Detailed Human Avatars from Monocular Video

Thiemo Alldieck[1,2] Marcus Magnor[1] Weipeng Xu[2] Christian Theobalt[2] Gerard Pons-Moll[2]
[1]Computer Graphics Lab, TU Braunschweig, Germany
[2]Max Planck Institute for Informatics, Saarland Informatics Campus, Germany
{alldieck,magnor}@cg.cs.tu-bs.de {wxu,theobalt,gpons}@mpi-inf.mpg.de

Figure 1: Our method creates a detailed avatar from a monocular video of a person turning around. Based on the SMPL model, we first compute a medium-level avatar, then add subject-specific details and finally generate a seamless texture.

Abstract

We present a novel method for high detail-preserving human avatar creation from monocular video. A parameterized body model is refined and optimized to maximally resemble subjects from a video showing them from all sides. Our avatars feature a natural face, hairstyle, clothes with garment wrinkles, and high-resolution texture. Our paper contributes facial landmark and shading-based human body shape refinement, a semantic texture prior, and a novel texture stitching strategy, resulting in the most sophisticated-looking human avatars obtained from a single video to date. Numerous results show the robustness and versatility of our method. A user study illustrates its superiority over the state-of-the-art in terms of identity preservation, level of detail, realism, and overall user preference.

1. Introduction

The automatic generation of personalized 3D human models is needed for many applications, including virtual and augmented reality, entertainment, teleconferencing, virtual try-on, biometrics or surveillance. A personal 3D human model should comprise all the details that make us different from each other, such as hair, clothing, facial details and shape. Failure to faithfully recover all details results in users not feeling identified with their self-avatar.

To address this challenging problem, researchers have used very expensive recording equipment including 3D and 4D scanners [64, 11, 50] or multi-camera studios with controlled lighting [68, 46]. An alternative is to use passive stereo reconstruction [27, 55] with a camera moving around the person, but the person has to maintain a static pose which is not feasible in practice. Using depth data as input, the field has seen significant progress in reconstructing accurate 3D body models [9, 81, 91] or free-form geometry [95, 53, 59, 24] or both jointly [77]. Depth cameras are however much less ubiquitous than RGB cameras.

Monocular RGB methods are typically restricted to prediciting the parameters of a statistical body model [58, 42, 60, 10, 5, 35]. To the best of our knowledge, the only exception is a recent method [3] that can reconstruct shape, clothing and hair geometry from a monocular video sequence of a person rotating in front of the camera. The basic idea is to fuse the information from frame-wise silhouettes into a canonical pose, and optimize a free-form shape regularized by the SMPL body model [50]. While this is a significant step in 3D human reconstruction from monocular video, the reconstructions are overly smooth, lack facial details and the textures are blurry. This results in avatars that do not fully retain the identity of the real subjects.

In this work, we extend [3] in several important ways to improve the quality of the 3D reconstructions and textures. Specifically, we incorporate information from facial landmark detectors, shape-from-shading, and we introduce a new algorithm to efficiently stitch partial textures coming from frames of the moving person. Since the person is moving, information (projection rays from face landmarks and normal fields from shading cues) can not be directly fused into a single reconstruction. Hence, we track the person's pose using SMPL [50]; then we apply an inverse pose transformation to frame-wise projection rays and normal fields

to fuse all the evidence in a canonical T-pose; in that space, we optimize a high-resolution shape regularized by SMPL. Precisely, with respect to previous work, our approach differs in four important aspects that allow us better preserve subject identity and details in the reconstructions:

Facial landmarks: Since the face is a crucial part of the body, we incorporate 2D facial landmark detections into the 3D reconstruction objective. To gain robustness against misdetections, we fuse temporal detections by transforming the landmark projection rays into the joint T-pose space.

Illumination and shape-from-shading: Shading is a strong cue to recover fine details such as wrinkles. Most shape-from-shading approaches focus on adding detail to static objects. Here, we perform shape-from-shading at every frame, obtaining frame-wise partial 3D normal fields that are then fused in T-pose space for final reconstruction.

Efficient texture stitching: Seamless stitching of partial textures from different camera views is particulary hard for moving articulated objects. To prevent blurry textures, one typically assigns the RGB value of one the views to each texture pixel (texel), while preserving spatial smoothness. Such assignment problem can be formulated as a multi-labeling assignment, where number possible labels grows with the number of views. Consequently, the computational time and memory becomes intractable for for a large number of labels – we define a novel *texture update energy function* which can be minimized efficiently with a graph cut for every new incoming view.

Semantic texture stitching: Aside from stitching artifacts, texture spilling is another common problem. For example texture that corresponds to the clothing often floods into the skin region. To minimize spilling we add an additional semantic term into the texture update energy. The term penalizes updating a texel with an RGB value that is unlikely under a part-based appearance distribution. This semantic appearance term significantly reduces spilling, and implicitly "connects" texels belonging to the same part.

The result is the most sophisticated method to obtain detailed 3D human shape reconstructions from single monocular video. Since metric based evaluations such as scan to mesh distances do not reflect the perceptual quality, we performed a user study to assess the improvement of our method. The results show that users prefer our avatars over state-of-the-art 89.64% of the times and they think our reconstructions are more detailed 95.72% of the times.

2. Related work

Modeling the human body is a long-standing problem in computer vision. Given a densely distributed multi-camera system, one can make use of multi-view stereo methods [43] for reconstructing the human body [27, 29, 40, 93]. More advanced systems allow reconstruction of body shape under clothing [90, 87, 85], joint shape, pose and clothing reconstruction [63], or capture body pose and facial expressions [41]. However, such setups are expensive and require complicated calibration.

Hence, monocular 3D reconstruction methods [54, 55] are appealing but require depth images from many view points around a static object and humans can not hold a static pose for a long time. Therefore, nonrigid deformation of the human body has to be taken into account. Many methods are based on depth sensors and require the subject to hold the same pose. For example, in [48, 20, 71, 89], the subject alternatively makes a certain pose and rotates in front of the sensor. Then, several depth snapshots taken from different view points are fused to generate a complete 3D model. Similarly, [78] proposes to use a turntable to rotate the subject to minimize pose variations. In contrast, the methods of [9, 81, 91] allow a user to move freely in front of the sensor. In recent years, real time nonrigid depth fusion has been achieved [53, 39, 74]. These methods usually maintain a growing template and consist of two alternating steps, i.e. a registration step, where the current template is aligned to the new frame, and a fusion step, where the observation in the new frame is merged to the template. However, these methods typically suffer from "phantom surfaces" artifacts during fast motion. In [77], this problem is alleviated by using SMPL to constraint tracking. Model based monocular methods [10, 23, 32, 5, 35, 66, 65] have recently been integrated with deep learning [58, 42, 60]. However, they are restricted to predicting the parameters of a statistical body model [50, 4, 36, 96, 64]. There are two exceptions, that recover clothing and shape from a single image [33, 18] but these methods require manual initialization of pose and clothing parameters. [3] is the first method capable of reconstructing full 3D shape and clothing geometry from a single RGB video. Users can freely rotate in front of the camera while roughly holding the A-pose. Unfortunately, this approach is restricted to recover only medium-level details. The fine-level details such as garment wrinkles, subtle geometry on the clothes and facial features, which are essential elements for preserving the identity information, are missing. Our goal is to recover the missing fine-level details of the geometry and improve the texture quality such that the appearance identity information can be faithfully recovered.

