Expressive Body Capture: 3D Hands, Face, and Body from a Single Image

Georgios Pavlakos[*1,2], Vasileios Choutas[*1], Nima Ghorbani[1], Timo Bolkart[1], Ahmed A. A. Osman[1], Dimitrios Tzionas[1], and Michael J. Black[1]

[1]MPI for Intelligent Systems, Tübingen, DE , [2] University of Pennsylvania, PA, USA
{gpavlakos, vchoutas, nghorbani, tbolkart, aosman, dtzionas, black}@tuebingen.mpg.de

Abstract

To facilitate the analysis of human actions, interactions and emotions, we compute a 3D model of human body pose, hand pose, and facial expression from a single monocular image. To achieve this, we use thousands of 3D scans to train a new, unified, 3D model of the human body, SMPL-X*, that extends SMPL with fully articulated hands and an expressive face. Learning to regress the parameters of SMPL-X directly from images is challenging without paired images and 3D ground truth. Consequently, we follow the approach of SMPLify, which estimates 2D features and then optimizes model parameters to fit the features. We improve on SMPLify in several significant ways: (1) we detect 2D features corresponding to the face, hands, and feet and fit the full SMPL-X model to these; (2) we train a new neural network pose prior using a large MoCap dataset; (3) we define a new interpenetration penalty that is both fast and accurate; (4) we automatically detect gender and the appropriate body models (male, female, or neutral); (5) our PyTorch implementation achieves a speedup of more than* $8\times$ *over Chumpy. We use the new method,* SMPLify-X*, to fit SMPL-X to both controlled images and images in the wild. We evaluate 3D accuracy on a new curated dataset comprising* 100 *images with pseudo ground-truth. This is a step towards automatic expressive human capture from monocular RGB data. The models, code, and data are available for research purposes at* https://smpl-x.is.tue.mpg.de.

1. Introduction

Humans are often a central element in images and videos. Understanding their posture, the social cues they communicate, and their interactions with the world is critical for holistic scene understanding. Recent methods have shown rapid progress on estimating the major body joints, hand joints and facial features in 2D [15, 31, 70]. Our interactions with the world, however, are fundamentally 3D and recent work has also made progress on the 3D estimation of the major joints and rough 3D pose directly from single images [10, 37, 59, 62].

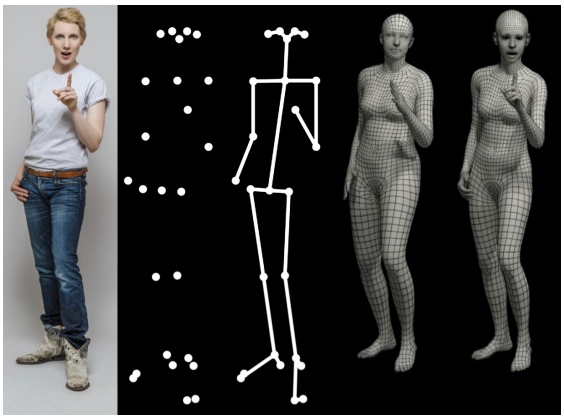

Figure 1: Communication and gesture rely on the *body* pose, *hand* pose, and *facial* expression, all *together*. The major joints of the body are not sufficient to represent this and current 3D models are not expressive enough. In contrast to prior work, our approach estimates a more detailed and expressive 3D model from a single image. From left to right: RGB image, major joints, skeleton, SMPL (female), SMPL-X (female). The hands and face in SMPL-X enable more *holistic* and *expressive* body capture.

To understand human behavior, however, we have to capture more than the major joints of the body – we need the full 3D surface of the body, hands and the face. There is no system that can do this today due to several major challenges including the lack of appropriate 3D models and rich 3D training data. Figure 1 illustrates the problem. The interpretation of expressive and communicative images is difficult using only sparse 2D information or 3D representations that lack hand and face detail. To address this problem, we need two things. First, we need a 3D model of the body that is able to represent the complexity of human faces, hands, and body pose. Second, we need a method to extract such a model from a single image.

Advances in neural networks and large datasets of manually labeled images have resulted in rapid progress in 2D human "pose" estimation. By "pose", the field often means

[*] equal contribution

Figure 2: We learn a new 3D model of the human body called *SMPL-X* that jointly models the human body, face and hands. We fit the female *SMPL-X* model with *SMPLify-X* to single RGB images and show that it captures a rich variety of *natural* and *expressive* 3D human poses, gestures and facial expressions.

the major joints of the body. This is not sufficient to understand human behavior as illustrated in Fig. 1. OpenPose [15, 60, 70] expands this to include the 2D hand joints and 2D facial features. While this captures much more about the communicative intent, it does not support reasoning about surfaces and human interactions with the 3D world.

Models of the 3D body have focused on capturing the overall shape and pose of the body, excluding the hands and face [2, 3, 6, 26, 48]. There is also an extensive literature on modelling hands [39, 53, 57, 58, 68, 69, 71, 74, 75] and faces [4, 9, 11, 13, 14, 43, 63, 76, 79] in 3D but in isolation from the rest of the body. Only recently has the field begun modeling the body together with hands [68], or together with the hands and face [36]. The Frank model [36], for example, combines a simplified version of the SMPL body model [48], with an artist-designed hand rig, and the FaceWarehouse [14] face model. These disparate models are stitched together, resulting in a model that is not fully realistic.

Here we learn a new, holistic, body model with face and hands from a large corpus of 3D scans. The new *SMPL-X* model (*SMPL eXpressive*) is based on SMPL and retains the benefits of that model: compatibility with graphics software, simple parametrization, small size, efficient, differentiable, etc. We combine SMPL with the FLAME head model [43] and the MANO hand model [68] and then register this combined model to 5586 3D scans that we curate for quality. By learning the model from data, we capture the natural correlations between the shape of bodies, faces and hands and the resulting model is free of the artifacts seen with Frank. The expressivity of the model can be seen in Fig. 2 where we fit SMPL-X to expressive RGB images, as well as in Fig. 4 where we fit SMPL-X to images of the public LSP dataset [33]. SMPL-X is freely available for research purposes.

