Robust 3D Self-portraits in Seconds

Zhe Li[1], Tao Yu[1], Chuanyu Pan[1], Zerong Zheng[1], Yebin Liu[1,2]
[1]Department of Automation, Tsinghua University, China
[2]Institute for Brain and Cognitive Sciences, Tsinghua University, China

Abstract

In this paper, we propose an efficient method for robust 3D self-portraits using a single RGBD camera. Benefiting from the proposed PIFusion and lightweight bundle adjustment algorithm, our method can generate detailed 3D self-portraits in seconds and shows the ability to handle subjects wearing extremely loose clothes. To achieve highly efficient and robust reconstruction, we propose PIFusion, which combines learning-based 3D recovery with volumetric non-rigid fusion to generate accurate sparse partial scans of the subject. Moreover, a non-rigid volumetric deformation method is proposed to continuously refine the learned shape prior. Finally, a lightweight bundle adjustment algorithm is proposed to guarantee that all the partial scans can not only "loop" with each other but also remain consistent with the selected live key observations. The results and experiments show that the proposed method achieves more robust and efficient 3D self-portraits compared with state-of-the-art methods.

1. Introduction

Human body 3D modeling, aiming at reconstructing the dense 3D surface geometry and texture of the subject, is a hot topic in both computer vision and graphics and is of great importance in the area of body measurement, digital content creation, virtual try-ons, etc. Traditional human body 3D modeling methods usually rely on experts for data capture and are therefore hard to use. Compared with traditional 3D scanning methods, 3D self-portrait methods, which allow users to capture their own portraits without any assistance, have significant potential for wide usage.

Current 3D self-portrait methods can be classified into 3 categories: learning-based methods, fusion-based methods, and bundle-adjustment-based methods. Learning-based methods mainly focus on 3D human recovery from a single RGB image ([13, 25]). Thus, the results are still far from accurate due to occlusions and depth ambiguities. Fusion-based methods reconstruct scene geometries in an incremental manner, so error accumulation is inevitable, especially for non-rigid scenarios [20], which is detrimental for loop closure reconstruction (e.g., 3D self-portraits). To suppress the accumulated error in incremental fusion, another branch of 3D self-portrait methods also utilizes bundle adjustment algorithms [17, 29, 7, 8, 30, 31]. The whole sequence is first segmented into several chunks, and then fusion methods are applied to each chunk to fuse a smooth partial scan. Finally, non-rigid bundle adjustment is used to "loop" all the partial scans simultaneously by non-rigid registration based on explicit loop closure correspondences and bundling correspondences. Although RGBD bundle adjustment methods have achieved state-of-the-art performance for 3D self-portraits, they still suffer from complicated hardware setups (e.g., relying on multiple sensors or electric turntables [29, 1, 2] or low efficiency [17, 5, 8, 7, 30, 31]).

Figure 1: Our system reconstructs a detailed and textured portrait after the subject self-rotates in front of an RGBD sensor.

One of our observations is that a good combination of non-rigid fusion and bundle adjustment should guarantee both efficiency and accuracy. However, non-rigid fusion methods (e.g., [20] etc.) usually suffer from heavy drifts and error accumulation during tracking, which limit their ability to generate accurate large partial scans. This limitation has led to the fact that previous bundle adjustment methods have to be conducted on considerably large numbers of small partial scans, which significantly increase the optimizing variables in the bundling step. For example, in [8], 40-50 small partial scans need to be bundled together, which takes approximately 5 hours.

To produce large and accurate partial scans by non-rigid fusion, a complete shape prior is necessary. To this end, we propose PIFusion, which utilizes learning-based 3D body recovery (PIFu [25]) as an inner layer in non-rigid fusion [20]. Specifically, in each frame, the inner layer generated by learning-based methods acts as a strong shape prior to improve the tracking accuracy and robustness, and the fused

mesh in return improves the accuracy of the inner layer by the proposed non-rigid volumetric deformation (Sec. 5.3). We also improve the original PIFu [25] by incorporating pixel-aligned depth features for more accurate and robust inner-layer generation (Fig. 3).

Another important observation is that to generate accurate portraits, all the partial scans produced by PIFusion should not only construct a looped model ([8, 17]) but also always remain consistent with the real-world observations, especially the depth point clouds and silhouettes. Instead of using the dense bundle method in [8, 30], we contribute a lightweight bundle adjustment method that involves live terms, key frame selection and joint optimization. Specifically, during each iteration, all the partial scans are not only optimized to "loop"' with each other in the reference frame but are also warped to fit each key input in live frames. The key frames are selected adaptively according to the proposed live depth/silhouette energies. This method further improves the bundle accuracy without losing efficiency.

In summary, by carefully designing the reconstruction pipeline, our method integrates all the advantages from learning, fusion and bundle adjustment methods while avoiding the disadvantages and finally enables efficient and robust 3D self-portraits using a single RGBD sensor.

The contributions can be summarized as follows:

- A new 3D self-portrait pipeline that leverages fusion, learning and bundle adjustment methods and achieves efficient and robust 3D self-portrait reconstruction using a single RGBD sensor.
- A new non-rigid fusion method, PIFusion, which combines a learning-based shape prior and a non-rigid volumetric deformation method to generate large and accurate partial scans.
- A lightweight bundle adjustment method that involves key frame selection and new live energy terms to jointly optimize the loop deformation in the reference frame, as well as the warp fields to live key frames, and finally improves the bundling accuracy without losing efficiency.

2. Related Work

2.1. Learning-based 3D Human Recovery

Learning-based 3D body reconstruction has become more and more popular in recent years. By "seeing" a large amount of ground truth 3D human models, current deep neural networks can infer plausible 3D bodies from various easy-to-obtain inputs, e.g., a single RGB image[13, 23, 15, 24, 9, 6, 38, 19, 25, 36]. For example, Kanazawa *et al.* [13], Omran *et al.* [23] and Kolotouros *et al.* [15] proposed to directly regress the parameters of a statistical body template from a single RGB image. Zhu *et al.* [38] and Alldieck *et al.* [6] took a step forward by deforming the body template according to shading and silhouettes in order to capture more surface details. To address the challenge of varying cloth topology, recent studies have explored many 3D surface representations for deep neural networks, including voxel grids[36], multi-view silhouettes[19], depth maps[9] and implicit functions[25]. Although these methods enable surprisingly convenient 3D human capture, they fail to generate detailed and accurate results due to occlusions and inherent depth ambiguities.

