GHUM & GHUML: Generative 3D Human Shape and Articulated Pose Models

Hongyi Xu Eduard Gabriel Bazavan Andrei Zanfir
William T. Freeman Rahul Sukthankar Cristian Sminchisescu

Google Research

{hongyixu, egbazavan, andreiz, wfreeman, sukthankar, sminchisescu}@google.com

Abstract

We present a statistical, articulated 3D human shape modeling pipeline, within a fully trainable, modular, deep learning framework. Given high-resolution complete 3D body scans of humans, captured in various poses, together with additional closeups of their head and facial expressions, as well as hand articulation, and given initial, artist designed, gender neutral rigged quad-meshes, we train all model parameters including non-linear shape spaces based on variational auto-encoders, pose-space deformation correctives, skeleton joint center predictors, and blend skinning functions, in a single consistent learning loop. The models are simultaneously trained with all the 3d dynamic scan data (over $60,000$ *diverse human configurations in our new dataset) in order to capture correlations and ensure consistency of various components. Models support facial expression analysis, as well as body (with detailed hand) shape and pose estimation. We provide fully trainable generic human models of different resolutions – the moderate-resolution GHUM consisting of 10,168 vertices and the low-resolution GHUML(ite) of 3,194 vertices –, run comparisons between them, analyze the impact of different components and illustrate their reconstruction from image data. The models will be available for research.*

1. Introduction

Human motion, action, and expression are of central practical importance, and subject to continuous focus, as well as creative capture in images and video. Immersive photography, augmented and virtual reality, and physical 3D space reasoning would be next. Consequently, models that can accurately represent the full body detail at the level of pose, shape, and facial expression, as well as hand manipulation are essential in order to capture and deeply analyze those subtle interactions that can only be fully understood in 3D. While considerable progress has been made in localizing human stick figures in images and video, and

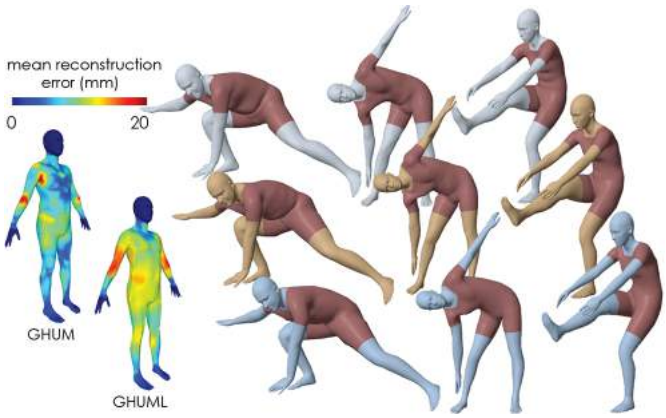

Figure 1. Illustration of accuracy of GHUM and GHUML on data from GHS3D, with heatmaps of both models on the left. Renderings show registrations of different body poses of a subject (grey, first row), as well as GHUM and GHUML reconstructions in second and third rows, respectively. Notice good level of detail capture for both models, with higher accuracy for GHUM.

– under certain conditions – lifting to equivalent 3D skeletons and basic shapes, the general quest for reconstructing accurate models of the the human body at the level of semantically meaningful surfaces, grounded in a 3D physical space, is still on.

The potential for model construction advances, at least in the medium term, appears to be at the incidence between intuitive physical and semantic human modeling, and large-scale datasets. While many expressive models for faces, hands and bodies have been constructed over time, most – if not all – were built in isolation rather than in the context of a full human body. Hence, inevitably, they did not take advantage of the large scale data analysis and model construction process that recently emerged in the context of deep learning. A number of recent full body models like Adam, Frank, or SMPL-X[14, 31], combine legacy components for face, body and hands, but usually focus on constructing a consistent, joint parameterization with proper scaling on top of already learnt components, rather than on training a full

