LiveCap: Real-Time Human Performance Capture From Monocular Video

MARC HABERMANN and WEIPENG XU, Max Planck Institute for Informatics
MICHAEL ZOLLHÖFER, Stanford University
GERARD PONS-MOLL and CHRISTIAN THEOBALT, Max Planck Institute for Informatics

Fig. 1. We propose the first real-time human performance capture approach that reconstructs dense, space-time coherent deforming geometry of people in their loose everyday clothing from just a single monocular RGB stream, e.g., captured by a webcam.

We present the first real-time human performance capture approach that reconstructs dense, space-time coherent deforming geometry of entire humans in general everyday clothing from just a single RGB video. We propose a novel two-stage analysis-by-synthesis optimization whose formulation and implementation are designed for high performance. In the first stage, a skinned template model is jointly fitted to background subtracted input video, 2D and 3D skeleton joint positions found using a deep neural network, and a set of sparse facial landmark detections. In the second stage, dense non-rigid 3D deformations of skin and even loose apparel are captured based on a novel real-time capable algorithm for non-rigid tracking using dense photometric and silhouette constraints. Our novel energy formulation leverages automatically identified material regions on the template to model the differing non-rigid deformation behavior of skin and apparel. The two resulting non-linear optimization problems per frame are solved with specially tailored data-parallel Gauss-Newton solvers. To achieve real-time performance of over 25Hz, we design a pipelined parallel architecture using the CPU and two commodity GPUs. Our method is the first real-time monocular approach for full-body performance capture. Our method yields comparable accuracy with off-line performance capture techniques while being orders of magnitude faster.

CCS Concepts: • **Computing methodologies → Computer graphics; Motion capture;**

Additional Key Words and Phrases: Monocular performance capture, 3D pose estimation, human body, non-rigid surface deformation

ACM Reference format:
Marc Habermann, Weipeng Xu, Michael Zollhöfer, Gerard Pons-Moll, and Christian Theobalt. 2019. LiveCap: Real-Time Human Performance Capture From Monocular Video. *ACM Trans. Graph.* 38, 2, Article 14 (March 2019), 17 pages.
https://doi.org/10.1145/3311970

Gerard Pons-Moll is funded by the Deutsche Forschungsgemeinschaft (DFG. German Research Foundation) - PO 2379/1-1.
This work was funded by the ERC Consolidator Grant 4DRepLy (770784).
Authors' addresses: M. Habermann, W. Xu, G. Pons-Moll, and C. Theobalt, Campus E1 4, Stuhlsatzenhausweg, 66123 Saarbrücken, Germany; emails: {mhaberma, wxu, gpons, theobalt}@mpi-inf.mpg.de; M. Zollhoefer, 353 Serra Mall, RM 386, Stanford, CA 943, USA; email: zollhoefer@cs.stanford.edu.

0730-0301/2019/03-ART14 $15.00
https://doi.org/10.1145/3311970

1 INTRODUCTION

Dynamic models of virtual human actors are key elements of modern visual effects for movies and games, and they are invaluable for believable, immersive virtual and augmented reality, telepresence, as well as 3D and free-viewpoint video. Such virtual human characters ideally feature high-quality, space-time coherent dense models of shape, motion, and deformation, as well as appearance of people, irrespective of physique or clothing style. Creating such models at high fidelity often requires many months of work of talented artists. To simplify the process, marker-less performance capture methods were researched to reconstruct at least parts of such models from camera recordings of real humans in motion.

Existing multi-camera methods capture human models at very good quality but often need dense arrays of video or depth cameras and controlled studios, struggle with complex deformations, and need pre-captured templates. Only few multi-view methods achieve real-time performance, but no real-time method for single RGB performance capture exists. Many applications in interactive VR and AR, gaming, virtual try-on (Hilsmann and Eisert 2009; Pons-Moll et al. 2017; Sekine et al. 2014), pre-visualization for visual effects, 3DTV, or telepresence (Orts-Escolano et al. 2016) critically depend on real-time performance capture. The use of complex camera arrays and studios restricted to indoor scenes presents a practical barrier to these applications. In daily use, systems should ideally require only one camera and work outdoors.

Under these requirements, performance capture becomes a much harder and much more underconstrained problem. Some methods have approached this challenge by using multiple (Collet et al. 2015; Dou et al. 2016; Wang et al. 2016) or a single low-cost consumer-grade depth (RGB-D) (Newcombe et al. 2015; Yu et al. 2017) camera for dense non-rigid deformation tracking. Although these methods are a significant step forward, RGB-D cameras are not as cheap and ubiquitous as color cameras, often have a limited capture range, do not work well under bright sunlight, and have limited resolution. Real-time human performance capture with a single color camera would therefore greatly enhance and simplify performance capture and further democratize its use, particularly in the aforementioned interactive applications of ever increasing importance. However, dense real-time reconstruction from one color view is even harder, and so today's best monocular methods only capture very coarse models, such as bone skeletons (Mehta et al. 2017; Sun et al. 2017).

In this article, we propose the first—to the best of our knowledge–real-time human performance capture method that reconstructs dense, space-time coherent deforming geometry of people in their loose everyday clothing from a single video camera. In a pre-processing step, the method builds a rigged surface and appearance template from a short video of the person in a static pose, on which regions of skin and pieces of apparel are automatically identified using a new multi-view segmentation that leverages deep learning. The template is fitted to the video sequence in a new coarse-to-fine two-stage optimization, whose problem formulation and implementation are rigorously designed for best accuracy at real-time performance. In its first stage, our new real-time skeleton pose optimizer fits the skinned template to (1) 2D and 3D skeleton joint positions found with a CNN, to (2) sparse detected facial landmarks, and (3) to the foreground silhouette.

In a second stage, dense non-rigid 3D deformations of even loose apparel is captured. To this end, we propose a novel real-time capable algorithm for non-rigid analysis-by-synthesis tracking from monocular RGB data. It minimizes a template-to-image alignment energy jointly considering distance-field–based silhouette alignment, dense photometric alignment, and spatial and temporal regularizers, all designed for real-time performance. The energy formulation leverages the shape template segmentation labels (obtained in the pre-processing stage) to account for the varying non-rigid deformation behavior of different clothing during reconstruction. The non-linear optimization problems in both stages are solved with specially tailored GPU-accelerated Gauss-Newton solvers. To achieve real-time performance of over 25Hz, we design a pipelined solver architecture that executes the first and the second stage on two GPUs in a rolling manner. Our approach captures high-quality models of humans and their clothing in real-time from a single monocular camera. We demonstrate intriguing examples of live applications in 3D video and virtual try-on. We show qualitatively and quantitatively that our method outperforms related monocular on-line methods and comes close to off-line performance capture approaches in terms of reconstruction density and accuracy.

In summary, our contributions are the following. First, we propose the first real-time system for monocular human performance capture. To achieve real-time performance, we not only made specific algorithmic design choices but also contributed several new algorithmic ideas, e.g., the adaptive material–based regularization and the displacement warping to guarantee high quality results under a tight real-time constraint. Second, we also show how to efficiently implement these design decisions by combining the compute power of two GPUs and the CPU in a pipelined architecture and how dense and sparse linear systems of equations can be efficiently optimized on the GPU. Third, to evaluate our approach on a wide range of data, we show high-quality results on an extensive new dataset of more than 20 minutes of video footage captured in 11 scenarios, which contain different types of loose apparel and challenging motions.

2 RELATED WORK

Performance capture methods typically use multi-view images or depth sensors. We focus here on approaches to capture 3D humans in motion and leave out the body of work on 2D pose and shape capture. Most monocular-based methods ignore clothing and are restricted to capturing the articulated motion and the undressed shape of the person. Considering that there are almost no works that do performance capture from monocular video, we focus here on multi-view and depth-based methods and approaches that capture pose and undressed shape from single images.

Multi-view. Many multi-view methods use stereo and shape from silhouette cues to capture the moving actor (Collet et al. 2015; Matusik et al. 2000; Starck and Hilton 2007; Waschbüsch et al. 2005) or reconstruct via multi-view photometric stereo (Vlasic et al. 2009). Provided with sufficient images, some methods directly non-rigidly deform a subject specific template mesh (Cagniart et al. 2010; Carranza et al. 2003; De Aguiar et al. 2008) or a volumetric shape representation (Allain et al. 2015; Huang et al. 2016). Such methods are free-form and can potentially capture arbitrary shapes (Mustafa et al. 2016), as they do not incorporate any skeletal constraints. Such flexibility comes at the cost of robustness. To mitigate this, some methods incorporate a skeleton in the template to constrain the motion to be nearly articulated (Gall et al. 2009; Liu et al. 2011; Vlasic et al. 2008). This also enables off-line performance capture from a stereo pair of cameras (Wu et al. 2013). Some systems combine reconstruction and segmentation to improve results (Bray et al. 2006; Brox et al. 2010; Liu et al. 2011; Wu et al. 2012). Such methods typically require a high-resolution scan of the person as input. To sidestep scanning, a parametric body model can be employed. Early models were based on simple geometric primitives (Metaxas and Terzopoulos 1993; Plänkers and Fua 2001; Sigal et al. 2004; Sminchisescu and Triggs 2003). Recent ones are more accurate, are more detailed, and are learned from thousands of scans (Anguelov et al. 2005; Hasler et al. 2010; Kadlecek et al. 2016; Kim et al. 2017; Loper et al. 2015; Park and Hodgins 2008; Pons-Moll et al. 2015). Capture approaches that use a statistical body model typically ignore clothing or treat it as noise (Balan et al. 2007) or explicitly estimate the shape under the apparel (Bălan and Black 2008; Yang et al. 2016; Zhang et al. 2017). The off-line human performance capture approach of Huang et al. (2017) fits the SMPL body model to 2D joint detections and silhouettes in multi-view data. Some of the recent off-line multi-view approaches jointly track facial expressions (Joo et al. 2018) and hands (Joo et al. 2018; Romero et al. 2017). Even these approaches do not reconstruct dynamic hair. To capture the geometry of the actor beyond

the body shape an option is to non-rigidly deform the base model to fit a scan (Zhang et al. 2017) or a set of images (Rhodin et al. 2016). The approach of Pons-Moll et al. (2017) can jointly capture body shape and clothing using separate meshes; very realistic results are achieved with this method, but it requires an expensive multi-view active stereo setup. All the aforementioned approaches require multi-view setups and are not practical for consumer use. Furthermore, none of the methods runs at real-time frame rates.

Depth based. With the availability of affordable depth camera sensors such as the Kinect, a large number of depth-based methods emerged. Recent approaches that are based on a single depth camera, such as KinectFusion, enable the reconstruction of 3D rigid scenes (Izadi et al. 2011; Newcombe et al. 2011) and also appearance models (Zhou and Koltun 2014) by incrementally fusing geometry in a canonical frame. The approach proposed in Newcombe et al. (2015) generalized KinectFusion to capture dynamic non-rigid scenes. The approach alternates non-rigid registration of the incoming depth frames with updates to the incomplete template, which is constructed incrementally. Such template-free methods (Guo et al. 2017; Innmann et al. 2016; Newcombe et al. 2011; Slavcheva et al. 2017) are flexible but are limited to capturing slow and careful motions. One way to make fusion and tracking more robust is by using a combination of a high frame rate/low-resolution and a low frame rate/high-resolution depth sensor (Guo et al. 2018), improved hardware and software components (Kowdle et al. 2018), multiple Kinects or similar depth sensors (Dou et al. 2016, 2017; Orts-Escolano et al. 2016; Ye et al. 2012; Zhang et al. 2014b), or multi-view data (Collet et al. 2015; Leroy et al. 2017; Prada et al. 2017) and registering new frames to a neighboring key frame; such methods achieve impressive reconstructions but do not register all frames to the same canonical template and require complicated capture setups. Another way to constrain the capture is to pre-scan the object or person to be tracked (De Aguiar et al. 2008; Ye et al. 2012; Zollhöfer et al. 2014), reducing the problem to tracking the non-rigid deformations. Constraining the motion to be articulated is also shown to increase robustness (Yu et al. 2017, 2018). Some works use simple human shape or statistical body models (Bogo et al. 2015; Helten et al. 2013; Wei et al. 2012; Weiss et al. 2011; Ye and Yang 2014; Zhang et al. 2014a), some of which exploit the temporal information to infer shape. Typically, a single shape and multiple poses are optimized to exploit the temporal information. Such approaches are limited to capture naked human shape or at best very tight clothing. Depth sensors are affordable and more practical than multi-view setups. Unfortunately, they have a high power consumption, do not work well under general illumination, and most of the media content is still in the format of 2D images and video. Furthermore, depth-based methods do not directly generalize to work with monocular video.

