SUPR: A Sparse Unified Part-Based Human Representation

Ahmed A. A. Osman[1], Timo Bolkart[1], Dimitrios Tzionas[2], and Michael J. Black[1]

[1] Max Planck Institute for Intelligent Systems, Tübingen, Germany
[2] University of Amsterdam
{aosman,tbolkart,black}@tuebingen.mpg.de,d.tzionas@uva.nl

Abstract. Statistical 3D shape models of the head, hands, and full body are widely used in computer vision and graphics. Despite their wide use, we show that existing models of the head and hands fail to capture the full range of motion for these parts. Moreover, existing work largely ignores the feet, which are crucial for modeling human movement and have applications in biomechanics, animation, and the footwear industry. The problem is that previous body part models are trained using 3D scans that are isolated to the individual parts. Such data does not capture the full range of motion for such parts, e.g. the motion of head relative to the neck. Our observation is that full-body scans provide important information about the motion of the body parts. Consequently, we propose a new learning scheme that jointly trains a full-body model and specific part models using a federated dataset of full-body and body-part scans. Specifically, we train an expressive human body model called SUPR (Sparse Unified Part-Based Representation), where each joint strictly influences a sparse set of model vertices. The factorized representation enables separating SUPR into an entire suite of body part models: an expressive head (SUPR-Head), an articulated hand (SUPR-Hand), and a novel foot (SUPR-Foot). Note that feet have received little attention and existing 3D body models have highly under-actuated feet. Using novel 4D scans of feet, we train a model with an extended kinematic tree that captures the range of motion of the toes. Additionally, feet deform due to ground contact. To model this, we include a novel non-linear deformation function that predicts foot deformation conditioned on the foot pose, shape, and ground contact. We train SUPR on an unprecedented number of scans: 1.2 million body, head, hand and foot scans. We quantitatively compare SUPR and the separate body parts to existing expressive human body models and body-part models and find that our suite of models generalizes better and captures the body parts' full range of motion. SUPR is publicly available for research purposes.

1 Introduction

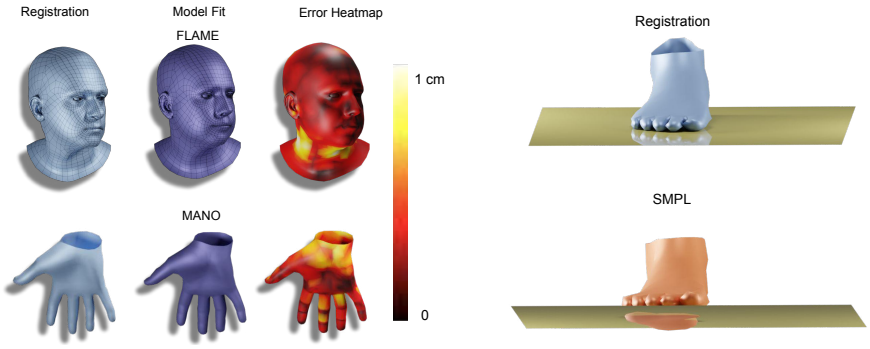
Generative 3D models of the human body and its parts play an important role in understanding human behaviour. Over the past two decades, numerous 3D

Fig. 1: **Expressive part-based human body model.** SUPR is a factorized representation of the human body that can be separated into a full suite of body part models.

models of the body [1,2,3,4,5,6,7,8,9], face [10,11,12,13,14,15,16,17] and hands [18,19,20,21,22,23] have been proposed. Such models enabled a myriad of applications ranging from reconstructing bodies [24,25,26], faces [27,28,29], and hands [30,31] from images and videos, modeling human interactions [32], generating 3D clothed humans [33,34,35,36,37,38,39], or generating humans in scenes [40,41,42]. They are also used as priors for fitting models to a wide range of sensory input measurements like motion capture markers [43,44] or IMUs [45,46,47].

Hand [21,48,22,49], head [12,13,49] and body [6,7] models are typically built independently. Heads and hands are captured with a 3D scanner in which a subject remains static, while the face and hands are articulated. This data is unnatural as it does not capture how the body parts move together with the body. As a consequence, the construction of head/hand models implicitly assumes a static body, and use a simple kinematic tree that fails to model the head/hand full degrees of freedom. For example, in Fig. 2a we fit the FLAME head model [13] to a pose where the subject is looking right and find that FLAME exhibits a significant error in the neck region. Similarly, we fit the MANO [21] hand model to a hand pose where the the wrist is fully bent downwards. MANO fails to capture the wrist deformation that results from the bent wrist. This is a systematic limitation of existing head/hand models, which can not be addressed by simply training on more data.

Another significant limitation of existing body-part models is the lack of an articulated foot model in the literature. This is surprising given the many applications of a 3D foot model in the design, sale, and animation of footwear. Feet are also critical for human locomotion. Any biomechanical or physics-based model must have realistic feet to be faithful. The feet on existing full body models like SMPL are overly simplistic, have limited articulation, and do not deform with contact as shown in Fig. 2b.



(a) Body part models boundary error. (b) SMPL ground penetration.

Fig. 2: **Body part models failure cases.** Left: Existing body part models such as the FLAME [13] head model and the MANO [21] hand model fail to capture the corresponding body part's shape through the full range of motion. Fitting FLAME to a subject looking left results in significant error in the neck region. Similarly, fitting MANO to hands with a bent wrist, results in significant error at the wrist region. Right: The foot of SMPL [6] fails to model deformations due to ground contact, hence penetrating the ground. Additionally, it has a limited number of joints to model the toes articulation.

In contrast to the existing approaches, we propose to jointly train the full human body and body part models together. We first train a new full-body model called SUPR, with articulated hands and an expressive head using a federated dataset of body, hand, head and foot scans. This joint learning captures the full range of motion of the body parts along with the associated deformation. Then, given the learned deformations, we separate the body model into body part models. To enable separating SUPR into compact individual body parts we learn a sparse factorization of the pose-corrective blend shape function as shown in the teaser Fig. 1. The factored representation of SUPR enables separating SUPR into an entire suite of models: SUPR-Head, SUPR-Hand and SUPR-Foot. A body part model is separated by considering all the joints that influence the set of vertices defined by the body part template mesh. We show that the learned kinematic tree structure for the head/hand contains significantly more joints than commonly used by head/hand models. In contrast to the existing body part models that are learned in isolation of the body, our training algorithm unifies many disparate prior efforts and results in a suite of models that can capture the full range of motion of the head, hands, and feet.

