

Learning to Reconstruct People in Clothing from a Single RGB Camera

Thiemo Alldieck^{1,2,*} Marcus Magnor¹ Bharat Lal Bhatnagar²
 Christian Theobalt² Gerard Pons-Moll²

¹Computer Graphics Lab, TU Braunschweig, Germany

²Max Planck Institute for Informatics, Saarland Informatics Campus, Germany

{alldieck,magnor}@cg.cs.tu-bs.de {bbhatnag,theobalt,gpons}@mpi-inf.mpg.de



Figure 1: We present a deep learning based approach to estimate personalized body shape, including hair and clothing, using a single RGB camera. The shapes shown above have been calculated using only 8 input images, and re-posed using SMPL.

Abstract

We present *Octopus*, a learning-based model to infer the personalized 3D shape of people from a few frames (1-8) of a monocular video in which the person is moving with a reconstruction accuracy of 4 to 5mm, while being orders of magnitude faster than previous methods. From semantic segmentation images, our *Octopus* model reconstructs a 3D shape, including the parameters of SMPL plus clothing and hair in 10 seconds or less. The model achieves fast and accurate predictions based on two key design choices. First, by predicting shape in a canonical T-pose space, the network learns to encode the images of the person into pose-invariant latent codes, where the information is fused. Second, based on the observation that feed-forward predictions are fast but do not always align with the input images, we predict using both, bottom-up and top-down streams (one per view) allowing information to flow in both directions. Learning relies only on synthetic 3D data. Once learned, *Octopus* can take a variable number of frames as input, and is able to reconstruct shapes even from a single image with an accuracy of 5mm. Results on 3 different datasets demonstrate the efficacy and accuracy of our approach. Code is available at [2].

1. Introduction

The automatic acquisition of detailed 3D human shape and appearance, including clothing and facial details is required for many applications such as VR/AR, gaming, virtual try-on, and cinematography.

A common way to acquire such models is with a scanner or a multi-view studio [3, 46]. The cost and size prevent the wide-spread use of such setups. Therefore, numerous works address capturing body shape and pose with more practical setups, e.g. from a low number of video cameras [60], or using one or more depth cameras, either specifically for the human body [9, 77, 84] or for general free-form surfaces [88, 51, 54, 37, 26, 70]. The most practical but also challenging setting is capturing from a single monocular RGB camera. Some methods attempt to infer the shape parameters of a body model from a single image [41, 53, 10, 24, 8, 32, 86, 39, 55], but reconstructed detail is constrained to the model shape space, and thus does not capture personalized shape detail and clothing geometry. Recent work [6, 5] estimates more detailed shape, including clothing, from a video sequence of a person rotating in front of a camera while holding a rough A-pose. While reconstructed models have high quality, the optimization approach takes around 2 minutes only for the shape component. More importantly, the main bottleneck is the pre-processing step, which requires fitting the SMPL model to each of the frame silhouettes using time-consuming non-

* Work partly conducted during an internship at the Real Virtual Humans group of Max Planck for Informatics.

linear optimization (≈ 120 min for 120 frames). This is impractical for many applications that require fast acquisition such as telepresence and gaming.

In this work, we address these limitations and introduce *Octopus*, a convolutional neural network (CNN) based model that learns to predict 3D human shapes in a canonical pose given a few frames of a person rotating in front of a *single camera*. *Octopus* predicts using both, bottom-up and top-down streams (one per view) allowing information to flow in both directions. It can make bottom-up predictions in 50ms per view, which are effectively refined top-down using the same images in 10s. Inference, both bottom-up and top-down, is performed fully-automatically using the same model. *Octopus* is therefore easy to use and more practical than previous work [6]. Learning only relies on synthetic 3D data, and on semantic segmentation images and keypoints derived from synthesized video sequences. Consequently, *Octopus* can be trained without paired data – real images with ground truth 3D shape annotations – which is very difficult to obtain in practice.

Octopus predicts SMPL body model parameters, which represent the undressed shape and the pose, plus additional 3D vertex offsets that model clothing, hair, and details beyond the SMPL space. Specifically, a CNN encodes F frames of the person (in different poses) into F latent codes that are fused to obtain a single shape code. From the shape code, two separate network streams predict the SMPL shape parameters, and the 3D vertex offsets in the canonical T-pose space, giving us the “unpose” shape or T-shape. Predicting the T-shape forces the F latent codes to be pose-invariant, which is necessary to fuse the shape information contained in each frame. *Octopus* also predicts a pose for each frame, which allows to “pose” the T-shape and render a silhouette to evaluate the overlap against the input images in a top-down manner during both training and inference. Specifically, since bottom-up models do not have a feedback loop, the feed-forward 3D predictions are correct but do not perfectly align with the input images. Consequently, we refine the prediction top-down by optimizing the F poses, the T-shape, and the vertex offsets to maximize silhouette overlap and joint re-projection error.

Experiments on a newly collected dataset (LifeScans), the publicly available PeopleSnapshot dataset [6], and on the dataset used in [9] demonstrate that our model infers shapes with a reconstruction accuracy of 4mm in less than 10 seconds. In summary, *Octopus* is faster than purely optimization-based fitting approaches such as [6], it combines the advantages of bottom-up and top-down methods in a single model, and can reconstruct detailed shapes and clothing from a few video frames. Examples of reconstruction results are shown in Fig. 1. To foster further research in this direction, we made *Octopus* available for research purposes [2].

2. Related Work

Methods for 3D human shape and pose reconstruction can be broadly classified as top-down or bottom-up. Top-down methods either fit a *free-form* surface or a statistical body model (*model-based*). Bottom-up methods directly infer a surface or body model parametrization from sensor data. We will review bottom-up and top-down methods for human reconstruction.

Top-down, free-form methods non-rigidly deform meshes [14, 22, 12] or volumetric shape representations [36, 4]. These methods are based on multi-view stereo reconstruction [42], and therefore require multiple RGB or depth cameras, which is a practical barrier for many applications. Using *depth cameras*, KinectFusion [38, 52] approaches reconstruct 3D scenes by incrementally fusing frame geometry, and appearance [85], in a canonical frame. Several methods build on KinectFusion for body scanning [64, 47, 82, 20]. The problem is that these methods require the person to stand still while the camera is turned around. DynamicFusion [51] generalized KinectFusion to non-rigid objects by combining non-rigid tracking and fusion. Although template-free approaches [52, 37, 65] are flexible, they can only handle very careful motions. Common ways to add robustness are pre-scanning the template [88], or using multiple kinects [26, 54] or multi-view [67, 44, 19]. These methods, however, do not register the temporal 3D reconstructions to the same template and focus on other applications such as streaming or telepresence [54]. Estimating shape by compensating for pose changes can be tracked back to Cheung *et al.* [17, 18], where they align visual hulls over time to improve shape estimation. To compensate for articulation, they merge shape information in a coarse voxel model. However, they need to track each body part separately and require multi-view input. All free-form works require multi-view input, depth cameras or cannot handle moving humans.

