Video Based Reconstruction of 3D People Models

Thiemo Alldieck[1] Marcus Magnor[1] Weipeng Xu[2] Christian Theobalt[2] Gerard Pons-Moll[2]

[1]Computer Graphics Lab, TU Braunschweig, Germany
[2]Max Planck Institute for Informatics, Saarland Informatics Campus, Germany
{alldieck,magnor}@cg.cs.tu-bs.de {wxu,theobalt,gpons}@mpi-inf.mpg.de

Figure 1: Our technique allows to extract for the first time accurate 3D human body models, including hair and clothing, from a single video sequence of the person moving in front of the camera such that the person is seen from all sides.

Abstract

This paper describes a method to obtain accurate 3D body models and texture of arbitrary people from a single, monocular video in which a person is moving. Based on a parametric body model, we present a robust processing pipeline to infer 3D model shapes including clothed people with 4.5mm reconstruction accuracy. At the core of our approach is the transformation of dynamic body pose into a canonical frame of reference. Our main contribution is a method to transform the silhouette cones corresponding to dynamic human silhouettes to obtain a visual hull in a common reference frame. This enables efficient estimation of a consensus 3D shape, texture and implanted animation skeleton based on a large number of frames. Results on 4 different datasets demonstrate the effectiveness of our approach to produce accurate 3D models. Requiring only an RGB camera, our method enables everyone to create their own fully animatable digital double, e.g., for social VR applications or virtual try-on for online fashion shopping.

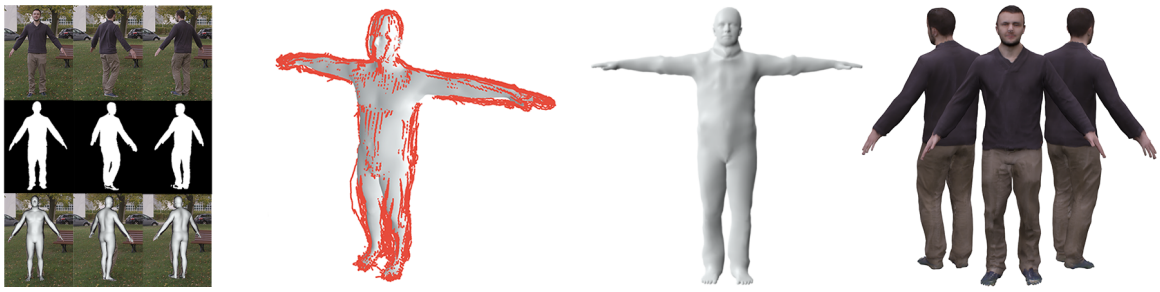
1. Introduction

A personalized realistic and animatable 3D model of a human is required for many applications, including virtual and augmented reality, human tracking for surveillance, gaming, or biometrics. This model should comprise the person-specific static geometry of the body, hair and clothing, alongside a coherent surface texture.

One way to capture such models is to use expensive active scanners. But size and cost of such scanners prevent their use in consumer applications. Alternatively, multi-view passive reconstruction from a dense set of static body pose images can be used [22, 46]. However, it is hard for people to stand still for a long time, and so this process is time-consuming and error-prone. Also, consumer RGB-D cameras can be used to scan 3D body models [39], but these specialized sensors are not as widely available as video. Further, all these methods merely reconstruct surface shape and texture, but no rigged animation skeleton inside. All aforementioned applications would benefit from the ability to automatically reconstruct a personalized movable avatar from monocular RGB video.

Despite remarkable progress in reconstructing 3D body models [6, 71, 81] or free-form surface [86, 44, 47, 21] from depth data, 3D reconstruction of humans in clothing from monocular video (without a pre-recorded scan of the person) has not been addressed before. In this work, we estimate the shape of people in clothing from a single video in which the person moves. Some methods infer shape parameters of a parametric body model from a single image [7, 20, 5, 27, 83, 34], but the reconstruction is limited to the parametric space and can not capture personalized shape detail and clothing geometry.

To estimate geometry from a video sequence, we could jointly optimize a single free-form shape constrained by a



a) b) c) d)

Figure 2. Overview of our method. The input to our method is an image sequence with corresponding segmentations. We first calculate poses using the SMPL model (a). Then we unpose silhouette camera rays (unposed silhouettes depicted in red) (b) and optimize for the subjects shape in the canonical T-pose (c). Finally, we are able to calculate a texture and generate a personalized blend shape model (d).

body model to fit a set of F images. Unfortunately, this requires to optimize F poses at once and more importantly it requires storing F models in memory during optimization which makes it computationally expensive and unpractical.

The key idea of our approach is to generalize visual hull methods [41] to monocular videos of people in motion. Standard visual hull methods capture a static shape from multiple views. Every camera ray through a silhouette point in the image casts a constraint on the 3D body shape. To make visual hulls work for monocular video of a moving person it is necessary to "undo" the human motion and bring it to a canonical frame of reference. In this work, the geometry of people (in wide or tight clothing) is represented as a deviation from the SMPL parametric body model [40] of naked people in a canonical T-pose; this model also features a pose-dependent non-rigid surface skinning. We first estimate an initial body shape and 3D pose at each frame by fitting the SMPL model to 2D detections similar to [37, 7]. Given such fits, we associate every silhouette point in every frame to a 3D point in the body model. We then transform every projection ray according to the inverse deformation model of its corresponding 3D model point; we call this operation unposing (Fig. 3). After unposing the rays for all frames we obtain a visual hull that constrains the body shape in a canonical T-pose. We then jointly optimize body shape parameters and free-form vertex displacements to minimize the distance between 3D model points and unposed rays. This allows us to efficiently optimize a single displacement surface on top of SMPL constrained to fit all frames at once, which requires storing only one model in memory (Fig. 2). Our technique allows for the first time extracting accurate 3D human body models, including hair and clothing, from a single video sequence of the person moving in front of the camera such that the person is seen from all sides.

Our results on several 3D datasets show that our method can reconstruct 3D human shape to a remarkable accuracy of 4.5 mm (even higher 3.1 mm with ground truth poses) despite monocular depth ambiguities. We provide our dataset and source code of our method for research purposes [1].

2. Related Work

Shape reconstruction of humans in clothing can be classified according to two criteria: (1) the type of sensor used and (2) the kind of template prior used for reconstruction. *Free-form* methods typically use multi-view cameras, depth cameras or fusion of sensors and reconstruct surface geometry quite accurately without using a strong prior on the shape. In more unconstrained and ambiguous settings, such as in the monocular case, a parametric body model helps to constrain the problem significantly. Here we review free-form and model-based methods and focus on methods for monocular images.

