MulayCap: Multi-layer Human Performance Capture Using A Monocular Video Camera

Zhaoqi Su, Weilin Wan, Tao Yu, Lingjie Liu, Lu Fang, Wenping Wang and Yebin Liu

Abstract—We introduce MulayCap, a novel human performance capture method using a monocular video camera without the need for pre-scanning. The method uses "multi-layer" representations for geometry reconstruction and texture rendering, respectively. For geometry reconstruction, we decompose the clothed human into multiple geometry layers, namely a body mesh layer and a garment piece layer. The key technique behind is a Garment-from-Video (GfV) method for optimizing the garment shape and reconstructing the dynamic cloth to fit the input video sequence, based on a cloth simulation model which is effectively solved with gradient descent. For texture rendering, we decompose each input image frame into a shading layer and an albedo layer, and propose a method for fusing a fixed albedo map and solving for detailed garment geometry using the shading layer. Compared with existing single view human performance capture systems, our "multi-layer" approach bypasses the tedious and time consuming scanning step for obtaining a human specific mesh template. Experimental results demonstrate that MulayCap produces realistic rendering of dynamically changing details that has not been achieved in any previous monocular video camera systems. Benefiting from its fully semantic modeling, MulayCap can be applied to various important editing applications, such as cloth editing, re-targeting, relighting, and AR applications.

Index Terms—Human Performance Capture, 3D Pose Estimation, Cloth Animation, Non-rigid Deformation, Intrinsic Decomposition.

1 Introduction

Human performance capture aims to reconstruct a temporally coherent representation of a person's dynamically deforming surface (i.e., 4D reconstruction). Despite the rapid progress in the study on 4D reconstruction using multiple RGB cameras or single RGB-D camera, using a single monocular video camera for robust and accurate 4D reconstruction remains an ultimate goal because it will provide a practical and convenient way of human performance capturing in general scenarios, thus enabling the adoption of human performance capturing technology in various consumer applications, such as augmented reality, computer animation, holography telepresence, biomechanics, virtual dressing, etc. However, this problem is highly challenging and ill-posed, due to the fast motion, complex cloth appearance, non-rigid deformations, occlusions and the lack of depth information.

Due to these difficulties, there have been few attempts using a single monocular RGB camera for human performance capture. The most recent works in [1] and [2] approach the problem by using a pre-scanned actor-specific template mesh, which requires extra labor and time to scan, making these methods hard to use for consumer applications or for human performance reconstruction using Internet videos. Moreover, these methods suffer from the limitation of using a single mesh surface to represent a human character, that is, the visible part of human skin and dressed cloth are not separated. As a consequence, common cloth-body interaction, such as layering and sliding, is poorly tracked and represented. Furthermore, once obtained from the pre-scanned template mesh, the reconstructed texture is fixed to the mesh over all the frames, resulting in unrealistic artifacts.

Without a pre-scanned model, human performance capture is a very difficult problem indeed, due to the need for resolving motion, geometry and appearance from the video frames simultaneously, without any prior knowledge about geometry and appearance. Regarding geometry, reconstruction of a free-form deformable surface from a single video is subject to ambiguity [3]. As for texture, it is hard to acquire a dynamic texture free of artifact. Specifically, complex non-rigid motions introduce spatially and temporally varying shading on the surface texture. Directly updating the observed texture on the garment template to represent the motion may introduce serious stitching artifacts, even with ideal and precise geometry models. While artifact-free texture mapping can be obtained by scanning a static key model followed by deforming it in a non-rigid manner for temporal reconstruction, the resultant appearance tends to be static and unnatural.

In this paper, we propose *MulayCap*, a multi-layer human performance capture approach using a monocular RGB video camera that achieves dynamic geometry and texture rendering without the need of an actor-specific pre-scanned template mesh. Here the 'Mulay' notation means that "multi-layer" representations are proposed for reconstructing geometry and texture, respectively. We use multi-layer representation in geometry reconstruction, which decomposes the clothed human into multiple geometric layers, namely a naked body mesh layer and a garment piece layer. In fact, two garment layers are used, one for the upper body clothing, such as a T-shirt, and the other for pants or trousers. The upper body clothing can also be generalized to include lady's dresses, as shown in Fig. 1, which uses the same 2D garment patterns as T-shirt, shown in Fig. 3,

- *Z. Su, T. Yu, L. Fang and Y. Liu are with Tsinghua University.*
- *W. Wan, L. Liu and W. Wang are with The University of Hong Kong.*
- *Corresponding authors: Yebin Liu*

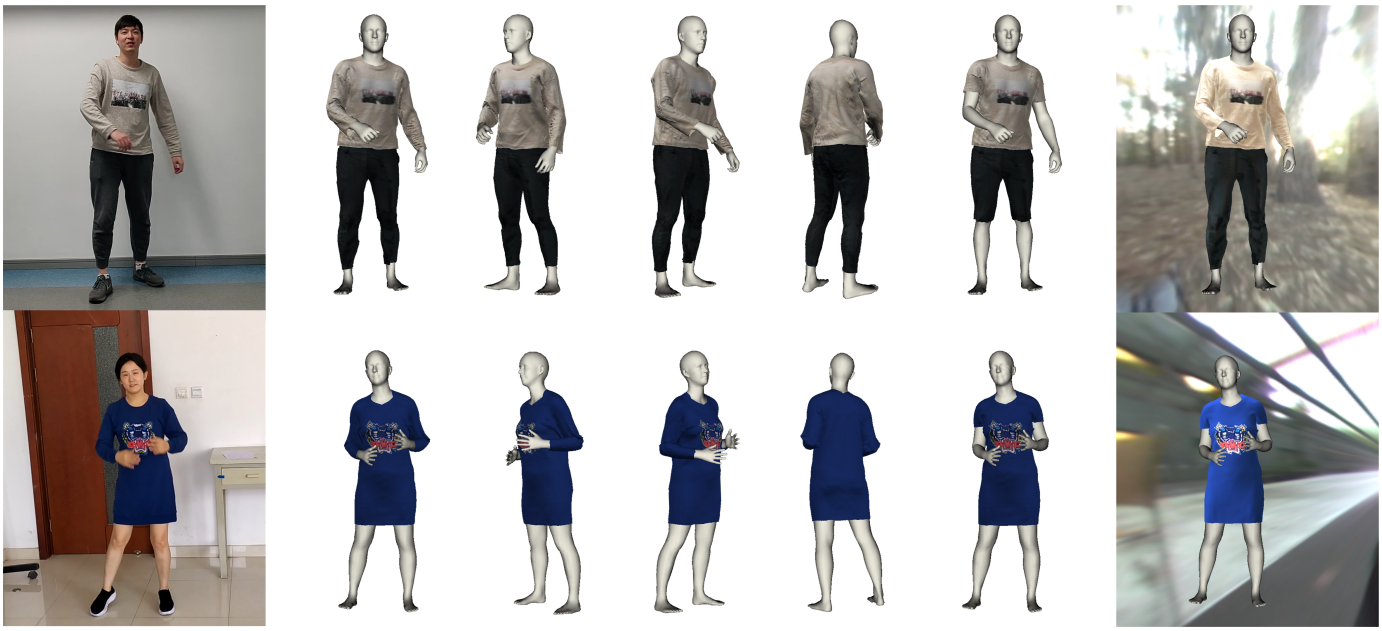


Fig. 1. Results generated by our MulayCap system from a monocular RGB video. From left to right: one of input images, four generated results (one in the reference view and three in different viewing directions), a cloth editing result, and a relighting result rendered under a novel lighting condition.

only different in parameters. To solve the garment modeling problem, we propose a Garment-from-Video (GfV) algorithm based cloth simulation. Specifically, the garment shape parameters serve as parameters for cloth simulation and optimized by minimizing the difference between simulated cloth model and the dressed garment observed in the input video. During optimization, to avoid the exhaustive and inefficient search for garment parameters, we use gradient descent minimization with a specified number of iterations. To further align the cloth simulation results with the input images, we apply a non-rigid deformation based on the shape and feature cues in each image. We demonstrate that our proposed garment simulation and optimization framework is capable of producing high quality and dynamic geometry details from a single video.

The multi-layer representation also works for dynamic texture reconstruction, in which the input video images are decomposed into albedo layers and shading layers for generating albedo atlas with geometry details on clothing. Specifically, each input image is first decomposed into an albedo image and a shading image, then the per-frame albedo is fused with the reconstructed garments to create a static and shading-free texture layer. The albedo layers serve to maintain a temporally coherent texture basis. To obtain a realistic dynamic shape of cloth, we use the shading image to solve for the garment geometry detail with a shape-from-shading method. Finally, by compositing the detailed geometry, albedo and lighting information, we produce high quality and dynamic textured human performance rendering, which preserves the spatial and temporal coherence of dynamic textures and the detail of dynamic wrinkles on clothes.

In a nutshell, we present a novel template-free approach, called *MulayCap*, for human performance capture with a single RGB camera. The use of the multi-layer representations enables more semantic modeling of human performance, in which the body, garment pieces, albedo, shading are separately modeled and elaborately integrated to produce high quality realistic results. This approach takes full advantage of high level vision priors from existing computer vision research to yield high quality reconstruction with light-weight input. In contrast with the existing human performance capture systems [4], [5], [6], our fully-semantic cloth and body reconstruction system facilitates more editing possibilities on the reconstructed human performances, such as relighting, body shape editing, cloth re-targeting, cloth appearance editing, etc., as will be shown later in the paper.

2 Related Work

While there are a large number of prominent works in human performance capture, we mainly review the works that are most related to our approach. We also summarize other related techniques including human shape and pose estimation as well as cloth simulation and capture.

Human Performance Capture The research on Human performance capture has been well studied for many decades in computer vision and computer graphics. Most of the existing systems adopt generative optimization approaches, which can be roughly categorized into multiview-RGB-based methods and depth-based methods according to the capture setups. On the other hand, based on the representations of the captured subject, generative human performance capture methods can be classified as free-form methods, template-based deformation methods and parametric-model-based deformation methods.

For multiview-RGB-based human performance capture methods, earlier researches focus on free-form dynamic reconstruction. These methods use multiview video input by leveraging shape-from-silhouette [7], [8], multiview stereo [9], [10], [11] or photometric stereo [12] techniques. [13] performs video-realistic interactive character animation from a

4D database captured in a multiple camera studio. Benefiting from the deep learning techniques, recent approaches try to minimize the number of used cameras (around 4) [14], [15]. Template-based deformation methods need pre-scanned templates of the subjects before motion tracking. They can generate topology consistent and temporally coherent model sequences. Such methods take advantage of the relatively accurate pre-scanned human geometry prior and use non-rigid surface deformation [16], [17] or skeleton-based skinning techniques [18], [19], [20], [21], [22] to match the multi-view silhouettes and the stereo cues. There have been few studies focusing on temporally coherent shape and pose capture using monocular RGB video sequences. Existing works include [1] and [2], where a pre-scanned textured 3D model is a pre-requisite for both of them. In their methods, 3D joint positions are optimized based on the CNN-based 2D and 3D joint detection results. Moreover, non-rigid surface deformation is incorporated to fit the silhouettes and photometric constraints for more accurate pose and surface deformation. In parametric-model-based deformation methods. The character specific models used in the methods above are replaced by parametric body models like [23], [24], [25], [26], [27], [28] to eliminate the pre-scanning efforts. However, parametric body models always have limited power to describe the real world detailed surface of the subject. Overall, as most of the template-based deformation methods regard the human surface as a single-piece of watertight geometry, various free-form garment motion and garment-body interactions cannot be described by the surface deformation, which also acts as a key preventer for high quality dynamic texture mapping.