2.2. 3D Human Using Fusion-based Methods

In fusion-based methods, given a noisy RGBD sequence, the scene geometry is first registered to each frame and then updated based on the observations. As a result, the noise in the depth map can be significantly filtered out and the scene can be completed in a incremental manner. The pioneer work in this direction is KinectFusion[21], which was designed for rigid scene scanning using a RGBD sensor. Thus, when scanning live targets like humans, the subjects are required to keep absolutely static in order to get accurate portraits, which is not consistent with the fact that humans are ultimately moving. To handle this problem, Zeng *et al.* [34] proposed a method for quasi-rigid fusion, but it still relies on rotating sensors for data capture, which is hard-to-use. DynamicFusion[20] extended KinectFusion and contributed the first non-rigid volumetric fusion method for real-time dynamic scene reconstruction. Following works [12, 26, 27, 10, 16, 32, 35] kept improving the performance of DynamicFusion by incorporating different types of motion priors or appearance information. For instance, based on the double-layer surface representation, DoubleFusion[33] achieved state-of-the-art performance for dynamic human body reconstruction (with implicit loop closure) using non-rigid fusion. However, constrained by the parametric inner layer representation, DoubleFusion has limited performance for reconstructing extremely wide clothes like long skirts and coats. Moreover, the A-pose requirement for system initialization complicates the portrait scanning process for more general poses.

2.3. 3D Self-portrait Using Bundle Adjustment

To suppress the accumulated error in incremental fusion, another branch of 3D self-portrait methods also utilizes bundle adjustment algorithms. Based on KinectFusion[21], Tong *et al.* [29] used 3 Kinects and a turntable for data capture and non-rigid bundle adjustment for portrait reconstruction. Cui *et al.* [7] achieved self-rotating portrait reconstruction via non-rigid bundle. However, the efficiency is low due to large partial scan numbers. Wang *et al.* [30] conducted bundle adjustment for all point sets without volumetric fusion, which leads to over-smoothed results. The method in [17] is a very related work to ours for it also fuses large partial scans for portrait reconstruction. However, it needs the subject to keep static during the partial scanning process, thus cannot handle self-rotating reconstructions.

Besides above RGBD methods, using a RGB (without depth) video of a rotating human to reconstruct a plausible portrait is also a practical direction. Alldieck *et al.* [5, 4, 3] used silhouette-based joint optimization and Zhu *et al.* [37] used multi-view stereo technologies. However, current methods in this direction still rely on offsetting parametric models to represent cloth, which inherently limits

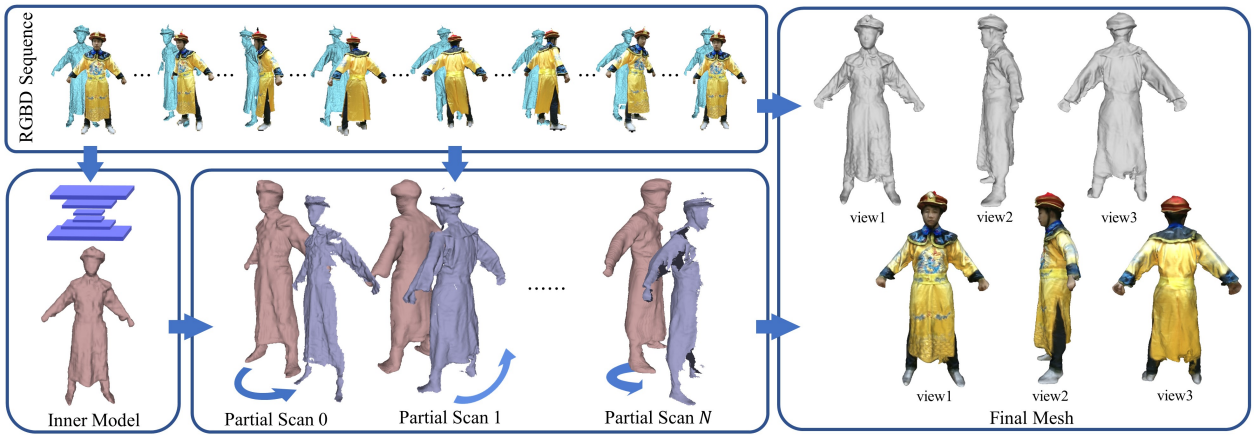


Figure 2: System pipeline. In the first frame, we utilize RGBD-PIFu to generate a roughly correct inner model as a prior. Then we perform PIFusion to generate large and accurate partial scans while the performer is turning around in front of the RGBD sensor. Finally, we conduct lightweight bundle adjustment to merge all the partial scans and generate an accurate and detailed 3D portrait.

their performance for more general clothed human reconstruction. Moreover, sparse feature points from RGB videos are not sufficient for detailed dense surface reconstruction.

3. Overview

As shown in Fig. 2, given an RGBD sequence with a naturally self-rotating motion of the subject, our system performs 3 steps sequentially:

1. **RGBD-PIFu:** In this step, we use a neural network to infer a roughly accurate model of the subject from the first RGBD frame.
2. **PIFusion:** For each frame, we first perform double-layer-based non-rigid tracking with the inferred model as the inner layer and then fuse the observations into the reference frame using the traditional non-rigid fusion method. Finally, non-rigid volumetric deformation is used to further optimize the inner model to improve both tracking and the fusion accuracy. The partial scans are then generated by splitting the whole sequence into several chunks and fusing each chunk separately.
3. **Lightweight bundle adjustment:** In each iteration, we first use key frame selection to select effective key frames to construct the live depth and silhouette terms. Then, joint optimization is performed to not only assemble all the partial scans in the reference frame but also optimize the warping fields to live key frames alternately.

