Fragmentation Guided Human Shape Reconstruction

Article *in* IEEE Access · March 2019

DOI: 10.1109/ACCESS.2019.2905879

4 authors, including:

Yingliang Zhang
ShanghaiTech University
10 PUBLICATIONS **30** CITATIONS
SEE PROFILE

Wei Yang
University of Delaware
15 PUBLICATIONS **69** CITATIONS
SEE PROFILE

Jingyi Yu
Michigan State University
141 PUBLICATIONS **2,514** CITATIONS
SEE PROFILE

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier XX.XXXX/ACCESS.201X.DOI

Fragmentation Guided Human Shape Reconstruction

YINGLIANG ZHANG[1,2,3], (Student Member, IEEE), XI LUO[1, 4], WEI YANG[4], AND JINGYI YU[1] (Member, IEEE)

[1]School of Information Science and Technology, ShanghaiTech University, Shanghai 201210, China
[2]Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, Shanghai 200050, China
[3]University of Chinese Academy of Sciences, Beijing 100049, China
[4]DGene Inc.

Corresponding author: Jingyi Yu (e-mail: yujingyi@shanghaitech.edu.cn).

This work is partially supported by the programs of STCSM (17XD1402900, 17JC1403800, 17511108201, 17511105805, 2015F0203-000-06), of SHMEC (2019-01-07-00-01-E00003), and of SHEITC (2018-RGZN-01011).

ABSTRACT We present a novel semantic-driven multi-view reconstruction technique for producing realistic 3D human models. Our approach borrows the fragmentation concept in Cubism style painting where human body is decomposed into semantically meaningful fragments for conveying space and movement. We first employ deep learning based skeleton estimation for warping a proxy human model under the canonical pose to the target multi-view input. It also conducts 3D fragment labeling on the warped model to separate different human body parts. Finally, we utilize the normal, depth, and fragment label of the proxy model as priors in the multi-view stereo reconstruction process. Comprehensive experiments have shown that our reconstruction technique outperforms the state-of-the-art methods in robustness and accuracy, especially near occlusion boundaries and on textureless regions. In particular, it manages to significantly reduce the "adhesive" artifacts commonly observed in MVS that incorrectly stitches different body parts.

INDEX TERMS 3D Reconstruction, human shape, semantic analysis, multi-view stereo.

I. INTRODUCTION

There has been an emerging demand on building high quality 3D models of real human. Applications are numerous, ranging from traditional motion analysis [1], to visual special effects [2], and to photorealistic avatars in virtual environments [3]. Reconstructing 3D human models, however, is inherently more challenging than the classic 3D object: human models are non-rigid and exhibit dynamic and complex occlusions between different body parts. More importantly, we observe human so frequently that slight visual artifacts significantly degrade visual realism of the final reconstruction.

Active sensing techniques by far can only produce low-resolution depth maps and subsequently low resolution and noisy 3D models [4]. Passive techniques such as multi-view stereo (MVS) rely heavily on the presence of textures. The lack of texture and the presence of repetitive textures can both lead to matching errors, as shown in Fig.8. MVS is particularly sensitive to occlusions: the occluded parts can produce outliers due to incorrect correspondence matching. For humans, even with the simplest anatomical model - head, neck, trunk, upper and lower limbs, a slight pose change can lead to drastically different occlusion patterns, e.g., a swinging arm can either occlude or be occluded by the torso, with or without touching the torso.

Occlusions not only cause incomplete reconstruction but also adhesive artifacts. Recall MVS recovers essentially 3D point clouds which need to be converted to meshes, e.g., using marching cubes [5]. Consequently, the occluder would be forced to stitch with the occludee (Fig.7) in the recovered mesh. Reconstruction artifacts are particularly severe near the occlusion boundaries and by far neither active nor passive techniques can effectively handle adhesive artifacts.

In this paper, we present a novel technique tailored for human body reconstruction that we call fragmentation-based reconstruction or FBR. The term fragmentation originates from Cubism style painting where artists such as Braque and Picasso decomposed human body into semantically meaningful fragments as a way of getting closest to the subject and conveying space and movement. Fig.I shows an extreme

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2905879, IEEE Access

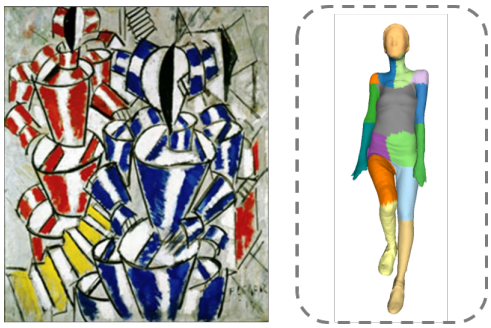


FIGURE 1. Left: One of the best known Cubism fragmentation is Fernand Léger's "Exit the Ballets Russes" where human body is decomposed to conic components. Right: We aims to employ 3D fragmentation for human body reconstruction.

example of Léger's "Exit the Ballets Russes" where the human form is rendered in fragments of cylinders and cones.

In a nutshell, FBR exploits semantic meaning of individual reconstruction components. While most existing semantic labeling is conducted on 2D images, fragmentation targets at 3D points. In this paper, we define fragmentation as the process of assigning a specific (semantic) label to every 3D point on human geometry. The availability of fragmentation benefits 3D reconstruction tasks in multiple aspects. First, it provides informative priors on the spatial configuration and thereby occlusions, to allow us to determine visible vs. occluded regions and adopt tailored solutions. Second, it mitigates the "adhesion" problem: we can enforce the occluder and the occludee be separated in the color matching step. Finally, fragmentation can accelerate the reconstruction process by significantly narrowing down the search space for correspondence matching.

Our FBR uses multi-view inputs captured by a dome composed of 80 DSLR cameras. We develop a learning-based technique to first compute the 3D skeleton and then warp a "canonical" human model captured under the neural pose onto the current pose. We partition the human body into 16 fragments according to the geodesic distance between points on the body and each 3D skeleton point. The canonical model also provides an initial estimate of the depth and normal as well as the fragment label. We present a simple but effective process to refine the 3D skeleton and then set out to conduct 3D reconstruction with the initial fragmentation as priors. Comprehensive experiments show that our technique significantly outperforms the state-of-the-art multi-view stereo approaches and improves reconstruction quality especially near the occlusion boundaries.

II. RELATED WORK

Most related to our work are the shape reconstruction works deployed for 3D human scanning.

a: Shape Reconstruction.

The literature on shape reconstruction is vast, and we only discuss the most relevant ones on human shape reconstruction. The past decade, there has been increasing attention on real human appearance and movement reconstruction. Most existing solutions can be categorized as passive vs. active reconstruction. The former uses multi-view camera systems [6] [7] and rely on correspondence matching, triangulation, and stereo matching. The latter uses active sensors [4], [8], [9] based on either structured light or time-of-flight range scanning and employs point cloud fusion methods. The most notable passive reconstruction methods are the streams of work by the CMU groups. Kanade [10] and Narayanan [11] strategically place 51 cameras in space and use multi-view stereo reconstruction to reconstruct 3D video sequence of captured person. Shape-from-silhouette based approaches [12] usually require more cameras to produce high-quality geometry and cannot recover concavities. More recent approaches [13]–[15] use 4D space-time representation and conduct high-dimensional shape reconstruction.