Depth-based methods are relatively more efficient as the 3D surface point clouds are provided directly. Many of the previous works in this field are free-form approaches, in which an in-completed template is gradually fused given continuous depth observations. Such free-from methods start from the fusion of a general dynamic scene [29], and have been improved by considering texture constraints [30], [31] and resolving topology changes [32], [33]. Multiple depth sensor based fusion approaches [34], [35], [36], [37] have been developed to improve the robustness and accuracy through registering multi-view depth streams. Besides free-form fusion based methods, performance capturing using template-based deformation is also a well studied area. [38], [39], [40], [41] leverage pre-scanned models to account for non-rigid surfaces, while in [42], [43], [44], [45] the performance capturing problem is decomposed into articulated motion tracking and shape adaption. [46] builds BUFF Dataset which contains high quality clothed 3D scan sequences of the human, and estimates the human body shape and pose from these sequences. There are also some fusion-based approaches combining articulated templates or articulated priors for robust motion tracking and surface fusion [47], [48], [49], [50], [51].

Recently, benefiting from the success of deep learning, discriminative approaches for single image human shape and pose detection catch lots of research attention. They have demonstrated that it is possible to estimate the human shape and pose using only a single RGB image by taking advantage of the parametric body models [23], [52]. [53] optimizes the body shape and pose parameters by minimizing the distance between the detected 2D joints from a CNN-based pose detector and the projected 3D joints of the SMPL model. Follow-up works extend this approach by predicting the 3D pose and shape parameters directly. [54] proposes a two-step deep learning framework, where the first step estimates key joints and silhouettes from input images, and the second step predicts the SMPL parameters. [55] estimates SMPL parameters through body part segmentation. [56] uses a 3D regression module to estimate SMPL parameters and weak camera parameters, and it incorporates an adversarial prior to discriminate unusual poses as well. [57] uses temporal information to estimate human poses in a video. [58] leverages both the idea from [56] and [53] and combine the structures from both for iteratively optimization of the human model. [59] uses a more expressive model SMPL-X for the human face and hands. Besides, There are also some deep learning approaches for estimating the whole human 3D model or frontal depth map from a single image without using parametric models [60], [61], [62], [63], [64].

Cloth Simulation and Capture The ultimate goal of cloth simulation and cloth capture is to generate realistic 3D cloth with its dynamics. Given a 3D cloth model with its physical material parameters, the task of cloth simulation is to simulate realistic cloth dynamics even under different kinds of cloth-object interactions. Classical force-based cloth simulation methods are derived from continuum mechanics [65], it can be a mass-spring system [66], [67], [68] or other more physically consistent models generated by the finite element method [69], [70]. These methods need to perform numerical time integration for simulating cloth dynamics, which include the more straightforward explicit Euler method [71] and other more stable implicit integration methods like implicit or semi-implicit Euler method [65], [72], [73]. The force-based cloth simulation methods can generate very realistic cloth dynamics benefiting from the physically consistent models. Note that in our MulayCap, cloth simulation is especially useful in dressing the naked body and generating plausible cloth dynamics when only 2D parametric cloth pieces and monocular color video are available. Since the highly accurate material modeling is not a requirement of our system, we use the method in [66] for simplicity and efficiency.

Different from cloth simulation, the cloth capture methods mainly concentrate on another problem: how to digitize the real world 3D model and even the real world dynamics of the cloth. For active methods, [74] custom designed the cloth with specific color patterns and [75] uses the custom designed active multi-spectral lighting for accurate cloth geometry and even material capture. However, the active methods are much more complex and may not generalize to off-the-shelf clothes. The passive methods are much more popular and have been developed using different kinds of information as input: multi-view rgb [5], [76], [77], 4D sequences [6], [78], RGBD [79], [80] or even single RGB [81], [82], [83], [84], [85], [86], [87]. Among these passive methods, [76], [77] focus on reconstructing real cloth geometries and even cloth wrinkle details using temporally coherent multi-view stereo algorithm and data driven approach. [5] utilizes the multi-view reconstructed 4D human performances to reconstruct a physically-animatable dressed avatar. [80] use a single RGBD camera to accomplish multi-layer human

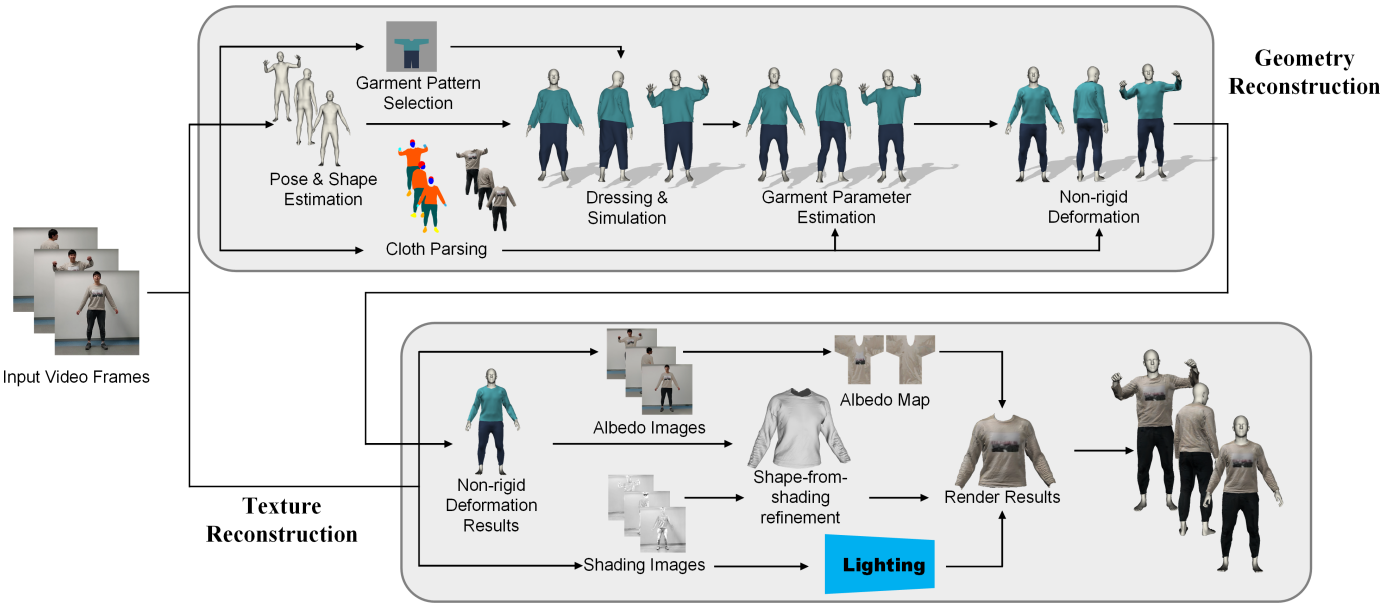


Fig. 2. The pipeline of MulayCap. Given the input monocular RGB video, the clothed human reconstruction is achieved by geometry and texture reconstruction. We first estimate the human pose and shape, and reconstruct the garment-based cloth based on the human model, then apply non-rigid cloth deformation based on semantic cloth segmentation result. The second step is to use albedo and shading images decomposed from the input frames to obtain cloth texture, geometry details and lighting, which are then combined for realistic rendering of the dynamic cloth.

performance capture, which benefits from a physics-based performance capture procedure. Given a high quality 4D sequence, [6] semantically digitizes the whole sequence and generates temporally coherent multi-layer meshes of both human body and the cloth, while [88] propose a multi-task learning framework for garment fashion landmark extraction and garment segmentation from an input image, and can generate garment shape and texture from a single image. [78] learns a cloth specific high frequency wrinkle model based on normal mapping and use the model to wrinkle the cloth under non-captured poses. [84] learns to reconstruct people in clothing with high accuracy, which uses a monocular video of a moving person as input. [86] build a real-time human performance capture system, which uses RGB video as input and can reconstruct space-time coherent deforming geometry of an entire human. [85] learns geometry details from texture map, and can infer a full body shape including cloth wrinkles, face and hair from a single RGB image. [87] uses silhouette information of a single RGB image to infer a textured 3D human model using deep generative models. [79] is a data driven approach, it first reconstructs a static textured model of the subject using RGBD sequence and then performs cloth parsing based on a pre-designed cloth database for static but semantic cloth capture; [81] improves [79] by using only a single image. [82] learns a specific cloth model from a large amount of cloth simulation results with different bodies and poses, and uses it to infer cloth geometry directly from a single color image. Benefiting from the parametric cloth models, [83] can reconstruct both body shape and physically plausible cloth from a single image. However, such method mainly focuses on static cloth reconstruction, thus the cloth dynamics can only be generated by simulation. While Our method can generate realistic cloth dynamic appearance given a video sequence. To conclude, on one hand, the data driven approaches above need either (captured or simulated) high quality 4D sequences or self-designed cloth databases as input, which is hard to obtain. Moreover, the generalization of such approaches remains challenging. On the other hand, current direct cloth capture approaches still need carefully designed setups or the heavy multi-view capture systems for high quality cloth capture. In our system, we use a data-driven approach to reconstruct static cloth models and propose a new direct cloth capture approach for capturing realistic cloth dynamic appearance from a monocular video footage. There are also many interesting applications of cloth simulation and capture. For example, [89] proposes a learning method which compiles 2D sketches, parametric cloth model and parametric body model into a shared space for interactive garment design, generation and fitting. One interesting work correlates to ours is [90], it mainly focuses on garment replacement (but not capture) given a monocular video footage and relies on manual intervention. However, our method is fully automatic and produces both realistic cloth capture and cloth replacement results.

Intrinsic Decomposition The objective of intrinsic decomposition is to decompose a raw image into the product of its reflectance and shading. Because the decomposition of raw images is insufficiently constrained, optimization based methods often tackle the problem by carefully designed priors [91], [92], [93], [94], while deep-learning based methods incorporate learning from ground truth decomposition of raw images [95], [96], [97]. There are also sequences based solutions, using propagation methods [98], [99], [100], or considering reflectance as static while shading changes over time [101], [102]. Methods proposed in [103], [104] leverage multi-view inputs to recover the scene geometry and estimate the environment lighting.