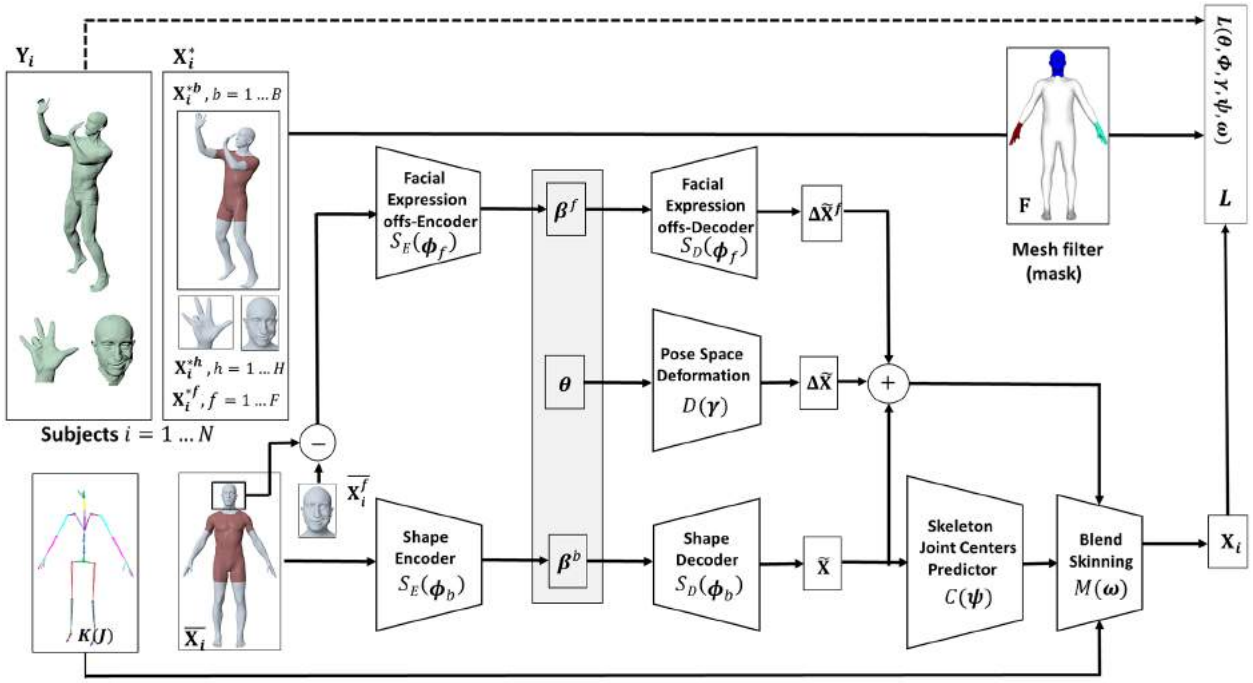

Figure 2. Overview of our end-to-end statistical 3D articulated human shape model construction. We are given a set of high-resolution 3D body scans including both 'A' – and arbitrary – poses exposing a variety of articulation and soft tissue deformations. Additionally, we also collect head closeup scans of detailed facial expressions and hand closeup scans to capture different gestures and object grabs. Body landmarks are automatically identified by rendering the photorealistic 3D reconstructions in multiple virtual viewpoints, detecting them in the generating images and triangulating. An artist designed full body articulated mesh is progressively registered to point clouds using losses that combine sparse landmark correspondences and dense iterative closest point (ICP) residuals (implemented as point scan to mesh facet distances), under as conformal as possible surface priors[41]. The model has non-linear shape spaces implemented as deep variational auto-encoders (VAEs) for the body $\boldsymbol{\phi}_b$, and offset VAEs for the facial expressions $\boldsymbol{\phi}_f$, and includes trainable pose-space deformation functions D, modulated by a skeleton $\mathbf{K}$ with J joints, centers predictor C, and blend skinning functions M. During training, all high-resolution scans of the same subjects (both full-body and closeups for face and hands) are used (see fig. 3), with residuals appropriately masked by the filter $\mathbf{F}$. For model construction, we use N captured subjects, with B full body scans, F closeup hand scans, and H closeup head scans. During learning, we alternate between minimizing the loss function w.r.t. pose estimates in each scan $\boldsymbol{\theta}$, and optimizing it with respect to the other model parameters $(\boldsymbol{\phi}, \boldsymbol{\gamma}, \boldsymbol{\psi}, \boldsymbol{\omega})$. In operation, *e.g.* for pose and shape estimation, the model is controlled by parameters $\boldsymbol{\alpha} = (\boldsymbol{\theta}, \boldsymbol{\beta})$, including kinematic pose $\boldsymbol{\theta}$ and VAE latent spaces for body shape and facial expressions $\boldsymbol{\beta} = (\boldsymbol{\beta^f}, \boldsymbol{\beta^b})$, with encoder-decoders given by $\boldsymbol{\phi} = (\boldsymbol{\phi}^f, \boldsymbol{\phi}^b)$.

body model, end-to-end, based on a large data repository. This makes it difficult to take full advantage of the structure in all data simultaneously, experiment with alternative representations for components or different losses, assess end impact, and innovate.

In this paper we propose an end-to-end learning pipeline to construct full body, statistical human shape and pose models capable of actuating facial expressions, as well as body and hand motion. We design end-to-end pipelines and unified loss functions based on deep learning, which allow for the simultaneous training of all model components, including non-linear shape spaces, pose-space deformation correctives, skeleton joint center estimators, and blend skinning functions in the context of minimal human skeleton parameterizations with anatomical joint angle constraints. The models are trained with high-resolution full body scans, as well as closeups of moving faces and hands, in order to capture maximum detail and ensure design consistency between body components. Our new collected 3D dataset of generic human shapes, GHS3D, consists of over 60,000 photo-realistic dynamic human body scans, and we also use over 4,000 full body scans from Caesar. We introduce both a moderate-resolution model, GHUM, and a specially designed (not down-sampled) low-resolution model GHUML, assess their relative performance for registration and constrained 3d surface fitting, under different linear and non-linear models (*e.g.* PCA or variational auto-encoders for body shape and facial expressions), and illustrate recovery of shape and pose from images.

Related work. There is a remarkable amount of work devoted to both constructing 3D articulated surface models for body parts, *i.e.* faces, hands and full bodies[2, 4, 10, 24, 37,

30, 11, 29, 16, 38, 3, 32, 7, 8, 21, 39], as well as, more recently, integrating them into complete, more expressive representations, as *e.g*. in Adam, Frank or SMPL-X[14, 31]. Many image and video-based pose and shape estimation methods have also been proposed[33, 27, 25, 40, 22, 1, 12, 26, 34, 23].

The Frank model[15] is based on a simplified version of the SMPL body[24], to which it connects an artist-designed hand rig, and the public FaceWarehouse head[8]. The combined asset has possibly inconsistent components grafted together, resulting in a model that may lack realism. In turn, SMPL-X attaches the FLAME[21] head to the SMPL-H (body and hand) model[35] and refits it to an additional set of 5,586 scans. However, since those full body scans have limited resolution for hands and faces, the authors use the original, pre-trained parameters of MANO and FLAME (pose space and pose corrective blendshapes of MANO[35] for the hands, and the expression space of FLAME[21], respectively), thus limiting the amount of data simultaneously used for learning the full model, and the potential realism attainable by jointly refining all parameters. In contrast to combining legacy components, we focus on using all high-resolution data simultaneously – both full body and closeup detail for faces and hands –, in order to construct low-res and high-res models where all parameters are refined end-to-end from the onset. This allows us to experiment with different resolutions, linear and non-linear shape spaces, loss functions, and assess their impact seamlessly for different tasks. Recent work focuses on building deep learning pipelines to predict articulated meshes from point clouds[19, 13]. These registration alternatives would be immediately applicable in our framework, although here we rely on direct optimization for registration with automatic landmark detection for accuracy, robustness, and generalization to virtually any pose and human datapoint scan.

Considerable work has been devoted to estimating 3D pose and shape from images acquired with one or several cameras or from video[33, 27, 25, 40, 22, 1, 12, 26, 9]. Several models rely on feed-forward pose and shape prediction based on different learning architectures, on pose prediction followed by pose and shape refinement to body joints of semantic body part segmentations, or on multicamera fusion[28, 43, 44, 20, 36, 5, 18]. Most shape priors come in the form of PCA as available in SMPL[24], Frank[15], or SMPL-X[31], and the pose priors are usually Gaussian Mixture Models [6], and more recently VAEs[31]. In contrast, our GHUM and GHUML rely on non-linear shape spaces constructed from deep variational autoencoders for body and facial deformation and on normalizing flow representations for skeleton (body and hand) kinematics[42]. Moreover our minimal skeleton parameterization supports the seamless integration of anatomical joint angle limits constraints during registration, learning and pose optimization, which reduces the search space, and makes estimates anatomically consistent and more robust.

While our primary goal in this paper is to introduce new end-to-end learnable 3D statistical articulated human body shape methodologies, the models we present are useful in connection with most work aiming to recover pose and shape from images. Moreover, by creating both a medium-resolution and a low-resolution model, we enable lightweight mobile applications of 3d human sensing, or approaches where different level of detail and run-time constraints could make it adequate to dynamically switch between models of different complexity.

