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

Tao Yu[1,2], Zerong Zheng[1], Kaiwen Guo[1,3], Jianhui Zhao[2], Qionghai Dai[1],
Hao Li[4], Gerard Pons-Moll[5], Yebin Liu[1,6]

[1]Tsinghua University, Beijing, China [2]Beihang University, Beijing, China [3]Google Inc
[4]University of Southern California / USC Institute for Creative Technologies
[5]Max-Planck-Institute for Informatics, Saarland Informatics Campus
[6]Beijing National Research Center for Information Science and Technology (BNRist)

Abstract

We propose DoubleFusion, a new real-time system that combines volumetric dynamic reconstruction with data-driven template fitting to simultaneously reconstruct detailed geometry, non-rigid motion and the inner 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 shape inside, and a gradually fused outer surface layer. A pre-defined node graph on the body surface 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 body shape is optimized online and forced to fit inside the outer surface layer. Overall, our method enables increasingly denoised, detailed and complete surface reconstructions, fast motion tracking performance and plausible inner body shape reconstruction in real-time. In particular, experiments show improved fast motion tracking and loop closure performance on more challenging scenarios.

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 [21, 12, 5, 6, 41, 22, 27, 11, 16, 30] 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 (iPhoneX, Google Tango, etc.), the recent trend focuses on using more and more practical setups like a single depth camera [45, 13, 3]. In particular, by combining non-rigid surface tracking and volumetric depth integration, DynamicFusion like approaches [28, 15, 14, 34] allow real-time dynamic scene reconstruction using a single depth camera without the requirement of pre-scanned model templates. Such systems are low cost, easy to set up and promising for popularization; however, they are still restricted to controlled slow motions. The challenges are occlusions (single view), computational resources (real-time), loop closure and no pre-scanned template model.

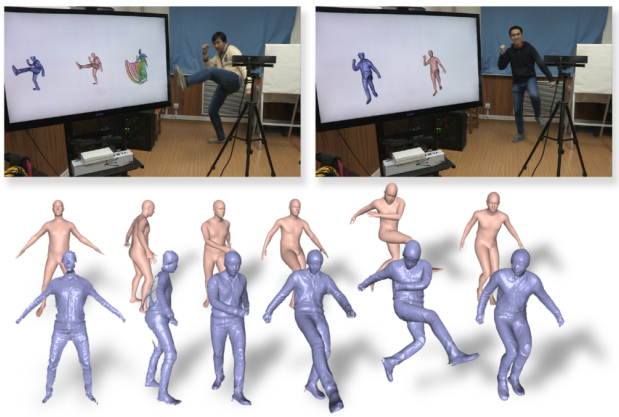

Figure 1: Our system and the real-time reconstructed results.

BodyFusion [43] is the most recent work in the direction of single-view real-time dynamic reconstruction; It shows that regularizing non-rigid deformations with a skeleton is beneficial to capture human performances. However, since the human joints are too sparse and it only uses the gradually fused surface for tracking, it fails during fast motions, especially when the surface is not yet complete. Moreover,

the skeleton embedding performance relies heavily on the initialization step and is fixed afterwards. Inaccurate skeleton embedding results in deteriorated tracking and deformation performance.

For human performance capture, besides the skeleton, body shape is also a very strong prior since it is loop closed 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. In addition, we make each layer benefit from each other. Based on the recent state-of-the-art body model SMPL [24], we propose a double-layer surface representation consisting of an outer surface layer, and an inner body layer for reconstruction and depth registration. 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. 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. 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, and without pre-scanning efforts. Compared to systems that only reconstruct the outer surface like BodyFusion [43], we demonstrate substantially improved performance in handling fast motions. In contrast to systems specialized to capture the inner body [3], our approach can handle people wearing casual clothing, and it works in real-time. To enable the above advantages, we make the following technical contributions in this paper.

- We propose the double-layer representation (Section 3.1) for high quality and realtime 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 like multiview systems.

- 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 reconstruction.

- Volumetric shape-pose optimization of the inner layer (Section 5). We fit the SMPL model parameters with the canonical model directly in the TSDF volume defined by the outer surface without searching correspondences. The optimized body shape and pose (skeleton embedding) in the canonical frame is beneficial for outer surface tracking.