On the active sensing front, Newcombe et al. [4] use a single RGB-D sensor and estimate a dense volumetric 6D motion field that warps the geometry into a live frame. [6], [9] fuse input from space apart depth sensors and generate temporally coherent geometry. These methods generate coarser geometry compared with multi-view camera system but can achieve real-time performance. It is important to note that both passive and active solutions suffer from what Kanade calls "the curse of more cameras": it may be counter-intuitive that more cameras (or depth sensors) may produce worse results. The reason behind it is that the calibration errors of the cameras can lead to severe space carving artifacts [16]: thin objects may be completely carved out due to slight miscalibration. In the active sensing case, these structures cannot be reliably fused and as results exhibit strong noise.

Our approach aims to leverage recent advances on human pose estimation and semantic human parts labeling. The former aims to approximate 2D or 3D poses of a single or multi persons from images. Recent successful solutions unanimously employ the learning-based techniques leveraging the availability of abundant training data [17]. The latter attempts to assign a meaningful human part label to each pixel in the image. Techniques such as the DeepLab models [18] make use of the fully connected pairwise Conditional Random Field (CRF) as an auxiliary post-processing step to refine the initial estimation. Overall human parts labeling is more challenging than pose estimation, as illustrated in the performance in robustness and accuracy on benchmark data.

There is also a trend on combining pose estimation and human parts labeling. Fang et al. proposed a regional multi-person pose estimation (RMPE) framework to improve 2D pose detection accuracy [19]. Cao et al. proposed a greedy bottom-up step to detect 2D poses using Part Affinity Fields (PAFs) [1]. These methods work on single images and perform well when images record entire bodies. To handle the pose estimation problem under the multi-view setup, Joo et al. built a Panoptic Studio system and proposed a skeletal representation to optimize with trajectory reassociation [20]. Our fragmentation scheme is closely related to these works, except that our goal is to resolve a volumetric labeling prob-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2905879, IEEE Access

lem. In particular, under the multi-view setup, our fragmentation labeling technique aims to produce semantic human part labels consist across the views.

III. PROXY GENERATION

The first step of our approach is to obtain a 3D proxy human model that we will later use for occlusion prediction and refinement. In our implementation, we capture a proxy human model under the "T" or "A" pose. This is because such poses exhibit very few occlusions under our dome camera settings: we evenly distribute the cameras over the spherical dome; with sufficient number of views (in our case, 80), standard multi-view reconstruction such as MVS can produce high quality reconstruction on even slightly occluded regions. Alternatively, one can generate the 3D proxy model by rotating a depth sensor around the body, e.g., via approaches such as dynamic fusion [4].

Next we conduct 3D skeleton estimation on the proxy model. Commercial tracking systems, e.g., OptiTrack, can produce highly accurate 3D skeletons but are expensive and require extra calibration equipment such as markers or special patterned clothes. We instead aim to develop an effective solution by fusing 2D skeletons from images captured from different viewpoints.

We first employ the learning-based technique, Alphapose [19]. Alphapose builds upon the Symmetric Spatial Transformer Network, Parametric Pose NonMaximum-Suppression and Pose-Guided Proposals Generator to detect poses and has shown great success on the MPII dataset [21]. For each 2D image, it estimates 16 image coordinates and their corresponding confidence scores. However, Alphapose assumes the image covers complete or nearly complete human body. In reality, our dome system uses cameras with a small field-of-view (FOV) so as to capture fine details of texture and geometry. This imposes significant challenges to Alphapose. Fig. 2(a) show the failure cases of 2D pose estimation.

To resolve the problem, we resort to multi-view reconstruction, in particular, the bundle-adjustment (BA) used in SfM. Our dome system can be pre-calibrated to obtain known and accurate camera intrinsics and extrinsics. We use the linear triangulation method to estimate the 3D skeleton coordinates [22]. To take into account the confidence scores from Alphapose, we integrate the scores into the bundle adjustment (BA) step to refine the 3D joint points and filter out the 2D estimations with low confidence. Specifically, given 2D images $X^1, X^2, ..., X^n$ with their corresponding estimated poses, we use $S_{X^i}(j) = (u_{ij}, v_{ij}, c_{ij}), i = 1, 2, ..., n, j = 1, 2, ..., 16$ to represent the skeleton point j on the image X^i, where u_{ij}, v_{ij} are the normalized image coordinates and c_{ij} is the confidence score. We modify classical BA as:

$$min \sum_{i=1}^{n} \sum_{j=1}^{16} c_{ij} \| M_{X^i} O_j - o_{ij} \| \quad (1)$$

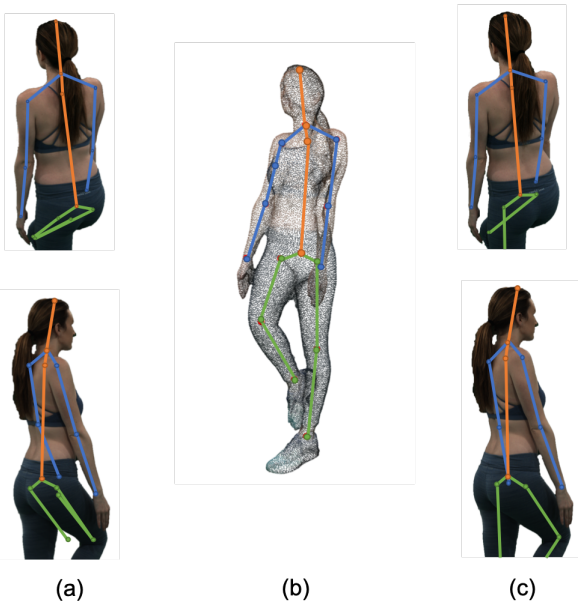


FIGURE 2. (a) Skeleton generated by Alphapose [19] in a single 2D image may be inaccurate due to occlusions or image cut-off. (b) We integrate skeleton estimations from all views to obtain a reliable 3D skeleton. (c) shows the reprojected 3D skeleton on the corresponding view using our approach.

where $o_{ij} = (u_{ij}, v_{ij})$ represents the 2D skeleton point on image X^i corresponding to a 3D pose point O_j, M_{X^i} is the projection matrix of X^i. We use the Levenberg-Marquardt algorithm to solve for above equation. Fig.2(b) show some sample 3D poses recovered using our approach.

Next, we conduct vertex skinning to attach each vertex to its corresponding bones, where each bone is composed of two skeleton points. There exist several skinning algorithms, such as Linear Blend Skinning(LBS), Dual Quaternion Blend Skinning(DQBS) [23], and most recently learning-based SMPL [24]. In our pipeline, we choose DQBS method for its simplicity. We compute the weights of each vertex to all bones using the Geodesic Voxel method [25].