Free-form methods reconstruct the moving geometry by deforming a mesh [12, 19, 10] or using a volumetric representation of shape [30, 2]. The advantage of these methods is that they allow reconstruction of general dynamic shapes provided that a template surface is available initially. While flexible, such approaches require high-quality *multi-view* input data which makes them impractical for many applications. Only one approach showed reconstruction of human pose and deforming cloth geometry from monocular video using a pre-captured shape template [74]. Using a *depth camera*, systems like KinectFusion [33, 45] allow reconstruction of 3D rigid scenes and also appearance models [82] by incrementally fusing geometry in a canonical frame. A number of methods adapt KinectFusion for human body scanning [58, 39, 79, 17]. The problem is that these methods require separate shots at different time instances. The person thus needs to stand still while the camera is turned around, or subtle pose changes need to be explicitly compensated. The approach in [44] generalized KinectFusion to non-rigid objects. The approach performs non-rigid registration between the incoming depth frames and a concurrently updated, initially incomplete, template. While general, such template-free approaches [45, 31, 60] are limited to slow and careful motions. One way to make fusion and tracking more robust is by using multiple kinects [21, 47] or multi-view [63, 38, 16]; such methods achieve impressive reconstructions but do not register all

frames to the same template and focus on different applications such as streaming or remote rendering for telepresence, e.g., in the holoportation project [47]. Pre-scanning the object or person to be tracked [86, 19] reduces the problem to tracking the non-rigid deformations. Some works are in-between free-form and model-based methods. In [23, 69] they pre-scan a template and insert a skeleton and in [78] they use a skeleton to regularize dynamic fusion. Our work is also related to the seminal work of [14, 15] where they align visual hulls over time to improve shape estimation. In the articulated case, they need to segment and track every body part separately and then merge the information together in a coarse voxel model; more importantly, they need multi-view input. In [35] they compensate for small motions of captured objects by de-blurring occupancy images but no results are shown for moving humans. In [85] they reconstruct the shape of clothed humans in outdoor environments from RGB video, requiring the subject to stand still. All these works use either multi-view systems, depth cameras or do not handle moving humans. In contrast, we use a single RGB video of a moving person, which makes the problem significantly harder as geometry can not be directly unwarped as it is done in depth fusion papers.

Model-based. Several works leverage a parametric body model for human pose and shape estimation from images [52]. Early models in computer vision were based on simple primitives [43, 24, 48, 59]. Recent ones are learned from thousands of scans of real people and encode pose, and shape deformations [4, 28, 40, 87, 51]. Some works reconstruct the body shape from *depth data* sequences [71, 29, 76, 81, 6] exploiting the temporal information. Typically, a single shape and multiple poses are optimized to exploit the temporal information. Using *multi-view* some works have shown performance capture outdoors [54, 55] by leveraging a sum of Gaussians body model [64] or using a pre-computed template [77]. A number of works are restricted to estimating the shape parameters of a body model [5, 25] from multiple views or single images with manually clicked points; silhouettes shading cues and color have been used for inference. Some works fit a body model to images using manual intervention [83, 34, 57] with the goal of image manipulation. Shape and clothing from a single image is recovered in [26, 13] but the user needs to click points in the image and select the clothing types from a database. In [36] they obtain shape from contour drawings. The advance in 2D pose detection [70, 11, 32] has made 3D pose and shape estimation possible in challenging scenarios. In [7, 37] they fit a 3D body model [40] to 2D detections; since only model parameters are optimized and these methods heavily rely on 2D detections, results tend to be close to the shape space mean. In [3] they add a silhouette term to reduce this effect.

Shape Under Clothing. The aforementioned methods ignore clothing or treat it as noise, but a number of works explicitly reason about clothing. Typically, these methods incorporate constraints such as the body should lie inside the clothing silhouette. In [5] they estimate body shape under clothing by optimizing model parameters for a set of images of the same person in different clothing. In [73, 75] they exploit temporal sequences of scans to estimate shape under clothing. Results are usually restricted to the (naked) model space. In [80] they estimate detailed shape under clothing from scan sequences by optimizing a free-form surface constrained by a body model. The approach in [50] jointly captures clothing geometry and body shape using separate meshes but requires 3D scan sequences as input. DoubleFusion [66] reconstructs clothing geometry and inner body shape from a single depth camera in real time.

Learning based. Only very few works predict human shape from images using learning methods since images annotated with ground truth shape, pose and clothing geometry are hardly available. A few exceptions are the approach of [20] that predicts shape from silhouettes using a neural network and [18] that predicts garment geometry from a single image. Predictions in [20] are restricted to model shape space and tend to look over-smooth; only garments seen in the dataset can be recovered in [18]. Recent works leverage 2D annotations to train networks for the task of 3D pose estimation [42, 53, 84, 65, 68, 56]. Such works typically predict a stick figure or bone skeleton only, and can not estimate body shape or clothing.

3. Method

Given a single monocular RGB video depicting a moving person, our goal is to generate a personalized 3D model of the subject, which consists of the shape of body, hair and clothing, a personalized texture map, and an underlying skeleton rigged to the surface. Non-rigid surface deformations in new poses are thus entirely skeleton-driven. Our method consists of 3 steps: 1) *pose reconstruction* (Sec. 3.2) 2) *consensus shape estimation* (Sec. 3.3) and 3) *frame refinement and texture map generation* (Sec. 3.4). Our main contribution is step 2), the consensus shape estimation; step 1) builds on previous work and step 3) to obtain texture and time-varying details is optional.

In order to estimate the consensus shape of the subject, we first calculate the 3D pose in each frame (Sec. 3.2). We extend the method of [7] to make it more robust and enforce better temporal coherence and silhouette overlap. In the second step, the *consensus shape* is calculated as detailed in Sec. 3.3. The consensus shape is efficiently optimized to maximally explain the silhouettes at each frame instance. Due to time-varying cloth deformations the posed consensus shape might be slightly misaligned with the frame silhouettes. Hence, in order to compute texture and capture

time-varying details, in step 3) deviations from the consensus shape are optimized per frame in a sliding window approach (Sec. 3.4). Given the refined frame-wise shapes we can compute the texture map. Our method relies on a foreground segmentation of the images. Therefore, we adopt the CNN based video segmentation method of [9] and train it with 3-4 manual segmentations per sequence. In order to counter ambiguities in monocular 3D human shape reconstruction, we use the SMPL body model [40] as starting point. In the following, we briefly explain how we adapt original SMPL body model for our problem formulation.

3.1. SMPL Body Model with Offsets

SMPL is a parameterized model of naked humans that takes 72 pose and 10 shape parameters and returns a triangulated mesh with $N = 6890$ vertices. The shape $\boldsymbol{\beta}$ and pose $\boldsymbol{\theta}$ deformations are applied to a base template $\mathbf{T}$, which in the original SMPL model corresponds to the statistical mean shape in the training scans $\mathbf{T}_\mu$:

$$M(\boldsymbol{\beta}, \boldsymbol{\theta}) = W(T(\boldsymbol{\beta}, \boldsymbol{\theta}), J(\boldsymbol{\beta}), \boldsymbol{\theta}, \mathbf{W}) \quad (1)$$

$$T(\boldsymbol{\beta}, \boldsymbol{\theta}) = \mathbf{T}_\mu + B_s(\boldsymbol{\beta}) + B_p(\boldsymbol{\theta}) \quad (2)$$

where W is a linear blend-skinning function applied to a rest pose $T(\boldsymbol{\beta}, \boldsymbol{\theta})$ based on the skeleton joints $J(\boldsymbol{\beta})$ and after pose-dependent deformations $B_p(\boldsymbol{\theta})$ and shape dependent deformations $B_p(\boldsymbol{\theta})$ are applied. Shape-dependent deformations $B_s(\boldsymbol{\beta})$ model subject identity. However the Principal Component shape space of SMPL was learned from scans of naked humans, so clothing and other personal surface detail cannot be modeled. In order to personalize the SMPL model, we simply add a set of auxiliary variables or offsets $\mathbf{D} \in \mathbb{R}^{3N}$ from the template:

$$T(\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{D}) = \mathbf{T}_\mu + B_s(\boldsymbol{\beta}) + B_p(\boldsymbol{\theta}) + \mathbf{D} \quad (3)$$

Such offsets $\mathbf{D}$ allow us to deform the model to better explain details and clothing. Offsets are optimized in step 2.

