STAR: Sparse Trained Articulated Human Body Regressor

Ahmed A. A. Osman, Timo Bolkart, and Michael J. Black

Max Planck Institute for Intelligent Systems, Tübingen, Germany
{aosman,tbolkart,black}@tuebingen.mpg.de

Abstract. The SMPL body model is widely used for the estimation, synthesis, and analysis of 3D human pose and shape. While popular, we show that SMPL has several limitations and introduce STAR, which is quantitatively and qualitatively superior to SMPL. First, SMPL has a huge number of parameters resulting from its use of global blend shapes. These dense pose-corrective offsets relate every vertex on the mesh to all the joints in the kinematic tree, capturing spurious long-range correlations. To address this, we define per-joint pose correctives and learn the subset of mesh vertices that are influenced by each joint movement. This sparse formulation results in more realistic deformations and significantly reduces the number of model parameters to 20% of SMPL. When trained on the same data as SMPL, STAR generalizes better despite having many fewer parameters. Second, SMPL factors pose-dependent deformations from body shape while, in reality, people with different shapes deform differently. Consequently, we learn shape-dependent pose-corrective blend shapes that depend on both body pose and BMI. Third, we show that the shape space of SMPL is not rich enough to capture the variation in the human population. We address this by training STAR with an additional 10,000 scans of male and female subjects, and show that this results in better model generalization. STAR is compact, generalizes better to new bodies and is a drop-in replacement for SMPL. STAR is publicly available for research purposes at http://star.is.tue.mpg.de.

1 Introduction

Human body models are widely used to reason about 3D body pose and shape in images and videos. While several models have been proposed [5–10, 30, 35, 37], SMPL [21] is currently the most widely use in academia and industry. SMPL is trained from thousands of 3D scans of people and captures the statistics of human body shape and pose. Key to SMPL's success is its compact and intuitive parametrization, decomposing the 3D body into pose parameters $\boldsymbol{\theta} \in \mathbb{R}^{72}$ corresponding to axis angle rotations of 24 joints and shape $\boldsymbol{\beta} \in \mathbb{R}^{10}$ capturing subject identity (the number of shape parameters can be as high as 300 but most research uses only 10). This makes it useful to reason about 3D human body pose and shape given sparse measurements, such as IMU accelerations [11, 12, 26], sparse mocap markers [22, 25] or 2D key points in images and videos [3, 14–16, 28, 33, 36, 38].

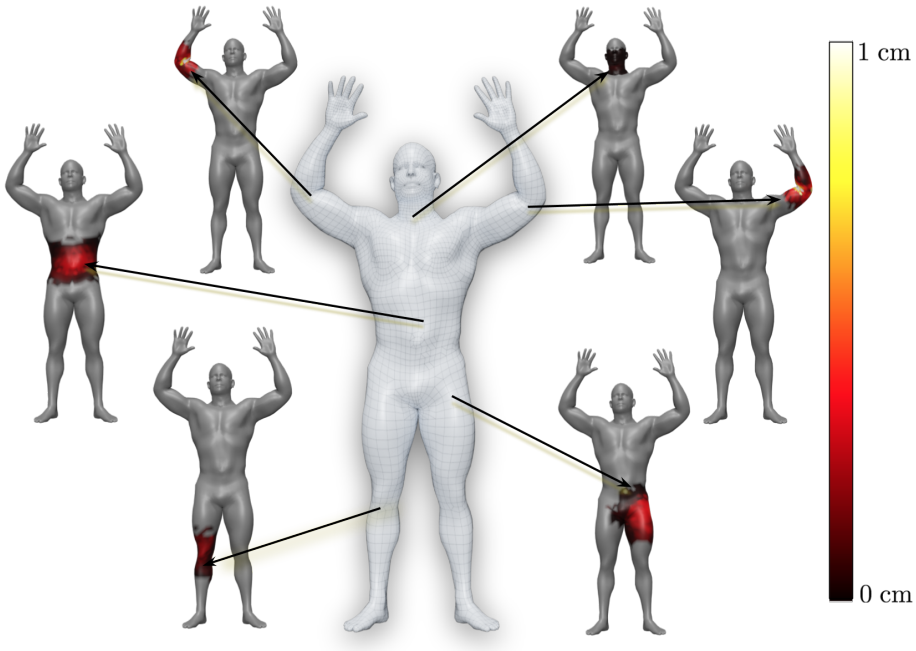
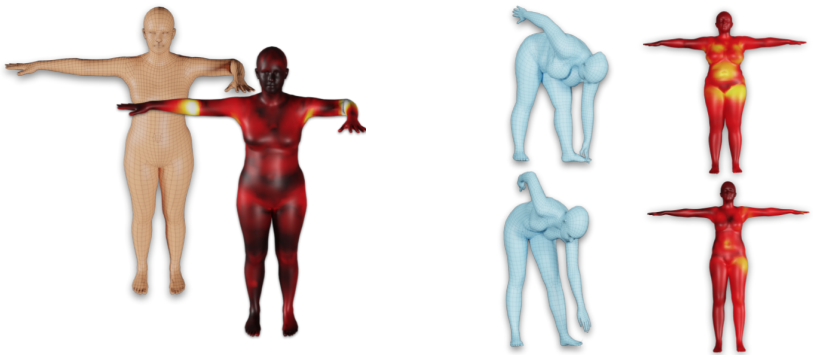


Fig. 1: **Sparse Local Pose Correctives:** STAR factorizes pose-dependent deformation into a set of sparse and spatially local pose-corrective blendshape functions, where each joint influences only a sparse subset of mesh vertices. The white mesh is STAR fit to a 3D scan of a professional body builder. The arrows point to joints in the STAR kinematic tree and the corresponding predicted corrective offset for the joint. The heat map encodes the magnitude of the corrective offsets. The joints have no influence on the gray mesh vertices.