Monocular 3D pose and shape estimation. Most methods to infer that 3D human motion from monocular images are based on convolutional neural networks (CNNs) and leverage 2D joint detections and predict 3D joint pose in the form of stick figures, e.g., Popa et al. (2017); Rogez et al. (2017); Sun et al. (2017); Tome et al. (2017); Zhou et al. (2017). Tekin et al. (2016) directly predict the 3D body pose from a rectified spatio-temporal volume of input frames. The approach of Tekin et al. (2017) learns to optimally fuse 2D and 3D image cues. These approaches do not capture the dense deforming shape. We also leverage a recent CNN-based 3D pose estimation method (Mehta et al. 2017), but we only employ it to regularize the skeletal motion estimation. Some works fit a (statistical) body surface model to images using substantial manual interaction (Guan et al. 2009; Jain et al. 2010; Rogge et al. 2014; Zhou et al. 2010) typically for the task of image manipulation. Shape and clothing is recovered in Chen et al. (2013) and Guo et al. (2012), but the user needs to click points in the image, select the clothing types from a database, and dynamics are not captured. Instead of clicked points, Kraevoy et al. (2009) propose to obtain the shape from contour drawings. With the advance of 2D joint detections, the works of Bogo et al. (2016), Kanazawa et al. (2018), and Lassner et al. (2017) fit a 3D body model (Loper et al. 2015) to them; since only model parameters are optimized, the results are constrained to the shape space. More recent work (Varol et al. 2018) directly regresses a coarse volumetric body shape. Correspondences from pixels of an input image to surface points on the SMPL body model can also be directly regressed (Güler et al. 2018). Capturing 3D non-rigid deformations from monocular video is very hard. In the domain of non-rigid structure from motion, model-free methods using rigidity and temporal smoothness priors can capture coarse 3D models of simple motions and medium-scale deformations (Garg et al. 2013; Russell et al. 2014). Some methods (Bartoli et al. 2015; Salzmann and Fua 2011; Yu et al. 2015) can non-rigidly track simple shapes and motions by off-line template fitting, but they were not shown to handle highly articulated fast body motions, including clothing, as we do. Specifically for faces, monocular performance capture methods were presented, e.g., in Cao et al. (2015) and Garrido et al. (2016). However, monocular full-body capture faces additional challenges due to more frequent (self-)occlusions and much more complex and diverse clothing and appearance. To the best of our knowledge, the only approach that has shown 3D performance capture of the human body including the non-rigid deformation of clothing from monocular video is the approach of Xu et al. (2018). Its space-time formulation can resolve difficult self-occluded poses at the expense of temporally oversmoothing the actual motion. But at more than 1 minute of runtime per frame, it is impractical for many applications, such as virtual try-on, gaming, or virtual teleportation. It is also challenged by starkly non-rigidly moving clothing. Reducing the processing time without compromising accuracy introduces challenges in formulation and implementation of model-based performance capture, which we address in this work. We present, for the first time, a real-time full-body performance capture system that just requires a monocular video as input. We show that it comes close in accuracy to the best off-line monocular and even multi-view methods while being orders of magnitude faster.

3 METHOD

The input to our method is a single color video stream. In addition, our approach requires a textured actor model, which we acquire in a pre-processing step (Section 3.1) from a monocular video sequence. From this input alone, our real-time human performance capture approach automatically estimates the articulated actor motion and the non-rigid deformation of skin and clothing coarse-to-fine in two subsequent stages per input frame. In the

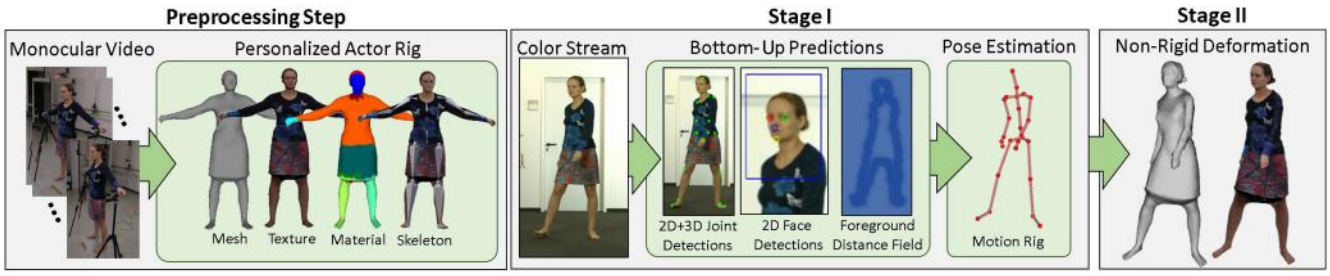

Fig. 2. Our real-time performance capture approach reconstructs dense, space-time coherent deforming geometry of people in loose everyday clothing from just a single RGB stream. A skinned template is jointly fit to background subtracted input video, 2D and 3D joint estimates, and sparse facial detections. Non-rigid 3D deformations of skin and even loose apparel are captured based on a novel real-time capable dense surface tracker.

first stage, we estimate the articulated 3D pose of the underlying kinematic skeleton. To this end, we propose an efficient way to fit the skeletal pose of the skinned template to 2D and 3D joint positions from a state-of-the-art CNN-based regressor to sparse detected face landmarks and to the foreground silhouette (Section 3.3). With this skeleton-deformed mesh and the warped non-rigid displacement of the previous frame as initialization, the second stage captures the surface deformation of the actor using a novel real-time template-to-image non-rigid registration approach (Section 3.4). We express non-rigid registration as an optimization problem consisting of a silhouette alignment term, a photometric term, and several regularization terms; the formulation and combination of terms in the energy is geared toward high efficiency at high accuracy despite the monocular ambiguities. The different components of our approach are illustrated in Figure 2. To achieve real-time performance, we tackle the underlying optimization problems based on dedicated data-parallel GPU optimizers (Section 4). In the following, we explain all components.

3.1 Actor Model Acquisition

Similar to many existing template-based performance capture methods, e.g., Allain et al. (2015), Cagniart et al. (2010), Gall et al. (2009), Vlasic et al. (2008), and Xu et al. (2018), we reconstruct an actor model in a pre-processing step. To this end, we take a set of M images $\mathcal{I}_{\text{rec}} = \{I_{\text{rec}_1}, \ldots, I_{\text{rec}_M}\}$ of the actor in a static neutral pose from a video captured while walking around the person, which covers the entire body. For all of our templates, we use around $M = 70$ images. With these images, we generate a triangulated template mesh $\hat{\mathbf{V}} \in \mathbb{R}^{N\times 3}$ (N denotes the number of vertices in the mesh) and the associated texture map of the actor using an image-based 3D reconstruction software.[1] We downsample the reconstructed geometry to a resolution of approximately $N = 5{,}000$ by using the Quadric Edge Collapse Decimation algorithm implemented in MeshLab.[2] The vertex colors of the template mesh $\mathbf{C} \in \mathbb{R}^{N\times 3}$ are transferred from the generated texture map. Then, skeleton joints and facial markers are manually placed on the template mesh resulting in a skeleton model. The template mesh is rigged to this skeleton model via dual quaternion skinning (Kavan et al. 2007), where the skinning weights are automatically computed using Blender[3] (other auto-rigging tools would be feasible). This allows us to deform the template mesh using the estimated skeletal pose parameters (Section 3.3). An important feature of our performance capture method is that we model material-dependent differences in deformation behavior, e.g., of skin and apparel during tracking (Section 3.4). To this end, we propose a new multi-view method to segment the template into one of seven non-rigidity classes. We first apply the state-of-the-art human parsing method of Gong et al. (2017) to each image in $\mathcal{I}_{\text{rec}}$ separately to obtain the corresponding semantic label images $\mathcal{L}_{\text{rec}} = \{L_{\text{rec}_1}, \ldots, L_{\text{rec}_M}\}$. The semantic labels $L \in \{1, \ldots, 20\}^N$ for all vertices $\mathbf{V}_i$ are computed based on their back-projection into all label images and a majority vote per vertex. The materials are binned into seven non-rigidity classes, each one having a different per-edge non-rigidity weight in the employed regularization term (Section 3.4). Those weights were empirically determined by visual observation of the deformation behavior under different weighting factors. The different classes and the corresponding non-rigidity weights are shown in Table 1. We use a very high weight for rigid body parts, e.g., the head, medium weights for the less rigid body parts, e.g., skin and tight clothing, and a low weight for loose clothing. We use a high rigidity weight for any kind of hairstyle, as we do not, similar to all other human performance capture approaches, consider and track hair dynamics. We map the per-vertex smoothness weights to per-edge non-rigidity weights $s_{i,j}$ by averaging the weights of vertex $\mathbf{V}_i$ and $\mathbf{V}_j$.

Table 1. The Employed Non-Rigidity Weights $s_{i,j}$

Class ID	Weight	Part/Apparel Type
1	1.0	Dress, coat, jumpsuit, skirt, background
2	2.0	Upper clothes
3	2.5	Pants
4	3.0	Scarf
5	50.0	Left leg, right leg, left arm, right arm, socks
6	100.0	Hat, glove, left shoe, right shoe,
7	200.0	Hair, face, sunglasses

3.2 Input Stream Processing

After the actor model acquisition step, our real-time performance capture approach works fully automatically and we do not rely on a careful initialization; e.g., it is sufficient to place the T-posed character model in the center of the frame. The input to our

[1] http://www.agisoft.com.
[2] http://www.meshlab.net.
[3] https://www.blender.org.

algorithm is a single color video stream from a static camera, e.g., a webcam. Thus, we assume the camera and world space to be the same. We calibrate the camera intrinsics using the Matlab calibration toolbox.[4] Our skeletal pose estimation and non-rigid registration stages rely on the silhouette segmentation of the input video frames. To this end, we leverage the background subtraction method of Zivkovic and van der Heijden (2006). We assume that the background is static, that its color is sufficiently different from the foreground, and a few frames of the empty scene are recorded before performance capture commences. We efficiently compute distance transform images I_{DT} from the foreground silhouettes, which are used in the skeletal pose estimation and non-rigid alignment step.

3.3 Skeletal Pose Estimation

We formulate skeletal pose estimation as a non-linear optimization problem in the unknown skeleton parameters $\mathcal{S}^*$:

$$\mathcal{S}^* = \mathrm{argmin}_{\mathcal{S}} E_{\mathrm{pose}}(\mathcal{S}). \tag{1}$$

The set $\mathcal{S} = \{\theta, \mathbf{R}, \mathbf{t}\}$ contains the joint angles $\theta \in \mathbb{R}^{27}$ of the J joints of the skeletal model, and the global pose $\mathbf{R} \in \mathbf{SO}(3)$ and translation $\mathbf{t} \in \mathbb{R}^3$ of the root. For pose estimation, we optimize an energy of the following general form:

$$\begin{aligned} E_{\mathrm{pose}}(\mathcal{S}) = {} & E_{\mathrm{2D}}(\mathcal{S}) + E_{\mathrm{3D}}(\mathcal{S}) + E_{\mathrm{silhouette}}(\mathcal{S}) + E_{\mathrm{temporal}}(\mathcal{S}) \\ & + E_{\mathrm{anatomic}}(\mathcal{S}). \end{aligned} \tag{2}$$

Here, E_{2D} and E_{3D} are alignment constraints based on regressed 2D and 3D joint positions, respectively. In addition, $E_{\mathrm{silhouette}}$ is a dense alignment term that fits the silhouette of the actor model to the detected silhouette in the input color images. At last, E_{temporal} and E_{anatomic} are temporal and anatomical regularization constraints that ensure that the speed of the motion and the joint angles stay in physically plausible ranges. To better handle fast motion, we initialize the skeleton parameters before optimization by extrapolating the poses of the last two frames in joint angle space based on an explicit Euler step. In the following, we explain each energy term in more detail.