Another branch of work in human body reconstruction is more focused on capturing the dynamic motion of the character. Works either recover articulated skeletal motion [76, 51, 30, 72, 2, 38], or surfaces with deformed clothing, usually called performance capture. In performance capture many approaches reconstruct a 3D model for each individual frame [75, 47, 19] or fuse a window of frames [59, 24]. However, these methods cannot generate a temporal coherent representation of the model, which is an important characteristic for many applications. To

solve this, methods register a common model to results of all frames [15], use volumetric representation for surface tracking [1, 37], or assume a pre-built static template. Again, most of those methods are based on multi-view images [21, 28, 62, 17, 67, 68]. There are attempts on reducing the number of cameras, such as the stereo method [82], single view depth based method [95] and the recent monocular RGB based method [86]. Note that the result of our method can be used as the initial template for above-mentioned template based performance capture methods.

Shape-from-shading is also highly related to our method. A comprehensive survey can be found in [92]. We only discuss the application of shape-from-shading in the context of human body modeling. Geometric details, e.g. folds in the non-textured region, are difficult to capture with silhouette or photometric information. In contrast, shape-from-shading captures such details [82, 83, 34]. There are also approaches for photometric stereo which recover the shape using controlled light stage setup [79].

Texture generation is an essential task for modeling a realistic virtual character, since a texture image can describe the material properties that cannot be modeled by the surface geometry. The key of a texture generation method is how to combine texture fragments created from different views. Many early works blend the texture fragments using weighted averaging across the entire surface [7, 22, 57, 61]. Others make use of mosaicing strategies, which yields sharper results [6, 45, 56, 69]. [44] is the first to formulate texture stitching as a graph cut problem. Such formulation has been commonly used in texture generation for multi-view 3D reconstruction. However, without accurately reconstructed 3D geometry and registered images, these methods usually suffer from blurring or ghosting artifacts. To this end, many methods focus on compensating registration errors [25, 8, 80, 26, 94]. In our scenario, the registration misalignment problem is even more severe, due to our challenging monocular nonrigid setting. Therefore, we propose to take advantage of semantic information to better constrain our problem.

3. Method

In this paper, our goal is to create a detailed avatar from an RGB video of a subject rotating in front of the camera. The focus lies hereby on fine-level details, that model a subject's identity and individual appearance. As shown in Fig. 2, our method reconstructs a textured mesh model in a coarse-to-fine manner, which consists of three steps: First we estimate a rough body shape of the subject, similar to [3], where the medium-level geometry of the clothing and skin is reconstructed. Then we add fine-level geometric details, such as garment wrinkles and facial features, based on shape-from-shading. Finally, we compute a seamless texture to capture the texel-level appearance details. In the following, we first describe our body shape model, and then discuss the details of our three steps.

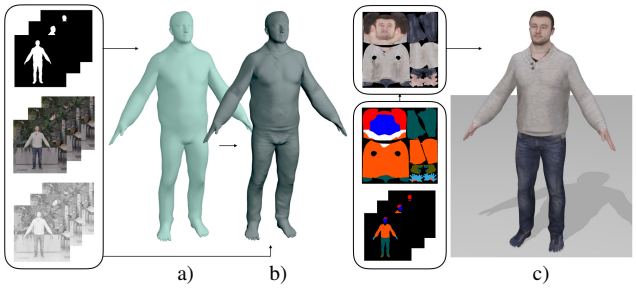

Figure 2. Our method 3-step method: We first estimate a medium level body shape based on segmentations (a), then we add details using shape-from-shading (b). Finally we compute a texture using a semantic prior and a novel graph cut optimization strategy (c).

3.1. Subdivided SMPL body model

Our method is based on the SMPL body model [50]. However, the original SMPL model is too coarse to model fine-level details such as garment wrinkles and fine facial features. To this end, we adapt the model as follows.

The SMPL model is a parameterized human body model described by a function of pose $\boldsymbol{\theta}$ and shape $\boldsymbol{\beta}$ returning $N = 6890$ vertices and $F = 13776$ faces. As SMPL only models naked humans, we use the extended formulation from [3] allowing offsets $\mathbf{D}$ from the template $\mathbf{T}$:

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

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

where W is a linear blend-skinning function applied to a rest pose $T(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{D})$ based on the skeleton joints $J(\boldsymbol{\beta})$ and after pose $B_p(\boldsymbol{\theta})$ and shape dependent $B_s(\boldsymbol{\beta})$ deformations. The inverse function $M^{-1}(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{D})$ *unposes* the model and brings the vertices back into the canonical T-pose. As we aim for fine details and a subject's identity, we further extent the formulation. As shown in Fig. 3, we subdivide every edge of the the SMPL model twice. Every new vertex is defined as:

$$\boldsymbol{v}_{N+e} = 0.5(\boldsymbol{v}_i + \boldsymbol{v}_j) + s_e \boldsymbol{n}_e, \quad (i,j) \in \mathcal{E}_e \quad (3)$$

where $\mathcal{E}$ defines the pairs of vertices forming an edge and $\boldsymbol{n}_e$ is the average normal between the normals of the vertex pair. $s \in \mathbf{s}$ defines the displacement in normal direction $\boldsymbol{n}_e$. $\boldsymbol{n}_e$ is calculated at initialization time in unposed space and can be posed according to W. The new finer model $M_f(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{D}, \mathbf{s})$ consists of $N = 110210$ vertices and $F = 220416$ faces. To recover the high-res smooth surface we calculate an initial set $\mathbf{s}_0 = \{s_0, \ldots, s_e\}$ by minimizing

$$\arg\min_{\mathbf{s}} \Big(\mathbf{L} M_f = \sum_{j \in \mathcal{N}(i)} w_{ij} (\boldsymbol{v}_i - \boldsymbol{v}_j) \Big) \quad (4)$$

where $\mathbf{L}$ is the Laplace matrix with cotangent weights w_{ij} and $\mathcal{N}(i)$ defines the neighbors around $\boldsymbol{v}_i$.