While SMPL is widely used it suffers from several drawbacks. SMPL augments traditional linear blend skinning (LBS) with pose-dependent corrective offsets that are learned from 3D scans. Specifically, SMPL uses a pose-corrective blendshape function $\mathcal{P}(\boldsymbol{\theta}) : \mathbb{R}^{|\boldsymbol{\theta}|} \rightarrow \mathbb{R}^{3N}$, where N is the number of mesh vertices. The function $\mathcal{P}$ predicts corrective offsets for every mesh vertex such that, when the model is posed, the output mesh looks realistic. The function $\mathcal{P}$ can be viewed as a fully connected layer (FC), that relates the corrective offsets of every mesh vertex to the elements of the part rotation matrices of all the body joints. This dense blendshape formulation has several drawbacks. First, it significantly inflates the number of model parameters to > 4.2 million, making SMPL prone to overfitting during training. Even with numerous regularization terms, the model learns spurious correlations in the training set, as shown in Figure 2a; moving one elbow causes a bulge in the other elbow.

This is problematic for graphics, model fitting, and deep learning. The dense formulation causes dense spurious gradients to be propagated through the model.



(a) (b)

Fig. 2: **SMPL Limitations**: Examples of some SMPL limitations. Heat maps illustrate the magnitude of the pose-corrective offsets. Fig. 2a highlights the spurious long-range correlations learned by the SMPL pose corrective blend shapes. Bending one elbow results in a visible bulge in the other elbow. Fig. 2b shows two subjects registrations (show in blue) with two different body shapes (High BMI) and (Low BMI). While both are in the same pose, the corrective offsets are different since body deformation are influenced by both body pose and body shape. The SMPL pose corrective offsets are the same regardless of body shape.

A loss on the mesh surface back propagates spurious gradients to geodesically distant joints. The existing formulation of the pose corrective blend shapes limits the model compactness and visual realism.

To address this, we create a new compact human body model, called **STAR** (Sparse Trained Articulated Regressor), that is more accurate than SMPL yet has sparse and spatially local blend shapes, such that a joint only influences a sparse set of vertices that are geodesically close to it. The original SMPL paper acknowledges the problem and proposes a model called SMPL-LBS-Sparse that restricts the pose corrective blend shapes such that a vertex is only influenced by joints with the highest skinning weights. SMPL-LBS-Sparse, however, is less accurate than SMPL.

Our key insight is that the influence of a body joint should be inferred from the training data. The main challenge is formalizing a model and training objective such that we learn meaningful joint support regions that are sparse and spatially local as shown in Figure 1. To this end we formalize a differentiable thresholding function based on the Rectified Linear Unit operator, **ReLU**, that learns to predict 0 activations for irrelevant vertices in the model. The output activations are used to mask the output of the joint blendshape regressor to only influence vertices with non-zero activations. This results in a sparse model of pose-dependent deformation.

We go further in improving the model compactness. SMPL uses a Rodrigues representation of the joints angles and has a separate pose-corrective regressor

for each element of the matrix, resulting in 9 regressors per joint. We switch to a quaternion representation with only 4 numbers per joint, with no loss in performance. This, in combination with the sparsity, means that STAR has 20% of the parameters of SMPL. We evaluate STAR by training it on different datasets. When we train STAR on the same data as SMPL, we find that it is more accurate on held-out test data. Note that the use of quaternions is an internal representation change from SMPL and transparent to users who can continue to use the SMPL pose parameters.

SMPL disentangles shape due to identity from shape due to pose. This is a strength because it results in a simple model with additive shape functions. It is also a weakness, however, because it cannot capture correlations between body shape and how soft tissue deforms with pose. To address this we extend the existing pose corrective formulation by regressing the correctives using both body pose $\boldsymbol{\theta}$ and body shape $\boldsymbol{\beta}$. Here we use the second principal component of the of the body shape space, which correlates highly with Body Mass Index (BMI). This change results in more realistic pose-based deformations.

SMPL is used in many fields such as apparel and healthcare because it captures the statistics of human body shape. The SMPL shape space was trained using the CAESAR database, which contains 1700 male and 2107 female subjects. CAESAR bodies, however, are distributed according to the US population in 1990 [32] and do not reflect global body shape statistics today. Additionally, CAESAR's capture protocol dressed all women in the same sports-bra-type top, resulting in a female chest shape that does not reflect the diversity of shapes found in real applications. We show that SMPL trained on CAESAR is not able to capture the variation in the more recent, and more diverse, SizeUSA dataset of 10,000 subjects (2845 male and 6436 female) [2], and vice versa. To address these problems, we train STAR from the combination of CAESAR and SizeUSA scans and show that the complementary information contained in both datasets enables STAR to generalize better to unseen body shapes.

We summarize our contributions by organizing them around impact areas where SMPL is currently used:

1. **Computer vision:** We propose a compact model that is 80% smaller than SMPL. We achieve compactness in two ways: First, we formalize sparse corrective blend shapes and learn the set of vertices influenced by each joint. Second, we use quaternion features for offset regression. While STAR is more compact than SMPL, it generalizes better on held-out test data.
2. **Graphics:** Non-local deformations make animation difficult because changing the pose of one body part affects other parts. Our local model fixes this problem with SMPL.
3. **Health:** Realistic avatars are important in health research. We increase realism by conditioning the pose corrective blend shapes on body shape. Bodies with different BMI produce different pose corrective blend shapes.
4. **Clothing Industry:** Accurate body shape matters for clothing. We use the largest training set to date to learn body shape and show that previous models were insufficient to capture the diversity of human shape.

The model is a drop-in replacement for SMPL, with the same pose and shape parametrization. We make the model with a 300 principal component shape space publicly available for research purposes at http://star.is.tue.mpg.de.

2 Related Work

There is a long literature on 3D modelling of the human body, either manually or using data-driven methods. We review the most related literature here with a focus on methods that learn bodies from data, pioneered by [4, 6].