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 and free-form 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.* [39] and Gall *et al.* [12] adopted a template with embedded skeleton driven by multi-view silhouettes and temporal feature constraints. Liu *et al.* [23] extended the method to handle multiple interacting performers. Some approaches [37, 32] use a random forest to predict correspondences to a template, and use them to fit the template to the depth data. Ye *et al.* [41] considered the case of multiple Kinects input. Ye *et al.* [42] adopted a similar skinned model to estimate shape and pose parameters using a single-view depth camera in real-time. For this kind of template, in order to achieve accurate tracking, skeleton embedding is usually done manually.

Besides templates with an embedded skeleton, some works adopted template based non-rigid surface deformation. Li *et al.* [17] utilized embedded deformation graph in Sumner *et al.* [35] to parameterize the pre-scanned template to produce locally as-rigid-as-possible deformation. Guo *et al.* [13] adopted an ℓ_0 norm constraint to generate articulate motions without explicitly embedded skeleton. Zollhöfer *et al.* [45] took advantage of massive parallelism of GPU to enable real-time performance of general non-rigid tracking.

For the aforementioned 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 [2] is one of the widely used model, it factorizes deformations into pose and shape components. SMPL [24] 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.* [9] adopted SCAPE to capture body motion using a single depth camera. Bogo *et al.* [3] extended SCAPE to capture detailed body shape with appearance. Bogo *et al.* [4] 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.* [44] they addressed this problem by estimating the inner shape and recovering surface details. Pons-Moll *et al.* [30] introduce ClothCap, which jointly estimates clothing geometry and body shape using separate meshes. In both [44] and [30], results are only shown for complete 4D scan sequences. Alldieck *et al.* [1] 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 [20], deformation graph [18], subspace deformation [40], articulate deformation [7, 8] and [29], 4D spatio-temporal surface [26] and [36], incompressible flows [33], animation cartography [38], quasi-rigid motions [19] and directional field [10].

Only in recent years, free-form capture methods with real-time performance have been proposed. DynamicFusion [28] proposed a hierarchical node graph structure and an approximate direct GPU solver to enable capturing non-rigid scenes in real-time. Guo *et al.* [14] 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.* [34] proposed a killing constraint for regularization. However, neither of methods demonstrated full body performance capture with natural motions. Fusion4D [11] setup a rig with 8 depth camera to capture dynamic scenes with challenging motions in real-time. BodyFusion [43], utilizes skeleton priors for human body reconstruction, but cannot handle challenging fast motions and cannot infer inner body shape.

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 surface of the performer. The outer layer are observable surface regions, such as clothing, visible body parts (e.g. face, hair), while the inner layer is a parametric human shape and skeleton model based on the skinned multi-person linear model (SMPL) [24]. Similar to previous work [28], 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 [28] 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 motions. 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 Fig. 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 labled 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 Fig. 2(d)(bottom).

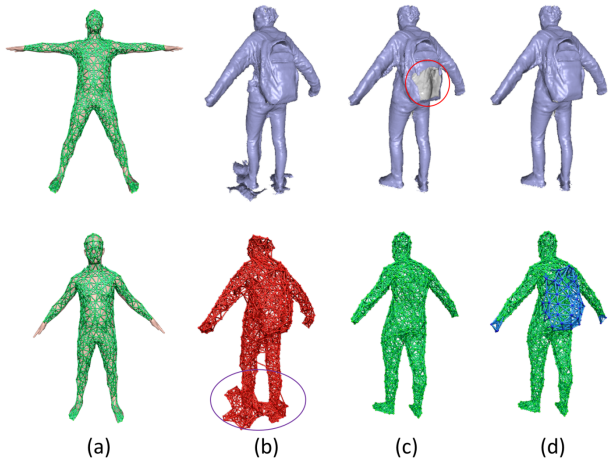

Figure 2: (a) Initialization of the on-body node graph. (b)(c)(d) Evaluation of the double node graph. The figure shows the geometry results and live node graph of (b) traditional free-form sampled node graph (red), (c) on-body node graph (green) only and (d) double node graph (with far-body nodes in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

3.2. Inner Body Model: SMPL

SMPL [24] 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 [24] for more details.