2. Overview

Given a training set of human body scans, represented as unstructured point clouds $\{\mathbf{Y} \in \mathbb{R}^{3P}\}$, where the number of points P varies, we learn a statistical human model $\mathbf{X}(\boldsymbol{\alpha}) \in \mathbb{R}^{3V}$ representing the variability of body shapes and natural deformation due to articulation. The body model $\mathbf{X}$ has consistent topology with V vertices, as specified by an artist-provided (rigged) template mesh, and $\boldsymbol{\alpha}$ are variables that control the body deformation as a result of both shape and articulation. As illustrated in fig. 2, to learn a data-driven human model from 3D scans $\mathbf{Y}$, we first register the body template to point clouds in order to obtain new meshes of the same topology, marked as $\{\mathbf{X}^* \in \mathbb{R}^{3V}\}$ (see Sup. Mat. for details on our registration methodology). We then feed the registered meshes $\mathbf{X}^*$ into an end-to-end

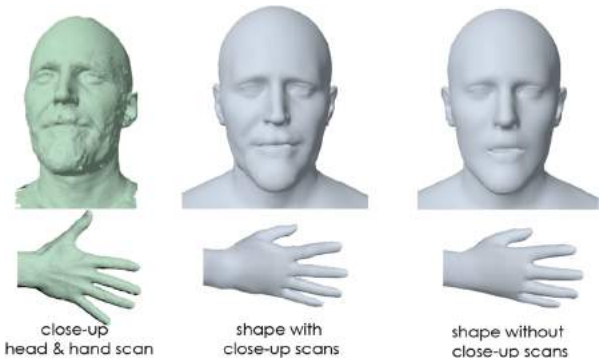

Figure 3. We estimate the full body shape at a neutral A pose by fusing the body scan and the closeup hand and head scans. Compared with body shape estimation from a single body scan, we can thus take advantage of additional head and hand shape detail.

training network where model parameters $\boldsymbol{\alpha}$ are adjusted to produce outputs that closely match the input as a result of both articulation and shape adjustment. In practice, we experimented with both direct model parameter adjustment to the point cloud via iterative closest point (ICP) losses (identical to the ones used for registration) or with alignment to the proxy meshes $\mathbf{X}^*$. Since our registration process is extremely accurate, we haven't noticed any significant differ-

ence between the two. In contrast, using target input meshes $\mathbf{X}^*$ of the same model topology, makes the process considerably faster and training losses are better behaved.

2.1. Human Model Representation

We represent the human model as an articulated mesh, specified by a skeleton $\mathbf{K}$ with J joints and skin deformed based on Linear Blending Skinning (LBS) to explicitly encode the motion of joints. In addition to skeletal articulated motion, we use nonlinear models to drive facial expressions. A model $\mathbf{X}$ with J joints can be formulated as $M(\boldsymbol{\alpha} = (\boldsymbol{\theta}, \boldsymbol{\beta}), \boldsymbol{\phi}, \boldsymbol{\gamma}, \boldsymbol{\psi}, \boldsymbol{\omega})$, or in detail, as

$$\mathbf{X}(\boldsymbol{\alpha}) = M(\boldsymbol{\theta}, \tilde{\mathbf{X}}(\boldsymbol{\beta}^b), \boldsymbol{\Delta}\tilde{\mathbf{X}}(\boldsymbol{\theta}), \boldsymbol{\Delta}\tilde{\mathbf{X}}^f(\boldsymbol{\beta}^f), C(\tilde{\mathbf{X}}), \boldsymbol{\omega}) \tag{1}$$

where $\tilde{\mathbf{X}}(\boldsymbol{\beta}^b) \in \mathbb{R}^{3V}$ is the identity-based rest shape in A-pose (fig. 2), with $\boldsymbol{\beta}^b$ a low-dimensional embedding vector encoding body shape variability (different low-dimensional representations including PCA or VAEs will be used); similarly, $\boldsymbol{\Delta}\tilde{\mathbf{X}}^f(\boldsymbol{\beta}^f)$, is the facial expression at neutral head pose controlled by low-dimensional latent code $\boldsymbol{\beta}^f$; $\mathbf{c} = C(\tilde{\mathbf{X}}) \in \mathbb{R}^{3J}$ are skeletal joint centers dependent on the body shape, $\boldsymbol{\theta} \in \mathbb{R}^{3\times(J+1)}$ is a vector of skeleton pose parameters consisting of (up to) 3 rotational DOFs in Euler angles for each joint and 3 translational variables at the root, $\boldsymbol{\omega} \in \mathbb{R}^{V\times I}$ are per-vertex skinning weights influenced by at most $I = 4$ (in our experiments) joints. FInally, pose-dependent corrective blend shapes $\boldsymbol{\Delta}\tilde{\mathbf{X}}(\boldsymbol{\theta})$ are added to the rest shape to correct for skinning artifacts. We initialize our human models, GHUM and GHUML, using artist-defined rigged template meshes ($V_{\text{ghum}} = 10,168, V_{\text{ghuml}} = 3,194, J = 63$), respectively and our pipeline will estimate all the parameters $(\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\gamma}, \boldsymbol{\psi}, \boldsymbol{\omega})$ while the mesh topology and the joint hierarchy $\mathbf{K}$ are considered fixed. The hierarchy is anatomically (minimally) parameterized in order to take advantage of bio-mechanical joint angle limits during optimization. Vertices $\mathbf{x}_i \in \mathbf{X}$ can be written as

$$\mathbf{x}_i = \sum_{j=1}^{I} \omega_{i,j} \mathbf{T}_j(\boldsymbol{\theta}, \mathbf{c}) \mathbf{T}_j(\bar{\boldsymbol{\theta}}, \mathbf{c})^{-1} \begin{bmatrix} \tilde{\mathbf{x}}_i + \boldsymbol{\Delta}\tilde{\mathbf{x}}_i + \boldsymbol{\Delta}\tilde{\mathbf{x}}_i^f \\ 1 \end{bmatrix} \tag{2}$$

$$\mathbf{T}_j(\boldsymbol{\theta}, \mathbf{c}) = \prod_{a \in \mathbf{K}(j)} \begin{bmatrix} \mathbf{R}_a(\theta_a) & \mathbf{c}_a \\ 0 & 1 \end{bmatrix} \in SE(3), \tag{3}$$

where $\mathbf{T}_j(\boldsymbol{\theta}, \mathbf{c})$ is the world transformation matrix for joint j, integrated by traversing the kinematic chain from the root to j. The transformation from rest to posed mesh is constructed by multiplying with the inverse of the world transformation matrix at rest pose $\bar{\boldsymbol{\theta}}$.