Intrinsic decomposition has been widely applied for identifying the true colors of objects and analyzing inter-

actions with lights in the scene. Researchers have presented various applications based on the progress in this field, such as material editing by shading modification and recolorization by defining transfer among the origin and the target reflectance [98]. Since most wrinkles and folds on the cloth mainly contribute to the shading effects of the input frames, intrinsic decomposition can be directly incorporated into our system to recover such details.

3 Method Overview

The main pipeline of our MulayCap consists of two modules, i.e., multi-layer based geometry reconstruction (see Sect. 4) and multi-layer based texturing rendering (see Sect. 5), as shown in Fig. 2. For the geometry module, we reconstruct a clothed human model for each input video frame. Each target clothed model contains separated geometry mesh layers for individual garments and the human body mesh model SMPL [52]. We first detect and optimize the human shape and pose parameters of SMPL model to get the body layer (see Sect. 4.1). We select the 2D garment pattern and automatically dress the human temporal body models using available cloth simulation methods [105] (see Sect. 4.2.1). After that, since the garments may not fit with the input images, we optimize the 2D garment shape parameters using all the 2D segmented garment pieces obtained by instance human parsing methods like [106] from the video frames (see Sect. 4.2.2). We name this garment shape optimization method based on cloth simulation as GfV. To further align the boundary in each temporal image, we refine the non-rigid deformation of the garments based on the silhouette information in each input image (see Sect. 4.2.3).

For the texture module, to achieve temporally dynamical and artifact-free texture updating, we composite a static albedo layer and a constantly updated geometry detail layer on the 3D garments. The garment albedo layer represents a clean and shadow-free texture while the geometry detail layer describes the dynamically changing wrinkles and shadows. First, based on the obtained clothed model sequence in the geometry module, we leverage the intrinsic decomposition method in [95] to decompose the input cloth images into albedo images and shading images. Multiple albedo images are then stitched and optimized on the 3D garment to form a static albedo layer (see Sect. 5.1). For the geometry detail, we further decompose the shading images into environment lighting and surface normal images (see Sect. 5.2). The normal images are then used to solve for surface details on the 3D garments (see Sect. 5.3). In this way, by using albedo images to render surface albedo, surface detail and environment lighting, we achieve realistic cloth rendering with temporally varying wrinkle details, without the side effect of stitching texture artifacts.

4 Multi Layer Geometry

To elaborate on the geometry reconstruction in MulayCap, we first describe the reconstruction of body meshes, followed by the dressing-on and optimization of the garment layers.

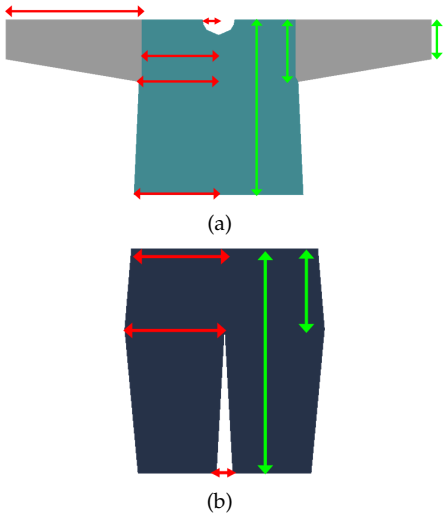


Fig. 3. 2D garment patterns for generating different clothes. (a) The 2D garment for upper cloth. (b) The 2D garment for pants. The red arrows and green arrows indicate the parameters for controlling the width and height of the clothes, respectively.

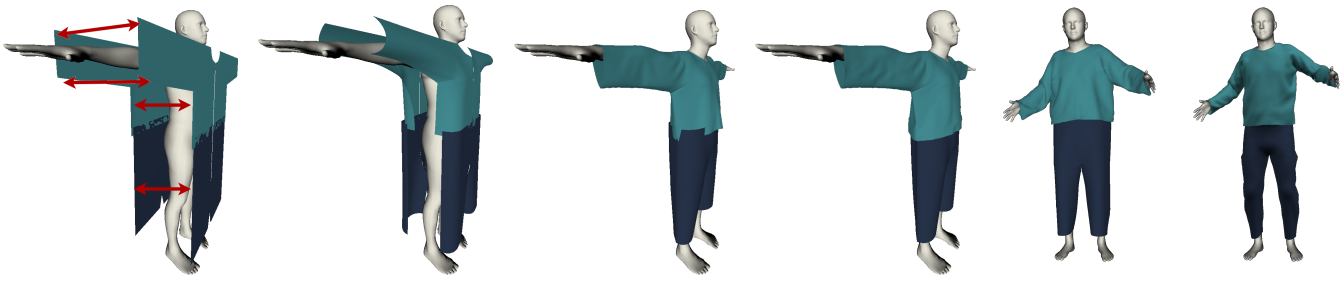
4.1 Body Estimation

We use SMPL [52] to track the human shape and pose in each frame. Specifically, We first use HMMR method [107] to estimate initial pose parameters $\mathcal{P}_i$ and shape parameters $\mathcal{S}_i$ for each frame $\mathcal{I}_i$. All the $\mathcal{S}_i$ are averaged to get consistent SMPL shape for the whole sequence. We apply temporal smoothing to the pose parameters of adjacent frames to alleviate errors and jitter effects, and replace those poses with drastic changes by interpolation of their adjacent poses. We also leverage the 2D joints of humans detect by OpenPose [108] to further fix the inaccurate pose detected by HMMR [107] and estimate the global translation of the human model, by constraining the 2D distance between the projected 3D joint position and the detect one.

4.2 Garment from Video (GfV)

4.2.1 Dressing

The cloth dressing task consists of two steps: a) garment pattern initialization; b) physical simulation. For reconstructing garment layers, multiple 2D template garment patterns are used for different initial 3D garment meshes. Two layers of garments are used for the upper body and the lower body, respectively, as shown in Fig. 3. The parameters are defined as the length of the green and red arrows in Fig. 3, which leads to 8 parameters for upper cloth and 5 parameters for pants. The parameters are defined in 2D garment patterns, inspired by the industry designing pattern of clothes. Each pattern is composed of a front piece and a back piece. The sizes of garment pieces are specified by the estimated body shape automatically. As shown in Fig. 4(e), the initialization step needs to guarantee the cloth is wide enough to be dressed on. Therefore, we utilize the length of the torso and legs of the SMPL model for setting the initial heights for the 2D patterns, and set initial width parameters according to the bust and waist measurement of the SMPL model with a scale factor of 1.5, for dressing on the body. After initialization, the 2D garment parameters can be optimized to fit



(a) (b) (c) (d) (e) (f)

Fig. 4. Garment dressing on T-posed body. (a) The initial state of the dressing process, defined by initial garment parameters. The red arrows indicates the attractive force defined on the unseamed garment vertexes. (b) The process of seaming the garment. (c) The cloth is seamed and worn on the body, simulated with gravity and body collision. (d) Resolving garment-garment collision. (e) Simulation result on one of the input frame based on initial garment parameters. (f) Results after the garment parameter optimization under the same pose of (e).

the real world observations by leveraging both 3D physics-based simulation and numerical differentiation, which is detailed in Sect 4.2.2.

To drape the template garments onto a human body mesh in the standard T pose, we introduce external forces to stitch the front and back piece using physics-based simulation, as shown in Fig. 4(a)-(e). The details of cloth dressing and simulation are described below.

For the sake of efficiency, we use the efficient and classical Force-Based Mass-Spring method of [105] for physics-based simulation, which treats the cloth mesh vertexes as particles connected by springs. For cloth simulation, the external forces applied to each garment vertex include gravity G and the friction between the human body and the cloth. The collision constraints are added between the cloth vertices and the human model, and also the cloth itself (e.g. T-shirt and pants) to avoid inter-penetrations. Specifically, for each cloth vertex v_i, we find its nearest SMPL vertex p_i and calculate the point-plane distance $n_i(v_i - p_i)$. If the point-plane distance is below zero, we assume that the collision constraint should be applied to v_i. The collision constraints between overlapping cloth working in a similar way. For the whole sequence, the physics-based simulation is conducted sequentially. Specifically, using the draping of the previous frame to initialize the draping of current frame.

The goal of the dressing step is to seam the two 2D pieces and stitch them as a complete 3D garment. We use the mean parameters of our parametric cloth model for the dressing step, the parameters will be further refined as described in Sect. 4.2.2. As shown in Fig. 4, to put different garment elements together on the T-posed human body, we apply attractive forces on particle pairs at the cutting edges. After about 300 rounds of simulation, a seam detection algorithm is performed to detect whether each unseamed particle-pairs is seamed successfully by detecting whether the distance between the vertex-pairs to be seamed are all smaller than a threshold. If not, it means that the initial garment parameters cannot fit the human body, and thus will be automatically re-adjusted. For instance, if the upper cloth cannot be seamed at the wrist part, the algorithm predicts that the cloth is too tight for the human model, and the corresponding garment parameters will be updated progressively until the cloth is successfully seamed, in this case the parameter for sleeve height will be increased gradually until the cloth is seamed well.

4.2.2 Garment Parameter Estimation

After the dressing step, we estimate the garment parameters with several simulation passes. In each pass, the cloth simulation is performed on the whole sequence according to the estimated body shape and poses from Sect. 4.1. The optimization is initialized with the garment geometry from the initial simulation pass.

The garment shape is optimized by minimizing the difference between the rendered simulation results and the input image frames. Here we use cloth boundaries in all the image frames as the main constraints for fitting. Given a set of garment shape parameters θ_C, we simulate, render and measure the following energy function $E_{garment}$

$$E_{garment} = E_{bd} + E_{reg}, \tag{1}$$

where $E_{reg} = ||\Delta\theta_C||^2$ is used to regularize the updates between iterations, and E_{bd} is used to maximize the matching between the cloth boundaries. For a garment C worn on the body, E_{bd} is defined as

$$\begin{aligned} E_{bd} &= \sum_k ||\mathcal{F}_k(\theta_C)||^2 \\ &\equiv \sum_k ||\mathcal{DT}(\mathcal{I}_{img}(C)) - \mathcal{DT}(\mathcal{I}_{rn}(\theta_C))||^2, \end{aligned} \tag{2}$$

where k is the frame index evenly sampled from the input frames, and $\mathcal{I}_{img}(C)$ is the segmented cloth from the input image obtained with the garment instance parsing method [106], $\mathcal{I}_{rn}(\theta_C)$ is the simulated and rendered cloth silhouette using garment parameters θ_C, and $\mathcal{DT}(\mathcal{I})$ represents the distance map of the silhouette boundary of image $\mathcal{I}$ and is defined as

$$\mathcal{DT}(\mathcal{I}) = max(0, min(\epsilon_{DT}, (\epsilon_{DT} - C(\mathcal{I})) + (\epsilon_{DT} - C(\bar{\mathcal{I}}))), \tag{3}$$

where ϵ_{DT} is a threshold set as 50. Here $C(\mathcal{I})$ and $C(\bar{\mathcal{I}})$ are the distance transform from silhouette of image $\mathcal{I}$ and its inverse image, respectively.