3.2. Medium-level body shape reconstruction

In recent work, a pipeline to recover a subject's body shape, hair and clothing in the same setup as ours has been presented [3]. They first select a number of key-frames ($K \approx 120$) evenly distributed over the sequence and segment them into foreground and background using a CNN [14]. Then they recover the 3D pose for each selected frame based on 2D landmarks [16]. At the core of their method they transform the silhouette cone of every key-frame back into the canonical T-pose of the SMPL model using the inverse formulation of SMPL. This allows efficient optimization of the body shape independent of pose. We follow their pipeline and optimize for the subjects body shape in unposed space. However, we notice that the face estimation of [3] is not accurate enough. This prevents us from further recovering fine-level facial features in the following steps, since precise face alignment is necessary for that. To this end, we propose a new objective for body shape estimation (dependency on parameters removed for clarity):

$$\arg\min_{\boldsymbol{\beta},\mathbf{D}} E_{\text{silh}} + E_{\text{face}} + E_{\text{regm}} \tag{5}$$

The silhouette term E_{silh} measures the distance between boundary vertices and silhouette rays. See [3] for details and regularization E_{regm}. The face alignment term E_{face} penalizes the distance between the 2D facial landmark detections and the 2D projection of 3D facial landmarks. We use OpenPose [73] to detect 2D facial landmarks for every key-frame. In order to incorporate the detections into the method, we establish a static mapping between landmarks and points on the mesh. Every landmark $\boldsymbol{l}$ is mapped to the surface via barycentric interpolation of neighboring vertices. During optimization, we measure the point to line distance between the landmark $\boldsymbol{l}$ on the model and the corresponding camera ray $\mathbf{r}$ describing the 2D landmark detection in unposed space:

$$\delta(\boldsymbol{l}, \mathbf{r}) = \boldsymbol{l} \times \boldsymbol{r}_n - \boldsymbol{r}_m \tag{6}$$

where $\mathbf{r} = (\boldsymbol{r}_m, \boldsymbol{r}_n)$ is given in Plucker coordinates. The face alignment term finally is:

$$E_{\text{face}} = \sum_{l,r \in \mathcal{L}} w_l \rho(\delta(\boldsymbol{l}_l, \mathbf{r}_r)) \tag{7}$$

where $\mathcal{L}$ defines the mapping between mesh points and landmarks, w is the confidence of the landmark given by the CNN and ρ is the Geman-McClure robust cost function. To speed up computation time, we use the coarse SMPL model formulation (Eq. 1) for the medium-level shape estimation.

3.3. Modeling fine-level surface details

In Sec. 3.2, we capture the medium-level details by globally integrating the silhouette information from all key-frames. Now our goal is to obtain fine-level surface details, which cannot be estimated from silhouette, based on shape-from-shading. Note that estimating shape-from-shading globally over all frames would lead to a smooth shape without details, due to fabric movement and misalignments. Thus, we first capture the details for a number of key-frames individually, and then incrementally merge the details into the model as new triangles become visible in a consecutive key-frame. We found that the number of key-frames can be lower than in the first step and choose $K = 60$. Now we describe how to capture the fine-level details for a single key-frame k based on shape-from-shading. To make this process robust, we estimate shading normals individually in a window around the key-frame and then jointly optimize for the surface.

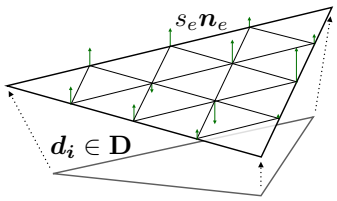

Figure 3. One face of the new SMPL formulation. The displacement field vectors $\boldsymbol{d}_*$ and the normal displacements $s_* \boldsymbol{n}_*$ form the subdivided surface.

Shape-from-shading: For each frame, we first decompose the image into reflectance I_r and shading I_s using the CNN based intrinsic decomposition method of [52]. The function $H_{\boldsymbol{c}}$ calculates the shading of a vertex with spherical harmonic components $\boldsymbol{c}$. We estimate spherical harmonic components $\boldsymbol{c}$ that minimize the difference between the simulated shading and the observed image shading I_s jointly for the given window of frames [84]:

$$\arg\min_{\boldsymbol{c}} \sum_{i \in \mathcal{V}} \left| H_{\boldsymbol{c}}(\boldsymbol{n}_i) - I_s(\mathbf{P}\boldsymbol{v}_i) \right|, \tag{8}$$

where $\mathcal{V}$ denotes the subset of visible vertices, i.e. the angle between the normal and the viewing direction is $0 < \alpha \leq \alpha_{\text{max}}$. $\mathbf{P}$ is the projection matrix. Having the scene illumination and the shading for every pixel, we can now estimate auxiliary normals $\tilde{\mathbf{N}} = \{\tilde{\boldsymbol{n}}_0, \ldots, \tilde{\boldsymbol{n}}_N\}$ for every vertex per frame:

$$\arg\min_{\tilde{\mathbf{N}}} E_{\text{grad}} + w_{\text{lapn}} E_{\text{lapn}}. \tag{9}$$

The Laplacian smoothness term $E_{\text{lapn}} = \mathbf{L}\tilde{\mathbf{N}}$ enforces the normals to be locally smooth. E_{grad} penalizes shading errors by calculating the difference between the gradient between a shaded vertex and its neighbors $\mathcal{N}$ and the image gradient at the projected vertex positions:

$$E_{\text{grad}} = \sum_{i \in \mathcal{V}} \sum_{j \in \mathcal{N}(i) \cap \mathcal{V}} ||\Delta_{H_{\boldsymbol{c}}}(\tilde{\boldsymbol{n}}_i, \tilde{\boldsymbol{n}}_j) - \Delta_{I_s}(\mathbf{P}\boldsymbol{v}_i, \mathbf{P}\boldsymbol{v}_j)||^2 \tag{10}$$

with $\Delta_f(a, b) = f(a) - f(b)$.

Surface reconstruction: In order to merge information about all estimated normals within the window, we trans-

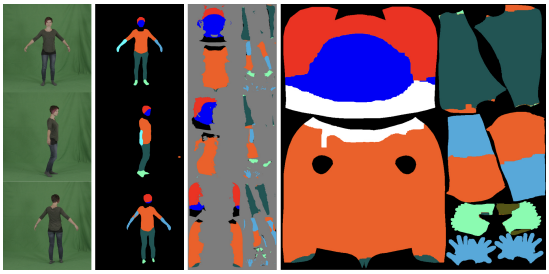

Figure 4. We calculate a semantic segmentation for every key frame. The semantic labels are mapped into texture space and combined into a semantic texture prior.