4. RGBD-PIFu

In this work, we extend pixel-aligned implicit functions (PIFu)[25] and propose RGBD-PIFu for 3D self-portrait inference from an RGBD image. PIFu is a spatially aligned representation for 3D surfaces. It is a level-set function f

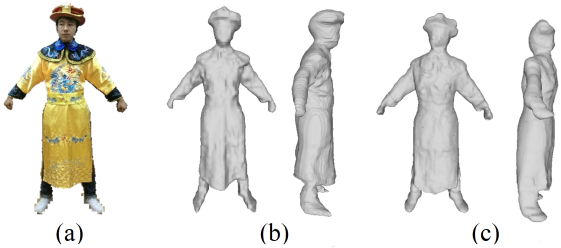


Figure 3: Comparison of RGBD-PIFu and PIFu [25]. (a) Reference color image; (b) RGBD-PIFu result; (c) PIFu result.

that defines the surface implicitly, e.g., $f(X) = 0, X \in \mathbb{R}^3$. In our RGBD-PIFu method, this function is expressed as a composite function f, which consists of a fully convolutional RGBD image encoder g and an implicit function h represented by multilayer perceptrons:

$$f(X; I) = h(G(x; I), X_z),\ X \in \mathbb{R}^3, \tag{1}$$

where I is the input RGBD image, $x = \pi(X)$ is the 2D projection of a 3D point X, $G(x; I)$ is the feature vector of x on the encoded feature map $g(I)$, and X_z is the depth value of X. Different from [25], our image encoder also encodes depth information, which forces the inner model to be consistent with the depth input, thus resolving the depth ambiguity problem and improving the reconstruction accuracy. The training loss is defined as the mean squared error:

$$\mathcal{L} = \frac{1}{n}\sum_{i=1}^{n} \left| f\left(X_i; I\right) - f^*\left(X_i\right)\right|^2, \tag{2}$$

where X_i is a sampled point, $f^*(X_i)$ is the ground-truth value, and n is the number of sampled points.

In the model inference stage, to avoid dense sampling of the implicit function as in [25], we utilize the depth input to

ignore empty regions and only perform uniform sampling of the implicit function in the invisible regions. The isosurface is extracted by the marching cube algorithm [18]. By incorporating depth features, our network is more robust and accurate than the original RGB-PIFu, thus producing a better mesh as the inner model for robust fusion performance, as shown in Fig. 3.

5. PIFusion

5.1. Initialization

In the first frame, we initialize the TSDF (truncated signed distance function) volume by direct depth map projection and then fit the inner model to the initialized TSDF volume. The deformation node graph ([28]) is then uniformly sampled on the inner model using geodesic distance, which is used to parameterize the non-rigid deformation of the fused surface and the inner model.

5.2. Double-layer Non-rigid Tracking

Given the inner model and the fused mesh (i.e., the double-layer surface) in the $(t-1)$-th frame, we need to deform them to track the depth map in the t-th frame. Different from DynamicFusion [20], an inner layer is used to assist non-rigid tracking. Hence, there are two types of correspondences: one is between the fused mesh (outer layer) and the depth observation, and the other is between the inner model (inner layer) and the depth observation. The energy function is then formulated as:

$$E_{\text{tracking}} = \lambda_{\text{outer}} E_{\text{outer}} + \lambda_{\text{inner}} E_{\text{inner}} + \lambda_{\text{smooth}} E_{\text{smooth}}, \quad (3)$$

where E_{outer} and E_{inner} are the energies of the two types of correspondences, E_{smooth} is a smooth term to regularize local as-rigid-as-possible deformations, and $\lambda_{\text{outer}}, \lambda_{\text{inner}}, \lambda_{\text{smooth}}$ are the term weights.

Outer and Inner Term The two terms measure the misalignment between the double layers and the depth map, and they have similar formulations:

$$E_{\text{outer/inner}} = \sum_{(\mathbf{v},\mathbf{u}) \in \mathcal{C}_{\text{outer/inner}}} \left| \hat{\mathbf{n}}_{\mathbf{v}}^{\top} (\hat{\mathbf{v}} - \mathbf{u}) \right|^2, \quad (4)$$

where $\mathcal{C}_{\text{outer}}$ and $\mathcal{C}_{\text{inner}}$ are two types of correspondence sets, and $(\mathbf{v}, \mathbf{u})$ is a correspondence pair; $\mathbf{v}$ is a vertex on the outer layer (fused mesh) or the inner layer (inner model), and $\mathbf{u}$ is the closest point to $\mathbf{v}$ on the depth map. Note that $\mathbf{v}$ is the coordinate in the reference frame, while $\hat{\mathbf{v}}$ and $\hat{\mathbf{n}}_{\mathbf{v}}$ are the position and normal of $\mathbf{v}$ in the live frame warped by its KNN nodes using dual quaternion blending:

$$\mathbf{T}(\mathbf{v}) = SE3 \left(\sum_{k \in \mathcal{N}(\mathbf{v})} w(k, \mathbf{v}) \mathbf{dq}_k \right), \quad (5)$$

where $\mathbf{dq}_k$ is the dual quaternion of the k-th node, $SE3(\cdot)$ maps a dual quaternion to the **SE**(3) space, $\mathcal{N}(\mathbf{v})$ are the KNN nodes of $\mathbf{v}$, $w(k, \mathbf{v}) = \exp(-\|\mathbf{v} - \mathbf{x}_k\|_2^2 / (2r^2))$ is the blending weight, $\mathbf{x}_k$ is the position of the k-th node, and r is the active radius.

Smooth Term The smooth term is defined on all edges of the node graph to guarantee local rigid deformation. This term is defined as

$$E_{\text{smooth}} = \sum_i \sum_{j \in \mathcal{N}(i)} \|\mathbf{T}_i \mathbf{x}_j - \mathbf{T}_j \mathbf{x}_j\|_2^2, \quad (6)$$

where $\mathbf{T}_i$ and $\mathbf{T}_j$ are the transformations associated with the i-th and j-th nodes, and $\mathbf{x}_i$ and $\mathbf{x}_j$ are the positions of the i-th and j-th nodes in the reference frame, respectively.