3. End-to-End Statistical Model Learning

In this section, we will provide an end-to-end neural network-based pipeline where we optimize the skinning weights $\boldsymbol{\omega}$, and learn a rest shape embedding $\boldsymbol{\beta}^b$, a facial expression embedding $\boldsymbol{\beta}^f$, identity shape-dependent joint centers estimator $C(\boldsymbol{\psi})$, pose-dependent blend shapes function $D(\boldsymbol{\gamma})$ given multi-subject and multi-pose surface meshes $\mathbf{X}^*$ registered to full body and close-up face and hand scans (fig. 3). As a result of ICP registration, we can easily formulate reconstruction losses using per-vertex Euclidean distance error under one-to-one correspondences as

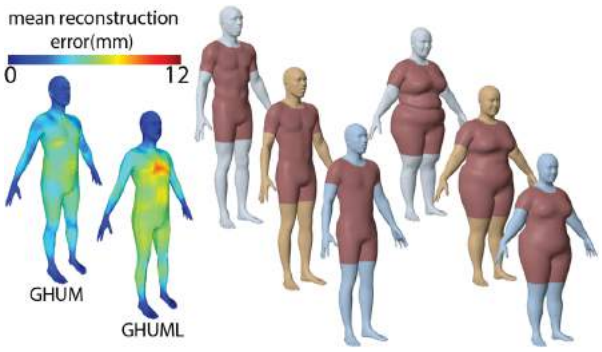

Figure 4. Evaluation on Caesar. Left: Per-vertex Euclidean error to the registration for GHUM and GHUML. Right: (top to bottom, registrations, GHUM and GHUML) VAE-based models can represent body shape very well. Compared to GHUML additional, *e.g.* muscle or waist, soft tissue detail is preserved by GHUM.

$$L_r(\mathbf{X}^*, \mathbf{X}(\boldsymbol{\alpha})) = \frac{1}{V} \sum_{i=1}^{V} \|\mathbf{F}_i(\mathbf{x}_i - \mathbf{x}_i^*)\|, \tag{4}$$

where $\mathbf{F}$ is a filter that accounts for different types of data (full body scans as opposed to closeups). In order to construct $\mathbf{X}(\boldsymbol{\alpha})$, we need to jointly estimate the pose $\boldsymbol{\theta}$ and the statistical shape parameters. We rely on block coordinate descent, alternating between estimation of pose parameters $\boldsymbol{\theta}$ under the current shape parameters $\boldsymbol{\beta}$, based on a BFGS layer, and updating the other model parameters with $\boldsymbol{\theta}$ fixed. We initialize skinning from the artist-provided defaults, all other parameters to 0. In the sequel, we detail how each sub-module updates the parameters $\boldsymbol{\alpha}$ of the global loss (4).

3.1. Variational Body Shape Autoencoder

We obtain multi-subject shape scans by registering our models to the Caesar dataset ($4,329$ subjects) as well as our captured scans in GHS3D, in neutral A-pose. For now, given rest shapes $\bar{\mathbf{X}}$ estimated for multiple subjects, we build a compact latent space for the body shape variation. Instead of simply building a PCA subspace, here we choose to represent body shape using a deep nonlinear variational autoencoder with a lower-dimensional latent code. Because we estimate mesh articulation, the input scans to our autoencoder $\bar{\mathbf{X}}$ are all well aligned at A-pose without significant perturbations from rigid transformations or pose articulation. The encoder and decoder are using parametric ReLU

activation functions, as they can model either an identity transformation or a standard ReLU, for certain parameters. As standard practice, the variational encoder will output a mean and a variance (μ, Σ), which will be transformed to the latent space through the re-parametrization trick [17], in order to obtain the sampled code $\boldsymbol{\beta}^b$. We choose a simple distribution, $\mathcal{N}(0, I)$, and integrate the Kullback-Leibler divergence in the loss function, to regularize the latent space

$$\tilde{\mathbf{X}}(\boldsymbol{\beta}^b) = \frac{1}{NB}\sum_{1}^{NB}\bar{\mathbf{X}} + S_D(\boldsymbol{\beta}^b) \tag{5}$$

$$\boldsymbol{\beta}^b = S_E\Big(\bar{\mathbf{X}} - \frac{1}{NB}\sum_{1}^{NB}\bar{\mathbf{X}}\Big) \tag{6}$$

where the encoder S_E captures the variance from the mean body shape into the latent vector $\boldsymbol{\beta}^b$ and the decoder S_D builds up the rest shape from $\boldsymbol{\beta}^b$ to match the input target rest shape. In particular, we initialize the first and last layers of the encoder and decoder, respectively, to the PCA subspace $\mathbf{U} \in \mathbb{R}^{3V\times L}$, where L is the dimensionality of the latent space. All other fully-connected layers are initialized to identity, including the PReLU units. We initialize the sub-matrix of log-variance entries to 0, and set the bias to a sufficiently large negative value. The network will thus, effectively initialize from the linear model, while keeping additional parameters to a minimum, compared to PCA.

3.2. Variational Facial Expression Autoencoder

The variational shape autoencoder can represent various body proportions, including the variances of face shapes. To additionally support complex facial expressions (as opposed to just anthropometric head and face variations at rest) we introduce additional facial modeling. We build the model from thousands of facial expression motion sequence scans available in GHS3D. In addition to a 3-DOF articulated jaw, two 2-DOF eyelids and two 2-DOF eyeballs, the parameters of the articulated joints on the head, including skinning weights and pose space deformation, will be updated together with the rest of pipeline. For facial motions caused by expression and not articulation, we build a nonlinear embedding $\boldsymbol{\beta}^f$ with the same network structure as the variational body shape autoencoder. The input to the VAE is a facial expression $\boldsymbol{\Delta}\bar{\mathbf{X}}^f \in \mathbb{R}^{3V^f}$ ($V^f = 2,056$ for GHUM and 596 for GHUML) at neutral head pose by removing all articulated joint motion (including neck, head, eyes and jaw). To un-pose the registered head mesh to neutral, we first fit the articulated joint motion $\boldsymbol{\theta}$ for the neutral head shape (without expression) that matches the registration as much as possible *c.f.* (4). The displacement field between the posed head and the registration is accounted to facial expressions, and before estimating it, we undo (unpose) the effect of articulated joint motion $\boldsymbol{\theta}$.

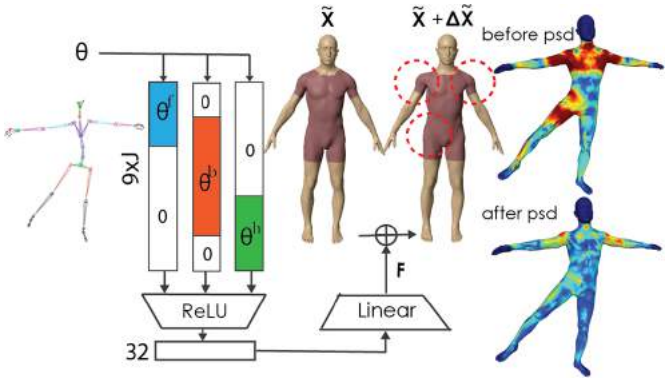

Figure 5. Pose space deformation architecture sketch and illustration showing the benefit of PSD, here around non-passive articulation points, *e.g.* right hip and thigh, as well as chest and armpits. For simplicity of illustration, here we use θ as the input feature, instead of $\mathbf{R}_i(\boldsymbol{\theta}_i) - \mathbf{R}_i(\bar{\boldsymbol{\theta}}_i)$.