IV. PROXY WARPING

Now that assume we capture a different set of images of the model under drastically different poses. Instead of directly recovering the geometry, we first warp the proxy model onto the target 3D skeleton under the new pose. We conduct the same process as Sec.III to obtain an initial estimate of the 3D skeleton of the target model.

To warp the model, we set up a hierarchy transformation tree for the extracted 3D bones, e.g., the transformation of parent bones should be applied to the children bones. Moreover, the bones embedded in the mesh have a fixed length. Therefore, when we warp each bone to a target position, we only need to compute rotation relative to its parent bone. Assume the initial direction of a bone is $\vec{a}$, and the target direction of the bone is $\vec{b}$, we determine the rotation axis $\vec{A_{ab}}$ and rotation angle α from these two directions as:

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2905879, IEEE Access

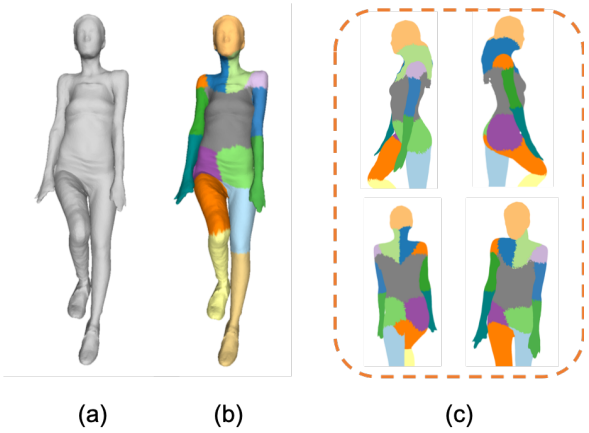


FIGURE 3. (a) shows the warped 3D canonical model. (b) shows the fragment labeling result with 16 labels. (c) projects the 3D model onto respective views to show 2D fragment labels.

$$\vec{A_{ab}} = \vec{a} \times \vec{b}, \quad \alpha = \arccos\left(\frac{\vec{a} \cdot \vec{b}}{\|\vec{a}\| \cdot \|\vec{b}\|}\right) \tag{2}$$

Once we compute the transformation of each bone using Eqn.2, we determine the final transformation for each vertex by combining the related bone transformations and the corresponding weights. We convert the transformation matrix into the dual-quaternion representation via a linear model as:

$$D\vec{Q}_i = \frac{\sum_{k \in \vec{B}} w_k \hat{\vec{dq}_k}}{\| \sum_{k \in \vec{B}} w_k \hat{\vec{dq}_k} \|} \tag{3}$$

where $D\vec{Q}_i$ is the final transformation in dual-quaternion format for i_{th} vertex, $\vec{B}$ is the set contains all bones, $\hat{\vec{dq}_k}$ represents the current transformation of k_{th} bone and w_k is the corresponding influence weight to the vertex. Eqn.3 can be viewed as the normalization operator of the dual-quaternion. Since each vertex on the proxy mesh has been attached to a specific set of bones through our skinning process, we can apply the corresponding sets of transformations derived above to each vertex to obtain an initial warping.

V. FRAGMENTATION LABELING

Next, we conduct fragmentation labeling on the warped 3D proxy model. Notice that the fragmentation process is different from skinning: the former assigns a unique fragmentation label to each vertex whereas the latter assigns multiple (weight-blended) bone segments to each vertex.

The problem is closely correlated to semantic labeling. Significant advances have been achieved on human body labeling on 2D images, e.g., learning-based [26] [27], superpixels-based [28], and graph-based [29] methods. However, such 2D labeling schemes are not directly applicable under our multi-view settings: individually labeling each input view does not guarantee label consistencies across views. We instead devise a 3D labeling scheme on the warped model and then project the labeled results onto each viewpoint. Such an approach not only guarantees consistency but also provides cues for consistent reconstruction as shown in Sec.VI-B.

Given 16 3D skeleton points (regarded as joints) and the warped model in previous steps, our goal is to find the relationships between each vertex on the model and the joints. Moreover, the warped model provides beneficial information for semantic segmentation that there exists neither adhesion nor holes on the warped model. This allows us to use the geodesic distance [30] to find the relationships and k nearest neighbors (KNN) to label each vertex. Compared with Euclidean distance, geodesic distance maintains robustness under non-rigid deformations and has shown great success on shape modeling [31].

Specifically, we first build a 3D net graph by treating the warped model as a 3D net graph: every vertex as a node and the corresponding edges as the weight. Next, we find the nearest nodes to represent joints. Since the joints (3D skeleton points) are not on the surface of the model, we first find 10 nearest nodes of each joint to represent it. Next, we compute the geodesic distance by setting the nearest nodes as seed nodes. Finally, we obtain the labeled model based on the geodesic distances. Specifically, we use KNN(k = 3) to find candidates and then vote for each node's final label by the candidates.

Fig.3 shows our 3D segmentation and the corresponding reprojection 2D results. Our method can generate 16 labels. Compared to the current human parts segmentation on 2D images [32], our method generates more consistent and accurate segmentation in our dome setting. The availability of fragmentation benefits 3D reconstruction by providing extra priors on the spatial configuration and thereby occlusions and helps us solve the "adhesion" problem.

VI. FRAGMENTATION-BASED RECONSTRUCTION

A. PRIORS FROM 3D PROXY

Finally, we use the warped proxy model and its fragmentation labels to conduct multi-view shape reconstruction. Given the prior 3D proxy with semantic labeling from Sec.V and the calibrated camera parameters of a specific view, we project the model onto the view to obtain the depth, normal and label maps. This can be easily achieved by reusing the rendering pipeline, e.g., under OpenGL. Due to slight misalignment on the warped model, the projected maps may still be inconsistent with the actual ones, e.g., if we compute the mask of the target model within each view. Therefore, we conduct a filtering process similar to the guided filter [33]. The final depth, normal, and label maps serve as priors for our reconstruction.

B. FRAGMENTATION-BASED MULTI-VIEW STEREO

Recent works [34], [35] of joint semantic estimation and surface reconstruction utilize the implicit volumetric representation and cast the semantic guided reconstruction to a volumetric labeling problem. Instead, we exploit the semantic prior at the multi-view stereo stage. We show our

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2905879, IEEE Access

approach provides higher accuracy and better completeness and scalability, and lower memory overhead compared to the volumetric method. Specifically, the warped 3D proxy provides priors on the spatial configuration and thereby occlusions, to allow us to determine visible vs. occluded regions and adopt tailored solutions.

To fully incorporate the fragmentation prior, we leverage the optimization framework proposed by [36], [37] for joint depth map estimation and semantic labeling. Our method sets out to jointly estimate the depth θ_l, normal n_l and semantic labels s_l for each pixel l in a reference image X^r given a set of source images $X^m, m = 1...M$ with known camera calibration parameters. We adopt a similar graphical model proposed by Zheng et.al. [36] and Schonberger et.al. [37], where we set the semantic label $\boldsymbol{S}$ as hidden variables, the images $\boldsymbol{X}$ as observations, depth $\boldsymbol{\theta}$ and normal $\boldsymbol{N}$ are model parameters. Solving for this problem equates to computing the maximum of the posterior probability $P(\boldsymbol{S}, \boldsymbol{\theta}, \boldsymbol{N}|\boldsymbol{X})$.

