accurate camera parameters, the basic model can improve the accuracy at the subpixel level.

3.3.2. Detailed MVS Algorithms

Despite the improved subpixel accuracy of the algorithm, the negative impact of body hair is also increased and incomplete reconstruction in low-texture skin regions is not addressed. This section describes the proposed algorithms in detail for addressing challenges of increased body hair disturbance and matching uncertainty in low-texture regions. Although the negative impact of body hair can be addressed in a post-processing step by applying standard geometry processing techniques, some operations, such as smoothing and hole filling, may lead to shrinkage of the model and accuracy loss. To preserve the original depth information inferred from the human body surface, we decide to address these artifacts during the step of estimating the point clouds of human body shapes, and meanwhile, the matching ambiguity in low-texture regions can be decreased. First, image preprocessing is applied to detect and blur the body hair regions in the captured images for the subsequent estimation process. Second, a local smoothness constraint is introduced to the basic model to smooth the marked regions affected by body hair. Finally, a hierarchical framework is constructed, and

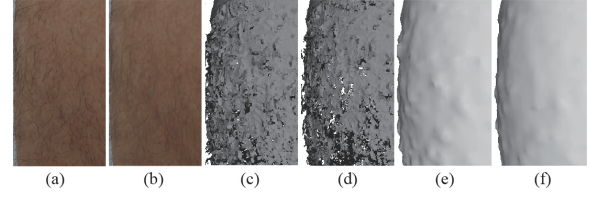


Figure 5: (a) Captured image. (b) Image after local blurring. The results of point clouds and meshes using the original image (c, e) and the blurred image (d, f), respectively.

the estimation of an image with a coarser scale is leveraged to further decrease the negative impact of body hair and to alleviate the matching ambiguity in low-texture skin regions.

(a) Image Preprocessing

Instead of directly utilizing the captured images as inputs, image preprocessing, including body hair detection and blurring, is applied to decrease the disturbing induced by body hair in the subsequent reconstructed shape. First, we detect the skin regions of the captured human images by a human parsing network [WSC*20] that is retrained using augmented data with four labels (background, head, skin, and clothes). After the human parsing stage, the skin regions of the body are labeled and segmented for subsequent hair detection. Second, we detect body hair with a series of image filters that are similar to those in previous work [BBN*12]. We transfer the captured RGB images to HSV space. To estimate the orientation of each hair pixel, the real part of a Gabor filter kernel K_α is used for convolution with the S and V channels of the images to produce a score for body hair oriented along the α direction. The orientation map $O(x, y) = \tilde{\alpha}$ is calculated by finding the best orientation $\tilde{\alpha}$ that yields the highest score $F(x, y) = |K_{\tilde{\alpha}} * V|_{(x,y)} + |K_{\tilde{\alpha}} * S|_{(x,y)}$ at pixel (x, y) among 18 different directions (one every 10 degrees). Aiming at obtaining a distinguishable hair mask, we apply a non-maximum suppression strategy [Can86] to suppress the artifacts in which the scores are not the local maxima in the direction orthogonal to the orientation. Hysteresis thresholding in [Can86] and morphology operations (an opening operation followed by a closed operation) are used to judge and enhance the edges of the hair mask. Finally, we invert the hair mask and employ a median filter to eliminate the detected pores for the purpose of obtaining an accurate hair mask M_0 .

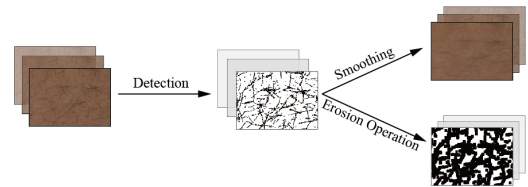


Figure 6: Overview of the image preprocessing approach.

Owing to the wide baseline between two views in a multi-view stereo system, large amounts of surrounding skin pixels are affected by hair occlusion. Removing these influenced regions via the guidance of the reconstruction mask may lead to abundant missing data

and incomplete results. Thus, it is not ideal to merely mask out these influenced regions to acquire satisfactory results. Similar to previous work [BBN*12], we first choose a gentle strategy that blurs the detected hair regions via an anisotropic Gaussian filter to improve the reconstruction quality to some extent. The orientation of the Gaussian filter is given by the orientation map O . The masked pixels in M_0 are blurred to a high spatial extent, and the skin regions not occluded by body hair are not changed. The improved results after blurring are given in Figure 5. Then, we employ an erosion operation with a 15×15 kernel, which is sufficient, to M_0 to indicate the surrounding regions affected by body hair. The acquired binary mask M_1 is introduced for guiding the subsequent depth estimation to process to give a smoothness constraint in the masked regions. A review of image preprocessing is shown in Figure 6.

(b) Optimization with A Local Smoothness Constraint

To further decrease the impact of body hair on the surrounding skin regions, a local smoothness constraint guided by the binary mask M_1 is introduced to the optimization of the basic model. Body hair pixels interfere with the correct local matching of surrounding skin regions due to their high discrimination in a fixed patch window. Conversely, under a global view, hair pixels, for which the ratio in the captured images is lower, become less distinguishable during the matching process. The negative impact of hair pixels can be decreased with a global method, but such a method is time-consuming and may smooth out other high-frequency details. Based on this observation, we only focus on the masked regions in M_1 that are affected by hair pixels, crop a local image patch χ_l^h for each pixel l^h in these regions, and introduce a smoothness constraint to decrease the local negative impact of body hair.

Dense matching is performed by minimizing the following energy functions consisting of a data term φ and a smoothness term ϕ for the image pixels affected by hair pixels:

$$E = \sum_{l^s} \varphi(l^s, \mathbf{v}_l) + \sum_{l^h} \varphi(l^h, \mathbf{v}_l) + \sum_{l^h} \sum_{r \in N_{l^h}} \phi(l^h, r, \mathbf{v}_l, \mathbf{v}_r), \quad (4)$$

where l^s denotes an unmasked pixel in M_1 , and N_{l^h} is the neighborhood set of the masked pixel l^h . The data term φ computes the local similarity between the reference patch and the corresponding source patch for the label $\mathbf{v}_l = (\theta_l, \mathbf{n}_l)$ of each pixel $l = \{l^s, l^h\}$. The smoothness term ϕ forces the planes around l^h to change smoothly and suppresses the tilt of hair planes. These two terms are defined as follows:

$$\varphi(l, \mathbf{v}_l) = \frac{1}{|S|} \sum_{m \in S} \xi_l^m(\theta_l^*, \mathbf{n}_l^*), \quad (5)$$

$$\phi(l^h, r, \mathbf{v}_l, \mathbf{v}_r) = \lambda(1 - \zeta_{l^h r}), \quad (6)$$

where the data term φ is obtained from Equation (1), and λ is a constant regularizer. We define $\zeta_{l^h r} = \exp(-\frac{d(p_{l^h}, f_r) + d(p_r, f_{l^h})}{2\sigma})$ where $d(p_{l^h}, f_r)$ is the distance between the reconstructed point p_{l^h} at l^h and the tangent plane f_r at r , and $d(p_r, f_{l^h})$ has a similar definition, respectively. The constant σ is preset to 0.1, which provides a strict constraint for making the two planes closer considering the actual human body size. Note that we only calculate the smoothness term for l^h in the local image patch χ_l^h , thus other regions maintain high accuracy and retain high-frequency details.