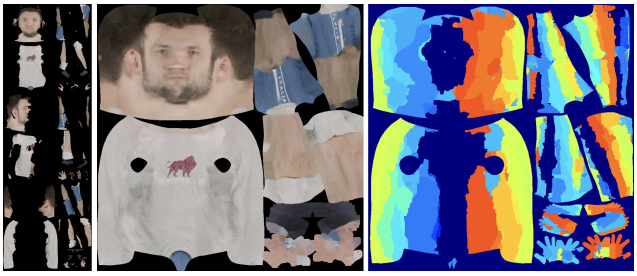

Figure 5. Based on part textures from key frames (left), we stitch a complete texture using graph-cut based optimization. The associated key frames for each texel are shown as colors on the right.

form the normals back into the canonical T-pose using the inverse pose function of SMPL M^{-1}. Then we optimize for the surface which explains the merged normals. Further, we include the silhouette term and face term of Sec. 3.2 to enforce the surface to be well aligned to the images. Specifically, we minimize:

$$\underset{\mathbf{D},\mathbf{s}}{\arg\min} \sum_{j \in \mathcal{C}} \left(\lambda_j E_{\text{silh},j} + \lambda_j w_{\text{face}} E_{\text{face},j}\right) + w_{\text{sfs}} E_{\text{sfs}} + E_{\text{regf}} \quad (11)$$

with weights w_* and $\lambda_j = 1$ for $j = k$ and $\lambda_j < 1$ otherwise. E_{silh} and E_{face} are evaluated over a number of control frames $\mathcal{C}$ and matches in E_{silh} are limited to vertices in the original SMPL model. The shape-from-shading term is defined as:

$$E_{\text{sfs}} = \sum_{f=k-m}^{k+m} \sum_{i \in \mathcal{V}} ||\boldsymbol{n}_i - \tilde{\boldsymbol{n}}_i^f||^2 \quad (12)$$

where k is the current key-frame and m specifies the window size, usually $m = 1$. $\tilde{\boldsymbol{n}}_i^f$ denotes the auxiliary normal of vertex i calculated from frame f. All normals are in T-pose space. E_{regf} regularizes the optimization as described in the following:

$$E_{\text{regf}} = w_{\text{match}} E_{\text{match}} + w_{\text{lap}} E_{\text{lap}} + w_{\text{struc}} E_{\text{struc}} + w_{\text{cons}} E_{\text{cons}} \quad (13)$$

E_{match} penalizes the discrepancy between two neighboring key-frames. Specifically, for a perfect estimation, the following assumption should hold: When warping a key-frame into a neighboring key-frame based on the warp-field described by the projected vertex displacement, the warped frame and the target frame should be similar. E_{match} describes this metric: First we calculate the described warp. Then we calculate warping errors based on optical flow [13]. Based on the sum of the initial warp-field and the calculated error, we establish a grid of correspondences between neighboring key-frames. Every correspondence c should be explained by a particular point of the mesh surface. We first find a candidate for every correspondence:

$$\underset{i \in \mathcal{V}}{\arg\min} \frac{\cos(\alpha_k^i)\delta(\boldsymbol{v}_i^k, \mathbf{r}_c^k) + \cos(\alpha_j^i)\delta(\boldsymbol{v}_i^j, \mathbf{r}_c^j)}{\cos(\alpha_k^i) + \cos(\alpha_j^i)} \quad (14)$$

where α_k^i is the viewing angle under which the vertex i has been seen in key-frame k and $\mathbf{r}_c^k$ is the projection ray of correspondence c in posed space of key-frame k. Then we minimize point to line distance in unposed space:

$$E_{\text{match}} = \sum_{i,c \in \mathcal{M}} \rho(\delta(\boldsymbol{v}_i, \mathbf{r}_c)) \quad (15)$$

where $\mathcal{M}$ is the set of matches established in Eq. 14.

The remaining regularization terms of Eq.13 are as follows: E_{lap} is the Laplacian smoothness term with anisotropic weights [84]. E_{struc} aims to keep the structure of the mesh by pruning edge length variations. E_{cons} prunes large deviations from the consensus shape.

We optimize using a *dog-leg* trust region method using the chumpy autodifferentiation framework. We alternate minimizing and finding silhouette point to line correspondences. Regularization is reduced step-wise.

3.4. Texture generation

A high quality texture image is an essential component for a realistic virtual character, since it can describe the material properties that cannot be modeled by the surface geometry. In order to obtain a sharp and seamless texture, we solve the texture stitching on a per texel level (Fig. 5), in contrast to that on a per face level as in other works [44]. In other words, our goal is to color each pixel in the texture image with a pixel value taken from one out of K key-frames. However, this makes the scale of our problem much larger, and therefore does not allow us to perform global optimization. To this end, we propose a novel texture merging method based on graph cut, which translates our problem to a series of binary labeling subproblems that can be efficiently solved. Furthermore, meshes and key-frames are not perfectly aligned. To reduce color spilling and artifacts caused by misalignments, we compute a semantic prior before stitching the final texture (Fig. 4).

Partial texture generation: For every key-frame, we first project all visible surface points to the frame and write the color at the projected position into the corresponding texture coordinates. In order to factor out the illumination in the texture images, we *unshade* the input images by dividing them with the shading images as used in Sec. 3.3. The partial texture calculation can easily be achieved using the

OpenGL rasterization pipeline. Apart from the partial color texture image, we calculate two additional texture maps for the merging step, i.e. the viewing-angle map and the semantic map. For the viewing-angle map, we compute the viewing angle α_k^t under which the surface point t has been seen in key-frame k.

The semantic prior is generated by re-projecting the human semantic segmentation to the texture space. Specifically, we first calculate a semantic label for every pixel in the input frames using a CNN based human parsing method [49]. Each frame is segmented into 10 semantic classes such as *hair*, *face*, *left leg* and *upper clothes*. Then the semantic information of all frames is fused into the global semantic map by minimizing for labeling $\boldsymbol{x}$:

$$\arg\min_{\boldsymbol{x}} \sum_{t=0}^{T} \varphi_t(x_t) + \sum_{t,q \in \mathcal{N}} \psi(x_t, x_q) \tag{16}$$

$$\varphi_t(x_t) = 1 - \frac{\sum_{k=0}^{K} X_k(\cos^2 \alpha_k^t)}{K} \tag{17}$$

Here φ is the energy term describing the compatibility of a label x with the texel t, where X_k returns the given value if the texel was labeled with x in view k and 0 otherwise. ψ gives the label compatibility of neighboring texels t and q. We solve Eq. 16 by multi-label graph-cut optimization with alpha-beta swaps [12]. While constructing the graph, we connect every texel not only with its neighbors in texture space but with all neighbors on the surface. In particular this means texels are connected across texture seams. To have a strong prior for the texture completion, we calculate Gaussian mixture models (GMM) of the colors in HSV space per label using the part-textures and corresponding labels.

