SUPR: A Sparse Unified Part-Based Human Representation
*** Supplementary Material ***

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

1 Overview

Paper Summary. We highlight drawbacks in widely used body part models. Existing head/hand models do not model the head/hand full range of motion. We address this with a new holistic learning scheme in which we train body parts models jointly with the body. To this end, we train SUPR an expressive human body model, where the each joint strictly influences a sparse set of the model vertices. This sparse factorization enables us to separate SUPR into a full suite of high fidelity body part models. We show that body part hand/head models learned jointly with the body are influenced by significantly more joints than existing part models. Additionally, we note that, despite the importance of the foot for locomotion, there is no existing foot part model and the feet of full body models are significantly under actuated. To address this, we learn a foot model from novel 4D scans and propose a novel function that relates the foot deformation to the foot shape, pose and ground contact.

1.1 Paper Content

In this Supplementary Material, we perform extensive ablation studies and evaluations to further explore the main paper's key contributions. The rest of the document is arranged as follows: in Section 2 we describe the federated training dataset of scans, including the foot scanner that enables us to model the foot deformations due to contact. The foot deformation architecture is described in detail in Section 3. We describe the SUPR training in Section 4. In Section 5 we further evaluate SUPR. Ablations for the SUPR-Foot network are introduced in Section 6. As introduced in the main paper, SUPR is based on spherical joints, which produce redundant degrees of freedom (DoF) for joints like those of the fingers. In Section 7 we describe a constrained version of the kinematic tree with fewer DoF. We provide a comparison between SUPR and existing expressive human body models in Section 8. We conclude by discussing limitations of our method in Section 9

2 Data

SUPR is trained on a federated dataset of 3D scans. In total 4 types of scanners are used: a full body scanner, a hand scanner, a head scanner and a foot scanner. All the scanners are 4D scanners, capturing high resolution dynamic sequences for each body part. We additionally leverage datasets that are either publicly available for research purposes or commercial datasets from private vendors. In this section we describe the scanning setup for each scanner and describe the external datasets. We discuss the foot scanner in Section 2.5 which was key to enable us to capture the foot including the toes and the foot soles. In the main paper we highlight that SUPR and the separated body part models were trained on 1.2 millions scans, a break down of the number of scans for each body part is discussed in Section 2.6.

2.1 Body Scanner

Human bodies deform in complex ways as a result of changes in body pose and body shape. To study and model minimally-clothed human body deformations, we use a 4D scanner that captures the full 3D human body shape at 60 frames per second (fps). The full-body scanner is custom built by 3dMD (Atlanta, GA). The system uses 22 pairs of stereo cameras, 22 color cameras, and speckle-light projectors. The speckle patterns allow accurate stereo reconstruction of 3D shape. This speckle pattern alternates at 120 fps with large white-light LED panels that provide a smooth nearly uniform illumination. The scanner outputs high resolution meshes with approximately 150,000 vertices. The high resolution meshes in addition to the high frame rate (60 fps) enable us to model the subtle deformations of the human body. The full body scanner scanning volume is sufficient to capture poses such as a full leg split by a ballerina, or a sitting or lying down poses.

The captured data contains a wide diversity of body shapes. The training scans include extreme body shapes such as body builders and anorexia nervosa patients. Furthermore, the data capture protocol include athletes such as a ballerina and a yoga expert. Additionally, since SUPR has a full expressive kinematic tree, including a fully articulated hand, jaw and an expressive head, we capture expressive sequences where subjects performed motions communicating emotions and intent. An overview of the full body training scans is shown in Figure 1

External datasets In addition to the scans from the 4D body scanner, we leverage a number of datasets of 3D human body scans. To capture the diversity of human body shape we use the CAESAR [9] and SizeUSA [1] datasets. The CAESAR database contains 1700 male and 2107 female subjects distributed according to the US population in 1990. A limitation of CAESAR's capture protocol is that all women subject were in sports-bra-type top. As a result of the bra type, the CAESAR female chest shape does not reflect the diversity

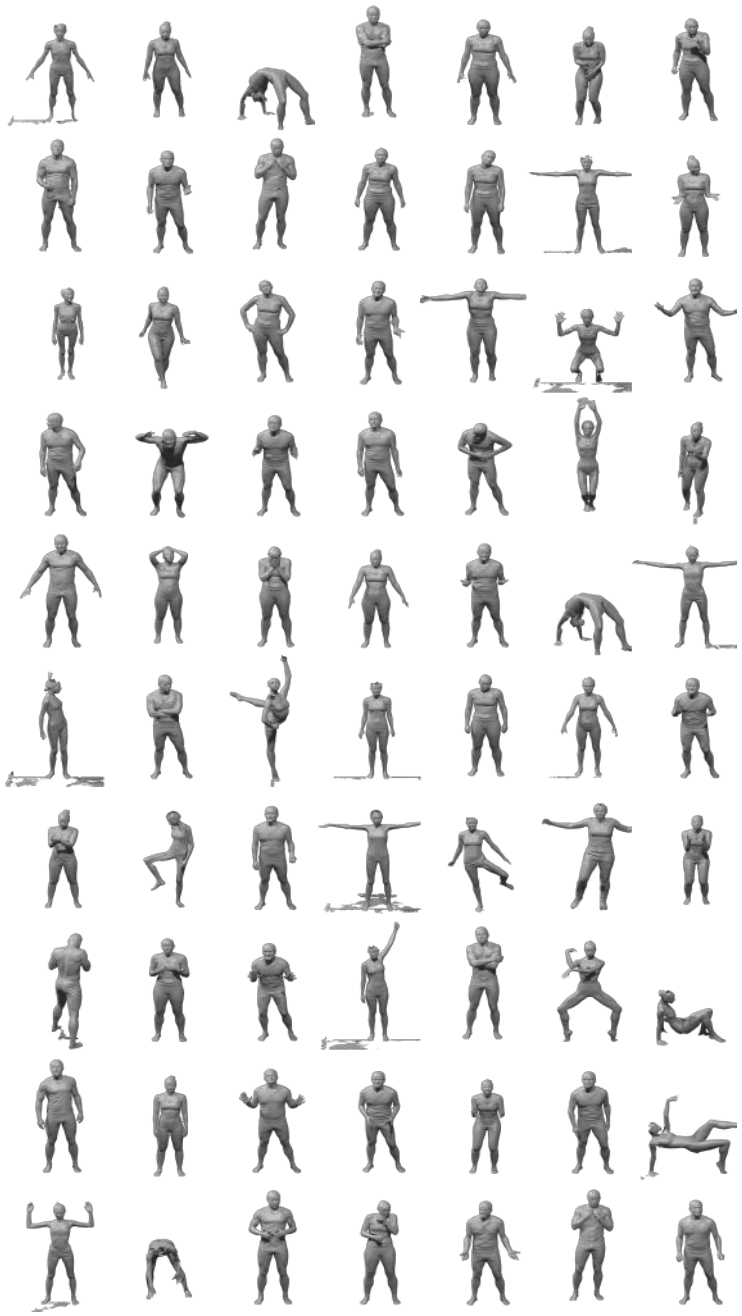


Fig. 1: Overview of the scans captured in the full body scanner. The scans are detailed and high resolution. Note however, the hands and the feet are poorly reconstructed, and the head resolution is not sufficient to capture subtle facial expressions.

of shapes found in real applications. We additionally use the SizeUSA dataset, which contains a richer diversity of body shapes and the female subjects wore a traditional bra. The SizeUSA dataset contains $10,000$ subjects (2845 male and 6436 female).