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 TSDF volume by projecting depth map into the volume. Then we use volumetric shape-pose optimization (see Sec. 5.2) to estimate initial shape parameters $\boldsymbol{\beta}_0$ and skeletal pose $\boldsymbol{\theta}_0$. After that, we initialize the double node graph using the on-body node graph and initial pose and shape as shown in Fig. 2(a)(bottom). We extract a triangle mesh from the volume using Marching Cube algorithm [25] 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*, *geometric fusion* and *volumetric shape-pose optimization* sequentially (Fig. 3). We briefly introduce the main components of the pipeline below:

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

Geometric fusion Similar to previous work [28], we non-rigidly integrate depth observation of multiple frames in a reference volume (Sec. 5.1). We also explicitly detect collided voxels to avoid erroneously fused geometry [14].

Volumetric shape-pose optimization After geometric fusion, the surface in the canonical frame gets more complete. We directly optimize the body shape and pose by using the fused signed distance field (Sec. 5.2) This step is very efficient because it does not require finding correspondences.

4. Joint Motion Tracking

There are two parameterizations in our motion tracking component, skeletal motions and non-rigid node deformations. Similar to the previous work [43], we adopt a binding term that constrains both motions to be consistent. Different from [43], we only enforce the binding term on on-body nodes to penalize non-articulated motions 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 [4] to use a statistic pose prior to prevent unnatural poses. The energy of joint optimization is then

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

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 fitting between the reconstructed double layer surface and depth map:

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

where $\mathcal{P}$ is the correspondence set; $\psi(\cdot)$ is the robust Geman-McClure penalty function; $(\mathbf{v}_c, \mathbf{u})$ is a correspondence pair; $\mathbf{u}$ is a sampled point on the depth map and its closest point $\mathbf{v}_c$ can be on either the body shape or fused surface. Correspondences on the body shape enable fast and robust tracking performance. $\tau_1(\mathbf{v}_c)$, $\tau_2(\mathbf{v}_c)$ and $\tau_3(\mathbf{v}_c)$ are correspondence indicator functions: $\tau_1(\mathbf{v}_c)$ equals to 1 only if $\mathbf{v}_c$ is on the fused surface; $\tau_2(\mathbf{v}_c)$ equals to 1 when $\mathbf{v}_c$ is on the body shape; $\tau_3(\mathbf{v}_c)$ equals to 1 when $\mathbf{v}_c$ is on the fused surface and its 4 nearest nodes (knn-nodes) of $\mathbf{v}_c$ are all on-body nodes. $\tilde{\mathbf{v}}_c$ 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

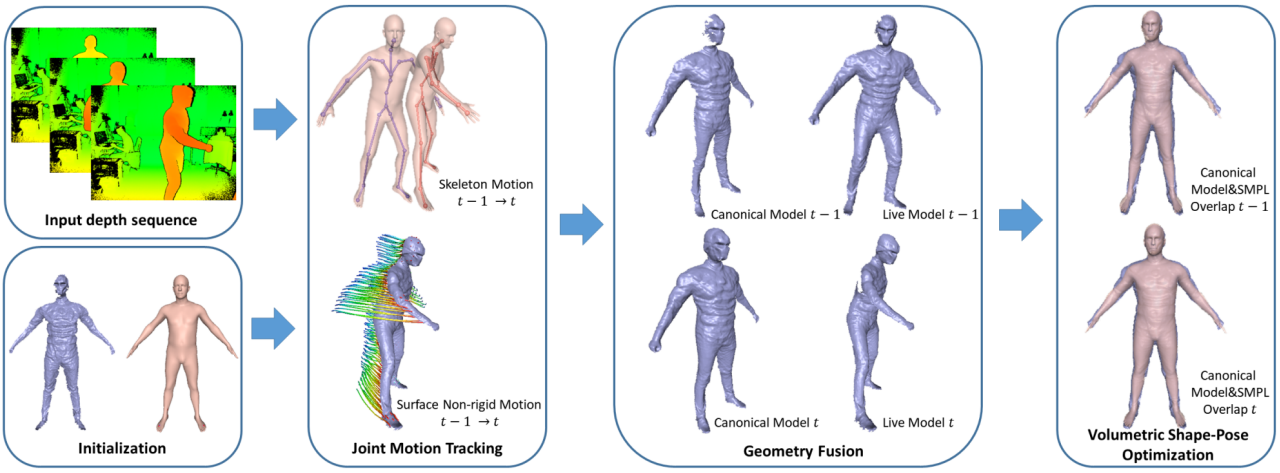

Figure 3: Our system pipeline. We first initialize our system using the first depth frame (Sec. 3.3). Then for each frame, we sequentially perform the next 3 steps: joint motion tracking (Sec. 4), geometric fusion (Sec. 5.1) and volumetric shape-pose optimization (Sec. 5.2).