Although the introduced smoothness term for l^h constrains the skin surfaces around the hair to be smooth, the additional computation cost for optimization over a continuous space is huge. Similar to [BRFK14], we transfer the calculation of the smoothness term over the continuous \mathbf{v}_r to a computation over a finite set. We define the optimal hypothesis set $H_l = \left\{ \left(\hat{\theta}_l^{opt(i)}, \hat{\mathbf{n}}_l^{opt(i)} \right) \right\}_{i=1}^K$, which is first randomly initialized, correspondingly to each pixel l . To maintain the parallel computation of the basic model, we consider integrating an optimizer for solving Equation (4) into the PatchMatch sequential propagation procedure. We observe that a 1D optimizer, dynamic programming (DP), has a sweep and update scheme that is similar to that of sequential propagation along the scanline, but that suffers from stripe artifacts. Thus, an improved DP algorithm integrated with the winner-take-all (WTA) results is proposed to acquire the optimal hypothesis set for the marked regions in M_1 . Equation (4) can be minimized in parallel along each row/column as follows:

$$\left\{ \left(\hat{\theta}_l^{opt(i)}, \hat{\mathbf{n}}_l^{opt(i)} \right) \right\}_{i=1}^K = \begin{cases} \arg \min_{\theta_l^*, \mathbf{n}_l^*} \frac{1}{|S|} \sum_{m \in S} \xi_l^m(\theta_l^*, \mathbf{n}_l^*), & if l = l^s \\ \arg \min_{\theta_l^*, \mathbf{n}_l^*} M(l, \theta_l^*, \mathbf{n}_l^*), & if l = l^h, \end{cases} \quad (7)$$

$$M(l, \theta_l^*, \mathbf{n}_l^*) = \frac{1}{|S|} \sum_{m \in S} \xi_l^m(\theta_l^*, \mathbf{n}_l^*) + \min_{(\hat{\theta}_{l-1}, \hat{\mathbf{n}}_{l-1}) \in H'_{l-1}} [M(l-1, \hat{\theta}_{l-1}, \hat{\mathbf{n}}_{l-1}) + \lambda(1 - \zeta_{l(l-1)})], \quad (8)$$

where $H'_l = \{H_l, (\hat{\theta}_l^{opt(0)}, \hat{\mathbf{n}}_l^{opt(0)})\}$ is the modified hypothesis label set, and the optimal pair $(\hat{\theta}_l^{opt(0)}, \hat{\mathbf{n}}_l^{opt(0)})$ estimated by the WTA method is introduced to H'_l and the candidate set for alleviating stripe artifacts. The candidate set of the pair $(\theta_l^*, \mathbf{n}_l^*)$ in Equation (3) is modified as follows:

$$\left\{ \left\{ \left(\theta_l^{(i)}, \mathbf{n}_l^{(i)} \right), \left(\theta_{l-1}^{pr(i)}, \mathbf{n}_{l-1}^{pr(i)} \right), \left(\theta_l^{md(i)}, \mathbf{n}_l^{md(i)} \right), \left(\theta_l^{(i)}, \mathbf{n}_l^{md(i)} \right), \left(\theta_l^{nd(i)}, \mathbf{n}_l^{nd(i)} \right), \left(\theta_l^{pr(i)}, \mathbf{n}_l^{pr(i)} \right), \left(\theta_l^{(i)}, \mathbf{n}_l^{pr(i)} \right) \right\}_{i=1}^K, \left(\hat{\theta}_l^{opt(0)}, \hat{\mathbf{n}}_l^{opt(0)} \right) \right\}. \quad (9)$$

During the optimization stage, as shown in Equation (7), we apply a divide-and-conquer strategy for the different pixels classified by the binary mask M_1 . For the skin pixels not affected by hair, we follow the WTA strategy by utilizing the bilateral NCC described in Section 3.3.1 to acquire the optimal hypothesis. For the pixels affected by hair, the smoothness constraint is introduced in the local patch χ_l^h and optimized by the efficient DP approach, which does

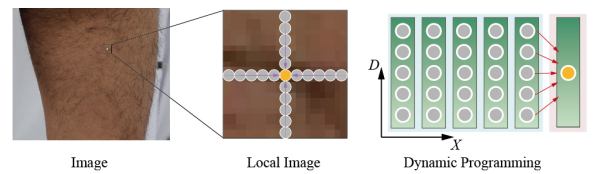


Figure 7: Optimization with the local smoothness constraint. The yellow circle indicates the current processing pixel which is affected by body hair. A local image patch is first cropped. During each sweep (purple arrows) of the sequential propagation procedure, the DP algorithm is applied to estimate the depth and normal of the current pixel.

not break the framework of parallel computation. This is shown in Figure 7. Thus, this strategy seldom changes the integral surface accuracy, recovers the natural smoothness of skin around body hair, and improves the precision of the algorithm for the target skin surface with body hair. Finally, these obtained K best pairs of each l are sorted by energy cost and the rank 1st pair is set to $(\hat{\theta}_l^{opt}, \hat{\mathbf{n}}_l^{opt})$.

In the end, the proposed local smoothness constraint decreases the negative impact of body hair and improves the quality of reconstruction in skin regions with body hair while maintaining high-accuracy results in other regions.

(c) Hierarchical Framework

Although the introduced local smoothness constraint can largely smooth out the rough skin surfaces caused by the impact of body hair, it is still difficult in cases where the subjects have thick body hair. In addition, the challenge of matching ambiguity in low-texture skin regions still cannot be handled due to the local matching in the PatchMatch MVS method. We observe that, when an image is downsampled, the image patch of skin under the same patch window becomes more discriminative, while the body hair within the patch is less distinguishable. This is because the richness of the skin texture is increased at a coarser scale with a more global view. Thus, both the matching ambiguity in low-texture skin regions and the negative impact of body hair can be alleviated by performing estimation at a coarser scale. Similar to previous works [XT19,LFYX19], we construct a hierarchical framework for utilizing coarser estimations to ease the matching problem in low-texture regions and the interference of skin caused by body hair.

As shown in Figure 1, a five-layer image pyramid is established with a downscaling factor of $\epsilon = 0.5$ for all images to fully utilize the coarser estimations. Random sampling with PatchMatch is used to initialize the hypothesis for the layer at the highest level. For other layers, the initialized hypothesis is set as the optimal hypothesis of each pixel upsampled via a joint bilateral upsampler [KCLU07] at the upper level. Thus, reliable estimations of the low-texture regions and the skin around body hair at the coarsest scale are propagated to the finest scale. We notice that as the propagation proceeds from the coarser scale, the optimization at the finer scale is accelerated and quickly converges. Based on this observation, in the first stage which involves photometric estimation, we set the number of iterations l^{S_5} at the coarsest scale to 5 and l^{S_i} at each scale S_i is decreased scale by scale until $l^{S_1} = 1$ at the finest scale. To obtain more accurate results, the second stage with geometric consistency is applied twice at each layer scale to refine the depth maps and normal maps. Finally, a median filter with a kernel with the same size as the window size is adopted.