Despite the 4D scanner high resolution output meshes, the output scans have have poorly reconstructed hands and foot. The foot sole is poorly reconstructed because it is always occluded by the glass platform. The full body scans are not suitable for learning head, hand and foot deformations.

2.2 Head Scans

The human head exhibits a range of highly dynamic deformations. When we refer to the head we mean the face, the back of the head including the scalp and the neck. The human head 3D deformations are due to facial expressions, jaw movement, head movement relative to the neck and body movement relative to the neck (for example when shrugging). We use a dedicated head scanner to complement the full body 4D scanner. The head scanner has a significantly higher number of cameras focused on the head region compared to the body scanner in Section 2.1. The scanning setup enables us to capture the subtle facial expressions. We note, however, that the head scanner has a limited scanning volume making it infeasible to capture the full range of motion of the human head relative to the body.

Similar to the full body scanner, the head scanner is a 4D scanner capturing high-resolution dynamic sequences. The scanner employs 6 pairs of stereo cameras to compute shape and geometry with the assistance of custom speckle projectors. It also includes 6 color cameras and white-light panels to capture texture. The data capturing protocol was designed by experts to capture subtle and extreme facial expressions, full movement of the jaw, in addition to neck movement poses such as looking up, down to the left or right.

2.3 Hand Scans

The reconstructed fingers in full-body scans are typically noisy and poorly reconstructed, as shown in Figure 1. To better capture the hands, we use the data from the MANO hand model [10]. These hand scans are used to learn the pose corrective blend shapes due to finger articulation. A sample of the captured hand scans is shown in Figure 3.

2.4 Foot Scanner

The human foot is a complex structure containing muscles, other soft tissue, and a quarter of the bones in the human skeleton. All existing human body models [2,7,6,8,11,4] use a highly simplified kinematic tree to model the feet with a limited number of joints. Such modes can not model the full range of

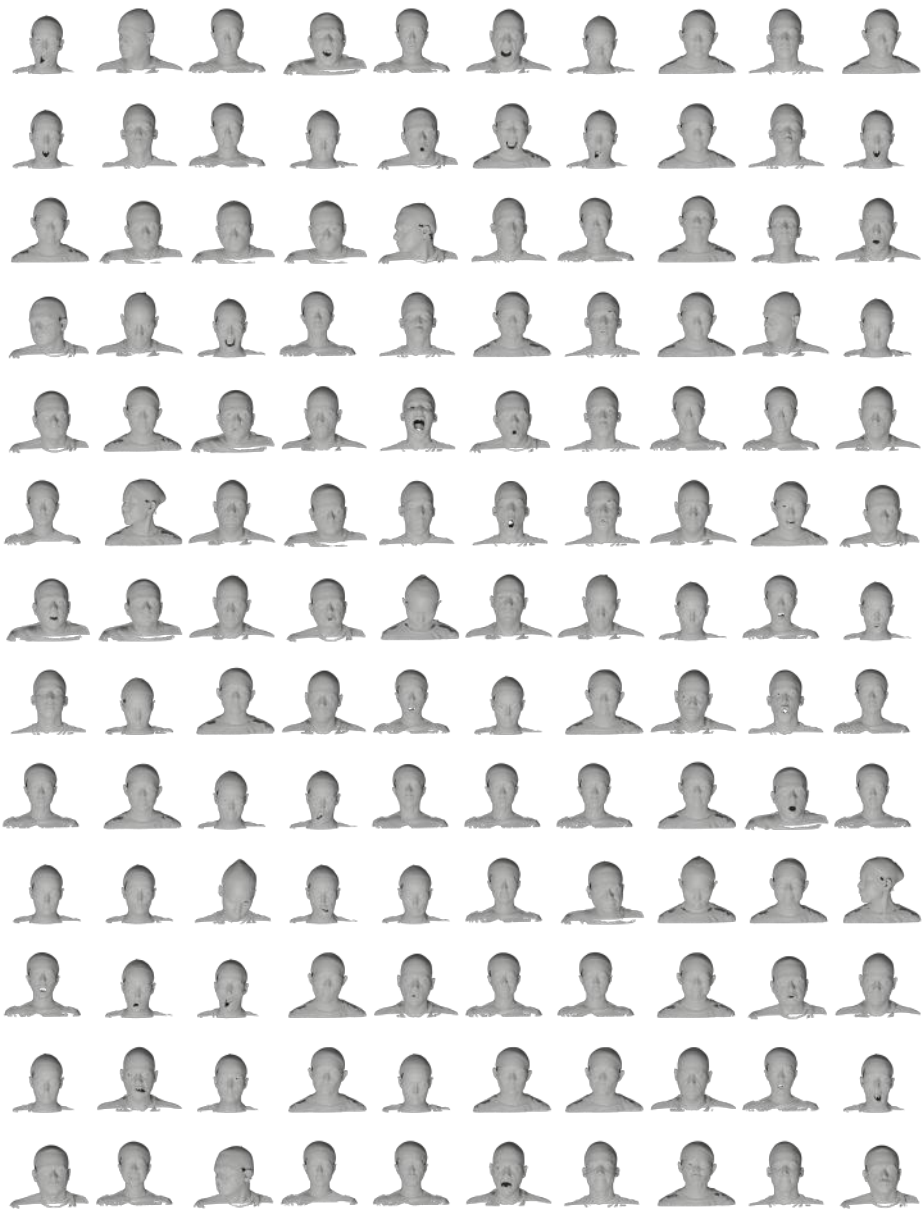


Fig. 2: **Head Scans** A sample of the head scans using in training SUPR.

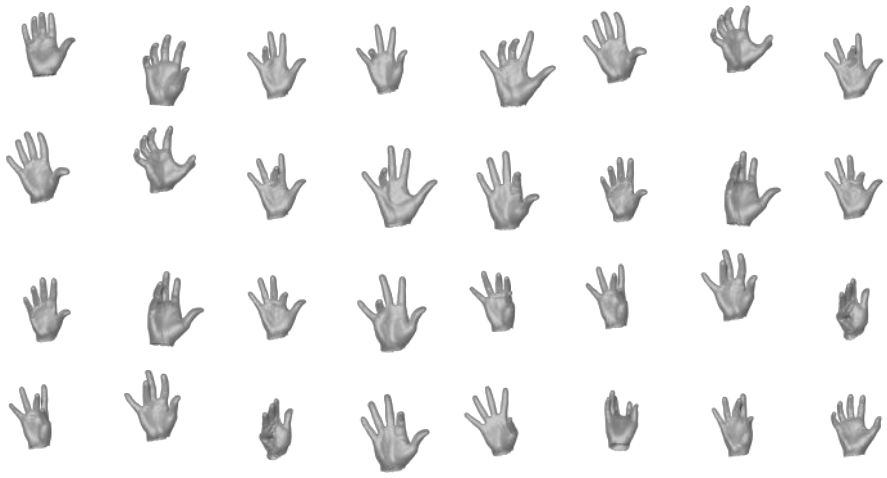


Fig. 3: **Hand Scans:** A sample of the hand scans used to train SUPR.

motion of the bones in the human foot, as highlighted in Figure 3 in the main paper. The kinematic tree of models like SMPL-X is not sufficient to capture the toe articulation. The commonly used pose deformation functions only define body deformations due to pose. As we discuss in the introduction section in the main paper, this is insufficient to capture the foot deformations due to ground contact.