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 motions 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

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

where we choose $\mu = 0.2$; we set $\delta_{\max} = 0.1$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. The binding term attaches on-body nodes to their nearest bones and helps to produce articulated deformations on the body. It is defined as

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

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 Eqn. 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 [43]. This term is then defined as

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

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 weight function in [43]. 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 [28].

Pose Prior Term Similar to [4], we include a pose prior penalizing the unnatural poses. It is defined as

$$E_{\text{prior}} = -\log\left(\sum_j \omega_j N(\boldsymbol{\theta}; \mu_j, \delta_j) \right). \tag{10}$$

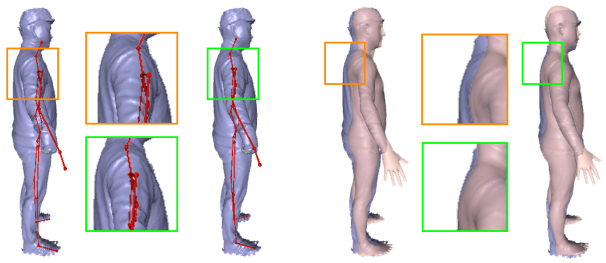

Figure 4: Illustration of volumetric shape-pose optimization. (a) skeleton embedding results before and after optimization. (b) shape-mesh overlap before and after optimization.

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 the optimization problem (Eqn. 3) using Iterative Closest Point (ICP) method. First we build a correspondence set $\mathcal{P}$ using the latest motion parameters; then we solve the non-linear least squares using Gauss-Newton method. We use a twist representation for both the bone and node transformations. Within each iteration of Gauss-Newton procedure, the transformations are approximated using one-order Taylor expansion around the latest values. Then we solve the resulting linear system using a custom designed highly efficient preconditioned conjugate gradient (PCG) solver on GPU [14, 11].

5. Volumetric Fusion & Optimization

5.1. Geometric Fusion

Similar to the previous non-rigid fusion works [28, 15, 14], we integrate the depth information into a reference volume. First, the voxels in the reference volume are warped to live frame according to current non-rigid warp field. Then, we calculate the PSDF value of each valid voxel and use it to update their TSDF values. We follow the work [14] to cope with collided voxels in live frame to prevent erroneous fusion results caused by collisions.

5.2. Volumetric Shape-Pose Optimization

After the non-rigid fusion, we have an updated surface in the canonical volume with more complete geometry. Since the initial shape and pose parameters $(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0)$ may not fit well with the new observation in the volume, as shown in Fig.4(a), we propose a novel algorithm that can efficiently optimize both of the shape parameters and initial embedding pose jointly in the canonical volume. The formulation of the energy is then

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

where E_{sdata} measures misalignment error in the reference volume; E_{sreg} is a temporal constraint that makes the new shape and poses parameters consistent with the previous ones. E_{pri} is the same as in Eqn. 3 to prevent unnatural poses. The novel volumetric data term is defined as

$$E_{\text{sdata}}(\boldsymbol{\beta}, \boldsymbol{\theta}) = \sum_{\bar{\mathbf{v}} \in \bar{\mathbf{T}}} \psi(\mathbf{D}(W(T(\bar{\mathbf{v}}; \boldsymbol{\beta}, \boldsymbol{\theta}); J(\boldsymbol{\beta}), \boldsymbol{\theta}))), \tag{12}$$

where $\mathbf{D}(\cdot)$ is a bilinear sampling function that takes a point in the canonical volume and returns interpolated TSDF. 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. $\mathbf{v} = T(\bar{\mathbf{v}}; \boldsymbol{\beta}, \boldsymbol{\theta})$ modifies $\bar{\mathbf{v}}$ by shape blend shape and pose blend shape; $W(\mathbf{v}; J(\boldsymbol{\beta}, \boldsymbol{\theta}), \boldsymbol{\theta})$ deforms $\mathbf{v}$ using linear blend skinning. The temporal regularization is defined as

$$E_{\text{sreg}}(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\beta}', \boldsymbol{\theta}') = \gamma_1 \|\boldsymbol{\beta} - \boldsymbol{\beta}'\|_2^2 + \gamma_2 \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2^2. \tag{13}$$

