Joint Optimization for Multi-Person Shape Models from Markerless 3D-Scans

Samuel Zeitvogel[1], Johannes Dornheim[1], and Astrid Laubenheimer[1]

Intelligent Systems Research Group (ISRG), Karlsruhe University of Applied Sciences, Germany
{samuel.zeitvogel,johannes.dornheim,astrid.laubenheimer}@hs-karlsruhe.de

Abstract. We propose a markerless end-to-end training framework for parametric 3D human shape models. The training of statistical 3D human shape models with minimal supervision is an important problem in computer vision. Contrary to prior work, the whole training process (i) uses a differentiable shape model surface and (ii) is trained end-to-end by jointly optimizing all parameters of a single, self-contained objective that can be solved with slightly modified off-the-shelf non-linear least squares solvers. The training process only requires a compact model definition and an off-the-shelf 2D RGB pose estimator. No pre-trained shape models are required. For training (iii) a medium-sized dataset of approximately 1000 low-resolution human body scans is sufficient to achieve competitive performance on the challenging FAUST surface correspondence benchmark. The training and evaluation code will be made available for research purposes to facilitate end-to-end shape model training on novel datasets with minimal setup cost.

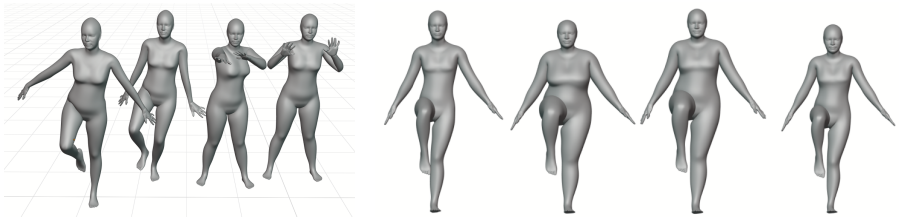
Keywords: Body shape, skinning, subdivision surfaces, blendshapes

1 Introduction

Statistical human shape models are a prerequisite for a wide variety of tasks such as shape completion, 3D virtual avatar generation, *e.g.* for virtual try-on, gaming, and markerless motion capture.

Conventional approaches [3,18,26] for human shape model training employ a two-stage process consisting of a template-to-scan registration step followed by a model parameter estimation step. In the first step, high-quality 3D scans of humans are registered to a common template mesh using additional supervision, *e.g.* hand-picked landmarks [3,18]. Once the scans are brought into correspondence, the registered meshes (registrations) are manually reviewed for errors. Correct registrations are then used for the training process to produce multi-person articulated human shape models of high quality.

Shape model training is a chicken-and-egg problem. High-quality registrations are best acquired with a good model while the training of a good model requires high-quality registrations. Alternatively, a bootstrapping approach [18] employs a weak model to regularize the registration process and train a better



(a) Pose space (b) Shape space

Fig. 1: Samples from our trained articulated morphable human shape model. A template with $N = 1326$ vertices produces realistic avatars. Shapes are acquired by changing pose-specific parameters (a) and shape-specific parameters (b).

model. This process can be repeated multiple times with alternating optimizers.

In contrast, recent advances in 3D hand model inference on noisy depth images indicate that joint, continuous optimization of data correspondences and model parameters is less likely to converge to bad local minima [33]. Similar work on joint shape model training for hands [21] and humans [37] exists. A differentiable surface model (*e.g.* subdivision surfaces) enables joint, continuous optimization. In this work we follow the best practices for differentiable shape model formulation [21, 27, 37], objective formulation and joint optimization [11, 12, 21, 36]. The training and evaluation code[1] will be made available for research purposes to facilitate end-to-end shape model training on novel datasets with minimal setup cost. Our contribution is threefold:

First, we propose a differentiable multi-person articulated human shape model (inspired by [21, 26, 37]) that can be trained using joint optimization without any 3D supervision. Differentiability is achieved using a Catmull-Clark subdivision surface module [9] with additional benefits: Only a low-poly base mesh is required to generate realistic 3D avatars (see Fig. 1) and the model parameter count is reduced considerably. Similar to the model proposed in [26], the resulting model is compatible with 3D-modelling software and can be computed efficiently.

Secondly, we formulate a single objective for model training that can be minimized with off-the-shelf nonlinear least squares solvers with minor modifications. We employ common best practices to deal with non-euclidean manifolds, robust cost functions, and discrete data-to-model correspondence updates.

Our third contribution is the application of the aforementioned differentiable multi-person shape model and the proposed optimization procedure to roughly 1000 markerless low-resolution point clouds. The advent of least squares solvers on the GPU [11] enables large scale joint optimization for multi-person shape model training which was previously only considered using alternating optimization

[1] `https://github.com/Intelligent-Systems-Research-Group/JOMS/`

methods. We evaluate the reconstruction quality of our approach and benchmark the competitive generalization quality on a challenging shape correspondence benchmark.

2 Related Work

2.1 Human Shape Models

Early work by Blanz and Vetter [4] introduces a morphable shape model for faces using a triangulated mesh structure. Registered meshes in the training set are distilled to a morphable model using principal components analysis (PCA).

Anguelov et al. [3] propose the popular morphable human shape model SCAPE. The model factorizes in a subject-specific shape model and a pose specific shape model. SCAPE does not model the vertex displacement directly but instead relies on triangle transformations. The triangle soup is realigned with an additional least squares estimation step. This approach is also not directly compatible with current software packages, 3D modelling software, and game engines. The required registered meshes are generated in a semi-supervised preprocessing step.

Hirshberg et al. [18] modify SCAPE and incorporate the registration process into the model using alternating optimization. This approach is enhanced by Bogo et al. [5] to incorporate texture information and deal with temporal information that arises by shapes in motion [6]. The Stitched Puppet method [38] transforms SCAPE into a probabilistic graphical model and fits the model to data using a particle-based optimization method.

Loper et al. [26] introduce the SMPL model. This approach produces shape models that are compatible with 3D modelling software. The quality of the model relies heavily on high-quality registrations. Our model formulation extends SMPL and we review a modified version in Section 3. SMPL is extended to faces [24], hands [30] and modelling infants [17]. Multiple models can be combined to construct a fully articulated morphable 3D human shape model [20, 27].

In contrast to bootstrapping approaches that require registrations, deep learning approaches find shape correspondences without an explicit model [16]. However, many deep learning approaches still rely on templates for training data generation. [10, 14, 15, 23] are trained on large scale synthetic datasets like SURREAL [34]. SURREAL is generated with the help of trained shape models.