Since the rendering results after cloth simulation are also determined by complex cloth-body interactions, the cloth vertex position cannot be simply formulated as a function of θ_C. To calculate the gradient of the energy term for Gauss-Newton iteration, we use a numerical differentiation strategy: given the garment parameters θ_C, we add a small value $\Delta\theta_C^i$ to its i^{th} element, then reset the cloth vertex-pair constraint based on the new garment parameters, and

perform the simulation again to calculate the energy for the new parameters as $\mathcal{F}_k(\theta_C + \Delta\theta_C^i)$. The gradient used for Gauss-Newton iteration is then calculated as:

$$\frac{\partial \mathcal{F}_k(\theta_C)}{\partial \theta_C^i} = \frac{\mathcal{F}_k(\theta_C + \Delta\theta_C^i) - \mathcal{F}_k(\theta_C - \Delta\theta_C^i)}{2\Delta\theta_C^i}. \quad (4)$$

To be more specific, when generating 3D garments under new parameters $\theta_C + \Delta\theta_C^i$ in the k^{th} frame, we first update the corresponding 2D garment patterns, for the garment parameters θ_C are defined on the 2D patterns directly as shown in Fig. 3. As we cannot change the shape of the garment mesh directly in the process of physics-based simulation, we find out that we can directly update the initial simulation status (including force and vertex-pair rest length calculated by 2D patterns) of the garment, and perform the simulation. In this way, we can generate simulated garment mesh with new parameters without interrupting the simulation or performing the cloth dressing again. Finally, after the simulation has reached a stable state under the new parameters, we can use it to calculate the energy function $\mathcal{F}_k(\theta_C + \Delta\theta_C^i)$. The update of θ_C is calculated using Gauss-Newton method. Note that $\Delta\theta_C^i$ is only used as a step value for numerical differentiation, which is not the update of θ_C in each iteration. The $\Delta\theta_C^i$ in our system is set to 0.01.

Our tests show that 25 iterations is generally enough for the above simulation-and-numerical-optimization method to converge and get plausible garment shape parameters. Fig. 4(f) illustrates the garment shape optimization result over Fig. 4(e).

4.2.3 Non-rigid Deformation Refinement

Note that the garment parameters solved in Sect. 4.2.2 provides only a rough estimation and the physics-based simulation cannot describe the subtle movement of the cloth, such as wrinkles, which are critical to realistic appearance modeling. We therefore refine the garment geometry using a non-rigid deformation approach to model dynamic cloth details. We up-sample the low resolution cloth mesh used for physics-based simulation to match the pixel resolution for detailed non-rigid alignment and subsequent geometry refinement. Here, we use garment boundary to determine the displacement of each vertex Δv_i^h in the high resolution garment mesh by minimizing the following energy function

$$E_{nonrigid} = E_{bd}^{(h)} + E_{smooth}^{(h)} + E_{reg}^{(h)}. \quad (5)$$

Here

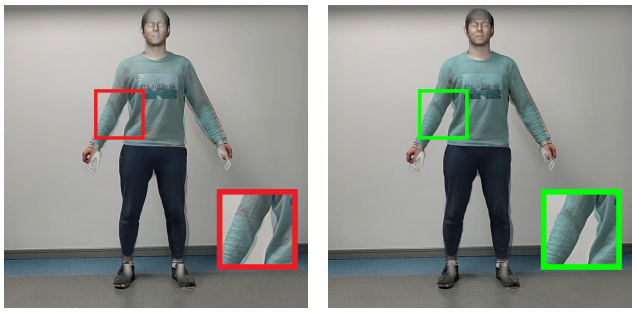
$$E_{bd}^{(h)} = \sum_i ||\mathcal{DT}(\mathcal{I}_{img}({v_i}^h + \Delta {v_i}^h)) - \mathcal{DT}(\mathcal{I}_{rn}({v_i}^h))||^2. \quad (6)$$

The smoothness term $E_{smooth}^{(h)}$ used to regularize the difference between displacements of the neighboring mesh vertexes:

$$E_{smooth}^{(h)} = \lambda_{nearby}^{(h)} \sum_i \sum_{j \in \mathcal{N}_i} ||\Delta {v_i}^h - \Delta {v_j}^h||^2. \quad (7)$$

The regularization term $E_{reg}^{(h)}$ is defined in the same way as E_{reg} to constrain the displacement magnitudes.

The energy function in 5 is minimized using the Gauss-Newton method. As the energy term is defined either on each high resolution mesh vertex, or between nearby vertexes, the energy matrix is sparse so that the conjugate gradient algorithm is used in each Gauss-Newton iteration. The improvement resulting from refinement by non-rigid deformation is shown in Fig. 5(a) and Fig. 5(b). See the zoom-in for detailed boundary overlays. It can be seen that the boundary overlay of rendered mesh with optimization in Fig. 5(b) is more accurate than that in Fig. 5(a).



(a) (b)

Fig. 5. The results of garment deformation refinement. Both (a) and (b) show the reconstructed garment meshes overlay on the input image. Please zoom in to see the overlapping boundaries. (a) The result before deformation refinement. (b) The result after deformation refinement.

5 Texture and Geometry Detail Refinement

In this section we will consider mapping texture to the detailed garment surface reconstructed in the preceding step. Note that directly texturing and updating would introduce serious stitching artifacts due to the spatially and temporally varying shadings and shadows. To obtain artifact-free and dynamically changing surface texture, we decompose the texture into the shading layer and the albedo layer, with the former for geometry detail refinement and global lighting, and the latter for generating a static albedo map for the cloth.

Specifically, for each input image frame $\mathcal{I}$, we use the CNN-based intrinsic decomposition method proposed in [95] to get a reflectance image $\mathcal{I}_F$ and a shading image $\mathcal{I}_S$. We then use $\mathcal{I}_F$ for garment albedo map calculation and $\mathcal{I}_S$ for dynamic geometry detail refinement and lighting estimation. All these three components (i.e. detailed garment geometry, albedo map and lighting) are then combined to produce realistic garment rendering.

5.1 Albedo Atlas Fusion

To generate a static albedo atlas on each 3D garment, we need to keep a optimized texture base for reducing stitching artifacts and maintaining spatially and temporally consistent texturing. Specifically, for each garment, we build a texture U-V coordinate domain according to the 2D garment designing pattern so that each vertex on a reconstructed garment mesh is assigned to a UV coordinate.

Our albedo fusion algorithm creates the albedo atlas based on the evenly sampled albedo images $\mathcal{I}_F$. As multiple albedo pixels on different albedo images may project to the same UV coordinate, a multi-image blending algorithm is needed for creating a high quality albedo atlas. We resolve

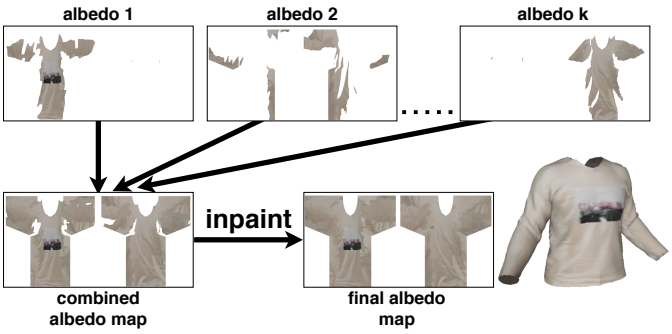


Fig. 6. The pipeline of generating the albedo atlas. We iteratively combine albedo maps from evenly sampled key-frames and inpaint the unseen areas.

this problem by simply using an as-good-as-possible albedo pixel from the multiple images for obtaining the albedo atlas. The selection strategy resembles the multiview texture mapping schemes [109], [110], which try to select the camera that minimizes the angle between the surface normal direction and the vertex-to-camera direction. To mitigate the mosaicing seams, we follow the MRF seam optimization method [111] to remove mosaic seams without affecting the fine details of the albedo. For areas unseen in the sequence, we inpaint [112] those areas to obtain a full albedo atlas. Fig. 6 illustrates the albedo atlas fusion pipeline.

5.2 Shading Decomposition

The goal for shading decomposition is to estimate an incident lighting $\mathcal{L}$ and normal images $\mathcal{I}_N$ from the shading image sequence $\mathcal{I}_S$. The incident lighting is then used for shape-from-shading based geometry refinement. Following the shading based surface refinement approach in [113], we use spherical harmonics to optimize the lighting $\mathcal{L}$ and the normal image $\mathcal{I}_N$ by minimizing the energy function

$$E_{shading} = \sum_p ||\mathcal{H}_M(\mathcal{L}, \mathcal{I}_N(p)) - \mathcal{I}_S(p)||^2, \tag{8}$$

which models the different the approximate shading $\mathcal{H}_M$ by spherical harmonics and the observed shading $\mathcal{I}_S$ over all the pixels p that belong to a garment in an image frame.

We get a normal map from the recovered garment surface for lighting optimization, which is a better initialization than uniform normal map initialization. To improve accuracy, we select multiple key frames to enrich the variance of surface normals as done for the albedo atlas generation step in Sect. 5.1, and estimate the lighting using the least square method over all the pixels in these frames. We select key frames in an iterative manner. Specifically, if the pose difference between the current frame and all the previously selected key frames is larger than a threshold, we add the current frame as a new key frame; the iteration is stopped when no new key frame can be added. In our experiments, we mainly focus on the torso movements and the threshold is set to be 0.5 for average angle-axis Euler distance among torso joints defined by SMPL model. To constrain the range of normal estimation, we regularize $E_{shading}$ and minimize the energy function E_{normal} based on the estimated lighting $\mathcal{L}$ by minimizing the following energy

$$E_{normal} = E_{shading} + E_{Lap} + E_{grad} + E_{norm}. \tag{9}$$

Here:

regularization term

$$E_{grad} = \lambda_{grad} \sum_p ||\Delta \mathbf{n}(p)||^2 \tag{10}$$

is used to constrain the updating step.

Laplacian term

$$E_{lap} = \lambda_{lap} \sum_p ||Avg_{p' \in N_p} \mathbf{n}(p') - \mathbf{n}(p)||^2 \tag{11}$$

is used to constrain the smoothness of the normal image, where N_p is the set of p's neighbor pixels.

normalization term

$$E_{norm} = \lambda_{norm} \sum_p ||\mathbf{n}(p)^T \mathbf{n}(p) - 1||^2 \tag{12}$$

is used to constrain the normal at every cloth pixel p to be normalized.