3.2. Pose Reconstruction

The approach in [7] optimizes SMPL model parameters to fit a set of 2D joint detections in the image. As with any monocular method, scale is an inherent ambiguity. To mitigate this effect, we take inspiration from [54] and extend [7] such that it jointly considers $P = 5$ frames and optimizes a single shape and $P = 5$ poses. Note that optimizing many more frames would become computationally very expensive and many models would have to be simultaneously stored in memory. Our experiments reveal that even when optimizing over $P = 5$ poses the scale ambiguity prevails. The reason is that pose differences induce additional 3D ambiguities which cannot be uniquely decoupled from global size, even on multiple frames [67, 61, 49]. Hence, if the height of the person is known, we incorporate it as constraint during optimization. If height is not known the shape reconstructions of our method are still accurate up to a scale factor (height estimation is roughly off by 2-5 cm). The output of initialization are SMPL model shape parameters $\boldsymbol{\beta}_0$ that we keep fixed during subsequent frame-wise pose estimation. In order to estimate 3D pose more reliably, we extend [7] by incorporating a silhouette term:

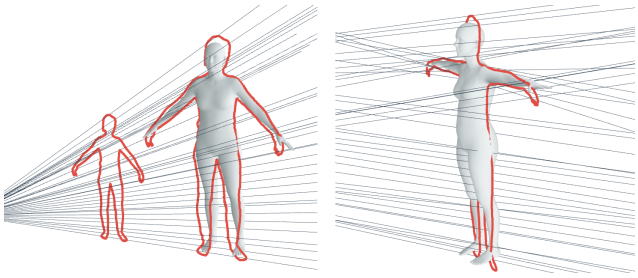


Figure 3. The camera rays that form the image silhouette (left) are getting unposed into the canonical T-pose (right). This allows efficient shape optimization on a single model for multiple frames.

$$E_{\text{silh}}(\boldsymbol{\theta}) = G(\mathbf{w}_{\text{o}}\mathbf{I}_{rn}(\boldsymbol{\theta})\mathbf{C} + \mathbf{w}_{\text{i}}(\mathbf{1} - \mathbf{I}_{rn}(\boldsymbol{\theta}))\bar{\mathbf{C}}) \quad (4)$$

with the silhouette image of the rendered model $\mathbf{I}_{rn}(\boldsymbol{\theta})$, distance transform of observed image mask $\mathbf{C}$ and its inverse $\bar{\boldsymbol{C}}$, weights $\mathbf{w}$. To be robust to local minima we optimize at 4 different levels of a Gaussian pyramid G. We further update the method to use state of the art 2D joint detections [11, 70] and a single-modal A-pose prior. We train the prior from SMPL poses fitted against body scans of people in A-pose. Further, we enforce a temporal smoothness and initialize the pose in a new frame with the estimated pose $\boldsymbol{\theta}$ in the previous frame. If the objective error gets too large, we re-initialize the tracker by setting the pose to zero. While optimization in batches of frames would be beneficial it slows down computation and we have not found significant differences in pose accuracy. The output of this step is a set of poses $\{\boldsymbol{\theta}_p\}_{p=1}^{F}$ for the F frames in the sequence.

3.3. Consensus Shape

Given the set of estimated poses we could jointly optimize a single refined shape matching all original F poses, which would yield a complex, non-convex optimization problem. Instead, we merge all the information into an unposed canonical frame, where refinement is computationally easier. At every frame a silhouette places a new constraint on the body shape; specifically, the set of rays going from the camera to the silhouette points define a constraint cone, see Fig. 3. Since the person is moving, the pose is changing. Our key idea is to *unpose* the cone defined by the projection rays using the estimated poses. Effectively, we invert the SMPL function for every ray. In SMPL, every vertex $\mathbf{v}$ deforms according to the following equation:

$$\mathbf{v}'_i = \sum_{k=1}^{K} w_{k,i} G_k(\boldsymbol{\theta}, J(\boldsymbol{\beta}))(\mathbf{v}_i + b_{s,i}(\boldsymbol{\beta}) + b_{P,i}(\boldsymbol{\theta})) \quad (5)$$

where G_k is the global transformation of joint k and $b_{s,i}(\boldsymbol{\beta}) \in \mathbb{R}$ and $b_{P,i}(\boldsymbol{\theta})$ are elements of $B_s(\boldsymbol{\beta})$ and $B_p(\boldsymbol{\theta})$ corresponding to $i-th$ vertex. For every ray $\mathbf{r}$ we find its closest 3D model point. From Eq. (5) it follows that the inverse transformation applied to a ray $\mathbf{r}$ corresponding to model point $\mathbf{v}'_i$ is

$$\mathbf{r} = \left(\sum_{k=1}^{K} w_{k,i} G_k(\boldsymbol{\theta}, J(\boldsymbol{\beta}))\right)^{-1} \mathbf{r}' - b_{P,i}(\boldsymbol{\theta}). \quad (6)$$

Doing this for every ray effectively unposes the silhouette cone and places constraints on a canonical T-pose, see Fig. 3. Unposing removes blend-shape calculations from the optimization problem and significantly reduces the memory foot-print of the method. Without unposing the vertex operations and the respective Jacobians would have to be computed for every frame at *every update* of the shape. Given the set of unposed rays for F silhouettes (we use $F = 120$ in all experiments), we formulate an optimization in the canonical frame

$$E_{\text{cons}} = E_{\text{data}} + w_{\text{lp}} E_{\text{lp}} + w_{\text{var}} E_{\text{var}} + w_{\text{sym}} E_{\text{sym}} \quad (7)$$

and minimize it with respect to shape parameters $\boldsymbol{\beta}$ of a template model and the vertex offsets $\mathbf{D}$ defined in Eq. 3. The objective E_{cons} consists of a data term E_{data} and three regularization terms $E_{\text{lp}}, E_{\text{var}}, E_{\text{sym}}$ with weights w_* that balance its influence.