2.2 Subdivision Surfaces

Training or fitting mesh models is usually done with variants of nonrigid iterative closest point (NICP) where the data to model correspondence finding and model training is performed in an alternating fashion. To use joint optimization of correspondences and model parameters a differentiable surface representation is required. One way to transform a polygonal mesh into a differentiable surface is the application of subdivision surfaces [9]. Transforming a polygonal mesh into a subdivision surface is easy and can be implemented as a post-processing step.

Cashman and Fitzgibbon [8] train animal shape models on segmented images using a template mesh with subdivision surfaces. Taylor et al. [32] learn a subdivision surface hand model from depth images and cast the whole training process as a single joint optimization problem. They incorporate subdivision surfaces to allow for continuous sliding of the corresponding surface points. Khamis et al. [21] build on this framework to train a morphable articulated hand model from depth images. Taylor et al. [33] use the subdivision surface hand model from [21] to build a hand tracker within a joint continuous optimization framework. Catmull Clark subdivision surfaces [9] have also been considered for human surface models in the context of surface reconstruction [19] and motion capture [37]. The major difference between [37] and our work is that in our work we train a *multi-person* shape model with the addition of shape blend-shapes [26]. The approach outlined in [37] requires existing sparse surface correspondences while our method describes a fully automatic model training pipeline. In contrast to [37], we also provide a quantitative evaluation to show the efficacy of our approach.

2.3 Joint Optimization

In general, alternating optimization is employed due to implementation simplicity and scalability. It is also used in variants of NICP which is in turn required when the surface is not differentiable. Current shape models are trained using forms of bootstrapping with human supervision in the loop.

Taylor et al. [32] compare alternating with joint optimization and report an increased convergence rate and a decreased reconstruction error using joint optimization. For optimization, they linearize the surface at the corresponding point with the tangent plane. The update in the tangent plane is applied to the underlying surface by traversing the mesh and transforming the update direction and magnitude between surface patches accordingly.

Robustified cost functions are prevalent in shape model training to deal with noisy 3D data. Zach [35] proposes a joint optimization scheme for robust bundle adjustment using lifting methods. Lifting methods introduce additional variables and circumvent alternating optimization. Zach and Bourmaud [36] deliver further insight into lifting and gradual refinement for bundle adjustment and recommend using the lifting method when a fast decrease in the objective is preferred and a sensible initial estimation of parameters is known.

Large scale optimization with nonlinear least squares objectives has been made more accessible through open-source optimizers such as Ceres [1] and Optlang [11]. DeVito et al. [11] introduce a GPU solver and show that using generic Gauss-Newton and Levenberg-Marquardt with conjugate gradient as the inner solver is competitive to handcrafted problem-specific solvers for many optimization problems that arise in computer graphics and computer vision.

3 Articulated Morphable Shape Model

The employed statistical shape model is a variation of SMPL [26] that enables joint optimization of all parameters. A differentiable surface model derived from

the SMPL pose deformation model is introduced by [37]. They use an articulated person model for joint optimization. The model we propose incorporates the multi-person aspects of [26] and the subdivision surfaces for smooth parametric shape modelling from [21] into the articulated human shape model from [37].

In order to make the following formalism easier to read, the supplementary material contains tables summarizing symbols.

Formalism: The model template consists of a base mesh with $N = 1326$ vertices and $F = 1324$ quadrilateral faces. The underlying skeleton consists of $K = 16$ joints. A root joint is added and the resulting $K + 1$ nodes are connected by K edges (often referred to as bones in the literature). A visualization of the body parts can be found in the supplementary material. Sample meshes can be instantiated from the statistical model using subject-specific parameters $\vec{\beta} \in \mathbb{R}^B$ with $B = 10$ and pose specific parameters $(R, \boldsymbol{m}) \in SO(3)^{K+1} \times \mathbb{R}^3$. $(R, \boldsymbol{m}) = (R_0, R_1, \ldots, R_K, \boldsymbol{m})$ is separated in global pose parameters R_0 (global rotation), $\boldsymbol{m}$ (global translation) and skeleton pose parameters $R_1, R_2, \cdots, R_K$. The mean shape is denoted by $\bar{\boldsymbol{T}} \in \mathbb{R}^{3N}$ (vectors in $\mathbb{R}^{3N}$ are interpreted as stacked x, y, z coordinates of N vectors in $\mathbb{R}^3$). Shape blend-shapes are denoted by $\mathcal{S} \in \mathbb{R}^{3N \times B}$ and corrective pose blend-shapes by $\mathcal{P} \in \mathbb{R}^{3N \times 9K}$. Each column in $\mathcal{S}$ denotes a shape blend-shape and the columns of $\mathcal{P}$ denote the $9K$ corrective pose blend-shapes. Shape blend-shapes are introduced to model varying shape between different subjects. In contrast to [26], the shape blend-shapes are not enforced to be orthogonal. Corrective pose blend-shape are incorporated to counteract artefacts (*e.g.* surface shrinking near joints during mesh articulation) when applying linear blend skinning Eq. (5). In contrast to [26], our skeleton always binds to the mean shape $\bar{\boldsymbol{T}}$ with the zero pose $(R^*, \boldsymbol{m}^*) = (I, I, \ldots, I, \boldsymbol{0}) \in SO(3)^{K+1} \times \mathbb{R}^3$.

Blend-Shape Application: Blend-shapes are applied to the mean shape $\bar{\boldsymbol{T}}$ using

$$\boldsymbol{T} = \bar{\boldsymbol{T}} + \mathcal{S}\vec{\beta} + \mathcal{P}\,\mathrm{vec}(R_1 - R_1^*, R_2 - R_2^*, \ldots, R_K - R_K^*), \tag{1}$$

where $I \in \mathbb{R}^{3\times3}$ is the identity matrix and $\mathrm{vec}(\cdot)$ flattens each argument and concatenates the vectors to form a single column vector of the required shape. The resulting vertex positions after the application of all linear blend-shapes are denoted by $\boldsymbol{T}$. We prescribe a fixed sparsity pattern to the corrective pose blend-shape matrix $\mathcal{P}$ which reduces the parameter count and prevents overfitting. In our implementation, we assign each vertex to two adjacent joints which leads to at most $2 \cdot 9 = 18$ nonzero entries per row in $\mathcal{P}$.