Several methods use deep learning to regress the parameters of SMPL from a single image [37, 59, 62]. To estimate a 3D body with the hands and face though, there exists no suitable training dataset. To address this, we follow the approach of SMPLify. First, we estimate 2D image features "bottom up" using OpenPose [15, 70, 77], which detects the joints of the body, hands, feet, and face features. We then fit the SMPL-X model to these 2D features "top down", with our method called *SMPLify-X*. To do so, we make several significant improvements over SMPLify. Specifically, we learn a new, and better performing, pose prior from a large dataset of motion capture data [50, 51] using a variational auto-encoder. This prior is critical because the mapping from 2D features to 3D pose is ambiguous. We also define a new (self-) interpenetration penalty term that is significantly more accurate and efficient than the approximate method in SMPLify; it remains differentiable. We train a gender detector and use this to automatically determine what body model to use, either male, female or gender neutral. Finally, one motivation for training direct regression methods to estimate SMPL parameters is that SMPLify is slow. Here we address this with a PyTorch implementation that is at least 8 times faster than the corresponding Chumpy implementation, by leveraging the computing power of modern GPUs. Examples of this SMPLify-X method are shown in Fig. 2.

To evaluate the accuracy, we need new data with full-body RGB images and corresponding 3D ground truth bodies. To that end, we curate a new evaluation dataset containing images of a subject performing a wide variety of poses, gestures and expressions. We capture 3D body shape using a scanning system and we fit the SMPL-X model to the scans. This form of pseudo ground-truth is accurate enough to enable quantitative evaluations for models of body, hands and faces together. We find that our model and method performs significantly better than related and less powerful models, resulting in natural and expressive results.

We believe that this work is a significant step towards *expressive* capture of bodies, hands and faces *together* from a single RGB image. We make available for research purposes the SMPL-X model, SMPLify-X code, trained networks, model fits, and the evaluation dataset at https://smpl-x.is.tue.mpg.de.

2. Related work

2.1. Modeling the body

Bodies, Faces and Hands. The problem of modeling the 3D body has previously been tackled by breaking the body into parts and modeling these parts separately. We focus on methods that learn statistical shape models from 3D scans.

Blanz and Vetter [9] pioneered this direction with their 3D morphable face model. Numerous methods since then have learned 3D face shape and expression from scan data; see [13, 81] for recent reviews. A key feature of such models is that they can represent different face shapes and a wide range of expressions, typically using blend shapes inspired by FACS [21]. Most approaches focus only on the face region and not the whole head. FLAME [43], in contrast, models the whole head, captures 3D head rotations, and also models the neck region; we find this critical for connecting the head and the body. None of these methods, model correlations in face shape and body shape.

The availability of 3D body scanners enabled learning of body shape from scans. In particular the CAESAR dataset [67] opened up the learning of shape [2]. Most early work focuses on body shape using scans of people in roughly the same pose. Anguelov et al. [6] combined shape with scans of one subject in many poses to learn a factored model of body shape and pose based on triangle deformations. Many models followed this, either using triangle deformations [16, 23, 26, 29, 64] or vertex-based displacements [3, 27, 48], however they all focus on modeling body shape and pose without the hands or face. These methods assume that the hand is either in a fist or an open pose and that the face is in a neutral expression.

Similarly, hand modeling approaches typically ignore the body. Additionally, 3D hand models are typically not learned but either are artist designed [71], based on shape primitives [53, 58, 69], reconstructed with multiview stereo and have fixed shape [8, 75], use non-learned per-part scaling parameters [19], or use simple shape spaces [74]. Only recently [39, 68] have learned hand models appeared in the literature. Khamis *et al*. [39] collect partial depth maps of 50 people to learn a model of shape variation, however they do not capture a pose space. Romero *et al*. [68] on the other side learn a parametric hand model (MANO) with both a rich shape and pose space using 3D scans of 31 subjects in up to 51 poses, following the SMPL [48] formulation.

Unified Models. The most similar models to ours are Frank [36] and SMPL+H [68]. Frank stitches together three different models: SMPL (with no pose blend shapes) for the body, an artist-created rig for the hands, and the FaceWarehouse model [14] for the face. The resulting model is not fully realistic. SMPL+H combines the SMPL body with a 3D hand model that is learned from 3D scans. The shape variation of the hand comes from full body scans, while the pose dependent deformations are learned from a dataset of hand scans. SMPL+H does not contain a deformable face.

We start from the publicly-available SMPL+H [52] and add the publicly-available FLAME head model [22] to it. Unlike Frank, however, we do not simply graft this onto the body. Instead we take the full model and fit it to 5586 3D scans and learn the shape and pose-dependent blend shapes. This results in a natural looking model with a consistent parameterization. Being based on SMPL, it is differentiable and easy to swap into applications that already use SMPL.

2.2. Inferring the body

There are many methods that estimate 3D faces from images or RGB-D [81] as well as methods that estimate hands from such data [80]. While there are numerous methods that estimate the location of 3D joints from a single image, here we focus on methods that extract a full 3D body mesh.

Several methods estimate the SMPL model from a single image [37, 41, 59, 62]. This is not trivial due to a paucity of training images with paired 3D model parameters. To address this, SMPLify [10] detects 2D image features "bottom up" and then fits the SMPL model to these "top down" in an optimization framework. In [41] these SMPLify fits are used to iteratively curate a training set of paired data to train a direct regression method. HMR [37] trains a model without paired data by using 2D keypoints and an adversary that knows about 3D bodies. Like SMPLify, NBF [59] uses an intermediate 2D representation (body part segmentation) and infers 3D pose from this intermediate representation. MonoPerfCap [78] infers 3D pose while also refining surface geometry to capture clothing. These methods estimate only the 3D pose of the body without the hands or face.

There are also many multi-camera setups for capturing 3D pose, 3D meshes (performance capture), or parametric

3D models [7, 20, 24, 30, 35, 46, 54, 66, 72]. Most relevant is the Panoptic studio [35] which shares our goal of capturing rich, expressive, human interactions. In [36], the Frank model parameters are estimated from multi-camera data by fitting the model to 3D keypoints and 3D point clouds. The capture environment is complex, using 140 VGA cameras for the body, 480 VGA cameras for the feet, and 31 HD cameras for the face and hand keypoints. We aim for a similar level of expressive detail but from a *single RGB image*.

3. Technical approach

In the following we describe SMPL-X (Section 3.1), and our approach (Section 3.2) for fitting SMPL-X to single RGB images. Compared to SMPLify [10], SMPLify-X uses a better pose prior (Section 3.3), a more detailed collision penalty (Section 3.4), gender detection (Section 3.5), and a faster PyTorch implementation (Section 3.6).