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 (Eqn. 12) and make the convergence faster. Then we solve the resulting energy using the same GPU-based Gauss-Newton solver as in Sec. 4. At last, we update the body shape and pose that embedded into the canonical frame and recalculate the motion field and the skeleton motions. After more surface observation is fused into the TSDF volume, the body shape and canonical body pose get more accurate. (Fig. 4(b)).

6. Results

In this section, we first report the performance and the main parameters of the system. Then we compare with previous state-of-the-art methods qualitatively and quantitatively. We also evaluate each of our main contributions. In Fig. 5, we demonstrate the results of our system. Note the various shapes, challenging motions and different types of cloth of the loop closed model that we can reconstructed.

6.1. Performance

DoubleFusion runs in real-time (running at 32ms per frame). The entire pipeline is implemented on one NVIDIA TITAN X GPU. Executing 6 ICP iterations, the joint motion tracking takes 21 ms. The geometric fusion takes 6 ms

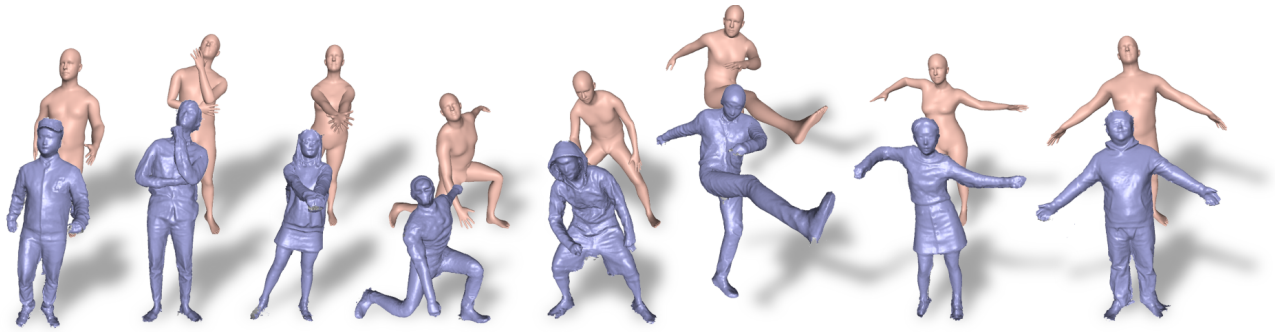

Figure 5: Example results reconstructed by our system.

and volumetric shape-pose optimization takes 3 ms. Prior to the joint motion tracking, we process the input depth frame using bilateral filtering, boundary outlier and floor plane removal. After volumetric shape-pose optimization, a triangulated mesh is extracted, non-rigidly transformed into camera coordinates and rendered on the frame. These two parts run asynchronously with the main pipeline, and runtime overhead is negligible with less than 1 ms. For all of our experiments, we choose $\lambda_{\text{data}} = 1.0$, $\lambda_{\text{bind}} = 1.0$, $\lambda_{\text{reg}} = 5.0$ and $\lambda_{\text{pri}} = 0.01$. 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 Fig. 2. The standard node graph construction scheme [35] uniformly samples all the nodes on the fused outer surface. The lack of semantic information results in wrong connections (connection between two legs) and erroneous fusion results as shown in Fig. 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 Fig. 2(c)) since it is out of the control area of on-body node graph. By using the proposed double node graph (Fig. 2(d)), we can get clean and complete results.

Joint motion tracking. In Fig. 6, we evaluate different components of the joint motion tracking step qualitatively. We eliminate non-rigid registration in Fig. 6(b) and (c). In Fig. 6(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 Eqn. 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 Fig. 6(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 we can get accurate pose and fusion results as shown in Fig. 6(d). We also evaluate the on-body correspondences separately in Fig. 7. Only using fused surface for tracking will easily get failed: will quickly fail when the left arm reappears with large motion

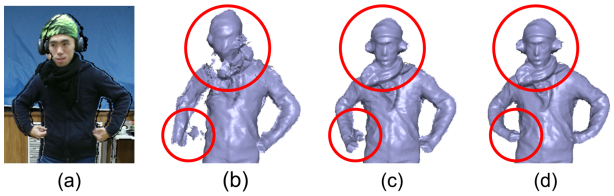

Figure 6: Evaluation of joint motion tracking. (a) reference color image. (b) results only using correspondences on body for skeleton tracking, without non-rigid registration; (c) searching correspondences on both body and fused surface for skeleton tracking, without non-rigid registration; (d) using full energy terms.