3.3. Skinning Model

Besides nonlinear shape and facial expression models, we rely on optimal skinning functions estimated from multi-subject and multi-pose mesh data. Specifically, we share the same data term as in (4) but now the optimization variables are parameters of the joint center predictor $C(\boldsymbol{\psi}) : \tilde{\mathbf{X}} \rightarrow \mathbf{c}$, pose-dependent corrections to body shape $D(\boldsymbol{\gamma}) : \boldsymbol{\theta} \rightarrow \Delta\tilde{\mathbf{X}}$, and skinning weights $\boldsymbol{\omega}$. A natural choice for the skeletonal joint centers is to place them at average positions on the ring of boundary vertices connecting two mesh components (segmentations) maximally influenced by a joint. The average of boundary vertices, $\bar{\mathbf{C}}\tilde{\mathbf{X}} \in \mathbb{R}^{3J}$, imposes that the skeleton lies in the convex hull of the mesh surface, thus adapting the center placement to different body proportions. However, we observe downgraded skinning quality when using such predictors. For better skinning, we keep the estimate $\bar{\mathbf{C}}$ but on top build a linear regressor $\boldsymbol{\Delta}\mathbf{C} : \mathbb{R}^{3V} \rightarrow \mathbb{R}^{3J}$ to learn joint center *corrections* given the body shape

$$\mathbf{c}(\tilde{\mathbf{X}}) = \bar{\mathbf{C}}\tilde{\mathbf{X}} + \boldsymbol{\Delta}\mathbf{C}\tilde{\mathbf{X}} \tag{7}$$

Instead of learning joint centers globally by pooling over all mesh vertices, we only estimate locally from those vertices skinned by the joint. This leads to considerably fewer trainable parameters going down from $3N\times 3J$ to $3N\times 3I$, with $I = 4$ in practice. We also encourage sparsity, through L_1 regularization, and also alignment of the bone directions to the template. To avoid singularities and prevent joint centers from moving outside the surface, we regularize the magnitude of center corrections $\|\boldsymbol{\Delta}\mathbf{C}\tilde{\mathbf{X}}\|_2$.

To correct skinning artifacts as a result of complex soft tissue deformation, we learn a data-driven pose-dependent corrector (PSD) $\boldsymbol{\Delta}\tilde{\mathbf{X}}(\boldsymbol{\theta})$ applied to the rest shape. We estimate a nonlinear mapping $\mathbf{D} : \mathbf{R}_i(\theta_i) - \mathbf{R}_i(\bar{\theta}_i) \in \mathbb{R}^{9J} \rightarrow$

$\boldsymbol{\Delta}\tilde{\mathbf{X}}(\boldsymbol{\theta}) \in \mathbb{R}^{3n}$. However, pose space corrections on a mesh vertex should intuitively be sourced from neighboring joints. We therefore use a fully-connected ReLU activated layer to extract a much more compact feature vector than the input (we use 32 units), from which we then linearly regress the pose space deformation. Moreover, our $\tilde{\mathbf{X}}(\boldsymbol{\theta})$ is sparse, and a joint can only generate local deformation correctives to its skinned mesh patch. Compared to the dense linear regressor in SMPL [24], our network produces similar quality deformations with considerably fewer ($17\times$ fewer) trainable parameters. We regularize the magnitude of pose space deformation to be small, preventing matching the targets by over-fitting through PSD corrections. This is implemented by a simple L_2 penalty as

$$L_p(\boldsymbol{\Delta}\tilde{\mathbf{X}}) = \|\boldsymbol{\Delta}\tilde{\mathbf{X}}(\boldsymbol{\theta})\|^2. \tag{8}$$

High-frequency local PSD is often undesirable and most likely due to overfitting. Therefore we encourage smooth pose space deformations with

$$L_s(\boldsymbol{\Delta}\tilde{\mathbf{X}}) = \sum_{i=1}^{V} \sum_{j \in \mathrm{N}(i)} \|l_{i,j}(\boldsymbol{\Delta}\tilde{\mathbf{x}}_i - \boldsymbol{\Delta}\tilde{\mathbf{x}}_j)\|^2, \tag{9}$$

where $\mathrm{N}(i)$ are the neighboring vertices to vertex i and $l_{i,j}$ are cotangent-based Laplacian weights.

Even with PSD regularizers and a reduced number of trainable weights, overfiting could still occur. Differently from SMPL or MANO [35], where pose space deformation were built specifically for only certain regions (body or hand), we construct a PSD model for the entire humanoid, trained jointly based on high-resolution body, hand and head data closeups. Consequently our body data has limited variation on hand and head motions, whereas head and hand data has no motion for the rest of the body. Hence, there is a large articulation space where all joints can move without an effect on the loss, which is undesirable. To prevent overfitting, we filter (mask) the input pose feature vector into 4 feature vectors, consisting of the head, body, left hand and right hand joints. Each feature vector will be taken into the *same* ReLU layer and we sum up the outputs before the next regressor (fig. 5). We formulate a loss

$$L_f(\boldsymbol{\Delta}\tilde{\mathbf{X}}) = \|\mathbf{F}\boldsymbol{\Delta}\tilde{\mathbf{X}} - \boldsymbol{\Delta}\tilde{\mathbf{X}}\|^2, \tag{10}$$

that enforces PSDs outside masked regions to be small, thus biasing the correctives produced by the network towards limited global impact. However, deformations of shared surface regions corresponding to areas at the interface between the head, hand, and the rest of body, are learnt from all relevant data.

To estimate skinning weights, at the end of the pipeline, we create a linear blending layer which, given poses $\boldsymbol{\theta}$ and pose-corrected rest shape with facial expression $\tilde{\mathbf{X}} + \Delta\tilde{\mathbf{X}} + \Delta\tilde{\mathbf{X}}^f$, outputs a posed mesh (2) controlled by trainable skinning weight parameters $\boldsymbol{\omega}$. Each skinned vertex is maximally influenced by $I = 4$ joints in the template. We also include a prior on $\boldsymbol{\omega}$ based on the initial artist painted values $\bar{\boldsymbol{\omega}}$, ensure that weights are spatially smooth, and per-vertex weight components are non-negative and normalized

$$L_\omega^s(\boldsymbol{\omega}) = \sum_{i=1}^{V} \sum_{j \in \mathrm{N}(i)} \sum_{k=1}^{I} \|l_{i,j}(\omega_{i,k} - \omega_{j,k})\|^2$$

$$L_\omega^i(\boldsymbol{\omega}) = \sum_{i=1}^{V} \sum_{k=1}^{I} \|\omega_{i,k} - \bar{\omega}_{i,k}\|^2$$

$$s.t. \quad \sum_{k=1}^{I} \omega_{i,k} = 1, \quad \omega_{i,k} \geq 0. \tag{11}$$