The local smoothness constraint is exploited by the last three layers to suppress hair pixels that are discriminative in high-resolution images while retaining the basic accuracy of the first two layers using the basic model. Instead of recomputing the WTA optimal pairs for each pixel of a coarser scale image at the current scale, the coarser estimation $(\hat{\theta}_l^{opt(0)}, \hat{\mathbf{n}}_l^{opt(0)})$ of each pixel at the second layer is upsampled and introduced to the DP optimization approach described in Section 3.3.2(b). For a smooth transition between the masked and unmasked pixels in M_1 , a median filter is utilized after

the optimization process with a local smoothness constraint at each layer.

In this way, both the negative impact of thick body hair and the matching uncertainty in low-texture skin regions are alleviated by the proposed hierarchical framework. In the end, the modified PatchMatch MVS method overcomes the problems regarding matching issues in the self-occluded and low-texture regions, improves the accuracy of the reconstruction results, and solves the challenge of the negative impact of body hair. Finally, the depth map fusion method implemented in [SZFP16] is adopted in our method for obtaining a whole human body shape point cloud.

3.4. Meshing

To further improve accuracy of the acquired point cloud, an outlier removal strategy [KKSZ09] is utilized to filter outliers while keeping the other points fixed. Although the basic model alleviates the self-occlusion problem in most regions, some nearly invisible regions in the images, such as the crotch, the axilla, and the bottom of the foot, are hard to reconstruct. To acquire a watertight mesh, we introduced a hole filling step to fill these missing data. Template-based deformation [ACP03] is applied to these nearly invisible regions that do not affect the personalities of human body shapes. Finally, we generate a high-fidelity human body mesh via screened Poisson surface reconstruction [KH13].

4. Experiments and Discussion

In this section, we first validate the effectiveness of the three individual parts of the proposed system, including the multi-view camera setup, the calibration process, and the dense reconstruction approach. Then, overall quantitative and qualitative evaluations are subsequently performed via a comparison with the state-of-the-art passive methods. Finally, we discuss a parameter analysis, as well as the superiority, and limitations of the proposed system.

Following previous work [RZY*20], the accuracy of the acquired mesh is evaluated by leveraging ground-truth anthropometry measurements obtained using thin sticky measuring tapes attached to the skin. The commonly used anthropometry measurements are

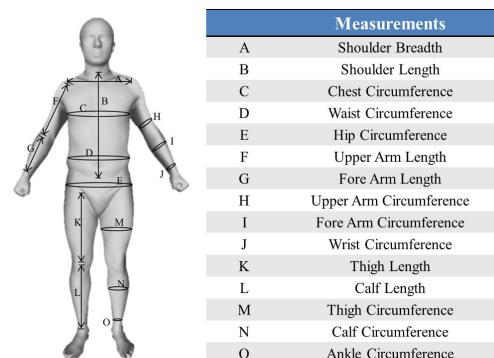
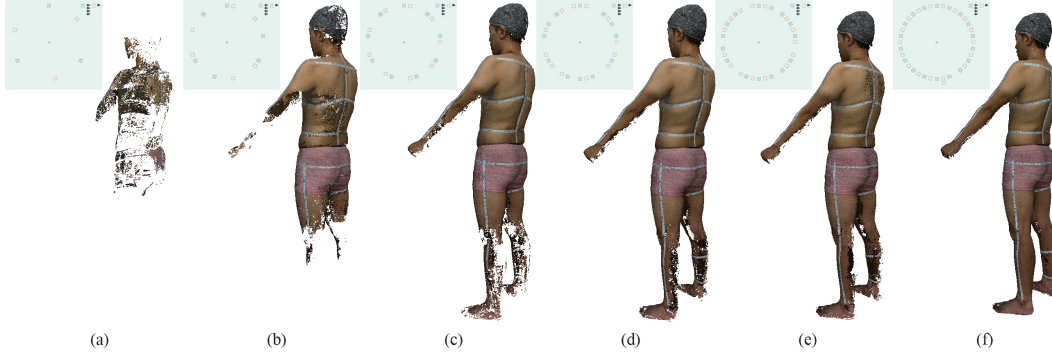


Figure 8: Illustration of anthropometry measurements.

Table 1: Comparisons of point cloud errors of different anthropometry measurements before and after using the refinement in the calibration.

	Method	Measurements														Avr	
		A	B	C	D	E	F	G	H	I	J	K	L	M	N		O
Male	w/o CR	2.49	1.51	0.95	2.74	2.46	1.19	1.25	1.32	0.78	0.93	1.10	0.55	1.95	0.62	0.63	1.36
	Ours	1.11	0.46	0.81	1.16	1.57	0.84	0.89	0.77	0.17	0.47	0.33	0.41	0.41	0.17	0.40	0.66
Female	w/o CR	1.73	1.82	2.67	2.95	2.72	1.35	1.27	1.31	1.71	0.65	0.70	0.59	1.61	1.01	0.66	1.52
	Ours	1.02	0.93	1.43	1.70	1.89	0.93	0.98	0.98	0.94	0.45	0.26	0.04	0.93	0.31	0.40	0.88

**Figure 9:** Results of point clouds using different camera setups. (a) 15 cameras. (b) 30 cameras. (c) 45 cameras. (d) 60 cameras. (e) 75 cameras. (f) 90 cameras.

shown in Figure 8. Different from taking manual distance measurements [RZY*20] of the skin surface along the tape contours, we apply a B-spline curve fitting [CC78] on the mesh and compute the arc length to acquire the reconstructed distance. The absolute differences between the ground truths and the actual measurements reflect the accuracy of the human mesh.

The proposed method is implemented in C++ with CUDA. All experiments are carried out on a PC with an Intel Core i7-6700K CPU at 4GHz, 64 GB of RAM, and two GeForce GTX 1080Ti GPUs.

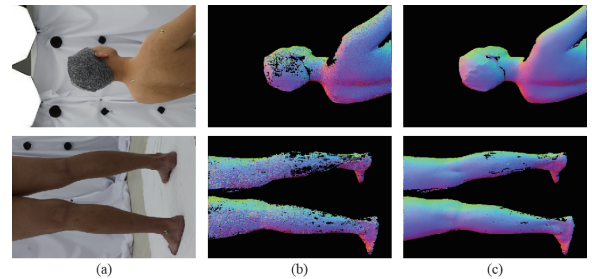
4.1. Validation of the Three Parts

Multi-View Camera Setup. The multi-view camera setup consisting of 90 high-resolution cameras offers adequate views for avoiding incomplete human body shapes caused by self-occlusion issues. To test the effectiveness of the proposed setup, the completeness of the captured point clouds is compared using 6 different camera setups. We first set 3 groups of cameras (4 cameras form a group) that are focused on the main torso and 3 cameras with wide-angle lenses for a wider view as the basic setup. Then, sets of 15 cameras are added incrementally in other setups. These setups and the corresponding obtained point clouds are illustrated in Figure 9. The reconstruction results show that the completeness increases with the increase in the number of cameras and that the proposed multi-view setup performs best.

Calibration. The proposed calibration pipeline using the encoded corners of calibration objects and human features improves the efficiency of calibration and the robustness of the system in terms of accuracy. For efficiency, we test the times of conducting calibration in our system and [RZY*20] which applied checkerboard calibration. To calibrate all cameras, [RZY*20] takes about

120 minutes while our method only about 15 minutes which improves by 8 times. Instead of repeated calibration used in [RZY*20] for each new model, the refinement strategy using human features further saves time and improves calibration efficiency. For accuracy, we reconstruct different point clouds using different camera parameters estimated with or without refinement (w/o CR). As shown in Figure 10(c), the acquired normal maps are finer after the addition of the refinement mechanism that exploits human features. The quantitative results are presented in Table 1. We apply a B-spline curve fitting to the acquired point clouds and compute the mean accuracy by comparing the measurements with the ground truth. It can be observed that the proposed calibration method decreases the anthropometry measurement errors and improves the overall accuracy.