Existing Foot Models. The key reason existing models fail to accurately model foot deformations is because the foot dynamic deformations are hard to capture. The only interesting exception is the work of Boppana et al. [3] discussed in Section 2 in the main paper, which is, to best of our knowledge, the first and only work that attempts to build a model of dynamic foot deformations. Bopanna et al. recognize the limitation of existing scanning solutions to model the human foot deformations and propose the DynaMo system. The Dynamo scanning setup is comprised of a treadmill surrounded by 7 Intel RealSense cameras. A total of 30 subjects were captured walking on a treadmill. A sample reconstructed mesh by the DynaMo system is shown in Figure 6b. We note here that the output scan is noisy and does not capture the foot sole. Bopanna et al. register the DynaMo scans to a high resolution template mesh and learns a PCA space on the registered meshes. The model they propose does not contain toes or a foot sole as shown in Figure 6. This is not surprising given the fidelity of the foot scans captured by the DynaMo system. In contrast to the Bopanna et al. model, SUPR-Foot contains an extensive kinematic tree with 13 joints per foot as shown in Figure 3c in the main paper. SUPR is the first articulated model of the human foot with a pose space that can be driven by bio-mechanics simulations of the

human foot for example, in addition to a shape space to capture the diversity of human foot shape.

2.5 Foot Scanner

SUPR goes beyond existing expressive human body models to model the human foot. To enable capturing the full range of the human foot deformations, we use a custom built scanner dedicated for the foot. The scanner is designed to be mechanically stable to capture dynamic poses such as walking, running or jumping. The output scans are high resolution and can capture the movement of the toes. The scanner floor is a transparent glass platform (which can support subjects up to 150 kg), which enables us to capture the foot sole deformation due to ground contact.

An overview of the foot scanner is shown in Figure 5. The scanner setup features a runway for the subjects to run or walk. In Figure 5b, we show raw scanner images, where the foot is visible from all views, including the foot sole. The scanner uses 10 pairs of stereo cameras, including dedicated cameras capturing the bottom of the foot. The frame rate of the scanner is 10 fps. The output scans contain on average $30,000$ points.

Data Capture Protocol. We capture a total of 30 subjects, 15 female and 15 male subjects with a total of $70,000$ scans. The data capture protocol is designed by experts to explore the space of human foot deformations. The capture protocol is divided into two main parts: 1) Non-Contact sequences 2) Contact Sequences. In the non-contact sequences, the subject foot is not in contact with the glass platform. The data capture protocol for such sequences is designed to explore the full degree of freedom of the toes and the ankle. In contact sequences the subject's foot is partially or in full contact with the glass platform. The contact sequences include motions such as walking/running and jumping. In total We capture 356 dynamic sequences which is the largest training dataset for human scans report in the literature. An overview of the captured scans is shown in Figure 4

Foot Shape Scans. The 30 subjects captured in the dynamic foot scanner do not represent the diversity of human foot shape. Accurate modeling of the human foot shape is crucial for the footwear industry. The feet in the CAESAR and SizeUSA scans, shown in Figure 7a, are noisy, missing, and are not good enough to learn a statistical model. To accurately model the diversity of the human foot scans, we acquired an additional $7,000$ high resolution foot scans from a private vendor. Figure 7 compares the curated high resolution foot scans in comparison to CAESAR and SizeUSA foot scans. In contrast to CAESAR and SizeUSA, the curated dataset of foot scans is significantly less noisy, with on average 10x the resolution of a foot scans from CAESAR/SizeUSA. The high resolution foot scans preserve the 3D geometry of the individual toes. We use this data in learning the the local shape space of SUPR-Foot.

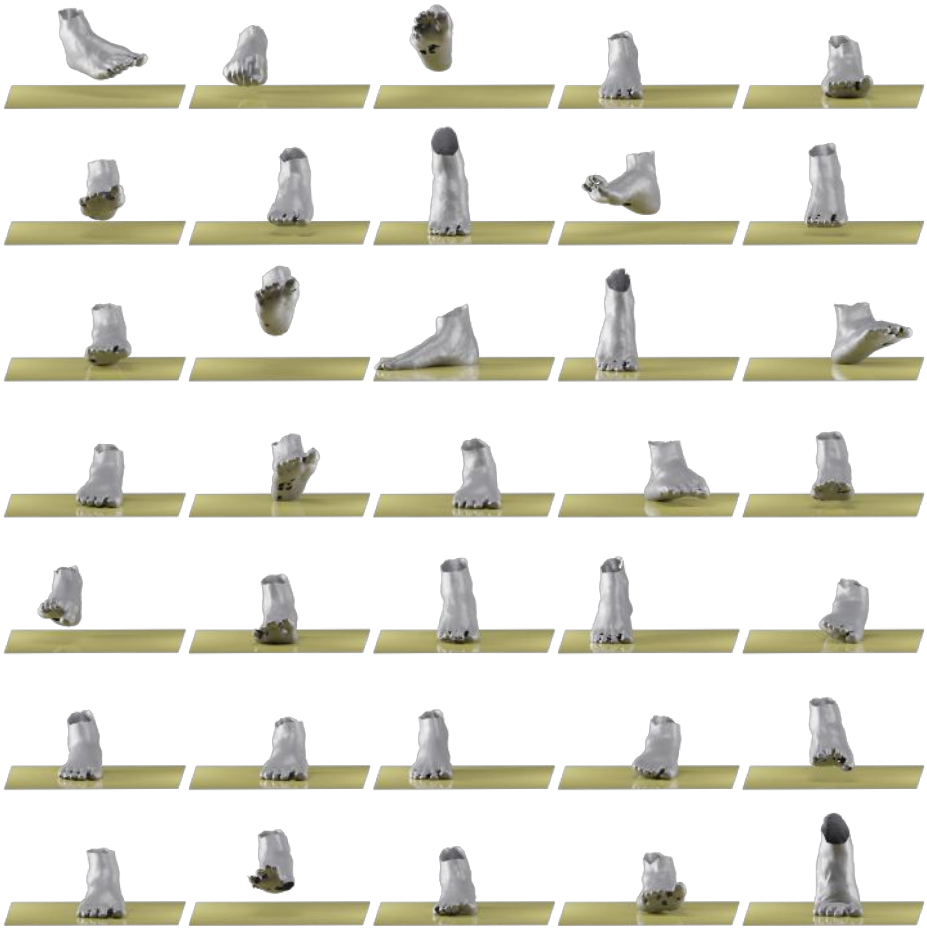
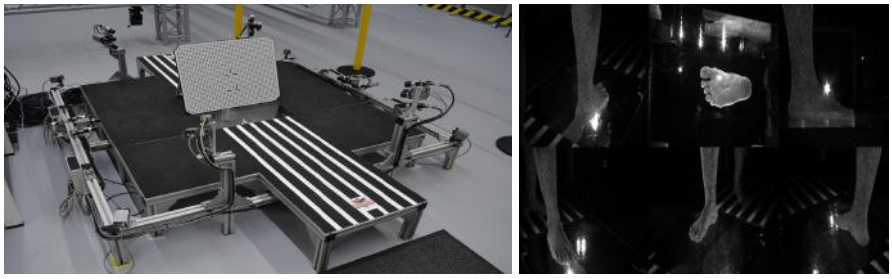
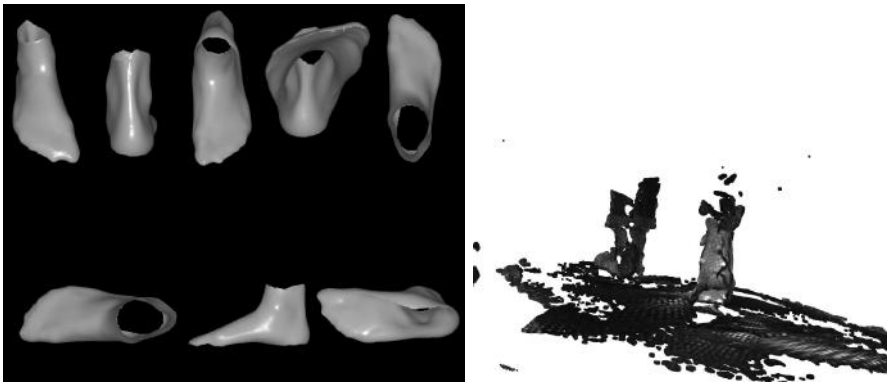