We solve Eq. 3 by the iterative closest point (ICP) algorithm and use the Gauss-Newton algorithm to solve the energy optimization problem. After tracking, we use the typical fusion method [20] to fuse the current depth observations and update the TSDF volume.

5.3. Non-rigid Volumetric Deformation

The initial inner model inferred by RGBD-PIFu is by no means accurate enough for double-layer surface tracking, and the correspondences between the inner model and the depth map may even reduce the tracking performance. To deal with this issue, inspired by [33], we conduct a non-rigid volumetric deformation algorithm to continue correcting the inner model by fitting it to the fused mesh (i.e., the 0-level set of the TSDF) in the reference volume. Moreover, the weight of the inner term, λ_{inner} in Eq. 3, is also designed to decrease along the ICP iterations to enable a more accurate outer surface fitting performance.

We utilize the initialized node graph to parameterize the non-rigid deformation of the inner model. Given the updated TSDF volume of the fused mesh, the energy function of non-rigid volumetric deformation is defined as:

$$E_{\text{vol}} = E_{\text{tsdf}} + \lambda_{\text{smooth}} E_{\text{smooth}}, \quad (7)$$

where E_{tsdf} measures the misalignment error between the inner model and the isosurface at threshold 0, and E_{smooth} is the same as Eq. 6. The TSDF term is defined as

$$E_{\text{tsdf}} = \sum_{\mathbf{v} \in \mathbf{T}} \left| \text{TSDF}(\hat{\mathbf{v}}) \right|^2, \quad (8)$$

where $\mathbf{T}$ is the initial inner model without non-rigid deformations in the reference frame, $\mathbf{v}$ is a vertex of $\mathbf{T}$, $\hat{\mathbf{v}}$ is the position warped by the KNN nodes of $\mathbf{v}$, TSDF$(\cdot)$ is a trilinear sampling function that takes a point in the reference frame and returns the interpolated TSDF value. By minimizing the squared sum of the TSDF values of all the vertices of the deformed inner model, the inner model will perfectly align with the fused mesh in the reference frame.

For the next frame, the corrected inner model is warped to the live frame to search for correspondences in the tracking step. This step provides more accurate correspondences and significantly improves the registration accuracy compared with directly warping the initial inner model.

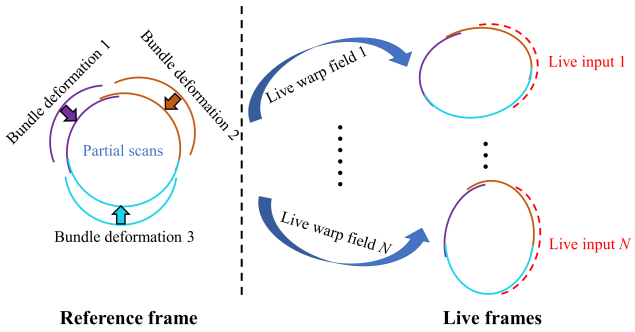


Figure 4: Illustration of bundle adjustment with joint optimization. The bundle deformations are optimized to "loop" these partial scans in the reference frame, while the live warp fields are optimized to deform the partial scans to fit live input.

5.4. Partial Scan Fusion

To guarantee that the following bundle adjustment is only conducted on a small number of partial scans, we fuse the partial scans within several large chunks of the whole sequence in the reference frame. Specifically, given a sequence of the performer turning around in front of the sensor, we calculate the orientation of the performer and then split the whole sequence into 5 chunks, which cover the front, back and two side views of the performer. Due to the accumulated error, the first and last partial scans that compose a loop may not align very well. The proposed lightweight bundle adjustment will resolve this problem and finally generate accurate 3D portraits.

6. Lightweight Bundle Adjustment

Regarding non-rigid bundle adjustment (BA), we argue that a well-looped model after typical BA is an accurate model. Our insight is that after BA, all the partial scans should not only construct a looped model in the reference frame but also be well fitted to all the live observations after non-rigid warping using the live warp fields. To this end, we propose an efficient algorithm to jointly optimize the bundle deformations (for loop closure reconstruction) and live warp fields (for live depth fitting), as shown in Fig. 4. Novel energy terms, including the live depth and silhouette energies, are incorporated to enforce the consistency between the warped partial scans and live depth inputs. However, optimizing the live warp fields corresponding to all the live frames in bundle adjustment will significantly decrease the efficiency. In practice, we found that performing live depth fitting on only several key frames is sufficient for generating accurate results. Therefore, we propose a key frame selection strategy to select effective key frames by sorting the live depth and silhouette energies.

6.1. Joint Optimization

After PIFusion, we can acquire N partial scans. As illustrated in Fig. 4, we first construct a node graph for each partial scan for describing the bundle deformation, which is then optimized using loop closure correspondences to deform the partial scan to "loop" with the others in the reference frame. Moreover, all the partial scans are deformed together to fit each live frame by optimizing the corresponding live warp field, which is similar to the non-rigid tracking in PIFusion. As a result, each partial scan has its own bundle deformation, and all the partial scans share live warp fields in common.

We solve the joint optimization problem by optimizing the bundle deformations and live warp fields alternately. In each iteration, both bundle deformations and live warp fields will be updated to minimize the total energy.

6.2. Key Frame Selection

To maintain the efficiency of our algorithm, we propose a key frame selection strategy for constructing efficient and effective live depth fitting terms. Specifically, we uniformly divide the whole sequence into K segments, and before each iteration of joint optimization, for each frame, we calculate two types of metrics: the geometric misalignment error and the silhouette error. The first metric is the misalignment between warped partial scans and the corresponding input depth point cloud. The silhouette error is calculated by first rendering a mask map in the camera view using all the warped partial scans and then calculating the difference between the rendered mask and the input silhouette. We then select the frames with the largest geometric misalignment error and silhouette error in each segment as depth key frames $\mathcal{K}_{\text{dep}}$ and silhouette key frames $\mathcal{K}_{\text{sil}}$, respectively.

