DoubleFusion: Real-time Capture of Human Performances with Inner Body Shapes from a Single Depth Sensor

Tao Yu, Jianhui Zhao, Zerong Zheng, Kaiwen Guo, Qionghai Dai, *Senior Member, IEEE,*
Hao Li, Gerard Pons-Moll, and Yebin Liu, *Member, IEEE,*

Abstract—We propose DoubleFusion, a new real-time system that combines volumetric non-rigid reconstruction with data-driven template fitting to simultaneously reconstruct detailed surface geometry, large non-rigid motion and the optimized human body shape from a single depth camera. One of the key contributions of this method is a double-layer representation consisting of a complete parametric body model inside, and a gradually fused detailed surface outside. A pre-defined node graph on the body parameterizes the non-rigid deformations near the body, and a free-form dynamically changing graph parameterizes the outer surface layer far from the body, which allows more general reconstruction. We further propose a joint motion tracking method based on the double-layer representation to enable robust and fast motion tracking performance. Moreover, the inner parametric body is optimized online and forced to fit inside the outer surface layer as well as the live depth input. Overall, our method enables increasingly denoised, detailed and complete surface reconstructions, fast motion tracking performance and plausible inner body shape reconstruction in real-time. Experiments and comparisons show improved fast motion tracking and loop closure performance on more challenging scenarios. Two extended applications including body measurement and shape retargeting show the potential of our system in terms of practical use.

Index Terms—Computer Vision, 3D Reconstruction, Real-time, Performance Capture.

1 Introduction

HUMAN performance capture has been a challenging research topic in computer vision and computer graphics for decades. The goal is to reconstruct a temporally coherent representation of the dynamically deforming surface of human characters from videos. Although array based methods [1], [2], [3], [4], [5], [6], [7], [8], [9], [10] using multiple video or depth cameras are well studied and have achieved high quality results, the expensive camera-array setups and controlled studios limit its application to a few technical experts. As depth cameras are increasingly popular in the consumer space (iPhone X/XS, Google Tango, etc.), the recent trend focuses on using more and more practical setups like a single depth camera [11], [12], [13]. In particular, by combining non-rigid surface tracking and volumetric depth integration, DynamicFusion like approaches [14], [15], [16], [17] achieved real-time dynamic scene reconstruction using a single depth camera without the requirement of pre-scanned templates. Such systems are low cost, easy to set up and promising for popularization; however, they are still restricted to controlled slow motion due to the very big solution space for general non-rigid surface tracking. The challenges are occlusions (single view), computational resources (real-time), loop closure and the lack of a pre-scanned template model.

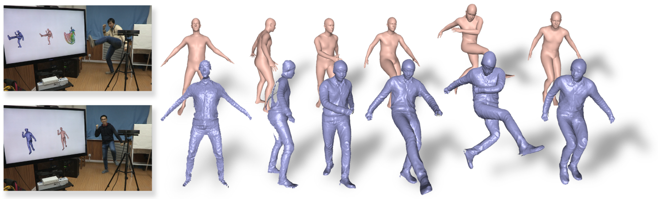


Fig. 1. Our system and the real-time reconstructed results.

Recent methods that use articulated motion priors ([18] and [19]) for real-time non-rigid fusion have achieved better results than those of [20] and [16]. These methods have shown that regularizing non-rigid deformations with a skeleton or articulated motion prior is beneficial for capturing human performance. Moreover, by explicitly utilizing human skeleton structure, [18] generates better human performance capture results than [19], which uses general articulated motion prior for general object reconstruction. However, the system of [18] fails during fast motion, especially when the surface is not yet complete, since the human joints are too sparse and only the gradually fused surface is available for tracking. Moreover, the skeleton embedding performance of [18] relies heavily on the initialization step because the embedding is fixed after initialization in their method. Inaccurate skeleton embedding leads to deteriorated motion tracking and surface skinning performance.

For human performance capture, besides the skeleton, body shape is also a very strong prior since it is loop closed

- *Tao Yu and Jianhui Zhao are with the School of Instrumentation and Optoelectronic Engineering, Beihang University, Beijing 100191, China.*
- *Yebin Liu, Qionghai Dai and Zerong Zheng are with the Broadband Network & Digital Media Lab, Department of Automation, Tsinghua University, Beijing 100084, China.*
- *Kaiwen Guo is with Google Inc, San Francisco, United States.*
- *Hao Li is with Institute for Creative Technologies, University of Southern California, Los Angeles, United States.*
- *Gerard Pons-Moll is with Max-Plunk Institute for Informatics, Saarbrucken, Germany.*
- *Corresponding author: Jianhui Zhao and Yebin Liu.*

Manuscript received October 11, 2018.

and complete. To fully take advantage of *both human shape and pose motion prior*, we propose "DoubleFusion": a single-view and real-time dynamic surface reconstruction system that simultaneously reconstructs general cloth geometry and inner body shape. Based on the recent state-of-the-art body model SMPL [21], we propose a double-layer surface representation consisting of an outer surface layer, and an inner body layer for reconstruction and depth registration. In addition, we make each layer benefit from each other. The observed outer surface is gradually fused and deformed while the shape and pose parameters of the inner body layer are also gradually optimized to fit inside the outer surface as well as the live depth input. On one hand, the inner body layer is a complete model that allows to find enough correspondences, especially when only partial surface is obtained; in addition, it places a constraint on where to fuse the geometry of the outer surface. On the other hand, the gradually fused outer surface provides increasingly more constraints to update the body shape and pose online. Note that since we optimize the inner body layer according to the outer fused surface, the estimated body shape in our pipeline may not be the real inner body shape of the subject (especially when capturing very loose cloth). Finally, the two layers are solved sequentially in real-time.

Overall, our proposed DoubleFusion system offers the new ability to simultaneously reconstruct the inner body shape and pose as well as the outer surface geometry and motion in real-time. This is achieved by using only a single depth camera without a separate pre-scanning phase, and only requiring an initialisation pose. Figure 1 shows the system setup and the real-time reconstruction results. Compared to systems that only reconstruct the outer surface like BodyFusion [18], we demonstrate substantially improved performance in handling fast motion. In contrast to systems specialized to capture the inner body [13], our approach can handle people wearing casual clothing, and it works in real-time. We make the following technical contributions in this paper to enable the above advantages.

- We propose the double-layer representation (Section 3.1) for high quality and real-time human performance capture. We define the double node graph that contains an on-body node graph and a far-body node graph. The double node graph enables better leverage of the human shape and pose prior, while still maintaining the ability to handle surface deformations that are far from the inner body surface. The double-layer representation may also be used in other human performance capture setups (e.g., multi-view systems) or other performance capture tasks (e.g., capturing animals).
- Joint motion tracking (Section 4). We introduce a method to jointly optimize for the pose of the inner body shape and the non-rigid deformation of the outer surface based on the double-layer representation. Feature correspondences on both the inner body shape and fused outer layer enable fast motion tracking performance and robust geometry fusion.
- Inner body optimization (Section 5.2). We optimize the parametric body model, according to the continuously updated truncated signed distance field (TSDF) in the canonical volume, without searching for vertex correspondences explicitly. Moreover, we also incorporate live depth input into the inner body optimization step. The optimized body shape and pose (joint positions) in the canonical frame (which is defined by the first depth frame) enables more accurate surface tracking and deformation results.

A preliminary version of this paper appeared in [22], which introduced an effective method for real-time human performance capture based on the proposed double-layer representation. The present work makes the following additional contributions. First, we propose dynamic detail deformation (Section 5.1), a simple yet efficient method for the recovery of the high frequency and dynamic details on the surface geometry (e.g., cloth wrinkles), which are seriously smoothed by the continuous fusion stragegy used in recent fusion-based dynamic 3D reconstruction methods (e.g., [14], [15], [16], [17] and [22]). Second, we introduce a new energy term for inner body optimization based on live depth input (Section 5.2). In the previous version, we only optimized the body according to the TSDF in the canonical volume. Although this method efficiently generates plausible inner body shapes, the actural body embedding in the canonical volume (which is also the skeleton embedding of the continuously fused canonical model) may not accurate due to the insufficient A-pose canonical volume information. For example, the positions of the knees are difficult to estimate under the A-pose since there are no bendings around the knees. In the current version, we incorporate live depth input and the constraints from all the live poses to improve the inner body reconstruction accuracy, especially inner body embedding in the canonical model as shown in Figure 7(left). Third, to better evaluate our method, we add two additional comparisons with state-of-the-art real-time non-rigid reconstruction methods (Section 6.3). Fourth, we include the analysis and comparison with the state-of-the-art learning-based methods that also reconstruct the body shape and pose (Section 2 and Section 6.3). Finally, we present two novel applications: (1) convenient human body measurement, and (2) body shape retargeting, which further demonstrate the practicability and effectiveness of our system (Section 7).

2 Related Work

In this work, we focus on capturing the dynamic geometry of human performer with detailed surface and personal body shape identity using a single depth sensor. The related methods can roughly divided into static template based, model-based, free-form and learning-based reconstruction methods.

Static template based dynamic reconstruction. For performance capture, some of the previous works leverage pre-scanned templates. Thus surface reconstruction is turned into a motion tracking and surface deformation problem. Vlasic *et al.* [23] and Gall *et al.* [2] adopted a template with embedded skeleton driven by multi-view silhouettes and temporal feature constraints. Liu *et al.* [24] extended the method to handle multiple interacting performers. Some approaches [25], [26] use a random forest to predict correspondences to a template, and use them to fit the template to the depth data. Ye *et al.* [5] considered the case of multiple Kinects input. Ye *et al.* [27] adopted a similar skinned model to estimate body shape and pose using a single depth camera in real-time.

Besides templates with an embedded skeleton, some works adopted template based non-rigid surface deformation. Li *et al.* [28] utilized embedded deformation graph in Sumner *et al.* [29] to parameterize the pre-scanned template to produce locally as-rigid-as-possible deformation. Guo *et al.* [12] adopted an ℓ_0 norm constraint to generate articulate motion without explicitly embedded skeleton. Zollhöfer *et al.* [11] took advantage of massive parallelism of GPU to enable real-time performance of general non-rigid tracking.

The aforementioned works require scanning a template step before capturing people with different identities or even the same performer with various apparels.

Model-based dynamic reconstruction. In addition to pre-scanned templates, many general body models have been proposed in the last decades. SCAPE [30] is a widely used model, it factorizes deformations into pose and shape components. SMPL [21] is a recent body model that represents shape and pose dependent deformations in an efficient linear formulation. Dyna [31] learned a low-dimensional subspace to represent soft-tissue deformations.

Many research works utilized these shape priors to enforce more general constraints to capture dynamic bodies. Chen *et al.* [32] adopted SCAPE to capture body motion using a single depth camera. Bogo *et al.* [13] extended SCAPE to capture detailed body shape with appearance. Bogo *et al.* [33] used SMPL to fit predicted 2D joint locations to estimate human shape and pose. However, neither SCAPE nor SMPL can represent arbitrary geometry of the performer wearing various apparels. In Zhang *et al.* [34] they addressed this problem by estimating the inner shape and recovering surface details. Pons-Moll *et al.* [10] introduce ClothCap, which jointly estimates clothing geometry and body shape using separate meshes. In both [34] and [10], results are only shown for complete 4D scan sequences. Alldieck *et al.* [35] reconstruct detailed shape including clothing from a monocular RGB video but the approach is off-line.

Free-form dynamic reconstruction. Free-form capture does not assume any geometric prior. For general non-rigid scenes, motion and geometry are closely coupled. In order to fuse regions visible in the future into a complete geometry, the algorithm needs to estimate non-rigid motion accurately. On the other hand, one needs accurate geometry to estimate motion accurately. In the last decades, many methods have been proposed to address free-form capture: linear variational deformation [36], deformation graph [37], subspace deformation [38], articulate deformation [39], [40] and [41], 4D spatio-temporal surface [42] and [43], incompressible flows [44], animation cartography [45], quasi-rigid motion [46] and directional field [47].

Only in recent years, free-form capture methods with real-time performance have been proposed. DynamicFusion [14] proposed a hierarchical node graph structure and an approximate direct GPU solver to enable capturing non-rigid scenes in real-time. Guo *et al.* [16] proposed a real-time pipeline that utilized shading information of dynamic scenes to improve non-rigid registration, meanwhile accurate temporal correspondences are used to estimate surface appearance. Innmann *et al.* [15] used SIFT features to improve tracking and Slavcheva *et al.* ([17] and [20]) leveraged the Killing constraint and variational level set method to handle topological changes and relatively fast non-rigid motion. However, none of these methods demonstrated full body performance capture with natural motion. Fusion4D [8] and Motion2Fusion [48] used multiple depth cameras to capture dynamic scenes with challenging motion in real-time. BodyFusion [18] utilized skeleton priors for human body reconstruction, while Li *et al.* [19] used an articulated motion prior generated from node-graph segmentation for general objects. Although they both achieved more robust motion tracking performance than [14] and [15], neither can handle challenging fast motion of the human body or guarantee plausible loop closure performance. Note that based on the proposed method, [49] has achieved more accurate motion tracking performance by combining multiple IMUs with the RGBD camera, and [50] has achieved more realistic cloth tracking results by utilizing physics-based cloth simulation.