SUPR goes beyond existing statistical body models to include a novel foot model. To do so, we extend the standard kinematic tree for the foot to allow more degrees of freedom. To train the model, we capture foot scans using a custom 4D foot scanner (see Sup. Mat.), where the foot is visible from all views, including the sole of the foot which is imaged through a glass plate. This uniquely allows

us to capture how the foot is deformed by contact with the ground. We then model this deformation as a function of body pose and contact.

We train SUPR on 1.2 million hand, head, foot, and body scans, which is an order of magnitude more data than the largest training dataset reported in the literature (60K GHUM [49]). The training data contains extreme body shapes such as anorexia patients and body builders. All subjects gave informed written consent for participation and the use of their data in statistical models. Capture protocols were reviewed by the local university ethics board.

We quantitatively compare SUPR and the individual body-part models to existing models including SMPL-X, GHUM, MANO, and FLAME. We find that SUPR is more expressive, is more accurate, and generalizes better. In summary our main contributions are: (1) A unified framework for learning both expressive body models and a suite of high-fidelity body part models. (2) A novel 3D articulated foot model that captures compression due to contact. (3) SUPR, a sparse expressive and compact body model that generalizes better than existing expressive human body models. (4) An entire suite of body part models for the head, hand and feet, where the model kinematic tree and pose deformation are learned instead of being artist defined. (5) The Tensorflow and a PyTorch implementations of all the models are publicly available for research purposes.

2 Related Work

Body Models: SCAPE [2] is the first 3D model to factor body shape into separate pose and a shape spaces. SCAPE is based on triangle deformations and is not compatible with existing graphics pipelines. In contrast, SMPL [6] is the first learned statistical body model compatible with game engines SMPL is a vertex-based model with linear blend skinning (LBS) and learned pose and shape corrective blendshapes. A key drawback of SMPL is that it relates the pose corrective blendshapes to the elements of the part rotations matrices of all the model joints in the kinematic tree. Consequently, it learns spurious long-range correlations in the training data. STAR [7] addresses many of the drawback of SMPL by using a compact representation of the kinematic tree based on quaternions and learning sparse pose corrective blendshapes where each joint strictly influences a sparse set of the model vertices. The pose corrective blendshape formulation in SUPR is based on STAR. Also related to our work, the *Stitched Puppet* [50] is a part-based model of the human body. The body is segmented into 16 independent parts with learned pose and shape corrective blendshapes. A pairwise stitching function fuses the parts, but leaves visible discontinuities. While SUPR is also part-based model, we start with a unified model and learn its segmentation into parts during training from a federated training dataset.

Expressive Body Models: The most related to SUPR are expressive body models such as Frank [51], SMPL-X [52], and GHUM & GHUML [49,53]. Frank [51] merges the body of SMPL [6] with the FaceWarehouse [12] face model and an artist-defined hand rig. Due to the fusion of different models learned in isolation, Frank looks unrealistic. SMPL-X [52] learns an expressive body model and fuses

the MANO hand model [21] pose blendshapes and the FLAME head model [13] expression space. However, since MANO and FLAME are learned in isolation of the body, they do not capture the full degrees of freedom of the head and hands. Thus, fusing the parameters results in artifacts at the boundaries. In contrast to the construction of Frank and SMPL-X, for SUPR, we start with a coherent full body model, trained on a federated dataset of body, hand, head and feet scans, then separate the model into individual body parts. Xu et al. [49] propose GHUM & GHUML, which are trained on a federated dataset of $60K$ head, hand and body scans and use a fully connected neural network architecture to predict the pose deformation. The GHUM model can not be separated into body parts as a result of the dense fully connected formulation that relates all the vertices to all the joints in the model kinematic tree. In contrast, the SUPR factorized representation of the pose space deformations enables seamless separation of the body into head/hand and foot models.

Head Models: There are many models of 3D head shape [54,55,56], shape and expression [10,11,12,14,15,16,17] or shape, pose and expression [13]. We focus here on models with a full head template, including a neck. The FLAME head model [13], like SMPL, uses a dense pose corrective blendshape formulation that relates all vertices to all joints. Xu et al. [49] also propose GHUM-Head, where the template is based on the GHUM head with a retrained pose dependant corrector network (PSD). Both GHUM-Head and FLAME are trained in isolation of the body and do not have sufficient joints to model the full head degrees of freedom. In contrast to the previous methods, SUPR-Head is trained jointly with the body on a federated dataset of head and body meshes, which is critical to model the head full range of motion. It also has more joints than GHUM-Head or FLAME, which we show is crucial to model the head full range of motion.

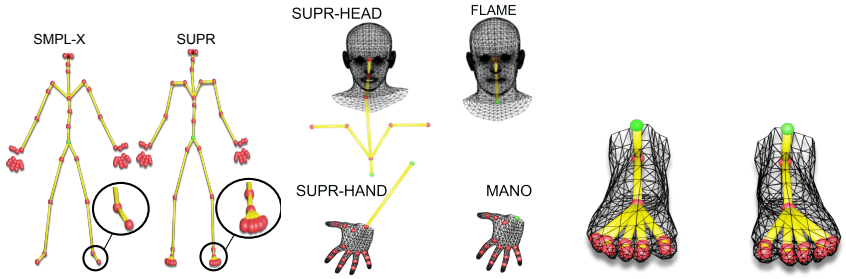
Hand Models: MANO [21] is widely use and is based on the SMPL formulation where the pose corrective blendshapes deformations are regularised to be local. The kinematic tree of MANO is based on spherical joints allowing redundant degrees of freedom for the fingers. Xu et al. [49] introduce the GHUM-Hand model where they separate the hands from the template mesh of GHUM and train a hand-specific pose-dependant corrector network (PSD). Both MANO and GHUM-Hand are trained in isolation of the body and result in implausible deformation around the wrist area. SUPR-Hand is trained jointly with the body and has a wrist joint which is critical to model the hands full range of motion.