6.3. Formulation

Different from other bundle adjustment algorithms ([8, 30]), we not only "loop"' these partial scans but also introduce live frame observations into the optimization procedure to improve the accuracy. The total energy function is defined as

$$\begin{aligned} E(\mathbf{W}_b^j, \mathbf{W}_l^i) = \lambda_{\text{loop}} E_{\text{loop}} + \lambda_{\text{depth}} E_{\text{depth}} \\ + \lambda_{\text{silhouette}} E_{\text{silhouette}} + \lambda_{\text{smooth}} E_{\text{smooth}}, \end{aligned} \tag{9}$$

where $\mathbf{W}_b^j$ is the bundle deformation corresponding to the j-th partial scan, $\mathbf{W}_l^i$ is the live warp field from the reference frame to the i-th key frame, and $E_{\text{loop}}, E_{\text{depth}}, E_{\text{silhouette}}$ and E_{smooth} are the energies of loop closure, live depth, live silhouette and smooth regularization terms, respectively.

In each iteration, we optimize the bundle deformation and live warp fields alternately to minimize Eq. 9. Note that after PIFusion, although the partial scans have already been well aligned with the live depth inputs, the live warp fields are still not accurate enough to guarantee all the fused partial scans to construct a loop in the reference frame directly. Thus, the bundle deformation in the reference frame will conflict with live depth fitting in the live frames without simultaneously optimizing the live warp fields.

Loop Term The loop term measures the amount of mis-

alignment among these partial scans and is defined as

$$E_{\text{loop}} = \sum_{\substack{i,j=1 \\ i \neq j}}^{N} \sum_{(\mathbf{v}_p, \mathbf{v}_q) \in \mathcal{C}_{i,j}} \left| \mathbf{W}_b^i(\mathbf{n}_p)^\top \left(\mathbf{W}_b^i(\mathbf{v}_p) - \mathbf{W}_b^j(\mathbf{v}_q) \right) \right|^2, \tag{10}$$

where N is the number of partial scans, $\mathcal{C}_{i,j}$ is the correspondence set between the i-th and j-th partial scans acquired by searching the closest points, $(\mathbf{v}_p, \mathbf{v}_q)$ is a correspondence pair, $\mathbf{v}_p$ and $\mathbf{v}_q$ are vertices on the i-th and j-th partial scans, respectively, $\mathbf{n}_p$ is the normal of $\mathbf{v}_p$ on the i-th partial scan, and $\mathbf{W}_b^i(\mathbf{v}_p)$ and $\mathbf{W}_b^i(\mathbf{n}_p)$ represent the position and normal warped by bundle deformation. This term forces all the partial scans to register with each other in the reference frame.

Live Depth Term This term measures the misalignment of all the partial scans with all the depth maps in $\mathcal{K}_{\text{depth}}$:

$$E_{\text{depth}} = \sum_{i=1}^{K} \sum_{j=1}^{N} \sum_{(\mathbf{v}, \mathbf{u}) \in \mathcal{D}_{j,i}} \left| \mathbf{W}_l^i(\mathbf{W}_b^j(\mathbf{n}))^\top \left(\mathbf{W}_l^i(\mathbf{W}_b^j(\mathbf{v})) - \mathbf{u} \right) \right|^2, \tag{11}$$

where $K = |\mathcal{K}_{\text{depth}}|$ is the number of key frames, $\mathcal{D}_{j,i}$ is the correspondence set between the j-th partial scan and the depth map in the i-th key frame, $(\mathbf{v}, \mathbf{u})$ is a correspondence pair, $\mathbf{v}$ is a vertex on the j-th partial scan, $\mathbf{u}$ is a point on the depth map, and $\mathbf{W}_l^i(\cdot)$ takes a point or normal in the reference frame as input and returns the warped position or normal in the i-th key frame. This term is designed to force partial scans to align with depth point clouds in $\mathcal{K}_{\text{depth}}$.

Live Silhouette Term This term measures the misalignment between the rendered mask of all the warped partial scans and the input mask of the body shape in the live frames. Similar to LiveCap[11], we preprocess the input mask using the distance transform. For the i-th key frame, we render a mask image of all partial scans deformed by the live warp field and then filter a boundary vertex set $\mathcal{B}_i$. We define the live silhouette term as

$$E_{\text{silhouette}} = \sum_{i=1}^{K} \sum_{j=1}^{N} \sum_{\mathbf{v}_j \in \mathcal{B}_i} d_j \cdot \left| I_{\text{DT}}^i(\pi(\mathbf{W}_l^i(\mathbf{W}_b^j(\mathbf{v}_j)))) \right|^2, \tag{12}$$

where $K = |\mathcal{K}_{\text{sil}}|$ is the number of key frames, $\mathbf{v}_j$ is a boundary vertex on the j-th partial scan (note that the boundary means that the vertex is projected near the boundaries of the rendered mask image rather than the boundary of this partial scan), $d_j \in \{-1, +1\}$ is an indicated value that indicates the correct direction in the distance field[11], I_{DT}^i is the distance-transformed image of the input mask and $\pi(\cdot)$ is the projection function. This term will deform the shape of the partial scans to fit with the input silhouettes.

The smooth term is defined similarly to Eq. 6. We solve Eq. 9 using the Gauss-Newton method. Within each iteration, we construct a large sparse system of linear equations and then utilize an efficient preconditioned conjugate gradient (PCG) solver on a GPU to obtain the updates.

6.4. Non-rigid Multi-texturing

After lightweight bundle adjustment, we fuse all the partial scans into a watertight mesh using Poisson reconstruction [14]. For each live frame, we project each visible vertex to the color image to retrieve a color value. After processing all the frames, we blend the retrieved color values according to the normal direction and obtain the final vertex color. Specifically, for vertex $\mathbf{v}_i$, we calculate its color $C_{\mathbf{v}_i}$ as a weighted average of the color values retrieved from all the live frames. The blending weight $\omega_{i,j}$ is defined as:

$$\omega_{i,j} = \begin{cases} 0 & , \mathbf{v}_i \text{ is invisible in the } j\text{-th frame} \\ \frac{|\mathbf{n}_{\mathbf{v}_i} \cdot \hat{z}|}{|\mathbf{n}_{\mathbf{v}_i}|} & , \mathbf{v}_i \text{ is visible in the } j\text{-th frame} \end{cases}, \tag{13}$$

where $\mathbf{n}_{\mathbf{v}_i}$ is the normal of $\mathbf{v}_i$ and $\hat{z}$ is the direction the camera is looking. To avoid oversmoothing, for each vertex, only the top 15% of weighted color values are blended.