To solve the problem, we can first compute the joint probability

$$\begin{aligned} P(\boldsymbol{X}, \boldsymbol{S}, \boldsymbol{\theta}, \boldsymbol{N}) = \prod_{m=1}^{M} \Big\{ \prod_{l=1}^{L} P(x_l^m|s_{l,t}^m, \theta_{l,t}, n_{l,t}) \cdot \\ \prod_{l=2}^{L} P(s_{l,t}^m|s_{l-1,t}^m, s_{l,t-1}^m) \Big\} \\ \underbrace{\prod_{m=1}^{M} P(s_1^m)}_{\mathbb{F}} \cdot \underbrace{\prod_{l=1}^{L} P(\theta_l, n_l)}_{\mathbb{G}} \end{aligned} \tag{4}$$

and then normalizing over $P(\boldsymbol{X})$. The spatial and temporal smoothness term $P(s_{l,t}^m|s_{l-1,t}^m, s_{l,t-1}^m)$ in Eqn. 4 enforce pairwise smoothness and reduce temporal oscillation during optimization on semantic labels, which we define as:

$$P(s_{l,t}^m|s_{l-1,t}^m, s_{l,t-1}^m) = P(s_{l,t}^m|s_{l-1,t}^m) \cdot P(s_{l,t}^m|s_{l,t-1}^m) \tag{5}$$

The spatial smoothness term has form:

$$P(s_{l,t}^m|s_{l-1,t}^m) = a_s(s_{l,t}^m, s_{l-1,t}^m) \tag{6}$$

where $a_s()$ is a predefined function to adjust the transition probability between different labels. We prefer to set the current label the same as its neighbors for spatially smoothness. And the transition probability between two adjacent labels in 3D mesh should be higher than those that are not. In this paper, we set $\{a_t, b_t\}$ to $\{0.7, 0.2\}$ if $s_l^m = s_{l-1}^m$; to $\{0.5, -0.3\}$ if $s_l^m \in \mathcal{N}(s_{l-1}^m)$, and to $\{0.3, -0.2\}$ otherwise. $\mathcal{N}(S)$ defines the set of all labels have connections to the label S. The final probability is normalized over each label S. The temporal smoothness term depends on optimization iteration t.

$$P(s_{l,t}^m|s_{l,t-1}^m) = [a_t(s_{l,t}^m, s_{l-1,t}^m) + b_t(s_{l,t}^m, s_{l-1,t}^m)\frac{t}{T}] * P_s(s_{l,t-1}) \tag{7}$$

$P_s(s_{l,t-1})$ is the prior computed from geodesic distance.

The emission probability $P(x_l^m|s_l^m, \theta_l, n_l)$ in Eqn. 4 is defined as

$$P(x_l^m|s_l^m, \theta_l, n_l) = \begin{cases} \frac{1}{NA} \exp(-\frac{[1-\rho_l^m(\theta_l, n_l)]^2}{2\sigma^2}), & \text{if } s_l^m = s_l^r \\ \frac{1}{N} u, & \text{otherwise} \end{cases} \tag{8}$$

where $A = \int_{-1}^{1} \exp(-\frac{(1-\rho)^2}{2\sigma^2}) d\rho$ and N is a constant. u is the uniform distribution. ρ_l^m is the semantic adaption of bilaterally weighted normalized cross correlation (NCC) which describes the photometric consistence between the reference patch ω^r and source patch ω^m at pixel l:

$$\rho_l^m = \frac{\mathbf{cov_s}(\omega^r, \omega^w)}{[\mathbf{cov_s}(\omega^r, \omega^w) \cdot \mathbf{cov_s}(\omega^r, \omega^w)]^{1/2}} \tag{9}$$

$\mathbf{cov_s}(x, y) = \mathbf{E_w}(x - \mathbf{E_w}(x))\mathbf{E_w}(y - \mathbf{E_w}(y))$. $\mathbf{E_w} = \sum_i w_i x_i / \sum_i w_i$. When computing NCC, we only include the pixels with the same semantic label with that of the center pixel by setting the weight to 0 when the labels do not match.

We use Zheng's [36] and Colmap's scheme to estimate the semantic label S and depth θ. We fix transition and emission probability and find the optimal semantic labels:

$$q_m^{opt}(s^m) \propto P(s_1^m) \prod_{l=2}^{L} P(s_l^m|s_{l-1}^m) \prod_{l=1}^{L} P(x_l^m|s_l^m, \theta_l^\star, n_l^\star) \tag{10}$$

The probability of $q(S_l^m)$ can be inferred efficiently by the forward-backward algorithm. To find the optimal depth θ, we use the Monte Carlo based method for depth estimation:

$$\theta_l^{opt} = \arg\min_{\theta_l} \frac{1}{|W|} \sum_{m \in W} (1 - \rho_l^m(\theta_l, n_l)) \tag{11}$$

W is a subset of source images sampled from distribution $P_s = q(s_l^m = s^r)/[\sum_{m=1}^{M} q(s_l^m = s_r)]$. q is an approximation of real post prior.

The skeleton-based warping provide useful information, including depth, normal and semantic labels, for reconstruction. We use a Gaussian distribution to describe the geometric prior $\mathbb{G}$ in Eqn. 4 as follows:

$$P(\theta_l, n_l) = P(\theta_l) * P(n_l) = \frac{1}{C} \exp(-\frac{\|\theta_l - \hat{\theta}_l\|^2}{2{\sigma_\theta}^2} - \frac{\|n_l - \hat{n}_l\|^2}{2{\sigma_n}^2}) \tag{12}$$

$\hat{\theta}_l$ and $\hat{n}_l$ are projected depth and normal prior of the wrapped model at pixel l in the reference view. For the projected semantic label, we use it as the initial state of our optimization. Fig.4 compares the recovered depth maps before and after labeling optimization. It shows that the label optimization reduces holes and correcting occlusion boundaries.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2905879, IEEE Access

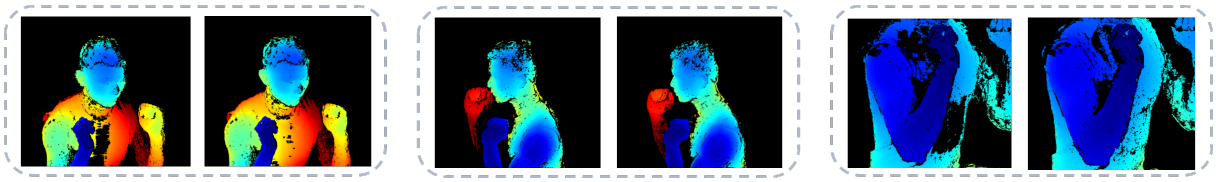


FIGURE 4. Comparisons on the recovered depth maps before (left) and after (right) labeling optimization. Notice how label optimization reduces holes and correcting occlusion boundaries.