Subject-specific Skeleton: To articulate the shape we use a skeleton that defines a forward kinematic tree. The subject-specific joint locations in the rest pose are given by

$$\boldsymbol{J} = \bar{\boldsymbol{J}} + \mathcal{J}\vec{\beta}, \tag{2}$$

where $\bar{\boldsymbol{J}} \in \mathbb{R}^{3K}$ are the joint locations corresponding to the mean shape $\bar{\boldsymbol{T}}$ and $\mathcal{J} \in \mathbb{R}^{3K\times B}$ are the skeleton basis shapes that are correlated with the respective shape blend-shapes $\mathcal{S}$. We follow [37] for the subject-specific skeleton

formalism instead of [26] which instead regresses from $\bar{\boldsymbol{T}}+\mathcal{S}\vec{\beta}$ to $\boldsymbol{J}$ by employing a sparse regression matrix with $3N \cdot 3K$ parameters. Our end-to-end optimization framework prohibits the use of sparsity inducing regularizers (as used by [26]). Additionally, only $(B+1)3K$ instead of $9NK$ parameters are introduced.

Kinematic Tree: We use the resulting joint locations and a prescribed skeleton topology to construct a transformation for each joint $1 \leq k \leq K$ in the skeleton. This transformation can then be applied to any $\boldsymbol{x} \in \mathbb{R}^3$. The transformation of $\boldsymbol{x}$ w.r.t. the joint indexed by k is denoted by $G'_k : \mathbb{R}^3 \times SO(3)^{K+1} \times \mathbb{R}^{3K} \to \mathbb{R}^3$ with

$$G'_k(\boldsymbol{x}; R, \boldsymbol{J}) = \begin{pmatrix} 1&0&0&0 \\ 0&1&0&0 \\ 0&0&1&0 \end{pmatrix} G_k \begin{pmatrix} \boldsymbol{x} - \boldsymbol{j}_k \\ 1 \end{pmatrix} \text{where} \tag{3}$$

$$G_0(R, \boldsymbol{J}) = \begin{pmatrix} R_0 & \mathbf{0} \\ \mathbf{0}^T & 1 \end{pmatrix} \text{ and } G_k(R, \boldsymbol{J}) = \begin{pmatrix} R_k & \boldsymbol{j}_k - \boldsymbol{j}_{A(k)} \\ \mathbf{0}^T & 1 \end{pmatrix} G_{A(k)} \quad \forall_{1 \leq k \leq K}. \tag{4}$$

The position of the joint indexed by k is $\boldsymbol{j}_k \in \mathbb{R}^3$, $A(k)$ denotes the ancestor joint index with respect to joint indexed by k and $\boldsymbol{j}_0 = \mathbf{0}$. Note that the $\boldsymbol{x} - \boldsymbol{j}_k$ transforms x from the joint indexed by k to the root node.

Linear Blend Skinning: The kinematic tree defined above is used to transform the vertex positions $\boldsymbol{T}$ to $\boldsymbol{T}'$ using linear blend skinning [22]:

$$\boldsymbol{t}'_i = \boldsymbol{m} + \sum_{k=1}^{K} w_{k,i} G'_k(\boldsymbol{t}_i; R, \boldsymbol{J}), \quad \boldsymbol{t}_i, \boldsymbol{t}'_i \in \mathbb{R}^3, \forall_{1 \leq i \leq N} \tag{5}$$

where $\boldsymbol{t}_i, \boldsymbol{t}'_i \in \mathbb{R}^3$ denote the i-th vertex of $\boldsymbol{T}$ and $\boldsymbol{T}'$ respectively and $w_{k,i} \in [0,1]$ is a linear blend skinning weight for vertex $\boldsymbol{t}_i$ transformed by G'_k. All blend skinning weights are denoted by $\mathcal{W} \in [0,1]^{K \times N}$. The blend skinning weights are constrained by

$$\sum_{k=1}^{K} w_{k,i} = 1 \tag{6}$$

for all vertices indexed by i. After linear blend skinning and global translation $\boldsymbol{m}$, the resulting vertex positions are denoted by $\boldsymbol{T}' \in \mathbb{R}^{3N}$.

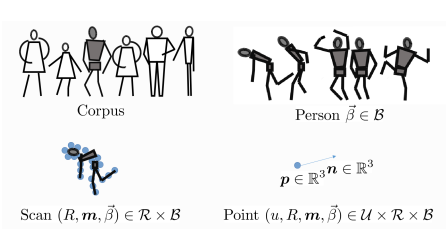
Complete Model Formulation: The whole mesh transformation is denoted by

$$\boldsymbol{T}' = M(R, \boldsymbol{m}, \vec{\beta}; \Theta), \tag{7}$$

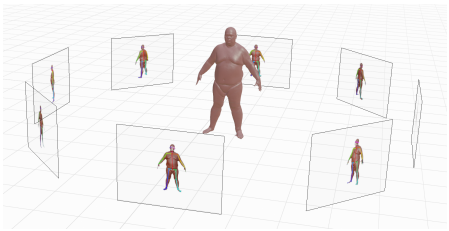
where $\Theta = (\bar{\boldsymbol{T}}, \mathcal{S}, \bar{\boldsymbol{J}}, \mathcal{J}, \mathcal{P}, \mathcal{W})$ denote the aggregated model parameters. We convert the quad mesh into a subdivision surface denoted by

$$S(u; \boldsymbol{T}') : \Omega \times \mathbb{R}^{3N} \to \mathbb{R}^3, \tag{8}$$

where $u = (s, t, f)$ denotes a point on the surface S in local 2D coordinates $(s,t) \in [0,1]^2$ with patch index $f \in \{1, 2, \cdots, F\}$. More precisely u consists of



(a) Hierarchical data setup



(b) OpenPose on synthetic images

Fig. 2: (a) Visualization of the hierarchical input data and latent variables. Top left: Depiction of different persons in the training corpus. Top right: A single person with shape parameters $\vec{\beta}$ shown in different body poses. Bottom left: A fitted model with subject-specific shape coefficients $\vec{\beta}$ and pose-specific parameters $(R, \boldsymbol{m})$. The 3D point cloud is indicated with filled blue circles. Bottom right: Zoom in on a single point in the point cloud with 3D position $\boldsymbol{p}$ and normal $\boldsymbol{n}$. The corresponding model surface point is parameterized by $(u, R, \boldsymbol{m}, \vec{\beta})$. Each datapoint induces a residual for the data term E_{Data}. (b) Keypoints: This figure shows an input scan from the Dynamic FAUST dataset [6] with 8 virtual cameras surrounding the scan in the center. Keypoints are extracted and visualized from each synthetic image using OpenPose [7].

a 2D bezier patch parameterization [13] and the respective patch index, so $u \in \Omega = [0, 1]^2 \times \{1, 2, \cdots, F\}$. We implement S approximately using the approach outlined in [25] due to its efficiency and simplicity. Implementation details are provided in the supplementary material.