5.3 Geometry refinement using shape-from-shading

So far, we have obtain the incident lighting of the scene. Then we use the shape-from-shading approach to refine the geometry detail using both lighting and shading image sequence $\mathcal{I}_S$ to represent wrinkles and folds for realistic rendering. Benefiting from the proposed multi-layer representation and physics-based optimization, we can obtain more reasonable garment geometry as the initial status for this geometric detail refinement step. To compute the per-vertex displacement of cloth, we first formulate its normal as follows,

$$n_k = norm(\sum_{j \in N_k} (v_j - v_k) \times (v_{j+1} - v_k)) \tag{13}$$

where N_k is the set of v_k's neighbors in clockwise order. Then we can formulate the energy term for the shape-from-shading based geometry refinement as follows,

$$E_{sfs} = E_{shading} + E_{nearby} + E_{grad} \tag{14}$$

where

$$E_{shading} = \sum_k ||\mathcal{H}_M(\mathcal{L}, n_k(\Delta \vec{v})) - \mathcal{I}_S(p)||^2 \tag{15}$$

where $\vec{v}$ represents the shifting of all the vertexes, E_{nearby} is the same as in Sect. 4.2.3, and E_{grad} is for the constraining updating magnitude. Noticing that our reconstructed human contains two layers, i.e., cloth layer and body layer, therefore, traditional shading-based geometry detail enhancement approaches [114] cannot be directly applied, as geometry refinement on the cloth layer may produce penetration or collision with the body layer. To solve this challenge, we propose an iteration optimization approach, by performing shape-from-shading detail enhancement, with physics-based collision detection and resolving in each iteration step. After each physics-based collision resolving, we guarantee penetration-free reconstruction results.

Fig. 7. The example result of geometry refinement. From left to right: input RGB image, garment geometry before detail refinement, garment geometry with shape refinement, two garment shapes from another point of view, rendered results without shape refinement procedure, our results.

As shown in Fig. 7, without geometry refinement, the garment geometry lacks details. Meanwhile, after geometry refinement, detail wrinkles and folds can be reconstructed from the input image, making the reconstructed geometry more accurate.

In order to improve the ability of geometry detail representation and time consistency of our dynamic garment reconstruction, we place the vertex displacement in its local coordinate system defined by its normal and neighbor vertexes, which is also used in estimating the displacement of invisible vertices. Note that the motion of each vertex can be decomposed into two parts, namely, the global garment motion by body-garment simulation (see Sect. 4.2.3) and the local details that cannot be described by simulation. For the invisible vertices, we assume that their local details remain unchanged. Specifically, we calculate the global rotation R of the vertex from its value in the visible frame according to the simulated normal orientation, and transform the vertex shifting according to the global rotation to maintain its local-coordinate parameters unchanged. For each temporal frame, we finally add a spatial smooth filtering over the boundaries between the visible and invisible regions, to mitigate the spatially inconsistent seam artifacts.

Finally, given a camera model, as the incident lighting, surface albedo and dynamic geometry details have already been obtained, we render the realistic clothed human performances using spherical harmonics rendering models [115].

6 Experiments

In our experiments, we use monocular RGB videos from both the internet and our own cameras containing casual human motions, including walking, playing soccer, speech, exercising, dancing, etc. The human clothing includes pants, trousers, long-sleeve/short-sleeve T-shirt and shirt, which are represented by our designated 2D garment patterns.

Besides human instance parsing and intrinsic decomposition, the main pipeline takes around 12 hours to process a sequence of 300 frames on a 3.4GHz Intel Xeon E3-1231 processor and an NVIDIA GeForce GTX 1070 GPU. Specifically, the pose and shape estimation takes approximately 15 minutes, garment parameter estimation takes 2 hours for 20 iterations of parameter optimization using every key frame and garment deformation refinement takes 10-12 seconds per frame. After obtaining the deformed mesh, the albedo atlas fusion step takes 10 minutes for cloth albedo generation, and geometry refinement takes 100-120 seconds per frame by using cuSPARSE toolkit.

6.1 Qualitative Results

To evaluate our method, Fig. 1, Fig. 8 and the supplemental video provide the reconstruction results of captured sequences from a monocular video camera, which show that our method is capable of generating plausible human performance capture results with detailed wrinkles and folds, as a benefit of the proposed decomposition-based geometry and albedo refinement method. Note that for the two

Fig. 8. Some reconstruction results in our test sequences. Each pair of result contains the original image on the left and the result on the right.

sequences in the bottom row of Fig. 8, the human characters only perform motions facing to the camera view, without capturing his/her back with turning motions. Nevertheless, our method still generates high quality results for these kinds of motions.

As the albedo map and dynamic geometry details for the cloth mesh are maintained during motion, we can generate free-viewpoint rendering results for the clothed human model. Fig. 9 shows the 360-degree free-viewpoint rendering of the human, where the cloth details are distinct in different viewpoints. Note that in the second and the last examples of Fig. 9, the person only shows his front in the whole sequence, but with cloth simulation, we can still render plausible results from the other unseen viewpoints.

6.2 Comparisons

We compared our human performance capture results with [116] and typical template-based deformation methods [38], [39] using a commercial RGBD camera, as shown in Fig. 10 and the supplemental video. The video avatar reconstruction method in [116] takes a single view video of human performance as input, and rectifies all the poses in the image frames to a T-pose for bundle optimization of shape. However, the subject needs to perform the restrictive movement to allow accurate shape reconstruction. So it fails to work for other more generate shapes, poses and dynamic textures, as shown in Fig. 10(a). In contrast, our method works robustly even when subjects perform more casual motions with natural cloth-body interaction and dynamic texture details.

Fig. 10(b) shows the comparison with typical template-based deformation approach [38], [39]. The result on the left is obtained by first fusing the geometry and texture using the DoubleFusion [49] system, followed by skeleton driven non-rigid surface deformation to align with the depth data and the silhouette. As shown in Fig. 10(b) and the video, the texture of such non-rigid reconstruction is static, so it cannot dynamically model changing surface details. In contrast, our method is able to capture the dynamical winkles and produce more plausible garment deformations.

We also make a comparison with a model-based approach [88] on our data. As their method takes only one picture as input, we also take only one picture and feed it into our pipeline for a fair comparison. Notice that [88] generates T-pose garment mesh only. As shown in Fig. 11, regarding to the garment geometry, [88] generates an over-smoothed surface of the garment without detailed wrinkles and folds, and the shape of the cloth does not fit the input image accurately. Meanwhile, our method successfully recovers the geometric details and produce much more realistic rendering results.

We also make a quantitative evaluation on BUFF Dataset [46] and compare MulayCap quantitatively with PIFu [60], which is deep learning method for reconstructing clothed human body from a single image, also without a pre-scanned template. The reconstruction results and the per-vertex average error is shown in Fig. 12. As shown in Fig. 12(a), benefiting from our multi-layer representation of the model and physics-based cloth simulation, we can generate high-frequency details of the cloth, both on the

Fig. 9. Free-viewpoint rendering of different human models. From left to right: input images, reconstructed models from the captured views, and the reconstructed models in two virtual views.

(a) (b)

Fig. 10. Comparison with [116] and [49]. (a) From left to right: [116] result, input frame, our result. (b) From left to right: result by a typical non-rigid surface deformation approach using a commercial depth camera [49], input frame, our result.

Fig. 11. Comparison with [88]. From left to right: input image, rendered garment and garment geometry generated by our method, rendered garment and garment geometry generated by [88]. Notice that [88] generates T-pose clothing output.

front and back. The pose estimated is also consistent with the input image. Meanwhile, although the model generated by PIFu [60] looks plausible from the front view, we can see that it actually generates a wrong pose of the human, also the texture on the back is not so vivid neither realistic. The comparison with PIFu [60] can be regarded as a typical comparison between the model-based methods and data-driven generative methods. Benefiting from other model-based methods like HMMR [57], we can generate more robust and accurate garment results. On the contrary, the implicit representation of PIFu [60] limits its ability of using model-based priors, leading to unrealistic human pose and texture generated.

As for quantitative experiments, we first put the model from both PIFu [60] and MulayCap into a consistent coordinate with the ground truth 3D model of BUFF Dataset [46], and then align the models with the ground truth one using ICP for solving the scale and relative transition of the models. The error is evaluated using the nearest-neighbor L2 distance. Fig. 12(b) shows that the per-vertex error of PIFu [60] is larger than MulayCap in most frames of an input video sequence rendered from BUFF Dataset [46], which shows that with our multi-layer human performance capture method, we can generate more accurate results than the one generated using an end-to-end network.

6.3 Applications

With our proposed multiple-layer modeling for human performance capture, our method produces fully-semantic

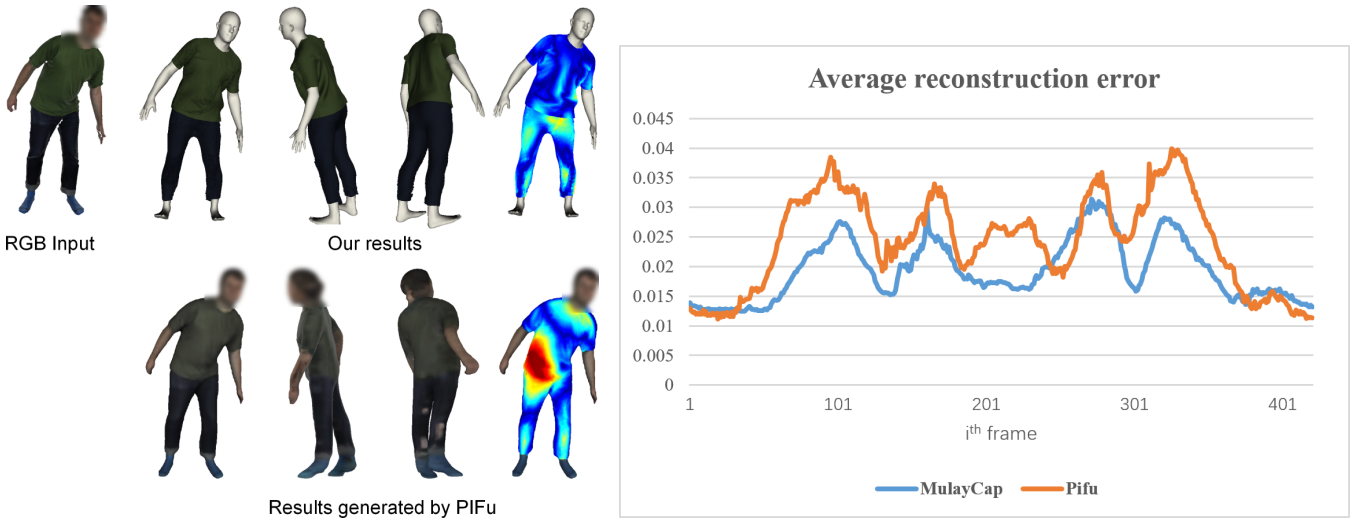


Fig. 12. Qualitative and quantitative comparison with PIFu [60] using rendered 4D model in BUFF Dataset [46] as input. (a) From left to right: rendered human model, reconstruction results from different viewpoints by MulayCap and PIFu [60], error map. (b) Quantitative comparison between two methods in one 4D sequential using per-vertex average error.

reconstruction and enables abundant editing possibilities in the following applications.

Garment Editing. Since in our method semantically models garments on the shape and texture, garment editing in terms of shape or texture can be achieved, as demonstrated in Fig. 13. Garment shape editing (upper row) allows the change of the length parameters of the T-shirt sleeve and the trousers so that the human performance of the same character with new clothing can be obtained. By combining a new albedo color of the cloth with the original shading results, we can render realistic color editing results for the reconstructed human performance as shown in the bottom row.