Fig. 4: **Foot Scans:** An Overview of the foot scans. The foot is full reconstructed including the toes and the foot sole.



(a) Scanner setup (b) Raw Scanner Images

Fig. 5: A 3dmD foot scanner using 10 pairs of stereo, including dedicated cameras capturing the bottom of the foot through a transparent glass platform. The scanner features a run way to capture dynamic sequences such as walking.



(a) Foot model. (b) A Scan from the DynaMo system.

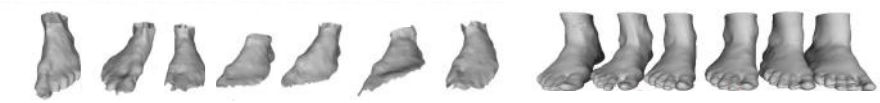
Fig. 6: Bopanna et al. [3] foot model based on Principal Component Analysis of dynamic foot scans. The model does not contains toes or a foot sole.

2.6 Training Scans Summary

SUPR and the separated body parts are trained on a total of 1.2 million scans. In Figure 8 we compare the scale of training datasets used to train body models in the literature. As Figure 8 highlights, the scale of the training data is an order of magnitude larger than the largest training dataset reported in the literature (60K, for the GHUM model). A breakdown of the number of scans captured by each body part is summarized in Table 1.

3 SUPR-Foot Network

In the main paper Section 3.3 we introduce the foot deformation function. A key contribution of our paper is the deformation function relating the foot deforma-



(a) Cropped Feet from full body scans (b) High resolution Foot Scans

Fig. 7: Comparing reconstructed feet from a full body scanner compared to the curated high resolution foot scans. We curate a total of $7,000$ high resolution foot scans. The curate scans have 10x the resolution of foot scans captured in a body scanner and preserve the individual toes geometry.

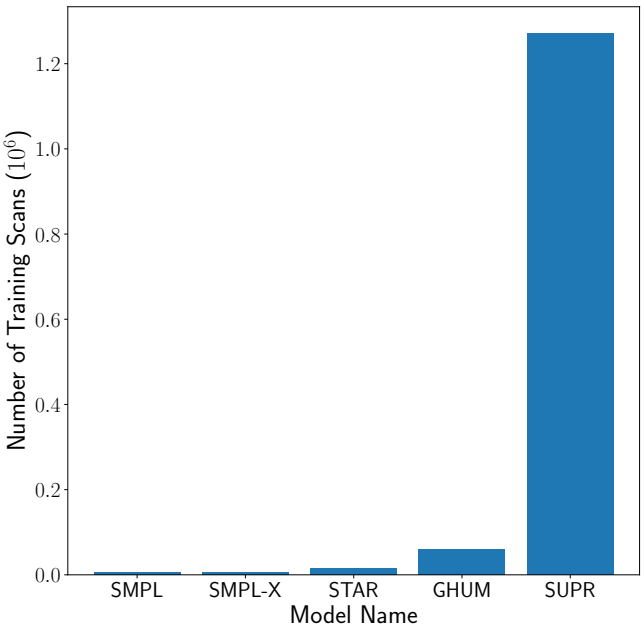


Fig. 8: A comparison between the scale of training datasets for recent human body models. SUPR is trained on a order of magnitude more data compared to the highest number of training scans report in the literature (GHUM 60k).

Body Part	Number of Scans
Body Scans	775,481
Head Scans	421,898
Foot Scans	69,257
Hand Scans	3,531
Total	1,270,167

Table 1: A breakdown of the number of scans for each body part.

tion to the foot pose, shape and contact parameters. In this section we describe the foot deformation network in detail.

Foot Contact The raw foot scans generated by the the foot scanner described in Section 2.5 does not provide per vertex contact labeling, describing whether a vertex in the scan is in contact with the glass platform. To estimate the per-vertex ground contact information we register all the scans to the foot template mesh $\overline{T}_{foot}$. We additionally estimate the ground plane for each dynamic sequence by fitting a plane to the glass platform scan points. A vertex $\vec{v} \in \overline{T}_{foot}$ is labelled in contact with the ground, if it the point-to-plane distance between the vertex and the ground plan is less than a threshold. We allow for a soft threshold when estimating contact since the scans have noise. The threshold used in the main paper is 0.1 mm.

3.1 Foot Deformation Network

The foot deformation network is an encoder-decoder architecture as described in Section 3.3 in the main paper. We train a deformation network for each foot separately. Below we describe the network for the right foot. We first introduce the notation we use:

- B_P: is the linear pose corrective blend shape described Equation 1 in the main paper.
- B_C: is the predicted deformations for the foot related to pose, contact and foot shape.
- $\vec{c}$: is a binary vector of which vertices are in contact with the glass platform.
- $\vec{z}$: is a latent code vector.
- $\vec{\theta}$: is a foot pose parameters.
- $\vec{\beta}$: is a foot shape parameters.
- $\vec{f}$: is a concatenated feature of the pose, shape and contact vector.
- LReLU: leaky rectified linear units with a slope of 0.1 for negative values.
- FC_m: fully connected layer with output dimension m.