Foot Models: Statistical shape models of the feet are less studied than those of the body, head, and hands. Conard et al. [57] propose a statistical shape model of the human foot, which is a PCA space learned from static foot scans. However, the human feet deform with motion and models learned from static scans can not capture the complexity of 3D foot deformations. To address the limitations of static scans, Boppana et al. [58] propose the DynaMo system to capture scans of the feet in motion and learn a PCA-based model from the scans. However, the DynaMo setup fails to capture the sole of the foot in motion. In contrast, to all prior work, SUPR-Foot contains a kinematic tree, a pose deformation space,

and a PCA shape space. We use a specialized 4D foot scanner, where the entire human foot is visible and accurately reconstructed, including the toes and the sole. Furthermore, we go beyond previous work to model the foot deformations resulting from ground contact, which was not possible before.

3 Model

We describe the formulation of SUPR in Section 3.1, followed by how we separate SUPR into body parts models in Section 3.2. Since existing body corrective deformation formulations fail to model foot deformations due to ground contact, we discuss a novel foot deformation network in Section 3.3.



(a) Body (b) Head/Hand (c) SUPR-Foot

Fig. 3: The kinematic tree of SUPR and the separated body part models.

3.1 SUPR

SUPR is a vertex-based 3D model with linear blend skinning (LBS) and learned blend shapes. The blend shapes are decomposed into 3 types: *Shape Blend Shapes* to capture the subject identity, *Pose-Corrective Blend Shapes* to correct for the widely-known LBS artifacts, and *Expression Blend Shapes* to model facial expressions. The SUPR mesh topology and kinematic tree are based on the SMPL-X topology. The template mesh contains $N = 10,475$ vertices and $K = 75$ joints. The SUPR kinematic tree is shown in Figure 3. In contrast to existing body models, the SUPR kinematic tree contains significantly more joints in the foot, ankle and toes as shown in Fig. 3a. Following the notation of SMPL, SUPR is defined by a function $M(\vec{\theta}, \vec{\beta}, \vec{\psi})$, where $\vec{\theta} \in \mathbb{R}^{75\times3}$ are the pose parameters corresponding to the individual bone rotations, $\vec{\beta} \in \mathbb{R}^{300}$ are the shape parameters corresponding to the subject identity, $\vec{\psi} \in \mathbb{R}^{100}$ are the expression parameters controlling facial expressions. Formally, SUPR is defined as

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

where the 3D body, $T_p(\vec{\theta}, \vec{\beta}, \vec{\psi})$, is transformed around the joints J by the linear-blend-skinning function $W(.)$, parameterized by the skinning weights $\mathcal{W} \in \mathbb{R}^{10475 \times 75}$. The cumulative corrective blend shapes term is defined as

$$T_p(\vec{\theta}, \vec{\beta}, \vec{\psi}) = \overline{T} + B_S(\vec{\beta}; \mathcal{S}) + B_P(\vec{\theta}; \mathcal{P}) + B_E(\vec{\psi}; \mathcal{E}), \tag{2}$$

where $\overline{T} \in \mathbb{R}^{10475 \times 3}$ is the template of the mean body shape, which is deformed by: $B_S(\vec{\beta}; \mathcal{S})$, the shape blend shape function capturing a PCA space of body shapes; $B_P(\vec{\theta}; \mathcal{P})$, the pose-corrective blend shapes that address the LBS artifacts; and $B_E(\vec{\psi}; \mathcal{E})$, a PCA space of facial expressions.

Sparse Pose Blend Shapes In order to separate SUPR into body parts, each joint should strictly influence a subset of the template vertices $\overline{T}$. To this end, we base the pose-corrective blend shapes $B_p(.)$ in Eq. 2 on the STAR model [7]. The pose-corrective blend shape function is factored into per-joint pose corrective blend shape functions

$$B_P(\vec{q}, \mathbf{K}, \mathbf{A}) = \sum_{j=1}^{K-1} B_P^j(\vec{q}_{ne(j)}; \mathbf{K}_j; A_j), \tag{3}$$

where the pose-corrective blend shapes are sum of $K-1$ sparse spatially-local pose-corrective blend-shape functions. Each joint-based corrective blend shape $B_P^j(.)$, predicts corrective offsets for a sparse set of the model vertices, defined by the learned joint activation weights $A_j \in \mathbb{R}^{10475}$. Each A_j is a sparse vector defining the sparse set of vertices influenced by the j^{th} joint blend shape $B_p^j(.)$. The joint corrective blend shape function is conditioned on the normalized unit quaternions $\vec{q}_{ne(j)}$ of the j^{th} joint's direct neighbouring joints' pose parameters. We note that the SUPR pose blend-shape formulation in Eq. 3 is not conditioned on body shape, unlike STAR, since the additional body-shape blend shape is not sparse and, hence, can not be factorized into body parts. Since the skinning weights in Eq. 1 and the pose-corrective blend-shape formulation in Eq. 3 are sparse, each vertex in the model is related to a small subset of the model joints. This sparse formulation of the pose space is key to separating the model into compact body part models.