Top-down, model-based methods exploit a parametric body model consisting of pose and shape [7, 33, 48, 89, 57, 40] to regularize the fitting process. Some *Depth-based* methods [77, 34, 79, 84, 9] exploit the temporal information by optimizing a single shape and multiple poses (jointly or sequentially). This leads to expensive optimization problems. Using *multi-view*, some works achieve fast performance [60, 61] at the cost of using a coarser body model based on Gaussians [68], or a pre-computed template [80]. Early RGB-based methods were restricted to estimating the parameters of a body model, and required multiple views [8] or manually clicked points [30, 86, 39, 63]. Shape and clothing have been recovered from RGB images [31, 15], depth [16], or scan data [56], but require manual intervention or clothing is limited to a pre-defined set of templates. In [78] a fuzzy vertex association from clothing

to body surface is introduced, which allows complex clothing modeled as body offsets. Some works are in-between free-form and model-based methods. In [27, 76], authors pre-scan a template and insert a skeleton, and in [70] authors combine the SMPL model with a volumetric representation to track the clothed human body from a depth camera.

Bottom-up. Learning of features for multi-view photo-consistency [45], and auto-encoders combined with visual hulls [28, 72] have shown to improve free-form performance capture. These works, however, require more than one camera view. Very few works learn to predict personalized human shape from images—lack of training data and the lack of a feedback loop between feed-forward predictions and the images makes the problem hard. Variants of random forests and neural networks have been used [24, 23, 25, 75] to regress shape from silhouettes. The problem here is that predictions tend to look over-smooth, are confined to the model shape space, and do not comprise clothing. Garments are predicted [21] from a single image, but a single model for every new garment needs to be trained, which makes it hard to use in practice. Recent pure bottom-up approaches to human analysis [50, 49, 58, 87, 69, 71, 62] typically predict shape represented as a coarse stick figure or bone skeleton, and can not estimate body shape or clothing.

Hybrid methods. A recent trend of works combines bottom-up and top-down approaches—a combination that has been exploited already in earlier works [66]. The most straightforward way is by fitting a 3D body model [48] to 2D pose detections [10, 43]. These methods, however, can not capture clothing and details beyond the model space. Clothing, hair and shape [6, 5] can be inferred by fusing dynamic silhouettes (predicted bottom-up) of a video to a canonical space. Even with good 2D predictions, these methods are susceptible to local minima when not initialized properly, and are typically slow. Furthermore, the 2D prediction network and the model fitting is de-coupled. Starting with a feed-forward 3D prediction, semantic segmentation, keypoints and scene constraints are integrated top-down in order to predict the pose and shape of multiple people [81]. Other recent works integrate the SMPL model, or a voxel representation [74], as a layer within a network architecture [41, 55, 53, 73]. This has several advantages: (i) predictions are constrained by a shape space of humans, and (ii) bottom-up 3D predictions can be verified top-down using 2D keypoints and silhouettes during training. However, the shape estimates are confined to the model shape space and tend to be close to the average. The focus of these works is rather on robust pose estimation, while we focus on personalized shapes. We also integrate SMPL within our architecture but our work is different in several aspects. First, our architecture fuses the information of several images of the same person in different poses. Second, our model incorporates a fast top-down component during training and

at test time. As a result, we can predict clothing, hair and personalized shapes using a single camera.

3. Method

The goal of this work is to create a 3D model of a subject from a few frames of a monocular RGB video, and in less than 10 seconds. The model should comprise body shape, hair, and clothing and should be animatable. We take inspiration from [6] and focus on the cooperative setting with videos of people rotating in front of a camera holding a rough A-pose – this motion is easy and fast to perform, and ensures that non-rigid motion of clothing and hair is not too large. In contrast to previous work [6], we aim for fast and fully automatic reconstruction. To this end, we train a novel convolutional neural network to infer a 3D mesh model of a subject from a small number of input frames. Additionally, we train the network to reconstruct the 3D pose of the subject in each frame. This allows us to refine the body shape by utilizing the decoder part of the network for instance-specific optimization (Fig. 2).

In Sec. 3.1 we describe the shape representation used in this work followed by its integration into the used predictor (Sec. 3.2). In Sec. 3.3 we explain the losses, that are used in the experiments. We conclude by describing the instance-specific top-down refinement of results (Sec. 3.4).

3.1. Shape representation

Similar to previous work [83, 6], we represent shape using the SMPL statistical body model [48], which represents the undressed body, and a set of offsets modeling instance specific details including clothing and hair.

SMPL is a function $M(\cdot)$ that maps pose θ and shape β to a mesh of $V = 6890$ vertices. By adding offsets \mathbf{D} to the template \mathbf{T} , we obtain a posed shape instance as follows:

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

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

where linear blend-skinning $W(\cdot)$ with weights \mathbf{W} , together with pose-dependent deformations $B_p(\theta)$ allow to pose the T-shape ($\mathbf{T} + B_s(\beta)$) based on its skeleton joints $J(\cdot)$. SMPL plus offsets, denoted as SMPL+D, is fully differentiable with respect to pose θ , shape β and free-form deformations \mathbf{D} . This allows us to directly integrate SMPL as a fixed layer in our convolutional architecture.

3.2. Model and data representation

Given a set of images $\mathcal{I} = \{\mathbf{I}_0, \dots, \mathbf{I}_{F-1}\}$ depicting a subject from different sides with corresponding 2D joints $\mathcal{J} = \{\mathbf{J}_0, \dots, \mathbf{J}_{F-1}\}$, we learn a predictor f_w^* that infers the body shape β , personal and scene specific body features \mathbf{D} , and 3D poses $\mathcal{P} = \{\theta_0, \dots, \theta_{F-1}\}$ along with 3D positions $\mathcal{T} = \{t_0, \dots, t_{F-1}\}$ for each image.

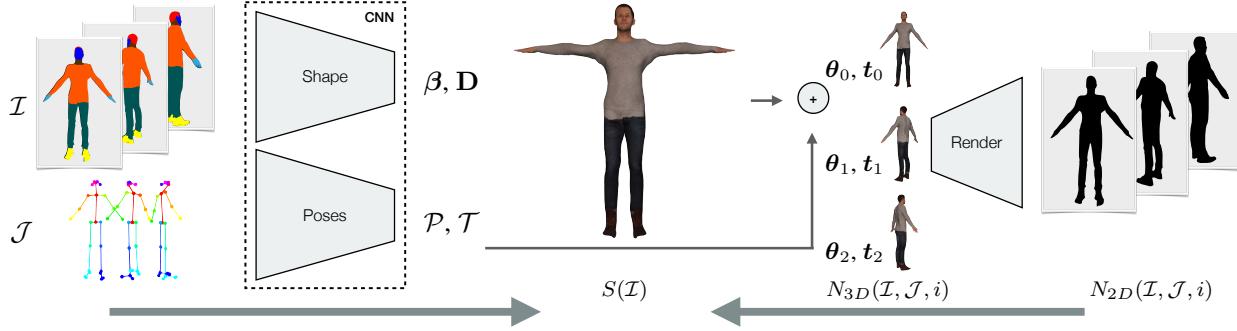


Figure 2. Overview of our method: Our novel CNN predicts 3D human shapes from semantic images in a canonical pose together with per-image pose information calculated from 2D joint detections (left to center). The pose information can be used to refine the shape via ‘render and compare’ optimization using the same predictor (right to center).

$f_w^* : (\mathcal{I}, \mathcal{J}) \mapsto (\beta, \mathbf{D}, \mathcal{P}, \mathcal{T})$ is a CNN parametrized by network parameters w .