We also weakly regularize the final skinned mesh $\mathbf{X}$ to be smooth by adding

$$L_m(\mathbf{X}) = \sum_{i=1}^{V} \sum_{j \in \mathrm{N}(i)} \|l_{i,j}(\mathbf{x}_i - \mathbf{x}_j)\|^2. \tag{12}$$

Pose Estimator. Given body shape estimates and current skinning parameters, we re-optimize poses $\boldsymbol{\theta}$ over the training set. To limit the search space, enforce consistency, and avoid unnatural local minimal, we leverage anatomical joint angle limits available with our anthropometric skeleton. The problem can be efficiently solved using an L-BFGS solver with box constraints, and gradients evaluated by (*e.g.* TensorFlow's) automatic differentiation.

Table 1. Registration error for GHUM and GHUML, on Caesar and GHS3D, with detail for faces, hands, and the rest of the body.

	ICP error (mm)		Chamfer distance (mm)	
Dataset	GHUM	GHUML	GHUM	GHUML
Caesar	0.265	0.465	19.13	31.84
body	0.371	0.725	20.76	33.64
head	0.442	0.519	10.12	12.38
hand	0.164	0.423	14.88	22.01

4. Experiments

Datasets. In addition to Caesar, which contains diverse body and face shapes ($4,329$ subjects), we also use multiple proprietary systems operating at 60Hz to capture 48 subjects (24 females and 24 males) with 55 body poses, 60 hand poses and 40 motion sequences of facial expressions.[1] The subjects have a BMI range from 17.5 to 39.2, height from 148 cm to 192 cm and are aged from 21 to 56. For all multi-pose data, we use 4 subjects for evaluation, and 4 subjects for testing, including a freestyle motion sequence containing poses generally not in the training set. Each face

[1] Subject data was collected in a lab setting with informed consent.

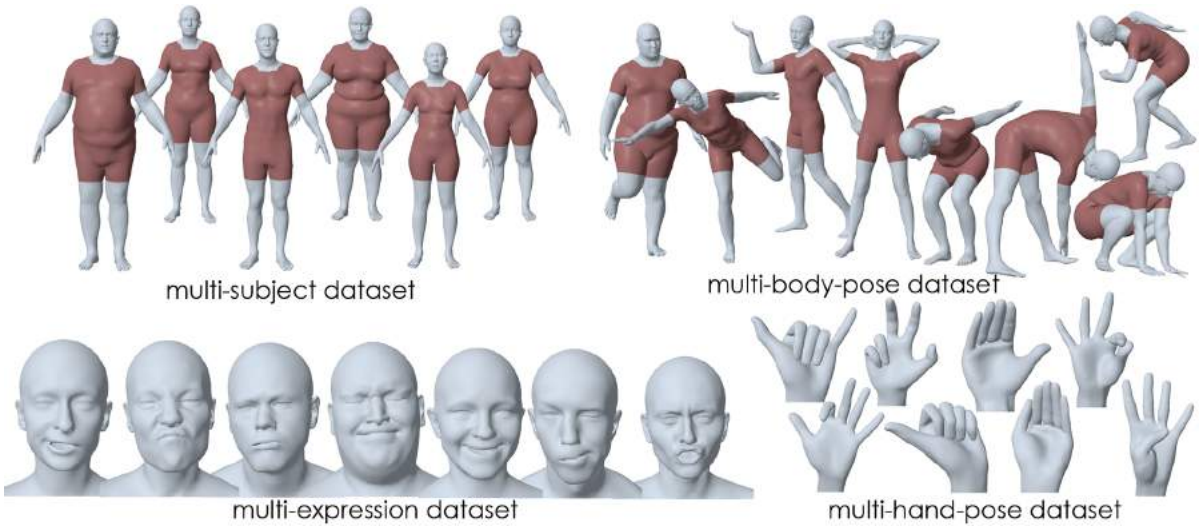

Figure 6. Sample registrations from Caesar (top left) as well as our GHS3D. Notice the quality of registration that captures subtle facial detail, and the soft tissue deformation of the other body parts as a result of articulation.

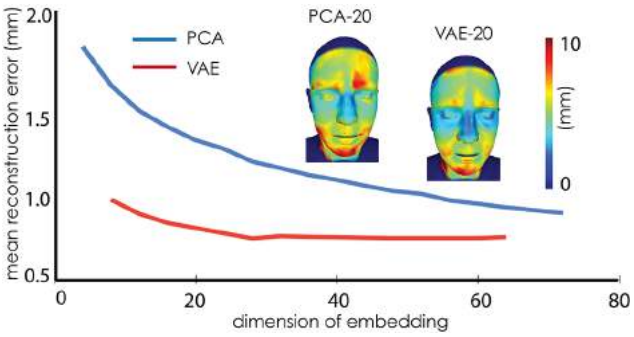

Figure 7. Analysis of VAE and PCA models illustrate the advantage of non-linear representations in the low-dimensional regime.

capture sequence starts from a neutral face to a designated facial expression and each sequence lasts about 2s. Registration samples from the data are shown in fig. 6.

Registration. In Table 1, we report registration to the point clouds using ICP and the (extended) Chamfer distance [19]. ICP error is measured as point-to-plane distance to the nearest registered mesh facet, whereas Chamfer distances are estimated point to point, bidirectionally. Registration has low error and preserves local point cloud detail (fig. 6).

Model Evaluation. We build both a full resolution and a low-resolution human model (GHUM and GHUML) using our end-to-end pipeline. Both models share the same set of skeleton joints but have $10,168$ vs. $3,194$ mesh vertices (with $1,932$ vs. 585 vertices for facial expressions). For both models, we evaluate the mean vertex-based Euclidean distances of meshes $\mathbf{X}$ to registrations $\mathbf{X}^*$ on testing data. Numbers are reported in Table 2 and visualizations are shown in figs. 1, 4, and 9. We compare the outputs of both models to registered meshes under their corresponding topology. Both models can closely represent a diversity of body shapes (modeled as VAEs, fig. 4), produce natural facial expressions (represented as facial VAEs), *c.f*. fig. 5 in Sup. Mat., and pose smoothly and naturally without noticeable skinning artifacts for a variety of shapes and poses (resulting from optimized skinning parameters, *c.f*. fig. 1).

Table 2. Mean vertex-based Euclidean reconstruction error from registration (mm).