Learning-based 3D body reconstruction. Learning-based 3D human body reconstruction has become a popular topic in recent years. Many works ([51], [52], [53], [54], [55], [56] and [57]) focus on inferring 3D human body shape and pose from a single RGB image or silhouettes. For example, Kanazawa *et al.* [51], Pavlakos *et al.* [52] and Omran *et al.* [56] integrated the SMPL model [21] within a deep neural network, and have shown the effectiveness of end-to-end frameworks for reconstructing a full 3D mesh of the human body from a single RGB image. Omran *et al.* [56] further demonstrated that, before lifting 2D to 3D, simplifying RGB images to semantic segmentations is beneficial. Additionally, they showed that when provided with a large amount of 2D annotations, only a small amount of 3D annotations are required for good performance. In contrast to mesh representations, Varol *et al.* [53] proposed a neural network for direct inference of volumetric body shape from a single image. By extending [33], Lassner *et al.* [54] generated a high-quality 3D body model for multiple human pose datasets, followed by training discriminative models with labels of 91 body landmark locations. They also validated the effectiveness of the 91-landmark pose estimator in terms of the accuracy of 3D human pose and shape optimization. Dibra *et al.* [55] used frontal and side silhouettes of the human body as input and inferenced the 3D body mesh directly using cross-modal neural networks and generative HKS descriptors. [57] used SMPL model as a complete-body-prior for volumetric inference of the 3D human body (including the real world geometry of the cloth). Benefiting from the rapid development of deep learning techniques and the large amount of available training data, these methods have achieved impressive body reconstruction results in a convenient and practical manner, even under challenging conditions. However, the main drawback of these methods is the temporally incoherent pose and shape reconstruction results when the methods are applied independently frame by frame. Moreover, most of the methods cannot achieve real-time performance. Finally, there is an additional line of learning-based methods, including [58] and [59], that have achieved realistic inference of cloth dynamics, but these methods still need either real captured high-quality

4D sequences or simulated cloth dynamics as training data to train specific models for different types of cloth.

3 Overview

3.1 Double-layer Surface Representation

The input to DoubleFusion is a depth stream captured from a single consumer-level depth sensor and the output is a double-layer representation of the performer. The outer layer are observable surface regions, such as clothing, visible body parts (e.g., face and hair), while the inner layer is a parametric human shape and skeleton model based on the skinned multi-person linear model (SMPL [21]). Similar to previous work [14], the motion of the outer surface is parametrized by a set of nodes. Every node deforms according to a rigid transformation. The *node graph* interconnects the nodes and constrain them to deform similarly. Unlike [14] that uniformly samples nodes on the newly fused surface, we pre-define an on-body node graph on the SMPL model, which provides a semantic and real prior to constrain non-rigid human motion. For example, it will prevent erroneous connections between body parts (e.g., connecting the legs). We uniformly sample on-body nodes and use geodesic distances to construct the predefined on-body node graph on the mean shape of SMPL model as shown in Figure 2(a)(top). The on-body nodes are inherently bound to skeleton joints in the SMPL model. Outer surface regions that are close to the inner body are bound to the on-body node graph. Deformations of regions far from the body cannot be accurately represented with the on-body graph. Hence, we additionally sample far-body nodes with a radius of $\delta = 5cm$ on the newly fused far-body geometry. A vertex is labeled as far-body when it is located further than $1.4 \times \delta cm$ from its nearest on-body node, which helps to make sure the sampling scheme is robust against depth noise and tracking failures. The double node graph is shown in Figure 2(d)(bottom).

3.2 Inner Body Model: SMPL

SMPL [21] is an efficient linear body model with $N = 6890$ vertices. SMPL incorporates a skeleton with $K = 24$ joints. Each joint has 3 rotational Degrees of Freedom (DoF). Including the global translation of the root joint, there are $3 \times 24 + 3 = 75$ pose parameters. Before posing, the body model $\bar{\mathbf{T}}$ deforms according to shape parameters $\boldsymbol{\beta}$ and pose parameters $\boldsymbol{\theta}$ to accommodate for different identities and non-rigid pose dependent deformations. Mathematically, the body shape $T(\boldsymbol{\beta}, \boldsymbol{\theta})$ is morphed according to

$$T(\boldsymbol{\beta}, \boldsymbol{\theta}) = \bar{\mathbf{T}} + B_s(\boldsymbol{\beta}) + B_p(\boldsymbol{\theta}) \tag{1}$$

where $B_s(\boldsymbol{\beta})$ and $B_p(\boldsymbol{\theta})$ are vectors of vertex offsets, representing shape blendshapes and pose blendshapes respectively. The posed body model $M(\boldsymbol{\beta}, \boldsymbol{\theta})$ is formulated as

$$M(\boldsymbol{\beta}, \boldsymbol{\theta}) = W(T(\boldsymbol{\beta}, \boldsymbol{\theta}), J(\boldsymbol{\beta}), \boldsymbol{\theta}, \mathcal{W}) \tag{2}$$

where $W(\cdot)$ is a general blend skinning function that takes the modified body shape $T(\boldsymbol{\beta}, \boldsymbol{\theta})$, pose parameters $\boldsymbol{\theta}$, joint locations $J(\boldsymbol{\beta})$ and skinning weights $\mathcal{W}$, and returns posed vertices. Since all parameters were learned from data, the model produces very realistic shapes in different poses. We use the open sourced SMPL model with 10 shape blendshapes. See [21] for more details.v

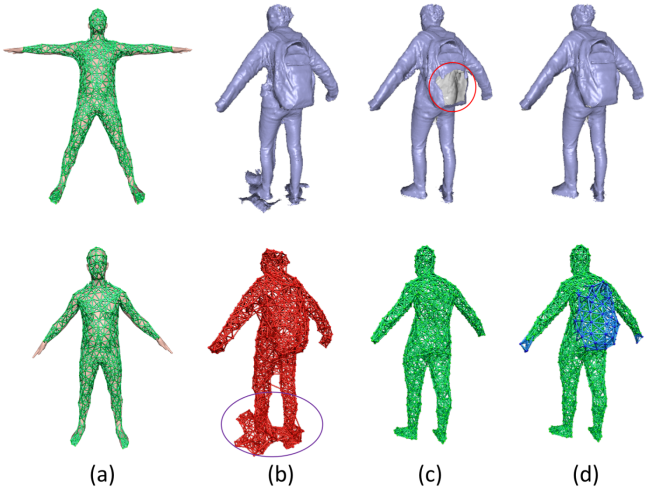


Fig. 2. (a) Initialization of the on-body node graph. (b,c,d) Evaluation of the double node graph. The figure shows the geometry results and live node graph of (b) traditional free-form sampled node graph (red), (c) on-body node graph (green) only and (d) double node graph (with far-body nodes in blue). Note that we render the inner surface of the geometry in gray in (c)(top).

3.3 Initialization

During capture, we assume a fixed camera position and treat camera movement as global scene rigid motion. In the initialization step, we require the performer to start with a rough A-pose. For the first frame, we initialize the TSDF in the canonical volume by projecting the depth map into the volume. Then we use the proposed inner body optimization method (Section 5.2) to estimate initial shape parameters $\boldsymbol{\beta}_0$ and pose parameters $\boldsymbol{\theta}_0$ of the parametric body model. After that, we initialize the double node graph using the on-body node graph and initial pose and shape as shown in Figure 2(a)(bottom). We extract a triangle mesh from the volume using Marching Cube algorithm [60] and sample additional *far-body nodes*. These nodes are used to parameterize non-rigid deformations far from inner body shape.

3.4 Main Pipeline

The main challenge to adopt SMPL in our pipeline is that initially the incomplete outer surface leads to difficult model fitting. Our solution is to continuously update the shape and pose in the canonical frame when more geometry is fused. Therefore, we propose a pipeline that executes *joint motion tracking*, *outer-layer geometry fusion* and *inner body optimization* sequentially (Figure 3). We briefly introduce each component of the pipeline below:

Joint Motion Tracking Given the current estimated parameters of body shape, we jointly optimize live body pose and the non-rigid deformations defined by the double node graph (Section 4). For the on-body nodes, we constrain the non-rigid deformations of them to follow skeletal motion. The far-body nodes are also optimized in the process but are not constrained by the skeleton.

Outer-Layer Geometry Fusion Similar to previous work [14], we non-rigidly integrate depth observation of multiple frames in the canonical volume (Section 5.1). We also explicitly detect collided voxels to avoid erroneously fused geometry [16].

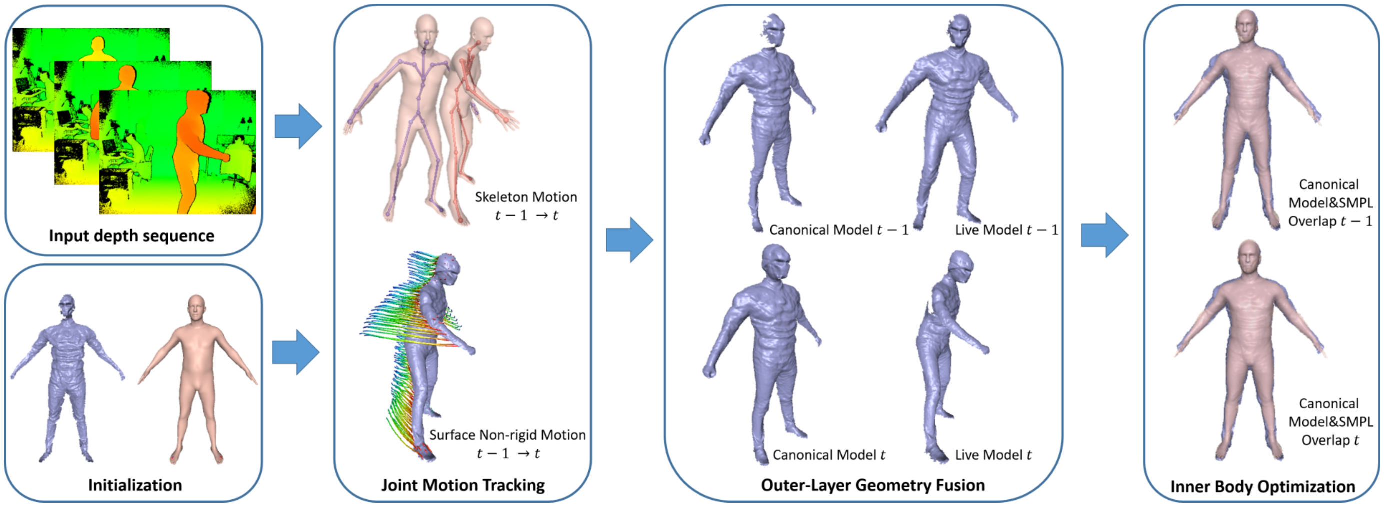


Fig. 3. The pipeline of our system. We first initialize the system using the first depth frame (Section 3.3). Then for each frame, we sequentially perform the next 3 steps: joint motion tracking (Section 4), outer-layer geometry fusion (Section 5.1) and inner body optimization (Section 5.2).

Inner Body Optimization After outer-layer geometry fusion, the surface of the canonical model gets more complete. We directly optimize the body shape and pose by using the updated TSDF in the canonical volume and the live depth input to generate more accurate inner body shape reconstruction results very efficiently (Section 5.2).

4 Joint Motion Tracking

There are two parameterizations in our motion tracking component, skeletal motion and non-rigid node deformations. Similar to the previous work [18], we adopt a binding term that constrains both motions to be consistent. Different from [18], we only enforce the binding term on on-body nodes to penalize non-articulated motion on on-body nodes. In contrast, far-body nodes have independent non-rigid deformations which are regularized to move like other nodes in the same graph structure. Besides geometric regularization, we also follow previous work [33] to use a statistic pose prior to prevent unnatural poses. The energy of joint optimization is then

$$E_{\mathrm{mot}} = \lambda_{\mathrm{data}} E_{\mathrm{data}} + \lambda_{\mathrm{bind}} E_{\mathrm{bind}} + \lambda_{\mathrm{reg}} E_{\mathrm{reg}} + \lambda_{\mathrm{pri}} E_{\mathrm{pri}}, \tag{3}$$

where E_{data}, E_{bind}, E_{reg} and E_{prior} are energies of data, binding, regularization and pose prior term respectively.