Input modalities. Images of humans are highly diverse in appearance, requiring large datasets of annotated images in the context of deep learning. Therefore, to abstract away as much information as possible while still retaining shape and pose signal, we build on previous work [29, 13] to simplify each RGB image to a semantic segmentation and 2D keypoint detections. This allows us to train the network using only synthetic data and generalize to real data.

Model parametrization. By integrating the SMPL+D model (Sec. 3.1) into our network formulation, we can utilize its mesh output in the training of f_w^* . Concretely, we supervise predicted SMPL+D parameters in three ways: Imposing a loss directly on the mesh vertices $M(\beta, \theta, \mathbf{D})$, on the predicted joint locations $J(\beta)$ and their projections on the image, and densely on a rendering of the mesh using a differential renderer [35].

The T-shape ($\mathbf{T} + B_s(\beta) + \mathbf{D}$) in Eq. 2 is now predicted from the set of semantic images \mathcal{I} with the function:

$$S(\mathcal{I}) = \mathbf{T} + B_s(f_w^\beta(\mathcal{I})) + f_w^\mathbf{D}(\mathcal{I}), \quad (3)$$

where f_w^* are the regressors to be learned. Similarly, the mesh posed $N_{3D}(\mathcal{I}, \mathcal{J}, i)$ is predicted from the image \mathbf{I}_i and 2D joints \mathcal{J}_i with the function:

$$N_{3D}(\mathcal{I}, \mathcal{J}, i) = W(P(\mathcal{I}, i), J(f_w^\beta(\mathcal{I})), f_w^{\theta_i}(\mathcal{I}, \mathcal{J}), \mathbf{W}) \quad (4)$$

$$P(\mathcal{I}, \mathcal{J}, i) = S(\mathcal{I}) + B_p(f_w^{\theta_i}(\mathcal{I}, \mathcal{J})), \quad (5)$$

from which the 3D Joints are predicted with the linear regressor J_{B25} :

$$N_{J_{3D}}(\mathcal{I}, \mathcal{J}, i) = J_{B25}(N_{3D}(\mathcal{I}, \mathcal{J}, i)) \quad (6)$$

J_{B25} has been trained to output 25 joint locations consistent with the BODY_25 [1] keypoint ordering. The estimated posed mesh N_{3D} can be rendered in uniform color with the image formation function $R(\cdot)$ parameterized by camera c :

$$N_{2D}(\mathcal{I}, \mathcal{J}, i) = R_c(N_{3D}(\mathcal{I}, \mathcal{J}, i)) \quad (7)$$

Similarly, we can project the the joints $N_{J_{3D}}$ to the image plane by perspective projection π :

$$N_{J_{2D}}(\mathcal{I}, \mathcal{J}, i) = \pi_c(N_{J_{3D}}(\mathcal{I}, \mathcal{J}, i)) \quad (8)$$

All these operations are differentiable, which we can conveniently use to formulate suitable loss functions.

3.3. Loss functions

Our architecture permits two sources of supervision: (i) 3D supervision (in our experiments, from synthetic data derived by fitting SMPL+D to static scans), and (ii) 2D supervision from video frames alone. In this section, we discuss different loss functions used to train the predictors f_w^* .

Losses on body shape and pose For a paired sample in the dataset $\{(\mathcal{I}, \mathcal{J}), (\beta, \mathbf{D}, \mathcal{P}, \mathcal{T})\}$ we use the following losses between our estimated model N_{3D} and the ground truth model $M(\cdot)$ scan:

- Per-vertex loss in the canonical T-pose $\mathbf{0}_\theta$. This loss provides a useful 3D supervision on shape independently of pose:

$$\mathcal{L}_S = \|S(\mathcal{I}) - M(\beta, \mathbf{0}_\theta, \mathbf{D})\|^2 \quad (9)$$

- Per-vertex loss in posed space. This loss supervises both pose and shape on the Euclidean space:

$$\mathcal{L}_{N_{3D}} = \sum_{i=0}^{F-1} \|N_{3D}(\mathcal{I}, \mathcal{J}, i) - M(\beta, \theta_i, \mathbf{D})\|^2 \quad (10)$$

- Silhouette overlap:

$$\mathcal{L}_{N_{2D}} = \sum_{i=0}^{F-1} \|R_c(N_{3D}(\mathcal{I}, \mathcal{J}, i)) - b(\mathbf{I}_i)\|^2, \quad (11)$$

where $b(\mathbf{I}_i)$ is the binary segmentation mask and R_c is the image formation function defined in Eq. 7. $\mathcal{L}_{N_{2D}}$ is a weakly supervised loss as it does not require 3D annotations and $b(\mathbf{I}_i)$ can be estimated directly from RGB images. In the experiments, we investigate whether such self-supervised loss can reduce the amount 3D supervision required (see 4.4). Additionally, we show that N_{2D} can be

used at test time to refine the bottom-up predictions and capture instance specific details in a top-down manner (see 3.4).

- Per-vertex SMPL undressed body loss:

The aforementioned losses only penalize the final SMPL+D 3D shape. It is useful to include an "undressed-body" (\hat{S}) loss to force the shape parameters β to be close to the ground truth

$$\mathcal{L}_{\hat{S}} = \|\hat{S}(\mathcal{I}) - M(\beta, \mathbf{0}_\theta, \mathbf{0}_D)\|^2 \quad (12)$$

$$\hat{S}(\mathcal{I}) = \mathbf{T} + B_s(f_w^\beta(\mathcal{I})), \quad (13)$$

where $\mathbf{0}_D$ are vectors of length 0. This also prevents that the offsets D explain the overall shape of the person.

Pose specific losses. In addition to the posed space $\mathcal{L}_{N_{3D}}$ and silhouette overlap $\mathcal{L}_{N_{2D}}$ losses, we train for the pose using a direct loss on the predicted parameters $\mathcal{L}_{\theta, t}$

$$\mathcal{L}_{\theta, t} = \sum_{i=0}^{F-1} \left(\|\mathbf{R}(f_w^{\theta_i}) - \mathbf{R}(\theta_i)\|^2 + \|f_w^{t_i} - t_i\|^2 \right), \quad (14)$$

where \mathbf{R} are vectorized rotation matrices of the 24 joints. Similar to [53, 43, 55], we use differentiable SVD to force the predicted matrices to lie on the manifold of rotation matrices. This term makes the pose part of the network converge faster.

Losses on joints. We further regularize the pose training by imposing a loss on the joints in Euclidean space:

$$\mathcal{L}_{J_{3D}} = \sum_{i=0}^{F-1} \|N_{J_{3D}}(\mathcal{I}, \mathcal{J}, i) - J_{B25}(M(\beta, \theta_i, D))\|^2 \quad (15)$$

Similar to the 2D image projection loss on model $\mathcal{L}_{N_{2D}}$ (Eq. 11), we also have a weakly supervised 2D joint projection loss $\mathcal{L}_{J_{2D}}$

$$\mathcal{L}_{J_{2D}} = \sum_{i=0}^{F-1} \|N_{J_{2D}}(\mathcal{I}, \mathcal{J}, i) - \pi_c(J_{B25}(M(\beta, \theta_i, D)))\|^2. \quad (16)$$

3.4. Instance-specific top-down optimization

The bottom-up predictions of the neural model can be refined top-down at test time to capture instance specific details. It is important to note that this step requires no 3D annotation as the network fine-tunes using only 2D data. Specifically, at the test time, given a subject's images \mathcal{I} and 2D joints \mathcal{J} we optimize a small set of layers in f_w^* using image and joint projection losses $\mathcal{L}_{N_{2D}}, \mathcal{L}_{J_{2D}}$ (see 4.1). By fixing most layers of the network and optimizing only latent layers, we find a compromise between the manifold of shapes learned by the network and new features, that have not been learned. We further regularize this step using Laplacian smoothness, face landmarks, and symmetry terms from [6, 5]. Table 1 illustrates the performance of the pipeline before and after optimization (see 4.2, 4.3).