4 Objective and Optimization

To avoid a cluttered multi-index formalism we introduce a number of sets to describe the hierarchical nature of the data and accompanying latent variables (see Fig 2a).

The training of the human shape requires the estimation $\hat{\Theta}$ of model parameters Θ from given point cloud measurements (scans) of different instances. Here, an instance denotes a 3D-scan of a specific subject in a specific pose. For n scans of Q individual subjects and P measurements per scan, we subsume all latent parameters in the triplet $\Gamma = (\mathcal{B}, \mathcal{R}, \mathcal{U})$, where $\mathcal{B}$ is the set of Q latent shape vectors $\vec{\beta} \in \mathcal{B}$, $\mathcal{R}$ is the set of n latent pose configurations $(R, \boldsymbol{m}) \in \mathcal{R}$ and $\mathcal{U}$ is the set of nP latent surface correspondences $u \in \mathcal{U}$. In total, this leads to an unknown variable count of

$$|\mathcal{B}| + |\mathcal{R}| + |\mathcal{U}| = QB + 3(K + 2)n + 2Pn. \tag{9}$$

The estimation of the model parameters Θ requires the estimation of the latent parameters denoted by $\hat{\Gamma}$. We cast the model learning process as a non-

linear least squares problem with unknowns Θ and Γ and propose the following cost function:

Data Term: The main functional is defined as

$$E_{\text{Data}} = \sum_{(\boldsymbol{p},u,R,\boldsymbol{m},\vec{\beta})\in\mathcal{I}_{\text{Data}}} \phi\left(||S(u,M(R,\boldsymbol{m},\vec{\beta};\Theta)) - \boldsymbol{p}||^2\right), \tag{10}$$

where $\boldsymbol{p} \in \mathbb{R}^3$ denotes a point on the 3D scan, $\mathcal{I}_{\text{Data}} \subset \mathbb{R}^3 \times \mathcal{U} \times \mathcal{R} \times \mathcal{B}$ contains all data points with latent variable dependencies and $\phi : \mathbb{R} \to \mathbb{R}$ with $\phi(r) = r^2/(r^2+\rho^2)$ is the robust Geman-McClure kernel.

Cost Function: Since the data term itself is not sufficient to lead to satisfactory results, additional regularization terms have to be added to constrain the solution space to minimizers which are restricted to non-degenerate shapes. To this end, we introduce additional prior information and regularization terms. Our complete cost function is

$$\begin{aligned} E = &\lambda_{\text{Data}}E_{\text{Data}} + \lambda_{\text{2D-joint}}E_{\text{2D-joint}} + \lambda_{\text{2D-surf}}E_{\text{2D-surf}} + \lambda_{\text{mean}}E_{\text{mean}} \\ &+\lambda_{\text{bshape}}E_{\text{bshape}} + \lambda_{\text{pshape}}E_{\text{pshape}} + \lambda_{\text{symm}}E_{\text{symm}} + \lambda_{\text{symm-skel}}E_{\text{symm-skel}} \\ &+\lambda_{\text{joint}}E_{\text{joint}} + \lambda_{\text{weights}}E_{\text{weights}} + \lambda_{\text{convex}}E_{\text{convex}} + \lambda_{\text{shape}}E_{\text{shape}} \\ &+\lambda_{\text{pose}}E_{\text{pose}} + \lambda_{\text{ground}}E_{\text{ground}} \end{aligned} \tag{11}$$

and consists of a weighted sum of squared error terms $E_\bullet$ and non-negative scalar hyperparameters $\lambda_\bullet$ that control the impact of each term. The values for the weights $\lambda_\bullet$ are listed in the supplementary material. The remaining section describes the regularization terms in detail.

2D Joint Term: When the model pose and shape is initialized far away from the pose and shape of the scan, the optimization of the data term is likely to end in a local minimum. In order to steer the pose and shape estimation in the right direction, we define a landmark term based on synthetic views of the scans. We apply OpenPose [7] on these scans for 2D keypoint extraction (see Fig. 2b). Those keypoints, which correspond to joints (and have a detection score above 0.5) are used as landmarks. We denote Π as the set of $|\Pi| = 8$ virtual cameras and add a landmark term that penalizes the distance between the model joints projected onto the virtual 2D images and the 2D landmarks:

$$E_{\text{2D-joint}} = \sum_{(\boldsymbol{Q},\pi,R,\boldsymbol{m},\vec{\beta})\in\mathcal{I}_{\text{2D-joint}}} \sum_{k=1}^{K} ||\pi(G'_k(\boldsymbol{j}_k,R,\boldsymbol{J}) + \boldsymbol{m}) - \boldsymbol{q}_k||^2 \tag{12}$$

where $\pi : \mathbb{R}^3 \to \mathbb{R}^2$ is a camera-specific projection that maps points in world coordinates to points in image coordinates. $\mathcal{I}_{\text{2D-joint}} \subset \mathbb{R}^{2K} \times \Pi \times \mathcal{R} \times \mathcal{B}$ contains the 2D-labels $\boldsymbol{Q} \in \mathbb{R}^{2K}$ with $\boldsymbol{q}_k \in \mathbb{R}^2$ for each joint k in an image with latent dependencies. Even though OpenPose was trained on natural labeled RGB images [7], the predictions work surprisingly well for non-photorealistic synthetic images. The camera setup and the synthetic images with keypoint estimations are depicted in Fig. 2b.