3.1. Unified model: SMPL-X

We create a unified model, called *SMPL-X*, for *SMPL eXpressive*, with shape parameters trained jointly for the face, hands and body. SMPL-X uses standard vertex-based linear blend skinning with learned corrective blend shapes, has $N = 10,475$ vertices and $K = 54$ joints, which includes joints for the neck, jaw, eyeballs and fingers. SMPL-X is defined by a function $M(\theta, \beta, \psi) : \mathbb{R}^{|\theta| \times |\beta| \times |\psi|} \rightarrow \mathbb{R}^{3N}$, parameterized by the pose $\theta \in \mathbb{R}^{3(K+1)}$ where K is the number of body joints in addition to a joint for global rotation. We decompose the pose parameters θ into: θ_f for the jaw joint, θ_h for the finger joints, and θ_b for the remaining body joints. The joint body, face and hands shape parameters are noted by $\beta \in \mathbb{R}^{|\beta|}$ and the facial expression parameters by $\psi \in \mathbb{R}^{|\psi|}$. More formally:

$$M(\beta, \theta, \psi) = W(T_p(\beta, \theta, \psi), J(\beta), \theta, \mathcal{W}) \quad (1)$$

$$T_P(\beta, \theta, \psi) = \bar{T} + B_S(\beta; \mathcal{S}) + B_E(\psi; \mathcal{E}) + B_P(\theta; \mathcal{P}) \quad (2)$$

where $B_S(\beta; \mathcal{S}) = \sum_{n=1}^{|\beta|} \beta_n \mathcal{S}_n$ is the shape blend shape function, β are linear shape coefficients, $|\beta|$ is their number, $\mathcal{S}_n \in \mathbb{R}^{3N}$ are orthonormal principle components of vertex displacements capturing shape variations due to different person identity, and $\mathcal{S} = \left[S_1, \ldots, S_{|\beta|}\right] \in \mathbb{R}^{3N \times |\beta|}$ is a matrix of all such displacements. $B_P(\theta; \mathcal{P}) : \mathbb{R}^{|\theta|} \rightarrow \mathbb{R}^{3N}$ is the pose blend shape function, which adds corrective vertex displacements to the template mesh $\bar{T}$ as in SMPL [47]:

$$B_P(\theta; \mathcal{P}) = \sum_{n=1}^{9K} (R_n(\theta) - R_n(\theta^*)) \mathcal{P}_n, \quad (3)$$

where $R : \mathbb{R}^{|\theta|} \rightarrow \mathbb{R}^{9K}$ is a function mapping the pose vector θ to a vector of concatenated part-relative rotation matrices, computed with the Rodrigues formula [12, 55, 65] and $R_n(\theta)$ is the n^{th} element of $R(\theta)$, θ^* is the pose vector of the rest pose, $\mathcal{P}_n \in \mathbb{R}^{3N}$ are again orthonormal principle components of vertex displacements, and $\mathcal{P} = [P_1, \ldots, P_{9K}] \in \mathbb{R}^{3N \times 9K}$ is a matrix of all pose blend shapes. $B_E(\psi; \mathcal{E}) = \sum_{n=1}^{|\psi|} \psi_n \mathcal{E}$ is the expression blend shape function, where $\mathcal{E}$ are principle components capturing variations due to facial expressions and ψ are PCA coefficients. Since 3D joint locations J vary between bodies of different shapes, they are a function of body shape $J(\beta) = \mathcal{J}\left(\bar{T} + B_S(\beta; \mathcal{S})\right)$, where $\mathcal{J}$ is a sparse linear regressor that regresses 3D joint locations from mesh vertices. A standard linear blend skinning function $W(.)$ [42] rotates the vertices in $T_p(.)$ around the estimated joints $J(\beta)$ smoothed by blend weights $\mathcal{W} \in \mathbb{R}^{N \times K}$.

We start with an artist designed 3D template, whose face and hands match the templates of FLAME [43] and MANO [68]. We fit the template to four datasets of 3D human scans to get 3D alignments as training data for SMPL-X. The shape space parameters, $\{\mathcal{S}\}$, are trained on 3800 alignments in an A-pose capturing variations across identities [67]. The body pose space parameters, $\{\mathcal{W}, \mathcal{P}, \mathcal{J}\}$, are trained on 1786 alignments in diverse poses. Since the full body scans have limited resolution for the hands and face, we leverage the parameters of MANO [68] and FLAME [43], learned from 1500 hand and 3800 head high resolution scans respectively. More specifically, we use the pose space and pose corrective blendshapes of MANO for the hands and the expression space $\mathcal{E}$ of FLAME.

The fingers have 30 joints, which correspond to 90 pose parameters (3 DoF per joint as axis-angle rotations). SMPL-X uses a lower dimensional PCA pose space for the hands such that $\theta_h = \sum_{n=1}^{|m_h|} m_{h_n} \mathcal{M}$, where $\mathcal{M}$ are principle components capturing the finger pose variations and m_h are the corresponding PCA coefficients. As noted above, we use the PCA pose space of MANO, that is trained on a large dataset of 3D articulated human hands. The total number of model parameters in SMPL-X is 119: 75 for the global body rotation and { body, eyes , jaw } joints, 24 parameters for the lower dimensional hand pose PCA space, 10 for subject shape and 10 for the facial expressions. Additionally there are separate male and female models, which are used when the gender is known, and a shape space constructed from both genders for when gender is unknown. SMPL-X is realistic, expressive, differentiable and easy to fit to data.

3.2. SMPLify-X: SMPL-X from a single image

To fit SMPL-X to single RGB images (*SMPLify-X*), we follow SMPLify [10] but improve every aspect of it. We formulate fitting SMPL-X to the image as an optimization problem, where we seek to minimize the objective function

$$E(\beta, \theta, \psi) = E_J + \lambda_{\theta_b} E_{\theta_b} + \lambda_{\theta_f} E_{\theta_f} + \lambda_{m_h} E_{m_h} + \lambda_\alpha E_\alpha + \lambda_\beta E_\beta + \lambda_\mathcal{E} E_\mathcal{E} + \lambda_\mathcal{C} E_\mathcal{C} \quad (4)$$

where θ_b, θ_f and m_h are the pose vectors for the body, face and the two hands respectively, and θ is the full set of optimizable pose parameters. The body pose parameters are a function $\theta_b(Z)$, where $Z \in \mathbb{R}^{32}$ is a lower-dimensional pose space described in Section 3.3. $E_J(\beta, \theta, K, J_{est})$ is the data term as described below, while the terms $E_{m_h}(m_h)$, $E_{\theta_f}(\theta_f)$, $E_\beta(\beta)$ and $E_\mathcal{E}(\psi)$ are simple L_2 priors for the hand pose, facial pose, body shape and facial expressions, penalizing deviation from the neutral state. Since the shape space of SMPL-X is scaled for unit variance, similarly to [68], $E_\beta(\beta) = \|\beta\|^2$ describes the Mahalanobis distance between the shape parameters being optimized and the shape distribution in the training dataset of SMPL-X. $E_\alpha(\theta_b) = \sum_{i \in (elbows, knees)} \exp(\theta_i)$ follows [10] and is a simple prior penalizing extreme bending only for elbows and knees. We further employ $E_{\theta_b}(\theta_b)$ that is a VAE-based body pose prior (Section 3.3), while $E_\mathcal{C}(\theta_{b,h,f}, \beta)$ is an interpenetration penalty (Section 3.4). Finally, λ denotes weights that steer the influence of each term in Eq. 4. We empirically find that an annealing scheme for λ helps optimization (Section 3.6).