FIGURE 5. Our multi-view human capture system uses 80 Canon DSLR cameras mounted on 4 rings surrounding the subject.

FIGURE 6. Sample synthetic data where we render a 3D human model from different viewpoints.

VII. EXPERIMENTS

To evaluate our FBR scheme, we conduct experiments on both synthetic and real data. All experiments are conducted on a computer with a Intel Xeon CPU 32GB memory and one Nvidia 1080ti graphics card. We compare our scheme with several state-of-the-art methods multi-view reconstruction methods such as PMVS [38], OpenMVS [39] and Colmap [37][1].

A. SYNTHETIC DATA

There are several open datasets available for multi-view stereo reconstruction, e.g., the Middlebury multi-view datasets and the new reconstruction datasets from [40]. However, they uniformly target at general scene reconstruction instead of focusing the specific task of human geometry reconstruction. The SMPL dataset from [24] is a public 3d human shape repository but these models do not contain textures that are the key assumption of multi-view reconstruction. We instead construct two new synthetic datasets for evaluating the accuracy and quality our FBR vs. the state-of-the-art. The ground truth consists of pairs of 3D model, one T or A pose and the second under general motion.

We set up a virtual multi-view capture system and render the models onto respective views as RGB images at a resolution of 3000×2000 for emulating a real capture setup. We use 80 virtual cameras with 45 degree field-of-view (FoV) facing towards the target human subject. The spatial distribution of cameras are uniform at four 360 circles which is identical to our real dome system. Moreover, we use diffuse texture and uniform lighting in our rendering. Fig.6 shows some samples of our rendered RGB images for testing.

Fig.7 compares reconstruction quality using our vs. classic algorithms. We include both point clouds and their corresponding Hausdorff distance field. For Hausdorff distance computation, we apply Screen-space Poisson Surface Reconstruction [41] on the recovered point clouds and then measure the distance between the recovered meshes and the ground truth ones. For better illustration, we use color coded error maps where blue, green and red correspond to small, median and large errors.

The top two rows illustrate the reconstruction results on the greeting scene. In this scene, human geometry exhibits few occlusion and therefore classical MVS methods can still produce reasonable results. Nonetheless, our method still outperforms the state-of-the-art especially near occlusion contours and on textureless regions. For example our FBR can better separate the geometry of the legs and the shoes. The bottom two rows compare the results on a more challenging choreographic pose where left arm occludes the torso and the right leg crosses a kneeing left leg. Our approach produces more complete geometry and better preserves silhouettes. This is because our pose analysis is able to predict occlusion patterns and subsequently reduces inaccurate matching near occlusion boundaries (e.g., arm vs. torso) and under similar textures (e.g., the crossing legs).

Table.1 shows the comparison on the time consumption and the vertices number on synthetic data. OpenMVS is the fastest MVS reconstruction. Our framework is developed based on a similar graphical model with [36] and Colmap. The depth, normal and labeling priors from 3D proxy guide our matching scheme in stereo process. And the fragmentation accelerate the searching process for corresponding matching by first matching labeling. Therefore, our algorithm is faster than Colmap.

[1]We use the official PMVS version provided by their research group, the latest OpenMVS from GitHub, and Colmap version 3.5 with CUDA support. We use the setting produces the best reconstruction results.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2905879, IEEE Access

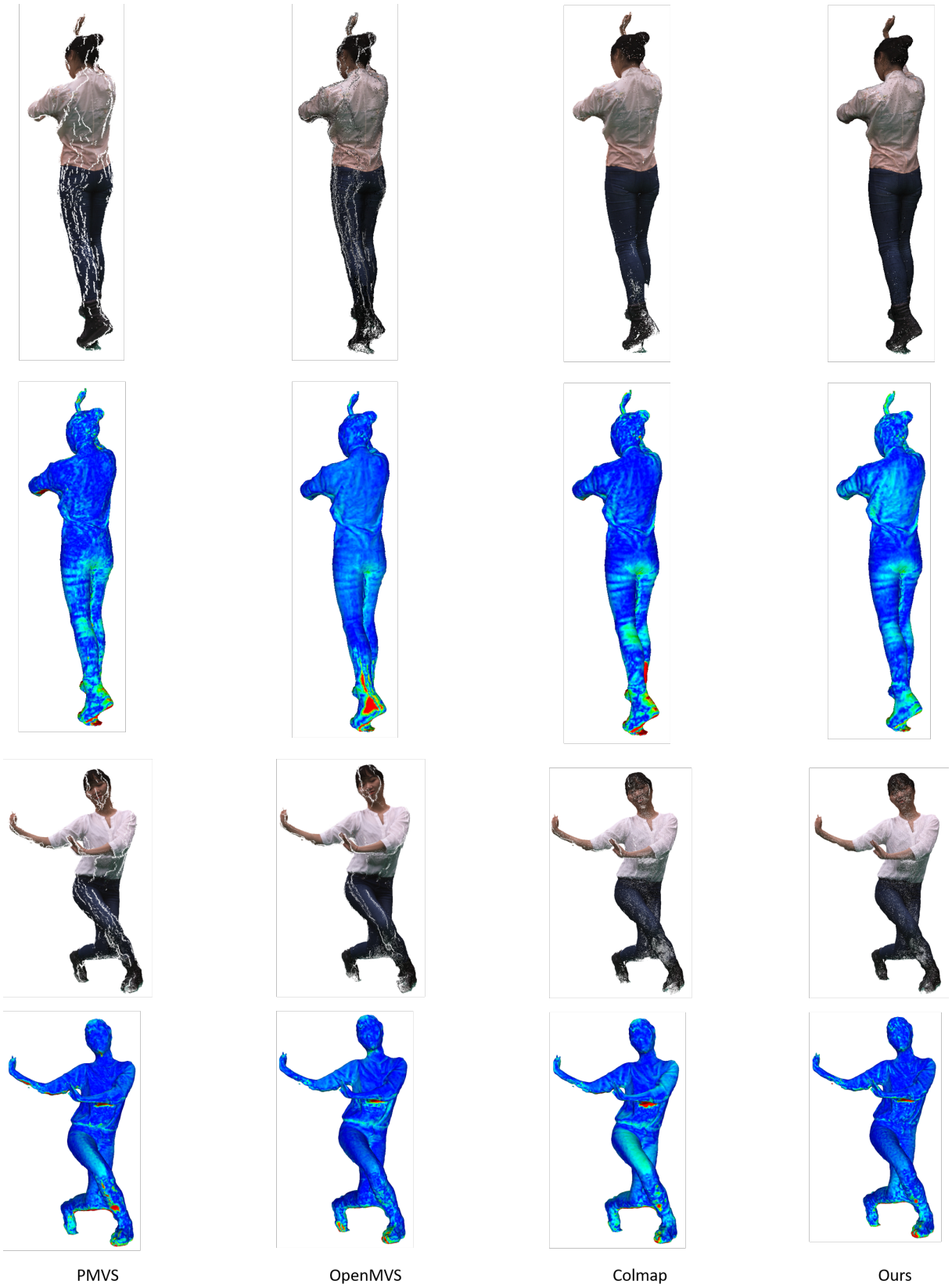


FIGURE 7. Reconstruction comparisons using different methods on a synthetic dataset. The top row shows the recovered raw point cloud and bottom shows the error map using the Hausdorff distance.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2905879, IEEE Access