Feature Representation The input to the network $\vec{f}$ is a concatentation feature representation of the foot pose, foot shape and contact. The foot pose representation is based on normalized unit quaternion representation defined by:

$$F(\vec{\theta}) = Q(\vec{\theta}) - Q(\vec{\theta^*}) \tag{1}$$

where $Q(.) : \mathbb{R}^3 \rightarrow \mathbb{R}^4$ is a function computing the quaterion representation of the input axis angle rotation, θ^* is the foot in the rest pose. The feature representation in Equation 1 will evaluate to 0 when the foot is in the rest pose. The foot template mesh $\overline{T}_{foot} \in \mathbb{R}^{266\times3}$ is a high dimensional representation to represent the foot shape. We represent the foot shape using the first two principal components which correspond to the foot length and foot volume. We experimented with different number of coefficients, and the first two PC component result in the lowest generalization error on the validation set. The state of the foot contact with the scene is represented using the $\vec{c}$. More formally the input feature to our network:

$$\{F(\vec{\theta}), \vec{\beta}_1, \vec{\beta}_2, \vec{c}\} \xrightarrow{\text{concat}} \vec{f} \in \mathbb{R}^{320}, \tag{2}$$

where $\vec{\beta}_1, \vec{\beta}_2$ are the first two PCA components and the *concat* operator is a standard vector concatenation operator.

3.2 Architecture

The architecture is an encoder-decoder fully-connected network, with non-linear activations based on LReLUs. Encoder:

$$\begin{aligned}\vec{f} \in \mathbb{R}^{320} &\rightarrow FC_{256} \rightarrow \\ &\rightarrow FC_{128} \rightarrow FC_{64} \rightarrow \\ &\rightarrow FC_{32} \rightarrow \vec{z} \in \mathbb{R}^{16}\end{aligned}$$

The dimensionality of the latent code $\vec{z}$ was chosen by grid search. We experimented with dimensionality 64, 32 and 16. A latent code with dimensionality 16 result in the lowest generalization error of the validation set. The decoder is described by:

$$\begin{aligned}\vec{z} \in \mathbb{R}^{16} &\rightarrow FC_{32} \rightarrow \\ &\rightarrow FC_{64} \rightarrow FC_{128} \rightarrow FC_{266} \rightarrow B_C\end{aligned}$$

where B_C is added to the linear blend shape B_P as shown in Equation 6 in the main paper.

4 Training

4.1 SUPR Training

SUPR pose corrective formulation is training is similar to STAR. The key difference is the pose corrective formulation of SUPR is not conditioned on body shape similar to STAR. The additional shape dependant blend shape is not sparse. The key reason we are able to separate SUPR is the fully sparse factorization of the pose blend shapes and the skinning weights as discussed in Section 3 in the main paper.

The SUPR pose corrective blend shapes are trained by minimizing the reconstruction loss between between the model prediction and the federated dataset of groundtruth registration. The SUPR pose blend shape parameter, namely the joint activation $\mathcal{A}$ and the pose corrective blend shapes $\mathcal{P}$ are trained by stochastic gradient descent. Since our data is based on 4D dynamic sequences, we first shuffle the data such that there is no similarity between subsequent frames. We use batches of size 32 to minimize the vertex-to-vertex loss given by:

$$\mathcal{L}_D = \frac{1}{B}\sum_{i=1}^{32} ||M(\theta^i) - R^i||_2. \tag{3}$$

where R^i is the ith groundtruth registration in the batch. Similar to STAR, we use an $L1$ penalty on the output of the joint activation $\mathcal{A}$,

$$\mathcal{L}_A = \lambda_c || \sum_{i=1}^{K-1} \phi_j(A_j)||, \tag{4}$$

where λ_c is a scalar constant. The full objective for the pose space is defined by Equation:

$$\mathcal{L} = \mathcal{L}_A + \mathcal{L}_D, \tag{5}$$

where Equation 5 is minimized with respect to the pose corrective regression weights $\mathbf{K_{1:80}}$, activation weights $\mathbf{A_{1:80}}$. We use a batch size $B = 32$ and the ADAM.

4.2 Body Part Training

Given the trained blend shapes, the body part pose space is separated as discussed in Section 3 in the main paper. We further train a local shape space for each of the separated body part models. The CAESAR head, hand of the CAESAR registrations are used to train a local shape space for SUPR-Head and SUPR-Head. The local shape space for the foot is trained on the curate high resolution foot scans, as the foot in the CAESAR scans were noisy. The percentage of explained variance as the number of shape components for each body part is shown in Figure 9.

4.3 Deformation Network Training

Given the learned linear blend shapes trained in Section 4.1, and the local shape space for the foot, we train the deformation network for the foot deformation described in Section 3 in the main paper. For training the deformation network we use both contact and non-contact foot registrations. The network is trained by minimizing the $L1$ loss between the model and the foot registrations:

$$\mathcal{L} = ||M(\theta^i, \vec{c}^i, \vec{\beta}^i) - R^i||. \tag{6}$$

The training loss is minized using stochastic gradient descent, where we used ADAM with batch size 32.

5 SUPR Ablation

5.1 STAR Evaluation

SUPR pose corrective blend shape formulation is based on the STAR pose corrective blend shape formulation as discussed in Section 3 in the main paper. For completeness we further evaluate SUPR against STAR. We note however that SUPR is expressive, with 1.5x more vertices and 3x more joints compared to STAR. We use the 3DBodyTex dataset and register the scans to the STAR template. A human expert validated all registrations. Similar to the evaluation Section 4.1, 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 both models are the pose parameters $\vec{\theta}$ and the shape parameters $\vec{\beta}$. We report the model generalization error in Figure 10.

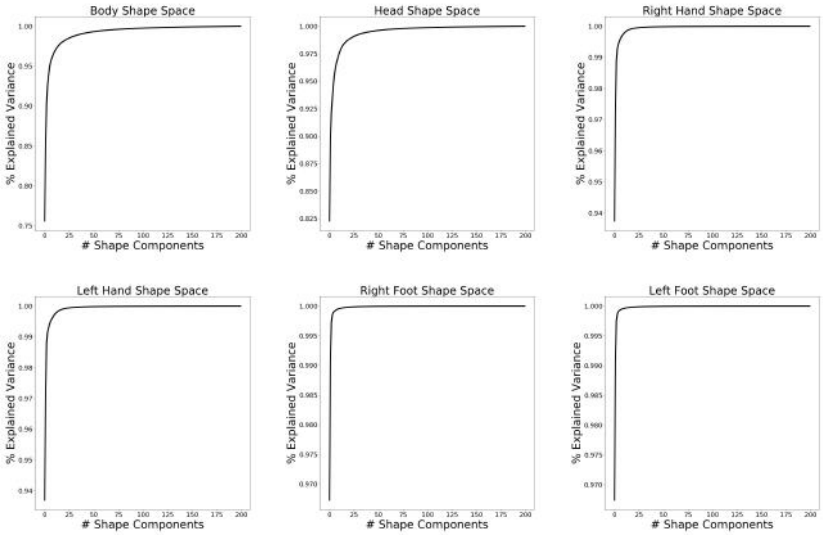
5.2 GHUM Body Parts Evaluation

Head Evaluation In Section 4.2 in the main paper we evaluate SUPR-Head against GHUM-Head. A sample qualitative comparison fits are shown in Fig. 12. Similar to FLAME, GHUM-Head has significant error around the neck region.