Linear Blend Skinning. Linear Blend Skinning (LBS), also known as Skeletal-Subspace Deformation (SSD) [23, 24], is the foundation for many existing body models because of its simplicity. With LBS, the mesh is rigged with an underling set of joints forming a kinematic tree where each mesh vertex v_i is associated with n body joints and corresponding skinning weights w_i. The transformations applied to each mesh vertex are a weighted function of the transformations of the associated n joints. The skinning weights are typically defined by an artist or learned from data. In SMPL [21] the skinning weights are initialized by an artist and fine tuned as part of the training process. Numerous works attempt to predict the skinning weights for arbitrary body meshes, e.g. [13, 20].

Pose Corrective Blend Shapes. Although LBS is widely used, it suffers from well known shortcomings, which several method have been proposed to address. Lewis [19] introduces the pose space deformation model (PSD) where LBS is complemented with corrective deformations. The deformations are in the form of corrective offsets added to the mesh vertices posed with LBS. The corrective deformations are related to the underlying kinematic tree pose. Weighted pose deformation (WPD) [18, 31] adds pose corrective offsets to the base template mesh in the canonical (rest) pose before posing it with LBS, such that final posed mesh is plausible. Typically, such correctives are artist defined in key poses. Given a new pose, a weighted combination of correctives from nearby key poses is applied. Allen et al. [4] are the first to learn such corrective offsets from 3D scans of human bodies.

Learned Models. The release of the CAESAR dataset of 3D scans [32] enabled researchers to begin training statistical models of body shape [5, 35]. SCAPE [6] is the first model to learn a factored representation of body shape and pose. SCAPE models body deformations due pose and shape as triangle deformations and has been extended in many ways [7–10, 29, 30]. SCAPE has several downsides, however. It requires a least-squares solver to create a valid mesh, has no explicit joints or skeletal structure, may not maintain limb lengths when posed, and is not compatible with graphics pipelines and game engines.

To address these issues, Loper et al. [21] introduced SMPL, which uses vertex-based corrective offsets. Like SCAPE, SMPL factors the body into shape dependent deformations and pose dependent deformations. SMPL is more accurate

than SCAPE when trained on the same data and is based on LBS, making it easier to use. SMPL is also the first model trained using the full CAESAR dataset [32], giving it a realistic shape space; previous methods used a subset of CAESAR or even smaller datasets.

SMPL models pose correctives as a linear function of the elements of the part rotation matrices. This results in 207 pose blend shapes with each one having a global effect. Instead, we train a non-linear model that is linear in the pose (for good generalization) but non-linear in the spatial extent (to make it local). We adopt a unit quaternion representation and reduce that number of blend shapes from 207 to 23. These functions are not based on a single joint but rather on groups of joints, giving more expressive power. We train the correctives using a non-linear function that encourages spatial sparsity in the blend shapes. This results in a model that is 80% smaller than SMPL and reduces long-range spurious deformations. Loper et al. [21] also proposed a sparse version of SMPL but found that it reduced accuracy. In contrast, when trained on the same data, STAR is more accurate than SMPL. Additionally, we show that CAESAR is not sufficient and we train on more body shape data (14,000 scans in total) than any previous model.

SMPL and SCAPE factor body shape and pose-dependent shape change, but ignore correlations between them. Several methods model this with a tensor representation [7, 9]. This allows them to vary muscle deformation with pose depending on the muscularity of the subject. Here we achieve similar effects while keeping the benefits of simple models like SMPL.

Sparse Pose Corrective Blend Shapes. Human pose deformations are largely local in nature and, hence, the pose corrective deformations should be similarly local. Kry et al. [17] introduce EigenSkin to learn a localized model of pose deformations. STAR is similar to EigenSkin in that it models localized joint support but, unlike EigenSkin we infer the joint support region from posed scan data without requiring a dedicated routine of manually posing joints. Neumann et al. [27], use sparse PCA to learn local and sparse deformations of pose-dependent body deformations but do not learn a function mapping body pose to these deformations. In contrast, STAR learns sparse and local pose deformations that are regressed directly from the body pose. Contemporaneous with our work, GHUM [37] builds on SMPL and its Rodrigues pose representation but reduces the pose parameters (including face and hands) to a 32-dimensional latent code. Pose correctives are linearly regressed from this latent representation with L1 sparsity, giving sparse correctives.

3 Model

STAR is a vertex-based LBS model complemented with a learned set of shape and pose corrective functions. Similar to SMPL, we factor the body shape into the subject's intrinsic shape and pose-dependent deformations. In STAR we define a pose corrective function for each joint, j, in the kinematic tree. In

contrast to SMPL, we condition the pose corrective deformation function on both body pose $\boldsymbol{\theta} \in \mathbb{R}^{|\boldsymbol{\theta}|}$ and shape $\boldsymbol{\beta} \in \mathbb{R}^{|\boldsymbol{\beta}|}$. Additionally, during training, we use a non-linear activation function, $\phi(.)$, that selects the subset of mesh vertices relevant to the joint j. The pose corrective blend shape function makes predictions only about a subset of the mesh vertices. We adopt the same notation used in SMPL [21]. We start with an artist defined template, $\overline{\boldsymbol{T}} \in \mathbb{R}^{3N}$ in the rest pose $\boldsymbol{\theta}^*$ (i.e. T-Pose) where $N = 6890$ is the number of mesh vertices. The model kinematic tree contains $K = 24$ joints, corresponding to 23 body joints in addition to a root joint. The template $\overline{\boldsymbol{T}}$ is then deformed by a shape corrective blend shape function B_S that captures the subjects identity and a function B_P that adds correctives offsets such that mesh looks realistic when posed.

Shape Blend Shapes. The shape blend shape function $B_S(\boldsymbol{\beta}; \mathcal{S}) : \mathbb{R}^{|\boldsymbol{\beta}|} \rightarrow \mathbb{R}^{3N}$ maps the identity parameters $\boldsymbol{\beta}$ to vertex offsets from the template mesh as

$$B_S(\boldsymbol{\beta}; \mathcal{S}) = \sum_{n=1}^{|\boldsymbol{\beta}|} \beta_n S_n, \tag{1}$$

where $\boldsymbol{\beta} = [\beta_1, \cdots, \beta_{|\beta|}]$ are the shape coefficients, and $\mathcal{S} = [S_1, \cdots, S_{|\beta|}] \in \mathbb{R}^{3N \times |\boldsymbol{\beta}|}$ are the principal components capturing the space of human shape variability. The shape correctives are added to the template:

$$\boldsymbol{V}_{shaped} = \overline{\boldsymbol{T}} + B_S(\boldsymbol{\beta}; \mathcal{S}), \tag{2}$$

where $\boldsymbol{V}_{shaped}$ contains the vertices representing the subject's physical attributes and identity.