Data Term The data term measures the fit between the reconstructed double-layer surface and the depth map:

$$\begin{aligned} E_{\mathrm{data}} = \sum_{(v_c,u)\in\mathcal{P}} & \tau_1(\mathbf{v}_c)\psi(\tilde{\mathbf{n}}_{v_c}^{\mathrm{T}}(\tilde{\mathbf{v}}_c - \mathbf{u})) + \\ & (\tau_2(\mathbf{v}_c) + \tau_3(\mathbf{v}_c))\psi(\hat{\mathbf{n}}_{v_c}^{\mathrm{T}}(\hat{\mathbf{v}}_c - \mathbf{u})), \end{aligned} \tag{4}$$

where $\mathcal{P}$ is the correspondence set, $\psi(t) = t^2/(2(1+t^2))$ is the robust Geman-McClure penalty function, and we approximate the t in the denominator as the initial data energy in each ICP step for simplicity. $(\mathbf{v}_c, \mathbf{u})$ is a correspondence pair, $\mathbf{u}$ is a sampled point on the depth map whose closest point $\mathbf{v}_c$ can be on either the inner body shape or fused outer surface, including the near-body surface (surface around the on-body node graph) and far-body surface (surface around the far-body node graph). Specifically, we use all the fused outer surface as candidates for reconstructing the non-rigid tracking data terms and use both the inner body shape and the near-body surface to reconstruct the skeleton tracking data terms. $\tau_1(\mathbf{v}_c)$, $\tau_2(\mathbf{v}_c)$ and $\tau_3(\mathbf{v}_c)$ are correspondence indicator functions: $\tau_1(\mathbf{v}_c)$ equals 1 only if $\mathbf{v}_c$ is on the fused outer surface; $\tau_2(\mathbf{v}_c)$ equals 1 when $\mathbf{v}_c$ is on the inner body shape; $\tau_3(\mathbf{v}_c)$ equals 1 when $\mathbf{v}_c$ is on the near-body surface. Correspondences on the inner body shape enable fast and robust skeleton tracking when $\tau_2(\mathbf{v}_c) = 1$, while correspondences on the fused outer surface provide more accurate skeleton alignment and non-rigid registration results when $\tau_1(\mathbf{v}_c) = 1$ and $\tau_3(\mathbf{v}_c) = 1$. $\tilde{\mathbf{v}}_c$ and $\tilde{\mathbf{n}}_{v_c}$ are the vertex position and normal warped by its knn-nodes using dual quaternion blending and defined as

$$\mathbf{T}(\mathbf{v}_c) = SE3\Big(\sum_{k\in\mathcal{N}(v_c)} \omega(k, v_c)\, \mathbf{dq}_k\Big), \tag{5}$$

where $\mathbf{dq}_j$ is the dual quaternion of jth node; $SE3(\cdot)$ maps a dual quaternion to $\mathbf{SE}(3)$ space; $\mathcal{N}(v_c)$ represents a set of node neighbors of $\mathbf{v}_c$; $\omega(k, v_c) = \exp(-\|\mathbf{v}_c - \mathbf{x}_k\|_2^2/(2r_k^2))$ is the influence weight of the kth node $\mathbf{x}_k$ to $\mathbf{v}_c$; we set the influence radius $r_k = 0.075$m for all nodes. $\hat{\mathbf{v}}_c$ and $\hat{\mathbf{n}_{v_c}}$ are the vertex position and its normal skinned by skeleton motion using linear blend skinning (LBS) and defined as

$$\begin{aligned} \mathbf{G}(\mathbf{v}_c) &= \sum_{i\in\mathcal{B}} w_{i,v_c}\, \mathbf{G}_i, \\ \mathbf{G}_i &= \prod_{k\in\mathcal{K}_i} \exp(\theta_k \hat{\xi}_k), \end{aligned} \tag{6}$$

where $\mathcal{B}$ is index set of bones; $\mathbf{G}_i$ is the cascaded rigid transformation of ith bone; w_{i,v_c} is the skinning weight associated with ith bone and point $\mathbf{v}_c$; $\mathcal{K}_i$ is parent indices of ith bone in the backward kinematic chain; $\exp(\theta_k \hat{\xi}_k)$ is the exponential map of the twist associated with kth bone. Note that the skinning weights of $\mathbf{v}_c$ is given by the weighted average of the skinning weights of its knn-nodes.

For each $\mathbf{u}$ on the depth map, we search for two types of correspondences on our double layer surface: $\mathbf{v}_t$ on the body shape and $\mathbf{v}_s$ on the fused surface. We choose the one that maximizes the following metric based on Euclidean distance and normal affinity

$$c = \underset{i\in\{t,s\}}{\operatorname{argmax}} \left(\left(1 - \frac{\|\mathbf{v}_i - \mathbf{u}\|_2}{\delta_{\max}}\right)^2 + \mu\, \tilde{\mathbf{n}}_{v_i}^{\mathrm{T}} \mathbf{n}_u \right), \quad (7)$$

where we choose $\mu = 0.2$; we set $\delta_{\max} = 0.1\mathrm{m}$ as the maximum radius used to search correspondences. We adopt two strategies for correspondence searching. To find correspondences between the depth map and the fused surface, we project the fused surface to 2D and then find correspondences within a local search window. For correspondences between the depth map and the body shape, we first find the nearest on-body node and then search for the nearest vertex around it. We eliminate the correspondences with distance bigger than δ_{max}. These two methods are efficient for real-time performance and avoid building complex space partitioning data structure on GPU.

Binding Term The binding term attaches on-body nodes to their nearest bones and helps to produce articulated deformations on the body. It is defined as

$$E_{\mathrm{bind}} = \sum_{i\in\mathcal{L}_s} \|\mathbf{T}_i \mathbf{x}_i - \hat{\mathbf{x}}_i\|_2^2, \quad (8)$$

where $\mathcal{L}_s$ is the index set of on-body nodes. $\hat{\mathbf{x}}_i$ is the node position skinned by LBS as defined in Equation 6.

Regularization Term The graph regularization is defined on all of the graph edges. This term is used to produce locally as-rigid-as-possible deformations. For on-body node graph, we decrease the effects of this regularization around joint regions by comparing the skinning weight vector of neighboring nodes as in [18]. This term is then defined as

$$E_{\mathrm{reg}} = \sum_i \sum_{j\in\mathcal{N}(i)} (1 - \rho(\|W_i - W_j\|_2^2)) \|\mathbf{T}_i \mathbf{x}_j - \mathbf{T}_j \mathbf{x}_j\|_2^2 \quad (9)$$

where $\mathbf{T}_i$ and $\mathbf{T}_j$ are transformation associated with ith and jth nodes; W_i and W_j are skinning weight vectors of these two nodes respectively; $\rho(\cdot)$ is the huber loss function with threshold 0.5. Around joint regions, if two neighbor nodes are on different body parts, the difference of the skinning weight vectors is large, and thus $\rho(\cdot)$ will decrease the effect of the regularization. This will help to produce articulated deformations of on-body node graph. For far-body node graph, we construct its regularization term similar to [14].

Pose Prior Term Similar to [33], we include a pose prior penalizing the unnatural poses. It is defined as

$$E_{\mathrm{prior}} = -\log\Big(\sum_j \omega_j N(\boldsymbol{\theta}; \mu_j, \delta_j)\Big). \quad (10)$$

This is formulated as a Gaussian Mixture Model (GMM), where ω_j, μ_j and δ_j is the mixture weight, the mean and the variance of jth Gaussian model.

We solve Equation 3 using Iterative Closest Point (ICP) method. We first construct the correspondence set $\mathcal{P}$ based on the latest motion, then solve the generated non-linear least squares problem using Gauss-Newton method. We use twist to represent the rigid transformations of both the joint and node. Within each iteration of Gauss-Newton method, the rigid transformations are approximated by one-order Taylor expansion around the latest transformation. The resulting linear system is then solved by a custom designed highly efficient preconditioned conjugate gradient (PCG) solver on GPU as in [16] and [8].

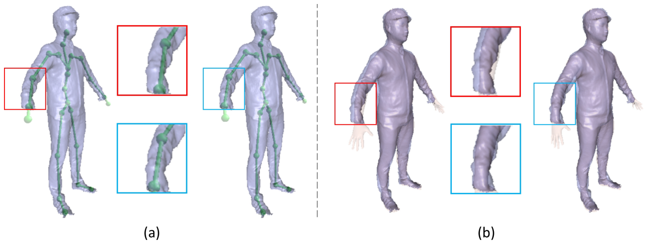


Fig. 4. Illustration of inner body optimization. (a) Skeleton embedding results before and after optimization. (b) Overlay of the body shape and outer surface before and after optimization.

5 Double-layer Surface Reconstruction

5.1 Outer-layer Geometry Fusion

Similar to the previous non-rigid fusion works [14], [15] and [16], we integrate the depth information into a canonical volume for reconstructing the outer surface. First, the voxels in the canonical volume are warped to live frame according to the current non-rigid warp field. Then, we calculate the PSDF values for the valid voxels and then update their TSDF values. We follow the work in [16] to cope with collided voxels in live frame to prevent erroneous fusion results caused by voxel collisions. We also constraint the fusion to those voxels that are within δ distance from any nodes.

Dynamic Detail Deformation One of the drawbacks of previous fusion-based methods is that they fail to reconstruct dynamic geometric details exhibited in the input depth sequence (e.g., the dynamic wrinkles of cloth). Classical TSDF fusion methods proposed in KinectFusion [61] and DynamicFusion [14] fuse all the depth frames using a relatively uniform integration weight to complete the surface and filter depth noise. This is an especially useful scheme for loop closure reconstruction since we can substantially decrease the accumulated tracking error by registering the fused (denoised) surface model to the noisy depth input (frame-to-model). By contrast, if the registration is performed between two continuous noisy depth frames, the tracking and loop closure performence of such systems will be severely degraded due to the incorporated depth noise as shown in Figure 8 of [61]. However, fusing all the depth frames together with a relatively uniform integration weight considerably smooths the dynamic geometric details, as shown in Figure 5(e). Moreover, although we can obtain dynamic geometric details of the cloth by simply assigning a higher integration weight to the current depth frame, the depth noise will also be substantially incorporated into the fusion and tracking steps, thus degrading system performance, especially the loop closure performance, as shown in Figure 5(f) (the left arm of the subject is not faithfully loop closed and is much thinner than other results). Other than that, the uniformly sampled sparse node graph used in recent real-time non-rigid reconstruction systems has limited freedom to describe the detailed non-rigid deformation of the cloth under the real-time budget.

To reconstruct vivid dynamic details while maintaining robust loop closure performance, we create a partial fusion volume that integrates only the lastest p depth frames. In each frame, the *"super depth"* map is generated by rendering the partial fused geometry, which is extracted from the partial fusion volume, as a vertex map. The map contains similar abundant dynamic geometric details as in a single depth frame but has much less noise. These noiseless details are then transferred onto the live surface through dynamic detail deformation, which deforms the live surface to align with the extracted super depth map. Note that the proposed dynamic detail deformation method is mainly used to recover those dynamic surface details that can be fused on the "super depth" but were smoothed out on the fully-fused-surface (which also means those detailed deformations that cannot be well described by the sparsely sampled node graph). Thus, this method will not substantially improve the surface tracking accuracy for the relatively rigid surfaces such as tight cloth and body skin. Moreover, although previous method [34] has evaluated the effectiveness of using the dynamic detailed surfaces under different poses for optimizing accurate body shapes, they still need to use a bundle optimization strategy after capturing the whole sequence and thus cannot achieve real-time performance. In our system, we only use the TSDF volume (but not the detailed surfaces after dynamic detail deformation) and the live depth input for inner body optimization for achieving real-time performance as explained in Section 5.2. To maintain the real-time performance of our system, our dynamic detail deformation is designed to avoid solving a large-scale nonlinear optimization problem, as in [10]. Instead, we stretch each vertex $\mathbf{v}_l$ on the live surface to its corresponding vertex $\mathbf{v}_s$ on the super depth map in an iterative and incremental manner. Specifically, in iteration t ($t > 1$), we have $\mathbf{v}_l^t = \mathbf{v}_l^{t-1} + s \cdot (w\Delta\mathbf{v}_l^t + \frac{1-w}{|N(v_l)|}\Sigma_{v_{n_i}\in N(v_l)}\Delta\mathbf{v}_{n_i}^t)$, where $N(v_l)$ denotes the neighbor vertex set of $\mathbf{v}_l$, $|N(v_l)|$ is the number of the neighbors, $\mathbf{v}_{n_i}$ is the ith neighbor of $\mathbf{v}_l$, s, which is set to 0.1, controls the deformation step in each iteration, w is the spatial smoothness weight, which is set to 0.6, and $\Delta\mathbf{v}_l^t = \mathbf{v}_s - \mathbf{v}_l^{t-1}$ is the vector from $\mathbf{v}_l^{t-1}$ to $\mathbf{v}_s$ and $\Delta\mathbf{v}_{n_i}^t$ is the vector from $\mathbf{v}_{n_i}^{t-1}$ to its corresponding super depth vertex. We search for the correspondences using the projective block searching scheme from [16] and perform 20 iterations in total. Note that for vertices with no corresponding super depth vertex, we simply set their correspondences as themselves. This incremental iterative stretching scheme is effective in our system since the source mesh (live surface) and the target mesh (super depth) are quite close spatially. As shown in Figure 5(d) and Figure 15, by using the proposed dynamic detail deformation approach, we can obtain accurate dynamic surface reconstruction results while maintaining robust capture performance.