Data Term measures the distance between vertices and rays. Point to line distances can be efficiently computed expressing rays using Plucker coordinates ($\mathbf{r} = \mathbf{r}_m, \mathbf{r}_n$). Given a set of correspondences $(\mathbf{v}_i, \mathbf{r}) \in \mathcal{M}$ the data term equals

$$E_{\text{data}} = \sum_{(\mathbf{v},\mathbf{r})\in\mathcal{M}} \rho(\mathbf{v} \times \mathbf{r}_n - \mathbf{r}_m) \quad (8)$$

where ρ is the Geman-McClure robust cost function, here applied to the point to line distance. Since the canonical pose parameters are all zero ($\boldsymbol{\theta} = \mathbf{0}$) it follows from Eq. 3 that vertex positions are a function of shape parameters and offsets $\mathbf{v}(\boldsymbol{\beta}_0, \mathbf{D}) = T_i(\boldsymbol{\beta}_0, \mathbf{D}) = (\mathbf{v}_{\mu,i} + b_{s,i}(\boldsymbol{\beta}_0) + \mathbf{d}_i)$, where $\mathbf{d}_i \in \mathbb{R}^3$ is the offset in $\mathbf{D}$ corresponding to vertex $\mathbf{v}_i$. In our notation, we remove the dependency on parameters for clarity. The remaining terms regularize the optimization.

Laplacian Term. We enforce smooth deformation by adding the Laplacian mesh regularizer [62]:

$$E_{\text{lp}} = \sum_{i=1}^{N} \tau_{l,i} ||L(\mathbf{v}_i) - \delta_i||^2 \quad (9)$$

where $\delta = L(\mathbf{v}(\boldsymbol{\beta}_0, \mathbf{0}))$ and L is the Laplace operator. The term forces the Laplacian of the optimized mesh to be similar to the Laplacian of the mesh at initialization (where offsets $\mathbf{D} = \mathbf{0}$).

Body Model Term. We penalize deviations of the reconstructed free-form vertices $\mathbf{v}(\boldsymbol{\beta}_0, \mathbf{D})$ from vertices explained by the SMPL model $\mathbf{v}(\boldsymbol{\beta}, \mathbf{0})$:

$$E_{\text{var}} = \sum_{i=1}^{N} \tau_{v,i} ||\mathbf{v}_i(\boldsymbol{\beta}_0, \mathbf{D}) - \mathbf{v}_i(\boldsymbol{\beta}, \mathbf{0})||^2 \quad (10)$$

Symmetry Term. Humans are usually axially symmetrical with respect to the Y-axis. Since the body model is nearly symmetric, we add a constraint on the offsets alone that enforces a symmetrical shape:

$$E_{\text{sym}} = \sum_{(i,j)\in\mathcal{S}} \tau_{s,i,j} \left|\left|[-1,1,1]^T \cdot \mathbf{d}_i - \mathbf{d}_j\right|\right|^2 \quad (11)$$

where $\mathcal{S}$ contains all pairs of Y-symmetric vertices. We phrase this as a soft-constraint to allow potential asymmetries in clothing wrinkles and body shapes. Since the refined consensus shape still has the mesh topology of SMPL, we can apply the pose-based deformation space of SMPL to simulate surface deformation in new skeleton poses.

Implementation Details. Body regions that are typically unclothed or where silhouettes are noisy (face, ears, hands, and feet) are more regularized towards the body model using per-vertex weights $\boldsymbol{\tau}$. We optimize E_{cons} using a "dog-leg" trust region method using the chumpy auto-differentiation framework. We alternate minimizing E_{cons} with respect to model parameters and offsets and finding point to line correspondences. We also re-initialize E_{lp}, E_{var}, E_{sym}. More implementation details and runtime metrics are given in the supplementary material.

3.4. Frame Refinement and Texture Generation

After calculating a *global* shape for the given sequence, we aim to capture the temporal variations. We adapt the energy in Eq. 7 to process frames sequentially. The optimization is initialized with the preceding frame and regularized with neighboring frames:

$$E_{\text{ref},j} = \sum_{j=f-m}^{f+m} \psi_j E_{\text{data},j} + w_{\text{var}} E_{\text{var},j} + w_{\text{lp}} E_{\text{lp},j} + w_{\text{last}} E_{\text{last},j} \quad (12)$$

where $\psi_j = 1$ for $j = k$ and $\psi_j = w_{\text{neigh}} < 1$ for neighboring frames. Hence, w_{neigh} defines the influence of neighboring frames and E_{last} regularizes the reconstruction to the result of the preceding frame. To create the texture, we warp our estimated canonical model back to each frame, back-project the image color to all visible vertices, and finally generate a texture image by calculating the median of the most orthogonal texels from all views. An example of keyframes we use for texture mapping and the resulting texture image is shown in Fig. 4.

Figure 4. We back-project the image color from several frames to all visible vertices to generate a full texture map.

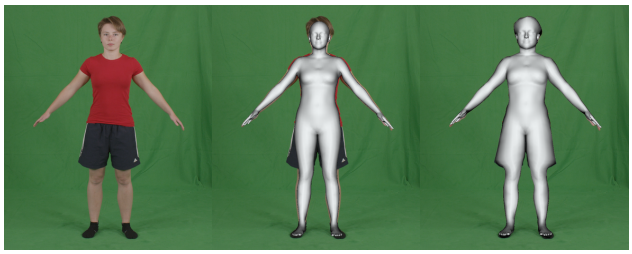


Figure 5. Comparison to the monocular model-based method [7] (left to right) input frame, SMPLify, consensus shape. To make a fair comparison we extended [7] to multiple views as well. Compared to pure model-based methods, our approach captures also medium level geometry details from a single RGB camera.

4. Experiments

We study the effectiveness of our method, qualitatively and quantitatively, in different scenarios. For quantitative evaluation, we used two publicly available datasets consisting of 3D scan sequences of humans in motion: with minimal clothing (MC) (DynamicFAUST [8]) and with clothing (BUFF [80]). Since these datasets were recorded without RGB sensors we simply render images of the scans using a virtual camera and use them as input. In order to evaluate our method on more varied clothing and backgrounds, we captured a new test dataset (People-Snapshot dataset), and present qualitative results. To the best of our knowledge, our method is the first approach that enables detailed human body model reconstruction in clothing from a single monocular RGB video without requiring a pre-scanned template or manually clicked points. Thus, there exist no methods with the same setting as ours. Hence, we provide a quantitative comparison to the state-of-the-art RGB-D based approach KinectCap [6] on their dataset. The image sequences and ground truth scans were provided by the authors of [6]. While reconstruction from monocular videos is much harder than from depth videos, a comparison is still informative. In all experiments, the method's parameters are set to two constant values, one set for clothed and one set for people in MC, which are empirically determined.