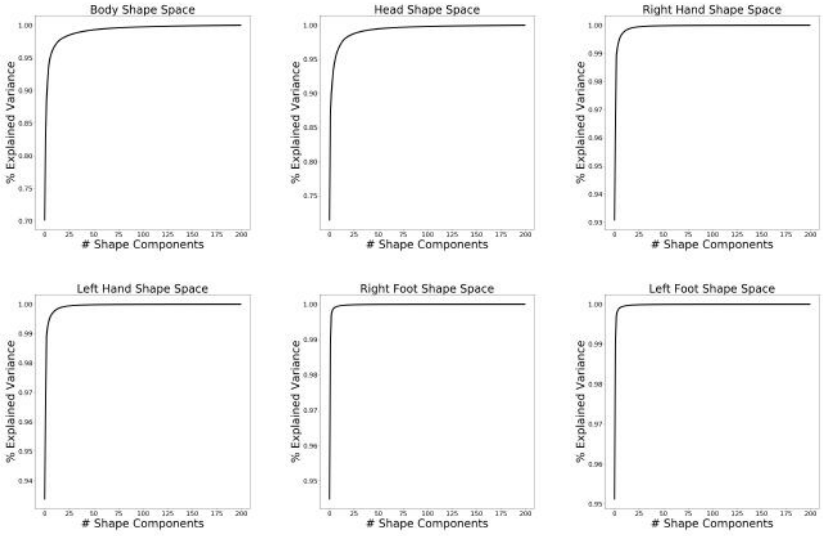
Hand Evaluation In Section 4.3 in the main paper we evaluate SUPR-Hand, against GHUM-Hand. A sample of the GHUM-Hand fits are shown in Figure 11. GHUM-Hand consistently has a high error in the wrist region compared to the the fingers.

6 SUPR-Foot Ablation

In the main paper Section 4.4 we evaluate SUPR-Foot against SMPL-X-Foot on a held out test set of contact and non-contact foot scans. We further break down the evaluation in Figure 13. We report the model mean absolute error as a function of the number of shape components used on non-contact frames in Figure 13a and contact frames in Figure 13b.



(a) Male Shape Space



(b) Female Shape Space

Fig. 9: Percentage of explained variance as a function of the number of shape components for SUPR and the separated body part models.

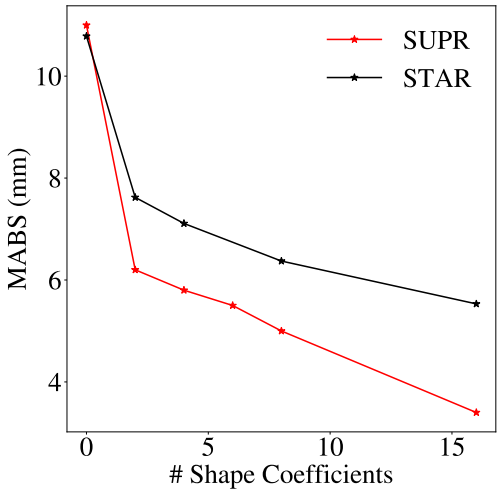


Fig. 10: Evaluation of SUPR against STAR on the 3DBodyTex dataset.

Fig. 11: **GHUM-Hand Evaluation :** Evaluating GHUM-Hand on the MANO test set. The top row are raw scans from the MANO test set, the second row is GHUM-Hand model fits with 10 shape components, while the bottom row is the corresponding error heatmap.

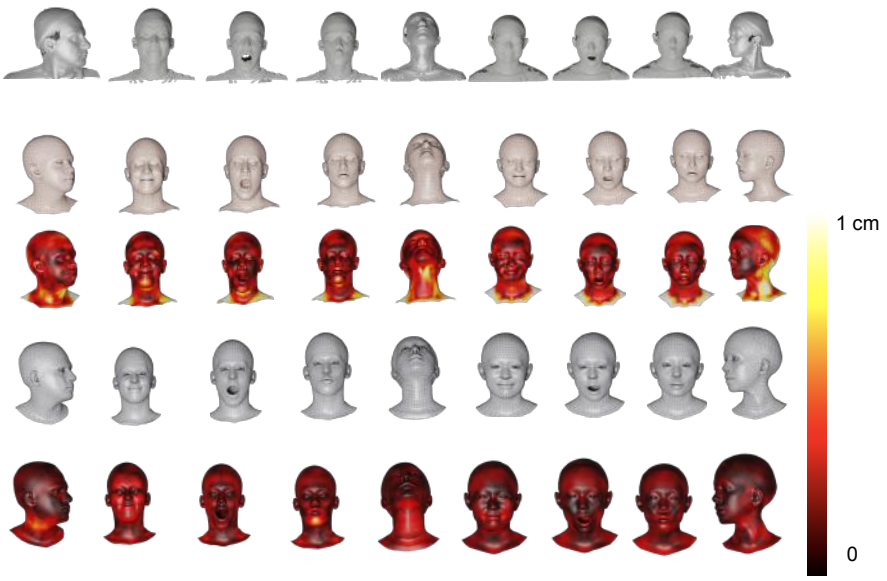
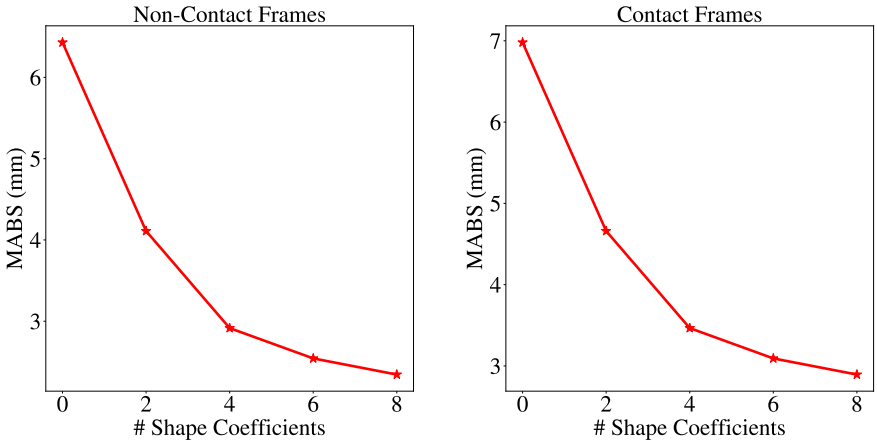


Fig. 12: **Head Evaluation:** Evaluating GHUM-Head and SUPR on a test containing head movement relative to the neck, jaw movement and extreme facial expressions. Both GHUM-Head and SUPR are fit with 16 shape and epxression components. The top raw correspond to head raw scans, second row correspond to GHUM-Head fits, third row correspond to the heatmap, fourth row correspond to SUPR-Head fits and the fifth row are the corresponding error heat maps.

Table 2: Foot Deformation Ablation Study. SUPR-Foot $_{lbs}$ corresponding to model with linear blend skinning, no additive correctives used. SUPR-Foot $_{lbs+l}$ correspond to lbs in addition to the linear correctives, SUPR-Foot $_{lbs+l+f(\theta)}$ is adding the non-linear deformation where the network is condition on pose only, SUPR-Foot $_{lbs+l+f(\theta,\vec{\beta})}$ where the network is conditioned on pose and shape information, while SUPR-Foot is the full model.