5.2 Inner Body Optimization

Due to the limited observation of the initial frame and the A-pose required for system initialization, the initialized shape and pose parameters $(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0)$ in the canonical volume may not accurately explain the later depth observations as shown in the left part of Figure 4(a) and (b). After the surface fusion step, we may have an updated outer surface in the canonical volume with more complete geometry. On the one hand, the updated geometry in the canonical volume can provide much more information under the initial A-pose to optimize the embedded body model. On the other hand, live input depth frames provide additional information on the inner body shape under different poses. Therefore, utilizing both the evolving outer surface and the live depth observations of different poses can substantially improve the accuracy of the inner body reconstruction and its canonical volume embedding. We propose a novel algorithm that can efficiently optimize the shape parameters $\boldsymbol{\beta}$ and canonical pose parameters $\boldsymbol{\theta}$ of the parametric body model jointly by means of the updated TSDF in the canonical volume and the live depth input in every frame. The formulation of the energy is then

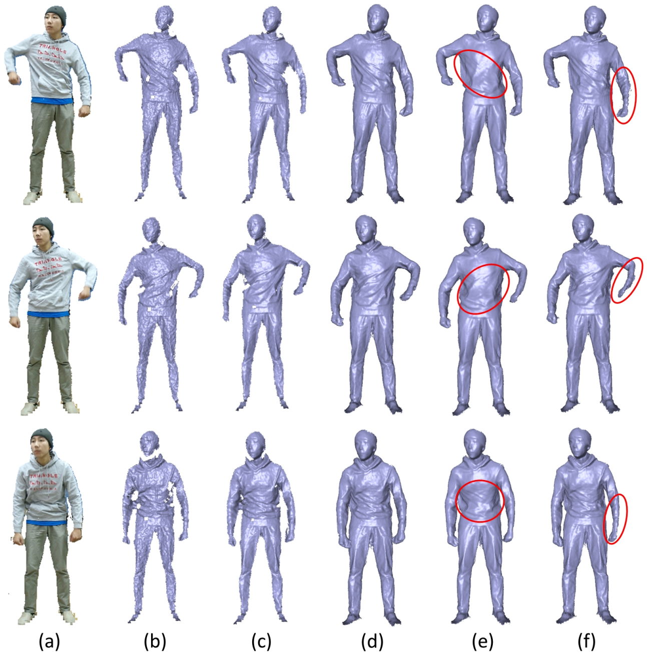


Fig. 5. Illustration of partial geometry fusion and dynamic detail deformation. (a,b) Color reference and depth input with different poses; the wrinkles on the cloth are significantly different from each other. (c) Partial fused geometry, which contains plausible dynamic geometric details and much less noise than the depth input. (d) Live geometry after dynamic detail deformation; the dynamic cloth wrinkles are faithfully reconstructed. (e) Live geometry without detailed deformation [22]; although the relatively static geometry details on the face and trousers are reconstructed, the dynamic cloth wrinkles cannot be reconstructed faithfully. (f) Live geometry reconstruction results by implementing a larger current integration weight (8) in the TSDF fusion step. Note that the left arm of the subject is not faithfully reconstructed due to the deteriorated loop closure performance.

$$E_{\text{shape}} = E_{\text{sdata}} + E_{\text{sreg}} + E_{\text{pri}}, \tag{11}$$

where E_{sdata} is a shape optimization data term based on the continuously updated canonical volume and the live depth input and E_{sreg} is a temporal constraint term that makes the new shape and canonical pose parameters consistent with the previous ones. E_{pri} is the same as in Equation 3 to prevent unnatural canonical poses. The novel shape optimization data term is defined as

$$E_{\text{sdata}} = E_{\text{volume}} + E_{\text{depth}}, \tag{12}$$

where E_{volume} measures the misalignment between the canonical body and the TSDF volume and E_{depth} measures the misalignment between the live body and the live depth input. An illustration of E_{volume} is shown in Figure 6. Note that although a complete canonical outer surface in the A-pose can provide a good constraint for body optimization,

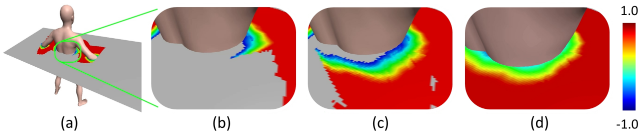


Fig. 6. Illustration of inner body optimization for the volume data term. (a) Canonical body model and a slice of the TSDF volume, with color-coded normalized TSDF value mapping from [-1, 1] to the HSV color space. Gray color represents regions without observations. (b,c,d) Temporal inner body optimization results around the waist. The inner body becomes increasingly accurate as more regions in the canonical volume are observed.

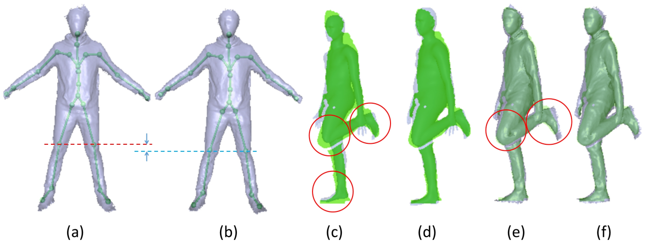


Fig. 7. Evaluation of the live-depth-based energy in the body optimization step. (a,b) Canonical model and inner body skeleton overlay (the skeleton embedding of the canonical model) without/with live-depth-based energy. (c,d) Live body and live depth silhouette (green) overlay without/with live-depth-based energy. (e,f) Live model and depth silhouette (green) overlay without/with live-depth-based energy.

the simple A-pose is not sufficient for accurate body shape optimization, especially for accurate body/skeleton embedding in the canonical volume. Therefore, we incorporate live depth input, which contains visual cues extracted from various poses, to construct a new energy term for accurate shape optimization. Specifically, we first calculate a $2D$ distance transform map for the live depth silhouette and then use the map to construct a silhouette term for shape optimization. Then, we calculate the $3D$ point-to-plane errors between the live body model and the valid live depth point cloud to obtain better $3D$ fitting results. Figure 7 shows that by incorporating the live-depth-based inner body optimization energy, we can obtain more accurate body shape reconstruction and skeleton embedding in the canonical frame (b,d) and, thus, more accurate large deformation results around the left knee (f). The quantitative evaluation of the live-depth-based energy term is shown in Figure 14. Note that in the surface tracking step, the fitting between the fused surface and the depth input mainly determines the non-rigid tracking accuracy, and only using the TSDF volume for body optimization will also generate robust surface tracking results as demonstrated in the preliminary version. However, for specific regions (such as the large bending regions around joints), a more accurate body shape (skeleton embedding) guarantees more accurate pose tracking results as shown in Figure 7 and the supplementary video. Moreover, for those cases that we cannot fuse a complete TSDF volume for body shape optimization (e.g., the subject never turns around in the whole sequence), the incorporation of the live depth energy will improve the surface tracking accuracy by estimating a more accurate body shape as shown in Figure 16 and Table 1.

The inner body optimization loss funciton is defined as

$$E_{\text{volume}}(\boldsymbol{\beta},\boldsymbol{\theta}) = \sum_{\bar{\mathbf{v}}\in\bar{\mathbf{T}}} \eta\cdot\psi(\mathbf{D}(W(T(\bar{\mathbf{v}};\boldsymbol{\beta},\boldsymbol{\theta});J(\boldsymbol{\beta}),\boldsymbol{\theta}))), \tag{13}$$

which is used to optimize canonical volume body embedding (body shape and canonical body pose) jointly. Where $\psi(\cdot)$ is the Geman-McClure penalty function as in Equation 4, $\mathbf{D}(\cdot)$ is a bilinear sampling function that takes a point in the TSDF volume and returns the interpolated TSDF value. Note that $\mathbf{D}(\cdot)$ returns valid distance values only when the knn-nodes of the given point are all on-body nodes; otherwise $\mathbf{D}(\cdot)$ returns 0. This prevents the body shape from incorrectly fitting exterior objects, e.g., the backpack a performer is wearing. η is a nonsymmetric weight for general capture: for all the sequences, we set $\eta = 1.5$ when $\mathbf{D}(\cdot) > 0$ (which means the inner body vertex is located outside the outer surface) and $\eta = 1.0$ when $\mathbf{D}(\cdot) <= 0$. By using this nonsymmetric weight, we are able to obtain more plausible inner body optimization results when people wear casual clothing. $\mathbf{v} = T(\bar{\mathbf{v}};\boldsymbol{\beta},\boldsymbol{\theta})$ modifies $\bar{\mathbf{v}}$ by shape blend shape and canonical body pose $\boldsymbol{\theta}$; $W(\mathbf{v};J(\boldsymbol{\beta},\boldsymbol{\theta}),\boldsymbol{\theta})$ deforms $\mathbf{v}$ using linear blend skinning.

The live-depth-based shape optimization energy is defined as

$$E_{\text{depth}}(\boldsymbol{\beta}) = \sum_{\bar{\mathbf{v}}\in\bar{\mathbf{T}}} \|\mathbf{DT}(\mathbf{P}(\tilde{\mathbf{v}}))\|_2^2 + \|\mathbf{dn}^T(\tilde{\mathbf{v}}-\mathbf{dv})\|_2^2, \tag{14}$$

which is used to optimize body shape according to the live depth input and the tracked body pose. Here, $\tilde{\mathbf{v}} = T(\bar{\mathbf{v}};\boldsymbol{\beta},\tilde{\boldsymbol{\theta}})$ modifies $\bar{\mathbf{v}}$ by shape blending and live pose $\tilde{\boldsymbol{\theta}}$, $\tilde{\mathbf{u}} = \mathbf{P}(\tilde{\mathbf{v}})$ projects live body vertex $\tilde{\mathbf{v}}$ onto the live depth image plane and acquires its $2D$ pixel coordinates $\tilde{\mathbf{u}}$. $\mathbf{DT}(\tilde{\mathbf{u}})$ is a sampling function that returns the distance transform value at $\tilde{\mathbf{u}}$ on the distance transform map using bilinear interpolation. $\mathbf{dv}$ and $\mathbf{dn}$ are the vertex position and normal of the live depth input, respectively.

The temporal regularization is defined as

$$E_{\text{sreg}}(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\beta}',\boldsymbol{\theta}') = \gamma_1\|\boldsymbol{\beta}-\boldsymbol{\beta}'\|_2^2 + \gamma_2\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|_2^2. \tag{15}$$

This term prevents the optimized shape and pose parameters $(\boldsymbol{\beta},\boldsymbol{\theta})$ from deviating the ones $(\boldsymbol{\beta}',\boldsymbol{\theta}')$ of the previous frame.

Note that $T(\bar{\mathbf{v}};\boldsymbol{\beta},\boldsymbol{\theta})$ includes both the pose and shape parameters, which makes $W(\mathbf{v};J(\boldsymbol{\beta},\boldsymbol{\theta}),\boldsymbol{\theta})$ a non-linear function. We find that generally the pose blend shape $B_p(\boldsymbol{\theta})$ in $T(\bar{\mathbf{v}};\boldsymbol{\beta},\boldsymbol{\theta})$ contributes much less to the modified body shape compared with the shape blend shape. Therefore we ignore the pose blend shape in $T(\bar{\mathbf{v}};\boldsymbol{\beta},\boldsymbol{\theta})$, and the resulting skinning formulation $W(T(\bar{\mathbf{v}};\boldsymbol{\beta});J(\boldsymbol{\beta},\boldsymbol{\theta}),\boldsymbol{\theta})$ becomes a linear function of $(\boldsymbol{\beta},\boldsymbol{\theta})$. This will generate a better energy landscape for the sampling based energy (Equation 13 and Equation 14) and make the convergence faster. Then we solve the resulting energy using the same GPU-based Gauss-Newton solver as in Section 4. At last, we update the body shape and pose that embedded into the TSDF volume and update the non-rigid and skeleton motion field. As shown in Figure 4(b), the reconstructed shape and pose (skeleton embedding) get more accurate by using the proposed inner body optimization method.