2D Surface Term: OpenPose not only provides keypoints that correspond to joints but also provides keypoints corresponding to landmarks on the surface of the scan (*e.g.* nose and ears). We propose an additional error term for such keypoints. The 2D surface landmark term

$$E_{\text{2D-surf}} = \sum_{(\boldsymbol{q}',u',\pi,R,\boldsymbol{m},\vec{\beta})\in\mathcal{I}_{\text{2D-surf}}} ||\pi\left(S(u', M(R,\boldsymbol{m},\vec{\beta};\Theta))\right) - \boldsymbol{q}'||^2, \tag{13}$$

encourages the optimization process to bring the model surface points at $u' \in \Omega$ close to the 2D annotations $\boldsymbol{q}' \in \mathbb{R}^2$ in the image space after the perspective projection π. $\mathcal{I}_{\text{2D-surf}} \subset \mathbb{R}^2 \times \Omega \times \Pi \times \mathcal{R} \times \mathcal{B}$ contains all 2D surface landmarks and links the relevant surface location, camera projection function and latent dependencies.

Smoothing Terms: Self-intersections for mesh-based model-fitting approaches have to be mitigated during optimization. Additionally, we have to deal with missing data (*e.g.* armpits and soles of the feet) and an interpolation scheme has to be adopted in such body regions. To this end, we correlate $\bar{\boldsymbol{T}}$, $\mathcal{S}$ and $\mathcal{P}$ so that vertex displacement corresponding to adjacent vertices in the graph of the mesh are similar. When interpreting these variables as vector fields on the mesh we prefer solutions where the vector fields are smooth. One way to encourage this behavior is by exploiting the linear, positive semidefinite Laplace-Beltrami operator from the template shape with vertex positions $\bar{\boldsymbol{T}}^{\text{init}} \in \mathbb{R}^{3N}$. We represent the linear discrete Laplace-Beltrami operator by the matrix $\Delta \in \mathbb{R}^{N\times N}$ and denote $||\boldsymbol{C}||_\Delta^2 = \boldsymbol{C}^T\Delta\boldsymbol{C}$ by slight abuse of notation for a provided vector $\boldsymbol{C} \in \mathbb{R}^N$. We estimate Δ with vertex positions $\bar{\boldsymbol{T}}^{\text{init}}$ using the approach from [2]. Our model parameters live mostly in $\mathbb{R}^{3N}$. We make use of Δ by applying it to the N different x, y and z components in the spirit of [31]. To this end, we employ the regularization terms

$$E_{\text{mean}} = ||\bar{\boldsymbol{T}}_x - \bar{\boldsymbol{T}}_x^{\text{init}}||_\Delta^2 + ||\bar{\boldsymbol{T}}_y - \bar{\boldsymbol{T}}_y^{\text{init}}||_\Delta^2 + ||\bar{\boldsymbol{T}}_z - \bar{\boldsymbol{T}}_z^{\text{init}}||_\Delta^2 \tag{14}$$

$$E_{\text{bshape}} = \sum_{j=1}^{B}\left(||\mathcal{S}_x^{(j)}||_\Delta^2 + ||\mathcal{S}_y^{(j)}||_\Delta^2 + ||\mathcal{S}_z^{(j)}||_\Delta^2\right) \tag{15}$$

$$E_{\text{pshape}} = \sum_{j=1}^{9K}\left(||\mathcal{P}_x^{(j)}||_\Delta^2 + ||\mathcal{P}_y^{(j)}||_\Delta^2 + ||\mathcal{P}_z^{(j)}||_\Delta^2\right) \tag{16}$$

where $\cdot_x$, $\cdot_y$, $\cdot_z$ refer to the x,y or z component from a vector in $\mathbb{R}^{3N}$ and $\cdot^{(j)}$ denotes the j-th column of the indexed matrix.

Symmetry Term: The body model is split into the left- and right-hand side along the y-z plane. We employ a symmetry term that encourages symmetric shapes with

$$\begin{aligned} E_{\text{symm}} = {} & ||\bar{\boldsymbol{T}}_x + \bar{\boldsymbol{T}}_x^{\text{mirror}}||^2 + ||\bar{\boldsymbol{T}}_y - \bar{\boldsymbol{T}}_y^{\text{mirror}}||^2 + ||\bar{\boldsymbol{T}}_z - \bar{\boldsymbol{T}}_z^{\text{mirror}}||^2 + \\ & \sum_{j=1}^{B}\left(||\mathcal{S}_x^{(j)} + \tilde{\mathcal{S}}_x^{(j)}||^2 + ||\mathcal{S}_y^{(j)} - \tilde{\mathcal{S}}_y^{(j)}||^2 + ||\mathcal{S}_z^{(j)} - \tilde{\mathcal{S}}_z^{(j)}||^2\right), \end{aligned} \tag{17}$$

where $\bar{\boldsymbol{T}}^{\text{mirror}}$ and $\tilde{\mathcal{S}}$ permute $\bar{\boldsymbol{T}}$ and $\mathcal{S}$, respectively, so that each vertex maps to its mirrored partner. A corresponding term exists for $\bar{\boldsymbol{J}}$ and $\mathcal{J}$ with

$$E_{\text{symm-skel}} = ||\bar{\boldsymbol{J}}_x + \bar{\boldsymbol{J}}_x^{\text{mirror}}||^2 + ||\bar{\boldsymbol{J}}_y - \bar{\boldsymbol{J}}_y^{\text{mirror}}||^2 + ||\bar{\boldsymbol{J}}_z - \bar{\boldsymbol{J}}_z^{\text{mirror}}||^2 + \sum_{j=1}^{B} \left(||\mathcal{J}_x^{(j)} + \tilde{\mathcal{J}}_x^{(j)}||^2 + ||\mathcal{J}_y^{(j)} - \tilde{\mathcal{J}}_y^{(j)}||^2 + ||\mathcal{J}_z^{(j)} - \tilde{\mathcal{J}}_z^{(j)}||^2 \right), \tag{18}$$

where $\bar{\boldsymbol{J}}^{\text{mirror}}$ and $\tilde{\mathcal{J}}$ are defined accordingly.

Skeleton Consistency Term: Our current formulation does not prevent the optimizer to move the joints to arbitrary locations outside of the surface. To counteract this behavior, we softly constrain the mean joint positions $\boldsymbol{J}$ and skeleton basis shapes $\mathcal{J}$ to deform with the mean shape $\bar{\boldsymbol{T}}$ and blend-shapes $\mathcal{S}$ at some specified vertices. We denote the set of manually specified vertex indices for the joint indexed by k with Ring_k (see *e.g.* [21] or the supplementary material). The joint error term is defined as

$$E_{\text{joint}} = \sum_{k=1}^{K} \left(||\boldsymbol{j}_k - \frac{1}{|\text{Ring}_k|} \sum_{i \in \text{Ring}_k} \bar{\boldsymbol{t}}_i||^2 + \sum_{b=1}^{B} ||\mathcal{J}_k^{(b)} - \frac{1}{|\text{Ring}_k|} \sum_{i \in \text{Ring}_k} \mathcal{S}_i^{(b)}||^2 \right), \tag{19}$$

where $\mathcal{S}_i^{(b)} \in \mathbb{R}^3$ denotes the i-th vertex position of the b-th blend-shape and $\mathcal{J}_k^{(b)} \in \mathbb{R}^3$ denotes the k-th joint of the b-th skeleton basis shape.