For the *data term* we use a re-projection loss to minimize the weighted robust distance between estimated 2D joints J_{est} and the 2D projection of the corresponding posed 3D joints $R_\theta(J(\beta))_i$ of SMPL-X for each joint i, where $R_\theta(\cdot)$ is a function that transforms the joints along the kinematic tree according to the pose θ. Following the notation of [10], the data term is $E_J(\beta, \theta, K, J_{est}) =$

$$\sum_{joint\ i} \gamma_i \omega_i \rho(\Pi_K(R_\theta(J(\beta))_i) - J_{est,i}) \tag{5}$$

where Π_K denotes the 3D to 2D projection with intrinsic camera parameters K. For the 2D detections we rely on the OpenPose library [15, 70, 77], which provides body, hands, face and feet keypoints jointly for each person in an image. To account for noise in the detections, the contribution of each joint in the data term is weighted by the detection confidence score ω_i, while γ_i are per-joint weights for annealed optimization, as described in Section 3.6. Finally, ρ denotes a robust Geman-McClure error function [25] for down weighting noisy detections.

3.3. Variational Human Body Pose Prior

We seek a prior over body pose that penalizes impossible poses while allowing possible ones. SMPLify uses an approximation to the negative log of a Gaussian mixture model trained on MoCap data. While effective, we find that the SMPLify prior is not sufficiently strong. Consequently, we train our body pose prior, VPoser, using a variational autoencoder [40], which learns a latent representation of human pose and regularizes the distribution of the latent code to be a normal distribution. We train our prior on the data released by [50, 51], namely pose parameters obtained by applying MoSh [47] on three publicly available human motion capture datasets: CMU [17], training set of Human3.6M [32], and the PosePrior dataset [1]. Our training and test data respectively consist of roughly 1M, and 65k poses, in rotation matrix representation. Details on the data preparation procedure is given in Sup. Mat.

The training loss of the VAE is formulated as:

$$\mathcal{L}_{total} = c_1\mathcal{L}_{KL} + c_2\mathcal{L}_{rec} + c_3\mathcal{L}_{orth} + c_4\mathcal{L}_{det1} + c_5\mathcal{L}_{reg} \tag{6}$$

$$\mathcal{L}_{KL} = KL(q(Z|R)||\mathcal{N}(0, I)) \tag{7}$$

$$\mathcal{L}_{rec} = ||R - \hat{R}||_2^2 \tag{8}$$

$$\mathcal{L}_{orth} = ||\hat{R}\hat{R}^{'} - I||_2^2 \tag{9}$$

$$\mathcal{L}_{det1} = |det(\hat{R}) - 1| \tag{10}$$

$$\mathcal{L}_{reg} = ||\phi||_2^2, \tag{11}$$

where $Z \in \mathbb{R}^{32}$ is the latent space of the autoencoder, $R \in SO(3)$ are 3×3 rotation matrices for each joint as the network input and $\hat{R}$ is a similarly shaped matrix representing the output. The Kullback-Leibler term in Eq. (7), and the reconstruction term in Eq. (8) follow the VAE formulation in [40], while their role is to encourage a normal distribution on the latent space, and to make an efficient code to reconstruct the input with high fidelity. Eq. (9) and (10) encourage the latent space to encode valid rotation matrices. Finally, Eq. (11) helps prevent over-fitting by encouraging smaller network weights ϕ. Implementation details can be found in Sup. Mat.

To employ VPoser in the optimization, rather than to optimize over θ_b directly in Eq. 4, we optimize the parameters of a 32 dimensional latent space with a quadratic penalty on Z and transform this back into joint angles θ_b in axis-angle representation. This is analogous to how hands are treated except that the hand pose θ_h is projected into a linear PCA space and the penalty is on the linear coefficients.

3.4. Collision penalizer

When fitting a model to observations, there are often self-collisions and penetrations of several body parts that are physically impossible. Our approach is inspired by SMPLify, that penalizes penetrations with an underlying collision model based on shape primitives, *i.e.* an ensemble of capsules. Although this model is computationally efficient, it is only a rough approximation of the human body.

For models like SMPL-X, that also model the fingers and facial details, a more accurate collision model in needed. To that end, we employ the detailed collision-based model for meshes from [8, 75]. We first detect a list of colliding triangles $\mathcal{C}$ by employing Bounding Volume Hierarchies (BVH) [73] and compute local conic 3D distance fields Ψ defined by the triangles $\mathcal{C}$ and their normals n. Penetrations are then penalized by the depth of intrusion, efficiently computed by

the position in the distance field. For two colliding triangles f_s and f_t, intrusion is bi-directional; the vertices v_t of f_t are the *intruders* in the distance field Ψ_{f_s} of the *receiver* triangle f_s and are penalized by $\Psi_{f_s}(v_t)$, and vice-versa. Thus, the collision term $E_{\mathcal{C}}$ in the objective (Eq. 4) is defined as

$$E_{\mathcal{C}}(\theta) = \sum_{(f_s(\theta), f_t(\theta)) \in \mathcal{C}} \left\{ \sum_{v_s \in f_s} \| - \Psi_{f_t}(v_s) n_s \|^2 + \sum_{v_t \in f_t} \| - \Psi_{f_s}(v_t) n_t \|^2 \right\}. \tag{12}$$

For technical details about Ψ, as well as details about handling collisions for parts with permanent or frequent self-contact we redirect the reader to [8, 75] and Sup. Mat.. For computational efficiency, we use a highly parallelized implementation of BVH following [38] with a custom CUDA kernel wrapped around a custom PyTorch operator.