Figure 3. Sample scans from the *LifeScans* dataset.

4. Experiments

The following section focuses on the evaluation of our method. In Sec. 4.1 we introduce technical details of the used dataset and network architecture. The following sections describe experiments for quantitative and qualitative evaluation as well as ablation and parameter analysis.

4.1. Experimental setup

Dataset. To alleviate the lack of paired data, we use 2043 static 3D scans of people in clothing. We purchased 163 scans from renderpeople.com and 54 from axyz-design.com. 1826 scans were kindly provided from Twin-dom (<https://web.twindom.com/>). Unfortunately, in the 2043 there is not enough variation in pose and shape to learn a model that generalizes. Hence, we generate synthetic 3D data by non-rigidly registering SMPL+D to each of the scans. This allows us to change the underlying body shape and pose of the scan using SMPL, see Fig. 3. Like [6], we focus on a cooperative scenario where the person is turning around in front of the camera. Therefore, we animate the scans with turn-around poses and random shapes and render video sequences from them. We call the resulting dataset *LifeScans*, which consists of rendered images *paired* with 3D animated scans in various shapes and poses. Since the static scans are from real people, the generated images are close to photo-realistic, see Fig 3. To prevent overfitting, we use semantic segmentation together with keypoints as intermediate image representation, which preserve shape and pose signatures while abstracting away appearance. This reduces the amount of appearance variation required for training. To be able to render synthetic semantic segmentation, we first render the *LifeScans* subjects from different viewpoints and segment the output with the method of [29]. Then we project the semantic labels back in the SMPL texture space and fuse different views using graph cut-based optimization. This final step enables full synthetic generation of paired training data.

Scale ambiguity. Scale is an inherent ambiguity in monocular imagery. Three factors determine the size of an object in an image: distance to the camera, camera intrinsics, and the size of the object. As it is not possible to decouple this ambiguity in a monocular set-up with moving objects, we fix two factors and regress one. In other works [53, 41, 55] authors have assumed fixed distance to the cam-

era. We cannot make this assumption, as we leverage multiple images of the same subject, where the distances to the camera may vary. Consequently, we fix the size of the subject to average body height. Precisely, we make SMPL height independent, by multiplying the model by 1.66m divided by the y-axis distance of vertices describing ankles and eyes. Finally, we fix the focal length to sensor height.

Network architecture. In the following we describe details of the convolutional neural network f_w^* . An overview is given in Fig. 4. The input to f_w^* is a set of 1080x1080px semantically segmented images \mathcal{I} and corresponding 2D joint locations \mathcal{J} . f_w^* encodes each image \mathbf{I}_i with a set of five, 3x3 convolutions with ReLU activations followed by 2x2 max-pooling operations into a pose invariant latent code l_i^{inv} . In our experiments we fixed the size of l_i^{inv} to 20. The *pose branch* maps both joint detections \mathbf{J}_i and output of the last convolutional layer to a vector of size 200 and finally to the pose-dependent latent code l_i^{pose} of size 100 via fully connected layers. The *shape branch* aggregates pose invariant information across images and computes mean l^{inv} . Note that this formulation allows us to aggregate pose-dependent and invariant information across an arbitrary and varying number of views. The shape branch goes on to predict SMPL shape parameters β and free-form deformations \mathbf{D} on the SMPL mesh. β is directly calculated from l^{inv} with a linear layer. In order to predict per-vertex offsets from the latent code l^{inv} , we use a four-step graph convolutional network with Chebyshev filters and mesh up-sampling layers similar to [59]. Each convolution is followed by ReLU activation. We prefer a graph convolutional network over a fully connected decoder due to memory constraints and in order to get structured predictions.

Training scheme. The proposed method, including rendering, is fully differentiable and end-to-end trainable. Empirically we found it better to train the pose branch before training the shape branch. Thereafter, we optimize the network end-to-end. We use a similar training schedule for our pose branch as [55], where we first train the network using losses on the joints and pose parameters ($\mathcal{L}_{J_{3D}}, \mathcal{L}_{\theta, t}$) followed by training using losses on the vertices and pose parameters ($\mathcal{L}_{N_{3D}}, \mathcal{L}_{\theta, t}$). We also experiment with various training schemes, and show that weakly supervised training can significantly reduce the dependence on 3D annotated data (see Sec. 4.4). For that experiment, we train the model with alternating full ($\mathcal{L}_S, \mathcal{L}_{\hat{S}}, \mathcal{L}_{N_{3D}}, \mathcal{L}_{J_{3D}}$) and weak supervision ($\mathcal{L}_{N_{2D}}, \mathcal{L}_{J_{2D}}$). During instance-specific optimization we keep most layers fixed and only optimize latent pose l^{pose} , latent shape l^{inv} and the last graph convolutional layer, that outputs free-form displacements \mathbf{D} .

4.2. Numerical evaluation

We quantitatively evaluate our method on a separated test set of the LifeScans dataset containing 55 subjects. We use

	Before optimization	After optimization
Full Pipeline	4.47 ± 4.45	4.00 ± 3.94
GT Poses	4.47 ± 4.41	3.17 ± 3.41

Table 1. Mean vertex error (mm) of 55 test samples computed on $F = 8$ input images. The *full method* with inferred poses produces comparable results to using *GT poses*. Both variants gain accuracy from subsequent optimization.

$F = 8$ semantic segmentation images and 2D poses as input and optimize the results for a maximum budget of 10 seconds. All results have been computed without intensive hyper-parameter tuning. To quantify shape reconstruction accuracy, we adjust the pose of the estimation to match the ground truth, following [83, 9]. This disentangles errors in pose from errors in shape and allows to quantify shape accuracy. Finally, we compute the bi-directional vertex to surface distance between scans and reconstructions. We report mean errors in millimeters (mm) across the test set in Tab. 1. We differentiate between *full method* and *ground truth (GT) poses*. Full method refers to our method as described in Sec. 4.1. The latter is a variant of our method that uses ground truth poses, which allows to study the effect of pose errors. In Fig. 5 we display subjects in the test set for both variants along with per-vertex error heatmaps. Visually the results look almost indistinguishable, which is corroborated by the fact that the numerical error increases only by ≈ 1 mm between GT and predicted pose models. This demonstrates the robustness of our approach. We show more examples with the corresponding texture for qualitative assessment in Fig. 1. The textures have been computed using graph cut-based optimization using semantic labels as described in [5].

4.3. Analysis of key parameters

Our method comes with two key hyper-parameters, namely number of input images F , and number of optimization steps. In the following section, we study these parameters and how they affect the performance of our approach. We also justify our design choices.

Fig. 7 illustrates the performance of our method with growing number of optimization steps. While the performance gain saturates at around 70 – 80 steps, we use 25 steps in following experiments as a compromise between accuracy and speed. For the case of $F = 8$ input images optimization for 25 steps takes ≈ 10 s on a single Volta V100 GPU. We believe 10s is a practical waiting time and a good compromise for many applications. Therefore we fix the time budget to 10s for the following experiments.