Retargeting. After the cloth shape and albedo have been generated for a sequence, we can retarget the clothing to other human bodies. Recall that the human models represented by SMPL model, which guarantees topology-consistency between different human models. So we can calculate a non-rigid warp field between the two human bodies with different shapes but the same pose, and adopt this warp field for cloth vertex mapping between the two models. The result is shown in Fig. 14, where two target body shapes are used for the retargeting application.

Relighting. Given albedo and detailed geometry with wrinkles and folds of the garment, we can generate relighting results for the captured sequence. As shown in Fig. 15, we put the character in four different environment illuminations and apply the relighting using spherical harmonic lighting coefficients generated by the cube-map texture. The albedo and geometry details are consistent in different lighting environments.

Augmented Reality. As we can automatically generate 4D human performance with only RGB video, it can be integrated into a real video for VR/AR applications. Given a video sequence of a particular scene as well as the camera positions and orientations in each frame, we can render the human performance at a particular location in the scene. With AR glasses such as Hololens, observers can see human performance in any viewpoint. The examples of such mixed-reality rendering are shown in Fig. 16 and the supplemental video.

Fig. 13. Garment editing results. The upper row is garment shape editing and the bottom row is for garment color editing. From left to right: input RGB frames, reconstructed results, results with shape editing and color editing.

7 Conclusion

In this paper, we present a novel method, called *MulayCap*, based on a multi-layer decomposition of geometry and texture for human performance capture using a single RGB video. Our method can generate novel free-view rendering of vivid cloth details and human motions from a casually captured video, either from the internet or video captured

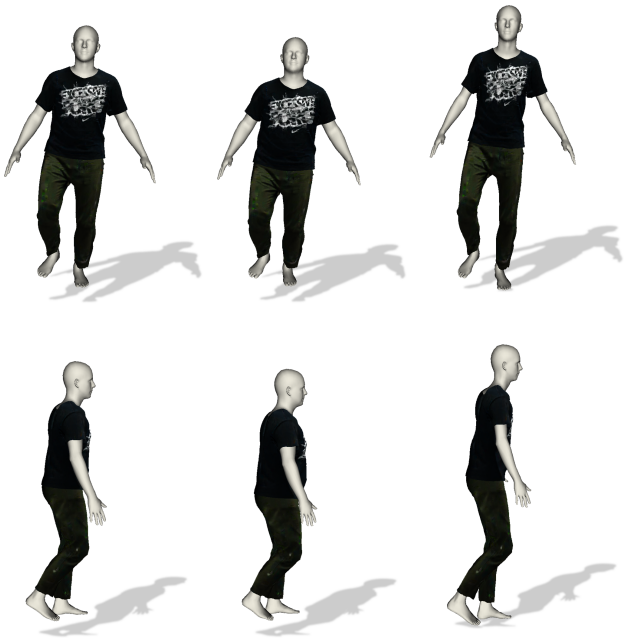


Fig. 14. Clothing retargeting between different human bodies. From left to right: reconstructed clothed human models, and two retargeting results using a taller thin body shape and a fatter body shape.

Fig. 15. Relighting results in four different environmental illumination maps from [117].

by the user. There are three main advantages of MulayCap: (1) it obviates the need for tedious human specific template scanning before real performance capture and still achieved high quality geometry reconstruction on the clothed human performances. This is made possible through the proposed GfV method based on cloth simulation techniques for estimating garment shape parameters by fitting the garment appearance to the input video sequence; (2) MulayCap achieves realistic rendering of the dynamically changing

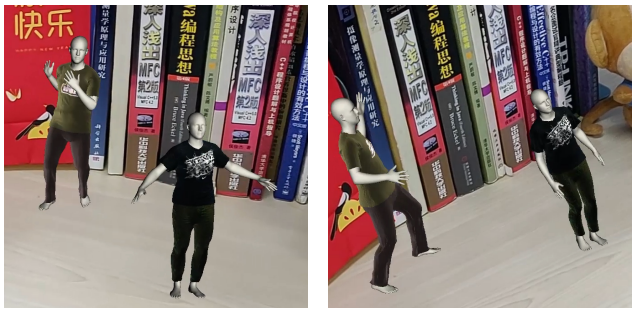


Fig. 16. Two frames of an augmented-reality application. We estimate the camera parameters using [118] and render the clothed human performance on the desk.

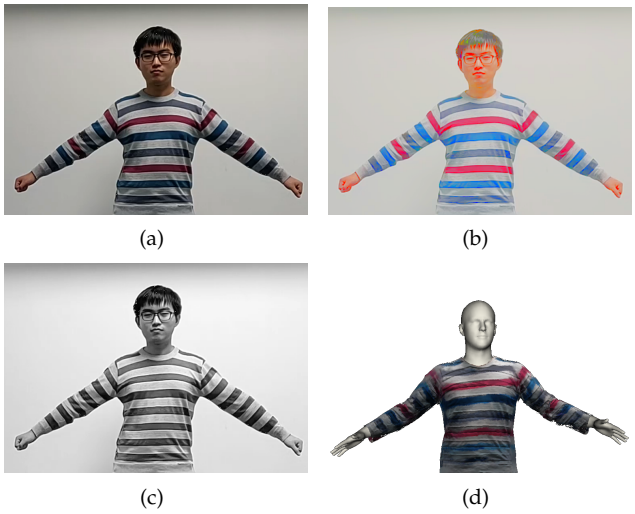


Fig. 17. Illustration of the failure case. (a) The input image. (b) The decomposed albedo image. (c) The decomposed shading image. (d) The rendered result.

details on the garments by using a novel technique of decoupling texture into albedo and shading layers. It is worth noting that such dynamically changing textures have not been demonstrated in any existing monocular human performance capture systems before; finally (3) benefiting from the fully semantic modeling in MulayCap, the reconstructed 4D performance naturally enables various important editing applications, such as cloth editing, re-targeting, relighting, etc.

Limitation and Discussion: MulayCap mainly focuses on the body and garment reconstruction, while the other semantic elements like head, facial expression, hand, skin and shoes would require extra efforts to be handled properly. Another deficiency is that the body motion still suffers from jittering effects, as the body shape parameters are difficult to be accurately *and* smoothly estimated from the video based on the available human shape and pose detection algorithms [107]. As a consequence, we cannot handle fast and extremely challenging motions, as the pose detection on challenging motions contains too many errors for cloth simulation and garment optimization. Also, although our system is robust for common cases, our cloth pattern cannot handle all possible clothes or clothes with non-common shapes.

In addition, in our pipeline, the qualities of the albedo

and shading image are crucial for the final rendering results, which may be affected by the performance of intrinsic decomposition methods to a certain extent. For garments with complex texture patterns such as the lattice T-shirt shown in Fig. 17, existing intrinsic decomposition methods can hardly produce accurate results. In our case, since the shading image extracted by [95] still contains much albedo information, the geometry detail solved by our system is messed with albedo information, as shown in Fig. 17. As most of the existing intrinsic decomposition methods are intended for general scenes, a novel intrinsic decomposition method particularly designed for garments may further improve the shading and albedo estimation in our task.

As for the future work, a more precise human performance capture including hands, skins, shoes, etc., as well as a variety of garment patterns like skirts, coats, etc. are promising directions to be explored. Along with the booming of single image human body estimation research [57], [119], research attentions can be directed on how to achieve jittering-free motion reconstruction to handle more challenging motions. Overall, we believe that our paper may inspire much follow-up research towards improving the quality of convenient and efficient human performance capture using a single monocular video camera, thus facilitating and promoting applications of consumer level human performance capture.

ACKNOWLEDGMENTS

The authors would like to thank Tsinghua University and The Hong Kong University of Science and Technology for supporting this work.

REFERENCES

[1] W. Xu, A. Chatterjee, M. Zollhöfer, H. Rhodin, D. Mehta, H. Seidel, and C. Theobalt, "Monoperfcap: Human performance capture from monocular video," *ACM Trans. Graph.*, vol. 37, no. 2, pp. 27:1–27:15, 2018.

[2] M. Habermann, W. Xu, M. Zollhöfer, G. Pons-Moll, and C. Theobalt, "Reticam: Real-time human performance capture from monocular video," *CoRR*, vol. abs/1810.02648, 2018.

[3] M. Gallardo, T. Collins, A. Bartoli, and F. Mathias, "Dense non-rigid structure-from-motion and shading with unknown albedos," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3884–3892.

[4] G. Li, C. Wu, C. Stoll, Y. Liu, K. Varanasi, Q. Dai, and C. Theobalt, "Capturing relightable human performances under general uncontrolled illumination," *Comput. Graph. Forum*, vol. 32, no. 2, pp. 275–284, 2013.

[5] C. Stoll, J. Gall, E. de Aguiar, S. Thrun, and C. Theobalt, "Video-based reconstruction of animatable human characters," *ACM Trans. Graph.*, vol. 29, no. 6, pp. 139:1–139:10, 2010.

[6] G. Pons-Moll, S. Pujades, S. Hu, and M. J. Black, "Clothcap: Seamless 4d clothing capture and retargeting," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 73:1–73:15, Jul. 2017.

[7] W. Matusik, C. Buehler, R. Raskar, S. J. Gortler, and L. McMillan, "Image-based visual hulls," in *ACM Trans. Graph. (Proc. SIGGRAPH)*, 2000, pp. 369–374.

[8] A. Mustafa, H. Kim, J. Guillemaut, and A. Hilton, "General dynamic scene reconstruction from multiple view video," in *ICCV*, 2015, pp. 900–908.

[9] J. Starck and A. Hilton, "Surface capture for performance-based animation," *IEEE Computer Graphics and Applications*, vol. 27, no. 3, pp. 21–31, 2007.

[10] Y. Liu, Q. Dai, and W. Xu, "A point-cloud-based multiview stereo algorithm for free-viewpoint video," *IEEE Trans. Vis. Comput. Graph.*, vol. 16, no. 3, pp. 407–418, 2010.

[11] M. Waschbüsch, S. Würmlin, D. Cotting, F. Sadlo, and M. H. Gross, "Scalable 3d video of dynamic scenes," *The Visual Computer*, vol. 21, no. 8-10, pp. 629–638, 2005.

[12] D. Vlasic, P. Peers, I. Baran, P. E. Debevec, J. Popovic, S. Rusinkiewicz, and W. Matusik, "Dynamic shape capture using multi-view photometric stereo," *ACM Trans. Graph.*, vol. 28, no. 5, pp. 174:1–174:11, 2009.

[13] D. Casas, M. Volino, J. Collomosse, and A. Hilton, "4d video textures for interactive character appearance," in *Eurographics*, April 2014.

[14] A. Gilbert, M. Volino, J. Collomosse, and A. Hilton, "Volumetric performance capture from minimal camera viewpoints," in *ECCV*, 2018, pp. 591–607.