Pose and Shape Corrective Blend Shapes. The output of the shape corrective blend shape function, $\boldsymbol{V}_{shaped}$, is further deformed by a pose corrective function. The pose corrective function is conditioned on both pose and shape and adds corrective offsets such that, when the mesh is posed with LBS, it looks realistic. We denote the kinematic tree unit quaternion vector as $\boldsymbol{q} \in \mathbb{R}^{96}$ (24 joints each represented with 4 parameters). The pose corrective function is denoted as $B_P(\boldsymbol{q}, \beta_2) \in \mathbb{R}^{|\boldsymbol{q}| \times 1} \rightarrow \mathbb{R}^{3N}$, where β_2 is the PCA coefficient of the second principal component, which highly correlates with the body mass index (BMI) as shown in Sup. Mat.. The STAR pose corrective function is factored into a sum of pose corrective functions:

$$B_P(\boldsymbol{q}, \beta_2; \mathcal{K}, \mathbf{A}) = \sum_{j=1}^{K-1} B_P^j(\boldsymbol{q}_{ne(j)}, \beta_2; \mathcal{K}_j, \boldsymbol{A}_j), \tag{3}$$

where a pose corrective function is defined for each joint in the kinematic tree excluding the root joint. The per-joint pose corrective function $B_P^j(\boldsymbol{q}_{ne(j)}, \beta_2; \mathcal{K}_j, \boldsymbol{A}_j)$ predicts corrective offsets given $\boldsymbol{q}_{ne(j)} \subset \boldsymbol{q}$, where $\boldsymbol{q}_{ne(j)}$ is a set containing the joint j and its direct neighbors in the kinematic tree. This formulation results in more powerful regressors compared to SMPL. $\mathcal{K}_j \in \mathbb{R}^{3N \times |\boldsymbol{q}_{ne(j)}|+1}$ is a linear

regressor weight matrix and $\boldsymbol{A}_j$ are the activation weights for each vertex, both of which are learned. Each pose corrective function, $B_P^j(\boldsymbol{q}_{ne(j)}, \beta_2)$, is defined as a composition of two functions, an activation function and a pose corrective regressor.

Activation Function. For each joint, j, we define a learnable set of mesh vertex weights, $\boldsymbol{A}_j = [w_j^1, \cdots, w_j^N] \in \mathbb{R}^N$, where $w_j^i \in \mathbb{R}$ denotes the weight of the i^{th} mesh vertex with respect to the j joint. The weight w_j^i for each vertex i is initialized as the reciprocal of the minimum geodesic to the set of vertices around joint j, normalized to the range $[0, 1]$. The weights are thresholded by a non-linear activation function, specifically a rectified linear unit (ReLU):

$$\phi(w_j^i) = \begin{cases} 0, & \text{if } w_j^i \leq 0, \\ w_j^i, & \text{otherwise,} \end{cases} \tag{4}$$

such that during training, vertices with a $w_j^i \leq 0$ have weight 0. The remaining set of vertices with $w_j^i > 0$ defines the support region of joint j.

Pose Corrective Regressor. The per-joint pose corrective function is defined as $P_j(\boldsymbol{q}_{ne(j)}) \in \mathbb{R}^{|\boldsymbol{q}_{ne(j)}|+1} \rightarrow \mathbb{R}^{3N}$, which regresses corrective offsets given the joint and its direct neighbors' quaternion values

$$P_j(\boldsymbol{q}_{ne(j)}, \beta_2; \mathcal{K}_j) = \mathcal{K}_j((\boldsymbol{q}_{ne(j)} - \boldsymbol{q}^*_{ne(j)})^T | \beta_2)^T, \tag{5}$$

where $\boldsymbol{q}^*_{ne(j)}$ is the vector of quaternion values for the set of joints $ne(j)$ in rest pose, and β_2 is concatenated to the quaternion difference vector. $\mathcal{K}_j \in \mathbb{R}^{3N \times |\boldsymbol{q}_{ne(j)}|+1}$ is the regression matrix for joint j's pose correctives offsets. The predicted pose corrective offsets in Equation 5 are masked by the joint activation function:

$$B_P^j(\boldsymbol{q}_{ne(j)}; \boldsymbol{A}_j, \mathcal{K}_j) = \phi(\boldsymbol{A}_j) \circ P_j(\boldsymbol{q}_{ne(j)}, \beta_2; \mathcal{K}_j), \tag{6}$$

where $\boldsymbol{X} \circ \boldsymbol{Y}$ is the element wise Hadamard product between the vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$. During training, vertices with zero activation with respect to joint j, will have no corrective offsets added to them. Therefore when summing the contribution of the individual joint pose corrective functions in Equation 3, each joint only contributes pose correctives to the vertices for which there is support.

Blend Skinning. Finally, the mesh with the added pose and shape corrective offsets is transformed using a standard skinning function $W(\overline{\boldsymbol{T}}, \boldsymbol{J}, \boldsymbol{\theta}, \mathcal{W})$ around the joints, $\boldsymbol{J} \in \mathbb{R}^{3K}$ and linearly smoothed by a learned set of blend weight parameters $\mathcal{W}$. The joint locations are intuitively influenced by the body shape and physical attributes. Similar to SMPL, the joints $\boldsymbol{J}(\boldsymbol{\beta}; \mathcal{J}, \overline{\boldsymbol{T}}, \mathcal{S}) = \mathcal{J}(\boldsymbol{V}_{shaped})$ are regressed from $\boldsymbol{V}_{shaped}$ by a sparse function $\mathcal{J} : \mathbb{R}^{3N} \rightarrow \mathbb{R}^{3K}$.