3.5. Deep Gender Classifier

Men and women have different proportions and shapes. Consequently, using the appropriate body model to fit 2D data means that we should apply the appropriate shape space. We know of no previous method that automatically takes gender into account in fitting 3D human pose. In this work, we train a gender classifier that takes as input an image containing the full body and the OpenPose joints, and assigns a gender label to the detected person. To this end, we first annotate through Amazon Mechanical Turk a large dataset of images from LSP [33], LSP-extended [34], MPII [5], MS-COCO[45], and LIP datset [44], while following their official splits for train and test sets. The final dataset includes 50216 training examples and 16170 test samples (see Sup. Mat.). We use this dataset to fine tune a pretrained ResNet18 [28] for binary gender classification. Moreover, we threshold the computed class probabilities, by using a class-equalized validation set, to obtain a good trade-off between discarded, correct, and incorrect predictions. We choose a threshold of 0.9 for accepting a predicted class, which yields 62.38% correct predictions, and 7.54% incorrect predictions on the validation set. At test time, we run the detector and fit the appropriate gendered model. When the detected class probability is below the threshold, we fit the gender-neutral body model.

3.6. Optimization

SMPLify employs Chumpy and OpenDR [49] which makes the optimization slow. To keep optimization of Eq. 4 tractable, we use PyTorch and the Limited-memory BFGS optimizer (L-BFGS) [56] with strong Wolfe line search. Implementation details can be found in Sup. Mat.

We optimize Eq. 4 with a multistage approach, similar to [10]. We assume that we know the exact or an approximate value for the focal length of the camera. Then we first estimate the unknown camera translation and global body orientation (see [10]). We then fix the camera parameters and optimize body shape, β, and pose, θ. Empirically, we found that an *annealing scheme* for the weights γ in the data term E_J (Eq. 5) helps optimization of the objective (Eq. 4) to deal with ambiguities and local optima. This is mainly motivated by the fact that small body parts like the hands and face have many keypoints relative to their size, and can dominate in Eq. 4, throwing optimization in a local optimum when the initial estimate is away from the solution.

In the following, we denote by γ_b the weights corresponding to the main body keypoints, γ_h the ones for hands and γ_f the ones for facial keypoints. We then follow three steps, starting with high regularization to mainly refine the global body pose, and gradually increase the influence of hand keypoints to refine the pose of the arms. After converging to a better pose estimate, we increase the influence of both hands and facial keypoints to capture expressivity. Throughout the above steps the weights $\lambda_\alpha, \lambda_\beta, \lambda_\mathcal{E}$ in Eq.4 start with high regularization that gradually lowers to allow for better fitting, The only exception is $\lambda_\mathcal{C}$ that gradually increases while the influence of hands gets stronger in E_J and more collisions are expected.

4. Experiments

4.1. Evaluation datasets

Despite the recent interest in more expressive models [36, 68] there exists no dataset containing images with ground-truth shape for bodies, hands and faces together. Consequently, we create a dataset for evaluation from currently available data through fitting and careful curation.

Expressive hands and faces dataset (EHF). We begin with the SMPL+H dataset [52], obtaining one full body RGB image per frame. We then align SMPL-X to the 4D scans following [68]. An expert annotator manually curated the dataset to select 100 frames that can be confidently considered pseudo ground-truth, according to alignment quality and interesting hand poses and facial expressions. The pseudo ground-truth meshes allow to use a stricter *vertex-to-vertex (v2v)* error metric [48, 62], in contrast to the common paradigm of reporting 3D joint error, which does not capture surface errors and rotations along the bones.

4.2. Qualitative & Quantitative evaluations

To test the effectiveness of SMPL-X and SMPLify-X, we perform comparisons to the most related models, namely SMPL [48], SMPL+H [68], and Frank [36]. In this direction we fit SMPL-X to the EHF images to evaluate both *qualitatively* and *quantitatively*. Note that we use *only* 1 image and 2D joints as input, while previous methods use *much more* information; *i.e.* 3D point clouds [36, 68] and joints [36]. Specifically [48, 68] employ 66 cameras and 34

Model	Keypoints	v2v error	Joint error
"SMPL"	Body	57.6	63.5
"SMPL"	Body+Hands+Face	64.5	71.7
"SMPL+H"	Body+Hands	54.2	63.9
SMPL-X	Body+Hands+Face	52.9	62.6

Table 1: Quantitative comparison of "SMPL", "SMPL+H" and SMPL-X, as described in Section 4.2, fitted with SMPLify-X on the EHF dataset. We report the mean vertex-to-vertex (v2v) and the standard mean 3D body (only) joint error in mm. The table shows that richer modeling power results in lower errors.

Version	v2v error
SMPLify-X	52.9
gender neutral model	58.0
replace Vposer with GMM	56.4
no collision term	53.5

Table 2: Ablative study for SMPLify-X on the EHF dataset. The numbers reflect the contribution of each component in overall accuracy.

projectors, while [36] employ more than 500 cameras.

We first compare to SMPL, SMPL+H and SMPL-X on the EHF dataset and report results in Table 1. The table reports *mean vertex-to-vertex (v2v)* error and *mean 3D body joint* error after Procrustes alignment with the ground-truth 3D meshes and body (only) joints respectively. To ease numeric evaluation, for this table only we "simulate" SMPL and SMPL+H with a SMPL-X variation with locked degrees of freedom, noted as "SMPL" and "SMPL+H" respectively. As expected, the errors show that the standard mean 3D joint error fails to capture accurately the difference in model expressivity. On the other hand, the much stricter v2v metric shows that enriching the body with finger and face modeling results in lower errors. We also fit SMPL with additional features for parts that are not properly modeled, *e.g.* finger features. The additional features result in an increasing error, pointing to the importance of richer and more expressive models. We report similar qualitative comparisons in Sup. Mat.

We then perform an ablative study, summarized in Table 2, where we report the *mean vertex-to-vertex (v2v)* error. SMPLify-X with a gender-specific model achieves 52.9 mm error. The gender neutral model is easier to use, as it does not need gender detection, but comes with a small compromise in terms of accuracy. Replacing VPoser with the GMM of SMPLify [10] increases the error to 56.4 mm, showing the effectiveness of VPoser. Finally, removing the collision term increases the error as well, to 53.5 mm, while also allowing for non physically plausible pose estimates.