Including more input views at test time can potentially improve the performance of the method. However, in practice, this means more data pre-processing and longer inference times. Fig. 8 illustrates the performance with different number of input images. Perhaps surprisingly, the performance saturates already at around 5 images before optimi-

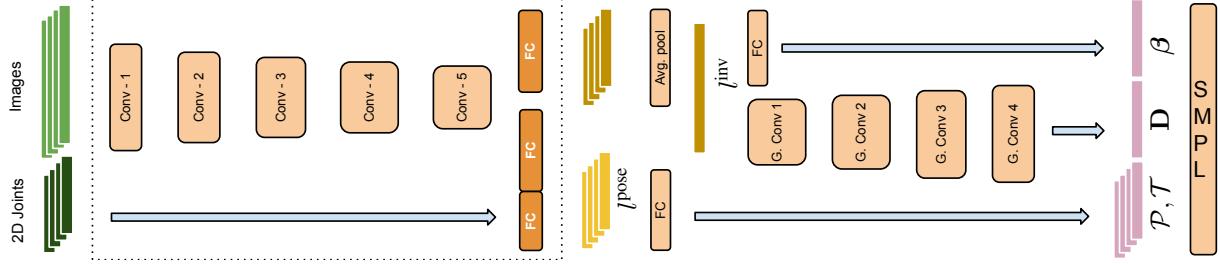


Figure 4. Network architecture: Our bottom-up inference network first encodes the input (semantically segmented images \mathcal{I} and 2D joints \mathcal{J}_{2D}) into a decoupled, pose dependent l^{pose} and pose invariant l^{inv} latent space. The pose branch subsequently infers per-frame pose and translation parameters \mathcal{P} and \mathcal{T} from l^{pose} . The shape branch infers the body shape β and free-form deformations \mathbf{D} in T-pose from l^{inv} . We use a graph convolution based decoder, to learn per-vertex offsets \mathbf{D} . The entire model is end-to-end trainable. The orange FC layers and the final graph convolution layer can be fine-tuned at test time to better model instance-specific details (see Sec. 3.4).

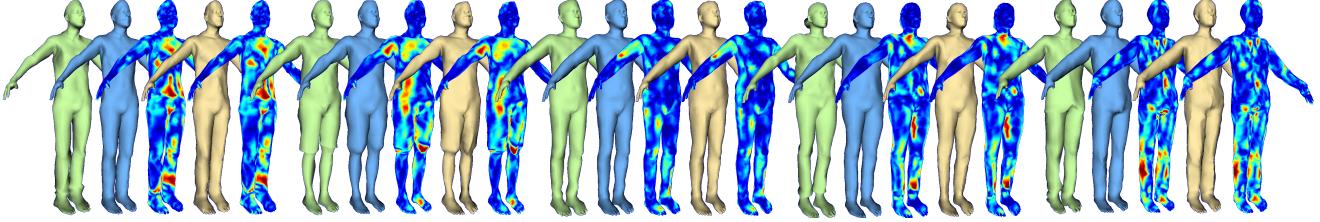


Figure 5. Results from LifeScans in comparison to ground truth shapes (green). We show results computed with ground truth poses (blue) and results of the full method (yellow) with corresponding error heatmaps with respect to ground truth shapes (red means $\geq 2\text{cm}$).

mization. After optimization, the error saturates at around 8 images. While more images potentially mean better supervision, we cannot see improved results for optimization on many images. This can be explained with the fixed time budget in this experiment, where more images mean fewer optimization steps. While we could potentially use fewer images, we found $F = 8$ views as a practical number of input views. This has the following reason: A calculated avatar should not only be numerically accurate but also visually appealing. Results based on more number of views show more fine details and most importantly allow accurate texture calculation.

4.4. Type of supervision

Since videos are easier to obtain than 3D annotations, we evaluate to which extent they can substitute full 3D supervision to train our network. To this end, we split the LifeScans dataset. One part is used for full supervision, the other part is used for weak supervision in form of image masks and 2D keypoints. All forms of supervision can be synthetically generated from the LifeScans dataset. We train f_w^* with 10%, 20%, 50%, and 100% full supervision and compare the performance on the test set in Tab. 2. In order to factor out the effect of problematic poses during the training, we used ground truth poses in this experiment. The results suggest that f_w^* can be trained with only minimal amount of full supervision, given strong pose predictions. The performance of the network decreases only slightly for less than 100% full supervision. Most interestingly, the results are almost identical for 10%, 20%, and 50% full super-

	Before optimization	After optimization
100%	4.47 ± 4.41	3.17 ± 3.41
50%	4.57 ± 4.52	3.19 ± 3.43
20%	4.74 ± 4.65	3.29 ± 3.53
10%	4.73 ± 4.56	3.46 ± 3.62

Table 2. Mean vertex error (mm) of 55 test samples with different amount of *full supervision* during training of the shape branch. f_w^* can be trained with only 10% full supervision with minimal accuracy loss.

vision. This experiment suggests that we could potentially improve performance by supervising our model with additionally recorded videos. We leave this for future work.

4.5. Qualitative results and comparisons

We qualitatively compare our method against the most relevant work [6] on their *PeopleSnapshot* dataset. While their method leverages 120 frames, we still use $F = 8$ frames for our reconstructions. For a fairer comparison, we optimize for $\approx 20\text{s}$ in this experiment. This is still several magnitudes faster than the 122min needed by [6]. Their method needs 2 minutes for shape optimization plus 1 minute per frame for the pose. In Fig. 6 we show side-by-side comparison to [6]. Our results are visually still on par while requiring a fraction of the data.

We also compare our method against [9], a RGB-D based optimization method. Their dataset displays subjects in minimal clothing rotating in front of the camera in T-pose. Unfortunately, the semantic segmentation network is not able to successfully segment subjects in minimal clothing. Therefore we slightly change the set-up for this experiment.

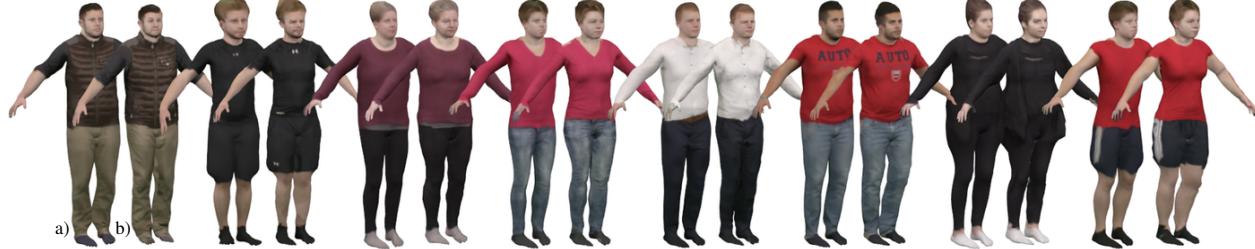


Figure 6. Comparison to the state-of-the-art optimization based method [6]. Their method (a) uses 120 frames, while ours (b) only uses 8 images and is several magnitudes faster.