4.1. Results on Rendered Images

We take all 9 sequences of 5 different subjects in the BUFF dataset and all 9 sequences of 9 subjects from the DynamicFaust dataset performing "Hip" movements, featuring strong fabric movement or soft tissue dynamics respectively. Each dynamic sequence consists of 300-800 frames. To simulate the subject rotating in front of a camera, we create a virtual camera at 2.5 meters away from the 3D scans of the subject. We rotate the camera in a circle around the person moving one time per sequence. The foreground masks are easily obtained from the alpha channel of the rendered images. For BUFF we render images with real dynamic textures; for DynamicFAUST since textures are not available we rendered shaded models. In Fig. 6, we show some examples of our reconstruction results on image sequences rendered from BUFF and DynamicFAUST scans. The complete results of all 9 sequences are provided in the supplementary material. To be able to quantitatively evaluate the reconstruction quality, we adjust the pose and scale of our reconstruction to match the ground truth body scans following [80, 6]. Then, we compute a bi-directional vertex to surface distance between our reconstruction and the ground truth geometry. Per-vertex errors (in millimeters) on all sequences are provided in Tab. 1. The heatmaps of per-vertex errors are shown in Fig. 6. As can be seen, our method yields accurate reconstruction on all sequences including personalized details. To study the importance of the pose estimation component, we report the accuracy of our method using *ground truth poses* versus using estimated poses *full method*. Ground truth poses were obtained by registering SMPL to the 3D scans. The results of the ablation evaluation are also shown in Fig. 6 and Tab. 1. We can see that our complete pipeline achieved comparable accuracy with the one using ground truth poses which demonstrates robustness. Results show that there is still room for improvement in 3D pose reconstruction.

4.2. Qualitative Results on RGB Images

We also evaluate our method on real image sequences. The People-Snapshot dataset consists of 24 sequences of 11 subjects varying a lot in height and weight. The sequences are captured with a fixed camera, and we ask the subjects to rotate while holding an A-pose. To cover a variety of clothing, lighting conditions and background, the subjects were captured with varying sets of garments and with three different background scenes: in the studio with green screen, outdoor, and indoor with complex dynamic background. Some examples of our reconstruction results are shown in Fig. 7 and Fig. 1. We show more example in the supplementary material and in the video. We can see that our method yields detailed reconstructions of similar quality as the results on rendered sequences, which demonstrates that our method generalizes well on the real world scenarios. The benefits of our method are further evidenced by overlaying the reposed final reconstruction on to the input images. As shown in Fig. 9, our reconstructions precisely overlay the body silhouettes in the input images.

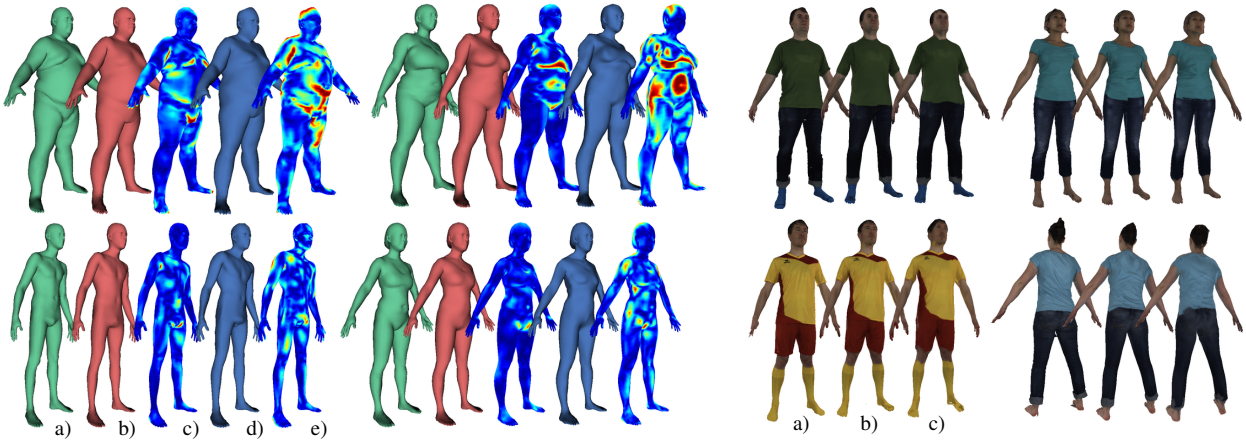


Figure 6. Our results on image sequences from BUFF and D-FAUST datasets. Left we show D-FAUST: (a) ground truth 3D scan, (b) consensus shape with ground truth poses (consensus-p), (c) consensus-p heatmap, (d) consensus shape (consensus), (e) consensus heat-map (blue means 0mm, red means $\geq$ 2cm). Right we show textured results on BUFF: (a) ground truth scan, (b) consensus-p (c) consensus.

Subject ID	*full method*	*GT poses*
50002	5.13 ±6.43	3.92 ±4.49
50004	4.36 ±4.67	2.95 ±3.11
50009	3.72 ±3.76	2.56 ±2.50
50020	3.32 ±3.04	2.27 ±2.06
50021	4.45 ±4.05	3.00 ±2.66
50022	5.71 ±5.78	2.96 ±2.97
50025	4.84 ±4.75	2.92 ±2.94
50026	4.56 ±4.83	2.62 ±2.48
50027	3.89 ±3.57	2.55 ±2.33

	Subject ID	*full method*	*GT poses*
t-shirt, long pants	00005	5.07 ±5.74	3.80 ±4.13
	00032	4.84 ±5.25	3.37 ±3.59
	00096	5.57 ±6.54	4.35 ±4.66
	00114	4.22 ±5.12	3.14 ±2.99
	03223	4.85 ±4.80	2.87 ±2.58
soccer outfit	00005	5.35 ±6.67	3.82 ±3.67
	00032	7.95 ±8.62	3.04 ±3.39
	00114	4.97 ±5.81	3.01 ±2.80
	03223	5.49 ±5.71	3.21 ±3.28

Subject ID		*Subject ID*	
00009	4.07 ±4.20	02909	3.94 ±4.80
00043	4.30 ±4.39	03122	3.21 ±2.85
00059	3.87 ±3.96	03123	3.68 ±3.22
00114	4.85 ±4.93	03124	3.67 ±3.31
00118	3.79 ±3.80	03126	4.89 ±6.12

Table 1. Numerical evaluation on 3 different datasets with ground truth 3D shapes. On D-FAUST and BUFF we rendered the ground truth scans on a virtual camera (see text), KinectCap already included images. We report for every subject the average surface to surface distance (see text). On BUFF, D-FAUST and KinectCap we achieve mean average errors of 5.37mm, 4.44mm, 3.97mm respectively. As expected best results are obtained using ground truth poses. Perhaps surprisingly, the results (3.40 mm for BUFF, 2.86 for D-FAUST) do not differ much from the average errors of the full pipeline. This demonstrates that our approach is robust to inaccuracies in 3D pose estimation.

4.3. Comparison with KinectCap

We compare our method to [6] on their collected dataset. Subjects were captured in both A-pose and T-poses in this dataset. Since T-poses (zero-pose in SMPL) are rather unnatural, they are not well captured in our general pose-prior. Hence, we adjust our pose prior to contain also T-poses. Note that their method relies on depth data, while ours only uses the RGB images. Notably, our method obtains comparable results qualitatively and quantitatively despite solving a much more ill-posed problem. This is further evidenced by the per-vertex errors in Tab. 1.