	PMVS		OpenMVS		Colmap		Ours	
Data	Greeting	Dance1	Greeting	Dance1	Greeting	Dance1	Greeting	Dance1
Vertices	939121	834436	1840697	1219151	410259	389419	420847	419892
Min.	11.36	10.14	2.42	2.04	21.662	21.560	19.187	19.416

TABLE 1. We record the time consumption of all methods and also the vertices number of the result models. Our method is faster than Colmap due to using the guide information from the 3D proxy. And the fragmentation narrows down the search space for corresponding matching. However, OpenMVS is the fastest reconstruction algorithm.

B. REAL DATA

For experiments on real data, we construct a large dome system that consists of 80 Canon 760D DSLR cameras where each camera can capture color images up to a resolution of 6000×4000. We evenly distribute the cameras along 4 parallel circles on a sphere as shown in Fig.5. Such a setup meant to distribute views in order to reduce occlusions. In our setup, we set the camera ISO to 100, shutter speed to $\frac{1}{13}$ and aperture to 11. We use uniformly lighting by using white LED lights. To main color consistencies, we disable auto white balancing. Since we aim to record dynamic data, we use 5 synchronization boxes to sync all 80 cameras. We pre-calibrate the camera intrinsic and extrinsic parameters using the regular SfM method [42] with a calibration target (a highly textured mannequin).

We compare reconstruction results on 3 dynamic motion datasets that exhibit drastically different body movements, i.e., Yoga, Kneeing and Dance2.

Fig.8 compares our reconstruction vs. the state-of-the-art method. For Yoga, the pose is close to A pose and the occlusion is less severe. The main challenge is on the textureless jogger worn by the model. Such textureless regions are particularly problematic to classical multi-view stereo: both PMVS and Colmap reconstructions contain large holes; OpenMVS reconstruction is very noisy. Our method in contrast effectively uses a 3D proxy as guidance. It ensures the completeness of the model and further provides a reliable depth and normal prior in the final reconstruction for individual body fragments.

The kneeing scene is more challenging. It exhibits very complex occlusion patterns and the shorts and the tank top are nearly textureless. PMVS and Colmap is only able to recover a sparse point cloud whereas OpenMVS recovers a slight better geometry with fewer holes although near occlusion boundaries OpenMVS incurs strong noise (e.g., between the torso and the left arm). Our FBR effectively reduces holes in the reconstruction while preserving clear boundaries better different body fragments, illustrating a key benefit of our fragmentation analysis.

The dancing scene exhibits complex occlusion patterns of the two hands as well as between the hands and the torso. Same as the other two scenes, our technique is able to recover much better quality geometry while preserving the occlusion boundaries across different body parts.

Table.2 shows the running time and the vertex number of the corresponding scenes. OpenMVS produces the fastest MVS reconstruction. Our approach is slightly slower but faster than Colmap. This is because the availability of the 3D proxy model narrows down the search range in multi-view stereo, a key factor on accelerating graphical model optimization.

Despite its effectiveness, our FRB technique is still unable to handle extremely textureless models, e.g., models wearing all black clothes, as shown in Fig.9. Our technique is able to recover 3D skeleton of the model but cannot recover the geometry. Even with our proxy model as guidance, the lack of textures across all body parts makes it difficult to reliably establish feature correspondences and hence depth estimation. State-of-the-art solutions also fail to recover 3D shape geometry.

VIII. CONCLUSIONS

We have presented a fragmentation-based multi-view human reconstruction technique. Our technique employs deep learning based skeleton estimation for warping a proxy human model under canonical poses to the target pose. It then conducts fragmentation labeling to separate different components of the human body. We further utilize the normal, depth, and semantic labels of warped proxy model as priors in the multi-view stereo reconstruction process. Comprehensive experiments have shown that our reconstruction technique outperforms the state-of-the-art methods in robustness and accuracy, especially near occlusion boundaries and on textureless regions. In particular, it manages to significantly reduce the adhesive artifacts commonly observed in MVS.

There are multiple future directions we plan to explore. Although our solution can handle dynamic multi-view video sequences, we currently handle each individual frame without considering their temporal coherence. In the future, we plan to add temporal/motion constraints to the reconstruction process to further reduce visual artifacts, e.g., caused by occlusions or textureless regions. Our current canonical T or A pose model is still constructed using multi-view reconstruction and requires users to clean up the data to ensure its quality. An alternative is to use the parametric human body such as SMPL [24] as the "canonical" model. A major limition of our approach is that it cannot handle interactions with objects, as it trains on human body alone. It may be possible to first separate the object from human using similar semantic analysis and then construct individual parts.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2905879, IEEE Access