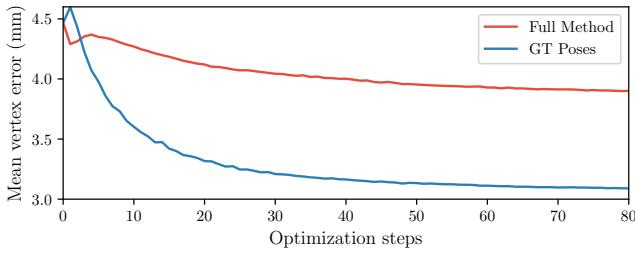


Figure 7. Error decrease of the test set with increased number of optimization steps computed on $F = 8$ input images.

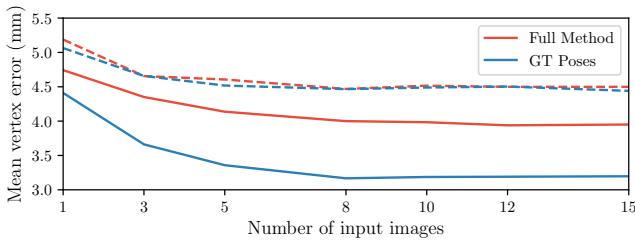


Figure 8. Error development on the test set with increased number of input views F before (dashed) and after optimization (solid). Optimization has been limited by a time budget of 10s allowing very few gradient steps for large numbers of views, which explains why the error plateaus for more than 8 views.

We segment their dataset using the semi-automatically approach [11] and re-train our predictor to be able to process binary segmentation masks. Additionally, we augment the LifeScans dataset with T-poses. We show side-by-side comparisons in Fig. 9. Again our results are visually similar, despite the use of less and only monocular data.

5. Discussion and Conclusion

We have proposed a novel method for automatic 3D body shape estimation from only 1 – 8 frames of a monocular video of a person moving. Our *Octopus* model predicts mesh-based pose invariant shape and per-image 3D pose from a flexible number of views. Experiments demonstrate that the feed-forward predictions are already quite accurate (4.5mm), but often lack detail and do not perfectly overlap with the input images. This motivates refining the estimates with top-down optimization against the input images. Refining brings the error down to 4mm and aligns the model with the input image silhouettes, which allows texture map-

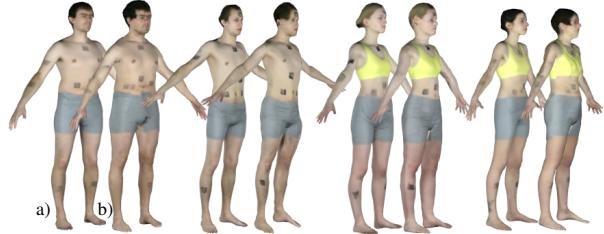


Figure 9. Comparison to the RGB-D method [9] (a). Our method (b) is visually on par, despite using only 8 RGB images as input.

ping. In summary, we improve over the state-of-the-art in the following aspects: Our method allows, for the first time, to estimate full body reconstructions of people in clothing in a fully automatic manner. We significantly reduce the number of needed images at test time, and compute the final result several magnitudes faster than state-of-the-art (from hours to seconds). Extensive experiments on the LifeScans dataset demonstrate the performance and influence of key parameters of the predictor. While our model is independent on the number of input images and can be refined for different numbers of optimization steps, we have shown that using 8 views and refining for 10 seconds are good compromises between accuracy and practicability. Qualitative results on two real-world datasets demonstrate generalization to real data, despite training from synthetic data alone.

Future work should enable the proposed method for scenarios where the subject is not cooperating, for example from Youtube videos, or legacy movie material. Furthermore, clothing with geometry far from the body, such as skirts and coats or hairstyles like ponytails will require a different formulation.

By enabling fully automatic 3D body shape reconstruction from a few images in only a few seconds, we prepare the ground for wide-spread acquisition of personalized 3D avatars. People are now able to quickly digitize themselves using only a webcam and can use their model for various VR and AR applications.

Acknowledgments

The authors gratefully acknowledge funding by Deutsche Forschungsgemeinschaft (DFG German Research Foundation) from projects MA2555/12-1 and 409792180. We would like to thank Twindom for providing us with the scan data. Another thanks goes to Verica Lazova for great help in data processing.