4.4. Surface Refinement Using Shading

As mentioned before, our method captures both body shape and medium level surface geometry. In contrast to pure model-based methods, we already add significant details (Fig. 5). Using existing shape from shading methods the reconstruction can be further improved by adding the finer level details of the surface, e.g. folding and wrinkles. Fig. 10 shows an example result of applying the shape from shading method of [72] to our reconstruction. This application further demonstrates the accuracy of our reconstruction, since such good result cannot be obtained without an accurate model-to-image alignment.

5. Discussion and Conclusions

We have proposed the first approach to reconstruct a personalized 3D human body model from a single video of a moving person. The reconstruction comprises personalized geometry of hair, body, and clothing, surface texture, and an underlying model that allows changes in pose and shape. Our approach combines a parametric human body model extended by surface displacements for refinement, and a novel method to morph and fuse the dynamic human silhouette cones in a common frame of reference. The fused cones merge the shape information contained in the video, allowing us to optimize a detailed model shape. Our algorithm not only captures the geometry and appearance of the surface, but also automatically rigs the body model with a kinematic skeleton enabling approximate pose-dependent surface deformation. Quantitative results demonstrate that our approach can reconstruct human body shape with an accuracy of 4.5mm and an ablation analysis shows robustness to noisy 3D pose estimates.

The presented method finds its limits in appearances that do not share the same topology as the body: long open hair or skirts can not be modeled as an offset from the body. Furthermore, we can only capture surface details that are seen

Figure 7. Qualitative results: since the reconstructed templates share the topology with the SMPL body model we can use SMPL to change the *pose and shape* of our reconstructions. While SMPL does not model clothing deformations the deformed templates look plausible and maybe of sufficient quality for several applications.

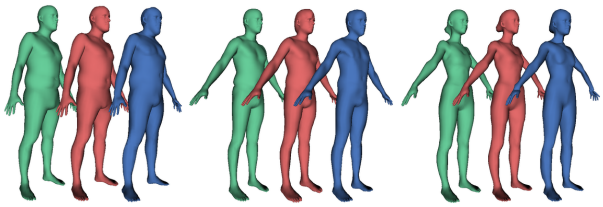


Figure 8. Comparison to the RGB-D based method of [6] (red) and ground truth scans (green). Our approach (blue) achieves similar qualitative results despite using a monocular video sequence as opposed to a depth camera. Their approach is more accurate numerically 2.54 mm versus 3.97 mm but our results are comparable despite using a single RGB camera.

Figure 9. Side-by-side comparison of our reconstructions (right) and the input images (left). As can be seen from the right side, our reconstructions precisely overlay on the input images. The reconstructed models rendered in a side view are shown at bottom right.

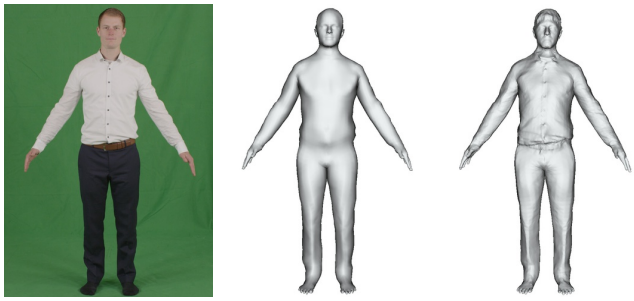


Figure 10. Our reconstruction can be further improved by adding the finer level details of the surface using shape from shading.

on the outline of at least one view. This means especially concave regions like armpits or inner thighs are sometimes not well handled. Strong fabric movement caused by fast skeletal motions will additionally result in decreased level of detail. In future work, we plan to incorporate illumination and material estimation alongside with temporally varying textures in our method to enable realistic rendering and video augmentation.

For the first time, our method can extract realistic avatars including hair and clothing from a moving person in a *monocular RGB video*. Since cameras are ubiquitous and low cost, people will be able to digitize themselves and use the 3D human models for VR applications, entertainment, biometrics or virtual try-on for online shopping. Furthermore, our method precisely aligns models with the images, which opens up many possibilities for image editing.

Acknowledgments

The authors gratefully acknowledge funding by the German Science Foundation from project DFG MA2555/12-1. We would like to thank Rudolf Martin and Juan Mateo Castrillon Cuervo for great help in data collection and processing. Another thanks goes to Federica Bogo and Javier Romero for providing their results for comparison.

References

[1] https://graphics.tu-bs.de/people-snapshot. 2

[2] B. Allain, J.-S. Franco, and E. Boyer. An Efficient Volumetric Framework for Shape Tracking. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 268–276, Boston, United States, 2015. IEEE. 2

[3] T. Alldieck, M. Kassubeck, B. Wandt, B. Rosenhahn, and M. Magnor. Optical flow-based 3d human motion estimation from monocular video. In *German Conf. on Pattern Recognition*, pages 347–360, 2017. 3

[4] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis. SCAPE: shape completion and animation of people. In *ACM Transactions on Graphics*, volume 24, pages 408–416. ACM, 2005. 3

[5] A. O. Bălan and M. J. Black. The naked truth: Estimating body shape under clothing. In *European Conf. on Computer Vision*, pages 15–29. Springer, 2008. 1, 3

[6] F. Bogo, M. J. Black, M. Loper, and J. Romero. Detailed full-body reconstructions of moving people from monocular RGB-D sequences. In *IEEE International Conf. on Computer Vision*, pages 2300–2308, 2015. 1, 3, 6, 7, 8

[7] 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 *European Conf. on Computer Vision*. Springer International Publishing, 2016. 1, 2, 3, 4, 6

[8] F. Bogo, J. Romero, G. Pons-Moll, and M. J. Black. Dynamic FAUST: Registering human bodies in motion. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2017. 6

[9] S. Caelles, K. Maninis, J. Pont-Tuset, L. Leal-Taixé, D. Cremers, and L. Van Gool. One-shot video object segmentation. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2017. 4

[10] C. Cagniart, E. Boyer, and S. Ilic. Probabilistic deformable surface tracking from multiple videos. In K. Daniilidis, P. Maragos, and N. Paragios, editors, *European Conf. on Computer Vision*, volume 6314 of *Lecture Notes in Computer Science*, pages 326–339, Heraklion, Greece, 2010. Springer. 2

[11] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2017. 3, 4

[12] J. Carranza, C. Theobalt, M. A. Magnor, and H.-P. Seidel. Free-viewpoint video of human actors. In *ACM Transactions on Graphics*, volume 22, pages 569–577. ACM, 2003. 2

[13] X. Chen, Y. Guo, B. Zhou, and Q. Zhao. Deformable model for estimating clothed and naked human shapes from a single image. *The Visual Computer*, 29(11):1187–1196, 2013. 3

[14] G. K. Cheung, S. Baker, and T. Kanade. Shape-from-silhouette of articulated objects and its use for human body kinematics estimation and motion capture. In *IEEE Conf. on Computer Vision and Pattern Recognition*, volume 1, pages I–I. IEEE, 2003. 3

[15] G. K. Cheung, S. Baker, and T. Kanade. Visual hull alignment and refinement across time: A 3d reconstruction algorithm combining shape-from-silhouette with stereo. In *IEEE Conf. on Computer Vision and Pattern Recognition*, volume 2, pages II–375. IEEE, 2003. 3