[15] Z. Huang, T. Li, W. Chen, Y. Zhao, J. Xing, C. LeGendre, L. Luo, C. Ma, and H. Li, "Deep volumetric video from very sparse multi-view performance capture," in *ECCV*, 2018, pp. 351–369.

[16] E. de Aguiar, C. Stoll, C. Theobalt, N. Ahmed, H. Seidel, and S. Thrun, "Performance capture from sparse multi-view video," *ACM Trans. Graph.*, vol. 27, no. 3, pp. 98:1–98:10, 2008.

[17] J. Carranza, C. Theobalt, M. A. Magnor, and H. Seidel, "Free-viewpoint video of human actors," *ACM Trans. Graph.*, vol. 22, no. 3, pp. 569–577, 2003.

[18] D. Vlasic, I. Baran, W. Matusik, and J. Popovic, "Articulated mesh animation from multi-view silhouettes," *ACM Trans. Graph.*, vol. 27, no. 3, pp. 97:1–97:9, 2008.

[19] J. Gall, C. Stoll, E. de Aguiar, C. Theobalt, B. Rosenhahn, and H. Seidel, "Motion capture using joint skeleton tracking and surface estimation," in *IEEE CVPR*, 2009, pp. 1746–1753.

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

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

[22] C. Wu, C. Stoll, L. Valgaerts, and C. Theobalt, "On-set performance capture of multiple actors with a stereo camera," *ACM Trans. Graph.*, vol. 32, no. 6, pp. 161:1–161:11, 2013.

[23] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis, "SCAPE: shape completion and animation of people," *ACM Trans. Graph.*, vol. 24, no. 3, pp. 408–416, 2005.

[24] M. Loper, N. Mahmood, and M. J. Black, "Mosh: motion and shape capture from sparse markers," *ACM Trans. Graph.*, vol. 33, no. 6, pp. 220:1–220:13, 2014.

[25] L. Sigal, A. Balan, and M. J. Black, "Combined discriminative and generative articulated pose and non-rigid shape estimation," in *Advances in neural information processing systems*, 2008, pp. 1337–1344.

[26] N. Hasler, H. Ackermann, B. Rosenhahn, T. Thormählen, and H. Seidel, "Multilinear pose and body shape estimation of dressed subjects from image sets," in *CVPR*, 2010, pp. 1823–1830.

[27] R. Plänkers and P. Fua, "Tracking and modeling people in video sequences," *Computer Vision and Image Understanding*, vol. 81, no. 3, pp. 285–302, 2001.

[28] D. Song, R. Tong, J. Chang, X. Yang, M. Tang, and J. Zhang, "3d body shapes estimation from dressed-human silhouettes," *Comput. Graph. Forum*, vol. 35, no. 7, pp. 147–156, 2016.

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

[30] M. Innmann, M. Zollhöfer, M. Nießner, C. Theobalt, and M. Stamminger, "Volumedeform: Real-time volumetric non-rigid reconstruction," in *ECCV*, 2016, pp. 362–379.

[31] K. Guo, F. Xu, T. Yu, X. Liu, Q. Dai, and Y. Liu, "Real-time geometry, albedo, and motion reconstruction using a single rgb-d camera," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 3, p. 32, 2017.

[32] M. Slavcheva, M. Baust, D. Cremers, and S. Ilic, "Killingfusion: Non-rigid 3d reconstruction without correspondences," in *IEEE CVPR*, 2017, pp. 5474–5483.

[33] M. Slavcheva, M. Baust, and S. Ilic, "SobolevFusion: 3D Reconstruction of Scenes Undergoing Free Non-rigid Motion," in *IEEE CVPR*, 2018.

[34] A. Collet, M. Chuang, P. Sweeney, D. Gillett, D. Evseev, D. Calabrese, H. Hoppe, A. Kirk, and S. Sullivan, "High-quality stream-

able free-viewpoint video," *ACM Trans. Graph*, vol. 34, no. 4, p. 69, 2015.

[35] S. Orts-Escolano, C. Rhemann, S. Fanello, W. Chang, A. Kowdle, Y. Degtyarev, D. Kim, P. L. Davidson, S. Khamis, M. Dou *et al.*, "Holoportation: Virtual 3d teleportation in real-time," in *UIST*. ACM, 2016, pp. 741–754.

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

[37] M. Dou, P. Davidson, S. R. Fanello, S. Khamis, A. Kowdle, C. Rhemann, V. Tankovich, and S. Izadi, "Motion2fusion: real-time volumetric performance capture," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 6, p. 246, 2017.

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

[39] K. Guo, F. Xu, Y. Wang, Y. Liu, and Q. Dai, "Robust non-rigid motion tracking and surface reconstruction using l0 regularization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3083–3091.

[40] G. Ye, Y. Liu, N. Hasler, X. Ji, Q. Dai, and C. Theobalt, "Performance capture of interacting characters with handheld kinects," in *Computer Vision–ECCV 2012*. Springer, 2012, pp. 828–841.

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

[42] T. Helten, M. Muller, H.-P. Seidel, and C. Theobalt, "Real-time body tracking with one depth camera and inertial sensors," in *ICCV*, 2013, pp. 1105–1112.

[43] M. Ye and R. Yang, "Real-time simultaneous pose and shape estimation for articulated objects using a single depth camera," in *CVPR*, 2014, pp. 2345–2352.

[44] F. Bogo, M. J. Black, M. Loper, and J. Romero, "Detailed full-body reconstructions of moving people from monocular rgb-d sequences," in *ICCV*, 2015, pp. 2300–2308.

[45] A. Walsman, W. Wan, T. Schmidt, and D. Fox, "Dynamic high resolution deformable articulated tracking," in *3D Vision (3DV), 2017 International Conference on*. IEEE, 2017, pp. 38–47.

[46] C. Zhang, S. Pujades, M. J. Black, and G. Pons-Moll, "Detailed, accurate, human shape estimation from clothed 3d scan sequences," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[47] T. Yu, K. Guo, F. Xu, Y. Dong, Z. Su, J. Zhao, J. Li, Q. Dai, and Y. Liu, "Bodyfusion: Real-time capture of human motion and surface geometry using a single depth camera," in *IEEE ICCV*, October 2017.

[48] C. Li, Z. Zhao, and X. Guo, "Articulatedfusion: Real-time reconstruction of motion, geometry and segmentation using a single depth camera," in *ECCV*, 2018, pp. 324–340.

[49] T. Yu, Z. Zheng, K. Guo, J. Zhao, Q. Dai, H. Li, G. Pons-Moll, and Y. Liu, "Doublefusion: Real-time capture of human performances with inner body shapes from a single depth sensor," in *IEEE CVPR*, June 2018.

[50] L. Xu, Z. Su, T. Yu, Y. Liu, and L. Fang, "Unstructuredfusion: Realtime 4d geometry and texture reconstruction using commercial rgbd cameras," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2019.

[51] Z. Zheng, T. Yu, H. Li, K. Guo, Q. Dai, L. Fang, and Y. Liu, "Hybridfusion: Real-time performance capture using a single depth sensor and sparse imus," in *The European Conference on Computer Vision (ECCV)*, September 2018.

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

[53] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black, "Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image," in *ECCV 2016*, 2016, pp. 561–578.

[54] G. Pavlakos, L. Zhu, X. Zhou, and K. Daniilidis, "Learning to estimate 3d human pose and shape from a single color image," *arXiv preprint arXiv:1805.04092*, 2018.

[55] M. Omran, C. Lassner, G. Pons-Moll, P. Gehler, and B. Schiele, "Neural body fitting: Unifying deep learning and model based human pose and shape estimation," in *International Conference on 3D Vision (3DV)*, sep 2018.

[56] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik, "End-to-end recovery of human shape and pose," in *IEEE CVPR*, 2018.

[57] A. Kanazawa, J. Zhang, P. Felsen, and J. Malik, "Learning 3d human dynamics from video," *CoRR*, vol. abs/1812.01601, 2018.

[58] N. Kolotouros, G. Pavlakos, M. J. Black, and K. Daniilidis, "Learning to reconstruct 3D human pose and shape via model-fitting in the loop," in *International Conference on Computer Vision*, Oct. 2019.

[59] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. A. Osman, D. Tzionas, and M. J. Black, "Expressive body capture: 3d hands, face, and body from a single image," in *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[60] S. Saito, , Z. Huang, R. Natsume, S. Morishima, A. Kanazawa, and H. Li, "Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization," *arXiv preprint arXiv:1905.05172*, 2019.

[61] B. L. Bhatnagar, G. Tiwari, C. Theobalt, and G. Pons-Moll, "Multi-garment net: Learning to dress 3d people from images," in *IEEE International Conference on Computer Vision (ICCV)*. IEEE, oct 2019.

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

[63] H. Zhu, X. Zuo, S. Wang, X. Cao, and R. Yang, "Detailed human shape estimation from a single image by hierarchical mesh deformation," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[64] S. Tang, F. Tan, K. Cheng, Z. Li, S. Zhu, and P. Tan, "A neural network for detailed human depth estimation from a single image," in *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[65] D. Terzopoulos, J. Platt, A. Barr, and K. Fleischer, "Elastically deformable models," in *ACM Siggraph Computer Graphics*, vol. 21, no. 4, 1987, pp. 205–214.

[66] X. Provot, "Deformation constraints in a mass-spring model to describe rigid cloth behavior," in *IN GRAPHICS INTERFACE*, 1995, pp. 147–154.

[67] K.-J. Choi and H.-S. Ko, "Stable but responsive cloth," in *ACM SIGGRAPH 2005 Courses*, 2005, p. 1.

[68] T. Liu, A. W. Bargteil, J. F. O'Brien, and L. Kavan, "Fast simulation of mass-spring systems," *ACM Transactions on Graphics (TOG)*, vol. 32, no. 6, p. 214, 2013.

[69] J. Bonet and R. D. Wood, *Nonlinear Continuum Mechanics for Finite Element Analysis*. Cambridge: Cambridge University Press, 1997.

[70] C. Jiang, T. Gast, and J. Teran, "Anisotropic elastoplasticity for cloth, knit and hair frictional contact," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 152:1–152:14, Jul. 2017.

[71] W. H. Press, *Numerical recipes 3rd edition: The art of scientific computing*. Cambridge university press, 2007.

[72] D. Baraff and A. Witkin, "Large steps in cloth simulation," in *Proceedings of the 25th annual conference on Computer graphics and interactive techniques*, 1998, pp. 43–54.

[73] S. Bouaziz, S. Martin, T. Liu, L. Kavan, and M. Pauly, "Projective dynamics: Fusing constraint projections for fast simulation," *ACM Trans. Graph.*, vol. 33, no. 4, pp. 154:1–154:11, Jul. 2014.

[74] R. White, K. Crane, and D. A. Forsyth, "Capturing and animating occluded cloth," in *ACM SIGGRAPH 2007 Papers*, ser. SIGGRAPH '07, 2007.