Model	Non-Contact v2v (mm) ↓	With-Contact v2v (mm) ↓
SUPR-Foot $_{lbs}$	5.235 ±0.126	6.691 ±1.369
SUPR-Foot $_{lbs+l}$	4.587 ±0.589	5.364 ±1.279
SUPR-Foot $_{lbs+l+f(\theta)}$	2.982 ±0.859	4.129 ±1.883
SUPR-Foot $_{lbs+l+f(\theta,\beta)}$	2.910 ±0.728	3.934 ±1.819
SUPR-Foot (ours)	2.753 ±0.821	3.122 ±1.462



(a) Non-Contact (b) Contact

Fig. 13: Evaluating SUPR-Foot on frames where the foot was not in contact with the glass platform shown in Figure 13a, and frames where the foot was partially or fully in contact with the glass platform in Figure 13b.

Deformation function: A key contribution of our work is introducing a novel deformation function which relates the foot deformations to the foot pose, shape and ground contact. We illustrate the influence of each term on the model generalization by ablating the foot deformation network described in Section 3 in the main paper. We retain variations of the deformation network from scratch and refit each model to the test set. We report the model $v2v$ error in Table. 2. The result clearly show the vertex to vertex error decreasing on the held out test set when adding each term in the foot deformation function across both the contact and non-contact frames.

7 Constrained SUPR

The SUPR kinematic tree introduced in Section 3 is based on spherical joints. Each spherical joint j is parameterized by $\vec{\theta_j} \in \mathbb{R}^3$. The spherical joints allow redundant degrees of freedom for some body parts such as the fingers. For the fingers, for example, the axes of rotation are not bone-aligned. In order to simply bend a finger we have to control 3 axis-angle rotations. This is problematic to use by animators and for architectures that regress hand pose parameters from images. In this section we describe a constrained version of SUPR that uses hinge/double hinge joints in contrast to spherical joints.

Constrained Kinematic Tree. The kinematic tree of the constrained version of SUPR (shown in Figure 14) uses hinge and double hinge joints. A hinge joint

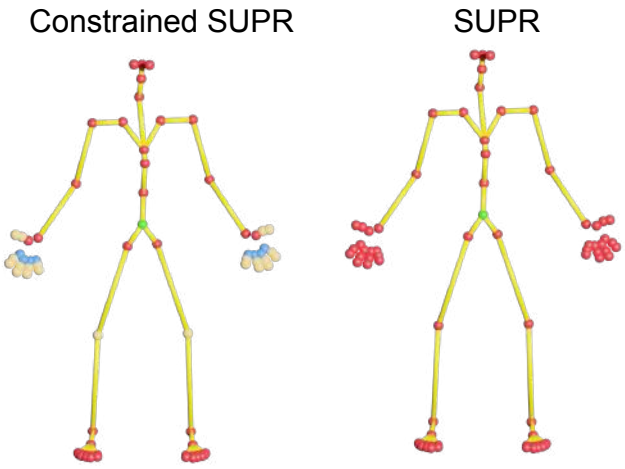


Fig. 14: **Constrained SUPR Kinematic Tree:** SUPR is based on spherical joints which allow redundant degrees of freedom for body parts such as the fingers. The constrained SUPR kinematic tree contains a mixture of joints: Spherical joints (shown in red), Hinge Joints (shown in beige) and double hinge joints (shown in blue).

is fully parameterized by an axis of rotation $\vec{a} \in \mathbb{R}^3$ and a pose parameter $\vec{\theta} \in \mathbb{R}$. A double hinge joint is defined by two axes of rotation and pose parameters $\vec{\theta} \in \mathbb{R}^2$. The axes of rotation for the hinge and double hinge joints are orthogonal to the bone. Therefore, to simply bend a finger in SUPR requires only controlling or regressing one or two scalars. This compact representation is convenient for artists, regression tasks and is more anatomically plausible.

Specifically, this version of SUPR is defined by Eq. 7:

$$M(\vec{\theta}, \vec{\beta}, \vec{\psi}) = W(T_p(\vec{\theta}, \vec{\beta}, \vec{\psi}), J(\vec{\beta}), AX, \vec{\theta}, \mathcal{W}), \tag{7}$$

where $AX \in \mathbb{R}^{30\times 3}$ is the axis of rotation matrix for the hinge and double hinge joints. The key difference between Equation 1 in the main paper and Equation 7 is the bone transformation rotation matrix. The rotation matrix for a hinge joint is a constrained rotation matrix, which only allows a single degree of freedom with respect to the axis of rotation $\vec{a}$. A constrained rotation matrix is defined by:

$$\begin{bmatrix} a_x^2 + c_\theta(1-a_x^2) & a_x a_y(1-c_\theta) + a_z s_\theta & a_x a_z(1-c_\theta) - a_y s_\theta \\ a_x a_y(1-c_\theta) - a_z s_\theta & a_y^2 + c_\theta(1-a_y^2) & a_y a_z(1-c_\theta) + a_z s_\theta \\ a_x a_z(1-c_\theta) + a_y s_\theta & a_y a_z(1-c_\theta) - a_x s_\theta & a_z^2 + c_\theta(1-a_z^2) \end{bmatrix}$$

where a_x,a_y,a_z are the x , y and z coordinates of the axis of rotation $\vec{a}$. c_θ and s_θ are $\cos(\theta)$ and $\sin(\theta)$ correspondingly.

The constrained version of SUPR only limits the bones' degrees of freedom, by constraining the rotation matrices of the corresponding joints. Therefore,

this is an additional functionality, which we will release with SUPR, that can be enabled or disabled by a user of regression model.

8 Model Comparison

SUPR is trained on a federated dataset of head, body and head registrations. As a consequence of the sparse factorization of the pose space, we are able to separate the model into body part models. A comparison between SUPR and existing body models is shown in Figure 15.

Model Name	Sparse Pose Deformations	Federated Training Data	Articulated Hands	Expressive Head	Part Based	Game Engine Compatiability	Publicly Available
SCAPE	✗	✗	✗	✗	✗	✗	✔
Stitched Puppet	✔	✗	✗	✗	✔	✗	✔
SMPL	✗	✗	✗	✗	✗	✔	✔
SMPL-H	✗	✔	✔	✗	✗	✔	✔
Frank	-	✗	✔	✔	✔	✔	✔
SMPL-X	✗	✔	✔	✔	✗	✔	✔
GHUM	✗	✔	✔	✔	✗	✗	✔
STAR	✔	✗	✗	✗	✗	✔	✔
BLSM	✗	✗	✗	✗	✗	✔	✗
SUPR	✔	✔	✔	✔	✔	✔	✔

Fig. 15: A comparison between SUPR and existing body models.

8.1 SUPR

Model	# Pose	# Joints	# Blendshapes
SUPR	225	75	296
SMPL-X [8]	165	55	486
GHUM [11]	124	63	-

Table 3: **Body Models Comparison:** Comparing existing expressive human body models according to the number of pose parameters, number of joints and number of pose corrective blendshapes.

SUPR is a compact model that is compatible with the existing gaming and animation industry standards. The number of parameters of SUPR compared to existing expressive human body models is summarised in Table 3.