Sparse 2D and 3D alignment constraint. For each input frame I, we estimate the 2D and 3D joint positions $\mathbf{P}_{\mathrm{2D},i} \in \mathbb{R}^2$ and $\mathbf{P}_{\mathrm{3D},i} \in \mathbb{R}^3$ of the J joints using the efficient deep skeleton joint regression network of the VNect algorithm (Mehta et al. 2017) trained with the original data of Mehta et al. (2017). However, with these skeleton-only joint detections, it is not possible to determine the orientation of the head. Therefore, we further augment the 2D joint predictions of Mehta et al. (2017) with a subset of the facial landmark detections of Saragih et al. (2009), which includes the eyes, nose, and chin. We incorporate the 2D detections $\mathbf{P}_{\mathrm{2D},i} \in \mathbb{R}^2$ based on the following re-projection constraint:

$$E_{\mathrm{2D}}(\mathcal{S}) = \lambda_{\mathrm{2D}} \sum_{i=1}^{J+4} \lambda_i \left\| \pi\left(p_{\mathbf{3D},i}(\theta, \mathbf{R}, \mathbf{t})\right) - \mathbf{P}_{\mathrm{2D},i} \right\|^2. \tag{3}$$

Here, $p_{\mathbf{3D},i}$ is the 3D position of the i-th joint/face marker of the used kinematic skeleton and $\pi : \mathbb{R}^3 \rightarrow \mathbb{R}^2$ is a full perspective projection that maps 3D space to the 2D image plane. Thus, this term enforces that all projected joint positions are close to their corresponding detections. λ_i are detection-based weights. We use $\lambda_i = 0.326$ for the facial landmarks and $\lambda_i = 1.0$ for all other detections to avoid that the head error dominates all other body parts. To resolve the inherent depth ambiguities of the re-projection constraint, we also employ the following 3D-to-3D alignment term between model joints $p_{\mathbf{3D},i}(\theta, \mathbf{R}, \mathbf{t})$ and 3D detections $\mathbf{P}_{\mathrm{3D},i}$:

$$E_{\mathrm{3D}}(\mathcal{S}) = \lambda_{\mathrm{3D}} \sum_{i=1}^{J} \left\| p_{\mathbf{3D},i}(\theta, \mathbf{R}, \mathbf{t}) - (\mathbf{P}_{\mathrm{3D},i} + \mathbf{t}') \right\|^2. \tag{4}$$

Here, $\mathbf{t}' \in \mathbb{R}^3$ is an auxiliary variable that transforms the regressed 3D joint positions $\mathbf{P}_{\mathrm{3D},i}$ from the root-centered local coordinate system to the global coordinate system. Note that the regressed 3D joint positions $\mathbf{P}_{\mathrm{3D},i}$ are in a normalized space. Therefore, we rescale the regressed skeleton according to the bone lengths of our parameterized skeleton model.

Dense silhouette alignment constraint. We enforce a dense alignment between the boundary of the skinned actor model and the detected silhouette in the input image. In contrast to the approach of Xu et al. (2018) that requires closest point computations, we employ a distance transform–based constraint for efficiency reasons. Once per frame, we extract a set of contour vertices $\mathcal{B}$ from the current deformed version of the actor model. Afterward, we enforce that all contour vertices align well to the interface between the detected foreground and background:

$$E_{\mathrm{silhouette}}(\mathcal{S}) = \lambda_{\mathrm{silhouette}} \sum_{i \in \mathcal{B}} b_i \cdot [I_{\mathrm{DT}}(\pi(\mathbf{V}_i(\theta, \mathbf{R}, \mathbf{t})))]^2. \tag{5}$$

Here, $\mathbf{V}_i$ is the i-th boundary vertex of the skinned actor model and the image I_{DT} stores the Euclidean distance transform with respect to the detected silhouette in the input image. The $b_i \in \{-1, +1\}$ are directional weights that guide the optimization to follow the right direction in the distance field. In the minimization of the term in Equation (5), the silhouette model points are pushed in the negative direction of the distance transform image gradient $\mathbf{z} = -\nabla_{xy} I_{DT} \in \mathbb{R}^2$. By definition, $\mathbf{z}$ points in the direction of the nearest *image silhouette* contour. If model points fall outside of the image silhouette, they will be dragged toward the nearest image silhouette contour as desired. When model points fall inside the image silhouette, however, there are two possibilities: (1) the model point normal $\mathbf{n}$ follows roughly the same direction as $\mathbf{z}$ or (2) it does not. In case (1), the normal at the nearest image silhouette point matches the direction of the model point normal. This indicates that $\mathbf{z}$ is a good direction to follow. In case (2), however, the normal at the nearest image silhouette point follows the opposite direction, indicating that $\mathbf{z}$ is pointing toward the wrong image silhouette contour (Figure 3). Therefore, in case (2), we follow the opposite direction $\mathbf{p} = -\mathbf{z}$ by setting $b_i = -1$. This is preferable over just following $\mathbf{n}$, because $\mathbf{n}$ is not necessarily pointing away from the wrong image silhouette contour. Mathematically, we consider that we are in case (2) when $\mathbf{n}^T \mathbf{z} < 0$. For all other cases, we follow the direction of $\mathbf{z}$ by setting $b_i = +1$.

Temporal stabilization. To mitigate temporal noise, we use a temporal stabilization constraint, which penalizes the change in joint

[4] http://www.vision.caltech.edu/bouguetj/calib_doc.

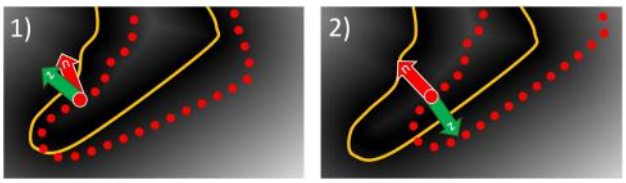

Fig. 3. The two cases in the silhouette alignment constraint. Target silhouette (yellow), model silhouette (red), negative gradient of the distance field $\mathbf{z}$ (green arrow), and the projected 2D normal $\mathbf{n}$ of the boundary vertex (red arrow).

position between the current and previous frame:

$$E_{\text{temporal}}(\mathcal{S}) = \lambda_{\text{temporal}} \sum_{i=1}^{J} \lambda_i \left\| p_{\text{3D},i}(\theta, \mathbf{R}, \mathbf{t}) - p_{\text{3D},i}^{t-1}(\theta, \mathbf{R}, \mathbf{t}) \right\|^2. \tag{6}$$

Here, the λ_i are joint-based temporal smoothness weights. We use $\lambda_i = 2.5$ for joints on the torso and the head, $\lambda_i = 2.0$ for shoulders, $\lambda_i = 1.5$ for the knees and elbows, and $\lambda_i = 1.0$ for the hands and feet.

Joint angle limits. The joints of the human skeleton have physical limits. We integrate this prior knowledge into our pose estimation objective based on a soft constraint on $\theta \in \mathbb{R}^{27}$. To this end, we enforce that all degrees of freedom stay within their anatomical limits $\theta_{\min} \in \mathbb{R}^{27}$ and $\theta_{\max} \in \mathbb{R}^{27}$:

$$E_{\text{anatomic}}(\mathcal{S}) = \lambda_{\text{anatomic}} \sum_{i=1}^{27} \Psi(\theta_i).$$

Here, $\Psi(x)$ is a quadratic barrier function that penalizes if a degree of freedom exceeds its limits:

$$\Psi(x) = \begin{cases} (x - \theta_{\max,i})^2, & \text{if } x > \theta_{\max,i} \\ (\theta_{\min,i} - x)^2, & \text{if } x < \theta_{\min,i} \\ 0, & \text{otherwise.} \end{cases}$$

This term prevents un-plausible human pose estimates.

3.4 Non-Rigid Surface Registration

The pose estimation step cannot capture realistic non-rigid deformations of skin and clothing that are not explained through skinning. The model therefore does not yet align with the image well everywhere, particularly in cloth and some skin regions. Hence, starting from the pose estimation result, we solve the following non-rigid surface tracking energy:

$$E_{\text{non-rigid}}(\mathbf{V}) = E_{\text{data}}(\mathbf{V}) + E_{\text{reg}}(\mathbf{V}). \tag{7}$$

The energy consists of several data terms E_{data} and regularization constraints E_{reg}, which we explain in the following. Our data terms are a combination of a dense photometric alignment term E_{photo} and a dense silhouette alignment term $E_{\text{silhouette}}$:

$$E_{\text{data}}(\mathbf{V}) = E_{\text{photo}}(\mathbf{V}) + E_{\text{silhouette}}(\mathbf{V}). \tag{8}$$

Dense photometric alignment. The photometric alignment term measures the re-projection error densely:

$$E_{\text{photo}}(\mathbf{V}) = \sum_{i \in \mathcal{V}} w_{\text{photo}} \left\| \sigma_c \left(I_{\text{Gauss}} \left(\pi \left(\mathbf{V}_i \right) \right) - \mathbf{C}_i \right) \right\|^2, \tag{9}$$

where $\mathbf{C}_i$ is the color of vertex $\mathbf{V}_i$ in the template model and $\sigma_c(\cdot)$ is a robust kernel that prunes wrong correspondences according to color similarity by setting residuals that are above a certain threshold to zero. More specifically, we project every visible vertex $\mathbf{V}_i \in \mathcal{V}$ to screen space based on the full perspective camera model π. The visibility is obtained based on the skinned mesh after the pose estimation step using depth buffering. To speed up convergence, we compute the photometric term based on a three-level pyramid of the input image. We perform one Gauss-Newton iteration on each level. We use the projected positions to sample a Gaussian blurred version I_{Gauss} of the input image I at the current timestep for more stable and longer range gradients. The Gaussian kernel sizes for the three levels are 15, 9, and 3 respectively.

Dense silhouette alignment. In addition to dense photometric alignment, we also enforce alignment of the projected 3D model boundary with the detected silhouette in the input image:

$$E_{\text{silhouette}}(\mathbf{V}) = w_{\text{silhouette}} \sum_{i \in \mathcal{B}} b_i \cdot [I_{\text{DT}}(\pi(\mathbf{V}_i))]^2. \tag{10}$$

After Stage I, we first update the model boundary $\mathcal{B}$ and consider all vertices $\mathbf{V}_i \in \mathcal{B}$. These boundary vertices are encouraged to match the zero iso-line of the distance transform image I_{DT} and thus be aligned with the detected input silhouette. The b_i are computed similar to the pose optimization step (see Section 3.3). Due to the non-rigid deformation that cannot be recovered by our pose estimation stage, in some cases the projection of the mesh from Stage I has a gap between body parts such as the arms and torso, whereas in the input image the gaps do not exist. To prevent image silhouettes being wrongly explained by multiple model boundaries, we project the posed model $\mathbf{V}^{\text{S}}$ into the current frame and compute a body part mask—derived from the skinning weights. We increase the extent of each body part by a dilation (maximum of 10 pixels, the torso has preference over the other parts) to obtain a conservative region boundary that closes the gaps mentioned previously. If a vertex $\mathbf{V}_i$ moves onto a region with a differing semantic label, we disable its silhouette term by setting $b_i = 0$. This drastically improves the reconstruction quality (Figure 4).

Our high-dimensional monocular non-rigid registration problem with only the data terms is ill posed. Therefore, we use regularization constraints:

$$E_{\text{reg}}(\mathbf{V}) = E_{\text{smooth}}(\mathbf{V}) + E_{\text{edge}}(\mathbf{V}) + E_{\text{velocity}}(\mathbf{V}) + E_{\text{acceleration}}(\mathbf{V}). \tag{11}$$

Here, E_{smooth} and E_{edge} are spatial smoothness priors on the mesh geometry, and E_{velocity} and $E_{\text{acceleration}}$ are temporal priors. In the following, we provide more details.

Spatial smoothness. The first prior on the mesh geometry is a spatial smoothness term with respect to the pose estimation result:

$$E_{\text{smooth}}(\mathbf{V}) = w_{\text{smooth}} \sum_{i=1}^{N} \sum_{j \in \mathcal{N}_i} \frac{s_{ij}}{|\mathcal{N}_i|} \| (\mathbf{V}_i - \mathbf{V}_j) - \left(\mathbf{V}_i^{\text{S}} - \mathbf{V}_j^{\text{S}} \right) \|^2. \tag{12}$$

Here, the $\mathbf{V}_i$ are the unknown optimal vertex positions and the $\mathbf{V}_i^{\text{S}}$ are vertex positions after skinning using the current pose estimation result of Stage I. s_{ij} are the semantic label based per-edge smoothness weights (see Section 3.1) that model material dependent non-rigidity. The energy term enforces that every edge

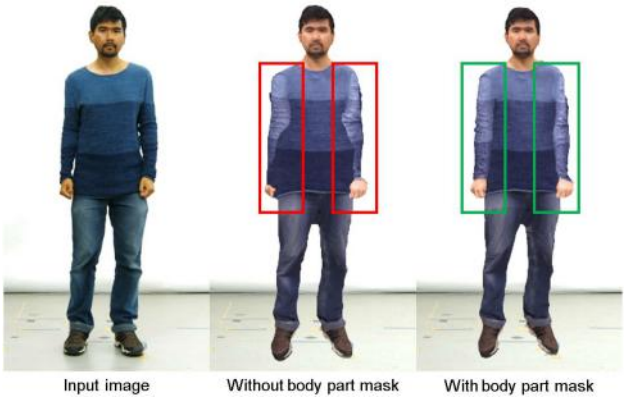

Fig. 4. Left: Input image. Middle: Textured reconstruction without using the body part mask. One can clearly see the artifacts since multiple model boundaries wrongly explain the silhouette of the arms. Right: Using the body part mask in the distance transform image, the foreground silhouette is correctly explained.

in the deformed model is similar to the un-deformed model in terms of its length and orientation. In addition to this surface smoothness term, we also enforce locally isometric deformations:

$$E_{\text{edge}}(\mathbf{V}) = w_{\text{edge}} \sum_{i=1}^{N} \sum_{j \in \mathcal{N}_i} \frac{s_{ij}}{|\mathcal{N}_i|} \left(\left\| \mathbf{V}_i - \mathbf{V}_j \right\| - \left\| \hat{\mathbf{V}}_i - \hat{\mathbf{V}}_j \right\| \right)^2, \quad (13)$$

where $\hat{\mathbf{V}}$ denotes the vertex position in the template's rest pose. We enforce that the edge length does not change much between the rest pose $\hat{\mathbf{V}}_i$ and the optimal unknown pose $\mathbf{V}_i$. Although this is similar to the first term, it enables us to penalize stretching independently of shearing.