To summarize, STAR is full defined by:

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

where T_P is defined as:

$$T_p(\boldsymbol{\beta}, \boldsymbol{\theta}) = \overline{\boldsymbol{T}} + B_S(\boldsymbol{\beta}) + B_P(\boldsymbol{q}, \beta_2), \tag{8}$$

where $\boldsymbol{q}$ is the quaternion representation of pose $\boldsymbol{\theta}$. The STAR model is fully parameterized by 72 (i.e. 24 * 3) pose parameters $\boldsymbol{\theta}$ in axis-angle representation, and up to 300 shape parameters $\boldsymbol{\beta}$.

3.1 Model Training

STAR training is similar to SMPL [22]. The key difference is the training of the pose corrective function in Equation 3. STAR pose corrective blend shapes are trained to minimize the *vertex-to-vertex* error between the model predictions and the ground-truth registrations where, in each iteration, the model parameters ($\mathcal{A}$,$\mathcal{K}$) are minimized by stochastic gradient descent across a batch of B registrations, denoted as $\boldsymbol{R} \in \mathbb{R}^{3N}$. The data term is given by:

$$\mathcal{L}_D = \frac{1}{B} \sum_{i=1}^{B} ||M(\boldsymbol{\beta}_i, \boldsymbol{\theta}_i) - \boldsymbol{R}_i||_2. \tag{9}$$

In addition to the data term we regularize the pose corrective regression weights ($\mathcal{K}$) with an $L2$ norm:

$$\mathcal{L}_B = \lambda_b \sum_{i=1}^{K-1} ||\mathcal{K}_i||_2, \tag{10}$$

where K is the number of joints in STAR and λ_b is a scalar constant. In order to induce sparsity in the activation masks $\phi(.)$, we use an $L1$ penalty

$$\mathcal{L}_A = \lambda_c || \sum_{i=1}^{K-1} \phi_j(\boldsymbol{A}_j)||_1, \tag{11}$$

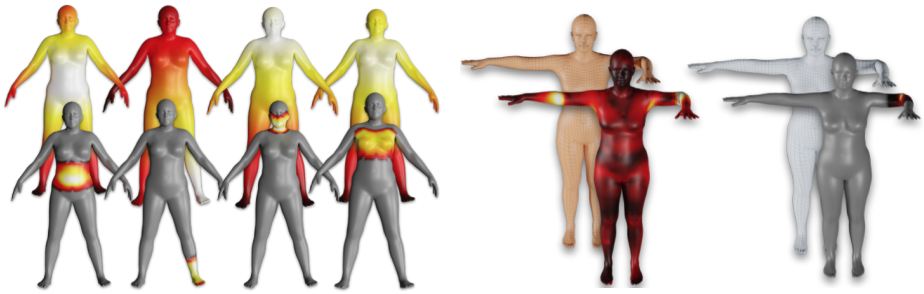
where λ_c is a scalar constant. Similar to SMPL we use a sparsity regularizer term on the skinning weights $\mathcal{W}$ and regularize the skinning weights to initial artist-defined skinning weights, $\mathcal{W}_{\text{prior}} \in \mathbb{R}^{N \times K}$:

$$\mathcal{L}_W = \lambda_p ||\mathcal{W} - \mathcal{W}_{\text{prior}}||_2 + \lambda_s ||\mathcal{W}||_1, \tag{12}$$

where λ_p and λ_s are scalar constants. To summarize the complete training objective is given by

$$\mathcal{L} = \mathcal{L}_D + \mathcal{L}_B + \mathcal{L}_A + \mathcal{L}_W. \tag{13}$$

The objective in Equation 13 is minimized with respect to the skinning weights $\mathcal{W}$, pose corrective regression weights $\mathcal{K}_{1:24}$, activation weights $\boldsymbol{A}_{1:24}$. We train the model iteratively. In each training iteration, we anneal the regularization parameters as described in the Sup. Mat.



(a) (b)

Fig. 3: **Spatially local and sparse pose corrective blend shapes**. (a) The top row shows a sample of the joints activation functions output before training and the bottom row shows the output after training (gray is zero). (b) shows SMPL (brown) and STAR (white) in the rest pose except for the left elbow, which is rotated. The heat map visualizes the corrective offsets for each model caused by moving this one joint. Note that unlike STAR, SMPL has spurious long-range displacements.

4 Experiments

4.1 Activation

Key to learning the sparse and spatially local pose corrective blend shapes are the joint activation functions introduced in Equation 4. During training the output of the activation functions becomes more sparse, limiting the number of vertices a joint can influence. Figure 3a summarizes a sample of the activation functions output before and after training. As a result of the output of the activation functions becoming more sparse, the number of model parameters decreases. By the end of training, the male model pose blend shapes contained 3.37×10^5 non-zero parameters and the female model contained 3.94×10^5 non-zero parameters compared to SMPL which has a dense pose corrective blendshape formulation with 4.28×10^6 parameters. At test time only the non-zero parameters need to be stored.

Figure 3b show a SMPL model bending an elbow resulting in a bulge in the other elbow, as a result of the pose corrective blend shapes learning long range spurious correlations from the training data. In contrast, STAR correctives are spatially local and sparse, this is a result of the learned local sparse pose corrective blend shape formulation of STAR.