Dense Reconstruction. The presented dense reconstruction method offers a universal and passive solution for high-quality human body point cloud acquisition. To validate the effectiveness

**Figure 10:** (a) Reference image. The results of normal maps before (b) and after (c) using the refinement operation in the calibration pipeline.

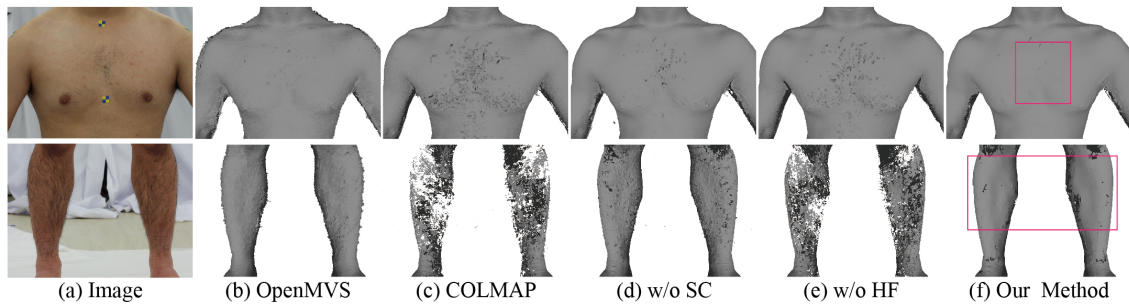


Figure 11: Qualitative point cloud comparison using different MVS methods.

Table 2: Quantitative point cloud comparisons using different MVS methods. The two parts of the point clouds that are seriously and weakly impacted by body hair are denoted as SIR and WIR, respectively.

	OpenMVS		COLMAP		w/o SC		w/o HF		Our Method	
	M	Std	M	Std	M	Std	M	Std	M	Std
SIR	2.24	1.36	1.67	0.73	3.49	0.91	2.70	0.84	4.73	0.67
WIR	1.93	1.32	2.58	1.19	3.21	1.21	3.08	0.99	4.40	1.00
ALL	2.10	1.35	2.10	1.07	3.36	1.07	2.88	0.93	4.57	0.85

of the dense reconstruction method, we choose 30 point clouds, which are divided into 16 and 14 point clouds that are seriously and weakly impacted by body hair, respectively, from 5 subjects and 6 fixed body parts, including the back, breast, abdomen, arm, thigh, and calf, to evaluate the quality of the point clouds. We compare our method with a hierarchical method (OpenMVS) [cdc], the baseline method (COLMAP) [SZFP16], our method without the local smoothness constraint (w/o SC), and our method without the hierarchical framework (w/o HF). The qualitative results are shown in Figure 11. The negative impact of body hair on the skin is eliminated by our method. Without the hierarchical framework, the completeness in low-texture skin regions is decreased. Moreover, we design a user study, that recruits 24 volunteers with professional experience to score the point clouds according to quality (decreasing from 5 to 1). The mean scores (M) and standard deviations (Std) are given in Table 2. The detailed instructions for the participants and the statistics (both mean and standard deviations) of each body part can be found in the supplementary material. Whether the effect of body hair is serious or slight, our method performs best and reconstructs the most realistic skin surface.

4.2. Overall Evaluation

We evaluate our system by comparing it with other state-of-the-art passive systems including two commercial software programs (RealityCapture [RC] and MetaShape [agi]), COLMAP [SF16, SZFP16], and a pipeline consisting of OpenMVG [ope] and OpenMVS [cdc]. Screened Poisson surface reconstruction [KH13] with the same octree depth is applied to all systems. There are several accuracy levels from the lowest to the highest in the steps of the workflow of RealityCapture and MetaShape. The highest accuracy level is set for each step in the evaluation. The comparison results

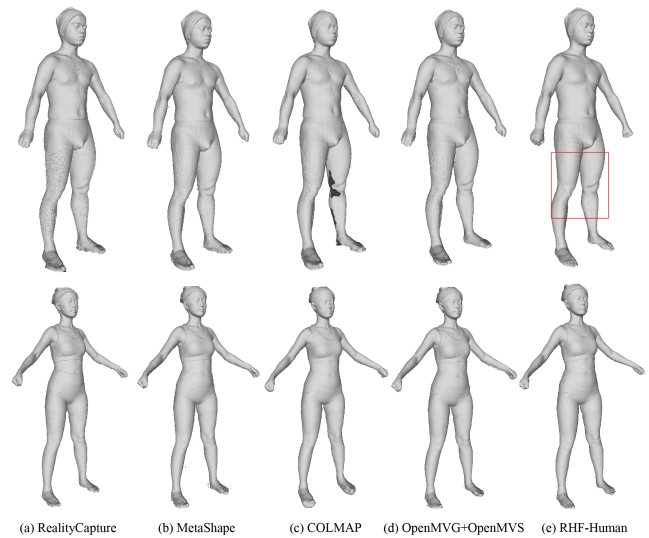


Figure 12: Qualitative mesh comparison of one male and one female using different passive systems.

of one male with more body hair and one female with less body hair are illustrated in Figure 12. We observe that our system decreases the negative impact of body hair on skin reconstruction and performs best. For subjects with less body hair, the performance of the proposed system is comparable to those of other state-of-the-art systems. The quantitative comparison between the accuracy leveraging anthropometry measurements obtained by the methods on meshes is shown in Table 3. The human body shapes acquired from our system are one-to-one in terms of proportion to the real human body with a measurement error of less than 1.5 mm on average, so the proposed method performs best with respect to accuracy. For timings comparisons, our system takes about 4.17 hours for generating a single model while RealityCapture takes about 2.2 hours, MetaShape about 8.15 hours, COLMAP about 4.5 hours, and the pipeline with OpenMVG and OpenMVS about 2.75 hours. All the methods are executed on the same PC. The run time of our system ranks 3rd and is acceptable for high-accuracy human body shape reconstruction with 90 cameras.

We capture 57 subjects with different genders, poses, and skin

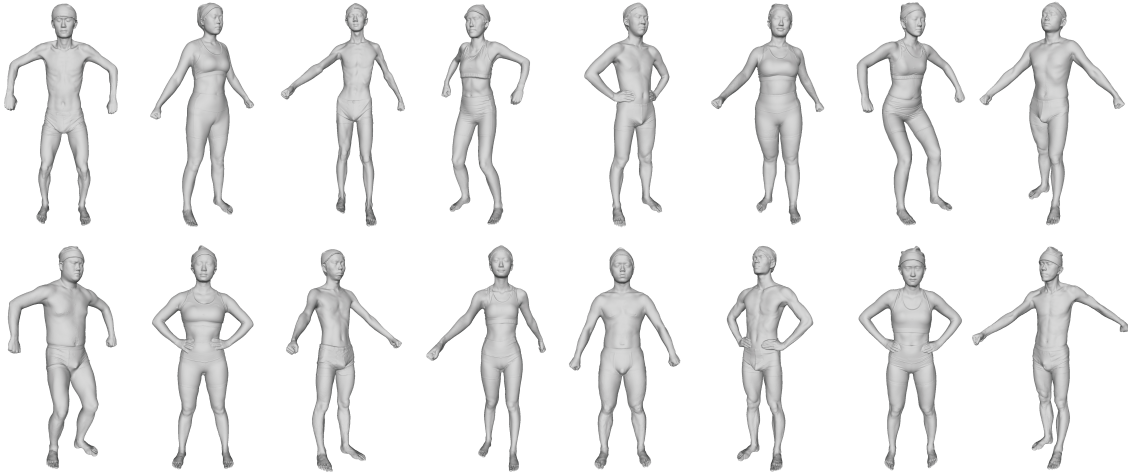


Figure 13: Mesh results obtained using the proposed system for various skin with different amounts of body hair, genders, and poses.