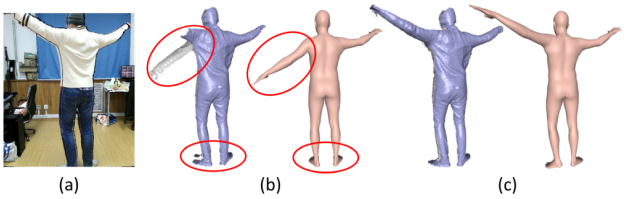

Figure 7: Evaluation of on-body correspondences. (a) reference color image (b) results only using fused surface for tracking. (c) results using both body and fused surface for tracking.

in the scene due to the lack of surface geometry as shown in Fig. 7(b). Using both surface and body shape for tracking will generate more plausible results as shown in Fig. 7(c).

Volumetric shape-pose optimization. We evaluate volumetric shape-pose optimization both qualitatively and quantitatively. To evaluate non-rigid tracking accuracy, in Fig. 8, 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 optimization as shown in Fig. 8(a), and (b-c) demonstrates the reconstructed shape-mesh overlap with and without optimization. In Fig. 9 and Fig. 10, we evaluate the accuracy of our reconstructed shape. We obtain ground truth undressed shape using laser scanner. Then we capture the same subject with clothing using DoubleFusion. As shown in Fig. 9, our reconstructed body shapes are plausible even though the subjects are dressed. Fig. 10 shows the average shape reconstruction error along the sequence.

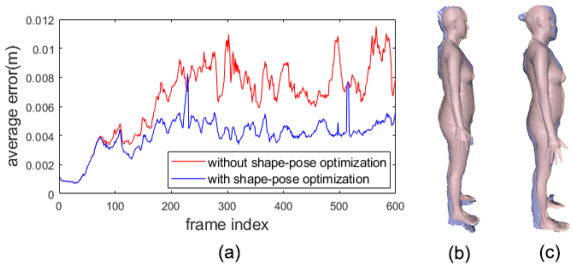

Figure 8: Evaluation of volumetric shape-pose optimization using non-rigid tracking accuracy. (a) average tracking error per frame, (b) reconstructed shape-mesh overlap with optimization, (c) reconstructed shape-mesh overlap without optimization.

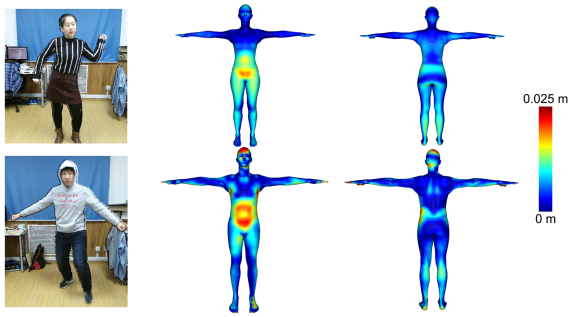

Figure 9: Per-vertex error of the reconstructed body shapes.

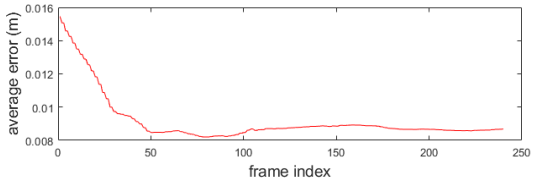

Figure 10: Evaluation of the body shape estimation accuracy of our online shape-pose optimization method.

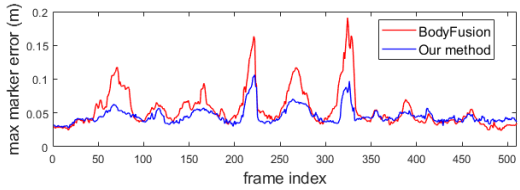

Figure 11: Comparison of tracking accuracy on sequence "szq".