Texture merging: Next, we calculate the complete texture by merging the partial textures. While keeping the same graph structure, the objective function is:

$$\arg\min_{\boldsymbol{u}} \sum_{t=0}^{T} \theta_t(u_t) + \sum_{t,q \in \mathcal{N}} \eta_{t,q}(u_t, u_q) \tag{18}$$

where the labeling $\boldsymbol{u}$ assigns every texel to a partial texture k. The first term seeks to find the best image for each texel:

$$\begin{aligned} \theta_t(k) = w_{\text{vis}} \sin^2 \alpha_k^t + w_{\text{gmm}} m(\mathbf{U}_k^t, x_t) \\ + w_{\text{face}} d(\mathbf{U}_k^t) + w_{\text{silh}} E_{\text{silh},k} \end{aligned} \tag{19}$$

with weights w_*. m returns the Mahalanobis distance between the color value for t in part-texture k given the semantic label x_t. d calculates the structural dissimilarity between the first and the given key-frame. d is only evaluated on texels belonging to the facial region and ensures consistent facial expression over the texture.

The smoothness-term η ensures similar colors for neighboring texels. For neighboring texels assigned to different key-frames $u_t \neq u_q$, while belonging to the same semantic region $x_t = x_q$, $\eta_{t,q}$ equals the gradient magnitude between the texel colors $||\mathbf{U}_{u_t}^t - \mathbf{U}_{u_q}^q||$.

Figure 6. Side-by-side comparisons of our reconstructions (b) and the input frame (a). As can be seen from (b), our method closely resembles the subject in the video (a).

Since the number of combinations in η is very high, it is computationally not feasible to solve Eq. 18 as a multi label graph-cut problem. Thus, we propose the following strategy for an approximate solution: We convert the multi-label problem to a binary labeling decision $b \in \{update, keep\}$. We initialize the texture with $\mathbf{M} = \mathbf{U}_0$. Then we randomly choose a key-frame k and test it against the current solution. The likelihood of selecting a key-frame is inversly proportional to its remaining silhouette error $E_{\text{silh},k}$ in order to favor well-aligned key-frames. Further, η is approximated with:

$$\eta_{t,q} = \begin{cases} \max(||\mathbf{M}^t - \mathbf{U}_k^q||, ||\mathbf{M}^q - \mathbf{U}_k^t||), & \text{if } b_t \neq b_q \wedge x_t = x_q \\ 0, & \text{otherwise} \end{cases} \tag{20}$$

Convergence is usually reached between $2K$ to $3K$ iterations. Finally, we cross-blend between different labels to reduce visible seams. The run-time per iteration on 1000×1000 px with Python code using a standard graph cut library is $\sim$2 sec. No attempts for run-time optimization have been made.

4. Experiments

We evaluate our method on two publicly available datasets: The People-Snapshot dataset [3] and the dataset used in [9]. To validate the perceived quality of our results we performed a user study.

4.1. Qualitative results and comparisons

We compare our method to the recent method of [3] on their People-Snapshot dataset. The approach of [3] is the only other monocular 3D person reconstruction method. The People-Snapshot dataset consists of 24 sequences of different subjects rotating in front of the camera while roughly holding an A-pose. In Fig. 6, we show some examples of our reconstruction results, which precisely overlay the subjects in the image. Note that the level of detail of

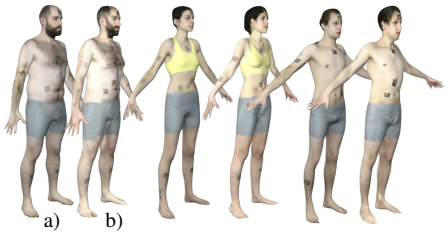

Figure 7. Our results (b) in comparison against the RGB-D method [9] (a). Note that the texture prior has not been used (see Sec. 4.1).

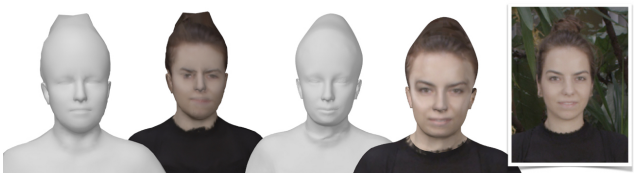

Figure 8. In comparison to the method of [3] (left), the faces in our results (right) have finer details in the mesh and closely resemble the subject in the photograph.

the input images is captured by our reconstructed avatars. In Fig. 11, we show side-by-side comparison to [3]. Our results (right) reconstruct the face better and preserve many more details, e.g. clothing wrinkles and t-shirt stamps.

Additionally, we compare against the state-of-the-art RGB-D method [9], also using their dataset of people in minimal clothing[1]. While their method relies on depth data, we only use the RGB video which makes the problem much harder. Despite this, as shown in Fig. 7, our results are comparable in quality to theirs.

4.2. Face similarity measure

One goal of our method was to preserve the individual appearance of subjects in their avatars. Since the face is crucial for this, we leverage facial landmarks detections and shape-from-shading. As seen in Fig. 11 our method adds a significant level of detail to the facial region in comparison to state-of-the-art. In Fig. 8 we show the same comparison also for untextured meshes. Our result closely resembles the subject in the photograph. To further demonstrate the effectiveness of our method for face similarity preservation, we perform the following experiment: FaceNet [70] is a deep network, that is trained to map from face images to an Euclidean space where distance corresponds to face similarity. We use FaceNet trained on the CASIA WebFace dataset [88] to measure the similarity between photos of the subjects in the People-snapshot dataset and their reconstructions. Two distinct subjects in the dataset have a mean similarity distance of 1.33 ± 0.13. Same subjects in different settings differ by 0.55 ± 0.18. Our reconstructions feature a mean distance of 0.99 ± 0.11 to their photo counterparts. Reconstructions of [3] perform significantly worse with a mean distance of 1.09 ± 0.15. While our reconstructions can be reliable identified using FaceNet, reconstructions of [3] have a similarity distance close to a distance of distinct people, making them less likely to be identified correctly.

[1]The deep learning based segmentation [31] only works for fully clothed people so we had to deactivate the semantic prior in this dataset.

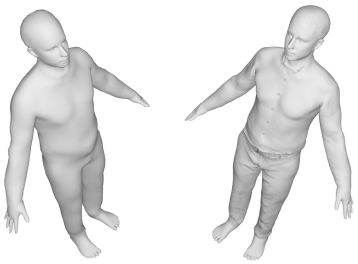

Figure 9. Comparison of a result of our method before (left) and after (right) applying shape-from-shading based detail enhancing.

Figure 10. The semantic prior for texture stitching successfully removes color spilling (left) in our final texture (right).

4.3. Ablation analysis

In the following we qualitatively demonstrate the effectiveness of further design choices of our method.

Shape-from-shading: In order to render the avatars under different illuminations, detailed geometry should be present in the mesh. In Fig. 9, we demonstrate the level of detail added to the meshes by shape-from-shading. While the mesh on the left only describes the low-frequency shape, our refined result on the right contains fine-grained details such as wrinkles and buttons.

Influence of the texture prior: In Fig. 10 we show the effectiveness of the semantic prior for texture stitching. While the texture on the left computed without the prior contains noticeable color spills on the arms and hands, the final texture on the right contains no color spills and less stitching artifacts along semantic boundaries.