Temporal smoothness. We also use temporal priors that favor temporally coherent non-rigid deformations. Similar to temporal smoothness in skeletal pose estimation, the first term,

$$E_{\text{velocity}}(\mathbf{V}) = w_{\text{velocity}} \sum_{i=1}^{N} \left\| \mathbf{V}_i - \mathbf{V}_i^{t-1} \right\|^2, \quad (14)$$

encourages small velocity, and the second term,

$$E_{\text{acceleration}}(\mathbf{V}) = w_{\text{acceleration}} \sum_{i=1}^{N} \left\| \mathbf{V}_i - 2\mathbf{V}_i^{t-1} + \mathbf{V}_i^{t-2} \right\|^2, \quad (15)$$

encourages small acceleration between adjacent frames.

Displacement warping. The non-rigid displacements $\mathbf{d}_i^{t-1} = \mathbf{V}_i^{t-1} - \mathbf{V}_i^{\mathrm{S},t-1} \in \mathbb{R}^3$ that are added to each vertex i after skinning are usually similar from frame $t-1$ to frame t. We warp $\mathbf{d}_i^{t-1}$ back to the rest pose by applying dual quaternion skinning with the inverse rotation quaternions given by the pose at time $t-1$. We refer to them as $\hat{\mathbf{d}}_i^{t-1}$. For the next frame t, we transform $\hat{\mathbf{d}}_i^{t-1}$ according to the pose at time t resulting in a skinned displacement $\mathbf{d}_i^{\mathrm{S},t}$ and initialize the non-rigid stage with $\mathbf{V}_i^t = \mathbf{V}_i^{\mathrm{S},t} + \mathbf{d}_i^{\mathrm{S},t}$. This jump-starts the non-rigid alignment step and leads to improved tracking quality. Similarly, we add $\mathbf{d}_i^{\mathrm{S},t}$ to the skinned actor model for more accurate dense silhouette alignment during the skeletal pose estimation stage.

Vertex snapping. After the non-rigid stage, the boundary vertices are already very close to the image silhouette. Therefore, we can robustly snap them to the closest silhouette point by walking on the distance transform along the negative gradient direction until the zero crossing is reached. Vertex snapping allows us to reduce the number of iteration steps, because if the solution is already close to the optimum, the updates of the solver become smaller, as is true for most optimization problems. Therefore, if the mesh is already close to the silhouette, we "snap" to the silhouette in a single step instead of requiring multiple iterations of Gauss-Newton. To obtain continuous results, non-boundary vertices are smoothly adjusted based on a Laplacian warp in a local neighborhood around the mesh contour.

4 DATA PARALLEL GPU OPTIMIZATION

The described pose estimation and non-rigid registration problems are non-linear optimizations based on an objective E with respect to unknowns $\mathcal{X}$, i.e., the parameters of the kinematic model $\mathcal{S}$ for pose estimation and the vertex positions $\mathbf{V}$ for non-rigid surface deformation. The optimal parameters $\mathcal{X}^*$ are found via energy minimization:

$$\mathcal{X}^* = \arg\min_{\mathcal{X}} E(\mathcal{X}). \quad (16)$$

In both capture stages, i.e., pose estimation (see Section 3.3) and non-rigid surface tracking (see Section 3.4), the objective E can be expressed as a sum of squares:

$$E(\mathcal{X}) = \sum_i \left[\mathbf{F}_i(\mathcal{X}) \right]^2 = \left\| \mathbf{F}(\mathcal{X}) \right\|_2^2. \quad (17)$$

Here, $\mathbf{F}$ is the error vector resulting from stacking all residual terms. We tackle this optimization at real-time rates using a data-parallel iterative Gauss-Newton solver that minimizes the total error by linearizing $\mathbf{F}$ and taking local steps $\mathcal{X}_k = \mathcal{X}_{k-1} + \boldsymbol{\delta}_k^*$ obtained by the solution of a sequence of linear sub-problems (normal equations):

$$\mathbf{J}^T(\mathcal{X}_{\mathrm{k}-1})\mathbf{J}(\mathcal{X}_{\mathrm{k}-1}) \cdot \boldsymbol{\delta}_k^* = -\mathbf{J}^T(\mathcal{X}_{\mathrm{k}-1})\mathbf{F}(\mathcal{X}_{\mathrm{k}-1}). \quad (18)$$

Here, $\mathbf{J}$ is the Jacobian of $\mathbf{F}$. Depending on the problems (pose estimation or non-rigid registration), the linear systems have a quite different structure in terms of dimensionality and sparsity. Thus, we use tailored parallelization strategies for each of the problems. Since we use Gauss-Newton instead of Levenberg-Marquardt, the residual does not have to be computed during the iterations, thus leading to faster runtimes, and consequently more iterations are possible within the tight real-time constraint.

Pose estimation. The normal equations of the pose optimization problem are small but dense, i.e, the corresponding system matrix is small, rectangular, and dense. Handling each non-linear Gauss-Newton step efficiently requires a specifically tailored parallelization and optimization strategy. First, in the beginning of each Gauss-Newton step, we compute the system matrix $\mathbf{J}^T\mathbf{J}$ and right-hand side $-\mathbf{J}^T\mathbf{F}$ in global memory on the GPU. Afterward, we ship the small system of size 36×36($36 = 3 + 3 + 27 + 3$, 3 DoFs for $\mathbf{R}$, 3 for $\mathbf{t}$, 27 for θ, and 3 for $\mathbf{t}'$) to the CPU and solve it based on QR decomposition. The strategy of splitting the computation to CPU

and GPU is similar to Tagliasacchi et al. (2015) in spirit. To compute $\mathbf{J}^T\mathbf{J}$ on the GPU, we first compute $\mathbf{J}$ fully in parallel and store it in device memory based on a kernel that launches one thread per matrix entry. We perform a similar operation for $\mathbf{F}$. $\mathbf{J}^T\mathbf{J}$ is then computed based on a data-parallel version of a matrix-matrix multiplication that exploits shared memory for high performance. The same kernel also directly computes $\mathbf{J}^T\mathbf{F}$. We launch several thread blocks per element of the output matrix/vector, which cooperate in computing the required dot products, e.g., between the i-th and j-th column of $\mathbf{J}$ or the i-th column of $\mathbf{J}$ and $\mathbf{F}$. To this end, each thread block computes a small subpart of the dot product based on a shared memory reduction. The per-block results are summed up based on global memory atomics. In total, we perform six Gauss-Newton steps, which turned out to be a good trade-off between accuracy and speed.

Non-rigid surface registration. The non-rigid optimization problem that results from the energy $E_{\text{non-rigid}}$ has a substantially different structure. It leads to a large sparse system of normal equations, i.e, the corresponding system matrix is sparse and has a low number of non-zeros per row. Similar to Innmann et al. (2016) and Zollhöfer et al. (2014), during GPU-based data-parallel preconditioned conjugate gradient (PCG), we parallelize over the rows (unknowns) of the system matrix $\mathbf{J}^T\mathbf{J}$ using one thread per block row (x-,y-, and z-entry of a vertex). Each thread collects and handles all non-zeros in the corresponding row. We use the diagonal of $\frac{1}{\mathbf{J}^T\mathbf{J}}$ as a pre-conditioner. We perform three Gauss-Newton steps and solve the linear system based on four PCG iterations, which turned out to be a good trade-off between accuracy and speed.

Pipelined implementation. To achieve real-time performance, we use a data-parallel implementation of our entire performance capture algorithm in combination with a pipeline strategy tailored for our problem. To this end, we run our approach in three threads on a PC with two GPUs. Thread 1 uses only the CPU, which is responsible for data pre-processing. Thread 2 computes the CNN-based human pose detection on the first graphics card, and thread 3 solves the pose optimization problem and estimates the non-rigid deformation on the second graphics card. Our distributed computation strategy induces a two-frame delay, but for most applications it is barely noticeable.

5 RESULTS

For all of our tests, we employ an Intel Core i7 with two GeForce GTX 1080Ti graphics cards. Our algorithm runs at around 25fps, which fulfills the performance requirement of many real-time applications. In all of our experiments, we use the same set of parameters that are empirically determined: $\lambda_{\text{2D}} = 460$, $\lambda_{\text{3D}} = 28$, $\lambda_{\text{silhouette}} = 200$, $\lambda_{\text{temporal}} = 1.5$, $\lambda_{\text{anatomic}} = 10^6$, $w_{\text{photo}} = 10{,}000$, $w_{\text{silhouette}} = 600$, $w_{\text{smooth}} = 10.0$, $w_{\text{edge}} = 30.0$, $w_{\text{velocity}} = 0.25$, and $w_{\text{acceleration}} = 0.1$. In the following, we first introduce our new dataset, evaluate our approach on several challenging sequences qualitatively and quantitatively, and compare to related methods. Then, we perform an ablation evaluation to study the importance of the different components of our approach. Finally, we demonstrate several live applications. More results are shown in our two supplementary videos, which in total show more than 20 minutes of performance capture results. We applied smoothing with a filter of window size 3 (stencil: $[0.15, 0.7, 0.15]$) to the trajectories of the vertex coordinates as a post-process for all video results except in the live setup.

5.1 Dataset

To qualitatively evaluate our method on a wide range of settings, we recorded several challenging motion sequences. These contain large variations in non-rigid clothing deformations, e.g., skirts and hooded sweaters, and fast motions like dancing and jumping jacks. In total, we captured more than 20 minutes of video footage split in 11 sequences with different sets of apparel each worn by one of seven subjects. All sequences were recorded with a Blackmagic video camera (30fps, 540×960 resolution). We provide semantically segmented, rigged, and textured templates; calibrated camera parameters; and an empty background image for all sequences. In addition, we provide the silhouettes from background subtraction, our motion estimates, and the non-rigidly deformed meshes. For eight of the sequences, we captured the subject from a reference view, which we will also make available, to evaluate the tracking quality. Figure 5 shows some of the templates and example frames of the captured sequences. All templates are shown in the supplementary video. We will make the full dataset publicly available.

5.2 Qualitative and Quantitative Results

In total, we evaluated our approach on our new dataset and five existing video sequences of people in different sets of apparel. In addition, we test our method with four subjects in a live setup (see Figure 1) with a low-cost webcam. Our method takes frames at 540×960 resolution as input. To better evaluate our non-rigid surface registration method, we used challenging loose clothing in these sequences, including skirts, dresses, hooded sweatshirts, and baggy pants. The sequences show a wide range of difficult motions (slow to fast, self-occlusions) for monocular capture. Additionally, we compare our approach to the state-of-the-art monocular performance capture method of Xu et al. (2018) on two of their sequences and on one of our new captured sequences.

Qualitative evaluation. In Figure 5, we show several frames from live performance capture results. We can see that our results precisely overlay the person in the input images. Note that body pose, head-orientation, and non-rigid deformation of loose clothing are accurately captured. Both the side-by-side comparison to RGB input and the accurate overlay with the reconstructed mesh show the high quality of the reconstruction. Also note that our reconstruction results match the images captured from a laterally displaced reference view, which is not used for tracking (see the supplemental video). This provides further evidence of the fidelity of our 3D performance capture results, also in depth, which shows that our formulation effectively meets the non-trivial underconstrained monocular reconstruction challenge. To evaluate the robustness of our method, we included many fast and challenging motions in our test set. As shown in Figure 6, even the fast 360° rotation (see the first row) and the jumping motion (see the second row) are successfully tracked. This illustrates the robustness of our algorithm and its efficient and effective combined consideration of sparse and dense image cues, as well as both learning- and

Fig. 5. Qualitative results. We show several live monocular performance capture results of entire humans in their loose everyday clothing. (a) The template models. (b) Input images to our method. (c) The corresponding results precisely overlay the person in the input images. Our results can be used to render realistic images (d) or free-viewpoint video (e).