References

- [1] <https://github.com/cmu-perceptual-computing-lab/openpose>. 4
- [2] <http://virtualhumans.mpi-inf.mpg.de/octopus/>. 1, 2
- [3] Naveed Ahmed, Edilson de Aguiar, Christian Theobalt, Marcus Magnor, and Hans-Peter Seidel. Automatic generation of personalized human avatars from multi-view video. In *Proc. of the ACM Symposium on Virtual Reality Software and Technology*, VRST '05, pages 257–260, New York, NY, USA, 2005. ACM. 1
- [4] Benjamin Allain, Jean-Sébastien Franco, and Edmond Boyer. An Efficient Volumetric Framework for Shape Tracking. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 268–276, Boston, United States, 2015. IEEE. 2
- [5] Thimo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Detailed human avatars from monocular video. In *International Conf. on 3D Vision*, sep 2018. 1, 3, 5, 6
- [6] Thimo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3D people models. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2018. 1, 2, 3, 5, 7, 8
- [7] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. SCAPE: shape completion and animation of people. In *ACM Transactions on Graphics*, volume 24, pages 408–416. ACM, 2005. 2
- [8] Alexandru O Bălan and Michael J Black. The naked truth: Estimating body shape under clothing. In *European Conf. on Computer Vision*, pages 15–29. Springer, 2008. 1, 2
- [9] Federica Bogo, Michael J. Black, Matthew Loper, and Javier 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, 8
- [10] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael 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, 3
- [11] Sergi Caelles, Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, and Luc Van Gool. One-shot video object segmentation. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2017. 8
- [12] Cedric Cagniart, Edmond Boyer, and Slobodan Ilic. Probabilistic deformable surface tracking from multiple videos. In Kostas Daniilidis, Petros Maragos, and Nikos Paragios, editors, *European Conf. on Computer Vision*, volume 6314 of *Lecture Notes in Computer Science*, pages 326–339, Heraklion, Greece, 2010. Springer. 2
- [13] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2017. 4
- [14] Joel Carranza, Christian Theobalt, Marcus A Magnor, and Hans-Peter Seidel. Free-viewpoint video of human actors. In *ACM Transactions on Graphics*, volume 22, pages 569–577. ACM, 2003. 2
- [15] Xiaowu Chen, Yu Guo, Bin Zhou, and Qinping Zhao. Deformable model for estimating clothed and naked human shapes from a single image. *The Visual Computer*, 29(11):1187–1196, 2013. 2
- [16] Xiaowu Chen, Bin Zhou, Feixiang Lu, Lin Wang, Lang Bi, and Ping Tan. Garment modeling with a depth camera. *ACM Transactions on Graphics*, 34(6):203, 2015. 2
- [17] German KM Cheung, Simon Baker, and Takeo Kanade. Shape-from-silhouette of articulated objects and its use for human body kinematics estimation and motion capture. In *IEEE Conf. on Computer Vision and Pattern Recognition*, volume 1, pages I–I. IEEE, 2003. 2
- [18] German KM Cheung, Simon Baker, and Takeo Kanade. Visual hull alignment and refinement across time: A 3d reconstruction algorithm combining shape-from-silhouette with stereo. In *IEEE Conf. on Computer Vision and Pattern Recognition*, volume 2, pages II–375. IEEE, 2003. 2
- [19] Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese, Hugues Hoppe, Adam Kirk, and Steve Sullivan. High-quality streamable free-viewpoint video. *ACM Transactions on Graphics*, 34(4):69, 2015. 2
- [20] Yan Cui, Will Chang, Tobias Nöll, and Didier Stricker. Kinectavatar: fully automatic body capture using a single kinect. In *Asian Conf. on Computer Vision*, pages 133–147, 2012. 2
- [21] R Daněk, Endri Dibra, C Öztireli, Remo Ziegler, and Markus Gross. Deepgarment: 3d garment shape estimation from a single image. In *Computer Graphics Forum*, volume 36, pages 269–280. Wiley Online Library, 2017. 3
- [22] Edilson De Aguiar, Carsten Stoll, Christian Theobalt, Naveed Ahmed, Hans-Peter Seidel, and Sebastian Thrun. Performance capture from sparse multi-view video. In *ACM Transactions on Graphics*, page 98, 2008. 2
- [23] Endri Dibra, Himanshu Jain, Cengiz Öztireli, Remo Ziegler, and Markus Gross. Hs-nets: Estimating human body shape from silhouettes with convolutional neural networks. In *International Conf. on 3D Vision*, pages 108–117, 2016. 3
- [24] Endri Dibra, Himanshu Jain, Cengiz Öztireli, Remo Ziegler, and Markus Gross. Human shape from silhouettes using generative hks descriptors and cross-modal neural networks. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2017. 1, 3
- [25] Endri Dibra, Cengiz Öztireli, Remo Ziegler, and Markus Gross. Shape from selfies: Human body shape estimation using cca regression forests. In *European Conf. on Computer Vision*, pages 88–104. Springer, 2016. 3
- [26] Mingsong Dou, Sameh Khamis, Yury Degtyarev, Philip Davidson, Sean Ryan Fanello, Adarsh Kowdle, Sergio Orts Escolano, Christoph Rhemann, David Kim, Jonathan Taylor, et al. Fusion4d: Real-time performance capture of challenging scenes. *ACM Transactions on Graphics*, 35(4):114, 2016. 1, 2
- [27] Juergen Gall, Carsten Stoll, Edilson De Aguiar, Christian Theobalt, Bodo Rosenhahn, and Hans-Peter 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
- [28] Andrew Gilbert, Marco Volino, John Collomosse, and Adrian Hilton. Volumetric performance capture from minimal camera viewpoints. In *European Conf. on Computer Vision*, 2018.
- [29] Ke Gong, Xiaodan Liang, Yicheng Li, Yimin Chen, Ming Yang, and Liang Lin. Instance-level human parsing via part grouping network. In *European Conf. on Computer Vision*, 2018. 4, 5
- [30] Peng Guan, Alexander Weiss, Alexandru O Bălan, and Michael J Black. Estimating human shape and pose from a single image. In *IEEE International Conf. on Computer Vision*, pages 1381–1388. IEEE, 2009. 2
- [31] Yu Guo, Xiaowu Chen, Bin Zhou, and Qinping Zhao. Clothed and naked human shapes estimation from a single image. *Computational Visual Media*, pages 43–50, 2012. 2
- [32] Nils Hasler, Hanno Ackermann, Bodo Rosenhahn, Thorsten Thormahlen, and Hans-Peter Seidel. Multilinear pose and body shape estimation of dressed subjects from image sets. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1823–1830. IEEE, 2010. 1
- [33] Nils Hasler, Carsten Stoll, Martin Sunkel, Bodo 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
- [34] Thomas Helten, Andreas Baak, Gaurav Bharaj, Meinard Muller, Hans-Peter Seidel, and Christian Theobalt. Personalization and evaluation of a real-time depth-based full body tracker. In *International Conf. on 3D Vision*, pages 279–286, Washington, DC, USA, 2013. 2
- [35] Paul Henderson and Vittorio Ferrari. Learning to generate and reconstruct 3d meshes with only 2d supervision. In *British Machine Vision Conference*, 2018. 4
- [36] Chun-Hao Huang, Benjamin Allain, Jean-Sébastien Franco, Nassir Navab, Slobodan Ilic, and Edmond Boyer. Volumetric 3d tracking by detection. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 3862–3870, 2016. 2
- [37] Matthias Innmann, Michael Zollhöfer, Matthias Nießner, Christian Theobalt, and Marc Stamminger. Volumedeform: Real-time volumetric non-rigid reconstruction. In *European Conf. on Computer Vision*, 2016. 1, 2
- [38] Shahram Izadi, David Kim, Otmar Hilliges, David Molyneaux, Richard Newcombe, Pushmeet Kohli, Jamie Shotton, Steve Hodges, Dustin Freeman, Andrew Davison, et al. Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera. In *ACM symposium on User interface software and technology*, pages 559–568. ACM, 2011. 2
- [39] Arjun Jain, Thorsten Thormählen, Hans-Peter Seidel, and Christian Theobalt. Moviereshape: Tracking and reshaping of humans in videos. In *ACM Transactions on Graphics*, volume 29, page 148. ACM, 2010. 1, 2
- [40] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 8320–8329, 2018. 2
- [41] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *IEEE Conf. on Computer Vision and Pattern Recognition*. IEEE Computer Society, 2018. 1, 3, 5
- [42] Reinhard Koch, Marc Pollefeys, and Luc Van Gool. Multi viewpoint stereo from uncalibrated video sequences. In *European conf. on computer vision*, pages 55–71. Springer, 1998. 2
- [43] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J Black, and Peter V Gehler. Unite the people: Closing the loop between 3d and 2d human representations. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2017. 3, 5
- [44] Vincent Leroy, Jean-Sébastien Franco, and Edmond Boyer. Multi-View Dynamic Shape Refinement Using Local Temporal Integration. In *IEEE International Conf. on Computer Vision*, Venice, Italy, 2017. 2
- [45] Vincent Leroy, Jean-Sébastien Franco, and Edmond Boyer. Shape reconstruction using volume sweeping and learned photoconsistency. In *European Conf. on Computer Vision*, pages 796–811. Springer, Cham, 2018. 3
- [46] Guannan Li, Chenglei Wu, Carsten Stoll, Yebin Liu, Kiran Varanasi, Qionghai Dai, and Christian Theobalt. Capturing relightable human performances under general uncontrolled illumination. In *Computer Graphics Forum (Proc. Eurographics)*, volume 32, pages 1–8, 2013. 1
- [47] Hao Li, Etienne Vouga, Anton Gudym, Linjie Luo, Jonathan T Barron, and Gleb Gusev. 3d self-portraits. *ACM Transactions on Graphics*, 32(6):187, 2013. 2
- [48] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics*, 34(6):248:1–248:16, 2015. 2, 3
- [49] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3d pose estimation from monocular rgb. In *International Conf. on 3D Vision*, sep 2018. 3
- [50] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. Vnect: Real-time 3d human pose estimation with a single rgb camera. *ACM Transactions on Graphics*, 36(4):44, 2017. 3
- [51] Richard A Newcombe, Dieter Fox, and Steven 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
- [52] Richard A Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J Davison, Pushmeet Kohi, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *IEEE International Symposium on Mixed and Augmented Reality*, pages 127–136, 2011. 2
- [53] Mohamed Omran, Christop Lassner, Gerard Pons-Moll, Peter Gehler, and Bernt Schiele. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *International Conf. on 3D Vision*, 2018. 1, 3, 5