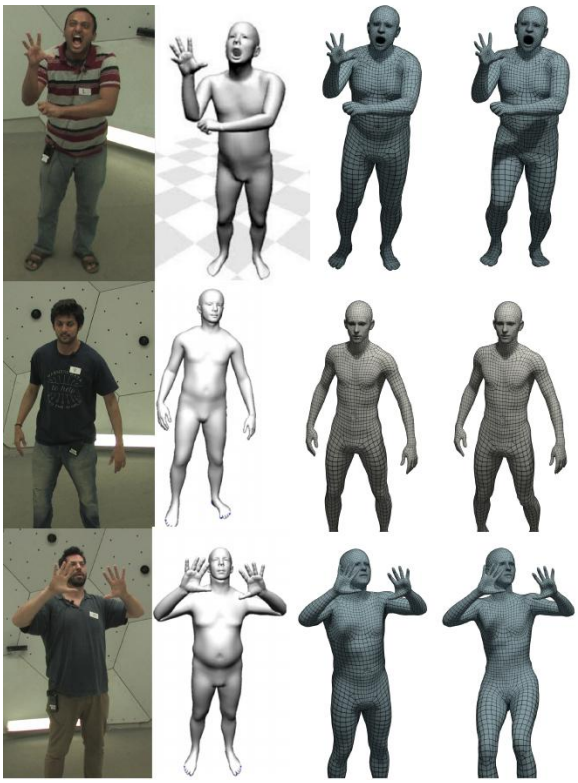

Figure 3: Qualitative comparison of our gender neutral model (top, bottom rows) or gender specific model (middle) against Frank [36] on some of their data. To fit Frank, [36] employ both 3D joints and point cloud, *i.e.* more than 500 cameras. In contrast, our method produces a realistic and expressive reconstruction using *only* 2D joints. We show results using the 3D joints of [36] projected in 1 camera view (third column), as well as using joints estimated from only 1 image (last column), to show the influence of noise in 2D joint detection. Compared to Frank, our SMPL-X does *not* have skinning artifacts around the joints, *e.g.* elbows.

The closest comparable model to SMPL-X is Frank [36]. Since Frank is not available to date, nor are the fittings to [18], we show images of results found online. Figure 3 shows Frank fittings to 3D joints *and* point clouds, *i.e.* using more than 500 cameras. Compare this with SMPL-X fitting that is done with SMPLify-X using *only* 1 RGB image with 2D joints. For a more direct comparison here, we fit SMPL-X to 2D projections of the 3D joints that [36] used for Frank. Although we use *much less* data, SMPL-X shows at least similar expressivity to Frank for both the face and hands. Since Frank does not use pose blend shapes, it suffers from skinning artifacts around the joints, *e.g.* elbows, as clearly seen in Figure 3. SMPL-X by contrast, is trained to include pose blend shapes and does not suffer from this. As a result it looks more *natural* and *realistic*.

Figure 4: Qualitative results of SMPL-X for the in-the-wild images of the LSP dataset [33]. A strong holistic model like SMPL-X results in *natural* and *expressive* reconstruction of bodies, hands and faces. Gray color depicts the gender-specific model for confident gender detections. Blue is the gender-neutral model that is used when the gender classifier is uncertain.

Figure 5: Comparison of the hands-only approach of [61] (middle) against our approach with the male model (right). Both approaches depend on OpenPose. In case of good detections both perform well (top). In case of noisy 2D detections (bottom) our holistic model shows increased robustness. (images cropped at the bottom in the interest of space)

To further show the value of a holistic model of the body, face and hands, in Fig. 5 we compare SMPL-X and SMPLify-X to the hands-only approach of [61]. Both approaches employ OpenPose for 2D joint detection, while [61] further depends on a hand detector. As seen in Fig. 5, in case of good detections both approaches perform nicely, though in case of noisy detections, SMPL-X shows increased robustness due to the context of the body. We further perform quantitative comparison after aligning the resulting fittings to EHF. Due to different mesh topology, for simplicity we use hand joints as pseudo ground-truth, and perform Procrustes analysis of each hand independently, ignoring the body. Panteleris *et al*. [61] achieve a mean 3D joint error of 26.5 mm, while SMPL-X has 19.8 mm.

Finally, we fit SMPL-X with SMPLify-X to some in-the-wild datasets, namely the LSP [33], LSP-extended [34] and MPII datasets [5]. Figure 4 shows some qualitative results for the LSP dataset [33]; see Sup. Mat. for more examples and failure cases. The images show that a strong holistic model like SMPL-X can effectively give *natural* and *expressive* reconstruction from everyday images.

5. Conclusion

In this work we present SMPL-X, a new model that *jointly* captures the body together with face and hands. We additionally present SMPLify-X, an approach to fit SMPL-X to a single RGB image and 2D OpenPose joint detections. We regularize fitting under ambiguities with a new powerful body pose prior and a fast and accurate method for detecting and penalizing penetrations. We present a wide range of qualitative results using images in-the-wild, showing the expressivity of SMPL-X and effectiveness of SMPLify-X. We introduce a curated dataset with pseudo ground-truth to perform quantitative evaluation, that shows the importance of more expressive models. In future work we will curate a dataset of in-the-wild SMPL-X fits and learn a regressor to directly regress SMPL-X parameters directly from RGB images. We believe that this work is an important step towards *expressive* capture of bodies, hands and faces *together* from an RGB image.

Acknowledgements: We thank Joachim Tesch for the help with Blender rendering and Pavel Karasik for the help with Amazon Mechanical Turk. We thank Soubhik Sanyal for the face-only baseline, Panteleris *et al*. from FORTH for running their hands-only method [61] on the EHF dataset, and Joo *et al*. from CMU for providing early access to their data [36].

Disclosure: MJB has received research gift funds from Intel, Nvidia, Adobe, Facebook, and Amazon. While MJB is a part-time employee of Amazon, his research was performed solely at, and funded solely by, MPI. MJB has financial interests in Amazon and Meshcapade GmbH.