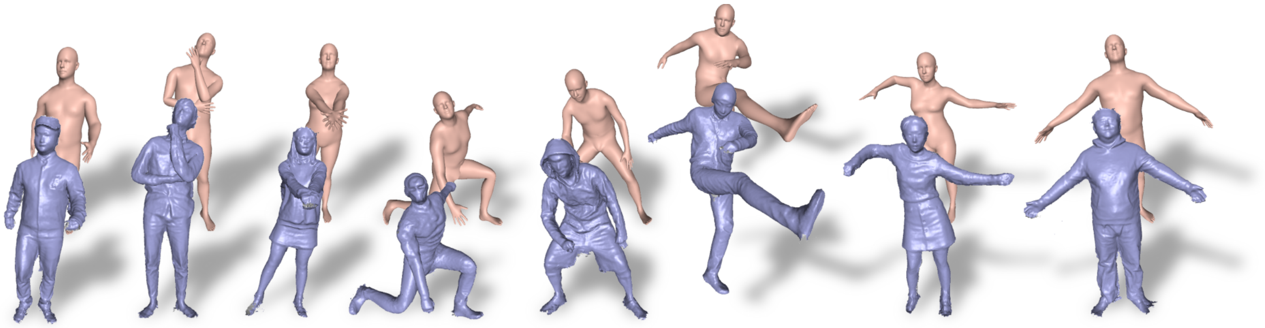


Fig. 8. Example results reconstructed by our system. Our system is capable of reconstructing different types of cloth and various body shapes.

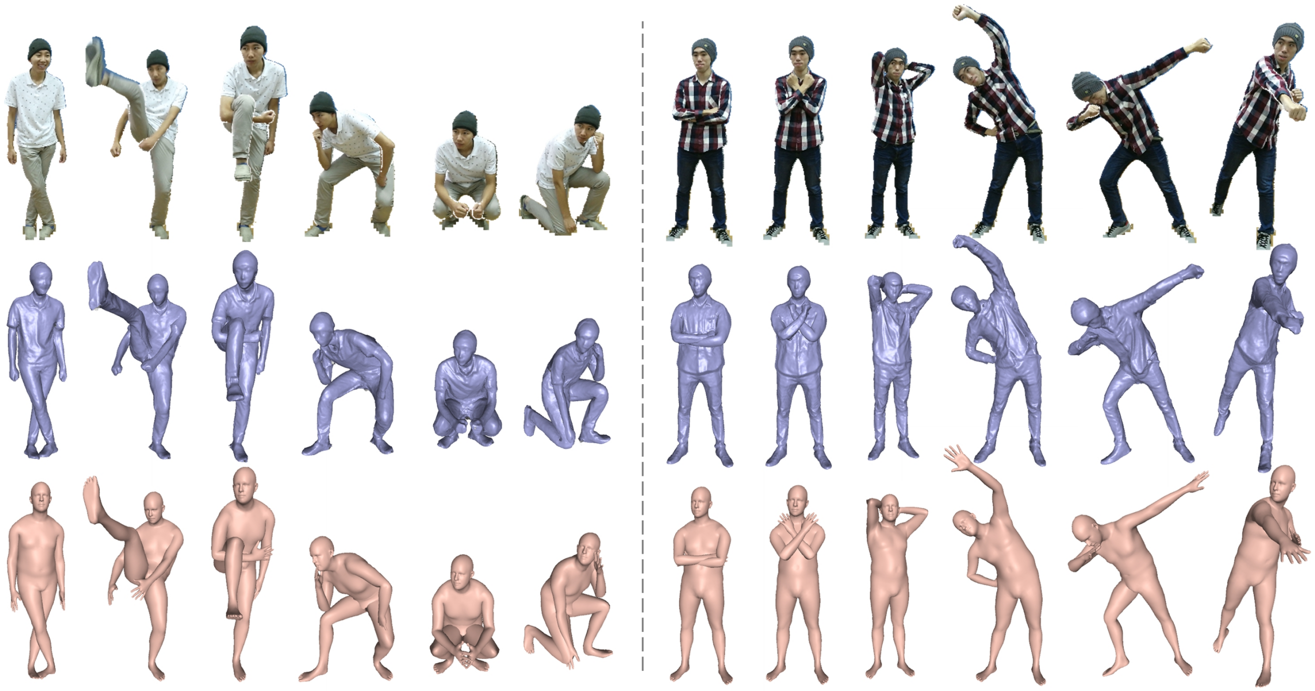


Fig. 9. Sequential reconstruction results on two sequences. The first row is the color reference (without background), and reconstruction results are presented in the last two rows. Our system is able to capture fast and complicated human motion as well as detailed surface geometries.

6 Results

In this section, we first report the performance and the main parameters of the system. Then we compare with current state-of-the-art methods qualitatively and quantitatively. We also evaluate each of our main contributions. In Figure 8, we demonstrate the reconstruction results of different subjects. Note the various shapes, challenging motion and different types of cloth of the loop closed model that we can reconstruct. In Figure 9, we demonstrate the sequential reconstruction results of two subjects. Note the complicated poses and detailed clothing geometries that we reconstructed.

6.1 Performance

DoubleFusion runs in real-time (running at 33ms per frame). The entire pipeline is implemented on one NVIDIA TITAN Xp GPU. Executing 6 Gauss-Newton iterations, the joint motion tracking takes 21.4 ms. The geometric fusion takes 6.1 ms and inner body optimization takes 3.4 ms. Prior to the joint motion tracking, we perform preprocessing for the input depth frame, which includes bilateral filtering, floor removal and distance transform calculation. After inner body optimization, a triangulated mesh is extracted, non-rigidly transformed into the live camera coordinates (using non-rigid warping and the proposed dynamic detail deformation approach) and rendered on the screen. The two parts above run asynchronously (in another CUDA stream) with the main pipeline, and the actural cost is less than 1 ms. For all of our experiments, we choose $\lambda_{\text{data}} = 1.0$, $\lambda_{\text{bind}} = 1.0$, $\lambda_{\text{reg}} = 5.0$, $\lambda_{\text{pri}} = 0.01$ and $p = 8$. For each vertex, we use its 4 nearest neighbors for warping; for each node, we use its 8 nearest neighbors to construct the node graph. The size of the voxel is set to 4 mm in each dimension.

6.2 Evaluation

Double Node Graph We evaluate the proposed double node graph in Figure 2. The standard node graph construction scheme [29] uniformly samples all the nodes on the fused outer surface. The lack of semantic information results in wrong connections (connection between two legs) and

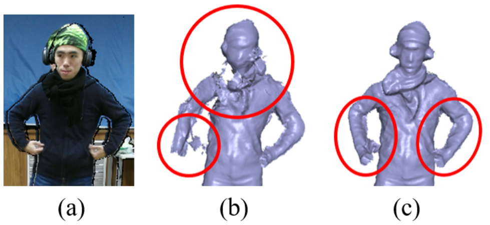


Fig. 10. Evaluation of joint motion tracking. (a) Reference color image. (b) Results using only correspondences on the body for skeleton tracking, without non-rigid registration. (c) Searching correspondences on both the body and fused surface for skeleton tracking, without non-rigid registration. (d) Using full energy terms.

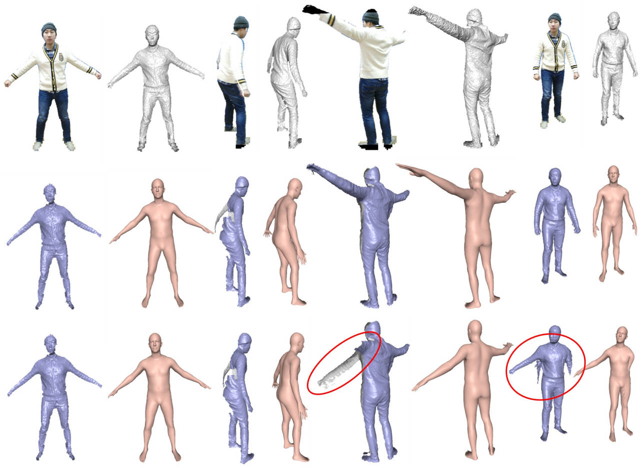


Fig. 11. Evaluation of on-body correspondences. The first row are reference color images and depth inputs. The second and third row are sequential double-layer surface reconstruction results with and without on-body correspondences, respectively.

erroneous fusion results as shown in Figure 2(b). Using the on-body node graph alone is limited to capturing relatively tight clothing (e.g. the incomplete geometry of the backpack in Figure 2(c)) since it is out of the control area of on-body node graph. By using the proposed double node graph (Figure 2(d)), we can get clean and complete results.

Joint motion tracking In Figure 10, we evaluate different components of the joint motion tracking step qualitatively. We eliminate non-rigid registration in Figure 10(b) and (c). In Figure 10(b), we only use correspondences on the body shape by setting $\tau_1(\mathbf{v}_c) \equiv 0, \tau_3(\mathbf{v}_c) \equiv 0$ in Equation 4. It shows that without detailed surface and non-rigid registration, although an approximate pose can be tracked, the fused surface is noisy and erroneous; In Figure 10(c), we use correspondences on both body shape and fused surface by setting $\tau_1(\mathbf{v}_c) \equiv 0$, the pose and fused surface get better but still contain artifacts. Only using all the energy terms can we get accurate motion and fusion reconstruction results as shown in Figure 10(d). We also evaluate the on-body correspondences separately in Figure 11, in which the performer starts with a rough A-pose and then turning around while waving his arms (please also refer to the supplementary video for more detailed evaluation). As shown in Figure 11, using only the fused surface for tracking will quickly lead to tracking failures when the left arm reappears with significant occluded motion (the 3rd row in the figure), and this is due to the lack of surface geometry for depth fitting. Using both the outer surface and inner body for tracking generates more plausible and complete results, as shown in the 2nd row. This evaluation also demonstrates that our method enables more robust human body reconstruction under unconstrained motion during self-turning-around.

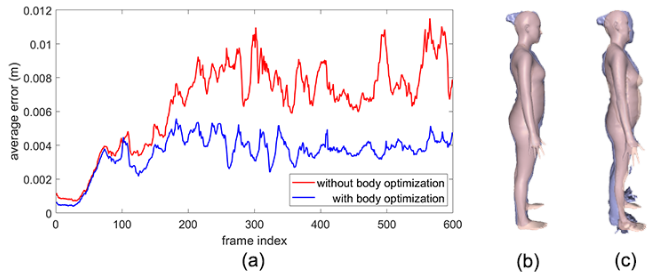


Fig. 12. Evaluation of inner body optimization according to non-rigid tracking accuracy. (a) Average tracking error per frame. (b) Reconstructed shape-mesh overlap with optimization. (c) Reconstructed shape-mesh overlap without optimization.

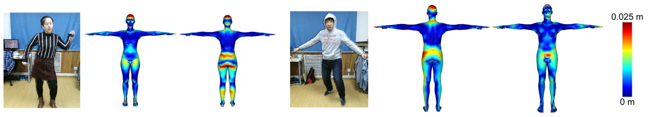


Fig. 13. Per-vertex error of the reconstructed body shapes.

Inner body optimization We evaluate inner body optimization both qualitatively and quantitatively. To evaluate the algorithm according to non-rigid tracking accuracy, in Figure 12, we use a public 4D sequence. We first render a single view depth sequence and then perform reconstruction using our system with/without optimization. The per-frame tracking error is calculated by averaging the point to plane error from the fused surface to the ground truth. We get better non-rigid tracking accuracy by using the body optimization as shown in Figure 12(a), and (b-c) demonstrates the reconstructed shape-mesh overlay with/without optimization. In Figure 13 and Figure 14, we evaluate the accuracy of the reconstructed body shape. We first obtain the ground truth undressed shape using laser scanner. Then we capture the same subject with clothing using our system. As shown in Figure 13, our reconstructed body shapes are plausible even though the subjects are dressed. Figure 14 shows the average error of shape reconstruction along the sequence with and without using the live-depth-based energy. Note that with the live-depth-based energy, we not only obtain a more accurate body shape at the

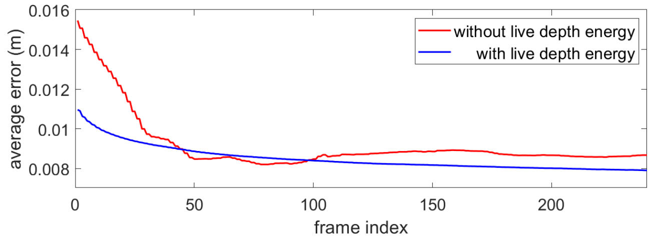


Fig. 14. Evaluation of the inner body optimization method and the live depth energy according to the shape reconstruction accuracy.

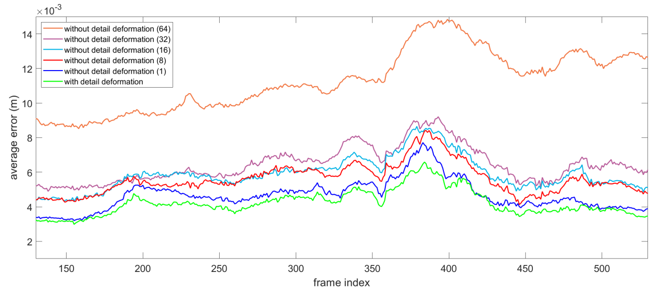


Fig. 15. Quantitative evaluation of dynamic detail deformation according to the surface reconstruction accuracy. The numbers-in-brackets in the legend represent current integration weights used in the TSDF fusion step. Note that all the results were generated by incorporating the live depth energy for inner body optimization.