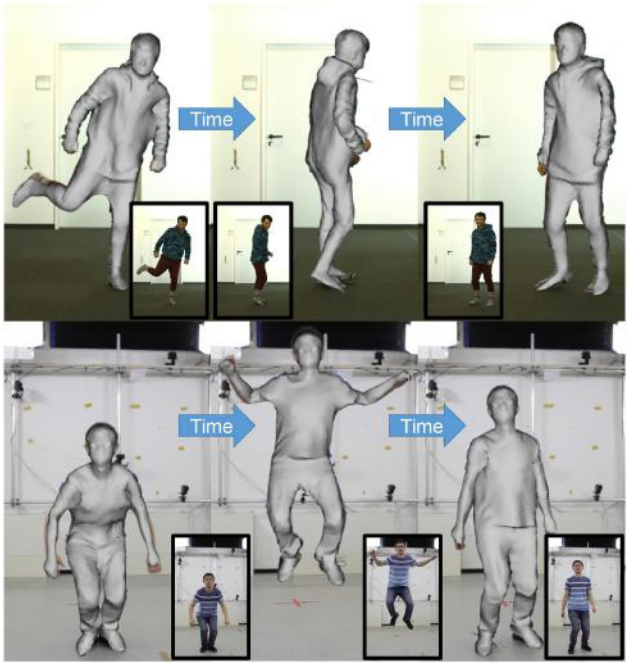

Fig. 6. Our real-time approach even tracks challenging and fast motions, such as jumping and a fast 360° rotation with high accuracy. The reconstructions overlay the input image well. For the complete sequence, we refer to the supplemental video.

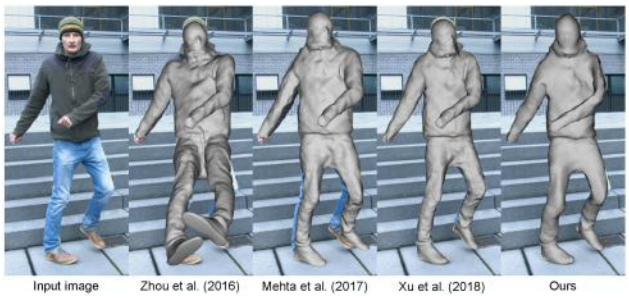

Fig. 7. Qualitative comparison to related monocular methods. The results of our approach overlay much better with the input than the skeleton-only results of Zhou et al. (2016) and Mehta et al. (2017). Our results come close in quality to the off-line approach of Xu et al. (2018).

model-based capture, which in this combination were not used in prior work, let alone in real time.

Comparison to related monocular methods. In Figure 7, we provide a comparison to three related state-of-the-art methods: the fundamentally off-line, monocular dense (surface-based) performance capture method of Xu et al. (2018), called *MonoPerfCap*, and two current monocular methods for 3D skeleton-only reconstruction, the 2D-to-3D lifting method of Zhou et al. (2016), and the real-time VNect algorithm (Mehta et al. 2017). For the latter two, we show the skinned rendering of our template using their skeleton pose. The test sequence is provided by Xu et al. (2018) with manually labeled ground truth silhouettes. Our method's results overlay much better with the input than the skeleton-only

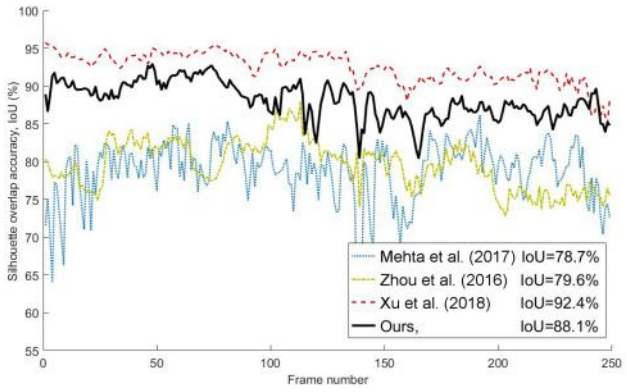

Fig. 8. Quantitative comparison to related monocular methods. In terms of the silhouette overlap accuracy (IoU), our method achieves better results and outperforms Zhou et al. (2016) and Mehta et al. (2017) by 8.5% and 9.4%, respectively. On average, our results are only 4.3% worse than the off-line approach of Xu et al. (2018), but our approach is orders of magnitude faster.

results of Zhou et al. (2016) and Mehta et al. (2017), confirming our much better reconstructions. Additionally, a quantitative comparison on this sequence in terms of the silhouette overlap accuracy (intersection over union [IoU]), presented in Figure 8, shows that our method achieves clearly better results and outperforms Zhou et al. (2016) and Mehta et al. (2017) by 8.5% and 9.4%, respectively. Using the same metric, our IoU is only 4.3% smaller than Xu et al. (2018), which is mainly caused by the fact that their foreground segmentation is more accurate than ours due to their more advanced but off-line foreground segmentation strategy (see Figure 9). However, note that our method is overall orders of magnitude faster than their algorithm, which takes more than 1 minute per frame, and our reconstructions are still robust to the noisy foreground segmentation. To compare against MonoPerfCap more thoroughly, we also compare against them on one of our sequences (see Section 5.1), which shows more challenging non-rigid dress deformations in combination with fast motions (see bottom rows of Figure 10). On this sequence, the accuracy of the foreground estimation is roughly the same, leading to the fact that our approach achieves an IoU of 86.86% (averaged over 500 frames), which is almost identical to the one of Xu et al. (2018) (86.89%). As shown in Figure 10, we achieve comparable reconstruction quality and overlay while being orders of magnitude faster. MonoPerfCap's window-based optimizer achieves slightly better boundary alignment and more stable tracking of some difficult, convolved, self-occluded poses but is much slower. Our reconstruction of head and feet is consistently better than Xu et al. (2018) due to the additional facial landmark alignment term and the better pose detector that we adopted. We provide a qualitative comparison showing highly challenging motions in the supplementary video.

Surface reconstruction accuracy. To evaluate our surface reconstruction error, also relative to multi-view methods, we use the *Pablo* sequence from the state-of-the-art multi-view template-based performance capture method of Robertini et al. (2016) (they also provide the template). As shown in Figure 11, our real-time monocular method comes very close in quality to the results of

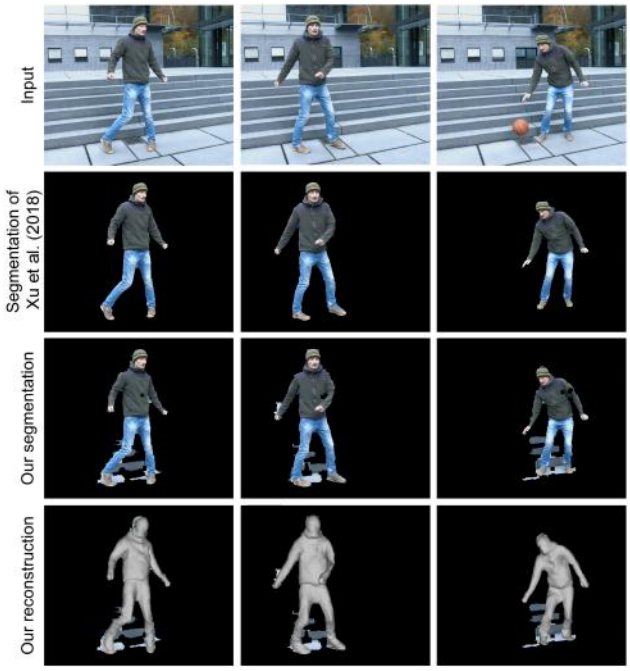

Fig. 9. Comparison of the foreground segmentation of Xu et al. (2018) and our method. Note that our silhouette estimates are less accurate than the ones of Xu et al. (2018). Nevertheless, our reconstruction results are robust to the noisy foreground estimates and look plausible.

the fundamentally off-line multi-view approach of Robertini et al. (2016) and the monocular off-line method of Xu et al. (2018). In addition, it clearly outperforms the monocular non-rigid capture method of Yu et al. (2015) and the rigged skeleton-only results of the 3D pose estimation methods of Zhou et al. (2016) and Mehta et al. (2017) (the latter two as described in the previous paragraph). This is further evidenced by our quantitative evaluation on per-vertex position errors (see Figure 12). We use the reconstruction results of Robertini et al. (2016) as the reference and show the per-vertex Euclidean surface error. Similar to Xu et al. (2018), we aligned the reconstruction of all methods to the reference meshes with a translation to eliminate the global depth offset. The method of Xu et al. (2018) achieves slightly better results in terms of surface reconstruction accuracy. Similar to our previous experiment (see Figure 9), we observed that our foreground estimates are slightly worse than the ones of Xu et al. (2018), which caused the lower accuracy.

Skeletal pose estimation accuracy. We also compare our approach in terms of joint position accuracy on the *Pablo* sequence against VNect (Mehta et al. 2017; Zhou et al. 2016) and MonoPerfCap (Xu et al. 2018). As the reference, we use the joint positions from the multi-view method of Robertini et al. (2016). We report the average per-joint 3D error (in millimeters) after aligning the per-frame poses with a similarity transform. As shown in Figure 13, our method outperforms the three other methods, most notably the skeleton-only methods (Mehta et al. 2017; Zhou et al. 2016). This shows that our combined surface and skeleton reconstruction also benefits 3D pose estimation quality in itself.

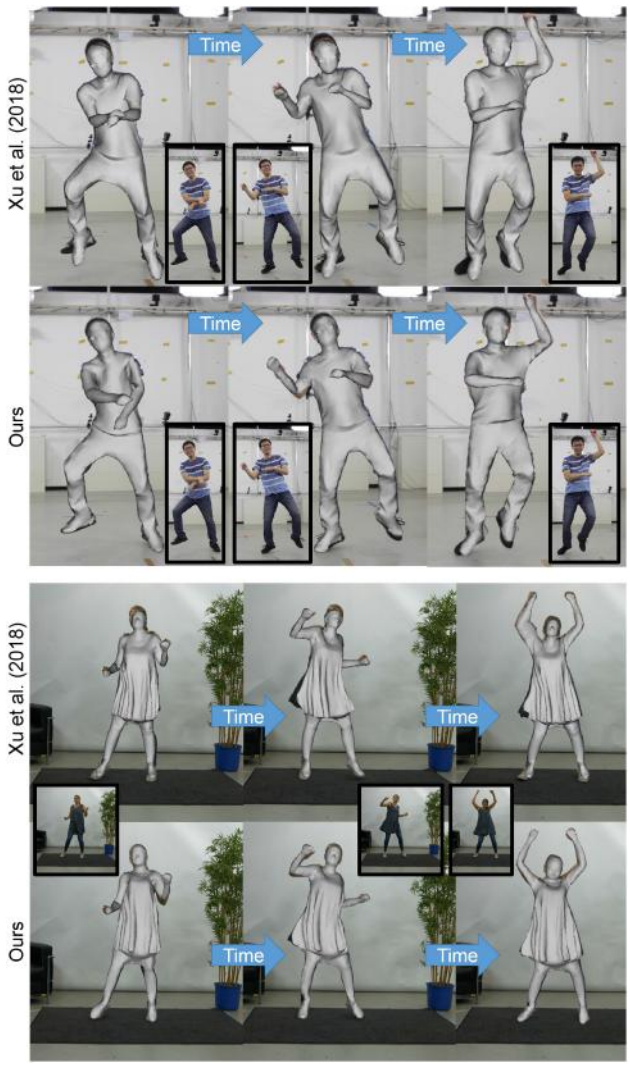

Fig. 10. Qualitative comparison to MonoPerfCap (Xu et al. 2018). We achieve comparable reconstruction quality and overlay while being orders of magnitude faster.

Ablation study. We first qualitatively evaluate the importance of all algorithmic components in an ablation study on a real video sequence. To this end, we compare the results of (1) our pose estimation without facial landmark alignment term and the silhouette term, which we refer to as $E_{2Dw/ofac e} + E_{3D}$; (2) our pose estimation without the silhouette term ($E_{2D} + E_{3D}$); (3) our complete pose estimation (E_{pose}); and (4) our full pipeline ($E_{pose} + E_{non-rigid}$). As shown in Figure 14, (1) the facial landmark alignment term significantly improves the head orientation estimation (red circles); (2) the misalignment of $E_{2D} + E_{3D}$ is corrected by our silhouette term in E_{pose} (yellow circles); and (3) the non-rigid deformation on the surface, which cannot be modeled by skinning, is accurately captured by our non-rigid registration method $E_{non-rigid}$ (blue circles). Second, we also quantitatively evaluated the importance of the terms on a sequence where high-quality reconstructions based on the multi-view performance capture results

Fig. 11. Qualitative comparisons of the surface reconstruction accuracy on the *Pablo* sequence. Our real-time monocular approach comes very close in quality to the results of the fundamentally off-line multi-view approach of Robertini et al. (2016) and the monocular off-line method of Xu et al. (2018). It clearly outperforms the monocular non-rigid capture method of Yu et al. (2015) and the rigged skeleton-only results of the 3D pose estimation methods of Zhou et al. (2016) and Mehta et al. (2017).