4.2 Model Generalization

While the learned activation masks are sparse and spatially local, which is good, it is equally important that the model still generalizes to unseen bodies. To this end, we evaluate the model generalization on held out test subjects. The test set

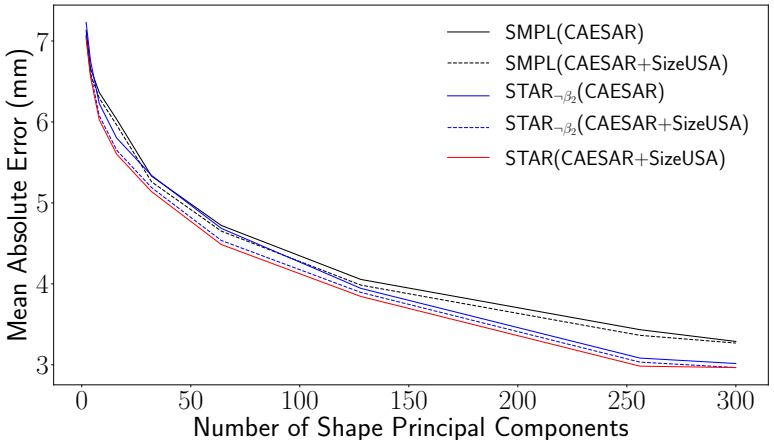


Fig. 4: **Generalization Accuracy**: Evaluating STAR and SMPL on unseen bodies. STAR$_{\neg\beta_2}$(CAESAR) is STAR trained on CAESAR with pose correctives depending on pose only (i.e. independent of β_2), STAR$_{\neg\beta_2}$(CAESAR+SizeUSA) is STAR trained on CAESAR and SizeUSA with pose corrective blend shapes depending on pose only, and STAR(CAEAR+SizeUSA) is STAR trained on CAEASAR and SizeUSA with pose and shape dependent pose corrective blend shapes.

we use contains the publicly available Dyna dataset [1] (the same evaluation set used in evaluating the SMPL model), in addition to the 3DBodyTex dataset [34] which contains static scans for 100 male and 100 female subjects in a diversity of poses. The total test set contains 570 registered meshes of 102 male subjects and 104 female subjects. We fit the models by minimizing the vertex to vertex mean absolute error (v2v), where the pose $\boldsymbol{\theta}$ and shape parameters $\boldsymbol{\beta}$ are the free optimization variables. We report the mean absolute error in (mm) as a function of the number of used shape coefficients in Figure 4. We first evaluate SMPL and STAR when they are both trained using the CAESAR dataset. In this evaluation both models are trained on the exact same pose and shape data. Since they both share the same topology and kinematic tree, differences in the fitting results are solely due to the different formulation of the two models. In Figure 4, STAR uniformly generalizes better than SMPL on the unseen test subjects. A sample qualitative comparison between SMPL and STAR fits is shown in Figure 5.

4.3 Extended Training Data

The CAESAR dataset is limited in its diversity, consequently limiting model generalization. Consequently, we extend the shape training database to include the SizeUSA database [2]. SizeUSA contains low quality scans of 2845 male and 6434 females with ages varying between 18 to 66+; a sample of the SizeUSA

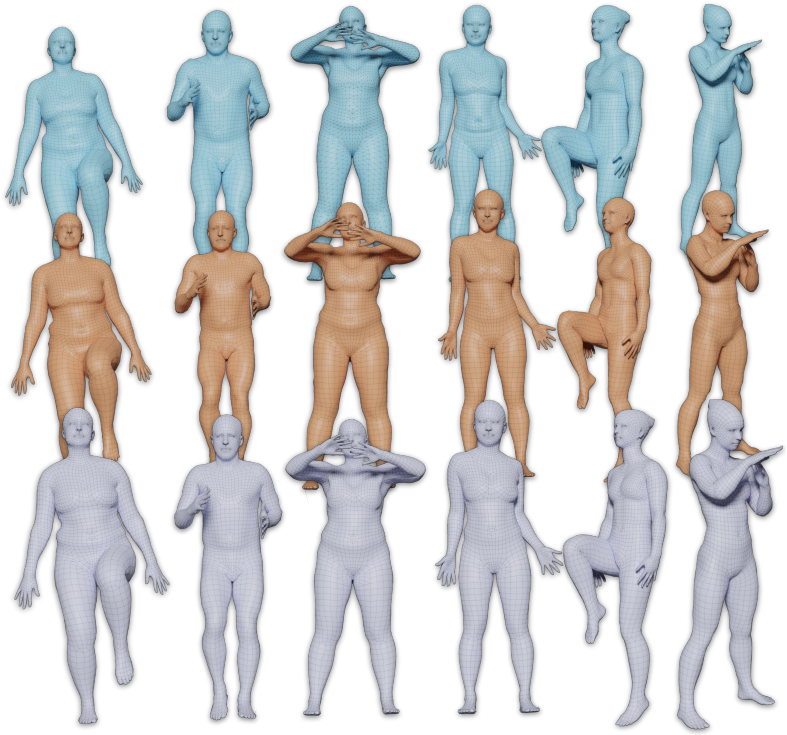
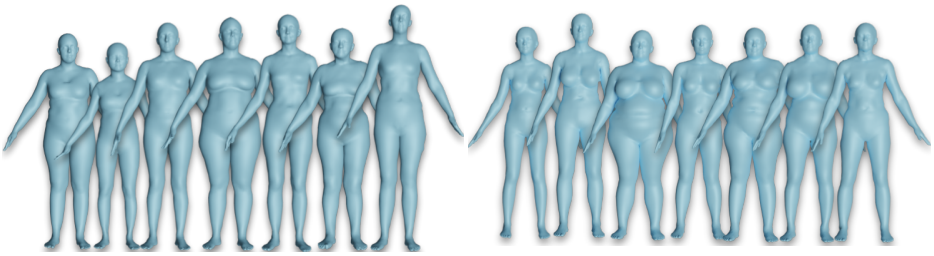


Fig. 5: **Qualitative Evaluation**: Comparison between SMPL and STAR. The ground truth registrations are shown in blue, the corresponding SMPL model fit meshes are shown in brown and STAR fits are shown in white. Here, both STAR and SMPL are trained on the CAESAR database.

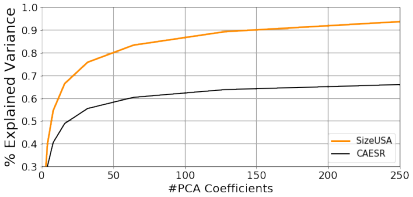
bodies compared to the CAESAR bodies are shown in Figure 6a and Figure 6b. We evaluate the generalization power of models trained separately on CEASER and SizeUSA. We do so by computing the percentage of explained variance of the SizeUSA subjects given a shape space trained on the CAESAR subjects, and vice versa. The results are shown in Figure 6 for the female subjects, the full analysis for both male and female subjects is shown in the Sup. Mat.. The key insight from this experiment is that a shape space trained on a single data set was not sufficient to explain the variance in the other data set. This suggests that training on both dataset should improve the model shape space expressiveness.