TABLE 1
Quantitative comparison of the tracking accuracy with BodyFusion [18] and the preliminary version of DoubleFusion [22] using the vicon dataset. Note that we use the sequence ”szq” in the dataset because it starts with a rough A-pose.

Method	[18]	[22]	DoubleFusion(Jrnl)
Maximum (mm)	57.4	44.6	41.4
Average (mm)	23.7	22.5	20.8

beginning (since the initial A-pose depth silhouette also contains sufficient shape information), but also reconstruct a more accurate body shape in the end. We also qualitatively evaluate the proposed live-depth-based energy of inner body optimization in Figure 7. As shown in the figure, more accurate body shape (canonical body/skeleton embedding) (b), skeleton tracking (d) and large deformation results can be reconstructed by means of the live-depth-based energy. Please refer to the supplementary video for qualitative evaluations on more sequences.

Dynamic Detail Deformation The quantitative evaluation of the proposed dynamic detail deformation approach is shown in Figure 15. We first render the ground truth depth sequence (which contains sufficient dynamic geometric details) from the Buff Dataset [34], which comprises high quality 4D performances of 5 clothed people and was captured by a commertial multi-view 3D reconstruction system. The 4D sequences in the Buff Dataset captured sufficient dynamic geometric details on the cloth, which is ideal for the quantitative evaluation of the dynamic detail reconstruciton methods. Then we add synthetic noise on the ground truth depth sequence according to the Kinect noise model in [62]and use the noise-added depth sequence as the system input for the evaluation. As shown in Figure 15, the reconstruction accuracy of the dynamic details is improved by using the proposed dynamic detail deformation approach. Moreover, by simply increasing current integration weight in the TSDF fusion step, the reconstruction accuracy is even worse since the depth noise fused on the surface will deteriorate the tracking and fusion accuracy. Please refer to the supplementary video for the more detailed visualization.

6.3 Comparison

Quantitative Comparison We compare our surface tracking quantitatively with BodyFusion [18] and DoubleFusion(Pre) [22] (the preliminary version) using the public vicon dataset [18] in Table 1 and Figure 16. Note that compared with the preliminary version [22], the extended algorithm (benefiting from the more accurate body shape) obtains smaller maximum error and average error.

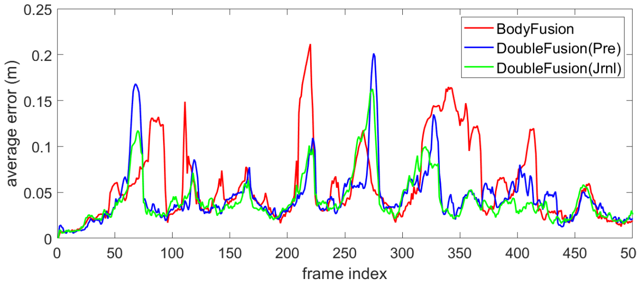


Fig. 16. Comparison of the tracking accuracy on the sequence ”szq” in BodyFusion [18] vicon dateset.

Qualitative Comparisons We qualitatively compare our method with 4 state-of-the-art single-view RGBD real-time non-rigid reconstruction methods [14], [18], [19], [20] in Figure 17. The approachs of [14] and [20] do not use any specific motion priors and are thus unable to handle very fast motion. Specifically, [14] tends to integrate erroneous surfaces when tracking fails while [20] loses the fused geometry around regions of tracking failure (e.g., hands and arms). [18] and [19] both incorporate articulated motion priors for non-rigid motion tracking. The difference between them is that [18] is focused on human reconstruction while [19] utilizes articulate motion priors for general objects. The generation of motion segmentation in [19] relies on reliable non-rigid motion tracking, so tracking fast motion is difficult, especially at the beginning. Moreover, the lack of body shape priors leads to deteriorated tracking and fusion performance in [18]. Note that our system still needs the subject to start capture with a rough A-pose, which is not a requisite of the other methods. Incorporating skeleton detection methods into the pipeline may overcome this limitation, as in [18], and we leave this step to future work. Figure 17 shows that benefiting from the proposed double-surface representation and joint motion tracking algorithm, our method achieves more robust tracking and loop closure performance than the other methods for human performance capture. Please see the supplementary video for more details.

Comparison against Learning-based Methods Most current methods for body shape and pose reconstruction, including [51], [52], [53], [54], [55] and [56], are based on machine learning techniques. These methods can infer plausible body shape and pose information from a single color image or silhouettes, which is convenient and promising. Among these methods, [51] is strong on pose reconstruction, while [53] is the state-of-the-art on shape reconstruction, benefiting from the proposed volumetric learning method. We qualitatively compare both of these techniques with our method for pose and shape reconstruction in Figure 18. Moreover, we quantitatively compare our method with [53] on shape reconstruction.

As shown in Figure 18, without incorporating temporal information into the inference network, both [51]

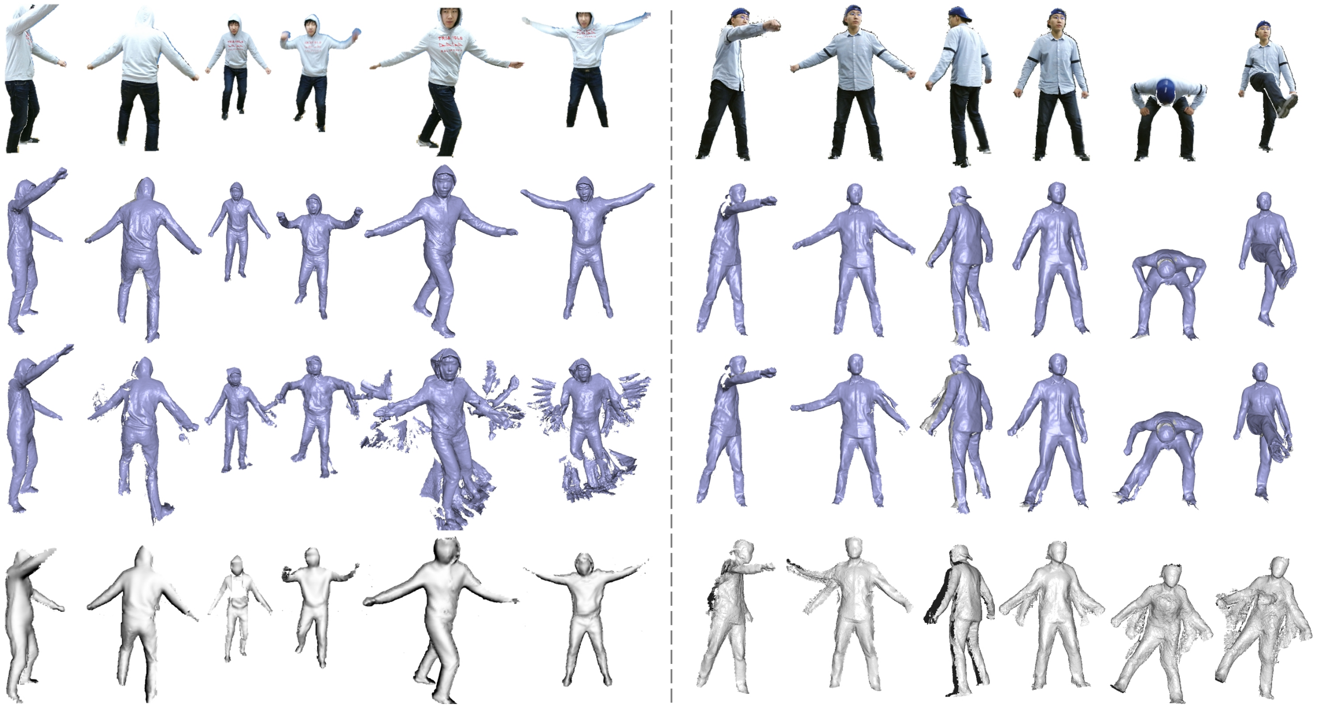


Fig. 17. Qualitative comparison with state-of-the-art real-time methods. The left column shows the comparison with [14] and [20] on the sequence ”Moonwalk”, in which the first row is the color reference and the second to fourth rows are the sequential reconstruction results of our method, the method in [14] and the method in [20], respectively. The right column shows the comparison with [18] and [19] on the sequence ”CVPR Live Demo 1”, in which the first row is the color reference and the second to fourth rows are the sequential reconstruction results of our method, the method in [18] and the method in [19], respectively.

and [53] produce temporally incoherent pose and shape reconstruction results. Moreover, since these methods do not consider depth information, they cannot reconstruct accurate 3D poses, which degrades their shape reconstruction performance. By contrast, our method generates temporally smooth and accurate shapes and poses even under challenging motion conditions.

For the quantitative comparison of shape reconstruction, we acquired the ground truth body shape by laser-scanning a naked body, as shown in Figure 13. Note that although [53] is designed for using only a single rgb image as input, we also test it on the image sequence for a more comprehensive comparison. Specifically, we first apply [53] to the sequence frame by frame and then average the inferred shape parameters from the whole sequence to obtain the final shape. To eliminate the impact of failure inferences for [53], we removed the frames of challenging poses when performing sequence-based shape reconstruction. The shape reconstruction accuracy is presented in Table 2. Our method achieves the highest accuracy because of the volumetric geometry fusion and accurate motion tracking performance. Table 2 also shows that our method can generate more accurate shapes and poses than [53] when only the initial A-pose depth frame is available. Moreover, when using the whole sequence for shape reconstruction, [53] achieves slightly worse accuracy than using only a single image because of the temporally incoherent shape reconstruction results and the inaccurate 3D poses. The results also demonstrate that shape inference under the initial front-facing A-pose is more accurate than that under other (especially side-facing) poses.

7 Applications

Our system can be used in many areas, including performance/motion capture, holoportation, 3D virtual try-on, VR/AR and gaming.

7.1 Body Measurement

One unique application of our system is convenient body measurement. Compared with traditional body measurement methods (e.g., measuring the body manually), available 3D scanner products (e.g., [63] and [64]) provide substantial information for accurate 3D body measurement. However, these products still require the subject to stand still on a turntable for approximately 15-30 seconds for data capture and several minutes for 3D body reconstruction and measurement. These requirements make these methods inconvenient and difficult for the consumer. Our system achieves convenient body measurement in real-time: the subject simply needs to turn around in front of a depth camera with a rough A-pose. Our method can perform body measurement either on the fused surface or on the optimized parametric (SMPL) model. Specifically, we predefine cross sections on different body parts of the SMPL model, where each contour produced by an intersection is shown in Figure 19. We measure circumferences of different body parts by calculating lengths of the contours. Additionally, we use the predefined cross sections on the SMPL model to cut the fused surface when performing body measurements, which is reasonable since the fitting between the SMPL and the fused surface has been optimized via the proposed

TABLE 2
Quantitative results of body shape reconstruction on the sequence ”Moonwalk” (frames 260 - 500) using our method and the approach of [53].

	Ours with single depth	*Ours with depth sequence*	*[53] with single rgb*	*[53] with rgb sequence*
Maximum Error (m)	0.0864	**0.0587**	0.1170	0.1433
Average Error (m)	0.0166	**0.0092**	0.0197	0.0210

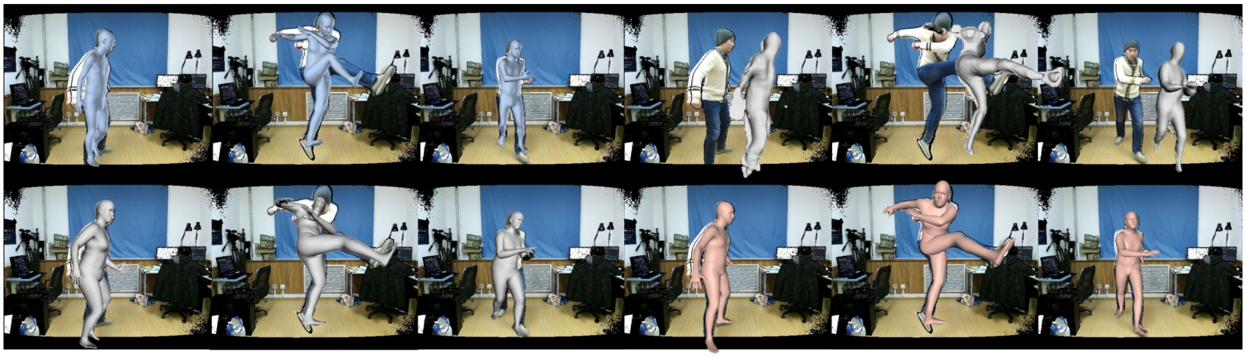


Fig. 18. Qualitative comparison with learning-based methods [51] and [53] on sequence ”Kicking” (the selected frames are 330, 1244 and 1339). The results of [51] are shown on the top left with blue color. The mesh inference results and corresponding SMPL models of [53] are shown with gray color on the top right and bottom left, respectively. Our SMPL model reconstruction results are shown on the bottom right with orange color.