Comparison with SMPL-X: SUPR has 30% fewer pose-corrective blendshapes, despite having significantly more joints compared to SMPL-X. This is because of the Quaternion-based representation, which is significantly more compact compared to the Rodrigues representation used by SMPL-X. However, despite SUPR's compactness, it uniformly generalizes better than SMPL-X. The shape space of SMPL-X is trained on the CAESAR dataset [9], while SUPR is trained on $15,000$ registrations from both CAESAR and SizeUSA [1]. The SizeUSA dataset contains a larger diversity of body shapes and, in addition, the female subjects wore a traditional bra, whereas, in the CAESAR dataset, the female subjects wore a sports bra. The pose space of SMPL-X is trained on 2000 full body registrations. In contrast, SUPR's pose space is trained on a federated dataset of 1.2 million registrations of head, hand, body and feet registrations.

SMPL-X's pose blendshape formulation is based on SMPL. As a result, SMPL-X suffers from the same drawbacks of SMPL, namely SMPL-X also learns false long range spurious correlations; e.g. bending one elbow results in a bulge in the other elbow.

Comparison with GHUM: The GHUM model [11] pose space deformation function (PSD) is modeled by a neural network, which is not compatible with the gaming and animation industry standards. SUPR's learned blendshapes are linearly related to the model pose parameters, and hence the formulation is full compatible with the gaming and animation industry standards. While both SUPR and GHUM are trained on a federated dataset, and the GHUM authors propose a separated suite of models (GHUM-Head and GHUM-Hand), there are key important differences. The GHUM shape space is trained only on the CAESAR data (5K subjects),while SUPR shape space is trained on both CAESAR and SizeUSA, for a combined total of 15K registrations. On the other hand, the pose space of GHUM is trained on a dataset of 60K head, hand and body registrations, while the SUPR pose space is trained on 1.2 million body, head, hand and feet registrations. SUPR is the first to train on dedicated foot registrations. This is crucial for modeling realistic foot deformations due to movement of the ankle or curling of the toes, since the feet are consistently poorly reconstructed in full-body scans.

The GHUM PSD formulation is a dense non-linear formulation, where all the joints are related to all the vertices using a VAE [5]. As a result the body pose-space formulation of GHUM can not be separated into compact body parts. To define separate body part models, the GHUM authors segment the mesh and re-train the PSD function of the separated parts. The proposed head and hand models for GHUM fail to capture the full degrees of freedom of th head. SUPR and the separated head/hand models are jointly trained once. In contrast to GHUM, the SUPR pose-space formulation is strictly sparse, where each joint only influences a sparse set of the model vertices. As a result, SUPR can be separated into a suite of compact models. The learned kinematic tree of SUPR-Head has significantly more joints (neck and shoulders).

All prior expressive human body models ignore the human foot. The kinematic tree contains an additional 24 joints for modeling the full range of motion of the ankle and the toes.

8.2 SUPR-Head

The SUPR-Head has a pose, shape and expression space. We train 3 head models: female, male and a gender neutral model. The pose blendshape function is a subset of the learned SUPR pose corrective blendshapes, which are also sparse and spatially local. A comparison between and existing full head models is shown in Table 4.

Model	# Pose	# Joints	# Blendshapes
SUPR-Head	29	10	40
FLAME [8]	12	4	36
GHUM-Head [11]	23	10	-

Table 4: **Head Models Comparison:** Comparing existing head models models according to the number of pose parameters, number of joints and number of pose corrective blendshapes.

8.3 SUPR-Hand

We train a single gender-neutral SUPR-Hand model. SUPR-Hand has a pose and shape space. A comparison between SUPR-Hand and existing hand models is shown in Table 5. In comparison to MANO, SUPR-Hand has an additional wrist joint, which is necessary to model the hand deformations as a result of the wrist movement. A comparison between SUPR-Hand and existing hand models is shown in Table 5.

Model	# Pose	# Joints	# Blendshapes
SUPR-Hand	90	30	120
MANO [8]	90	30	270
GHUM-Hand [11]	18	36	-

Table 5: **Hand Models Comparison:** Comparing existing hand models according to the number of pose parameters, number of joints and number of pose corrective blendshapes.

8.4 SUPR-Foot

We train a male, female and neutral models for SUPR-Foot. SUPR-Foot is the first publicly-available articulated model of the human foot. We propose a novel deformation function that relates the foot deformation to the foot pose, foot shape and foot contact. SUPR-Foot shape space is trained on $7,000$ high-resolution foot scans that capture the diversity of the human foot shape variation. The pose space is trained on $57,231$ high-resolution scans that capture the foot sole deformations due to ground contact.

9 Limitation

A limitation of our method is that model training relies on registering a template mesh to the scans. While registration is automatic, the resulting data needs curation by an expert to detect any failures or artifacts. Registering 1.2M hand, head, body, and foot scans is time-consuming and labor-intensive. Registration remains a bottleneck for training body models on large datasets.

A key limitation when evaluating SMPL-X and GHUM against SUPR is that they are both trained on propriety data that is not publicly available to the research community. Additionally, the training code of SMPL-X and GHUM is also not publicly available for research purposes. Therefore, direct comparisons between the body models on the same data is difficult. Nevertheless, we are the first to evaluate all expressive body models on a publicly available test benchmark.

References

1. SizeUSA dataset. https://www.tc2.com/size-usa.html (2017) 2, 21
2. Anguelov, D., Srinivasan, P., Koller, D., Thrun, S., Rodgers, J., Davis, J.: SCAPE: Shape Completion and Animation of PEople. ACM TOG **24**(3), 408–416 (2005) 4
3. Boppana, A., Anderson, A.P.: Dynamic foot morphology explained through 4d scanning and shape modeling. Journal of Biomechanics **122**, 110465 (2021) 6, 9
4. Joo, H., Simon, T., Sheikh, Y.: Total capture: A 3D deformation model for tracking faces, hands, and bodies. In: CVPR. pp. 8320–8329 (2018) 4
5. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013) 21
6. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. ACM Transactions on Graphics, (Proc. SIGGRAPH Asia) **34**(6), 248:1–248:16 (Oct 2015) 4
7. Osman, A.A.A., Bolkart, T., Black, M.J.: STAR: Sparse trained articulated human body regressor. In: ECCV. pp. 598–613 (2020) 4
8. 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 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2019) 4, 20, 22

9. Robinette, K.M., Blackwell, S., Daanen, H., Boehmer, M., Fleming, S., Brill, T., Hoeferlin, D., Burnsides, D.: Civilian American and European Surface Anthropometry Resource (CAESAR) final report. Tech. Rep. AFRL-HE-WP-TR-2002-0169, US Air Force Research Laboratory (2002) 2, 21
10. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: Modeling and capturing hands and bodies together. ACM Transactions on Graphics, (Proc. SIGGRAPH Asia) **36**(6), 245:1–245:17 (Nov 2017), http://doi.acm.org/10.1145/3130800.3130883 4
11. Xu, H., Bazavan, E.G., Zanfir, A., Freeman, W.T., Sukthankar, R., Sminchisescu, C.: GHUM & GHUML: Generative 3D human shape and articulated pose models. In: CVPR. pp. 6184–6193 (2020) 4, 20, 21, 22