3.2 Body Part Models

In traditional body part models like FLAME and MANO, the kinematic tree is designed by an artist and the models are learned in isolation of the body. In contrast, here the pose-corrective blend shapes of the hand (SUPR-Hand), head (SUPR-Head) and foot (SUPR-Foot) models are trained jointly with the body on a federated dataset. The kinematic tree of each part model is inferred from SUPR rather than being artist defined. To separate a body part, we first define the subset of mesh vertices of the body part $\overline{T}_{bp}$ from the SUPR template $\overline{T}_{bp} \in \overline{T}$. Since the learned SUPR skinning weights and pose-corrective blend shapes are

strictly sparse, any subset of the model vertices $\overline{T}_{bp}$ is strictly influenced by a subset of the model joints. More formally, a joint $\vec{j}$ is deemed to influence a body part defined by the template $\overline{T}_{bp}$ if:

$$\mathbb{I}(T_{bp}, \vec{j}) = \begin{cases} 1 & \text{if } \sum \mathcal{W}(\overline{T}_{bp}, \vec{j}) \neq 0 \text{ or } \sum A_j(\overline{T}_{bp}) \neq 0 \\ 0 & \text{othewise,} \end{cases} \tag{4}$$

where $\mathbb{I}(.,.)$ is an indicator function, $\mathcal{W}(\overline{T}_{bp}, \vec{j})$ is a subset of the SUPR learned skinning weights matrix, where the rows are defined by the vertices of $\overline{T}_{bp}$, the columns correspond to the j^{th} joint, $\vec{j}$, $A_j(\overline{T}_{bp})$ corresponds to the learned activation for the j^{th} joint and the rows defined by vertices $\overline{T}_{bp}$. The indicator function $\mathbb{I}$ returns 1 if a joint $\vec{j}$ has non-zero skinning weights or a non-zero activation for the vertices defined by $\overline{T}_{bp}$. Therefore the set of joints J_{bp} that influences the template $\overline{T}_{bp}$ is defined by:

$$J_{bp} = \Big\{ \mathbb{I}(\overline{T}_{bp}, j) = 1 \ \ \forall \ \ j \in \{1, \ldots, K\} \Big\}. \tag{5}$$

The kinematic tree defined for the body part models in Eq. 5 is implicitly defined by the learned skinning weights $\mathcal{W}$ and the per joint activation weights A_j. The resulting kinematic tree of the separated models is shown in Fig. 3b. Surprisingly, the head is influenced by substantially more joints than in the artist-designed kinematic tree used in FLAME. Similarly, SUPR-Hand has an additional wrist joint compared to MANO. We note here that the additional joints in SUPR-Head and SUPR-Hand are outside the head/hand mesh. The additional joints for the head and the hand are beyond the scanning volume of a body part head/hand scanner. This means that it is not possible to learn the influence of the shoulder and spine joints on the neck from head scans alone.

The skinning weights for a separated body are defined by $\mathcal{W}_{bp} = \mathcal{W}(\overline{T}_{bp}, J_{bp})$, where $\mathcal{W}(\overline{T}_{bp}, J_{bp})$ is the subset of the SUPR skinning weights defined by the rows corresponding to the vertices of $\overline{T}_{bp}$ and the columns defined by J_{bp}. Similarly, the pose corrective blendshapes are defined by $B_{bp} = B_p(\overline{T}_{bp}, J_{bp})$ where $B_p(\overline{T}_{bp}, J_{bp})$ corresponds to a subset of SUPR pose blend shapes defined by the vertices of $\overline{T}_{bp}$ and the quaternion features for the set of joints J_{bp}. The skinning weights $\mathcal{W}_{bp}$ and blendshapes B_{bp} are based on the SUPR learned blend shapes and skinning weights, which are trained on a federated dataset that explores each body part's full range of motion relative to the body. We additionally train a joint regressor $\mathcal{J}_{bp}$, to regress the joints $\mathcal{J}_{\text{bp}} : \overline{T}_{bp} \rightarrow J_{bp}$. We learn a local body part shape space $B_S(\vec{\beta}_{bp}; \mathcal{S}_{bp})$, where $\mathcal{S}_{bp}$ is the body part PCA shape components. For the head, we use the SUPR learned expression space $B_E(\psi; \mathcal{E})$.

3.3 Foot deformation Network

The linear pose-corrective blend shapes in Eq. 2 and Eq. 3 relate the body deformations to the body pose only. However, the human foot deforms as a function

of pose, shape and ground contact. To model this, we add a foot deformation network.

The foot body part model, separated from SUPR, is defined by the pose parameters $\vec{\theta}_{bp} \in \vec{\theta}$, corresponding to the ankle and toe pose parameters in addition to $\vec{\beta}_{bp}$, the PCA coefficients of the local foot shape space. We extend the pose blend shapes in Eq. 2 to include a deep corrective deformation term for the foot vertices defined by $\overline{T}_{foot} \in \overline{T}$. With a slight abuse of notation, we will refer to the deformation function $T_p(\vec{\theta}, \vec{\beta}, \vec{\psi})$ in Eq. 2 as T_p for simplicity. The foot deformation function is defined by:

$$T'_p(\vec{\theta}, \vec{\beta}, \vec{c}) = T_p + \vec{m} \circ B_F(\vec{\theta}_{foot}, \vec{\beta}_{foot}, \vec{c}; \mathcal{F}), \tag{6}$$

where $\vec{m} \in \{0, 1\}^{10475}$ is a binary vector with ones corresponding to the foot vertices and 0 elsewhere. $B_F(.)$ is a multilayer perceptron-based deformation function parameterized by $\mathcal{F}$, conditioned on the foot pose parameters $\vec{\theta}_{foot}$, foot shape parameters $\vec{\beta}_{foot}$ and foot contact state $\vec{c}$. The foot contact state variable is a binary vector $\vec{c} \in \{0, 1\}^{266}$ defining the contact state of each vertex in the foot template mesh, a vertex is represented by a 1 if it is in contact with the ground, and 0 otherwise. The Hadamard product between $\vec{m}$ and $B_F(.)$ ensures the network $B_F(.)$ strictly predicts deformations for the foot vertices only.

Implementation details. The foot contact deformation network is based on an encoder-decoder architecture. The input feature, $\vec{f} \in \mathbb{R}^{320}$, to the encoder is a concatenated feature of the foot pose, shape and contact vector. The foot pose is represented with a normalised unit quaternion representation, shape is encoded with the first two PCA coefficients of the local foot shape space. The input feature $\vec{f}$ is encoded into a latent vector $\vec{z} \in \mathbb{R}^{16}$ using fully connected layers with a leaky LReLU as an activation function with a slope of 0.1 for negative values. The latent embedding $\vec{z}$ is decoded to predict deformations for each vertex using fully connected layers with LReLU activation. The full architecture is described in detail in Supp. Mat. We train male, female and a gender-neutral versions of SUPR and the separated body part models. Training details are discussed in Supp. Mat.