Table 3: Comparisons between the reconstruction errors of different anthropometry measurements. RC, MS, COL, OMV, RHF denote RealityCapture, MetaShape, COLMAP, and a pipeline consisting of OpenMVG and OpenMVS, the proposed system, respectively.

M	Male					Female				
	RC	MS	COL	OMV	RHF	RC	MS	COL	OMV	RHF
A	8.96	8.92	7.38	6.89	2.24	6.31	7.10	5.93	8.34	1.83
B	5.42	2.35	7.51	6.06	1.23	3.52	7.06	3.71	6.04	0.32
C	8.27	10.05	14.17	18.56	1.43	10.66	10.29	14.41	26.16	1.22
D	10.97	8.20	15.48	17.87	1.69	5.78	8.42	6.62	8.15	1.79
E	9.53	8.69	19.74	22.76	2.36	8.51	7.28	20.89	14.60	2.5
F	2.73	3.49	5.67	8.06	0.91	3.93	2.51	5.09	6.41	1.02
G	3.27	2.94	5.31	5.03	1.06	1.54	2.31	5.68	4.87	0.23
H	3.03	3.22	5.32	10.06	1.33	2.16	4.01	7.82	8.33	1.14
I	2.43	4.17	9.11	8.26	0.78	3.98	4.37	5.58	13.85	1.01
J	4.75	2.59	4.79	5.74	0.75	1.89	2.30	7.46	5.10	0.99
K	1.72	5.00	10.59	8.01	1.16	3.56	1.46	3.73	5.11	2.17
L	4.19	4.31	9.89	9.43	0.43	2.30	4.21	6.60	5.51	0.63
M	8.46	8.53	7.47	8.74	0.61	5.25	3.80	8.93	8.16	0.54
N	3.73	3.74	9.36	9.16	0.7	4.62	5.01	8.73	5.61	0.15
O	1.84	4.82	7.64	10.48	0.86	3.92	4.03	10.92	9.15	1.24
Avr	5.29	5.40	9.30	10.34	1.17	4.53	4.95	8.14	9.02	1.12

with different amounts of body hair and construct a human body shape dataset with 227 watertight meshes. The results of the whole dataset can be found in the supplementary material. Partially reconstructed human body shapes are shown in Figure 13. The number of vertices in the acquired human body shapes is approximately 700 thousand, which is finer than most human body shapes obtained from the model-based methods and passive learning-based methods. It can be validated that the proposed system provides a general solution for handling different subjects with different genders, poses, and skin with different amounts of body hair.

4.3. Discussion

Parameter Analysis. To decrease the impact of the introduced smoothness constraint on the unmasked regions in M_1 , the size of the local image patch χ_i^h for each masked pixel l^h is set to a small radius of 5, which is the same as the matching window ra-

Table 4: The execution time of each step in the pipeline.

	Calibration	Dense Reconstruction	Meshing	All
Time (min.)	15	220	15	250

dius. We observe that a higher λ coefficient in the smoothness constraint yields a greater strength for smoothing the rough surface affected by body hair. Thus, we utilize a strong smoothness constraint with $\lambda = 15$. Similar to [BRFK14], using more hypotheses yields a converged solution with slightly lower energy but increases the computational complexity. In practice, we find that choosing $K = 3$ provides satisfactory results. As such, we use this value for all the experiments. All other parameters are the same as the default values adopted in [SF16, SZFP16].

Superiority. The experimental results demonstrate that the proposed system overcomes the weakness of previously developed passive methods and achieves a comparable level of accuracy (within 1.5 mm on average) to that of the mainstream methods. In each component, this system offers a practical and universal solution for reconstructing human body shapes. First, the single-shot capture process alleviates the capture issue with regard to the non-rigid human body. The multi-view setup is easy to adapt for commercial use. Second, the refinement mechanism using human features during calibration can be used in all studio environments to avoid harming the accuracy with slight camera movements and to save time by not repeating the calibration process for each new subject. Finally, the proposed dense reconstruction method achieves excellent performance for different subjects and acquires high-fidelity skin surfaces while excluding the negative impacts of hair, which is beneficial for applications in anthropometry and health care.

Limitations. As for versatility, the human body capture and reconstruction are limited to young East Asians, various skin tones for different ethnicities are lacking. Some methods in our system, i.e., body hair detection, may fail based on the input of other skin

tones. Regarding accuracy, some high-frequency details around body hair (i.e., nipples and the belly button) may be smoothed out because the mask M_1 covers the affected regions, which are not limited by body hair. In terms of efficiency, we count and analyze the total computational time of our system from the moment of acquiring body images to obtaining the reconstructed model finally and find that the most time-consuming step is the dense reconstruction. As Table 4 shown, our system takes about 250 minutes overall but the time of dense reconstruction takes over 80%. Although the hierarchical framework accelerates the optimization process, the introduction of a local smoothness constraint increases the computational complexity of the algorithm. We will consider a more elaborate optimization strategy to further improve the performance of our approach.

5. Conclusions

The proposed multi-view passive system for human body shape reconstruction overcomes the main challenges encountered by passive-vision methods, including calibration accuracy and stereo matching in self-occluded and low-texture skin regions. An integrated and self-correcting model is provided as a universal solution for high-precision human body shape acquisition, as well as to lay a foundation for practical applications and a new generation of passive-vision systems.

In the future, we will conduct in-depth research in the following aspects: For versatility, we will capture a wider scope of human body shapes with various skin tones and will find a universal solution to reconstruct these models; For efficiency, we will explore a more efficient optimizer to further improve the performance and extend the method to broader applications, such as dynamic human body shape reconstruction. Meanwhile, we will also explore the powerful 3D geometry processing techniques, which can effectively improve the quality of output generated by a simplified multi-view system; For textured models, subsequent research on texture mapping technique in our system will be conducted to recover the texture of human body shapes.

Acknowledgement. This work was jointly supported by the National Natural Science Foundation of China under Grants Nos. 61732015, 61932018 and 61472349.