4.4. User study

Finally, we conducted a user study in order to validate the visual fidelity of our results. Each participant was asked four questions about 6 randomly chosen results out of the 24 reconstructed subjects in People-Snapshot dataset. The avatars shown to each participant and the questions asked were randomized. In every question the participants had to decide between our method, and the method of [3]. The four question were:

– Which avatar preserves the identity of the person in the image better? (*identity*)

	Identitiy	Details	Realism	Preference
Textured Avatars	83.12 %	-	92.27 %	89.64 %
Untextured Avatars	65.70 %	95.72 %	89.73 %	-

Table 1. Results of the user study. Percentage of answers where users preferred our method over [3]. We asked for four different aspects. See Sec. 4.4 for details.

– Which avatar has more detail? (*detail*)
– Which avatar looks more real to you? (*realism*)
– Which avatar do you like better? (*preference*)

We presented the users renderings of the meshes in consistent pose and illumination. The users were allowed to zoom into the images. At questions *identity* and *realism* we showed the participants either textured or untextured meshes. For *identity* comparison we additionally showed a photo of the subject next to the renderings. When asking for *detail* we only showed untextured meshes, and when asking for *preference* we only showed textured results. Additionally, we asked for the level of experience with 3D data (*None*, *Beginner*, *Proficient*, *Expert*). 74 people participated in our online survey, covering the whole range of expertise.

The results of the study are summarized in Table 1. The participants clearly preferred our results in all scenarios over current state-of-the-art. Admittedly, when asked about identity preservation in untextured meshes, users preferred our method, but this time only 65.70%. Further inspection of the results shows that users with high experience with 3D data think our method preserves the identity better with 90.48% versus 60.49% for novice users. We hypothesize that unexperienced users find it more difficult to recognize people from 3D meshes without textures. Most importantly, by a large margin, our results are perceived as more realistic (92.27%), preserve more details (95.72%) and where preferred 89.64% of the times.

5. Discussion and Conclusion

We have proposed a novel method to create highly detailed personalized avatars from monocular video. We improve over the state-of-the-art in several important aspects: Our optimization scheme allows to integrate face landmark detections and shape-from-shading from multiple frames. Experiments demonstrate that this results in better face reconstruction and better identity preservation. This is also confirmed by our user study, which shows that people think our method preserves identity better 83.12% of the times, and capture more details 95.72% of the times.

We introduced a new texture stitching binary optimization, which allows us to efficiently merge the appearance of multiple frames into a single coherent texture. The optimization includes a semantic texture term that incorporates appearance models for each semantic segmentation part. Results demonstrate that the common artifact of color spilling from skin to clothing or viceversa gets reduced.

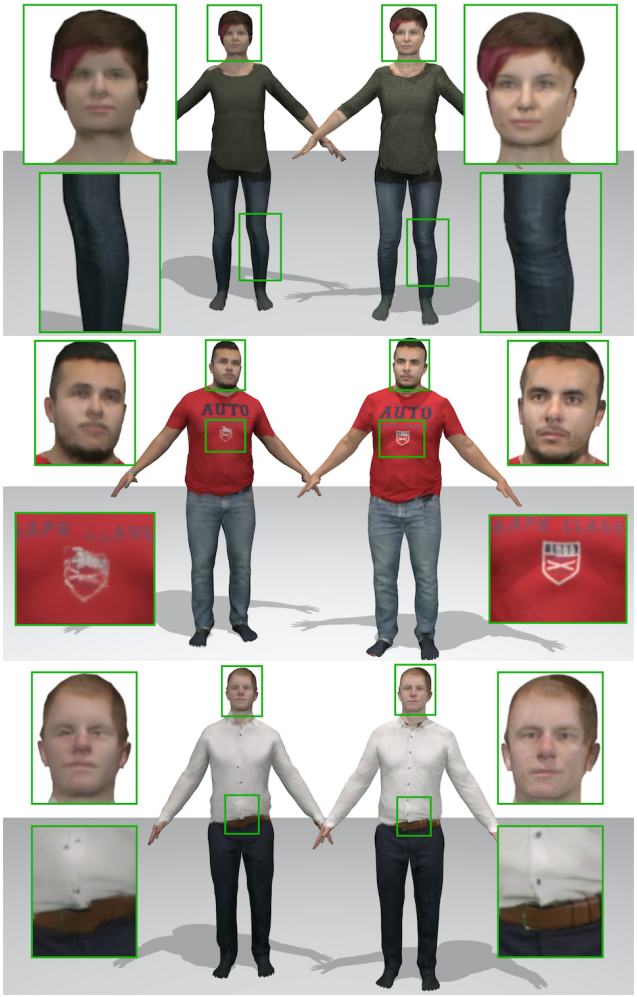

Figure 11. In comparison to the method of [3] (left), our results (right) look much more natural and have finer details.

We have argued for a method to capture the subtle, but very important details to make avatars look *realistic*. Indeed *details matter*, the user study shows that users think our results are more realistic than the state of the art 92.7% of the times, and prefer our avatars 89.64% of the times.

Future work should address capture of subjects wearing clothing with topology different from the body, including skirts and coats. Furthermore, to obtain full texturing, subjects have to be seen from all sides – it may be possible to infer occluded appearance using sufficient training data. Another avenue to explore is reconstruction in an uncooperative setting, e.g. from online videos of people.

Having cameras all around us, we can now serve the growing demand for personalized avatars in virtual and augmented reality applications e.g. in the fields of entertainment, communication or e-commerce.

Acknowledgments: The authors gratefully acknowledge funding by the German Science Foundation from project DFG MA2555/12-1.

References

[1] B. Allain, J.-S. Franco, and E. Boyer. An Efficient Volumetric Framework for Shape Tracking. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 268–276, Boston, United States, 2015. IEEE. 3

[2] T. Alldieck, M. Kassubeck, B. Wandt, B. Rosenhahn, and M. Magnor. Optical flow-based 3d human motion estimation from monocular video. In *German Conf. on Pattern Recognition*, pages 347–360, 2017. 2

[3] T. Alldieck, M. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll. Video based reconstruction of 3d people models. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2018. 1, 2, 3, 4, 6, 7, 8

[4] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis. SCAPE: shape completion and animation of people. In *ACM Transactions on Graphics*, volume 24, pages 408–416. ACM, 2005. 2

[5] A. O. Bălan and M. J. Black. The naked truth: Estimating body shape under clothing. In *European Conf. on Computer Vision*, pages 15–29. Springer, 2008. 1, 2

[6] A. Baumberg. Blending images for texturing 3d models. In *British Machine Vision Conference*, volume 3, page 5. Citeseer, 2002. 3

[7] F. Bernardini, I. M. Martin, and H. Rushmeier. High-quality texture reconstruction from multiple scans. *IEEE Transactions on Visualization and Computer Graphics*, 7(4):318–332, 2001. 3