4 Experiments

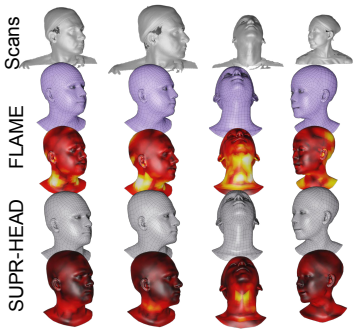
Our goal is to evaluate the generalization of SUPR and the separated head, hand, and foot model to unseen test subjects. We first evaluate the full SUPR body model against existing state of the art expressive human body models SMPL-X and GHUM (Section 4.1), then we evaluate the separated SUPR-Head model against existing head models FLAME and GHUM-Head (Section 4.2), and compare the hand model to GHUM-Hand and MANO (Section 4.3). Finally, we evaluate the SUPR-Foot (Section 4.4).

4.1 Full-Body Evaluation

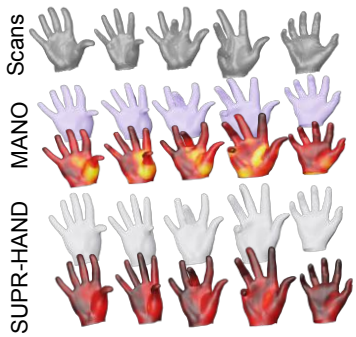
We use the publicly available 3DBodyTex dataset [59], which includes 100 male and 100 female subjects. We register the GHUM template and the SMPL-X tem-

(a) 3DBodyTex scans.

(b) Model Fits

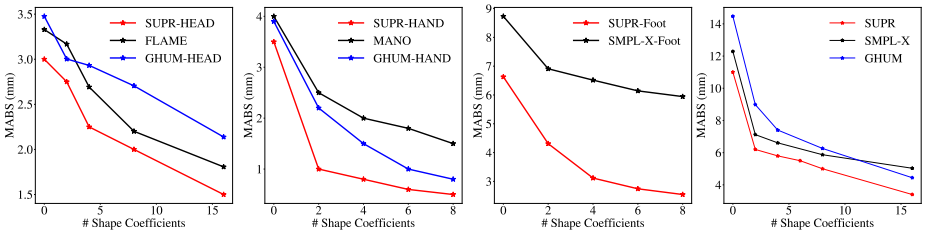


(c) Head Evaluation



(d) Hand Evaluation

Fig. 4: **Qualitative Evaluation:** We evaluate SUPR and the separated body part models against baselines. We use the 3DBodyTex dataset in Fig. 4a to evaluate GHUM, SMPL-X and SUPR in Fig. 4b using 16 shape components. We evaluate SUPR-Head against FLAME in Fig. 4c using 16 shape components and SUPR-Hand against MANO in Fig. 4d using 8 shape components.



(a) Head Evaluation (b) Hand Evaluation (c) Foot Evaluation (d) Body Evaluation

Fig. 5: **Quantitative Evaluation:** Evaluating the generalization of the separated head, hand and foot model from SUPR against existing body part models: GHUM-HEAD and FLAME for the head (Fig. 5a), GHUM-HAND and MANO (Fig. 5b). We report the *vertex-to-vertex* error (*mm*) as a function of the number of the shape coefficients used when fitting each model to the test set.

plate to all the scans; note SMPL-X and SUPR share the same mesh topology. We visually inspected all registered meshes for quality control. Given registered meshes, we fit each model by minimizing the vertex-to-vertex loss (*v2v*) between the model surface and the corresponding registration. The free optimization parameters for all models are the pose parameters $\vec{\theta}$ and the shape parameters $\vec{\beta}$. Note that, for fair comparison with GHUM, we only report errors for up to 16 shape components since this is the maximum in the GHUM release. SUPR includes 300 shape components that would reduce the errors significantly.

We follow the 3DBodyTex evaluation protocol and exclude the face and the hands when reporting the mean absolute error (*mabs*). We report the mean absolute error of each model on both male and female registrations. For the GHUM model, we use the PCA-based shape and expression space. We report the model generalization error in Fig. 5d and show a qualitative sample of the model fits in Fig. 4b. SUPR uniformly exhibits a lower error than SMPL-X and GHUM.

4.2 Head Evaluation

The head evaluation test set contains a total of 3 male and 3 female subjects, with sequences containing extreme facial expression, jaw movement and neck movement. As for the full body, we register the GHUM-Head model and the FLAME template to the test scans, and use these registered meshes for evaluation. For the GHUM-Head model, we use the linear PCA expression and shape space. We evaluate all models using a standard *v2v* objective, where the optimization free variables are the model pose, shape parameters, and expression parameters. We use 16 expression parameters when fitting all models. For GHUM-Head we exclude the internal head geometry (corresponding to a tongue-like structure) when reporting the *v*2*v* error. Fig. 5a shows the model generalization as a function of the number of shape components. We show a sample of the model fits in

Fig. 4c. Both GHUM-Head and FLAME fail to capture head-to-neck rotations plausibly, despite each featuring a full head mesh including a neck. This is clearly highlighted by the systematic error around the neck region in Fig. 4c. In contrast, SUPR-Head captures the head deformations and the neck deformations plausibly and uniformly generalizes better.