[16] A. Collet, M. Chuang, P. Sweeney, D. Gillett, D. Evseev, D. Calabrese, H. Hoppe, A. Kirk, and S. Sullivan. High-quality streamable free-viewpoint video. *ACM Transactions on Graphics*, 34(4):69, 2015. 2

[17] Y. Cui, W. Chang, T. Nöll, and D. Stricker. Kinectavatar: fully automatic body capture using a single kinect. In *Asian Conf. on Computer Vision*, pages 133–147. Springer, 2012. 2

[18] R. Daněřek, E. Dibra, C. Öztireli, R. Ziegler, and M. Gross. Deepgarment: 3d garment shape estimation from a single image. In *Computer Graphics Forum*, volume 36, pages 269–280. Wiley Online Library, 2017. 3

[19] E. De Aguiar, C. Stoll, C. Theobalt, N. Ahmed, H.-P. Seidel, and S. Thrun. Performance capture from sparse multi-view video. In *ACM Transactions on Graphics*, volume 27, page 98. ACM, 2008. 2, 3

[20] E. Dibra, H. Jain, C. Oztireli, R. Ziegler, and M. Gross. Human shape from silhouettes using generative hks descriptors and cross-modal neural networks. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 4826–4836, 2017. 1, 3

[21] 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, 2

[22] S. Fuhrmann, F. Langguth, and M. Goesele. Mve-a multi-view reconstruction environment. In *EUROGRAPHICS Workshops on Graphics and Cultural Heritage*, pages 11–18, 2014. 1

[23] 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 Conf. on Computer Vision and Pattern Recognition*, pages 1746–1753. IEEE, 2009. 3

[24] D. M. Gavrila and L. S. Davis. 3-d model-based tracking of humans in action: a multi-view approach. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 73–80. IEEE, 1996. 3

[25] P. Guan, A. Weiss, A. O. Bălan, and M. J. Black. Estimating human shape and pose from a single image. In *IEEE International Conf. on Computer Vision*, pages 1381–1388. IEEE, 2009. 3

[26] Y. Guo, X. Chen, B. Zhou, and Q. Zhao. Clothed and naked human shapes estimation from a single image. *Computational Visual Media*, pages 43–50, 2012. 3

[27] N. Hasler, H. Ackermann, B. Rosenhahn, T. Thormahlen, and H.-P. Seidel. Multilinear pose and body shape estisma-tion of dressed subjects from image sets. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1823–1830. IEEE, 2010. 1

[28] N. Hasler, C. Stoll, M. Sunkel, B. Rosenhahn, and H.-P. Seidel. A statistical model of human pose and body shape. In *Computer Graphics Forum*, volume 28, pages 337–346, 2009. 3

[29] T. Helten, A. Baak, G. Bharaj, M. Muller, H.-P. Seidel, and C. Theobalt. Personalization and evaluation of a real-time depth-based full body tracker. In *International Conf. on 3D Vision*, pages 279–286, Washington, DC, USA, 2013. 3

[30] C.-H. Huang, B. Allain, J.-S. Franco, N. Navab, S. Ilic, and E. Boyer. Volumetric 3d tracking by detection. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 3862–3870, 2016. 2

[31] M. Innmann, M. Zollhöfer, M. Nießner, C. Theobalt, and M. Stamminger. Volumedeform: Real-time volumetric non-rigid reconstruction. In *European Conf. on Computer Vision*, 2016. 2

[32] E. Insafutdinov, M. Andriluka, L. Pishchulin, S. Tang, E. Levinkov, B. Andres, and B. Schiele. Arttrack: Articulated multi-person tracking in the wild. In *IEEE Conf. on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, 2017. IEEE. 3

[33] S. Izadi, D. Kim, O. Hilliges, D. Molyneaux, R. Newcombe, P. Kohli, J. Shotton, S. Hodges, D. Freeman, A. Davison, et al. Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera. In *ACM symposium on User interface software and technology*, pages 559–568. ACM, 2011. 2

[34] A. Jain, T. Thormählen, H.-P. Seidel, and C. Theobalt. Moviereshape: Tracking and reshaping of humans in videos. In *ACM Transactions on Graphics*, volume 29, page 148. ACM, 2010. 1, 3

[35] S. M. Khan and M. Shah. Reconstructing non-stationary articulated objects in monocular video using silhouette information. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008. 3

[36] V. Kraevoy, A. Sheffer, and M. van de Panne. Modeling from contour drawings. In *Eurographics Symposium on Sketch-Based interfaces and Modeling*, pages 37–44. ACM, 2009. 3

[37] 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*, 2017. 2, 3

[38] V. Leroy, J.-S. Franco, and E. Boyer. Multi-View Dynamic Shape Refinement Using Local Temporal Integration. In *IEEE International Conf. on Computer Vision*, Venice, Italy, 2017. 2

[39] 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. 1, 2

[40] M. M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics*, 34(6):248:1–248:16, 2015. 2, 3, 4

[41] W. Matusik, C. Buehler, R. Raskar, S. J. Gortler, and L. McMillan. Image-based visual hulls. In *Annual Conf. on Computer Graphics and Interactive Techniques*, pages 369–374. ACM Press/Addison-Wesley Publishing Co., 2000. 2

[42] D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei, H.-P. Seidel, W. Xu, D. Casas, and C. Theobalt. Vnect: Real-time 3d human pose estimation with a single rgb camera. *ACM Transactions on Graphics*, 36(4):44, 2017. 3

[43] D. Metaxas and D. Terzopoulos. Shape and nonrigid motion estimation through physics-based synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(6):580–591, 1993. 3

[44] R. A. Newcombe, D. Fox, and S. M. Seitz. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 343–352, 2015. 1, 2

[45] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohi, J. Shotton, S. Hodges, and A. Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *IEEE International Symposium on Mixed and Augmented Reality*, pages 127–136. IEEE, 2011. 2

[46] R. A. Newcombe, S. J. Lovegrove, and A. J. Davison. Dtam: Dense tracking and mapping in real-time. In *IEEE International Conf. on Computer Vision*, pages 2320–2327, 2011. 1

[47] S. Orts-Escolano, C. Rhemann, S. Fanello, W. Chang, A. Kowdle, Y. Degtyarev, D. Kim, P. L. Davidson, S. Khamis, M. Dou, et al. Holoportation: Virtual 3d teleportation in real-time. In *Symposium on User Interface Software and Technology*, pages 741–754. ACM, 2016. 1, 2, 3

[48] R. Plankers and P. Fua. Articulated soft objects for video-based body modeling. In *IEEE International Conf. on Computer Vision*, number CVLAB-CONF-2001-005, pages 394–401, 2001. 3

[49] G. Pons-Moll, D. J. Fleet, and B. Rosenhahn. Posebits for monocular human pose estimation. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 2345–2352, Columbus, Ohio, USA, 2014. 4

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

[51] G. Pons-Moll, J. Romero, N. Mahmood, and M. J. Black. Dyna: a model of dynamic human shape in motion. *ACM Transactions on Graphics*, 34:120, 2015. 3