Method	BodyFusion[43]	Ours
Maximum Error (m)	0.0554	0.0458
Average Error (m)	0.0277	0.0221

Table 1: Average numerical errors on the entire sequence.

6.3. Comparison

We compare our tracking accuracy with BodyFusion [43] using their public vicon dataset. DoubleFusion obtains smaller per-frame max error (Fig. 11), and smaller average error (Tab. 1), especially during fast motions.

We qualitatively compare our method with two real-time state-of-the-art methods [28, 43]. [28] uses general non-rigid registration method without any prior, while [43] takes advantage of a human skeletal constraint for better tracking ability. Fig. 12 shows that our method achieves improved tracking and loop closure performance than other methods. Please see the supplementary video for more details.

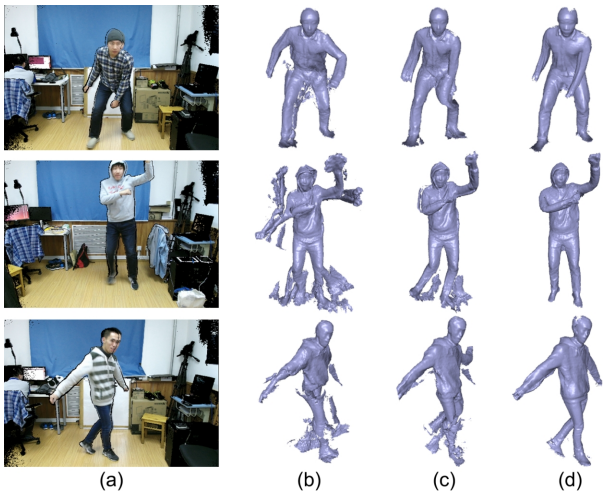

Figure 12: Comparison. (a) reference color image. (b)(c)(d), results of DynamicFusion[28], BodyFusion[43] and our method.

7. Discussion

Limitations Our system tends to over-estimate body size when users wear thick clothing, and reconstruction of very wide cloth remains challenging. We cannot handle geometry separations of the outer surface, this could be addressed incorporating the key-volume update method in [11]. Our current system can not handle human-object interactions, which we plan to address in future work.

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. Based on the proposed double surface representation, our system achieved better non-rigid tracking and surface loop closure performance than state-of-the-art methods. Moreover, the real-time reconstructed inner body shapes are visually plausible. We believe the robustness and accuracy of our approach will enable many applications, especially in AR/VR, gaming, entertainment and even virtual try-on as we also reconstruct the underlying body shape. For the first time, with DoubleFusion, users can easily digitize themselves.

Acknowledgements This work is supported by the National key foundation for exploring scientific instrument of China No.2013YQ140517; NKBRP of China No.2014CB744201; the National NSF of China grant No.61522111, No.61531014, No.61233005; Shenzhen Peacock Plan KQTD20140630115140843; Changjiang Scholars and Innovative Research Team in University, No.IRT_16R02; Google Faculty Research Award; the Okawa Foundation Research Grant; the U.S. Army Research Laboratory under contract W911NF-14- D-0005.

References

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

[2] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis. Scape: Shape completion and animation of people. *ACM Transactions on Graphics*, 24(3):408–416, July 2005. 2

[3] 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. 1, 2, 3

[4] 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*, Lecture Notes in Computer Science. Springer International Publishing, 2016. 3, 4, 5

[5] D. Bradley, T. Popa, A. Sheffer, W. Heidrich, and T. Boubekeur. Markerless garment capture. In *ACM Transactions on Graphics*, volume 27, page 99. ACM, 2008. 1

[6] 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.*, 32(3):402–415, 2010. 1

[7] W. Chang and M. Zwicker. Range scan registration using reduced deformable models. In *CGF*, volume 28, pages 447–456. Wiley Online Library, 2009. 3

[8] W. Chang and M. Zwicker. Global registration of dynamic range scans for articulated model reconstruction. *ACM Transactions on Graphics*, 30(3):26, 2011. 3

[9] 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*, 22(8):2000–2011, 2016. 3

[10] M. Dou, H. Fuchs, and J.-M. Frahm. Scanning and tracking dynamic objects with commodity depth cameras. In *IEEE ISMAR*, 2013. 3

[11] 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*, 35(4):114, 2016. 1, 3, 6, 8