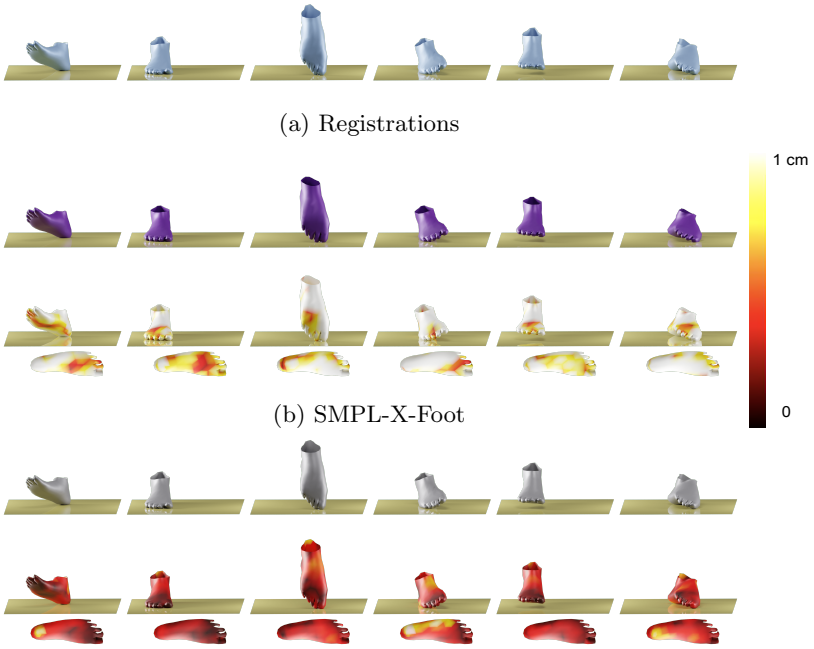
4.3 Hand Evaluation

We use the publicly available MANO test set [21]. Since both SUPR-Hand and MANO share the same topology, we used the MANO test registrations provided by the authors to evaluate both models. To evaluate GHUM-Hand, we register the model to the MANO test set. However, the GHUM-Hand features a hand and an entire forearm, therefore to register GHUM-Hand we selected vertices on the model corresponding to the hand and only register that hand part of the model to the MANO scans. We fit all models to the corresponding registrations using a standard *v2v* loss. For GHUM-Hand, we fit the model only to the selected hand vertices. The optimization free variables are the model pose and shape parameters. Fig. 5b shows generalization as a function of the number of shape parameters, where SUPR-Hand uniformly exhibits a lower error compared to both MANO and GHUM-Hand. A sample qualitative evaluation of MANO and SUPR-Hand is shown in Fig. 4d. In addition to a lower overall fitting error, SUPR-Hand has a lower error around the wrist region than MANO.

4.4 Foot Evaluation

We evaluate SUPR-Foot generalization on a test set of held-out subjects. The test set contains 120 registrations for 5 subjects that explore the foot's full range of motion, such as ankle and toe movements. We extract the foot from the SMPL-X body model as a baseline and refer to it as SMPL-X-Foot. We register the SUPR-Foot template to the test scans and fit the SUPR-Foot and SMPL-X-Foot to the registrations using a standard *v2v* objective. For SUPR-Foot, the optimization free variables are the model pose and shape parameters, while for SMPL-X-Foot the optimization free variables are the foot joints and the SMPL-X shape parameters. We report the models' generalization as a function of the number of shape components in Fig. 5c. A sample of the model fits are shown in Fig. 6. SUPR-Foot better captures the degrees of freedom of the foot, such as moving the ankle, curling the toes, and contact deformations.

Dynamic Evaluation We further evaluate the foot deformation network on a dynamic sequence shown in Fig. 7. Fig. 7a shows raw scanner footage of a subject performing a body rocking movement, where they lean forward then backward effectively changing the body center of mass. We visualise the corresponding SUPR-Foot fits and a heat map of the magnitude of predicted deformations in Fig. 7b. When the subject is leaning backward and the center of mass is directly above the ankle, the soft tissue at heel region of the foot deforms due to contact.



(a) Registrations

(b) SMPL-X-Foot

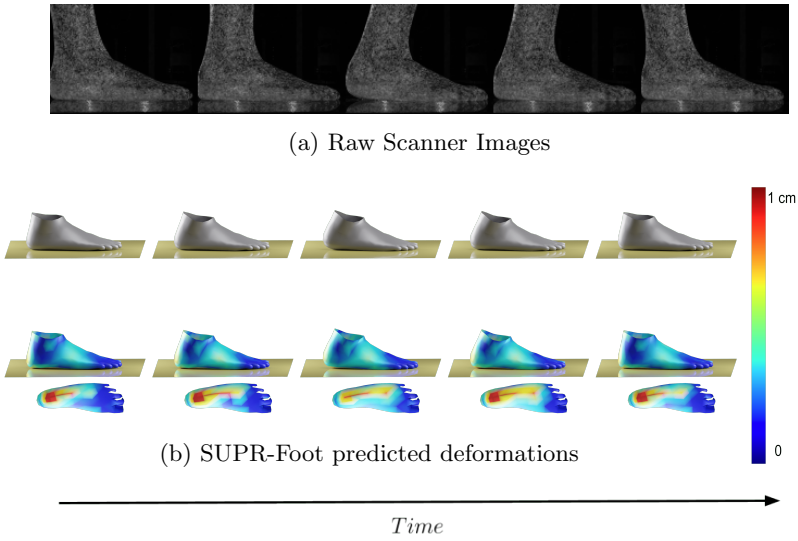
(c) SUPR-Foot

Fig. 6: Evaluating SUPR-Foot against SMPL-X-Foot.

The SUPR-Foot network predicts significant deformations localised around the heel region compared to the rest of the foot. However, when the subject leans forward the center of mass is above the toes, consequently the soft tissue at the heel is less compressed. The SUPR-Foot predicted deformations shift from the heel towards the front of the foot.