References

- [ACP03] ALLEN B., CURLLESS B., POPOVIC Z.: The space of human body shapes: reconstruction and parameterization from range scans. *ACM Trans. Graph.* 22, 3 (2003), 587–594. doi:10.1145/882262.882311. 2, 7
- [agi] Metashape. <https://www.agisoft.com/>. 9
- [AMB*19] ALLDIECK T., MAGNOR M. A., BHATNAGAR B. L., THEOBALT C., PONS-MOLL G.: Learning to reconstruct people in clothing from a single RGB camera. In *Proc. CVPR '19* (2019), pp. 1175–1186. doi:10.1109/CVPR.2019.00127. 2
- [ASK*05] ANGUELOV D., SRINIVASAN P., KOLLER D., THRUN S., RODGERS J., DAVIS J.: SCAPE: shape completion and animation of people. *ACM Trans. Graph.* 24, 3 (2005), 408–416. doi:10.1145/1073204.1073207. 2
- [AWLB17] ACHENBACH J., WALTERMATE T., LATOSCHIK M. E., BOTSCH M.: Fast generation of realistic virtual humans. In *Proc. VRST '17* (2017), ACM, pp. 12:1–12:10. doi:10.1145/3139131.3139154. 2
- [BBB*10] BEELER T., BICKEL B., BEARDSLEY P. A., SUMNER B., GROSS M. H.: High-quality single-shot capture of facial geometry. *ACM Trans. Graph.* 29, 4 (2010), 40:1–40:9. doi:10.1145/1778765.1778777. 1, 3
- [BBLR15] BOGO F., BLACK M. J., LOPER M., ROMERO J.: Detailed full-body reconstructions of moving people from monocular RGB-D sequences. In *Proc. ICCV '15* (2015), pp. 2300–2308. doi:10.1109/ICCV.2015.265. 1
- [BBN*12] BEELER T., BICKEL B., NORIS G., BEARDSLEY P. A., MARSCHNER S., SUMNER R. W., GROSS M. H.: Coupled 3d reconstruction of sparse facial hair and skin. *ACM Trans. Graph.* 31, 4 (2012), 117:1–117:10. doi:10.1145/2185520.2185613. 5, 6
- [BKL*16] BOGO F., KANAZAWA A., LASSNER C., GEHLER P. V., ROMERO J., BLACK M. J.: Keep it SMPL: automatic estimation of 3d human pose and shape from a single image. In *Proc. ECCV '16* (2016), vol. 9909, pp. 561–578. doi:10.1007/978-3-319-46454-1_34. 2
- [BRFK14] BESSE F., ROTHER C., FITZGIBBON A. W., KAUTZ J.: PMBP: patchmatch belief propagation for correspondence field estimation. *Int. J. Comput. Vis.* 110, 1 (2014), 2–13. doi:10.1007/s11263-013-0653-9. 6, 10
- [BRLB14] BOGO F., ROMERO J., LOPER M., BLACK M. J.: FAUST: dataset and evaluation for 3d mesh registration. In *Proc. CVPR '14* (2014), pp. 3794–3801. doi:10.1109/CVPR.2014.491. 2
- [BSB*07] BALAN A. O., SIGAL L., BLACK M. J., DAVIS J. E., HAUSSECKER H. W.: Detailed human shape and pose from images. In *Proc. CVPR '07* (2007). doi:10.1109/CVPR.2007.383340. 2
- [BSFG09] BARNES C., SHECHTMAN E., FINKELSTEIN A., GOLDMAN D. B.: Patchmatch: a randomized correspondence algorithm for structural image editing. *ACM Trans. Graph.* 28, 3 (2009), 24. doi:10.1145/1531326.1531330. 3
- [Can86] CANNY J. F.: A computational approach to edge detection. *IEEE Trans. Pattern Anal. Mach. Intell.* 8, 6 (1986), 679–698. doi:10.1109/TPAMI.1986.4767851. 5
- [CC78] CATMULL E., CLARK J.: Recursively generated b-spline surfaces on arbitrary topological meshes. *Computer-aided design* 10, 6 (1978), 350–355. 8
- [CCS*15] COLLET A., CHUANG M., SWEENEY P., GILLET T. S., EVSEEV D., CALABRESE D., HOPPE H., KIRK A. G., SULLIVAN S.: High-quality streamable free-viewpoint video. *ACM Trans. Graph.* 34, 4 (2015), 69:1–69:13. doi:10.1145/2766945. 1
- [cdc] Openmvs. <https://github.com/cdcseacave/openMVS>. 9
- [CVHC08] CAMPBELL N. D. F., VOGIATZIS G., HERNÁNDEZ C., CIPOLLA R.: Using multiple hypotheses to improve depth-maps for multi-view stereo. In *Proc. ECCV '08* (2008), vol. 5302, pp. 766–779. doi:10.1007/978-3-540-88682-2_58. 3
- [DS15] DONG J., SOATTO S.: Domain-size pooling in local descriptors: DSP-SIFT. In *Proc. CVPR '15* (2015), pp. 5097–5106. doi:10.1109/CVPR.2015.7299145. 4
- [ES04] ESTEBAN C. H., SCHMITT F.: Silhouette and stereo fusion for 3d object modeling. *Comput. Vis. Image Underst.* 96, 3 (2004), 367–392. doi:10.1016/j.cviu.2004.03.016. 3
- [FH15] FURUKAWA Y., HERNÁNDEZ C.: Multi-view stereo: A tutorial. *Found. Trends Comput. Graph. Vis.* 9, 1-2 (2015), 1–148. doi:10.1561/06000000052. 3
- [FP09] FURUKAWA Y., PONCE J.: Carved visual hulls for image-based modeling. *Int. J. Comput. Vis.* 81, 1 (2009), 53–67. doi:10.1007/s11263-008-0134-8. 3
- [FP10] FURUKAWA Y., PONCE J.: Accurate, dense, and robust multiview stereopsis. *IEEE Trans. Pattern Anal. Mach. Intell.* 32, 8 (2010), 1362–1376. doi:10.1109/TPAMI.2009.161. 3