- [54] Sergio Orts-Escalano, Christoph Rhemann, Sean Fanello, Wayne Chang, Adarsh Kowdle, Yury Degtyarev, David Kim, Philip L Davidson, Sameh Khamis, Mingsong 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
- [55] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas 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, 3, 5, 6
- [56] Gerard Pons-Moll, Sergi Pujades, Sonny Hu, and Michael Black. ClothCap: Seamless 4D clothing capture and retargeting. *ACM Transactions on Graphics*, 36(4), 2017. 2
- [57] Gerard Pons-Moll, Javier Romero, Naureen Mahmood, and Michael J Black. Dyna: a model of dynamic human shape in motion. *ACM Transactions on Graphics*, 34:120, 2015. 2
- [58] Alin-Ionut Popa, Mihai Zanfir, and Cristian Sminchisescu. Deep multitask architecture for integrated 2d and 3d human sensing. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2017. 3
- [59] Anurag Ranjan, Timo Bolkart, Soubhik Sanyal, and Michael J. Black. Generating 3D faces using convolutional mesh autoencoders. In *European Conf. on Computer Vision*, pages 725–741, 2018. 6
- [60] Helge Rhodin, Nadia Robertini, Dan Casas, Christian Richardt, Hans-Peter Seidel, and Christian Theobalt. General automatic human shape and motion capture using volumetric contour cues. In *European Conf. on Computer Vision*, pages 509–526. Springer, 2016. 1, 2
- [61] Nadia Robertini, Dan Casas, Helge Rhodin, Hans-Peter Seidel, and Christian Theobalt. Model-based outdoor performance capture. In *International Conf. on 3D Vision*, 2016. 2
- [62] Gregory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. Lcr-net: Localization-classification-regression for human pose. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2017. 3
- [63] Lorenz Rogge, Felix Klose, Michael Stengel, Martin Eisenmann, and Marcus Magnor. Garment replacement in monocular video sequences. *ACM Transactions on Graphics*, 34(1):6, 2014. 2
- [64] Ari Shapiro, Andrew Feng, Ruizhe Wang, Hao Li, Mark Bolles, Gerard Medioni, and Evan Suma. Rapid avatar capture and simulation using commodity depth sensors. *Computer Animation and Virtual Worlds*, 25(3-4):201–211, 2014. 2
- [65] Miroslava Slavcheva, Maximilian Baust, Daniel Cremers, and Slobodan Ilic. Killingfusion: Non-rigid 3d reconstruction without correspondences. In *IEEE Conf. on Computer Vision and Pattern Recognition*, volume 3, page 7, 2017. 2
- [66] Cristian Sminchisescu, Atul Kanaujia, and Dimitris Metaxas. Learning joint top-down and bottom-up processes for 3d visual inference. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1743–1752. IEEE, 2006. 3
- [67] Jonathan Starck and Adrian Hilton. Surface capture for performance-based animation. *IEEE Computer Graphics and Applications*, 27(3), 2007. 2
- [68] Carsten Stoll, Nils Hasler, Juergen Gall, Hans-Peter Seidel, and Christian 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
- [69] Xiao Sun, Jiaxiang Shang, Shuang Liang, and Yichen Wei. Compositional human pose regression. In *IEEE International Conf. on Computer Vision*, volume 2, 2017. 3
- [70] Yu Tao, Zerong Zheng, Kaiwen Guo, Jianhui Zhao, Dai Quionhai, Hao Li, G. Pons-Moll, and Yebin Liu. Double-fusion: 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, 3
- [71] Denis Tome, Chris Russell, and Lourdes Agapito. Lifting from the deep: Convolutional 3d pose estimation from a single image. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2017. 3
- [72] Matthew Trumble, Andrew Gilbert, Adrian Hilton, and John Collomosse. Deep autoencoder for combined human pose estimation and body model upscaling. In *European Conf. on Computer Vision*, sep 2018. 3
- [73] Hsiao-Yu Tung, Hsiao-Wei Tung, Ersin Yumer, and Katerina Fragkiadaki. Self-supervised learning of motion capture. In *Advances in Neural Information Processing Systems*, pages 5236–5246, 2017. 3
- [74] Gü̈l Varol, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid. Bodynet: Volumetric inference of 3d human body shapes. In *European Conf. on Computer Vision*, 2018. 3
- [75] Gü̈l Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2017. 3
- [76] Daniel Vlasic, Ilya Baran, Wojciech Matusik, and Jovan Popović. Articulated mesh animation from multi-view silhouettes. In *ACM Transactions on Graphics*, volume 27, page 97. ACM, 2008. 3
- [77] Alexander Weiss, David Hirshberg, and Michael 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
- [78] Jinlong Yang, Jean-Sébastien Franco, Franck Hétroy-Wheeler, and Stefanie Wuhrer. Analyzing clothing layer deformation statistics of 3d human motions. In *European Conf. on Computer Vision*, pages 237–253, 2018. 2
- [79] Mao Ye and Ruigang Yang. Real-time simultaneous pose and shape estimation for articulated objects using a single depth camera. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 2345–2352, 2014. 2
- [80] Rui Yu, Chris Russell, Neill D. F. Campbell, and Lourdes Agapito. Direct, dense, and deformable: Template-based non-rigid 3d reconstruction from rgb video. In *IEEE International Conf. on Computer Vision*, 2015. 2
- [81] Andrei Zanfir, Elisabeta Marinou, and Cristian Sminchisescu. Monocular 3d pose and shape estimation of multiple people in natural scenes—the importance of multiple scene constraints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2148–2157, 2018. 3

- [82] Ming Zeng, Jiaxiang Zheng, Xuan Cheng, and Xinguo Liu. Templateless quasi-rigid shape modeling with implicit loop-closure. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 145–152, 2013. [2](#)
- [83] Chao Zhang, Sergi Pujades, Michael Black, and Gerard Pons-Moll. Detailed, accurate, human shape estimation from clothed 3D scan sequences. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2017. [3](#), [6](#)
- [84] Qing Zhang, Bo Fu, Mao Ye, and Ruigang 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](#)
- [85] Qian-Yi Zhou and Vladlen Koltun. Color map optimization for 3d reconstruction with consumer depth cameras. *ACM Transactions on Graphics*, 33(4):155, 2014. [2](#)
- [86] Shizhe Zhou, Hongbo Fu, Ligang Liu, Daniel Cohen-Or, and Xiaoguang Han. Parametric reshaping of human bodies in images. In *ACM Transactions on Graphics*, volume 29, page 126. ACM, 2010. [1](#), [2](#)
- [87] Xingyi Zhou, Qixing Huang, Xiao Sun, Xiangyang Xue, and Yichen Wei. Towards 3d human pose estimation in the wild: A weakly-supervised approach. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 398–407, 2017. [3](#)
- [88] Michael Zollhöfer, Matthias Nießner, Shahram Izadi, Christoph Rehmann, Christopher Zach, Matthew Fisher, Chenglei Wu, Andrew Fitzgibbon, Charles Loop, Christian Theobalt, et al. Real-time non-rigid reconstruction using an rgbd camera. *ACM Transactions on Graphics*, 33(4):156, 2014. [1](#), [2](#)
- [89] Silvia Zuffi and Michael 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](#)