References

[1] Ijaz Akhter and Michael J. Black. Pose-conditioned joint angle limits for 3D human pose reconstruction. In *CVPR*, 2015. 5

[2] Brett Allen, Brian Curless, and Zoran Popović. The space of human body shapes: Reconstruction and parameterization from range scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH)*, 22(3):587–594, 2003. 2, 3

[3] Brett Allen, Brian Curless, Zoran Popović, and Aaron Hertzmann. Learning a correlated model of identity and pose-dependent body shape variation for real-time synthesis. In *ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, SCA '06, pages 147–156. Eurographics Association, 2006. 2, 3

[4] Brian Amberg, Reinhard Knothe, and Thomas Vetter. Expression invariant 3D face recognition with a morphable model. In *International Conference on Automatic Face Gesture Recognition*, 2008. 2

[5] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2D human pose estimation: New benchmark and state of the art analysis. In *CVPR*, 2014. 6, 8

[6] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. SCAPE: Shape Completion and Animation of PEople. *ACM Transactions on Graphics, (Proc. SIGGRAPH)*, 24(3):408–416, 2005. 2, 3

[7] Luca Ballan and Guido Maria Cortelazzo. Marker-less motion capture of skinned models in a four camera set-up using optical flow and silhouettes. In *International Symposium on 3D Data Processing, Visualization and Transmission (3DPVT)*, 2008. 4

[8] Luca Ballan, Aparna Taneja, Juergen Gall, Luc Van Gool, and Marc Pollefeys. Motion capture of hands in action using discriminative salient points. In *ECCV*, 2012. 3, 5, 6

[9] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3D faces. In *SIGGRAPH*, pages 187–194, 1999. 2, 3

[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 *ECCV*, 2016. 1, 3, 4, 5, 6, 7

[11] James Booth, Anastasios Roussos, Allan Ponniah, David Dunaway, and Stefanos Zafeiriou. Large scale 3D morphable models. *IJCV*, 126(2-4):233–254, 2018. 2

[12] Christoph Bregler, Jitendra Malik, and Katherine Pullen. Twist based acquisition and tracking of animal and human kinematics. *International Journal of Computer Vision (IJCV)*, 56(3):179–194, 2004. 4

[13] Alan Brunton, Augusto Salazar, Timo Bolkart, and Stefanie Wuhrer. Review of statistical shape spaces for 3D data with comparative analysis for human faces. *CVIU*, 128(0):1–17, 2014. 2, 3

[14] Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. Facewarehouse: A 3D facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3):413–425, 2014. 2, 3

[15] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2D pose estimation using part affinity fields. In *CVPR*, 2017. 1, 2, 5

[16] Yinpeng Chen, Zicheng Liu, and Zhengyou Zhang. Tensor-based human body modeling. In *CVPR*, 2013. 3

[17] CMU. CMU MoCap dataset. 5

[18] Total Capture Dataset. http://domedb.perception.cs.cmu.edu. 7

[19] Martin De La Gorce, David J. Fleet, and Nikos Paragios. Model-based 3D hand pose estimation from monocular video. *PAMI*, 33(9):1793–1805, 2011. 3

[20] Quentin Delamarre and Olivier D. Faugeras. 3D articulated models and multiview tracking with physical forces. *CVIU*, 81(3):328–357, 2001. 4

[21] P. Ekman and W. Friesen. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, 1978. 3

[22] models FLAME website: dataset and code. http://flame.is.tue.mpg.de. 3

[23] Oren Freifeld and Michael J. Black. Lie bodies: A manifold representation of 3D human shape. In *ECCV*, 2012. 3

[24] 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 *CVPR*, 2009. 4

[25] Stuart Geman and Donald E. McClure. Statistical methods for tomographic image reconstruction. In *Proceedings of the 46th Session of the International Statistical Institute, Bulletin of the ISI*, volume 52, 1987. 5

[26] Nils Hasler, Carsten Stoll, Martin Sunkel, Bodo Rosenhahn, and Hans-Peter Seidel. A statistical model of human pose and body shape. *Computer Graphics Forum*, 28(2):337–346, 2009. 2, 3

[27] Nils Hasler, Thorsten Thormählen, Bodo Rosenhahn, and Hans-Peter Seidel. Learning skeletons for shape and pose. In *Proceedings of the 2010 ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games*, I3D '10, pages 23–30, New York, NY, USA, 2010. ACM. 3

[28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6

[29] David A. Hirshberg, Matthew Loper, Eric Rachlin, and Michael J. Black. Coregistration: Simultaneous alignment and modeling of articulated 3D shape. In *ECCV*, 2012. 3

[30] Yinghao Huang, Federica Bogo, Christoph Lassner, Angjoo Kanazawa, Peter V. Gehler, Javier Romero, Ijaz Akhter, and Michael J. Black. Towards accurate marker-less human shape and pose estimation over time. In *3DV*, 2017. 4

[31] Eldar Insafutdinov, Leonid Pishchulin, Bjoern Andres, Mykhaylo Andriluka, and Bernt Schiele. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *ECCV*, 2016. 1

[32] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *PAMI*, 36(7):1325–1339, 2014. 5

[33] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *BMVC*, 2010. 2, 6, 8

[34] Sam Johnson and Mark Everingham. Learning effective human pose estimation from inaccurate annotation. In *CVPR*, 2011. 6, 8

[35] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *ICCV*, 2015. 4

[36] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3D deformation model for tracking faces, hands, and bodies. In *CVPR*, 2018. 2, 3, 4, 6, 7, 8

[37] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018. 1, 2, 3

[38] Tero Karras. Maximizing parallelism in the construction of BVHs, Octrees, and K-d trees. In *Proceedings of the Fourth ACM SIGGRAPH / Eurographics Conference on High-Performance Graphics*, pages 33–37, 2012. 6

[39] Sameh Khamis, Jonathan Taylor, Jamie Shotton, Cem Keskin, Shahram Izadi, and Andrew Fitzgibbon. Learning an efficient model of hand shape variation from depth images. In *CVPR*, 2015. 2, 3

[40] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014. 5

[41] 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 *CVPR*, 2017. 3

[42] John P. Lewis, Matt Cordner, and Nickson Fong. Pose space deformation: A unified approach to shape interpolation and skeleton-driven deformation. In *ACM Transactions on Graphics (SIGGRAPH)*, pages 165–172, 2000. 4

[43] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics (TOG)*, 36(6):194, 2017. 2, 3, 4

[44] Xiaodan Liang, Chunyan Xu, Xiaohui Shen, Jianchao Yang, Si Liu, Jinhui Tang, Liang Lin, and Shuicheng Yan. Human parsing with contextualized convolutional neural network. In *ICCV*, 2015. 6

[45] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014. 6

[46] Yebin Liu, Juergen Gall, Carsten Stoll, Qionghai Dai, Hans-Peter Seidel, and Christian Theobalt. Markerless motion capture of multiple characters using multiview image segmentation. *PAMI*, 35(11):2720–2735, 2013. 4

[47] Matthew Loper, Naureen Mahmood, and Michael J Black. MoSh: Motion and shape capture from sparse markers. *ACM Transactions on Graphics (TOG)*, 33(6):220, 2014. 4, 5

[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, (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, Oct. 2015. 2, 3, 6