Blend-Skinning Term: We encourage blend-weights $\mathcal{W}$ that are close to an initial estimate $\mathcal{W}^{\text{init}}$ (see the supplementary materials for details) with

$$E_{\text{weights}} = ||\mathcal{W} - \mathcal{W}^{\text{init}}||_F^2. \tag{20}$$

Convex Combination Term: We introduce another term that softly encourages the convex combination constraint Eq. (6) for the blend-skinning weights $\mathcal{W}$ via

$$E_{\text{convex}} = \sum_{i=1}^{N} \left(1 - \sum_{k=1}^{K} w_{k,i} \right)^2. \tag{21}$$

Shape Regularization Term: We encourage the blend-shape coefficients $\vec{\beta}$ to be small using

$$E_{\text{shape}} = \sum_{\vec{\beta} \in \mathcal{B}} ||\vec{\beta}||^2. \tag{22}$$

Pose Regularization Term: We want to discourage unrealistic human postures with the addition of simple pose regularization term

$$E_{\text{pose}} = \sum_{(R, \boldsymbol{m}) \in \mathcal{R}} \sum_{k=1}^{K} ||R_k - \bar{R}_k||_F^2 \tag{23}$$

where $\bar{R}_k \in SO(3)$ denotes the mean rotation of joint k. These are also introduced as variables during optimization and denoted by $\bar{R}$.

Ground Plane Consistency Term: To counteract missing 3D measurements on the soles of the feet we employ a ground plane term to compensate for missing data. We assume a known ground plane perpendicular to the y-axis and offset at height $H \in \mathbb{R}$. We penalize uniformly sampled model surface points $u'' \in \Omega$ that fall below the ground plane via

$$E_{\text{ground}} = \sum_{(u'',R,\boldsymbol{m},\vec{\beta})\in\mathcal{I}_{\text{ground}}} \Big(S(u'', M(R,\boldsymbol{m},\vec{\beta};\Theta))_y - H\Big)^2, \tag{24}$$

where $\mathcal{I}_{\text{ground}} \subset \Omega \times \mathcal{R} \times \mathcal{B}$ contains sampled surface points below the ground plane with latent dependencies for each scan and $\cdot_y$ denotes the y-coordinate of a point in $\mathbb{R}^3$.

Optimization: The objective defined in Eq. (11) is cast as a nonlinear least squares problem and minimized using a truncated variation of Levenberg-Marquardt, where the normal equations are solved approximately using the conjugate gradient method. This leads to the local minimizers

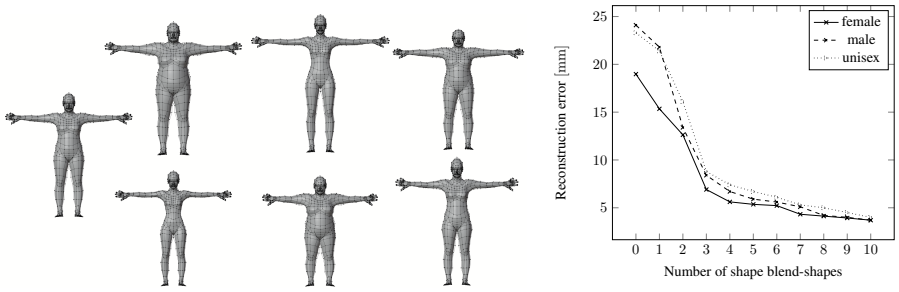
$$\hat{\Theta}, \hat{R} = \arg\min_{\Theta,\bar{R}} \min_{\Gamma} E(\Gamma;\Theta,\bar{R}). \tag{25}$$

Optimization is performed using the Optlang [11] framework. Several aspects of the optimization procedure require special attention (*e.g.* discrete correspondences updates) and are discussed in the supplementary material. We use a varying hyperparamter schedule for $\lambda_\bullet$, encouraging a rough alignment of body pose before relaxing the constraints on $\mathcal{S}, \mathcal{W}$ and $\mathcal{P}$. At the start of the optimization we mainly rely on detected keypoints. In the later stages, the keypoints are discarded and E_{data} takes precedence.

5 Experimental Evaluation

5.1 Training

We train a female, a male and a unisex model. For training, we test the limits of our approach and use the entire GPU memory (11 GB). We train the models on $n = 911$ scans each where the first 250 scans are acquired from the D-FAUST dataset [6]. D-FAUST contains 10 subjects (5 female and 5 male) and spans a wide variety of body poses. We sample 50 scans per person from this dataset by clustering the pose space with k-means on multi-view OpenPose 3D joint predictions to cover a diverse set of poses. The remaining 661 scans are randomly chosen from the CAESAR dataset [29]. We subsample all 911 scans to 20,000 data points each. This results in a total of $Q = 666$ distinct individuals for each model in the training process. One training process simultaneously optimizes more than $3.6 \cdot 10^7$ latent parameters plus $|\Theta| + 3K$ model parameters (the count of latent parameters Eq. (9) increases with the training set size). This setup requires up to three days of training on an NVIDIA RTX 2080 TI.



(a) Principal components (b) Reconstruction error

Fig. 3: (a) Visualization of our trained female shape model, from left to right: Mean shape and the first three PCA components sorted by explained variance in descending order ($\pm 3\sigma$ for the first two principal components and $\pm 5\sigma$ for the third principal component). The model without applied subdivision is overlayed as a wireframe mesh. (b) After training with $B = 10$ shape blend-shapes, we compute the shape and skeleton for each person in the training set and create new blend-shapes using PCA. We vary the number of principal components from 0 to 10 and record the average reconstruction error for each setting.

Fig. 4: Qualitative model fit results on the FAUST dataset: This figure shows 15 different poses in total of all 10 different subjects. The scans are shown in red and the model fits are shown in light blue.