[75] G. J. Brostow, C. Hernández, G. Vogiatzis, B. Stenger, and R. Cipolla, "Video normals from colored lights," *TPAMI*, vol. 33, no. 10, pp. 2104–2114, October 2011.

[76] D. Bradley, T. Popa, A. Sheffer, W. Heidrich, and T. Boubekeur, "Markerless garment capture," *ACM Trans. Graphics (Proc. SIGGRAPH)*, vol. 27, no. 3, p. 99, 2008.

[77] T. Popa, Q. Zhou, D. Bradley, V. Kraevoy, H. Fu, A. Sheffer, and W. Heidrich, "Wrinkling captured garments using space-time data-driven deformation," *Computer Graphics Forum (Proc. Eurographics)*, vol. 28, no. 2, pp. 427–435, 2009.

[78] Z. Lahner, D. Cremers, and T. Tung, "Deepwrinkles: Accurate and realistic clothing modeling," in *ECCV*, September 2018.

[79] X. Chen, B. Zhou, F. Lu, L. Wang, L. Bi, and P. Tan, "Garment modeling with a depth camera," *ACM Trans. Graph.*, vol. 34, no. 6, pp. 203:1–203:12, Oct. 2015.

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

[81] B. Zhou, X. Chen, Q. Fu, K. Guo, and P. Tan, "Garment Modeling from a Single Image," *Computer Graphics Forum*, 2013.

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

[83] S. Yang, Z. Pan, T. Amert, K. Wang, L. Yu, T. Berg, and M. C. Lin, "Physics-inspired garment recovery from a single-view image," *ACM Trans. Graph.*, vol. 37, no. 5, pp. 170:1–170:14, Nov. 2018.

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

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

[86] M. Habermann, W. Xu, , M. Zollhoefer, G. Pons-Moll, and C. Theobalt, "Livecap: Real-time human performance capture from monocular video," *ACM Transactions on Graphics, (Proc. SIGGRAPH)*, jul 2019.

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

[88] Y. Xu, Y. Shanglin, S. Wei, and T. Li, "3d virtual garment modeling from rgb images," in *Mixed and augmented reality (ISMAR), 2019 IEEE international symposium on*. IEEE, 2019.

[89] T. Y. Wang, D. Ceylan, J. Popovic, and N. J. Mitra, "Learning a shared shape space for multimodal garment design," *ACM Trans. Graph.*, vol. 37, no. 6, pp. 1:1–1:14, 2018.

[90] L. Rogge, F. Klose, M. Stengel, M. Eisemann, and M. Magnor, "Garment replacement in monocular video sequences," *ACM Trans. Graph.*, vol. 34, no. 1, pp. 6:1–6:10, Dec. 2014.

[91] J. T. Barron and J. Malik, "Shape, illumination, and reflectance from shading," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 8, pp. 1670–1687, 2015.

[92] Q. Zhao, P. Tan, Q. Dai, L. Shen, E. Wu, and S. Lin, "A closed-form solution to retinex with nonlocal texture constraints," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 7, pp. 1437–1444, 2012.

[93] S. Bell, K. Bala, and N. Snavely, "Intrinsic images in the wild," *ACM Transactions on Graphics (TOG)*, vol. 33, no. 4, p. 159, 2014.

[94] E. Garces, A. Muñoz, J. Lopez-Moreno, and D. Gutierrez, "Intrinsic images by clustering," *Comput. Graph. Forum*, vol. 31, no. 4, pp. 1415–1424, 2012.

[95] T. Nestmeyer and P. V. Gehler, "Reflectance adaptive filtering improves intrinsic image estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, no. 3, 2017, p. 4.

[96] M. Janner, J. Wu, T. D. Kulkarni, I. Yildirim, and J. Tenenbaum, "Self-supervised intrinsic image decomposition," in *Advances in Neural Information Processing Systems*, 2017, pp. 5936–5946.

[97] Z. Li and N. Snavely, "Learning intrinsic image decomposition from watching the world," *arXiv preprint arXiv:1804.00582*, 2018.

[98] G. Ye, E. Garces, Y. Liu, Q. Dai, and D. Gutierrez, "Intrinsic video and applications," *ACM Trans. Graph.*, vol. 33, no. 4, pp. 80:1–80:11, Jul. 2014.

[99] N. Bonneel, K. Sunkavalli, J. Tompkin, D. Sun, S. Paris, and H. Pfister, "Interactive intrinsic video editing," *ACM Transactions on Graphics (TOG)*, vol. 33, no. 6, p. 197, 2014.

[100] A. Meka, M. Zollhöfer, C. Richardt, and C. Theobalt, "Live intrinsic video," *ACM Transactions on Graphics (TOG)*, vol. 35, no. 4, p. 109, 2016.

[101] Y. Matsushita, S. Lin, S. B. Kang, and H.-Y. Shum, "Estimating intrinsic images from image sequences with biased illumination," in *European Conference on Computer Vision*. Springer, 2004, pp. 274–286.

[102] P.-Y. Laffont and J.-C. Bazin, "Intrinsic decomposition of image sequences from local temporal variations," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 433–441.

[103] P.-Y. Laffont, A. Bousseau, and G. Drettakis, "Rich intrinsic image decomposition of outdoor scenes from multiple views," *IEEE transactions on visualization and computer graphics*, vol. 19, no. 2, pp. 210–224, 2013.

[104] S. Duchêne, C. Riant, G. Chaurasia, J. Lopez-Moreno, P.-Y. Laffont, S. Popov, A. Bousseau, and G. Drettakis, "Multi-view intrinsic images of outdoors scenes with an application to relighting," *ACM Transactions on Graphics*, p. 16, 2015.

[105] X. Provot *et al.*, "Deformation constraints in a mass-spring model to describe rigid cloth behaviour," in *Graphics interface*. Canadian Information Processing Society, 1995, pp. 147–147.

[106] K. Gong, X. Liang, Y. Li, Y. Chen, M. Yang, and L. Lin, "Instance-level human parsing via part grouping network," in *ECCV*, 2018, pp. 805–822.

[107] A. Kanazawa, J. Y. Zhang, P. Felsen, and J. Malik, "Learning 3d human dynamics from video," in *Computer Vision and Pattern Recognition (CVPR)*, 2019.

[108] Z. Cao, T. Simon, S. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *IEEE CVPR*, 2017, pp. 1302–1310.

[109] M. Eisemann, B. De Decker, M. Magnor, P. Bekaert, E. De Aguiar, N. Ahmed, C. Theobalt, and A. Sellent, "Floating textures," in *Computer graphics forum*, vol. 27, no. 2. Wiley Online Library, 2008, pp. 409–418.

[110] C. Buehler, M. Bosse, L. McMillan, S. Gortler, and M. Cohen, "Unstructured lumigraph rendering," in *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*. ACM, 2001, pp. 425–432.

[111] V. Lempitsky and D. Ivanov, "Seamless mosaicing of image-based texture maps," in *CVPR*. IEEE, 2007, pp. 1–6.

[112] A. Telea, "An image inpainting technique based on the fast marching method," *Journal of graphics tools*, vol. 9, no. 1, pp. 23–34, 2004.

[113] C. Wu, B. Wilburn, Y. Matsushita, and C. Theobalt, "High-quality shape from multi-view stereo and shading under general illumination," in *CVPR*, 2011, pp. 969–976.

[114] C. Wu, K. Varanasi, Y. Liu, H. Seidel, and C. Theobalt, "Shading-based dynamic shape refinement from multi-view video under general illumination," in *IEEE ICCV*, 2011, pp. 1108–1115.

[115] R. Basri and D. Jacobs, "Lambertian reflectance and linear subspaces," in *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, vol. 2. IEEE, 2001, pp. 383–390.

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

[117] P. Debevec, "Rendering synthetic objects into real scenes: Bridging traditional and image-based graphics with global illumination and high dynamic range photography," in *SIGGRAPH*, 1998, pp. 189–198.

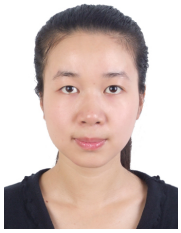
[118] S. Fuhrmann, F. Langguth, and M. Goesele, "Mve – a multi-view reconstruction environment," in *European Conference on Computer Vision (ECCV)*, September 2014.

[119] D. Xiang, H. Joo, and Y. Sheikh, "Monocular total capture: Posing face, body, and hands in the wild," *CoRR*, vol. abs/1812.01598, 2018.

Zhaoqi Su received the B.S. degree in Department of Physics, Tsinghua University, Beijing, China, in 2017. He is currently pursuing the Ph.D. degree in the Department of Automation, Tsinghua University, Beijing, China.

Weilin Wan received the B.S. degree in computer science from the University of Washington, Seattle, USA, in 2018. He is working toward his PhD degree in the Department of Computer Science, the University of Hong Kong. His research interests include computer graphics, machine learning, and robotics.

Tao Yu is a post-doctoral researcher at Tsinghua University. He received the B.S. degree in Measurement and Control from Hefei University of Technology, China, in 2012, and the Ph.D. degree in instrumental science from Beihang University, China. His current research interests include computer vision and computer graphics.



Lingjie Liu is a post-doctoral researcher at the Graphic, Vision & Video group of Max Planck Institute for Informatics in Saarbrücken, Germany. She received the B.E. degree from the Huazhong University of Science and Technology in 2014 and the PhD degree from the University of Hong Kong in 2019. Her research interests include 3D reconstruction, human performance capture and video synthesis. She has received Hong Kong PhD Fellowship Award (2014) and Lise Meitner Fellowship Award (2019).

Lu Fang is currently an Associate Professor at Tsinghua University. She received her Ph.D in Electronic and Computer Engineering from HKUST in 2011, and B.E. from USTC in 2007, respectively. Dr. Fang's research interests include image/video processing, vision for intelligent robot, and computational photography. Dr. Fang serves as TC member in Multimedia Signal Processing Technical Committee (MMSP-TC) in IEEE Signal Processing Society.

Wenping Wang is currently a Chair Professor at the Department of Computer Science, the University of Hong Kong. His research interests include computer graphics, visualization, and geometric computing. He has made fundamental research contributions in collision detection, shape modeling and analysis, mesh generation, and architectural geometry. He is journal associate editor of Computer Aided Geometric Design (CAGD), Computers and Graphics (CAG), IEEE Transactions on Visualization and Computer Graphics (TVCG, 2008-2012), Computer Graphics Forum (CGF), IEEE Computer Graphics and Applications, and IEEE Transactions on Computers. He received the Outstanding Researcher Award of the University of Hong Kong in 2013. He received John Gregory Award in 2017 for contributions in geometric modeling and computing. He is an IEEE Fellow.

Yebin Liu is currently an associate professor at Tsinghua University. He received the B.E. degree from the Beijing University of Posts and Telecommunications, China, in 2002, and the PhD degree from the Automation Department, Tsinghua University, Beijing, China, in 2009. He was a research fellow in the Computer Graphics Group of the Max Planck Institute for Informatik, Germany, in 2010. His research areas include computer vision, computer graphics and computational photography.