[49] Matthew M Loper and Michael J Black. OpenDR: An approximate differentiable renderer. In *ECCV*, 2014. 6

[50] Matthew M. Loper, Naureen Mahmood, and Michael J. Black. MoSh: Motion and shape capture from sparse markers. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 33(6):220:1–220:13, Nov. 2014. 2, 5

[51] N. Mahmood, N. Ghorbani, Troje N.F., G. Pons-Moll, and M.J. Black. AMASS: Archive of motion capture as surface shapes. In *preparation*, 2019. 2, 5

[52] MANO, models SMPL+H website: dataset, and code. http://mano.is.tue.mpg.de. 3, 6

[53] Stan Melax, Leonid Keselman, and Sterling Orsten. Dynamics based 3D skeletal hand tracking. In *Graphics Interface*, pages 63–70, 2013. 2, 3

[54] Thomas B. Moeslund, Adrian Hilton, and Volker Krüger. A survey of advances in vision-based human motion capture and analysis. *CVIU*, 104(2):90–126, 2006. 4

[55] Richard M. Murray, Li Zexiang, and S. Shankar Sastry. *A Mathematical Introduction to Robotic Manipulation*. CRC press, 1994. 4

[56] Jorge Nocedal and Stephen J Wright. *Nonlinear Equations*. Springer, 2006. 6

[57] Markus Oberweger, Paul Wohlhart, and Vincent Lepetit. Training a feedback loop for hand pose estimation. In *ICCV*, 2015. 2

[58] Iason Oikonomidis, Nikolaos Kyriazis, and Antonis A. Argyros. Efficient model-based 3D tracking of hand articulations using Kinect. In *BMVC*, 2011. 2, 3

[59] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter V Gehler, and Bernt Schiele. Neural body fitting: Unifying deep learning and model-based human pose and shape estimation. In *3DV*, 2018. 1, 2, 3

[60] OpenPose. https://github.com/CMU-Perceptual-Computing-Lab/openpose. 2

[61] Paschalis Panteleris, Iason Oikonomidis, and Antonis Argyros. Using a single RGB frame for real time 3D hand pose estimation in the wild. In *WACV*, 2018. 8

[62] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3D human pose and shape from a single color image. In *CVPR*, 2018. 1, 2, 3, 6

[63] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3D face model for pose and illumination invariant face recognition. In *2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 296–301, 2009. 2

[64] Gerard Pons-Moll, Javier Romero, Naureen Mahmood, and Michael J. Black. Dyna: A model of dynamic human shape in motion. *ACM Transactions on Graphics, (Proc. SIGGRAPH)*, 34(4):120:1–120:14, July 2015. 3

[65] Gerard Pons-Moll and Bodo Rosenhahn. *Model-Based Pose Estimation*, chapter 9, pages 139–170. Springer, 2011. 4

[66] Helge Rhodin, Jörg Spörri, Isinsu Katircioglu, Victor Constantin, Frédéric Meyer, Erich Müller, Mathieu Salzmann, and Pascal Fua. Learning monocular 3D human pose estimation from multi-view images. In *CVPR*, 2018. 4

[67] Kathleen M. Robinette, Sherri Blackwell, Hein Daanen, Mark Boehmer, Scott Fleming, Tina Brill, David Hoeferlin, and Dennis Burnsides. Civilian American and European Surface Anthropometry Resource (CAESAR) final report. Technical Report AFRL-HE-WP-TR-2002-0169, US Air Force Research Laboratory, 2002. 3, 4

[68] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics (TOG)*, 2017. 2, 3, 4, 5, 6

[69] Tanner Schmidt, Richard Newcombe, and Dieter Fox. DART: Dense articulated real-time tracking. In *RSS*, 2014. 2, 3

[70] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *CVPR*, 2017. 1, 2, 5

[71] Srinath Sridhar, Antti Oulasvirta, and Christian Theobalt. Interactive markerless articulated hand motion tracking using RGB and depth data. In *ICCV*, 2013. 2, 3

[72] Jonathan Starck and Adrian Hilton. Surface capture for performance-based animation. *IEEE computer graphics and applications*, 27(3), 2007. 4

[73] Matthias Teschner, Stefan Kimmerle, Bruno Heidelberger, Gabriel Zachmann, Laks Raghupathi, Arnulph Fuhrmann, Marie-Paule Cani, François Faure, Nadia Magnenat-Thalmann, Wolfgang Strasser, and Pascal Volino. Collision detection for deformable objects. In *Eurographics*, 2004. 5

[74] Anastasia Tkach, Mark Pauly, and Andrea Tagliasacchi. Sphere-meshes for real-time hand modeling and tracking. *ACM Transactions on Graphics (TOG)*, 35(6), 2016. 2, 3

[75] Dimitrios Tzionas, Luca Ballan, Abhilash Srikantha, Pablo Aponte, Marc Pollefeys, and Juergen Gall. Capturing hands in action using discriminative salient points and physics simulation. *IJCV*, 118(2):172–193, 2016. 2, 3, 5, 6

[76] Daniel Vlasic, Matthew Brand, Hanspeter Pfister, and Jovan Popović. Face transfer with multilinear models. *ACM transactions on graphics (TOG)*, 24(3):426–433, 2005. 2

[77] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *CVPR*, 2016. 2, 5

[78] Weipeng Xu, Avishek Chatterjee, Michael Zollhöfer, Helge Rhodin, Dushyant Mehta, Hans-Peter Seidel, and Christian Theobalt. Monoperfcap: Human performance capture from monocular video. *ACM Transactions on Graphics (TOG)*, 37(2):27, 2018. 3

[79] Fei Yang, Jue Wang, Eli Shechtman, Lubomir Bourdev, and Dimitri Metaxas. Expression flow for 3D-aware face component transfer. *ACM Transactions on Graphics (TOG)*, 30(4):60, 2011. 2

[80] Shanxin Yuan, Guillermo Garcia-Hernando, Björn Stenger, Gyeongsik Moon, Ju Yong Chang, Kyoung Mu Lee, Pavlo Molchanov, Jan Kautz, Sina Honari, Liuhao Ge, et al. Depth-based 3D hand pose estimation: From current achievements to future goals. In *CVPR*, 2018. 3

[81] Michael Zollhöfer, Justus Thies, Pablo Garrido, Derek Bradley, Thabo Beeler, Patrick Pérez, Marc Stamminger, Matthias Nießner, and Christian Theobalt. State of the art on monocular 3D face reconstruction, tracking, and applications. *Computer Graphics Forum*, 37(2):523–550, 2018. 3