Dataset	Caesar	GHS3D $\rightarrow$ body	face	hand
GHUM	2.81	5.21	2.96	2.22
GHUML	3.27	6.32	3.28	2.81

GHUM vs GHUML. The low resolution model preserves the global features of the body shape and correctly skins the body and facial motion. Compared with GHUM, we observe that GHUML loses some detail for lip deformations, muscle bulges at the arms and fingers, and wrinkles due to fat tissue. Performance-wise, GHUML is $2.0\times$ faster, in feed-forward evaluation mode, than GHUM.

VAE Evaluation. For body shape, our VAE supports both a 16-dim and a 64-dim latent representation where the former has $1.72\times$ higher reconstruction error (our report is based on a 16-dim representation). We use a 20-dim embedding for our facial expression VAE. Fig. 7 shows the reconstruction error of facial expressions as a function of the latent dimension, for both VAE and PCA. The 20-dimensional VAE has a reconstruction error similar to the one that uses 96 linear PCA bases, at the cost of $1.4\times$ slower performance.

GHUM vs SMPL. In fig. 8, we evaluate the skinning quality of GHUM and SMPL, for multiple subjects and poses, total of $1,100$ scans. We have different mesh and skeleton

Figure 8. From left to right, registration, GHUM, and SMPL. GHUM produces skinning with fewer pelvis artefacts for this motion sequence (0.76 mm lower error on average).

typologies from SMPL and SMPL does not have hand and facial joints. We therefore take a captured motion sequence (all the poses, not in our training dataset) from GHS3D, and register the captured sequence with SMPL and GHUM mesh respectively. We use one-to-one point-to-plane Euclidean distance for error calculations (to avoid sensitivity to surface sliding during registration), and we only evaluate error on the body (minus face and hands) for fair comparison with SMPL. GHUM's mean reconstruction error is 4.23 mm whereas SMPL has 4.96 mm error.

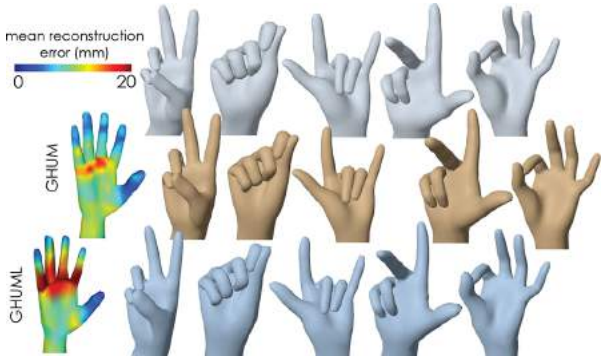

Figure 9. Evaluation and rendering as in fig.1 with emphasis on the hand reconstruction of GHUM and GHUML. Notice additional deformation detail around the flexion region of the palm preserved by GHUM over GHUML. See Sup. Mat. for facial expressions.

3D Pose and Shape Reconstruction from Monocular Images. We also illustrate image reconstruction using GHUM. The kinematic prior (for hands and the rest of the body, excluding the face) is based on normalizing flows and has been trained using Human3.6M, CMU, and GHS3D [42]. We do not use an image predictor for pose and shape, but initialize at 6 different kinematic configurations and optimize $\boldsymbol{\alpha}$ parameters under anatomical joint angle limits. As loss we use the skeleton joints reprojection error and a semantic body-part alignment *c.f*. [6, 43]. We show results in fig. 10, see Sup. Mat for more.

Application Use Cases: Our construction of GHUM/L models is motivated by the breadth of transformative, immersive 3D applications, that would become possible, including clothing virtual apparel try-on, fitness, personal well-being, health or rehabilitation, AR and VR for improved communication or collaboration, special effects, human-computer interaction or gaming, among others. In contrast, applications like visual surveillance and person identification would not be effectively supported currently, given that model's output does not provide sufficient detail or resolution for these purposes. The same is true for the creation of potentially adversely-impacting deepfakes, as an appearance model or a joint audio-visual model are not included to support photorealistic visual and voice synthesis.

Figure 10. Monocular 3D human pose and shape reconstruction with GHUM by relying on non-linear pose and shape optimization under a semantic body part alignment loss.

5. Conclusions

We have presented GHUM and GHUML(ite), two new generative 3D human shape and pose models of both moderate resolution ($10,168$ vertices) and low-resolution ($3,194$ vertices), respectively. The models are trained based on a new dataset, GHS3D, of over $60,000$ human scans, containing both full-body and closeups for faces and hands. We present a new end-to-end deep learning framework that supports – for the the first time, and based on all data simultaneously – the combined training of all model component parameters including non-linear shape spaces, pose-space deformation correctives, skeleton joint center estimators, and surface blend skinning functions. We run extensive experiments in the low-resolution and medium-resolution regime for both registration and constrained articulated 3D shape fitting and illustrate 3D pose and shape estimation from monocular images. A perhaps surprising conclusion is that appropriately trained, a low resolution nonlinear model of about $3,000$ vertices could have surprisingly good human shape representation capacity. Models will be made available for research.

Acknowledgements: We thank Elisabeta Oneata, Alin Popa, Mihai Zanfir, and Ana Padurariu for their outstanding support with data collection and processing.

References

[1] CMU graphics lab motion capture database. 2009. http://mocap.cs.cmu.edu/.

[2] Brett Allen, Brian Curless, Brian Curless, and Zoran Popović. The space of human body shapes: Reconstruction and parameterization from range scans. *ACM Trans. Graphics*, 2003.

[3] Brian Amberg, Reinhard Knothe, and Thomas Vetter. Expression invariant 3d face recognition with a morphable model. In *FG*, 2008.

[4] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. Scape: shape completion and animation of people. *ACM Trans. Graphics*, 2005.

[5] Abdallah Benzine, Bertrand Luvison, Quoc Cuong Pham, and Catherine Achard. Deep, robust and single shot 3d multi-person human pose estimation from monocular images. In *ICIP*, 2019.

[6] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it SMPL: Automatic estimation of 3d human pose and shape from a single image. In *ECCV*, 2016.

[7] Alan Brunton, Augusto Salazar, Timo Bolkart, and Stefanie Wuhrer. Comparative analysis of statistical shape spaces. *CoRR*, abs/1209.6491, 2012.

[8] Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. Facewarehouse: A 3d facial expression database for visual computing. *IEEE TVCG*, 2014.

[9] Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Three-dimensional reconstruction of human interactions. In *CVPR*, 2020.

[10] Nils Hasler, Carsten Stoll, Martin Sunkel, Bodo Rosenhahn, and Hans-Peter Seidel. A statistical model of human pose and body shape. *Computer Graphics Forum*, 2009.

[11] Nikolaos Kyriazis Iason Oikonomidis and Antonis Argyros. Efficient model-based 3d tracking of hand articulations using kinect. In *BMVC*, 2011.

[12] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3d human sensing in natural environments. *PAMI*, 2014.