Model Subspace Evaluation: We analyze the factorization of shape and pose parameters of our models qualitatively. Therefore, we keep the pose parameters $(R, \boldsymbol{m})$ fixed to a non-canonical pose and vary $\vec{\beta}$ to see if a change in $\vec{\beta}$ leads to obvious changes in posture. Results of this experiment such as shown in Fig. 1b indicate, that shape and pose are factored correctly. In particular changes in $\vec{\beta}$ do not lead to changes in pose.

Principal Components: We orthogonalize the trained shape blend-shapes $\mathcal{S}$ of our female model using PCA and visualize the first three principal components (see Fig. 3a). The first two principal components (PC) correspond to more or less correlated variations in body weight and height. This result is similar to [3] and [28] where the first two components describe variations in gender in addition to weight and height. It differs from the female model in SMPL [26], where the first PC clearly corresponds to body size and the second PC to body weight. We identify two different explanations: (I) Our training set consists of less (approx.

one third of [26]) and different (*e.g.* European vs. North American CAESAR dataset) individuals. Therefore, our training set is very likely to represent another statistical distribution. (II) Missing texture information can lead to sliding of correspondences along the surface [5] which has a negative impact on the model training process. We can not easily analyze (I) due to memory constraints on the GPU and the non-availability of the multi-pose dataset. We analyze (II) by benchmarking our approach on the FAUST correspondence challenge in Section 5.2.

Model Capacity: We evaluate the reconstruction error by incrementally adding shape blend-shapes to our model (see Fig. 3). After training with $B = 10$ shape blend-shapes, we compute the shape and skeleton for each person in the training set and create new blend-shapes using PCA. We vary the number of principal components from 0 to 10 and record the average reconstruction error for each setting (see Fig. 3b). No retraining is performed. The residuals in E_{Data} are interpreted as the reconstruction error and the euclidean distance $||\cdot||_2^2$ is computed for each data point. This error is estimated by sampling $20,000$ points from each scan, computing the distance to the model and averaging the error over all sampled measurements and scans. This error is not sufficient to evaluate the quality of the model but it can give further insight into the training process and model quality. The reconstruction error reduction seems to taper off when using 7 or more principal components. The average reconstruction error for all 10 principal components is 3.7 mm.

5.2 Inference

For inference we keep $\hat{\Theta}$ and $\hat{R}$ fixed and estimate $\hat{\Gamma} = \arg\min_{\Gamma} E(\Gamma; \hat{\Theta}, \hat{R})$. For evaluation we fit our female, male and unisex models on the FAUST dataset [5] which consists of 10 different persons with 30 scans each.

Qualitative Results: The model-fitting results on 10% of the scans of the FAUST dataset are depicted in Fig. 4. The 2D keypoint terms lead to convincing rough alignments. Coarse body proportions are faithfully reconstructed. Few coarse alignment errors occur due to insufficient pose estimates, mostly stemming from touching body parts or from erroneous head alignment by OpenPose when the scan is not facing at least one of the 8 virtual cameras.

Quantitative Results: We evaluate our approach quantitatively using the FAUST intra-subject challenge and inter-subject challenge. The challenges consist of 100 scan pairs where points on the source scan have to be mapped to corresponding points on the target scan. We evaluate registration results on FAUST using the provided evaluation platform. The average error on the intra-subject challenge and inter-subject challenge for our female model are 2.353 cm and 3.234 cm respectively. We compare our results to the current state of the art (see Table 1): In the intra-subject challenge, our approach performs comparably to the listed methods. On the inter-subject challenge our approach is on par with the

Table 1: Quantitative results on the FAUST dataset (average error in cm) and prerequisites for model training that were used to achieve the reported performance.

	Intra-subject	Inter-subject	Dependencies
FAUST [5]	0.7	1.1	strong pose prior
Stitched Puppet [38]	1.568	3.126	SCAPE [3]
3D-CODED [15] (unsup)	N/A	4.835	SURR. [34], SMPL [26]
3D-CODED [15] (sup)	1.985	2.878	SURR. [34], SMPL [26]
Deprelle et al. [10]	1.626	2.578	SURR. [34], 3D-CODED [15]
LBS-AE [23]	2.161	4.079	None
Halimi et al. [16]	2.51	N/A	None
Female only (Ours)	2.353	3.234	OpenPose [7]
Male only (Ours)	2.387	3.519	OpenPose [7]
Unisex (Ours)	2.304	3.525	OpenPose [7]
Female and male (Ours)	2.301	3.422	OpenPose [7]

Stitched Puppet [38] approach which, in contrast to our method, was trained by sampling a pre-trained, strong multi-person shape model. The various variants of the 3D-CODED methods perform worse in their unsupervised settings [15] or require large labeled synthetic training data to outperform model-based approaches [10,15]. Finally, we clearly outperform LBS-AE [23] on the inter-subject challenge. For Halimi et al. [16] no publicly available inter-subject result exists.

6 Discussion and Conclusion

We show that articulated multi-person shape model training can be addressed within a single objective where all parameters are jointly optimized. The proposed method is markerless in the sense that no handcrafted landmarks are required and no pre-existing shape model is required, which might implicitly incorporate expensively generated correspondences. Instead, the landmark term of our objective deals with untextured 3D scans in combination with the output of an off-the-shelf 2D keypoint detector with comparatively low accuracy.

The limiting factor of our approach is the memory of the GPU, which restricts the scalability in terms of the resolution and the amount of training data in comparison to alternating or stochastic optimization methods. On the other hand, the presented results show, that despite the limitations in terms of resolution and variability of the training data, the achieved accuracy in the FAUST correspondence challenge is comparable to strong human shape models that have higher resolution and rely on (semi)-supervised training schemes.

Acknowledgment

We thank A. Bender for the data setup figure. We thank J. Wetzel and N. Link for technical discussion. This work was supported by the German Federal Ministry of Education and Research (BMBF) under Grant 13FH025IX6.