5 Conclusion

We present a novel training algorithm for jointly learning high-fidelity expressive full-body and body parts models. We highlight a critical drawback in existing body part models such as FLAME and MANO, which fail to model the full range of motion of the head/hand. We identify that the issue stems from the current practice in which body parts are modeled with a simplified kinematic tree in isolation from the body. Alternatively, we propose a holistic approach where the body and body parts are jointly trained on a federated dataset that contains the body parts' full range of motion relative to the body. Additionally, we point out the lack of any articulated foot model in the literature and show that the feet of existing full-body models do not have enough joints to model the full range of motion of the foot. Using 4D scans, we learn a foot model with a novel pose-corrective deformation formulation that is conditioned on the foot pose, its shape, and ground contact information. We train SUPR



(a) Raw Scanner Images

(b) SUPR-Foot predicted deformations

Fig. 7: **Dynamic Evaluation**: Evaluating the SUPR-Foot predicted deformations on a dynamic sequence where the subject leans backward and forward, effectively shifting their center of mass.

with a federated dataset of 1.2 million scans of the body, hands, and feet. The sparse formulation of SUPR enables separating the model into an entire suite of body-part models. Surprisingly, we show that the head and hand models are influenced by significantly more joints than commonly used in existing models. We thoroughly compare SUPR and the separated models against SMPL-X, GHUM, MANO and FLAME and show that the models uniformly generalize better and have a significantly lower error when fitting test data. The pose-corrective blendshapes of SUPR and the separated body part models are linearly related to the kinematic tree pose parameters, therefore our new formulation is fully compatible with the existing animation and gaming industry standards. A Tensorflow and PyTorch implementation of SUPR and the separated head (SUPR-Head), hand (SUPR-Hand) and the foot (SUPR-Foot) models is publicly available for research purposes.

Acknowledgments: The authors thank the MPI-IS Tübingen members of the data capture team since 2012 for capturing the data used to train SUPR: S. Polikovsky, A. Keller, E. Holderness, J. Márquez, T. Alexiadis, M. Höschle, M. Landry, G. Henz, M. Safroshkin, M. Landry, T. McConnell, T. Bauch and B. Pellkofer for the IT support. The authors thank M. Safroshkin and M. Landry for configuring the foot scanner. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Ahmed A. A. Osman. This work was done when DT was at MPI.

MJB Disclosure: https://files.is.tue.mpg.de/black/CoI_ECCV_2022.txt

References

1. Brett Allen, Brian Curless, Brian Curless, and Zoran Popović. The space of human body shapes: Reconstruction and parameterization from range scans. *ACM TOG*, 22(3):587–594, 2003. 2
2. D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis. SCAPE: Shape Completion and Animation of PEople. *ACM TOG*, 24(3):408–416, 2005. 2, 4
3. Yinpeng Chen, Zicheng Liu, and Zhengyou Zhang. Tensor-based human body modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 105–112, 2013. 2
4. Nils Hasler, Carsten Stoll, Martin Sunkel, Bodo Rosenhahn, and Hans-Peter Seidel. A statistical model of human pose and body shape. *Comput. Graph. Forum*, 28(2):337–346, 2009. 2
5. David A. Hirshberg, Matthew Loper, Eric Rachlin, and Michael J. Black. Coregistration: Simultaneous alignment and modeling of articulated 3D shape. In *European Conference on Computer Vision*, volume 7577, pages 242–255, 2012. 2
6. 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, October 2015. 2, 3, 4
7. Ahmed A. A. Osman, Timo Bolkart, and Michael J. Black. STAR: Sparse trained articulated human body regressor. In *ECCV*, pages 598–613, 2020. 2, 4, 7
8. Leonid Pishchulin, Stefanie Wuhrer, Thomas Helten, Christian Theobalt, and Bernt Schiele. Building statistical shape spaces for 3D human modeling. *PR*, 67:276–286, 2017. 2
9. Haoyang Wang, Riza Alp Guler, Iasonas Kokkinos, George Papandreou, and Stefanos Zafeiriou. BLSM: A bone-level skinned model of the human mesh. In *ECCV*, pages 1–17, 2020. 2
10. Brian Amberg, Reinhard Knothe, and Thomas Vetter. Expression invariant 3D face recognition with a morphable model. pages 1–6, 2008. 2, 5
11. Alan Brunton, Timo Bolkart, and Stefanie Wuhrer. Multilinear wavelets: A statistical shape space for human faces. In *ECCV*, pages 297–312, 2014. 2, 5
12. 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, 4, 5
13. 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, (Proc. SIGGRAPH Asia)*, 36(6), 2017. 2, 3, 5
14. Ruilong Li, Karl Bladin, Yajie Zhao, Chinmay Chinara, Owen Ingraham, Pengda Xiang, Xinglei Ren, Pratusha Prasad, Bipin Kishore, Jun Xing, et al. Learning formation of physically-based face attributes. In *CVPR*, pages 3410–3419, 2020. 2, 5
15. Anurag Ranjan, Timo Bolkart, Soubhik Sanyal, and Michael J. Black. Generating 3D faces using convolutional mesh autoencoders. In *ECCV*, pages 725–741, 2018. 2, 5
16. Haotian Yang, Hao Zhu, Yanru Wang, Mingkai Huang, Qiu Shen, Ruigang Yang, and Xun Cao. FaceScape: a large-scale high quality 3D face dataset and detailed riggable 3D face prediction. In *CVPR*, pages 601–610, 2020. 2, 5