- [FRS17] FENG A. W., ROSENBERG E. S., SHAPIRO A.: Just-in-time, viable, 3-d avatars from scans. *Comput. Animat. Virtual Worlds* 28, 3-4 (2017), e1769. doi:10.1002/cav.1769. 3
- [GLS15] GALLIANI S., LASINGER K., SCHINDLER K.: Massively parallel multiview stereopsis by surface normal diffusion. In *Proc. ICCV '15* (2015), pp. 873–881. doi:10.1109/ICCV.2015.106. 3
- [GMMM14] GARRIDO-JURADO S., MUÑOZ-SALINAS R., MADRID-CUEVAS F. J., MARÍN-JIMÉNEZ M. J.: Automatic generation and detection of highly reliable fiducial markers under occlusion. *Pattern Recognit.* 47, 6 (2014), 2280–2292. doi:10.1016/j.patcog.2014.01.005. 3
- [GVCH18] GILBERT A., VOLINO M., COLLOMOSSE J. P., HILTON A.: Volumetric performance capture from minimal camera viewpoints. In *Proc. ECCV '18* (2018), vol. 11215, pp. 591–607. doi:10.1007/978-3-030-01252-6_35. 3
- [GWBB09] GUAN P., WEISS A., BALAN A. O., BLACK M. J.: Estimating human shape and pose from a single image. In *Proc. ICCV '09* (2009), pp. 1381–1388. doi:10.1109/ICCV.2009.5459300. 1, 2
- [HLC*18] HUANG Z., LI T., CHEN W., ZHAO Y., XING J., LEGENDRE C., LUO L., MA C., LI H.: Deep volumetric video from very sparse multi-view performance capture. In *Proc. ECCV '18* (2018), vol. 11220, pp. 351–369. doi:10.1007/978-3-030-01270-0_21. 3
- [Hua17] HUANG Y.: Towards accurate marker-less human shape and pose estimation over time. In *Proc. 3DV '17* (2017), pp. 421–430. doi:10.1109/3DV.2017.00055. 2
- [JLT*15] JOO H., LIU H., TAN L., GUI L., NABBE B. C., MATTHEWS I. A., KANADE T., NOBUHARA S., SHEIKH Y.: Panoptic studio: A massively multiview system for social motion capture. In *Proc. ICCV '15* (2015), pp. 3334–3342. doi:10.1109/ICCV.2015.381. 3
- [KBJM18] KANAZAWA A., BLACK M. J., JACOBS D. W., MALIK J.: End-to-end recovery of human shape and pose. In *Proc. CVPR '18* (2018), pp. 7122–7131. doi:10.1109/CVPR.2018.00744. 1, 2
- [KCLU07] KOPF J., COHEN M. F., LISCHINSKI D., UYTENDAELE M.: Joint bilateral upsampling. *ACM Trans. Graph.* 26, 3 (2007), 96. doi:10.1145/1276377.1276497. 7
- [KH13] KAZHDAN M. M., HOPPE H.: Screened poisson surface reconstruction. *ACM Trans. Graph.* 32, 3 (2013), 29:1–29:13. doi:10.1145/2487228.2487237. 7, 9
- [KKSZ09] KRIEGEL H., KRÖGER P., SCHUBERT E., ZIMEK A.: Loop: local outlier probabilities. In *Proc. CIKM '09* (2009), pp. 1649–1652. doi:10.1145/1645953.1646195. 7
- [KPZK17] KNAPITSCH A., PARK J., ZHOU Q., KOLTUN V.: Tanks and temples: benchmarking large-scale scene reconstruction. *ACM Trans. Graph.* 36, 4 (2017), 78:1–78:13. doi:10.1145/3072959.3073599. 1
- [LCK*21] LIU Z., CAO Y., KUANG Z., KOBELT L., HU S.: High-quality textured 3d shape reconstruction with cascaded fully convolutional networks. *IEEE Trans. Vis. Comput. Graph.* 27, 1 (2021), 83–97. doi:10.1109/TVCG.2019.2937300. 1, 2
- [LDX10] LIU Y., DAI Q., XU W.: A point-cloud-based multiview stereo algorithm for free-viewpoint video. *IEEE Trans. Vis. Comput. Graph.* 16, 3 (2010), 407–418. doi:10.1109/TVCG.2009.88. 3
- [LFB17] LEROY V., FRANCO J., BOYER E.: Multi-view dynamic shape refinement using local temporal integration. In *Proc. ICCV '17* (2017), IEEE Computer Society, pp. 3113–3122. doi:10.1109/ICCV.2017.336. 1
- [LFB18] LEROY V., FRANCO J., BOYER E.: Shape reconstruction using volume sweeping and learned photoconsistency. In *Proc. ECCV '18* (2018), vol. 11213, pp. 796–811. doi:10.1007/978-3-030-01240-3_48. 3
- [LFYX19] LIAO J., FU Y., YAN Q., XIAO C.: Pyramid multi-view stereo with local consistency. *Comput. Graph. Forum* 38, 7 (2019), 335–346. doi:10.1111/cgf.13841. 3, 7
- [LMR*15] LOPER M., MAHMOOD N., ROMERO J., PONS-MOLL G., BLACK M. J.: SMPL: a skinned multi-person linear model. *ACM Trans. Graph.* 34, 6 (2015), 248:1–248:16. doi:10.1145/2816795.2818013. 2
- [LMS16] LIU A. J., MARSCHNER S., SNAVELY N.: Caliber: Camera localization and calibration using rigidity constraints. *Int. J. Comput. Vis.* 118, 1 (2016), 1–21. doi:10.1007/s11263-015-0866-1. 3
- [LQ05] LHUILLIER M., QUAN L.: A quasi-dense approach to surface reconstruction from uncalibrated images. *IEEE Trans. Pattern Anal. Mach. Intell.* 27, 3 (2005), 418–433. doi:10.1109/TPAMI.2005.44. 3
- [LRK*17] LASSNER C., ROMERO J., KIEFEL M., BOGO F., BLACK M. J., GEHLER P. V.: Unite the people: Closing the loop between 3d and 2d human representations. In *Proc. CVPR '17* (2017), pp. 4704–4713. doi:10.1109/CVPR.2017.500. 2
- [LVG*13] LI H., VOUGA E., GUDYM A., LUO L., BARRON J. T., GUSEV G.: 3d self-portraits. *ACM Trans. Graph.* 32, 6 (2013), 187:1–187:9. doi:10.1145/2508363.2508407. 2
- [MBR*00] MATUSIK W., BUEHLER C., RASKAR R., GORTLER S. J., McMILLAN L.: Image-based visual hulls. In *Proc. SIGGRAPH '00* (2000), pp. 369–374. doi:10.1145/344779.344951. 3
- [NH98] NEAL R. M., HINTON G. E.: A view of the em algorithm that justifies incremental, sparse, and other variants. In *Learning in Graphical Models*, vol. 89, 1998, pp. 355–368. doi:10.1007/978-94-011-5014-9_12. 4
- [NSH*19] NATSUME R., SAITO S., HUANG Z., CHEN W., MA C., LI H., MORISHIMA S.: Siclope: Silhouette-based clothed people. In *Proc. CVPR '19* (2019), pp. 4480–4490. doi:10.1109/CVPR.2019.00461. 3
- [ope] Openmvg. <https://github.com/openMVG>. 9
- [PRMB15] PONS-MOLL G., ROMERO J., MAHMOOD N., BLACK M. J.: Dyna: a model of dynamic human shape in motion. *ACM Trans. Graph.* 34, 4 (2015), 120:1–120:14. doi:10.1145/2766993.1, 2
- [RC] Realitycapture. <https://www.capturingreality.com/>. 9
- [Rem04] REMONDINO F.: 3-d reconstruction of static human body shape from image sequence. *Comput. Vis. Image Underst.* 93, 1 (2004), 65–85. doi:10.1016/j.cviu.2003.08.006. 3
- [RM19] ROMANONI A., MATTEUCCI M.: TAPA-MVS: textureless-aware patchmatch multi-view stereo. In *Proc. ICCV '19* (2019), pp. 10412–10421. doi:10.1109/ICCV.2019.01051. 3
- [RZY*20] RAN Q., ZHOU K., YANG Y., KANG J., ZHU L., TANG Y., FENG J.: High-precision human body acquisition via multi-view binocular stereopsis. *Comput. Graph.* 87 (2020), 43–61. doi:10.1016/j.cag.2020.01.003. 1, 3, 4, 7, 8
- [SCD*06] SEITZ S. M., CURLESS B., DIEBEL J., SCHARSTEIN D., SZELISKI R.: A comparison and evaluation of multi-view stereo reconstruction algorithms. In *Proc. CVPR '06* (2006). doi:10.1109/CVPR.2006.19. 3
- [SF16] SCHÖNBERGER J. L., FRAHM J.: Structure-from-motion revisited. In *Proc. CVPR '16* (2016), pp. 4104–4113. doi:10.1109/CVPR.2016.445. 3, 4, 9, 10
- [SH07] STARCK J., HILTON A.: Surface capture for performance-based animation. *IEEE Computer Graphics and Applications* 27, 3 (2007), 21–31. doi:10.1109/MCG.2007.68. 3
- [SHN*19] SAITO S., HUANG Z., NATSUME R., MORISHIMA S., LI H., KANAZAWA A.: Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proc. ICCV '19* (2019), pp. 2304–2314. doi:10.1109/ICCV.2019.00239. 3
- [SSG*17] SCHÖPS T., SCHÖNBERGER J. L., GALLIANI S., SATTLER T., SCHINDLER K., POLLEFEYS M., GEIGER A.: A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Proc. CVPR '17* (2017), pp. 2538–2547. doi:10.1109/CVPR.2017.272. 1