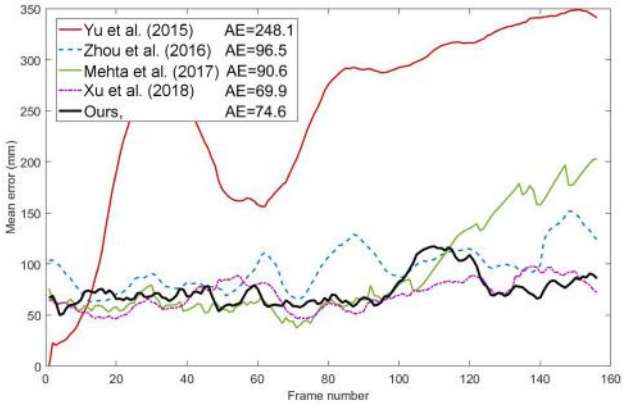

Fig. 12. Quantitative comparison of the surface reconstruction accuracy on the *Pablo* sequence. Our real-time monocular approach comes very close in quality to the results of the monocular off-line method of Xu et al. (2018). It clearly outperforms the monocular non-rigid capture method of Yu et al. (2015) and the rigged skeleton-only results of the 3D pose estimation methods of Zhou et al. (2016) and Mehta et al. (2017).

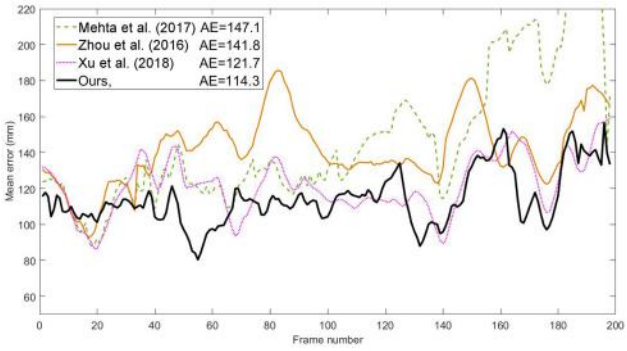

Fig. 13. Comparison of the skeletal pose estimation accuracy in terms of average per-joint 3D error on the *Pablo* sequence. Our method outperforms the three other methods, most notably the skeleton-only methods of Mehta et al. (2017) and Zhou et al. (2016).

of De Aguiar et al. (2008) are used as ground truth. The mean vertex position error shown in Figure 15 clearly demonstrates the consistent improvement by each of the algorithmic components of our approach. The non-rigid alignment stage obtains on average better results than the pose-only alignment. Since non-rigid deformations are most of the time concentrated in certain areas, e.g., a skirt, and at certain frames when articulated motion takes place, we also measure the per-frame and per-vertex improvement of the proposed non-rigid stage. To this end, we measure the improvement of $(E_{pose} + E_{non-rigid})$ over (E_{pose}) by computing the per-vertex error of the pose only results minus the per-vertex error of our method. Consequently, positive means that our method is better than the pose-only deformation. As demonstrated in Figure 16, the non-rigid stage significantly improves the reconstruction of the skirt and the arm. The improvement is especially noticeable for frames where the deformation of the skirt significantly differs from the static template model, as such motion cannot be handled by the pose-only step. On the same dataset, we also evaluated the influence of (1) the warping of the non-rigid displacement of the previous frame, (2) the proposed body part masks used in the dense silhouette alignment, and (3) the proposed vertex snapping. Those algorithmic changes respectively lead to 2.4%, 1.7%, and 1.7% improvement in average 3D vertex error, which sums up to a total improvement of 5.8%. The importance of our material-based non-rigid deformation adaptation strategy is shown in Figure 17. With constantly low non-rigidity weights ($s_{i,j} = 2.0$) in all regions, the deformation of the skirt is well reconstructed, but the head is severely distorted (left). In contrast, with high global non-rigidity weights ($s_{i,j} = 50.0$), the head shape is preserved, but the skirt cannot be tracked reliably (middle). Our new semantic weight

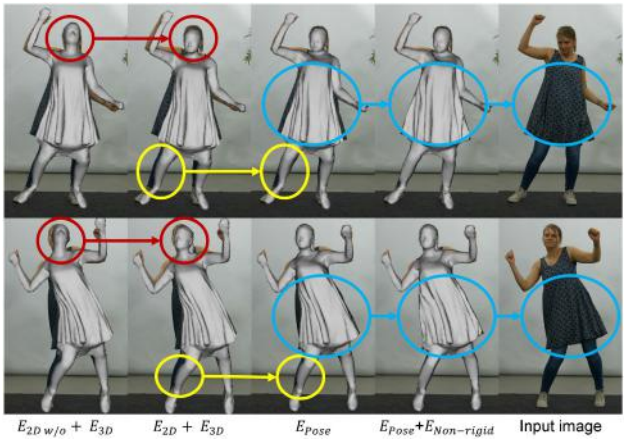

Fig. 14. Ablation study. First, the facial landmark alignment term significantly improves the head orientation estimation (red circles). Second, the misalignment of $E_{2D} + E_{3D}$ is corrected by our silhouette term in E_{pose} (yellow circles). Third, the non-rigid deformation on the surface, which cannot be modeled by skinning, is accurately captured by our non-rigid registration method $E_{non-rigid}$ (blue circles).

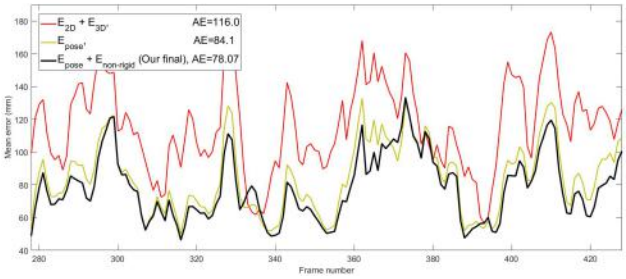

Fig. 15. Ablation study. The mean vertex position error clearly demonstrates the consistent improvement by each of the algorithmic components of our approach. Our full approach consistently obtains the lowest error.

adaptation strategy enables the reconstruction of both regions with high accuracy and leads to the best results (right).

5.3 Applications

Our monocular real-time human performance capture method can facilitate many applications that depend on real-time capture: interactive VR and AR, human-computer interaction, previsualization for visual effects, 3D video, or telepresence. We exemplify this through two application demonstrators. In Figure 18, we show that our method allows live free-viewpoint video rendering and computer animation of the performance captured result from just single color input. This illustrates the potential of our method in several of the aforementioned live application domains. In Figure 19, we demonstrate a real-time virtual try-on application based on our performance capture method. We replace the texture corresponding to the trousers on the template and visualize the tracked result in real time. With such a system, the users can see themselves in clothing variants in real time with live feedback, which could potentially be used in VR or even AR online shopping.

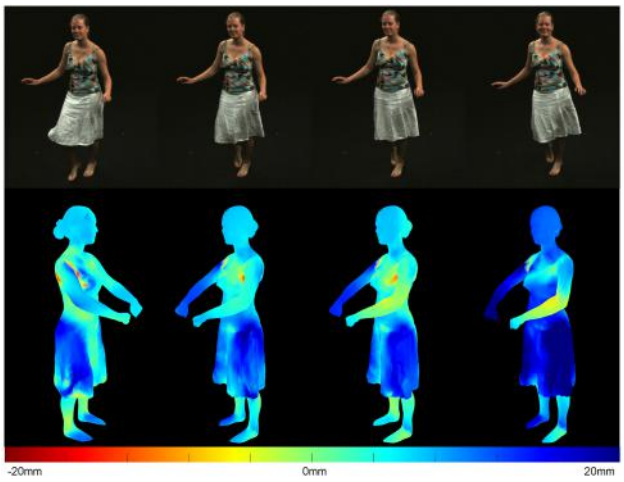

Fig. 16. Improvement of the non-rigid stage ($E_{pose} + E_{non-rigid}$) over pose-only deformation (E_{pose}). Top row: Four monocular input images. Bottom row: For each image, we show the per-vertex error of the pose only results minus the per-vertex error of our method. Consequently, negative means that only the pose is better and is colored in red. Positive means that our method is better and is colored in blue. As expected, our method achieves the most improvement on the non-rigid skirt part, which is around 20mm for the shown frames.

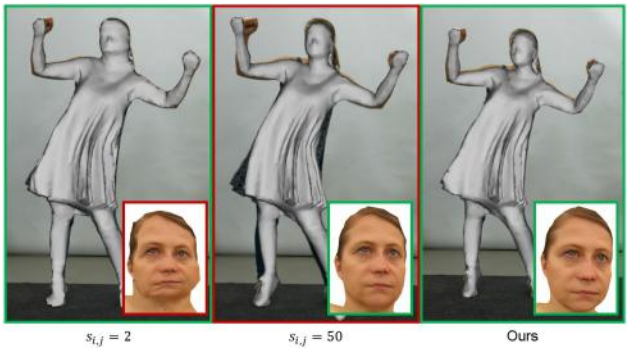

Fig. 17. Importance of our material-based non-rigid deformation adaptation strategy. With a low global regularization weight, the deformation of the skirt is well reconstructed but the head is distorted (left). A high deformation weight preserves the shape of the head but prevents tracking of the skirt motion (middle). Our new semantic weight adaptation strategy enables the reconstruction of both regions with high accuracy and leads to the best results (right).

Fig. 18. Free-viewpoint video rendering results using our approach.

Fig. 19. Live virtual try-on application based on our approach.

6 DISCUSSION AND LIMITATIONS

We have demonstrated compelling real-time full-body human performance capture results using a single consumer-grade color camera. Our formulation combines constraints used, individually, in different previous image-based reconstruction methods. But the specific combination we employ embedded in a hierarchical real-time approach is new and enables, for the first time, real-time monocular performance capture. Further, our formulation geared rigorously for real-time use differs from the related but off-line MonoPerfCap (Xu et al. 2018) method in several ways. In Stage I, the facial landmarks, as well as the displacement warping, which is also added during pose tracking, improve the pose accuracy of our real-time method. Further, we track the pose per frame instead of a batch-based formulation, which reduces the computation time and allows faster motions. Further improvement in terms of efficiency are achieved by our GPU-based pose solver. In Stage II, our dense photometric term that adds constraints for non-boundary vertices and our adaptive material–based regularization improve reconstruction quality. Our non-rigid fitting stage is faster due to the more efficient combination of spatial regularizers that requires a much smaller number of variables than the as-rigid-as-possible regularizer. We directly solve for the vertex displacements instead of estimating the embedded graph rotations/translations. We found that this formulation is better suited for a parallel implementation on the GPU and also gives a more flexible representation. Due to our real-time constraint, we make use of an efficient distance transform–based representation, instead of the ICP-based approach that requires expensive search of correspondences between the model boundary and the image silhouettes. Our experiments show that our method achieves a similar reconstruction quality compared to the off-line performance capture approach of Xu et al. (2018) while being orders of magnitude faster.

Nonetheless, our approach is subject to some limitations (Figure 20). Due to the ambiguities that come along with monocular performance capture, we rely on an accurate template acquisition because reconstruction errors and mislabeled part segmentations in the template itself cannot be recovered during tracking. Further, we cannot handle topological changes that are too far from the template, e.g., removing of some clothes and deformations along the camera viewing axis can only be partially recovered by our photometric term. The latter point could be addressed by an additional term that involves shading and illumination estimation. As is common for learning methods, the underlying 3D joint regression deep network fails for extreme poses not seen in training. Our model fitting can often, but not always, correct such wrong estimates, which produces glitches in the tracking results. However, our performance capture approach robustly recovers from such situations (see the top part of Figure 20). Considering that our method uses foreground/background segmentation, strong shadows and shading effects, objects with similar color to the performer, and changing illumination situations can cause suboptimal segmentation, thus leading to noisy data association in the silhouette alignment term, which manifests itself as high-frequency jitter. Our approach is robust to some degree of mis-classifications but can get confused by big segmentation outliers. This could be alleviated in the future by incorporating more sophisticated background segmentation strategies, e.g., based on deep neural networks. Strong changes in shading or shadows, specular materials, or non-diffuse lighting can also negatively impact the color alignment term. A joint optimization for scene illumination and material properties could alleviate this problem. Even though we carefully orchestrated the components of our method to achieve high accuracy and temporal stability in this challenging monocular setting, even under non-trivial occlusions, extensive (self-)occlusion are still fundamentally difficult. Our estimates for occluded parts will be less accurate than with multi-view methods due to the lack of image evidence. Although pose and silhouette plausibly constrain the back side of the body, fully occluded limbs may have incorrect poses. Additional learned motion priors could further resolve such ambiguous situations. Fortunately, our approach recovers as soon as the difficult occlusions are gone (see the bottom part of Figure 20).