[8] S. Bi, N. K. Kalantari, and R. Ramamoorthi. Patch-based optimization for image-based texture mapping. *ACM Transactions on Graphics*, 36(4), 2017. 3

[9] F. Bogo, M. J. Black, M. Loper, and J. Romero. Detailed full-body reconstructions of moving people from monocular RGB-D sequences. In *IEEE International Conf. on Computer Vision*, pages 2300–2308, 2015. 1, 2, 6, 7

[10] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *European Conf. on Computer Vision*. Springer International Publishing, 2016. 1, 2

[11] F. Bogo, J. Romero, G. Pons-Moll, and M. J. Black. Dynamic FAUST: Registering human bodies in motion. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2017. 1

[12] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(11):1222–1239, 2001. 6

[13] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert. High accuracy optical flow estimation based on a theory for warping. In *European Conf. on Computer Vision*, pages 25–36. Springer, 2004. 5

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

[15] C. Cagniart, E. Boyer, and S. Ilic. Probabilistic deformable surface tracking from multiple videos. In K. Daniilidis, P. Maragos, and N. Paragios, editors, *European Conf. on Computer Vision*, volume 6314 of *Lecture Notes in Computer Science*, pages 326–339, Heraklion, Greece, 2010. Springer. 3

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

[17] J. Carranza, C. Theobalt, M. A. Magnor, and H.-P. Seidel. Free-viewpoint video of human actors. In *ACM Transactions on Graphics*, volume 22, pages 569–577. ACM, 2003. 3

[18] X. Chen, Y. Guo, B. Zhou, and Q. Zhao. Deformable model for estimating clothed and naked human shapes from a single image. *The Visual Computer*, 29(11):1187–1196, 2013. 2

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

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

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

[22] P. E. Debevec, C. J. Taylor, and J. Malik. Modeling and rendering architecture from photographs: A hybrid geometry- and image-based approach. In *Annual Conf. on Computer Graphics and Interactive Techniques*, pages 11–20. ACM, 1996. 3

[23] E. Dibra, H. Jain, C. Oztireli, R. Ziegler, and M. Gross. Human shape from silhouettes using generative hks descriptors and cross-modal neural networks. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 4826–4836, 2017. 2

[24] M. Dou, S. Khamis, Y. Degtyarev, P. Davidson, S. R. Fanello, A. Kowdle, S. O. Escolano, C. Rhemann, D. Kim, J. Taylor, et al. Fusion4d: Real-time performance capture of challenging scenes. *ACM Transactions on Graphics*, 35(4):114, 2016. 1, 2

[25] 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*, volume 27, pages 409–418. Wiley Online Library, 2008. 3

[26] Y. Fu, Q. Yan, L. Yang, J. Liao, and C. Xiao. Texture mapping for 3d reconstruction with rgb-d sensor. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2018. 3

[27] S. Fuhrmann, F. Langguth, and M. Goesele. Mve-a multi-view reconstruction environment. In *Eurographics Workshops on Graphics and Cultural Heritage*, pages 11–18, 2014. 1, 2

[28] J. Gall, C. Stoll, E. De Aguiar, C. Theobalt, B. Rosenhahn, and H.-P. Seidel. Motion capture using joint skeleton tracking and surface estimation. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1746–1753. IEEE, 2009. 3

[29] S. Galliani, K. Lasinger, and K. Schindler. Massively parallel multiview stereopsis by surface normal diffusion. In

IEEE International Conf. on Computer Vision, pages 873–881, 2015. 2

[30] D. M. Gavrila and L. S. Davis. 3-d model-based tracking of humans in action: a multi-view approach. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 73–80. IEEE, 1996. 2

[31] K. Gong, X. Liang, X. Shen, and L. Lin. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2017. 7

[32] P. Guan, A. Weiss, A. O. Bălan, and M. J. Black. Estimating human shape and pose from a single image. In *IEEE International Conf. on Computer Vision*, pages 1381–1388. IEEE, 2009. 2

[33] Y. Guo, X. Chen, B. Zhou, and Q. Zhao. Clothed and naked human shapes estimation from a single image. *Computational Visual Media*, pages 43–50, 2012. 2

[34] B. Haefner, Y. Quéau, T. Möllenhoff, and D. Cremers. Fight ill-posedness with ill-posedness: Single-shot variational depth super-resolution from shading. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 164–174, 2018. 3

[35] N. Hasler, H. Ackermann, B. Rosenhahn, T. Thormahlen, and H.-P. Seidel. Multilinear pose and body shape estismation of dressed subjects from image sets. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1823–1830. IEEE, 2010. 1, 2

[36] N. Hasler, C. Stoll, M. Sunkel, B. Rosenhahn, and H.-P. Seidel. A statistical model of human pose and body shape. In *Computer Graphics Forum*, volume 28, pages 337–346, 2009. 2

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

[38] Y. Huang, F. Bogo, C. Classner, A. Kanazawa, P. V. Gehler, I. Akhter, and M. J. Black. Towards accurate markerless human shape and pose estimation over time. In *International Conf. on 3D Vision*, 2017. 2

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

[40] M. Jancosek and T. Pajdla. Multi-view reconstruction preserving weakly-supported surfaces. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 3121–3128. IEEE, 2011. 2

[41] H. Joo, T. Simon, and Y. Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8320–8329, 2018. 2

[42] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik. End-to-end recovery of human shape and pose. In *IEEE Conf. on Computer Vision and Pattern Recognition*. IEEE Computer Society, 2018. 1, 2

[43] R. Koch, M. Pollefeys, and L. Van Gool. Multi viewpoint stereo from uncalibrated video sequences. In *European conference on computer vision*, pages 55–71. Springer, 1998. 2

[44] V. Lempitsky and D. Ivanov. Seamless mosaicing of image-based texture maps. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1–6. IEEE, 2007. 3, 5

[45] H. P. Lensch, W. Heidrich, and H.-P. Seidel. A silhouette-based algorithm for texture registration and stitching. *Graphical Models*, 63(4):245–262, 2001. 3

[46] V. Leroy, J.-S. Franco, and E. Boyer. Multi-view dynamic shape refinement using local temporal integration. In *IEEE International Conf. on Computer Vision*, 2017. 1

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

[48] H. Li, E. Vouga, A. Gudym, L. Luo, J. T. Barron, and G. Gusev. 3d self-portraits. *ACM Transactions on Graphics*, 32(6):187, 2013. 2

[49] X. Liang, S. Liu, X. Shen, J. Yang, L. Liu, J. Dong, L. Lin, and S. Yan. Deep human parsing with active template regression. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 37(12):2402–2414, Dec 2015. 6

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

[51] D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei, H.-P. Seidel, W. Xu, D. Casas, and C. Theobalt. Vnect: Real-time 3d human pose estimation with a single rgb camera. *ACM Transactions on Graphics*, 36(4):44, 2017. 2