7. Results

In this section, we first report the system performance and our implementation. Then, we compare our method with current state-of-the-art works. Finally, we evaluate the core parts of our system. In Fig. 5, we demonstrate several 3D portraits acquired by our system.

7.1. Performance and Implementation

Our 3D self-portrait system is very efficient. The whole pipeline is implemented on one NVIDIA Geforce RTX 2080Ti GPU. The initialization that generrates the inner model by RGBD-PIFu and initializing PIFusion takes almost 10 seconds. PIFusion runs in real-time (at 30 ms per frame). For each frame, the tracking, volumetric deformation and fusion take 20 ms, 3 ms and 6 ms, respectively.

Similar to [25], we adapt a stacked hourglass network [22] as the image encoder, and the implicit function is presented by a MLP with 257, 1024, 512, 256, 128, and 1 neurons in each layer. We render the Twindom dataset (`https://web.twindom.com/`) to acquire depth and color images and utilize 3500 images to train this network. When training, the batch size is 4, the learning rate is 1×10^{-3}, and the number of epochs is 28. The training procedure takes one day on one RTX 2080Ti GPU.

In the tracking of PIFusion, the number of ICP iterations is 5 per frame, and we set $\lambda_{\text{outer}} = 1.0$, $\lambda_{\text{inner}} = 1.0$, and $\lambda_{\text{smooth}} = 5.0$, while λ_{inner} will decrease linearly as the iteration continues. For each vertex, we use its 4 nearest neighbors for non-rigid deformation, and the number of neighbors of each node is 8. In bundle adjustment, the numbers of partial scans and key frames are both 5, we set $\lambda_{\text{loop}} = 1.0$, $\lambda_{\text{depth}} = 0.5$, $\lambda_{\text{silhouette}} = 0.001$ and $\lambda_{\text{smooth}} = 2.0$, and the number of iterations is 25. This procedure takes only 15 seconds, and texturing takes 1 second.

7.2. Comparison

Comparison with Fusion Methods We compared our fusion method PIFusion with DynamicFusion [20] and DoubleFusion [33] using sequences captured by a Kinect V2

Figure 5: Examples of 3D portraits acquired by our system.

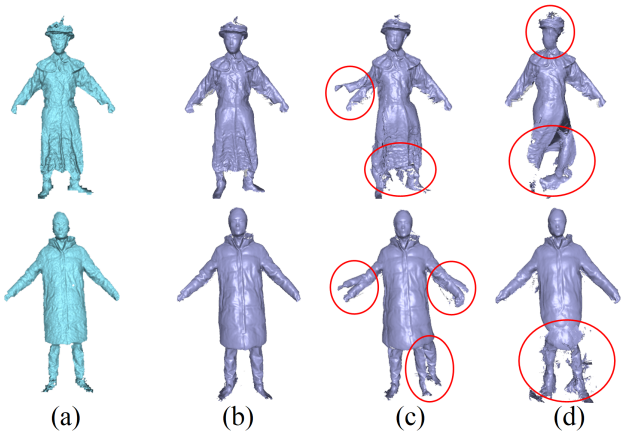


Figure 6: Comparison of the proposed PIFusion, DynamicFusion [20] and DoubleFusion [33] methods. (a) Reference depth input; (b), (c) and (d) are the results of PIFusion, DynamicFusion and DoubleFusion, respectively.

sensor. Fig. 6 demonstrates that our method improves the tracking and implicit loop-closure performance compared to the other methods, especially for subjects wearing loose clothes. Note that for this experiment, we use PIFusion to fuse the whole sequence without generating partial scans.

Comparison with Bundle Adjustment Methods We compare our method with the state-of-the-art non-rigid bundle adjustment method Wang *et al.* [30]. As shown in Fig. 7, our method achieves much more detailed and accurate 3D self-portraits than [30]. Moreover, as mentioned in the related work section of [30], although the results are plausible, [17] requires the subject to remain static several times during scanning, thus complicating the scanning process.

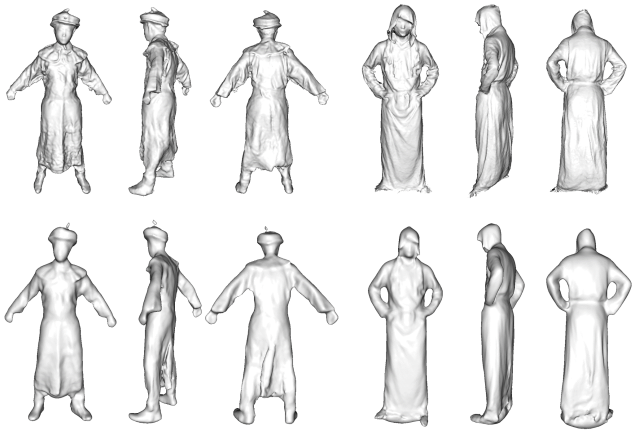


Figure 7: Comparison of our method (the top row) with the method proposed by Wang *et al.* [30] (the bottom row).

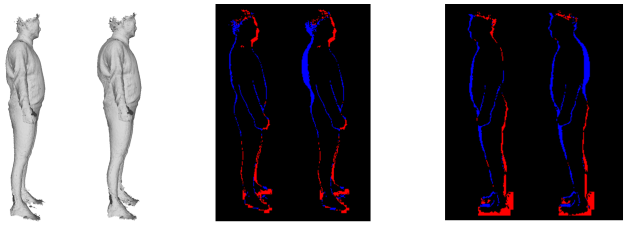
7.3. Evaluation

Ablation Studies on Energy Terms

– Inner and Outer Terms in PIFusion Without the inner term, PIFusion will degenerate into DynamicFusion [20], which suffers from heavy drifts and tracking errors (Fig. 6). Moreover, the lack of the outer term makes the final reconstruction accuracy fully depend on the accuracy of the shape prior, which is usually not accurate enough.