We retrain train both STAR and SMPL on the combined CAESAR and SizeUSA datasets an evaluate the model generalization on the held out test set as a function of the number of shape coefficient used as shown in Figure 4. Training on both CAESAR and SizeUSA results in both SMPL and STAR generalizing better than when trained only on CAESAR. We further note that STAR still uniformly generalizes better than SMPL when both models are trained on the combined CAESAR and SizeUSA dataset. Importantly STAR is more accurate

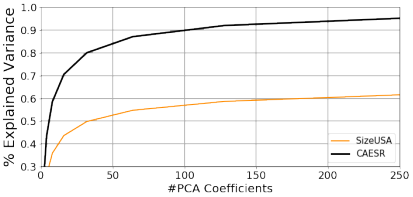


(a) Sample Females: CAESAR

(b) Sample Females: SizeUSA



(c) Shape Space Trained on SizeUSA



(d) Shape Space Trained on CAESAR

Fig. 6: **Explained Variance:** The percentage of explained variance of SizeUSA and CAESAR subjects when shape space is trained on SizeUSA is shown in Figure 6c and when the shape space is trained on CAESAR subjects in Figure 6d.

than SMPL despite the fact that uses many fewer parameters. Finally we extend the pose corrective blend shapes of STAR to be conditioned on both body pose and body shape and evaluate the model on the held out set. This results in a further improvement in the model generalization accuracy that, while modest, is consistent.

5 Discussion

STAR has 93 pose corrective blend shapes compared to 207 in SMPL and is 80% smaller than SMPL. It is surprising that it is able to uniformly perform better than SMPL when trained on the same data. This highlights the fact that the local and sparse assumptions of the pose corrective blend shapes is indeed realistic a priori knowledge that should be incorporated in any body model. Importantly, having fewer parameters means that STAR is less likely to overfit, even though our non-linear model makes training more difficult.

For SMPL, the authors report that enforcing sparsity of the pose corrective blend shapes resulted into worse results than SMPL. We take a different approach and learn the sparse set of vertices relevant to a joint from data. The key strength of our approach is that it is learned from data.

We are able to learn spatially local and sparse joint support regions due to two key implementation details: The initialization of the vertex weight $\boldsymbol{A}_j$

with the normalized inverse of geodesic distance to a joint. Secondly, the pose corrective blend shapes for each joint are regressed from local pose information, corresponding to the joint and its direct neighbors in the kinematic tree; this is a richer representation than SMPL. These two factors together with the sparsity inducing $L1$ norm on the activation weights, act as an inductive bias to learn a sparse set of vertices that are geodesically local to a joint.

The sparse pose correctives formulation reduces the number of parameters and regularizes the model, preventing it from learning spurious long range correlations from the training data. Since each vertex is only influenced by a limited number of joints in the kinematic tree, the gradients propagated through the model are sparse and the derivative of a vertex with respect to a geodesically distant joint is 0, which is not the case in the SMPL.

6 Conclusion

We have introduced STAR, which has fewer parameters than SMPL yet is more accurate and generalizes better to unseen bodies when trained on the same data. Our key insight is that human pose deformation is local and sparse. While this observation is not new, our formulation is. We define a non-linear (ReLU) activation function for each joint and train the model from data to estimate both the linear corrective pose blend shapes and the activation region on the mesh that these joints influence. We kept what is popular with SMPL while improving on it in every sense. STAR has only 20% of the pose corrective parameters of SMPL. Our training method and localized model fixes a key problem of SMPL– the spurious, long-range, correlations that result in non-local deformations. Such artifacts make SMPL unappealing for animators. Moreover, we show that, while SMPL is trained from thousands of scans, human bodies are more varied than the CAESAR dataset. More training scans results in a better model. Finally we make pose-corrective blend shapes depend on body shape, producing more realistic deformations. We make STAR available for research with 300 shape principal components. It can be swapped in for SMPL in any existing application since the pose and shape parameterization is the same to the user. Future work work should extend this approach to the SMPL-X model which includes an expressive face and hands.

Acknowledgments: The authors thank N. Mahmood for insightful discussions and feedback, and the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting A. A. A. Osman. The authors would like to thank Joachim Tesch, Muhammed Kocabas, Nikos Athanasiou, Nikos Kolotouros and Vassilis Choutas for their support and fruitful discussions.

Disclosure: In the last five years, MJB has received research gift funds from Intel, Nvidia, Facebook, and Amazon. He is a co-founder and investor in Meshcapade GmbH, which commercializes 3D body shape technology. While MJB is a part-time employee of Amazon, his research was performed solely at, and funded solely by, MPI.