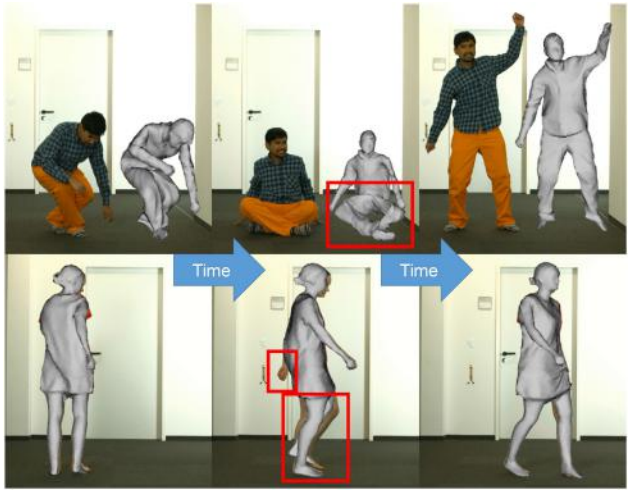

Fig. 20. Failure cases. Top row: The underlying 3D joint regression deep network can fail for extreme poses not seen in training, which can produce glitches in the tracking results. Our model fitting can often, but not always, correct such wrong estimates. However, our performance capture approach robustly recovers from such situations. Bottom row: Our estimates for occluded parts will be less accurate than with multi-view methods due to the lack of image evidence. Although pose and silhouette plausibly constrain the back side of the body, fully occluded limbs may have incorrect poses.

7 CONCLUSION

We have presented the first monocular real-time human performance capture approach that reconstructs dense, space-time

coherent deforming geometry of entire humans in their loose everyday clothing. Our novel energy formulation leverages automatically identified material regions on the template to differentiate between different non-rigid deformation behaviors of skin and various types of apparel. We tackle the underlying non-linear optimization problems at real time based on a pipelined implementation that runs two specially tailored data-parallel Gauss-Newton solvers, one for pose estimation and one for non-rigid tracking, at the same time. We deem our approach as a first step toward general real-time capture of humans from just a single view, which is an invaluable tool for believable, immersive virtual and augmented reality, telepresence, virtual try-on, and many more exciting applications that the future will bring to our homes. An interesting direction for future work is the joint estimation of human motion, facial expression, hand pose, and hair dynamics from a single monocular camera.

REFERENCES

Benjamin Allain, Jean-Sébastien Franco, and Edmond Boyer. 2015. An efficient volumetric framework for shape tracking. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'15)*. IEEE, Los Alamitos, CA, 268–276. DOI: https://doi.org/10.1109/CVPR.2015.7298623

Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. 2005. SCAPE: Shape completion and animation of people. *ACM Transactions on Graphics* 24, 3 (2005), 408–416.

Alexandru O. Bălan and Michael J. Black. 2008. The naked truth: Estimating body shape under clothing. In *Proceedings of the European Conference on Computer Vision*. 15–29.

Alexandru O. Balan, Leonid Sigal, Michael J. Black, James E. Davis, and Horst W. Haussecker. 2007. Detailed human shape and pose from images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'07)*. 1–8.

A. Bartoli, Y. Gérard, F. Chadebecq, T. Collins, and D. Pizarro. 2015. Shape-from-template. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37, 10 (Oct. 2015), 2099–2118. DOI: https://doi.org/10.1109/TPAMI.2015.2392759

Federica Bogo, Michael J. Black, Matthew Loper, and Javier Romero. 2015. Detailed full-body reconstructions of moving people from monocular RGB-D sequences. In *Proceedings of the International Conference on Computer Vision (ICCV'15)*. 2300–2308.

Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. 2016. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *Proceedings of the European Conference on Computer Vision (ECCV'16)*.

Matthieu Bray, Pushmeet Kohli, and Philip H. S. Torr. 2006. Posecut: Simultaneous segmentation and 3D pose estimation of humans using dynamic graph-cuts. In *Proceedings of the European Conference on Computer Vision*. 642–655.

Thomas Brox, Bodo Rosenhahn, Juergen Gall, and Daniel Cremers. 2010. Combined region and motion-based 3D tracking of rigid and articulated objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32, 3 (2010), 402–415.

Cedric Cagniart, Edmond Boyer, and Slobodan Ilic. 2010. Free-form mesh tracking: A patch-based approach. In *Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'10)*. IEEE, Los Alamitos, CA, 1339–1346.

Chen Cao, Derek Bradley, Kun Zhou, and Thabo Beeler. 2015. Real-time high-fidelity facial performance capture. *ACM Transactions on Graphics* 34, 4 (July 2015), Article 46, 9 pages.

Joel Carranza, Christian Theobalt, Marcus A. Magnor, and Hans-Peter Seidel. 2003. Free-viewpoint video of human actors. *ACM Transactions on Graphics* 22, 3 (July 2003), 569–577.

Xiaowu Chen, Yu Guo, Bin Zhou, and Qinping Zhao. 2013. Deformable model for estimating clothed and naked human shapes from a single image. *Visual Computer* 29, 11 (2013), 1187–1196.

Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese, Hugues Hoppe, et al. 2015. High-quality streamable free-viewpoint video. *ACM Transactions on Graphics* 34, 4 (2015), 69.

Edilson De Aguiar, Carsten Stoll, Christian Theobalt, Naveed Ahmed, Hans-Peter Seidel, and Sebastian Thrun. 2008. Performance capture from sparse multi-view video. *ACM Transactions on Graphics* 27, 3 (Aug. 2008), Article 98.

Mingsong Dou, Philip Davidson, Sean Ryan Fanello, Sameh Khamis, Adarsh Kowdle, Christoph Rhemann, Vladimir Tankovich, et al. 2017. Motion2Fusion: Real-time volumetric performance capture. *ACM Transactions on Graphics* 36, 6 (Nov. 2017), Article 246, 16 pages.

Mingsong Dou, Sameh Khamis, Yury Degtyarev, Philip Davidson, Sean Ryan Fanello, Adarsh Kowdle, Sergio Orts Escolano, et al. 2016. Fusion4D: Real-time performance capture of challenging scenes. *ACM Transactions on Graphics* 35, 4 (2016), 114.

Juergen Gall, Carsten Stoll, Edilson De Aguiar, Christian Theobalt, Bodo Rosenhahn, and Hans-Peter Seidel. 2009. Motion capture using joint skeleton tracking and surface estimation. In *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'09)*. IEEE, Los Alamitos, CA, 1746–1753.

R. Garg, A. Roussos, and L. Agapito. 2013. Dense variational reconstruction of non-rigid surfaces from monocular video. In *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition*. 1272–1279. DOI: https://doi.org/10.1109/CVPR.2013.168

Pablo Garrido, Michael Zollhoefer, Dan Casas, Levi Valgaerts, Kiran Varanasi, Patrick Perez, and Christian Theobalt. 2016. Reconstruction of personalized 3D face rigs from monocular video. *ACM Transactions on Graphics* 35, 3 (2016), Article 28, 15 pages.

Ke Gong, Xiaodan Liang, Dongyu Zhang, Xiaohui Shen, and Liang Lin. 2017. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'17)*.

Peng Guan, Alexander Weiss, Alexandru O Bălan, and Michael J. Black. 2009. Estimating human shape and pose from a single image. In *Proceedings of the 2009 IEEE 12th International Conference on Computer Vision (ICCV 09)*. 1381–1388.

Riza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. 2018. DensePose: Dense human pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'18)*.

Kaiwen Guo, Jonathan Taylor, Sean Fanello, Andrea Tagliasacchi, Mingsong Dou, Philip Davidson, Adarsh Kowdle, et al. 2018. TwinFusion: High framerate non-rigid fusion through fast correspondence tracking. In *Proceedings of the 2018 International Conference on 3D Vision (3DV'18)*. DOI: https://doi.org/10.1109/3DV.2018.00074

Kaiwen Guo, Feng Xu, Tao Yu, Xiaoyang Liu, Qionghai Dai, and Yebin Liu. 2017. Real-time geometry, albedo, and motion reconstruction using a single RGB-D camera. *ACM Transactions on Graphics* 36, 3 (2017), 32.

Yu Guo, Xiaowu Chen, Bin Zhou, and Qinping Zhao. 2012. Clothed and naked human shapes estimation from a single image. In *Proceedings of the 1st International Conference on Computational Visual Media (CVM'12)*. 43–50.

Nils Hasler, Hanno Ackermann, Bodo Rosenhahn, Thorsten Thormählen, and Hans-Peter Seidel. 2010. Multilinear pose and body shape estimation of dressed subjects from image sets. In *Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'10)*. IEEE, Los Alamitos, CA, 1823–1830.

Thomas Helten, Meinard Muller, Hans-Peter Seidel, and Christian Theobalt. 2013. Real-time body tracking with one depth camera and inertial sensors. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV'13)*.

Anna Hilsmann and Peter Eisert. 2009. Tracking and retexturing cloth for real-time virtual clothing applications. In *Proceedings of the 4th International Conference on Computer Vision/Computer Graphics CollaborationTechniques (MIRAGE'09)*. 94–105. DOI: https://doi.org/10.1007/978-3-642-01811-4_9

C.-H. Huang, B. Allain, J.-S. Franco, N. Navab, S. Ilic, and E. Boyer. 2016. Volumetric 3D tracking by detection. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'16)*.

Yinghao Huang, Federica Bogo, Christoph Lassner, Angjoo Kanazawa, Peter V. Gehler, Javier Romero, Ijaz Akhter, et al. 2017. Towards accurate marker-less human shape and pose estimation over time. In *Proceedings of the 2017 International Conference on 3D Vision (3DV'17)*.

Matthias Innmann, Michael Zollhöfer, Matthias Nießner, Christian Theobalt, and Marc Stamminger. 2016. VolumeDeform: Real-time volumetric non-rigid reconstruction. In *Computer Vision—ECCV 2016*. Springer, 17.

Shahram Izadi, David Kim, Otmar Hilliges, David Molyneaux, Richard Newcombe, Pushmeet Kohli, Jamie Shotton, et al. 2011. KinectFusion: Real-time 3D reconstruction and interaction using a moving depth camera. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology (UIST'11)*. ACM, New York, NY, 559–568.

Arjun Jain, Thorsten Thormählen, Hans-Peter Seidel, and Christian Theobalt. 2010. MovieReshape: Tracking and reshaping of humans in videos. *ACM Transactions on Graphics* 29, 6 (2010), Article 148. DOI: https://doi.org/10.1145/1866158.1866174

Hanbyul Joo, Tomas Simon, and Yaser Sheikh. 2018. Total capture: A 3D deformation model for tracking faces, hands, and bodies. arXiv:1801.01615.

Petr Kadlecek, Alexandru-Eugen Ichim, Tiantian Liu, Jaroslav Krivanek, and Ladislav Kavan. 2016. Reconstructing personalized anatomical models for physics-based body animation. *ACM Transactions on Graphics* 35, 6 (Nov. 2016), Article 213.

Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. 2018. End-to-end recovery of human shape and pose. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'18)*.

Ladislav Kavan, Steven Collins, Jiří Žára, and Carol O'Sullivan. 2007. Skinning with dual quaternions. In *Proceedings of the 2007 Symposium on Interactive 3D Graphics and Games*. ACM, New York, NY, 39–46.

Meekyoung Kim, Gerard Pons-Moll, Sergi Pujades, Sungbae Bang, Jinwwok Kim, Michael Black, and Sung-Hee Lee. 2017. Data-driven physics for human soft tissue animation. *ACM Transactions on Graphics* 36, 4 (July 2017), Article 54. http://dx.doi.org/10.1145/3072959.3073685

Adarsh Kowdle, Christoph Rhemann, Sean Fanello, Andrea Tagliasacchi, Jonathan Taylor, Philip Davidson, Mingsong Dou, et al. 2018. The need 4 speed in real-time dense visual tracking. *ACM Transactions on Graphics* 37, 6 (Nov. 2018), Article 220. DOI: https://doi.org/10.1145/3272127.3275062

Vladislav Kraevoy, Alla Sheffer, and Michiel van de Panne. 2009. Modeling from contour drawings. In *Proceedings of the 6th Eurographics Symposium on Sketch-Based Interfaces and Modeling*. ACM, New York, NY, 37–44.

Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J. Black, and Peter V. Gehler. 2017. Unite the people: Closing the loop between 3D and 2D human representations. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'17)*.

Vincent Leroy, Jean-Sébastien Franco, and Edmond Boyer. 2017. Multi-view dynamic shape refinement using local temporal integration. In *Proceedings of the IEEE International Conference on Computer Vision*. https://hal.archives-ouvertes.fr/hal-01567758

Yebin Liu, Carsten Stoll, Juergen Gall, Hans-Peter Seidel, and Christian Theobalt. 2011. Markerless motion capture of interacting characters using multi-view image segmentation. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'11)*. IEEE, Los Alamitos, CA, 1249–1256.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. 2015. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics* 34, 6 (Nov. 2015), Article 248.

Wojciech Matusik, Chris Buehler, Ramesh Raskar, Steven J. Gortler, and Leonard McMillan. 2000. Image-based visual hulls. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*. ACM, New York, NY, 369–374.

Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, et al. 2017. VNect: Real-time 3D human pose estimation with a single RGB camera. *ACM Transactions on Graphics* 36, 4 (2017), 14. DOI: https://doi.org/10.1145/3072959.3073596

Dimitris Metaxas and Demetri Terzopoulos. 1993. Shape and nonrigid motion estimation through physics-based synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 15, 6 (June 1993), 580–591.