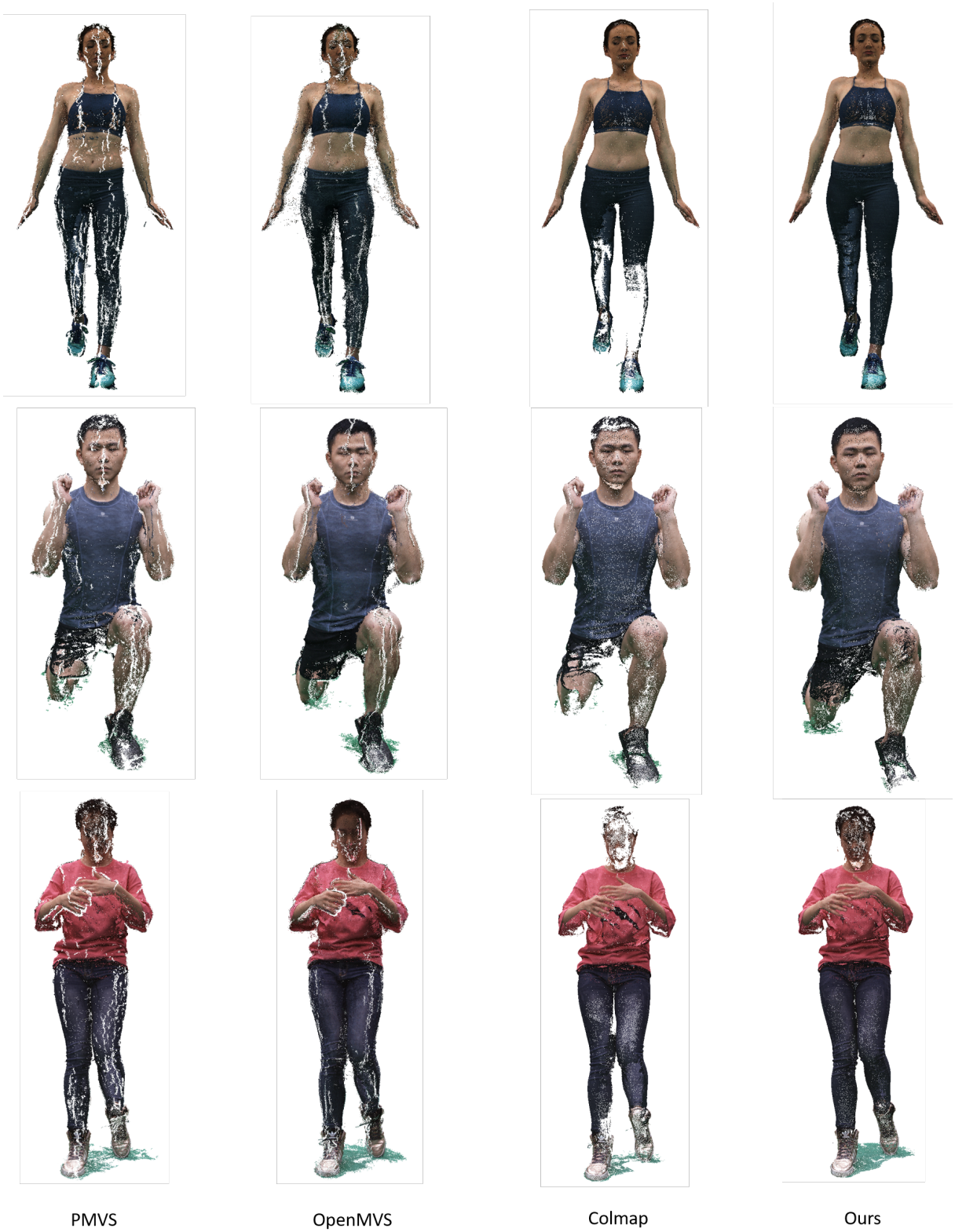


FIGURE 8. Comparisons on the recovered raw 3D point cloud using different methods on real captured data. Notice our results significantly reduce holes and generate much cleaner reconstruction near occlusion edges of different body components.

	PMVS			OpenMVS			Colmap			Ours		
Data	Yoga	Kneeing	Dance2	Yoga	Kneeing	Dance2	Yoga	Kneeing	Dance2	Yoga	Kneeing	Dance2
Vertices	874494	746614	845067	516080	1365245	1975654	673954	513331	553864	776183	811983	723064
Min.	11.23	9.46	12.01	2.42	2.23	3.04	18.19	16.61	21.62	15.76	14.428	19.06

TABLE 2. This table shows the time consumption of all methods and the vertices number of the result models. Compared to Colmap, our method is faster due to our input 3D proxy priors and labeling information. However, OpenMVS is the fastest reconstruction algorithm.

FIGURE 9. A failure case using our technique where the subject wears completely textureless clothes. Our technique can recover 3D skeleton but fails to recover 3D geometry.

REFERENCES

[1] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in CVPR, 2017.

[2] P. Debevec, T. Hawkins, C. Tchou, H.-P. Duiker, W. Sarokin, and M. Sagar, "Acquiring the reflectance field of a human face," in Proceedings of the 27th annual conference on Computer graphics and interactive techniques. ACM Press/Addison-Wesley Publishing Co., 2000, pp. 145–156.

[3] itSeez3D, "Avatar sdk," https://avatarsdk.com/.

[4] R. A. Newcombe, D. Fox, and S. M. Seitz, "Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 343–352.

[5] W. E. Lorensen and H. E. Cline, "Marching cubes: A high resolution 3d surface construction algorithm," in ACM siggraph computer graphics, vol. 21, no. 4. ACM, 1987, pp. 163–169.

[6] M. Dou, S. Khamis, Y. Degtyarev, P. Davidson, S. R. Fanello, A. Kowdle, S. O. Escolano, C. Rhemann, D. Kim, J. Taylor, P. Kohli, V. Tankovich, and S. Izadi, "Fusion4d: Real-time performance capture of challenging scenes," ACM Trans. Graph., vol. 35, no. 4, pp. 114:1–114:13, Jul. 2016. [Online]. Available: http://doi.acm.org/10.1145/2897824.2925969

[7] M. Mikhnevich and P. Hebert, "Shape from silhouette under varying lighting and multi-viewpoints," in 2011 Canadian Conference on Computer and Robot Vision. IEEE, 2011, pp. 285–292.

[8] T. Yu, K. Guo, F. Xu, Y. Dong, Z. Su, J. Zhao, J. Li, Q. Dai, and Y. Liu, "Bodyfusion: Real-time capture of human motion and surface geometry using a single depth camera," in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 910–919.

[9] S. Orts-Escolano, C. Rhemann, S. Fanello, W. Chang, A. Kowdle, Y. Degtyarev, D. Kim, P. L. Davidson, S. Khamis, M. Dou, V. Tankovich, C. Loop, Q. Cai, P. A. Chou, S. Mennicken, J. Valentin, V. Pradeep, S. Wang, S. B. Kang, P. Kohli, Y. Lutchyn, C. Keskin, and S. Izadi, "Holoportation: Virtual 3d teleportation in real-time," in Proceedings of the 29th Annual Symposium on User Interface Software and Technology, ser. UIST '16. New York, NY, USA: ACM, 2016, pp. 741–754. [Online]. Available: http://doi.acm.org/10.1145/2984511.2984517

[10] T. Kanade, P. Rander, and P. Narayanan, "Virtualized reality: Constructing virtual worlds from real scenes," IEEE multimedia, vol. 4, no. 1, pp. 34–47, 1997.

[11] P. Narayanan, P. W. Rander, and T. Kanade, "Constructing virtual worlds using dense stereo," in Sixth International Conference on Computer Vision (IEEE Cat. No. 98CH36271). IEEE, 1998, pp. 3–10.

[12] K. Cheung, S. Baker, and T. Kanade, "Shape-from-silhouette of articulated objects and its use for human body kinematics estimation and motion capture," in 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings., vol. 1. IEEE, 2003, pp. I–I.

[13] J. Süßmuth, M. Winter, and G. Greiner, "Reconstructing animated meshes from time-varying point clouds," in Computer Graphics Forum, vol. 27, no. 5. Wiley Online Library, 2008, pp. 1469–1476.

[14] M. Wand, B. Adams, M. Ovsjanikov, A. Berner, M. Bokeloh, P. Jenke, L. Guibas, H.-P. Seidel, and A. Schilling, "Efficient reconstruction of non-rigid shape and motion from real-time 3d scanner data," ACM Transactions on Graphics (TOG), vol. 28, no. 2, p. 15, 2009.

[15] E. De Aguiar, C. Stoll, C. Theobalt, N. Ahmed, H.-P. Seidel, and S. Thrun, Performance capture from sparse multi-view video. ACM, 2008, vol. 27, no. 3.

[16] K. N. Kutulakos and S. M. Seitz, "A theory of shape by space carving," International journal of computer vision, vol. 38, no. 3, pp. 199–218, 2000.