References

1. Agarwal, S., Mierle, K., Others: Ceres solver. `http://ceres-solver.org`
2. Alexa, M., Wardetzky, M.: Discrete laplacians on general polygonal meshes. In: ACM Transactions on Graphics (TOG). vol. 30, p. 102. ACM (2011)
3. Anguelov, D., Srinivasan, P., Koller, D., Thrun, S., Rodgers, J., Davis, J.: Scape: shape completion and animation of people. In: ACM transactions on graphics (TOG). vol. 24, pp. 408–416. ACM (2005)
4. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3d faces. In: Proceedings of the conference on Computer graphics and interactive techniques (SIGGRAPH). vol. 99, pp. 187–194 (1999)
5. Bogo, F., Romero, J., Loper, M., Black, M.J.: Faust: Dataset and evaluation for 3d mesh registration. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3794–3801 (2014)
6. Bogo, F., Romero, J., Pons-Moll, G., Black, M.J.: Dynamic faust: Registering human bodies in motion. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6233–6242 (2017)
7. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
8. Cashman, T.J., Fitzgibbon, A.W.: What shape are dolphins? building 3d morphable models from 2d images. Transactions on pattern analysis and machine intelligence (PAMI) **35**(1), 232–244 (2012)
9. Catmull, E., Clark, J.: Recursively generated b-spline surfaces on arbitrary topological meshes. Computer-aided design **10**(6), 350–355 (1978)
10. Deprelle, T., Groueix, T., Fisher, M., Kim, V., Russell, B., Aubry, M.: Learning elementary structures for 3d shape generation and matching. In: Advances in Neural Information Processing Systems (NIPS). pp. 7433–7443 (2019)
11. DeVito, Z., Mara, M., Zollöfer, M., Bernstein, G., Theobalt, C., Hanrahan, P., Fisher, M., Nießner, M.: Opt: A domain specific language for non-linear least squares optimization in graphics and imaging. ACM Transactions on Graphics 2017 (TOG) (2017)
12. Engel, J., Koltun, V., Cremers, D.: Direct sparse odometry. Transactions on pattern analysis and machine Intelligence (PAMI) **40**(3), 611–625 (March 2018)
13. Farin, G.E., Farin, G.: Curves and surfaces for CAGD: a practical guide. Morgan Kaufmann (2002)
14. Genova, K., Cole, F., Maschinot, A., Sarna, A., Vlasic, D., Freeman, W.T.: Unsupervised training for 3d morphable model regression. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8377–8386 (2018)
15. Groueix, T., Fisher, M., Kim, V.G., Russell, B.C., Aubry, M.: 3d-coded: 3d correspondences by deep deformation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 230–246 (2018)
16. Halimi, O., Litany, O., Rodola, E., Bronstein, A.M., Kimmel, R.: Unsupervised learning of dense shape correspondence. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4370–4379 (2019)
17. Hesse, N., Pujades, S., Romero, J., Black, M.J., Bodensteiner, C., Arens, M., Hofmann, U.G., Tacke, U., Hadders-Algra, M., Weinberger, R., Müller-Felber, W., Schroeder, A.S.: Learning an infant body model from RGB-D data for accurate full body motion analysis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). Springer (2018)

18. Hirshberg, D.A., Loper, M., Rachlin, E., Black, M.J.: Coregistration: Simultaneous alignment and modeling of articulated 3d shape. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 242–255. Springer (2012)
19. Jaimez, M., Cashman, T.J., Fitzgibbon, A., Gonzalez-Jimenez, J., Cremers, D.: An efficient background term for 3d reconstruction and tracking with smooth surface models. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
20. Joo, H., Simon, T., Sheikh, Y.: Total capture: A 3d deformation model for tracking faces, hands, and bodies. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8320–8329 (2018)
21. Khamis, S., Taylor, J., Shotton, J., Keskin, C., Izadi, S., Fitzgibbon, A.: Learning an efficient model of hand shape variation from depth images (June 2015)
22. Lewis, J.P., Cordner, M., Fong, N.: Pose space deformation: a unified approach to shape interpolation and skeleton-driven deformation. In: Proceedings of the conference on Computer graphics and interactive techniques. pp. 165–172 (2000)
23. Li, C.L., Simon, T., Saragih, J., Póczos, B., Sheikh, Y.: Lbs autoencoder: Self-supervised fitting of articulated meshes to point clouds. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11967–11976 (2019)
24. Li, T., Bolkart, T., Black, M.J., Li, H., Romero, J.: Learning a model of facial shape and expression from 4d scans. ACM Transactions on Graphics (TOG) **36**(6), 194 (2017)
25. Loop, C., Schaefer, S.: Approximating catmull-clark subdivision surfaces with bicubic patches. ACM Transactions on Graphics (TOG) **27**(1), 8 (2008)
26. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. ACM transactions on graphics (TOG) **34**(6), 248 (2015)
27. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3d hands, face, and body from a single image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
28. Pishchulin, L., Wuhrer, S., Helten, T., Theobalt, C., Schiele, B.: Building statistical shape spaces for 3d human modeling. Pattern Recognition **67**, 276–286 (2017)
29. Robinette, K.M., Daanen, H., Paquet, E.: The caesar project: a 3-d surface anthropometry survey. In: International Conference on 3-D Digital Imaging and Modeling. pp. 380–386. IEEE (1999)
30. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: Modeling and capturing hands and bodies together. ACM Transactions on Graphics (TOG) **36**(6), 245 (2017)
31. Sorkine, O., Alexa, M.: As-rigid-as-possible surface modeling. In: Symposium on Geometry processing. vol. 4 (2007)
32. Taylor, J., Stebbing, R., Ramakrishna, V., Keskin, C., Shotton, J., Izadi, S., Hertzmann, A., Fitzgibbon, A.: User-specific hand modeling from monocular depth sequences. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 644–651 (2014)
33. Taylor, J., Bordeaux, L., Cashman, T., Corish, B., Keskin, C., Sharp, T., Soto, E., Sweeney, D., Valentin, J., Luff, B., et al.: Efficient and precise interactive hand tracking through joint, continuous optimization of pose and correspondences. ACM Transactions on Graphics (TOG) **35**(4), 143 (2016)
34. Varol, G., Romero, J., Martin, X., Mahmood, N., Black, M.J., Laptev, I., Schmid, C.: Learning from synthetic humans. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)

35. Zach, C.: Robust bundle adjustment revisited. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 772–787. Springer (2014)
36. Zach, C., Bourmaud, G.: Iterated lifting for robust cost optimization. In: Proceedings of the British Machine Vision Conference (BMVC) (2017)
37. Zeitvogel, S., Laubenheimer, A.: Towards end-to-end 3d human avatar shape reconstruction from 4d data. In: International Symposium on Electronics and Telecommunications (ISETC). pp. 1–4. IEEE (2018)
38. Zuffi, S., Black, M.J.: The stitched puppet: A graphical model of 3D human shape and pose. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3537–3546 (2015)