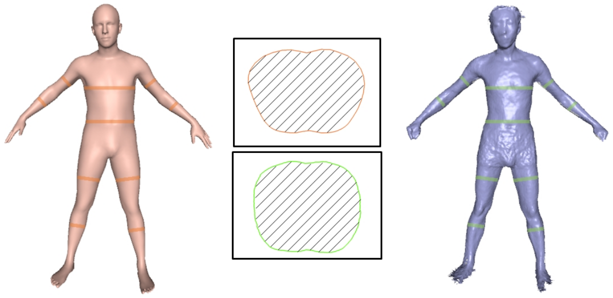


Fig. 19. Illustration of body measurement on the SMPL model and the fused surface. The red lines and green lines indicate the positions of different cross sections predefined on the SMPL model. The two cross sections shown in the middle are cross sections of the chest on the optimized SMPL model (top) and the fused surface (bottom).

”inner body optimization method” in Section 5.2. However, in this specific application, since the subjects all wear underwear, there is no reason to constrain the optimized SMPL model to be located inside the fused surface, so we set $\eta \equiv 1.0$ for the volume data term in Equation 13 for constraining the optimized SMPL model to perfectly align with the fused surface.

We captured 4 subjects wearing underwear for evaluation. Table 3 compares the body measurement results of our system with the hand measurement results from skilled people. The average and maximum error of the fused-surface measurement are lower than those of the parametric-body measurement, especially when measuring the chest and waist. Although the accuracy is not as high as that of professional measurement machines, this is the first method that achieves passive self-turning body measurement in real-time using only a single depth camera. The body measurement accuracy can be improved by combining bundle adjustment methods and color features into our pipeline to generate more accurate loop closed 3D body surfaces.

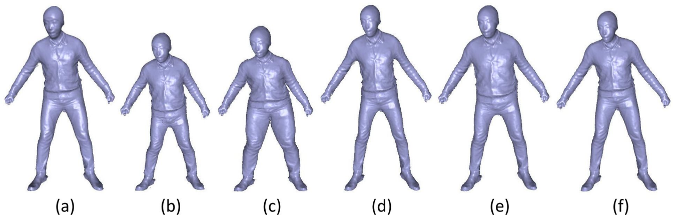


Fig. 20. Detailed outer surface shape retargeting results. (a) Reconstructed detailed geometry of a clothed human body. (b-f) Shape retargeting results for different body shapes.

7.2 Shape retargeting

Body shape retargeting for 4D sequences is especially useful when demonstrating subjects under new body shapes and synthesizing virtual datasets for human body analysis. In DoubleFusion, joint reconstruction of the outer surface and inner body shape enables parametric shape retargeting. Specifically, after obtaining the target shape parameters, we can generate the target body shape which has the same mesh topology as the source body shape. Then, we use all the vertices on the inner body as deformation nodes and estimate a smooth non-rigid deformation graph that can deform the source body shape to the target body shape. Finally, we apply the estimated deformation to the valid outer surface to generate the shape retargeting results of the outer surface geometry. Figure 20 demonstrates the shape retargeting application of our system.

8 Limitations

The reconstruction of very wide cloth remains challenging, as shown in Figure 21(a) (the reconstruction results of a long skirt) for two main reasons. First, free-form large non-rigid deformations of clothing does not strictly follow the human body motion prior. Moreover, it is too complicated for

TABLE 3
Body Measurement Results of 4 Subjects (YT, AL, ZMJ and WLZ). The measured circumferences and the measurement errors of different body parts are shown in the 1st, 2nd and 3rd subrow for each subject. Where the "Hand", "SMPL" and "Fused" means measurement results by hand, on the optimized SMPL model and on the fused surface, respectively. The last two rows show the maximum and average measurement error of the measurement results on the optimized SMPL model and on the fused surface. The unit is meter(m) for all the results.

		Upper arm	Lower arm	Upper leg	Lower leg	Chest	Waist
YT	Hand	0.275	0.235	0.505	0.396	0.970	0.803
	SMPL/error	0.304/0.029	**0.238/0.003**	0.547/0.042	0.382/-0.014	**0.973/0.003**	0.884/0.081
	Fused/error	**0.282/0.007**	0.256/0.021	**0.521/0.016**	**0.385/0.011**	0.960/0.010	**0.832/0.029**
AL	Hand	0.275	0.220	0.535	0.375	0.935	0.870
	SMPL/error	**0.296/0.021**	**0.227/0.007**	0.554/0.029	**0.378/0.003**	0.974/0.039	0.897/0.027
	Fused/error	0.299/0.024	0.251/0.031	**0.533/-0.002**	0.398/0.023	**0.954/0.019**	**0.896/0.026**
ZMJ	Hand	0.361	0.290	0.592	0.447	1.165	1.092
	SMPL/error	0.376/0.015	0.277/-0.012	**0.625/0.033**	0.427/-0.020	**1.178/0.013**	1.131/0.039
	Fused/error	**0.373/0.012**	**0.283/-0.007**	0.630/0.038	**0.455/0.009**	1.180/0.015	**1.104/0.012**
WLZ	Hand	0.291	0.241	0.570	0.395	0.890	0.860
	SMPL/error	0.314/0.023	0.244/0.003	0.564/-0.006	**0.389/-0.006**	0.998/0.108	0.922/0.062
	Fused/error	**0.303/0.012**	**0.241/0.000**	**0.567/-0.003**	0.411/0.016	**0.961/0.071**	**0.908/0.048**
SMPL	max/avg ($L1$)	0.029/0.022	**0.012/0.006**	0.042/0.028	**0.020/0.011**	0.108/0.041	0.081/0.052
Fused	max/avg ($L1$)	**0.024/0.014**	0.031/0.015	**0.038/0.015**	0.023/0.015	**0.071/0.029**	**0.048/0.029**

general non-rigid tracking methods to track such large free-form non-rigid deformation. Note that previous real-time fusion-based methods and even offline template-based non-rigid surface tracking methods are also unable to reconstruct satisfactory deformations under such conditions.

Furthermore, our system tends to overestimate body size when users wear thick clothes. Utilizing existing data-driven and learning-based techniques to handle cloth detection and reconstruction may produce better results. In addition, our system cannot handle geometric separations of the outer surface, which could be addressed by incorporating the key-volume update method from [8] or the variational-level-set-based method from [20]. Our system does not handle hair reconstruction very well because of the low depth quality of the hair regions of current depth sensors. Using color images as a reference to improve depth quality or using generative models for 3D hair generation may produce better results. Due to the lack of observations around the hands and face under the single-view full-body setup, we cannot capture detailed hand motion and facial expressions, which remains a challenging task for single-view full-body performance capture.

Moreover, our method cannot handle human-object interactions, as illustrated in Figure 21(b)(c). Future work will explore combining learning-based detection, classification and 3D reconstruction methods into our pipeline as a promising direction to overcome these limitations and obtain better results. For example, by utilizing [51], we can initialize our system from any pose and recover our system from severe tracking failures.

9 Conclusion

In this paper, we have demonstrated the first method for real-time reconstruction of both clothing and inner body shape from a single depth sensor. By incorporating the parametric human body model into the traditional single-view non-rigid RGBD reconstruction pipeline, we achieve accurate and robust single-view real-time human performance capture. More importantly, we have demonstrated that the free-form fusion-based methods and parametric-model-based methods can be used together and even benefit each other: With the assistance of parametric models, free-form fusion-based methods achieve more robust fast motion tracking and surface fusion performance. Furthermore, the fused detailed surface provides much more accurate observations for optimizing accurate parametric models. This idea can be used in other RGBD reconstruction areas, such as dynamic detailed face reconstruction, hand reconstruction, animal reconstruction and the reconstruction of other specific types of objects. On the basis of the proposed double-layer representation, our system achieves better non-rigid tracking and surface loop closure performance than that of state-of-the-art methods. Moreover, the real-time reconstructed inner body shapes are accurate and plausible. We believe the robustness and accuracy of our approach will enable many applications. In conclusion, with DoubleFusion users can easily digitise themselves.

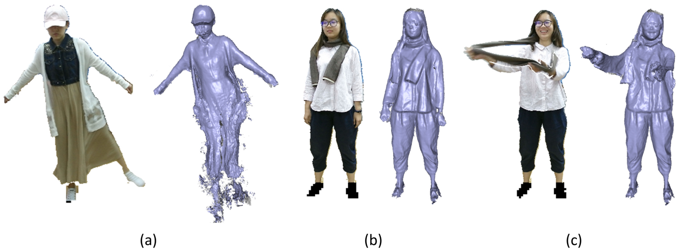


Fig. 21. Limitations: reconstruction results of loose cloth (long skirt)(a), and geometry separations (removing scarf from body)(b,c).

Acknowledgments

This work is supported by the National Natural Science Foundation of China NO.61827805, NO.61861166002, NO.61531014 and NO.51574012; Changjiang Scholars and Innovative Research Team in University of China, NO.IRT_16R02; Google Faculty Research Award; The Okawa Foundation Research Grant; The U.S. Army Research Laboratory under contract W911NF-14-D-0005. Gerard Pons-Moll is funded by the Deutsche Forschungsgemeinschaft(DFG. German Research Foundation)409792180.

References

[1] Y. Liu, Q. Dai, and W. Xu, "A point-cloud-based multiview stereo algorithm for free-viewpoint video," *IEEE Transactions on Visualization and Computer Graphics*, vol. 16, no. 3, pp. 407–418, May 2010. [Online]. Available: http://dx.doi.org/10.1109/TVCG.2009.88

[2] J. Gall, C. Stoll, E. De Aguiar, C. Theobalt, B. Rosenhahn, and H.-P. Seidel, "Motion capture using joint skeleton tracking and surface estimation," in *IEEE CVPR*, 2009.

[3] D. Bradley, T. Popa, A. Sheffer, W. Heidrich, and T. Boubekeur, "Markerless garment capture," in *ACM Transactions on Graphics*, vol. 27, no. 3. ACM, 2008, p. 99.

[4] T. Brox, B. Rosenhahn, J. Gall, and D. Cremers, "Combined region and motion-based 3d tracking of rigid and articulated objects," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 3, pp. 402–415, 2010.

[5] G. Ye, Y. Liu, N. Hasler, X. Ji, Q. Dai, and C. Theobalt, "Performance capture of interacting characters with handheld kinects," in *IEEE ECCV*, 2012.

[6] Y. Liu, J. Gall, C. Stoll, Q. Dai, H. Seidel, and C. Theobalt, "Motion capture of multiple characters using multiview image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 11, pp. 2720–2735, 2013.

[7] A. Mustafa, H. Kim, J. Guillemaut, and A. Hilton, "General dynamic scene reconstruction from multiple view video," in *IEEE ICCV*, 2015.

[8] M. Dou, S. Khamis, Y. Degtyarev, P. Davidson, S. R. Fanello, A. Kowdle, S. O. Escolano, C. Rhemann, D. Kim, J. Taylor *et al.*, "Fusion4d: real-time performance capture of challenging scenes," *ACM Transactions on Graphics*, vol. 35, no. 4, p. 114, 2016.

[9] V. Leroy, J.-S. Franco, and E. Boyer, "Multi-view dynamic shape refinement using local temporal integration," in *IEEE ICCV*, 2017.

[10] G. Pons-Moll, S. Pujades, S. Hu, and M. Black, "ClothCap: Seamless 4D clothing capture and retargeting," *ACM Transactions on Graphics, (Proc. SIGGRAPH)*, vol. 36, no. 4, 2017. [Online]. Available: http://dx.doi.org/10.1145/3072959.3073711

[11] M. Zollhöfer, M. Nießner, S. Izadi, C. Rehmann, C. Zach, M. Fisher, C. Wu, A. Fitzgibbon, C. Loop, C. Theobalt *et al.*, "Real-time non-rigid reconstruction using an RGB-D camera," *ACM Transactions on Graphics*, vol. 33, no. 4, p. 156, 2014.

[12] K. Guo, F. Xu, Y. Wang, Y. Liu, and Q. Dai, "Robust non-rigid motion tracking and surface reconstruction using l0 regularization," in *IEEE ICCV*, 2015.

[13] F. Bogo, M. J. Black, M. Loper, and J. Romero, "Detailed full-body reconstructions of moving people from monocular RGB-D sequences," in *IEEE ICCV*, 2015.

[14] R. A. Newcombe, D. Fox, and S. M. Seitz, "Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time," in *IEEE CVPR*, 2015.

[15] M. Innmann, M. Zollhöfer, M. Nießner, C. Theobalt, and M. Stamminger, "Volumedeform: Real-time volumetric non-rigid reconstruction," in *IEEE ECCV*, 2016.

[16] 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*, 2017.

[17] M. Slavcheva, M. Baust, D. Cremers, and S. Ilic, "KillingFusion: Non-rigid 3D Reconstruction without Correspondences," in *IEEE CVPR*, 2017.

[18] 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 ICCV*, 2017.

[19] C. Li, Z. Zhang, and X. Guo, "Articulatedfusion: Real-time reconstruction of motion, geometry and segmentation using a single depth camera," in *ECCV*, 2018.