References

1. Dyna dataset. http://dyna.is.tue.mpg.de/ (2015), accessed: 2015-05-15
2. SizeUSA dataset. https://www.tc2.com/size-usa.html (2017)
3. Alldieck, T., Magnor, M., Xu, W., Theobalt, C., Pons-Moll, G.: Detailed human avatars from monocular video. In: International Conference on 3D Vision (3DV). pp. 98–109 (2018)
4. Allen, B., Curless, B., Popović, Z.: Articulated body deformation from range scan data. ACM Transactions on Graphics, (Proc. SIGGRAPH) **21**(3), 612–619 (2002)
5. Allen, B., Curless, B., Popović, Z.: The space of human body shapes: Reconstruction and parameterization from range scans. ACM Transactions on Graphics, (Proc. SIGGRAPH) **22**(3), 587–594 (2003)
6. Anguelov, D., Srinivasan, P., Koller, D., Thrun, S., Rodgers, J., Davis, J.: SCAPE: Shape Completion and Animation of PEople. ACM Transactions on Graphics, (Proc. SIGGRAPH) **24**(3), 408–416 (2005)
7. Chen, Y., Liu, Z., Zhang, Z.: Tensor-based human body modeling. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 105–112 (2013)
8. Freifeld, O., Black, M.J.: Lie Bodies: A manifold representation of 3D human shape. In: European Conference on Computer Vision (ECCV). pp. 1–14 (2012)
9. Hasler, N., Stoll, C., Sunkel, M., Rosenhahn, B., Seidel, H.P.: A statistical model of human pose and body shape. Computer Graphics Forum **28**(2), 337–346 (2009)
10. Hirshberg, D., Loper, M., Rachlin, E., Black, M.: Coregistration: Simultaneous alignment and modeling of articulated 3D shape. In: European Conference on Computer Vision (ECCV). pp. 242–255 (2012)
11. Huang, Y., Bogo, F., Lassner, C., Kanazawa, A., Gehler, P.V., Romero, J., Akhter, I., Black, M.J.: Towards accurate marker-less human shape and pose estimation over time. In: International Conference on 3D Vision (3DV). pp. 421–430 (2017)
12. Huang, Y., Kaufmann, M., Aksan, E., Black, M.J., Hilliges, O., Pons-Moll, G.: 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 (2018)
13. Jacobson, A., Baran, I., Kavan, L., Popović, J., Sorkine, O.: Fast automatic skinning transformations. ACM Transactions on Graphics (TOG) **31**(4), 1–10 (2012)
14. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7122–7131 (2018)
15. Kocabas, M., Athanasiou, N., Black, M.J.: VIBE: Video inference for human body pose and shape estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5253–5263 (2020)
16. Kolotouros, N., Pavlakos, G., Black, M.J., Daniilidis, K.: Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 2252–2261 (2019)
17. Kry, P.G., James, D.L., Pai, D.K.: Eigenskin: real time large deformation character skinning in hardware. In: Proceedings of the 2002 ACM SIGGRAPH/Eurographics symposium on Computer animation. pp. 153–159 (2002)
18. Kurihara, T., Miyata, N.: Modeling deformable human hands from medical images. In: Proceedings of the 2004 ACM SIGGRAPH/Eurographics symposium on Computer animation. pp. 355–363 (2004)

19. Lewis, J.P., Cordner, M., Fong, N.: Pose space deformation: A unified approach to shape interpolation and skeleton-driven deformation. In: Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques. pp. 165–172. SIGGRAPH '00 (2000)
20. Liu, L., Zheng, Y., Tang, D., Yuan, Y., Fan, C., Zhou, K.: NeuroSkinning: Automatic skin binding for production characters with deep graph networks. ACM Transactions on Graphics (TOG) **38**(4), 1–12 (2019)
21. 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 (2015)
22. Loper, M.M., Mahmood, N., Black, M.J.: MoSh: Motion and shape capture from sparse markers. ACM Transactions on Graphics, (Proc. SIGGRAPH Asia) **33**(6), 220:1–220:13 (2014)
23. Magnenat-Thalmann, N., Laperrire, R., Thalmann, D.: Joint-dependent local deformations for hand animation and object grasping. In: In Proceedings on Graphics interface. Citeseer (1988)
24. Magnenat-Thalmann, N., Thalmann, D.: Human body deformations using joint-dependent local operators and finite-element theory. Tech. rep., EPFL (1990)
25. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: AMASS: Archive of motion capture as surface shapes. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 5442–5451 (2019)
26. von Marcard, T., Rosenhahn, B., Black, M., Pons-Moll, G.: Sparse inertial poser: Automatic 3D human pose estimation from sparse IMUs. Computer Graphics Forum 36(2), Proceedings of the 38th Annual Conference of the European Association for Computer Graphics (Eurographics) pp. 349–360 (2017)
27. Neumann, T., Varanasi, K., Wenger, S., Wacker, M., Magnor, M., Theobalt, C.: Sparse localized deformation components. ACM Transactions on Graphics (TOG) **32**(6), 1–10 (2013)
28. Pavlakos, G., Zhu, L., Zhou, X., Daniilidis, K.: Learning to estimate 3D human pose and shape from a single color image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 459–468 (2018)
29. Pishchulin, L., Wuhrer, S., Helten, T., Theobalt, C., Schiele, B.: Building statistical shape spaces for 3D human modeling. Pattern Recognition **67**, 276–286 (2017)
30. Pons-Moll, G., Romero, J., Mahmood, N., Black, M.J.: Dyna: A model of dynamic human shape in motion. ACM Transactions on Graphics, (Proc. SIGGRAPH) **34**(4), 120:1–120:14 (2015)
31. Rhee, T., Lewis, J.P., Neumann, U.: Real-time weighted pose-space deformation on the gpu. Computer Graphics Forum **25**(3), 439–448 (2006)
32. 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)
33. Rueegg, N., Lassner, C., Black, M.J., Schindler, K.: Chained representation cycling: Learning to estimate 3D human pose and shape by cycling between representations. In: Conference on Artificial Intelligence (AAAI-20) (2020)
34. Saint, A., Ahmed, E., Cherenkova, K., Gusev, G., Aouada, D., Ottersten, B.: 3DBodyTex: Textured 3D body dataset. In: International Conference on 3D Vision (3DV). pp. 495–504 (2018)
35. Seo, H., Cordier, F., Magnenat-Thalmann, N.: Synthesizing animatable body models with parameterized shape modifications. In: ACM SIGGRAPH/Eurographics Symposium on Computer Animation. pp. 120–125. SCA '03 (2003)

36. Tan, J., Budvytis, I., Cipolla, R.: Indirect deep structured learning for 3D human body shape and pose prediction. In: Proceedings of the British Machine Vision Conference (BMVC). pp. 4–7 (2017)
37. 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: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6184–6193 (2020)
38. Zanfir, A., Marinoiu, E., Sminchisescu, C.: Monocular 3D pose and shape estimation of multiple people in natural scenes-the importance of multiple scene constraints. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2148–2157 (2018)