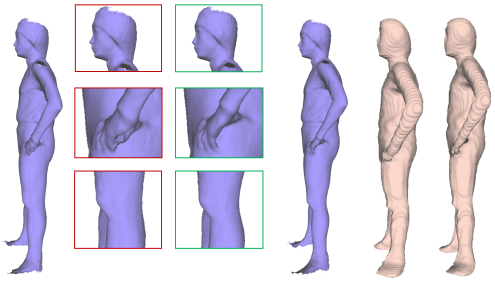
– Live Silhouette Term in Bundle Adjustment Fig. 8 demonstrates that the live silhouette term could deform partial scans to be consistent with input silhouettes, thus further improving the accuracy of the optimized partial scans.

Non-rigid Volumetric Deformation We evaluate the non-rigid volumetric deformation qualitatively, as shown in



(a) (b) (c)

Figure 8: Evaluation of the live silhouette term. (a) Optimized partial scans with (left) and without (right) the live silhouette term, (b) the mask error maps in the 1st key frame with (left) and without (right) the live silhouette term, (c) mask error maps in the 3rd key frame with (left) and without (right) the live silhouette term (non-black pixels represent errors).



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

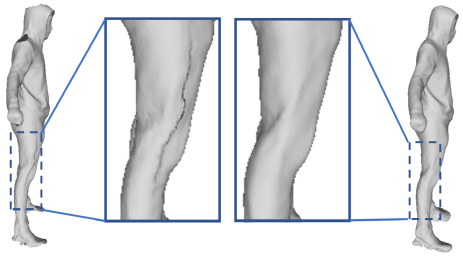
Figure 9: Evaluation of the non-rigid volumetric deformation step. Fused mesh without (a) and with (b) volumetric deformation; (c) the original inner model generated by RGBD-PIFu; (d) the inner model after volumetric deformation.

Fig. 9. The results demonstrate that without volumetric deformation, the fused geometry will highly depend on the original inner model generated by RGBD-PIFu. With non-rigid volumetric deformation, the outer observations are introduced to update the inner model in each frame. This step will alleviate the errors brought by the inner model and improve the accuracy of our reconstruction results.

Loop Closure We compare the reconstructed result after lightweight bundle adjustment with the mesh completely fused using PIFusion in Fig. 10. This comparison demonstrates that PIFusion still suffers from loop-closure problems, especially in the cases of challenging motion (very articulated motion) and an inaccurate initial inner model. With the help of bundle adjustment, we can obtain more accurate portraits efficiently than the fusion methods.

Joint Optimization We evaluate the joint optimization by the total energy of Eq. 9 in each iteration. Fig. 11 demonstrates that joint optimization of both the bundle deformation and live warp fields can achieve a lower optimum than the methods without joint optimization.

Body Measurement We quantitatively evaluated the accuracy of our results on body measurements. To evaluate the measurement error, we first utilized a laser scanner to obtain the ground-truth shape of a tight-clothed subject and then scanned the subject again using our system. Tab. 1 shows the measurement results of several body parts, which illustrates that our method acquires more accurate results than the state-of-the-art fusion-based body reconstruction method, DoubleFusion[33]. Moreover, the proposed lightweight bundle adjustment method is effective in further improving the accuracy of the final reconstruction.



(a) (b)

Figure 10: Evaluation of the loop closure. (a) Fused mesh by PIFusion; (b) the reconstructed mesh after bundle adjustment.

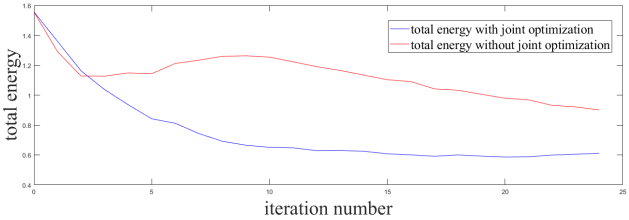


Figure 11: The total energy in each bundle adjustment iteration.

Method	chest	waist	right knee
DoubleFusion[33]	98.7	92.5	43.1
PIFusion	97.6	87.2	40.7
Bundle Adjustment	94.6	84.5	39.6
Ground Truth	91.2	79.7	37.7

Table 1: Evaluation of body measurements on case "lz": the circumference of some body parts (cm).

8. Discussion

Conclusion In this paper, we have proposed a new method for robust and efficient 3D self-portrait reconstruction from a single RGBD camera. We propose PIFusion, a novel volumetric non-rigid fusion method constrained by a learned shape prior, for generating large and accurate partial scans. More importantly, the proposed lightweight bundle adjustment method not only guarantees the generation of a looped model in the reference frame but also ensures the alignment with live key observations, which further improves the accuracy of the final portrait without losing efficiency. In conclusion, with the proposed method, users can conveniently obtain detailed and accurate 3D self-portraits in seconds.

Limitations Our method still relies on the completeness of the shape prior provided by RGBD-PIFu. Specifically, if the inferred shape prior loses some body parts, the final reconstruction may also lose these parts. Moreover, if some cases (e.g., object interactions) are not included in the training dataset of RGBD-PIFu, they may also not be well handled. However, growing the deformation node graph according to live observations may solve these problems.

Acknowledgments This paper is supported by the NSFC No.61827805, No.61531014 and No.61861166002.

References

[1] https://texel.graphics/.

[2] https://www.shapify.me/.

[3] T. Alldieck, M. Magnor, B. L. Bhatnagar, C. Theobalt, and G. Pons-Moll. Learning to reconstruct people in clothing from a single RGB camera. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, jun 2019.

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

[5] T. Alldieck, M. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll. Video based reconstruction of 3d people models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[6] T. Alldieck, G. Pons-Moll, C. Theobalt, and M. Magnor. Tex2shape: Detailed full human body geometry from a single image. In *IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2019.