[52] G. Pons-Moll and B. Rosenhahn. *Model-Based Pose Estimation*, chapter 9, pages 139–170. Springer, 2011. 3

[53] A.-I. Popa, M. Zanfir, and C. Sminchisescu. Deep multitask architecture for integrated 2d and 3d human sensing. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2017. 3

[54] H. Rhodin, N. Robertini, D. Casas, C. Richardt, H.-P. Seidel, and C. Theobalt. General automatic human shape and motion capture using volumetric contour cues. In *European Conf. on Computer Vision*, pages 509–526. Springer, 2016. 3, 4

[55] N. Robertini, D. Casas, H. Rhodin, H.-P. Seidel, and C. Theobalt. Model-based outdoor performance capture. In *International Conf. on 3D Vision*, 2016. 3

[56] G. Rogez, P. Weinzaepfel, and C. Schmid. Lcr-net: Localization-classification-regression for human pose. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2017. 3

[57] L. Rogge, F. Klose, M. Stengel, M. Eisemann, and M. Magnor. Garment replacement in monocular video sequences. *ACM Transactions on Graphics*, 34(1):6, 2014. 3

[58] A. Shapiro, A. Feng, R. Wang, H. Li, M. Bolas, G. Medioni, and E. Suma. Rapid avatar capture and simulation using commodity depth sensors. *Computer Animation and Virtual Worlds*, 25(3-4):201–211, 2014. 2

[59] L. Sigal, S. Bhatia, S. Roth, M. J. Black, and M. Isard. Tracking loose-limbed people. In *IEEE Conf. on Computer Vision and Pattern Recognition*, volume 1, pages I–421. IEEE, 2004. 3

[60] M. Slavcheva, M. Baust, D. Cremers, and S. Ilic. Killingfusion: Non-rigid 3d reconstruction without correspondences. In *IEEE Conf. on Computer Vision and Pattern Recognition*, volume 3, page 7, 2017. 2

[61] C. Sminchisescu and B. Triggs. Kinematic jump processes for monocular 3d human tracking. In *IEEE Conf. on Computer Vision and Pattern Recognition*, volume 1, pages I–I. IEEE, 2003. 4

[62] O. Sorkine, D. Cohen-Or, Y. Lipman, M. Alexa, C. Rössl, and H.-P. Seidel. Laplacian surface editing. In *Eurographics/ACM SIGGRAPH symposium on Geometry processing*, pages 175–184. ACM, 2004. 5

[63] J. Starck and A. Hilton. Surface capture for performance-based animation. *IEEE Computer Graphics and Applications*, 27(3), 2007. 2

[64] C. Stoll, N. Hasler, J. Gall, H.-P. Seidel, and C. Theobalt. Fast articulated motion tracking using a sums of gaussians body model. In *IEEE International Conf. on Computer Vision*, pages 951–958. IEEE, 2011. 3

[65] X. Sun, J. Shang, S. Liang, and Y. Wei. Compositional human pose regression. In *IEEE International Conf. on Computer Vision*, volume 2, 2017. 3

[66] Y. Tao, Z. Zheng, K. Guo, J. Zhao, D. Quionhai, H. Li, G. Pons-Moll, and Y. Liu. Doublefusion: Real-time capture of human performance with inner body shape from a depth sensor. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2018. 3

[67] C. J. Taylor. Reconstruction of articulated objects from point correspondences in a single uncalibrated image. In *IEEE Conf. on Computer Vision and Pattern Recognition*, volume 1, pages 677–684. IEEE, 2000. 4

[68] D. Tome, C. Russell, and L. Agapito. Lifting from the deep: Convolutional 3d pose estimation from a single image. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2017. 3

[69] 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. 3

[70] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2016. 3, 4

[71] A. Weiss, D. Hirshberg, and M. J. Black. Home 3d body scans from noisy image and range data. In *IEEE International Conf. on Computer Vision*, pages 1951–1958. IEEE, 2011. 1, 3

[72] C. Wu, B. Wilburn, Y. Matsushita, and C. Theobalt. High-quality shape from multi-view stereo and shading under general illumination. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 969–976, 2011. 7

[73] S. Wuhrer, L. Pishchulin, A. Brunton, C. Shu, and J. Lang. Estimation of human body shape and posture under clothing. *Computer Vision and Image Understanding*, 127:31–42, 2014. 3

[74] W. Xu, A. Chatterjee, M. Zollhoefer, H. Rhodin, D. Mehta, H.-P. Seidel, and C. Theobalt. Monoperfcap: Human performance capture from monocular video. In *ACM Transactions on Graphics*, 2018. 2

[75] J. Yang, J.-S. Franco, F. Hétroy-Wheeler, and S. Wuhrer. Estimation of Human Body Shape in Motion with Wide Clothing. In *European Conf. on Computer Vision*, Amsterdam, Netherlands, 2016. 3

[76] M. Ye and R. Yang. Real-time simultaneous pose and shape estimation for articulated objects using a single depth camera. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 2345–2352, 2014. 3

[77] R. Yu, C. Russell, N. D. F. Campbell, and L. Agapito. Direct, dense, and deformable: Template-based non-rigid 3d reconstruction from rgb video. In *IEEE International Conf. on Computer Vision*, 2015. 3

[78] T. Yu, K. Guo, F. Xu, Y. Dong, Z. Su, J. Zhao, J. Li, Q. Dai, and Y. Liu. Bodyfusion: Real-time capture of human motion and surface geometry using a single depth camera. In *IEEE International Conf. on Computer Vision*, pages 910–919, 2017. 3

[79] M. Zeng, J. Zheng, X. Cheng, and X. Liu. Templateless quasi-rigid shape modeling with implicit loop-closure. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 145–152, 2013. 2

[80] C. Zhang, S. Pujades, M. Black, and G. Pons-Moll. Detailed, accurate, human shape estimation from clothed 3D scan sequences. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2017. 3, 6

[81] Q. Zhang, B. Fu, M. Ye, and R. Yang. Quality dynamic human body modeling using a single low-cost depth camera. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 676–683. IEEE, 2014. 1, 3

[82] Q.-Y. Zhou and V. Koltun. Color map optimization for 3d reconstruction with consumer depth cameras. *ACM Transactions on Graphics*, 33(4):155, 2014. 2

[83] S. Zhou, H. Fu, L. Liu, D. Cohen-Or, and X. Han. Parametric reshaping of human bodies in images. In *ACM Transactions on Graphics*, volume 29, page 126. ACM, 2010. 1, 3

[84] X. Zhou, Q. Huang, X. Sun, X. Xue, and Y. Wei. Towards 3d human pose estimation in the wild: A weakly-supervised approach. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 398–407, 2017. 3

[85] H. Zhu, Y. Liu, J. Fan, Q. Dai, and X. Cao. Video-based outdoor human reconstruction. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(4):760–770, 2017. 3

[86] 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, 3

[87] S. Zuffi and M. J. Black. The stitched puppet: A graphical model of 3d human shape and pose. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 3537–3546. IEEE, 2015. 3