[17] A. Toshev and C. Szegedy, "Deeppose: Human pose estimation via deep neural networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 1653–1660.

[18] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," arXiv preprint arXiv:1412.7062, 2014.

[19] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "RMPE: Regional multi-person pose estimation," in ICCV, 2017.

[20] L. T. L. G. B. N. I. M. T. K. S. N. Hanbyul Joo, Hao Liu and Y. Sheikh, "Panoptic studio: A massively multiview system for social motion capture," 2015.

[21] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2d human pose estimation: New benchmark and state of the art analysis," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2014.

[22] R. I. Hartley and A. Zisserman, Multiple View Geometry in Computer Vision, 2nd ed. Cambridge University Press, ISBN: 0521540518, 2004.

[23] L. Kavan, S. Collins, C. OâĂŹSullivan, and J. Zara, "Dual quaternions for rigid transformation blending," Trinity College Dublin, Tech. Rep. TCD-CS-2006-46, 2006.

[24] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A skinned multi-person linear model," ACM Trans. Graphics (Proc. SIGGRAPH Asia), vol. 34, no. 6, pp. 248:1–248:16, Oct. 2015.

[25] O. Dionne and M. de Lasa, "Geodesic voxel binding for production character meshes," in Proceedings of the 12th ACM SIGGRAPH/Eurographics Symposium on Computer Animation. ACM, 2013, pp. 173–180.

[26] X. Liang, S. Liu, X. Shen, J. Yang, L. Liu, J. Dong, L. Lin, and S. Yan, "Deep human parsing with active template regression," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 37, no. 12, pp. 2402–2414, 2015.

[27] X. Liang, C. Xu, X. Shen, J. Yang, J. Tang, L. Lin, and S. Yan, "Human parsing with contextualized convolutional neural network," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 1, pp. 115–127, 2017.

[28] Z. Lu, Z. Fu, T. Xiang, P. Han, L. Wang, and X. Gao, "Learning from weak and noisy labels for semantic segmentation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 3, pp. 486–500, 2017.

[29] X. Liang, X. Shen, J. Feng, L. Lin, and S. Yan, "Semantic object parsing with graph lstm," european conference on computer vision, pp. 125–143, 2016.

[30] M. Lanthier, A. Maheshwari, and J. Sack, "Approximating weighted shortest paths on polyhedral surfaces," pp. 485–486, 1997.

[31] M. Dou, P. Davidson, S. R. Fanello, S. Khamis, A. Kowdle, C. Rhemann, V. Tankovich, and S. Izadi, "Motion2fusion: Real-time volumetric performance capture," ACM Trans. Graph., vol. 36, no. 6, pp. 246:1–246:16, Nov. 2017. [Online]. Available: http://doi.acm.org/10.1145/3130800.3130801

[32] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3431–3440, 2015.

[33] K. He, J. Sun, and X. Tang, "Guided image filtering," in Proceedings of the 11th European Conference on Computer Vision: Part I, ser. ECCV'10. Berlin, Heidelberg: Springer-Verlag, 2010, pp. 1–14. [Online]. Available: http://dl.acm.org/citation.cfm?id=1886063.1886065

[34] C. Häne, C. Zach, A. Cohen, and M. Pollefeys, "Dense semantic 3d reconstruction," IEEE transactions on pattern analysis and machine intelligence, vol. 39, no. 9, pp. 1730–1743, 2017.

[35] V. Vineet, O. Miksik, M. Lidegaard, M. Nießner, S. Golodetz, V. A. Prisacariu, O. Kähler, D. W. Murray, S. Izadi, P. Pérez et al., "Incremental dense semantic stereo fusion for large-scale semantic scene reconstruction," in 2015 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2015, pp. 75–82.

[36] E. Zheng, E. Dunn, V. Jojic, and J.-M. Frahm, "Patchmatch based joint view selection and depthmap estimation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1510–1517.

[37] J. L. Schönberger, E. Zheng, J.-M. Frahm, and M. Pollefeys, "Pixelwise view selection for unstructured multi-view stereo," in European Conference on Computer Vision. Springer, 2016, pp. 501–518.

[38] Y. Furukawa and J. Ponce, "Accurate, dense, and robust multiview stereopsis," IEEE transactions on pattern analysis and machine intelligence, vol. 32, no. 8, pp. 1362–1376, 2010.

[39] cdcseacave, "Openmvs," http://cdcseacave.github.io/openMVS/.

[40] T. Schöps, J. L. Schönberger, S. Galliani, T. Sattler, K. Schindler, M. Pollefeys, and A. Geiger, "A Multi-View Stereo Benchmark with High-Resolution Images and Multi-Camera Videos," in Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

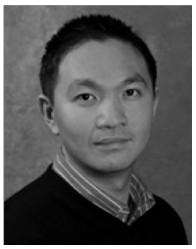
[41] M. Kazhdan and H. Hoppe, "Screened poisson surface reconstruction," ACM Transactions on Graphics (ToG), vol. 32, no. 3, p. 29, 2013.

[42] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4104–4113.

YINGLIANG ZHANG received the BE degree of communication engineering from Ningbo University, China, in 2014. He is pursuing the Ph.D degree of computer science from ShanghaiTech University, China. His research interests include image-based 3D reconstruction, light field rendering, and light field reconstruction.

XI LUO received her BE degree of communication engineering from Shandong University, China, in 2016. Now she is a Ph.D. student of computer science from ShanghaiTech University, China. She joined the DGene. Corp (Prev. Plex-VR) as a research intern in July 2017. Her research interests include computer vision and computer graphics, especially 3D reconstruction and virtual fitting.

WEI YANG received the BEng and MS degrees from the Huazhong University of Science and Technology and Harbin Institute of Technology respectively, and the PhD degree from the University of Delaware (UDel) in Dec 2017. He joined the DGene. Corp (Prev. Plex-VR) as a research scientist in Mar 2018. His research interests include computer vision and computer graphics, with special focus in computational photography and 3D reconstruction.



JINGYI YU received the B.S. degree from the California Institute of Technology, Pasadena, CA, USA, in 2000, and the Ph.D. degree from the Massachusetts Institute of Technology, Cambridge, MA, USA, in 2005. He is currently an Professor with the School of Information Science and Technology, ShanghaiTech University, Shanghai, China, and an Associate Professor with the Department of Computer and Information Sciences and the Department of Electrical and Computer Engineering, University of Delaware, Newark, DE, USA. His current research interests include computer vision and computer graphics, in particular, computational cameras and displays. Prof. Yu has served as the Program Chair of the 2011 Workshop on Omnidirectional Vision and Camera Networks, General Chair of the 2008 International Workshop on Projector-Camera Systems, and Area and Session Chair of the 2011 International Conference on Computer Vision. He was a recipient of the NSF CAREER Award and the AFOSR YIP Award. He is an Editorial Board Member of the IEEE Transactions on Pattern Analysis and Machine Intelligence, The Visual Computer Journal, and Machine Vision and Application.