[7] Y. Cui, W. Chang, T. Nöll, and D. Stricker. Kinectavatar: Fully automatic body capture using a single kinect. In J.-I. Park and J. Kim, editors, *Computer Vision - ACCV 2012 Workshops*, pages 133–147, Berlin, Heidelberg, 2013. SPRINGER Berlin Heidelberg.

[8] M. Dou, J. Taylor, H. Fuchs, A. Fitzgibbon, and S. Izadi. 3d scanning deformable objects with a single rgbd sensor. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 493–501, Boston, 2015. IEEE.

[9] V. Gabeur, J. Franco, X. Martin, C. Schmid, and G. Rogez. Moulding humans: Non-parametric 3d human shape estimation from single images. *CoRR*, abs/1908.00439, 2019.

[10] K. Guo, F. Xu, T. Yu, X. Liu, Q. Dai, and Y. Liu. Real-time geometry, albedo and motion reconstruction using a single rgbd camera. *ACM Transactions on Graphics*, 36(3):32:1–32:13, 2017.

[11] M. Habermann, W. Xu, M. Zollhöfer, G. Pons-Moll, and C. Theobalt. Livecap: Real-time human performance capture from monocular video. *ACM Transactions on Graphics*, 38(2):14:1–14:17, Mar. 2019.

[12] M. Innmann, M. Zollhöfer, M. Nießner, C. Theobalt, and M. Stamminger. Volumedeform: Real-time volumetric non-rigid reconstruction. In *European Conference on Computer Vision (ECCV)*, volume 9912, pages 362–379, Amsterdam, 2016. SPRINGER.

[13] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik. End-to-end recovery of human shape and pose. In *IEEE Conference on Computer Vision and Pattern Regognition (CVPR)*, pages 7122–7131, Salt Lake City, 2018. IEEE.

[14] M. Kazhdan, M. Bolitho, and H. Hoppe. Poisson surface reconstruction. In *Proceedings of the Fourth Eurographics Symposium on Geometry Processing*, SGP '06, pages 61–70, Aire-la-Ville, Switzerland, Switzerland, 2006. Eurographics Association.

[15] N. Kolotouros, G. Pavlakos, and K. Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *CVPR*, 2019.

[16] C. Li, Z. Zhang, and X. Guo. Articulatedfusion: Real-time reconstruction of motion, geometry and segmentation using a single depth camera. In *European Conference on Computer Vision (ECCV)*, pages 324–40, Munich, 2018. SPRINGER.

[17] H. Li, E. Vouga, A. Gudym, L. Luo, J. T. Barron, and G. Gusev. 3d self-portraits. *ACM Transactions on Graphics*, 32(6):187:1–187:9, 2013.

[18] W. E. Lorensen and H. E. Cline. Marching cubes: A high resolution 3d surface construction algorithm. In *ACM SIGGRAPH*, pages 163–169, New York, NY, USA, 1987. ACM.

[19] R. Natsume, S. Saito, Z. Huang, W. Chen, C. Ma, H. Li, and S. Morishima. Siclope: Silhouette-based clothed people. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[20] R. A. Newcombe, D. Fox, and S. M. Seitz. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 343–352, Boston, 2015. IEEE.

[21] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohli, J. Shotton, S. Hodges, and A. Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 127–136, Washington, DC, USA, 2011. IEEE Computer Society.

[22] A. Newell, K. Yang, e. B. Deng, Jia", J. Matas, N. Sebe, and M. Welling. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision (ECCV)*, pages 483–499, Cham, 2016. SPRINGER.

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

[24] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. A. Osman, D. Tzionas, and M. J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, Long Beach, 2019. IEEE.

[25] S. Saito, , Z. Huang, R. Natsume, S. Morishima, A. Kanazawa, and H. Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[26] M. Slavcheva, M. Baust, D. Cremers, and S. Ilic. KillingFusion: Non-rigid 3D Reconstruction without Correspondences. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5474–5483, Honolulu, 2017. IEEE.

[27] M. Slavcheva, M. Baust, and S. Ilic. Sobolevfusion: 3d reconstruction of scenes undergoing free non-rigid motion. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2646–2655, Salt Lake City, June 2018. IEEE.

[28] R. W. Sumner, J. Schmid, and M. Pauly. Embedded deformation for shape manipulation. *ACM Transactions on Graphics*, 26(3), July 2007.

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

[30] K. Wang, G. Zhang, and S. Xia. Templateless non-rigid reconstruction and motion tracking with a single rgb-d camera. *IEEE Transactions on Image Processing*, 26(12):5966–5979, Dec 2017.

[31] S. Wang, X. Zuo, C. Du, R. Wang, J. Zheng, and R. Yang. Dynamic non-rigid objects reconstruction with a single rgb-d sensor. *Sensors (Basel, Switzerland)*, 18, 03 2018.

[32] 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 *IEEE International Conference on Computer Vision (ICCV)*, pages 910–919, Venice, 2017. IEEE.

[33] T. Yu, Z. Zheng, K. Guo, J. Zhao, Q. Dai, H. Li, G. Pons-Moll, and Y. Liu. Doublefusion: Real-time capture of human performances with inner body shapes from a single depth sensor. In *IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, pages 7287–7296, Salt Lake City, June 2018. IEEE.

[34] M. Zeng, J. Zheng, X. Cheng, and X. Liu. Templateless quasi-rigid shape modeling with implicit loop-closure. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013.

[35] Z. Zheng, T. Yu, H. Li, K. Guo, Q. Dai, L. Fang, and Y. Liu. Hybridfusion: Real-time performance capture using a single depth sensor and sparse imus. In *European Conference on Computer Vision (ECCV)*, pages 389–406, Munich, Sept 2018. SPRINGER.

[36] Z. Zheng, T. Yu, Y. Wei, Q. Dai, and Y. Liu. Deephuman: 3d human reconstruction from a single image. In *IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[37] H. Zhu, Y. Liu, J. Fan, Q. Dai, and X. Cao. Video-based outdoor human reconstruction. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(4):760–770, Apr. 2017.

[38] H. Zhu, X. Zuo, S. Wang, X. Cao, and R. Yang. Detailed human shape estimation from a single image by hierarchical mesh deformation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.