- [SSSJ20] SAITO S., SIMON T., SARAGIH J. M., JOO H.: Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Proc. CVPR '20* (2020), pp. 81–90. doi:10.1109/CVPR42600.2020.00016.3
- [SZFP16] SCHÖNBERGER J. L., ZHENG E., FRAHM J., POLLEFEYS M.: Pixelwise view selection for unstructured multi-view stereo. In *Proc. ECCV '16* (2016), vol. 9907, pp. 501–518. doi:10.1007/978-3-319-46487-9_31.3,4,5,7,9,10
- [TMHF99] TRIGGS B., MCLAUCHLAN P. F., HARTLEY R. L., FITZGIBBON A. W.: Bundle adjustment - A modern synthesis. In *Proc. ICCV '99* (1999), vol. 1883, pp. 298–372. doi:10.1007/3-540-44480-7_21.4
- [TNM09] TUNG T., NOBUHARA S., MATSUYAMA T.: Complete multi-view reconstruction of dynamic scenes from probabilistic fusion of narrow and wide baseline stereo. In *Proc. ICCV '09* (2009), pp. 1709–1716. doi:10.1109/ICCV.2009.5459384.3
- [TTC*19] TANG S., TAN F., CHENG K., LI Z., ZHU S., TAN P.: A neural network for detailed human depth estimation from a single image. In *Proc. ICCV '19* (2019), IEEE, pp. 7749–7758. doi:10.1109/ICCV.2019.00784.3
- [TZL*12] TONG J., ZHOU J., LIU L., PAN Z., YAN H.: Scanning 3d full human bodies using kinects. *IEEE Trans. Vis. Comput. Graph.* 18, 4 (2012), 643–650. doi:10.1109/TVCG.2012.56.1,2
- [VBMP08] VLASIC D., BARAN I., MATUSIK W., POPOVIC J.: Articulated mesh animation from multi-view silhouettes. *ACM Trans. Graph.* 27, 3 (2008), 97. doi:10.1145/1360612.1360696.3
- [VETC07] VOGIATZIS G., ESTEBAN C. H., TORR P. H. S., CIPOLLA R.: Multiview stereo via volumetric graph-cuts and occlusion robust photo-consistency. *IEEE Trans. Pattern Anal. Mach. Intell.* 29, 12 (2007), 2241–2246. doi:10.1109/TPAMI.2007.70712.3
- [VPB*09] VLASIC D., PEERS P., BARAN I., DEBEVEC P. E., POPOVIC J., RUSINKIEWICZ S., MATUSIK W.: Dynamic shape capture using multi-view photometric stereo. *ACM Trans. Graph.* 28, 5 (2009), 174. doi:10.1145/1618452.1618520.3
- [WHB11] WEISS A., HIRSHBERG D. A., BLACK M. J.: Home 3d body scans from noisy image and range data. In *Proc. ICCV '11* (2011), pp. 1951–1958. doi:10.1109/ICCV.2011.6126465.2
- [WLDW11] WU C., LIU Y., DAI Q., WILBURN B.: Fusing multiview and photometric stereo for 3d reconstruction under uncalibrated illumination. *IEEE Trans. Vis. Comput. Graph.* 17, 8 (2011), 1082–1095. doi:10.1109/TVCG.2010.224.3
- [WSC*20] WANG J., SUN K., CHENG T., JIANG B., DENG C., ZHAO Y., LIU D., MU Y., TAN M., WANG X., ET AL.: Deep high-resolution representation learning for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* (2020), 1–1. doi:10.1109/TPAMI.2020.2983686.5
- [WVT12] WU C., VARANASI K., THEOBALT C.: Full body performance capture under uncontrolled and varying illumination: A shading-based approach. In *Proc. ECCV '12* (2012), vol. 7575, pp. 757–770. doi:10.1007/978-3-642-33765-9_54.3
- [WWMT11] WU C., WILBURN B., MATSUSHITA Y., THEOBALT C.: High-quality shape from multi-view stereo and shading under general illumination. In *Proc. CVPR '11* (2011), pp. 969–976. doi:10.1109/CVPR.2011.5995388.3
- [XLS*20] XU Z., LIU Y., SHI X., WANG Y., ZHENG Y.: MARMVS: matching ambiguity reduced multiple view stereo for efficient large scale scene reconstruction. In *Proc. CVPR '20* (2020), pp. 5980–5989. doi:10.1109/CVPR42600.2020.00602.3
- [XT19] XU Q., TAO W.: Multi-scale geometric consistency guided multi-view stereo. In *Proc. CVPR '19* (2019), pp. 5483–5492. doi:10.1109/CVPR.2019.00563.3,7
- [XT20] XU Q., TAO W.: Planar prior assisted patchmatch multi-view stereo. In *Proc. AAAI '20* (2020), pp. 12516–12523. doi:10.1609/aaai.v34i07.6940.3
- [YFHW16] YANG J., FRANCO J., HÉTROU-WHEELER F., WUHRER S.: Estimation of human body shape in motion with wide clothing. In *Proc. ECCV '16* (2016), vol. 9908, pp. 439–454. doi:10.1007/978-3-319-46493-0_27.2
- [YLL*18] YAO Y., LUO Z., LI S., FANG T., QUAN L.: Mvsnet: Depth inference for unstructured multi-view stereo. In *Proc. ECCV '18* (2018), vol. 11212, pp. 785–801. doi:10.1007/978-3-030-01237-3_47.3
- [YWK20] YAN S., WIRTA J., KÄMÄRÄINEN J.: Anthropometric clothing measurements from 3d body scans. *Mach. Vis. Appl.* 31, 1-2 (2020), 7. doi:10.1007/s00138-019-01054-4.2
- [ZDJF14] ZHENG E., DUNN E., JOJIC V., FRAHM J.: Patchmatch based joint view selection and depthmap estimation. In *Proc. CVPR '14* (2014), pp. 1510–1517. doi:10.1109/CVPR.2014.196.4
- [Zha00] ZHANG Z.: A flexible new technique for camera calibration. *IEEE Trans. Pattern Anal. Mach. Intell.* 22, 11 (2000), 1330–1334. doi:10.1109/34.888718.3
- [ZLNW19] ZHAO T., LI S., NGAN K. N., WU F.: 3-d reconstruction of human body shape from a single commodity depth camera. *IEEE Trans. Multim.* 21, 1 (2019), 114–123. doi:10.1109/TMM.2018.2844087.2
- [ZPBP17] ZHANG C., PUJADES S., BLACK M. J., PONS-MOLL G.: Detailed, accurate, human shape estimation from clothed 3d scan sequences. In *Proc. CVPR '17* (2017). doi:10.1109/CVPR.2017.582.2