17. Daniel Vlasic, Matthew Brand, Hanspeter Pfister, and Jovan Popovic. Face transfer with multilinear models. *ACM TOG*, 24(3):426–433, 2005. 2, 5
18. 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2540–2548, 2015. 2
19. Dominik Kulon, Haoyang Wang, Riza Alp Güler, Michael M. Bronstein, and Stefanos Zafeiriou. Single image 3D hand reconstruction with mesh convolutions. In *BMVC*, page 45, 2019. 2
20. Iason Oikonomidis, Nikolaos Kyriazis, and Antonis A. Argyros. Efficient model-based 3D tracking of hand articulations using kinect. In *BMVC*, pages 1–11, 2011. 2
21. Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6):245:1–245:17, November 2017. 2, 3, 5, 12
22. Breannan Smith, Chenglei Wu, He Wen, Patrick Peluse, Yaser Sheikh, Jessica K. Hodgins, and Takaaki Shiratori. Constraining dense hand surface tracking with elasticity. *ACM TOG*, 39(6):219:1–219:14, 2020. 2
23. Anastasia Tkach, Mark Pauly, and Andrea Tagliasacchi. Sphere-meshes for real-time hand modeling and tracking. *ACM TOG*, 35(6):222:1–222:11, 2016. 2
24. Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7122–7131, 2018. 2
25. Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. *arXiv preprint arXiv:1912.05656*, 2019. 2
26. Nikos Kolotouros, Georgios Pavlakos, Michael J. Black, and Kostas Daniilidis. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In *ICCV*, pages 2252–2261, 2019. 2
27. Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. Joint 3D face reconstruction and dense alignment with position map regression network. In *ECCV*, pages 534–551, 2018. 2
28. Ayush Tewari, Florian Bernard, Pablo Garrido, Gaurav Bharaj, Mohamed Elgharib, Hans-Peter Seidel, Patrick Pérez, Michael Zollhöfer, and Christian Theobalt. FML: Face Model Learning from Videos. In *CVPR*, pages 10812–10822, 2019. 2
29. Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael Black. Learning to regress 3D face shape and expression from an image without 3D supervision. In *CVPR*, pages 7763–7772, 2019. 2
30. Adnane Boukhayma, Rodrigo de Bem, and Philip H. S. Torr. 3D hand shape and pose from images in the wild. In *CVPR*, pages 10843–10852, 2019. 2
31. Yana Hasson, Gül Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *CVPR*, pages 11807–11816, 2019. 2
32. Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Three-dimensional reconstruction of human interactions. In *CVPR*, pages 7214–7223, 2020. 2
33. Thiemo Alldieck, Gerard Pons-Moll, Christian Theobalt, and Marcus Magnor. Tex2Shape: Detailed full human body geometry from a single image. In *ICCV*, pages 2293–2303, 2019. 2

34. Christoph Lassner, Gerard Pons-Moll, and Peter V. Gehler. A generative model of people in clothing. In *ICCV*, pages 853–862, 2017. 2
35. Qianli Ma, Jinlong Yang, Anurag Ranjan, Sergi Pujades, Gerard Pons-Moll, Siyu Tang, and Michael J. Black. Learning to dress 3D people in generative clothing. In *CVPR*, pages 6468–6477, 2020. 2
36. Chao Zhang, Sergi Pujades, Michael Black, and Gerard Pons-Moll. Detailed, accurate, human shape estimation from clothed 3D scan sequences. In *CVPR*, pages 5484–5493, 2017. 2
37. Gerard Pons-Moll, Sergi Pujades, Sonny Hu, and Michael Black. ClothCap: Seamless 4D clothing capture and retargeting. *ACM TOG*, 36(4):73:1–73:15. 2
38. Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. Multi-garment net: Learning to dress 3D people from images. In *ICCV*, pages 5419–5429, 2019. 2
39. Chaitanya Patel, Zhouyingcheng Liao, and Gerard Pons-Moll. TailorNet: Predicting clothing in 3D as a function of human pose, shape and garment style. In *CVPR*, pages 7363–7373, 2020. 2
40. Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Andrei Zanfir, and Cristian Sminchisescu. Human synthesis and scene compositing. In *AAAI*, pages 12749–12756, 2020. 2
41. Yan Zhang, Mohamed Hassan, Heiko Neumann, Michael J Black, and Siyu Tang. Generating 3D people in scenes without people. In *CVPR*, pages 6194–6204, 2020. 2
42. Siwei Zhang, Yan Zhang, Qianli Ma, Michael J. Black, and Siyu Tang. PLACE: Proximity learning of articulation and contact in 3D environments. 2020. 2
43. 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, November 2014. 2
44. Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *ICCV*, pages 5442–5451, 2019. 2
45. Timo von Marcard, Gerard Pons-Moll, and Bodo Rosenhahn. Human pose estimation from video and IMUs. *IEEE TPAMI*, 38(8):1533–1547, 2016. 2
46. 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. pages 421–430, 2017. 2
47. Yinghao Huang, Manuel Kaufmann, Emre Aksan, Michael J. Black, Otmar Hilliges, and Gerard Pons-Moll. Deep inertial poser: Learning to reconstruct human pose from sparse inertial measurements in real time. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 37:185:1–185:15, November 2018. Two first authors contributed equally. 2
48. Gyeongsik Moon, Takaaki Shiratori, and Kyoung Mu Lee. Deephandmesh: A weakly-supervised deep encoder-decoder framework for high-fidelity hand mesh modeling. In *European Conference on Computer Vision (ECCV)*, 2020. 2
49. Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. GHUM & GHUML: Generative 3D human shape and articulated pose models. In *CVPR*, pages 6184–6193, 2020. 2, 4, 5
50. Silvia Zuffi and Michael J Black. The stitched puppet: A graphical model of 3d human shape and pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3537–3546, 2015. 4

51. Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3D deformation model for tracking faces, hands, and bodies. In *CVPR*, pages 8320–8329, 2018. 4
52. Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. 4
53. Andrei Zanfir, Eduard Gabriel Bazavan, Hongyi Xu, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Weakly supervised 3d human pose and shape reconstruction with normalizing flows. In *European Conference on Computer Vision*, pages 465–481. Springer, 2020. 4
54. Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3D faces. In *SIGGRAPH*, volume 99, pages 187–194, 1999. 5
55. James Booth, Anastasios Roussos, Allan Ponniah, David Dunaway, and Stefanos Zafeiriou. Large scale 3D morphable models. *IJCV*, 126(2-4):233–254, 2018. 5
56. 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. Ieee, 2009. 5
57. Bryan P Conrad, Michael Amos, Irene Sintini, Brian Robert Polasek, and Peter Laz. Statistical shape modelling describes anatomic variation in the foot. *Footwear Science*, 11(sup1):S203–S205, 2019. 5
58. Abhishektha Boppana and Allison P Anderson. Dynamic foot morphology explained through 4d scanning and shape modeling. *Journal of Biomechanics*, 122:110465, 2021. 5
59. Alexandre Saint, Eman Ahmed, Kseniya Cherenkova, Gleb Gusev, Djamila Aouada, Bjorn Ottersten, et al. 3DBodyTex: Textured 3D body dataset. pages 495–504, 2018. 9