[52] T. Nestmeyer and P. V. Gehler. Reflectance adaptive filtering improves intrinsic image estimation. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1771–1780. IEE, 2017. 4

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

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

[55] R. A. Newcombe, S. J. Lovegrove, and A. J. Davison. Dtam: Dense tracking and mapping in real-time. In *IEEE International Conf. on Computer Vision*, pages 2320–2327, 2011. 1, 2

[56] W. Niem and J. Wingbermuhle. Automatic reconstruction of 3d objects using a mobile monoscopic camera. In *Proc. Inter. Conf. on Recent Advances in 3-D Digital Imaging and Modeling*, pages 173–180. IEEE, 1997. 3

[57] E. Ofek, E. Shilat, A. Rappoport, and M. Werman. Multiresolution textures from image sequences. *IEEE Computer Graphics and Applications*, 17(2):18–29, Mar. 1997. 3

[58] 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 Conf. on 3D Vision*, 2018. 1, 2

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

[60] G. Pavlakos, L. Zhu, X. Zhou, and K. Daniilidis. Learning to estimate 3D human pose and shape from a single color image. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2018. 1, 2

[61] F. Pighin, J. Hecker, D. Lischinski, R. Szeliski, and D. H. Salesin. Synthesizing realistic facial expressions from photographs. In *ACM SIGGRAPH 2006 Courses*, page 19. ACM, 2006. 3

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

[63] G. Pons-Moll, S. Pujades, S. Hu, and M. Black. ClothCap: Seamless 4D clothing capture and retargeting. *ACM Transactions on Graphics*, 36(4), 2017. 2

[64] G. Pons-Moll, J. Romero, N. Mahmood, and M. J. Black. Dyna: a model of dynamic human shape in motion. *ACM Transactions on Graphics*, 34:120, 2015. 1, 2

[65] G. Pons-Moll and B. Rosenhahn. *Model-Based Pose Estimation*, chapter 9, pages 139–170. Springer, 2011. 2

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

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

[68] N. Robertini, D. Casas, E. De Aguiar, and C. Theobalt. Multi-view performance capture of surface details. *International Journal of Computer Vision*, pages 1–18, 2017. 1, 3

[69] C. Rocchini, P. Cignoni, C. Montani, and R. Scopigno. Multiple textures stitching and blending on 3d objects. In *Rendering Techniques 99*, pages 119–130. Springer, 1999. 3

[70] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 815–823, 2015. 7

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

[72] L. Sigal, S. Bhatia, S. Roth, M. J. Black, and M. Isard. Tracking loose-limbed people. In *IEEE Conf. on Computer Vision and Pattern Recognition*, volume 1, pages I–421. IEEE, 2004. 2

[73] T. Simon, H. Joo, I. Matthews, and Y. Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2017. 4

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

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

[76] C. Stoll, N. Hasler, J. Gall, H.-P. Seidel, and C. Theobalt. Fast articulated motion tracking using a sums of gaussians body model. In *IEEE International Conf. on Computer Vision*, pages 951–958. IEEE, 2011. 2

[77] Y. Tao, Z. Zheng, K. Guo, J. Zhao, D. Quionhai, H. Li, G. Pons-Moll, and Y. Liu. Doublefusion: Real-time capture of human performance with inner body shape from a depth sensor. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2018. 1, 2

[78] J. Tong, J. Zhou, L. Liu, Z. Pan, and H. Yan. Scanning 3d full human bodies using kinects. *IEEE Transactions on Visualization and Computer Graphics*, 18(4):643–650, 2012. 2

[79] D. Vlasic, P. Peers, I. Baran, P. Debevec, J. Popović, S. Rusinkiewicz, and W. Matusik. Dynamic shape capture using multi-view photometric stereo. In *ACM Transactions on Graphics*, volume 28, page 174. ACM, 2009. 3

[80] M. Waechter, N. Moehrle, and M. Goesele. Let there be color! large-scale texturing of 3d reconstructions. In *European Conf. on Computer Vision*, pages 836–850. Springer, 2014. 3

[81] A. Weiss, D. Hirshberg, and M. J. Black. Home 3d body scans from noisy image and range data. In *IEEE International Conf. on Computer Vision*, pages 1951–1958. IEEE, 2011. 1, 2

[82] C. Wu, C. Stoll, L. Valgaerts, and C. Theobalt. On-set performance capture of multiple actors with a stereo camera. *ACM Transactions on Graphics*, 32(6):161, 2013. 3

[83] C. Wu, K. Varanasi, Y. Liu, H.-P. Seidel, and C. Theobalt. Shading-based dynamic shape refinement from multi-view video under general illumination. In *IEEE International Conf. on Computer Vision*, pages 1108–1115. IEEE, 2011. 3

[84] C. Wu, B. Wilburn, Y. Matsushita, and C. Theobalt. High-quality shape from multi-view stereo and shading under general illumination. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 969–976, 2011. 4, 5

[85] S. Wuhrer, L. Pishchulin, A. Brunton, C. Shu, and J. Lang. Estimation of human body shape and posture under clothing. *Computer Vision and Image Understanding*, 127:31–42, 2014. 2

[86] W. Xu, A. Chatterjee, M. Zollhoefer, H. Rhodin, D. Mehta, H.-P. Seidel, and C. Theobalt. Monoperfcap: Human performance capture from monocular video. *ACM Transactions on Graphics*, 2018. 3

[87] J. Yang, J.-S. Franco, F. Hétroy-Wheeler, and S. Wuhrer. Estimation of Human Body Shape in Motion with Wide Clothing. In *European Conf. on Computer Vision*, Amsterdam, Netherlands, 2016. 2

[88] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014. 7

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

[90] C. Zhang, S. Pujades, M. Black, and G. Pons-Moll. Detailed, accurate, human shape estimation from clothed 3D scan sequences. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2017. 2

[91] Q. Zhang, B. Fu, M. Ye, and R. Yang. Quality dynamic human body modeling using a single low-cost depth camera. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 676–683. IEEE, 2014. 1, 2

[92] R. Zhang, P.-S. Tsai, J. E. Cryer, and M. Shah. Shape-from-shading: a survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(8):690–706, 1999. 3

[93] E. Zheng, E. Dunn, V. Jojic, and J.-M. Frahm. Patchmatch based joint view selection and depthmap estimation. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1510–1517, 2014. 2

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

[95] M. Zollhöfer, M. Nießner, S. Izadi, C. Rehmann, C. Zach, M. Fisher, C. Wu, A. Fitzgibbon, C. Loop, C. Theobalt, et al. Real-time non-rigid reconstruction using an rgb-d camera. *ACM Transactions on Graphics*, 33(4):156, 2014. 1, 3

[96] S. Zuffi and M. J. Black. The stitched puppet: A graphical model of 3d human shape and pose. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 3537–3546. IEEE, 2015. 2