Armin Mustafa, Hansung Kim, Jean-Yves Guillemaut, and Adrian Hilton. 2016. Temporally coherent 4D reconstruction of complex dynamic scenes. In *Proceedings of the 29th IEEE Conference on Computer Vision and Pattern Recognition (CVPR'16)*. 4660–4669. DOI: https://doi.org/10.1109/CVPR.2016.504

Richard A. Newcombe, Dieter Fox, and Steven M. Seitz. 2015. DynamicFusion: Reconstruction and tracking of non-rigid scenes in real-time. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'15)*.

Richard A. Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J. Davison, Pushmeet Kohi, et al. 2011. KinectFusion: Real-time dense surface mapping and tracking. In *Proceedings of the 2011 10th International Symposium on Mixed and Augmented Reality (ISMAR'11)*. IEEE, Los Alamitos, CA, 127–136.

Sergio Orts-Escolano, Christoph Rhemann, Sean Fanello, Wayne Chang, Adarsh Kowdle, Yury Degtyarev, David Kim, et al. 2016. Holoportation: Virtual 3D teleportation in real-time. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. ACM, New York, NY, 741–754.

Sang Il Park and Jessica K. Hodgins. 2008. Data-driven modeling of skin and muscle deformation. *ACM Transactions on Graphics* 27, 3 (Aug. 2008), Article 96.

Ralf Plänkers and Pascal Fua. 2001. Tracking and modeling people in video sequences. *Computer Vision and Image Understanding* 81, 3 (2001), 285–302.

Gerard Pons-Moll, Sergi Pujades, Sonny Hu, and Michael Black. 2017. ClothCap: Seamless 4D clothing capture and retargeting. *ACM Transactions on Graphics* 36, 4 (July 2017), Article 73. http://dx.doi.org/10.1145/3072959.3073711

Gerard Pons-Moll, Javier Romero, Naureen Mahmood, and Michael J. Black. 2015. Dyna: A model of dynamic human shape in motion. *ACM Transactions on Graphics* 34, 4 (2015), 120.

Alin-Ionut Popa, Mihai Zanfir, and Cristian Sminchisescu. 2017. Deep multitask architecture for integrated 2D and 3D human sensing. In *Proceedings of the 2017IEEE Conference on Computer Vision and Pattern Recognition (CVPR'17)*.

Fabián Prada, Misha Kazhdan, Ming Chuang, Alvaro Collet, and Hugues Hoppe. 2017. Spatiotemporal atlas parameterization for evolving meshes. *ACM Transactions on Graphics* 36, 4 (2017), 58.

Helge Rhodin, Nadia Robertini, Dan Casas, Christian Richardt, Hans-Peter Seidel, and Christian Theobalt. 2016. General automatic human shape and motion capture using volumetric contour cues. In *Computer Vision—ECCV 2016*. Lecture Notes in Computer Science, Vol. 9909. Springer, 509–526.

Nadia Robertini, Dan Casas, Helge Rhodin, Hans-Peter Seidel, and Christian Theobalt. 2016. Model-based outdoor performance capture. In *Proceedings of the 2016 4th International Conference on 3D Vision (3DV'16)*.

Gregory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. 2017. LCR-Net: Localization-classification-regression for human pose. In *Proceedings of the 2017 IIEEE Conference on Computer Vision and Pattern Recognition (CVPR'17)*.

Lorenz Rogge, Felix Klose, Michael Stengel, Martin Eisemann, and Marcus Magnor. 2014. Garment replacement in monocular video sequences. *ACM Transactions on Graphics* 34, 1 (2014), 6.

Javier Romero, Dimitrios Tzionas, and Michael J. Black. 2017. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics* 36, 6 (Nov. 2017), Article 245, 17 pages. http://doi.acm.org/10.1145/3130800.3130883

Chris Russell, Rui Yu, and Lourdes Agapito. 2014. *Video Pop-Up: Monocular 3D Reconstruction of Dynamic Scenes*. Springer, Cham, Switzerland, 583–598. DOI: https://doi.org/10.1007/978-3-319-10584-0_38

Mathieu Salzmann and Pascal Fua. 2011. Linear local models for monocular reconstruction of deformable surfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33, 5 (2011), 931–944. DOI: https://doi.org/10.1109/TPAMI.2010.158

J. M. Saragih, S. Lucey, and J. F. Cohn. 2009. Face alignment through subspace constrained mean-shifts. In *Proceedings of the 2009 IEEE 12th International Conference on Computer Vision*. 1034–1041. DOI: https://doi.org/10.1109/ICCV.2009.5459377

M. Sekine, K. Sugita, F. Perbet, B. Stenger, and M. Nishiyama. 2014. Virtual fitting by single-shot body shape estimation. In *Proceedings of the International Conference on 3D Body Scanning Technologies*. 406–413.

Leonid Sigal, Sidharth Bhatia, Stefan Roth, Michael J. Black, and Michael Isard. 2004. Tracking loose-limbed people. In *Proceedings of the 2004 IEEE Computer Scoeity Conference on Computer Vision and Pattern Recognition (CVPR'04)*, Vol. 1. IEEE, Los Alamitos, CA.

Miroslava Slavcheva, Maximilian Baust, Daniel Cremers, and Slobodan Ilic. 2017. KillingFusion: Non-rigid 3D reconstruction without correspondences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'17)*, Vol. 3. 7.

Cristian Sminchisescu and Bill Triggs. 2003. Kinematic jump processes for monocular 3D human tracking. In *Proceedings. of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol. 1. IEEE, Los Alamitos, CA, I–69.

Jonathan Starck and Adrian Hilton. 2007. Surface capture for performance-based animation. *IEEE Computer Graphics and Applications* 27, 3 (2007), 21–31.

Xiao Sun, Jiaxiang Shang, Shuang Liang, and Yichen Wei. 2017. Compositional human pose regression. In *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV'17)*.

Andrea Tagliasacchi, Matthias Schroeder, Anastasia Tkach, Sofien Bouaziz, Mario Botsch, and Mark Pauly. 2015. Robust articulated-ICP for real-time hand tracking. *Computer Graphics Forum* 34, 5 (2015), Article 5.

Bugra Tekin, Pablo Márquez-Neila, Mathieu Salzmann, and Pascal Fua. 2017. Learning to fuse 2D and 3D image cues for monocular body pose estimation. In *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV'17)*. IEEE, Los Alamitos, CA, 3961–3970.

B. Tekin, A. Rozantsev, V. Lepetit, and P. Fua. 2016. Direct prediction of 3D body poses from motion compensated sequences. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'16)*. 991–1000.

Denis Tome, Chris Russell, and Lourdes Agapito. 2017. Lifting from the deep: Convolutional 3D pose estimation from a single image. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition*.

Gül Varol, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid. 2018. BodyNet: Volumetric inference of 3D human body shapes. In *Proceedings of the 2018 15th European Conference on Computer Vision (ECCV'18)*.

Daniel Vlasic, Ilya Baran, Wojciech Matusik, and Jovan Popović. 2008. Articulated mesh animation from multi-view silhouettes. *ACM Transactions on Graphics* 27, 3 (2008), Article 97.

Daniel Vlasic, Pieter Peers, Ilya Baran, Paul Debevec, Jovan Popović, Szymon Rusinkiewicz, and Wojciech Matusik. 2009. Dynamic shape capture using multi-view photometric stereo. *ACM Transactions on Graphics* 28, 5 (2009), 174.

Ruizhe Wang, Lingyu Wei, Etienne Vouga, Qixing Huang, Duygu Ceylan, Gerard Medioni, and Hao Li. 2016. Capturing dynamic textured surfaces of moving targets. In *Proceedings of the European Conference on Computer Vision (ECCV'16)*.

Michael Waschbüsch, Stephan Würmlin, Daniel Cotting, Filip Sadlo, and Markus Gross. 2005. Scalable 3D video of dynamic scenes. *Visual Computer* 21, 8-10 (2005), 629–638.

X. Wei, P. Zhang, and J. Chai. 2012. Accurate realtime full-body motion capture using a single depth camera. *ACM Transactions on Graphics* 31, 6 (2012), Article 188, 12 pages.

Alexander Weiss, David Hirshberg, and Michael J. Black. 2011. Home 3D body scans from noisy image and range data. In *Proceedings of the 2011 13th International Conference on Computer Vision (ICCV'11)*. IEEE, Los Alamitos, CA, 1951–1958.

Chenglei Wu, Carsten Stoll, Levi Valgaerts, and Christian Theobalt. 2013. On-set performance capture of multiple actors with a stereo camera. *ACM Transactions on Graphics* 32, Article 161, 11 pages. DOI: https://doi.org/10.1145/2508363.2508418

Chenglei Wu, Kiran Varanasi, and Christian Theobalt. 2012. Full body performance capture under uncontrolled and varying illumination: A shading-based approach. In *Proceedings of the 2012 European Conference on Computer Vision (ECCV'12)*. 757–770.

Weipeng Xu, Avishek Chatterjee, Michael Zollöfer, Helge Rhodin, Dushyant Mehta, Hans-Peter Seidel, and Christian Theobalt. 2018. MonoPerfCap: Human

performance capture from monocular video. *ACM Transactions on Graphics* 37, 2 (July 2018), Article 27.

Jinlong Yang, Jean-Sébastien Franco, Franck Hétroy-Wheeler, and Stefanie Wuhrer. 2016. Estimation of human body shape in motion with wide clothing. In *Proceedings of the 2016 European Conference on Computer Vision (ECCV'16)*.

Genzhi Ye, Yebin Liu, Nils Hasler, Xiangyang Ji, Qionghai Dai, and Christian Theobalt. 2012. Performance capture of interacting characters with handheld Kinects. In *Computer Vision—ECCV 2012*. Lecture Notes in Computer Science, Vol. 7573. Springer, 828–841. DOI: https://doi.org/10.1007/978-3-642-33709-3_59

Mao Ye and Ruigang Yang. 2014. Real-time simultaneous pose and shape estimation for articulated objects using a single depth camera. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2345–2352.

Rui Yu, Chris Russell, Neill D. F. Campbell, and Lourdes Agapito. 2015. Direct, dense, and deformable: Template-based non-rigid 3D reconstruction from RGB video. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV'15)*.

Tao Yu, Kaiwen Guo, Feng Xu, Yuan Dong, Zhaoqi Su, Jianhui Zhao, Jianguo Li, et al. 2017. BodyFusion: Real-time capture of human motion and surface geometry using a single depth camera. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV'17)*.

Tao Yu, Zerong Zheng, Kaiwen Guo, Jianhui Zhao, Qionghai Dai, Hao Li, Gerard Pons-Moll, et al. 2018. DoubleFusion: Real-time capture of human performances with inner body shapes from a single depth sensor. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'18)*. IEEE, Los Alamitos, CA.

Chao Zhang, Sergi Pujades, Michael Black, and Gerard Pons-Moll. 2017. Detailed, accurate, human shape estimation from clothed 3D scan sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'17)*.

Peizhao Zhang, Kristin Siu, Jianjie Zhang, C. Karen Liu, and Jinxiang Chai. 2014b. Leveraging depth cameras and wearable pressure sensors for full-body kinematics and dynamics capture. *ACM Transactions on Graphics* 33, 6 (2014), 14.

Qing Zhang, Bo Fu, Mao Ye, and Ruigang Yang. 2014a. Quality dynamic human body modeling using a single low-cost depth camera. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Los Alamitos, CA, 676–683.

Qian-Yi Zhou and Vladlen Koltun. 2014. Color map optimization for 3D reconstruction with consumer depth cameras. *ACM Transactions on Graphics* 33, 4 (2014), 155.

Shizhe Zhou, Hongbo Fu, Ligang Liu, Daniel Cohen-Or, and Xiaoguang Han. 2010. Parametric reshaping of human bodies in images. *ACM Transactions on Graphics* 29, 4 (2010), 126.

Xingyi Zhou, Qixing Huang, Xiao Sun, Xiangyang Xue, and Yichen Wei. 2017. Towards 3D human pose estimation in the wild: A weakly-supervised approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 398–407.

Xiaowei Zhou, Menglong Zhu, Spyridon Leonardos, Konstantinos G Derpanis, and Kostas Daniilidis. 2016. Sparseness meets deepness: 3D human pose estimation from monocular video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4966–4975.

Zoran Zivkovic and Ferdinand van der Heijden. 2006. Efficient adaptive density estimation per image pixel for the task of background subtraction. *Pattern Recognition Letters* 27, 7 (May 2006), 773–780. DOI: https://doi.org/10.1016/j.patrec.2005.11.005

Michael Zollhöfer, Matthias Nießner, Shahram Izadi, Christoph Rhemann, Christopher Zach, Matthew Fisher, Chenglei Wu, et al. 2014. Real-time non-rigid reconstruction using an RGB-D camera. *ACM Transactions on Graphics* 33, 4 (July 2014), Article 156.

Received September 2018; revised January 2019; accepted January 2019