[12] 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. 1, 2

[13] 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. 1, 2

[14] 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. 1, 3, 4, 6

[15] M. Innmann, M. Zollhöfer, M. Nießner, C. Theobalt, and M. Stamminger. Volumedeform: Real-time volumetric non-rigid reconstruction. In *IEEE ECCV*, 2016. 1, 3, 6

[16] V. Leroy, J.-S. Franco, and E. Boyer. Multi-view dynamic shape refinement using local temporal integration. In *IEEE ICCV*, 2017. 1

[17] H. Li, B. Adams, L. J. Guibas, and M. Pauly. Robust single-view geometry and motion reconstruction. In *ACM Transactions on Graphics*, volume 28, page 175. ACM, 2009. 2

[18] H. Li, R. W. Sumner, and M. Pauly. Global correspondence optimization for non-rigid registration of depth scans. In *CGF*, volume 27, pages 1421–1430. Wiley Online Library, 2008. 3

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

[20] M. Liao, Q. Zhang, H. Wang, R. Yang, and M. Gong. Modeling deformable objects from a single depth camera. In *IEEE ICCV*, 2009. 3

[21] 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*, 16(3):407–418, May 2010. 1

[22] 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.*, 35(11):2720–2735, 2013. 1

[23] 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. 2

[24] 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)*, 34(6):248:1–248:16, Oct. 2015. 2, 3, 4

[25] W. E. Lorensen and H. E. Cline. Marching cubes: A high resolution 3d surface construction algorithm. In *ACM SIGGRAPH*, pages 163–169, New York, NY, USA, 1987. ACM. 4

[26] N. J. Mitra, S. Flöry, M. Ovsjanikov, N. Gelfand, L. J. Guibas, and H. Pottmann. Dynamic geometry registration. In *SGP*, pages 173–182, 2007. 3

[27] A. Mustafa, H. Kim, J. Guillemaut, and A. Hilton. General dynamic scene reconstruction from multiple view video. In *IEEE ICCV*, 2015. 1

[28] R. A. Newcombe, D. Fox, and S. M. Seitz. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *IEEE CVPR*, 2015. 1, 3, 4, 5, 6, 8

[29] Y. Pekelny and C. Gotsman. Articulated object reconstruction and markerless motion capture from depth video. In *CGF*, volume 27, pages 399–408. Wiley Online Library, 2008. 3

[30] G. Pons-Moll, S. Pujades, S. Hu, and M. Black. ClothCap: Seamless 4D clothing capture and retargeting. *ACM Transactions on Graphics, (Proc. SIGGRAPH)*, 36(4), 2017. 1, 3

[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)*, 34(4):120:1–120:14, Aug. 2015. 2

[32] G. Pons-Moll, J. Taylor, J. Shotton, A. Hertzmann, and A. Fitzgibbon. Metric regression forests for correspondence estimation. *International Journal of Computer Vision*, pages 1–13, 2015. 2

[33] 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*, 27(5):110, 2008. 3

[34] M. Slavcheva, M. Baust, D. Cremers, and S. Ilic. KillingFusion: Non-rigid 3D Reconstruction without Correspondences. In *IEEE CVPR*, 2017. 1, 3

[35] R. W. Sumner, J. Schmid, and M. Pauly. Embedded deformation for shape manipulation. SIGGRAPH '07, New York, NY, USA, 2007. ACM. 2, 7

[36] J. Süßmuth, M. Winter, and G. Greiner. Reconstructing animated meshes from time-varying point clouds. In *CGF*, volume 27, pages 1469–1476. Blackwell Publishing Ltd, 2008. 3

[37] 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. 2

[38] 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*, 31(2):12, 2012. 3

[39] D. Vlasic, I. Baran, W. Matusik, and J. Popović. Articulated mesh animation from multi-view silhouettes. In *ACM Transactions on Graphics*, volume 27, page 97. ACM, 2008. 2

[40] 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*, 28(2):15, 2009. 3

[41] 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. 1, 2

[42] M. Ye and R. Yang. Real-time simultaneous pose and shape estimation for articulated objects using a single depth camera. In *IEEE CVPR*, 2014. 2

[43] 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. 1, 2, 3, 4, 5, 8

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

[45] 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*, 33(4):156, 2014. 1, 2