[20] M. Slavcheva, M. Baust, and S. Ilic, "Sobolevfusion: 3d reconstruction of scenes undergoing free non-rigid motion," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

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

[22] 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 *The IEEE International Conference on Computer Vision and Pattern Recognition(CVPR)*. IEEE, June 2018.

[23] D. Vlasic, I. Baran, W. Matusik, and J. Popović, "Articulated mesh animation from multi-view silhouettes," in *ACM Transactions on Graphics*, vol. 27, no. 3. ACM, 2008, p. 97.

[24] Y. Liu, C. Stoll, J. Gall, H.-P. Seidel, and C. Theobalt, "Markerless motion capture of interacting characters using multi-view image segmentation," in *IEEE CVPR*, 2011.

[25] J. Taylor, J. Shotton, T. Sharp, and A. Fitzgibbon, "The vitruvian manifold: Inferring dense correspondences for one-shot human pose estimation," in *IEEE CVPR*, 2012.

[26] G. Pons-Moll, J. Taylor, J. Shotton, A. Hertzmann, and A. Fitzgibbon, "Metric regression forests for correspondence estimation," *International Journal of Computer Vision*, pp. 1–13, 2015.

[27] M. Ye and R. Yang, "Real-time simultaneous pose and shape estimation for articulated objects using a single depth camera," in *IEEE CVPR*, 2014.

[28] H. Li, B. Adams, L. J. Guibas, and M. Pauly, "Robust single-view geometry and motion reconstruction," in *ACM Transactions on Graphics*, vol. 28, no. 5. ACM, 2009, p. 175.

[29] R. W. Sumner, J. Schmid, and M. Pauly, "Embedded deformation for shape manipulation," *ACM Trans. Graph.*, vol. 26, no. 3, Jul. 2007. [Online]. Available: http://doi.acm.org/10.1145/1276377.1276478

[30] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis, "Scape: Shape completion and animation of people," *ACM Transactions on Graphics*, vol. 24, no. 3, pp. 408–416, Jul. 2005. [Online]. Available: http://doi.acm.org/10.1145/1073204.1073207

[31] G. Pons-Moll, J. Romero, N. Mahmood, and M. J. Black, "Dyna: A model of dynamic human shape in motion," *ACM Transactions on Graphics, (Proc. SIGGRAPH)*, vol. 34, no. 4, pp. 120:1–120:14, Aug. 2015.

[32] Y. Chen, Z.-Q. Cheng, C. Lai, R. R. Martin, and G. Dang, "Realtime reconstruction of an animating human body from a single depth camera," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 8, pp. 2000–2011, 2016.

[33] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black, "Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image," in *IEEE ECCV*, ser. Lecture Notes in Computer Science. Springer International Publishing, 2016.

[34] C. Zhang, S. Pujades, M. Black, and G. Pons-Moll, "Detailed, accurate, human shape estimation from clothed 3D scan sequences," in *IEEE CVPR*, 2017.

[35] T. Alldieck, M. Magnor, C. Theobalt, and G. Pons-Moll, "Video-based reconstruction of 3d people models," *IEEE CVPR*, 2018.

[36] M. Liao, Q. Zhang, H. Wang, R. Yang, and M. Gong, "Modeling deformable objects from a single depth camera," in *IEEE ICCV*, 2009.

[37] H. Li, R. W. Sumner, and M. Pauly, "Global correspondence optimization for non-rigid registration of depth scans," in *CGF*, vol. 27, no. 5. Wiley Online Library, 2008, pp. 1421–1430.

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

[39] W. Chang and M. Zwicker, "Range scan registration using reduced deformable models," in *CGF*, vol. 28, no. 2. Wiley Online Library, 2009, pp. 447–456.

[40] W. Chang and M. Zwicker, "Global registration of dynamic range scans for articulated model reconstruction," *ACM Transactions on Graphics*, vol. 30, no. 3, p. 26, 2011.

[41] Y. Pekelny and C. Gotsman, "Articulated object reconstruction and markerless motion capture from depth video," in *CGF*, vol. 27, no. 2. Wiley Online Library, 2008, pp. 399–408.

[42] N. J. Mitra, S. Flöry, M. Ovsjanikov, N. Gelfand, L. J. Guibas, and H. Pottmann, "Dynamic geometry registration," in *SGP*, 2007, pp. 173–182.

[43] J. Süßmuth, M. Winter, and G. Greiner, "Reconstructing animated meshes from time-varying point clouds," in *CGF*, vol. 27, no. 5. Blackwell Publishing Ltd, 2008, pp. 1469–1476.

[44] A. Sharf, D. A. Alcantara, T. Lewiner, C. Greif, A. Sheffer, N. Amenta, and D. Cohen-Or, "Space-time surface reconstruction using incompressible flow," *ACM Transactions on Graphics*, vol. 27, no. 5, p. 110, 2008.

[45] A. Tevs, A. Berner, M. Wand, I. Ihrke, M. Bokeloh, J. Kerber, and H.-P. Seidel, "Animation cartography-intrinsic reconstruction of

shape and motion," *ACM Transactions on Graphics*, vol. 31, no. 2, p. 12, 2012.

[46] H. Li, E. Vouga, A. Gudym, L. Luo, J. T. Barron, and G. Gusev, "3d self-portraits," *ACM Transactions on Graphics*, vol. 32, no. 6, p. 187, 2013.

[47] M. Dou, H. Fuchs, and J.-M. Frahm, "Scanning and tracking dynamic objects with commodity depth cameras," in *IEEE ISMAR*, 2013.

[48] 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

[49] 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)*, Sept 2018.

[50] T. Yu, Z. Zheng, Y. Zhong, J. Zhao, Q. Dai, G. Pons-Moll, and Y. Liu, "Simulcap : Single-view human performance capture with cloth simulation," in *The IEEE International Conference on Computer Vision and Pattern Recognition(CVPR)*. IEEE, June 2019.

[51] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik, "End-to-end recovery of human shape and pose," in *Computer Vision and Pattern Regognition (CVPR)*, 2018.

[52] G. Pavlakos, L. Zhu, X. Zhou, and K. Daniilidis, "Learning to estimate 3D human pose and shape from a single color image," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[53] G. Varol, D. Ceylan, B. Russell, J. Yang, E. Yumer, I. Laptev, and C. Schmid, "Bodynet: Volumetric inference of 3d human body shapes," in *The European Conference on Computer Vision (ECCV)*, September 2018.

[54] C. Lassner, J. Romero, M. Kiefel, F. Bogo, M. J. Black, and P. V. Gehler, "Unite the people: Closing the loop between 3d and 2d human representations," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017. [Online]. Available: http://up.is.tuebingen.mpg.de

[55] E. Dibra, H. Jain, A. C. Öztireli, R. Ziegler, and M. H. Gross, "Human shape from silhouettes using generative hks descriptors and cross-modal neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, July 21-26, 2017*, 2017.

[56] 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)*, sep 2018.

[57] Z. Zheng, T. Yu, Y. Wei, Q. Dai, and Y. Liu, "Deephuman: 3d human reconstruction from a single image," in *arXiv*, March 2019.

[58] Z. Lahner, D. Cremers, and T. Tung, "Deepwrinkles: Accurate and realistic clothing modeling," in *The European Conference on Computer Vision (ECCV)*, September 2018.

[59] R. Danerek, E. Dibra, A. C. ztireli, R. Ziegler, and M. Gross, "DeepGarment: 3D Garment Shape Estimation from a Single Image," *Computer Graphics Forum*, 2017.

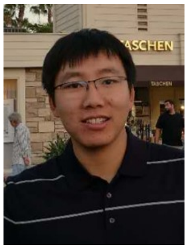
[60] W. E. Lorensen and H. E. Cline, "Marching cubes: A high resolution 3d surface construction algorithm," in *ACM SIGGRAPH*. New York, NY, USA: ACM, 1987, pp. 163–169. [Online]. Available: http://doi.acm.org/10.1145/37401.37422

[61] 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 *ISMAR*, ser. ISMAR '11. Washington, DC, USA: IEEE Computer Society, 2011, pp. 127–136. [Online]. Available: http://dx.doi.org/10.1109/ISMAR.2011.6092378

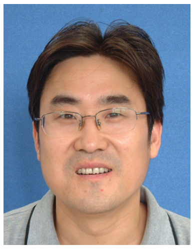
[62] C. V. Nguyen, S. Izadi, and D. Lovell, "Modeling kinect sensor noise for improved 3d reconstruction and tracking," in *2012 Second International Conference on 3D Imaging, Modeling, Processing, Visualization Transmission*, Oct 2012, pp. 524–530.

[63] "https://nakedlabs.com."

[64] "http://www.visbodyfit.com."

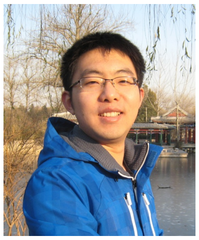


Tao Yu received the B.S. degree in Measurement and Control from Hefei University of Technology, China, in 2012. He is currently working toward Ph.D. degree in instrumental science in Beihang University, China. His current research interests include computer vision and computer graphics.



Jianhui Zhao received the BS degree in Electronic Engineering from Beihang University in 1992, and the MS and PhD degree in Precision Instrument and Machinery from Beihang University in 1995 and 2002, respectively. Since 1995, he has been with the faculty of Beihang University, Beijing, China, and is currently a professor and the director of the Department of Precision Instrument and Machinery. He has been working in Virginia Tech as a visiting scholar from 2009 to 2010. His research areas include spacecraft navigation and control, intelligent testing technologies and computer vision.

Zerong Zheng received the B.S. degree in Department of Automation, Tsinghua University in July 2018. He is currently pursuing Ph.D. degree in Department of Automation, Tsinghua University, advised by Prof. Yebin Liu. . His current research interests include computer vision and computer graphics.



Kaiwen Guo received the B.S. degree in control science and engineering from Northeastern University, China, in 2011, and the Ph.D. degree in control science and engineering from Tsinghua University in 2017. He then joined Google as a research scientist. His research interests include computer vision and graphics related to real-time and photorealistic capture of dynamic scene.

Qionghai Dai received the BS degree in mathematics from Shanxi Normal University, China, in 1987, and the ME and PhD degrees in computer science and automation from Northeastern University, China, in 1994 and 1996, respectively. Since 1997, he has been with the faculty of Tsinghua University, Beijing, China, and is currently a professor and the director of the Broadband Networks and Digital Media Laboratory. His research areas include video communication, computer vision, and graphics. He is a senior member of the IEEE. Prof. Dai is a member of the Chinese Academy of Engineering.

Hao Li is CEO/Co-Founder of Pinscreen, assistant professor of Computer Science at the University of Southern California, and the director of the Vision and Graphics Lab at the USC Institute for Creative Technologies. Hao's work in Computer Graphics and Computer Vision focuses on digitizing humans and capturing their performances for immersive communication and telepresence in virtual worlds. His research involves the development of novel geometry processing, data-driven, and deep learning algorithms. He is known for his seminal work in non-rigid shape alignment, real-time facial performance capture, hair digitization, and dynamic full body capture. He was previously a visiting professor at Weta Digital, a research lead at Industrial Light & Magic / Lucasfilm, and a postdoctoral fellow at Columbia and Princeton Universities. He was named top 35 innovator under 35 by MIT Technology Review in 2013 and was also awarded the Google Faculty Award, the Okawa Foundation Research Grant, as well as the Andrew and Erna Viterbi Early Career Chair. He won the Office of Naval Research (ONR) Young Investigator Award in 2018. Hao obtained his PhD at ETH Zurich and his MSc at the University of Karlsruhe (TH).

Gerard Pons-moll obtained his Degree in superior Telecommunications Engineering from the Technical University of Catalonia (UPC) in 2008. From 2007 to 2008 he was at Northeastern University in Boston USA working on medical image analysis. He received his Ph.D from the Leibniz University of Hannover in 2014. In 2012 he was a visiting researcher at University of Toronto and a research intern at the computer vision group at Microsoft Research Cambridge. From 11/2013-11/2017 he worked as a postdoc and Research Scientist at the Max Planck Institute for Intelligent Systems in Tuebingen, Germany. Since 11/2017 he is heading the research group ”Real Virtual Humans” at the Max Planck for Informatics (MPII) in Saarbrcken. His research lies at the intersection between computer vision, computer graphics and machine learning – with special focus on analyzing people in videos, and creating virtual human models by ”looking” at real ones.

Yebin Liu received the BE degree from Beijing University of Posts and Telecommunications, China, in 2002, and the PhD degree from the Department of Automation, Tsinghua University, Beijing, China, in 2009. He has been working as a research fellow at the computer graphics group of the Max Planck Institute for Informatik, Germany, in 2010. He has received the NSFC Excellent Young Scholar Award. He is currently an associate professor in the Department of Automation, Tsinghua University. His research areas include computer vision, computer graphics and computational photography. He is a member of IEEE (2012).