[13] Haiyong Jiang, Jianfei Cai, and Jianmin Zheng. Skeleton-aware 3d human shape reconstruction from point clouds. In *ICCV*, 2019.

[14] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *CVPR*, 2018.

[15] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *CVPR*, 2018.

[16] Isinsu Katircioglu, Bugra Tekin, Mathieu Salzmann, Vincent Lepetit, and Pascal Fua. Learning latent representations of 3d human pose with deep neural networks. *IJCV*, 2018.

[17] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[18] 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*, 2019.

[19] Chun-Liang Li, Tomas Simon, Jason Saragih, Barnabas Poczos, and Yaser Sheikh. Lbs autoencoder: Self-supervised fitting of articulated meshes to point clouds. In *CVPR*, 2019.

[20] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In *CVPR*, 2019.

[21] Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graphics*, 2017.

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

[23] Yebin Liu, Carsten Stoll, Juergen Gall, Hans-Peter Seidel, and Christian Theobalt. Markerless motion capture of interacting characters using multi-view image segmentation. In *CVPR*, 2011.

[24] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics*, 2015.

[25] Diogo C Luvizon, David Picard, and Hedi Tabia. 2d/3d pose estimation and action recognition using multitask deep learning. In *CVPR*, 2018.

[26] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *ICCV*, 2019.

[27] Julieta Martinez, Rayat Hossain, Javier Romero, and James J. Little. A simple yet effective baseline for 3d human pose estimation. In *ICCV*, 2017.

[28] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3D pose estimation from monocular RGB. In *3DV*, 2018.

[29] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. Vnect: Real-time 3d human pose estimation with a single rgb camera. *ACM Trans. Graphics*, 2017.

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

[31] 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 *CVPR*, 2019.

[32] Stylianos Ploumpis, Haoyang Wang, Nick Pears, William A. P. Smith, and Stefanos Zafeiriou. Combining 3d morphable models: A large scale face-and-head model. In *CVPR*, 2019.

[33] Alin-Ionut Popa, Mihai Zanfir, and Cristian Sminchisescu. Deep multitask architecture for integrated 2d and 3d human sensing. In *CVPR*, 2017.

[34] Helge Rhodin, Nadia Robertini, Dan Casas, Christian Richardt, Hans-Peter Seidel, and Christian Theobalt. Gen-

eral automatic human shape and motion capture using volumetric contour cues. In *ECCV*, 2016.

[35] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Trans. Graphics*, 2017.

[36] Kai Su, Dongdong Yu, Zhenqi Xu, Xin Geng, and Changhu Wang. Multi-person pose estimation with enhanced channel-wise and spatial information. In *CVPR*, 2019.

[37] Jonathan Taylor, Lucas Bordeaux, Thomas Cashman, Bob Corish, Cem Keskin, Toby Sharp, Eduardo Soto, David Sweeney, Julien Valentin, Benjamin Luff, Arran Topalian, Erroll Wood, Sameh Khamis, Pushmeet Kohli, Shahram Izadi, Richard Banks, Andrew Fitzgibbon, and Jamie Shotton. Efficient and precise interactive hand tracking through joint, continuous optimization of pose and correspondences. *ACM Trans. Graphics*, 2016.

[38] Dimitrios Tzionas, Luca Ballan, Abhilash Srikantha, Pablo Aponte, Marc Pollefeys, and Juergen Gall. Capturing hands in action using discriminative salient points and physics simulation. *IJCV*, 2016.

[39] Fei Yang, Jue Wang, Eli Shechtman, Lubomir Bourdev, and Dimitri Metaxas. Expression flow for 3d-aware face component transfer. In *ACM Trans. Graphics*, 2011.

[40] Wei Yang, Wanli Ouyang, Xiaolong Wang, Jimmy Ren, Hongsheng Li, and Xiaogang Wang. 3d human pose estimation in the wild by adversarial learning. In *CVPR*, 2018.

[41] Yusuke Yoshiyasu, Wan-Chun Ma, Eiichi Yoshida, and Fumio Kanehiro. As-conformal-as-possible surface registration. *Computer Graphics Forum*, 2014.

[42] Andrei Zanfir, Eduard Gabriel Bazavan, Hongyi Xu, Bill Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Weakly supervised 3d human pose and shape reconstruction with normalizing flows. *arXiv preprint arXiv:2003.10350*, 2020.

[43] Andrei Zanfir, Elisabeta Marinoiu, and Cristian Sminchisescu. Monocular 3d pose and shape estimation of multiple people in natural scenes - the importance of multiple scene constraints. In *CVPR*, 2018.

[44] Andrei Zanfir, Elisabeta Marinoiu, Mihai Zanfir, Alin-Ionut Popa, and Cristian Sminchisescu. Deep network for the integrated 3d sensing of multiple people in natural images. In *NIPS*, 2018.

GHUM & GHUML: Generative 3D Human Shape and Articulated Pose Models

Supplementary Material

Hongyi Xu Eduard Gabriel Bazavan Andrei Zanfir
William Freeman Rahul Sukthankar Cristian Sminchisescu
Google Research
{hongyixu, egbazavan, andreiz, wfreeman, sukthankar, sminchisescu}@google.com

In this supplementary material we describe our scalable registration process that combines automatic landmark detection with triangulation as well as Iterative Closest Point (ICP) optimization under as conformal as possible (ACAP) surface priors. We also provide additional illustration and quantitative insight for the optimization of skinning parameters, interpolation in the latent VAE space for face expressions and body shapes, as well as an evaluation of reconstruction accuracy for hands. We also illustrate the retargeting process of Human3.6M to our GHUM models and describe the optimization criteria.

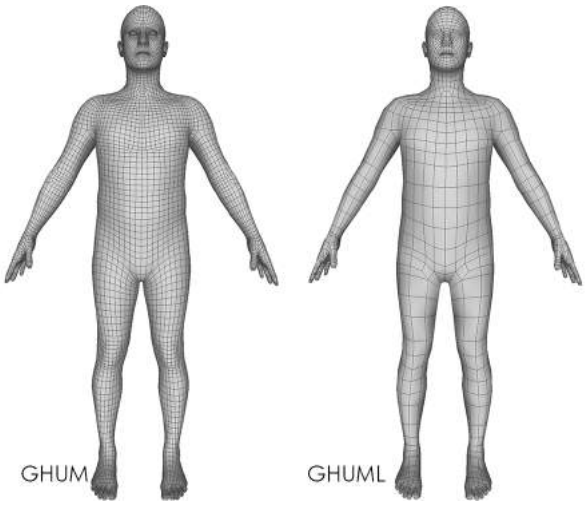

Figure 1. Template quad-mesh for GHUM ($10,168$ vertices and $10,166$ faces) and GHUML ($3,194$ vertices and $3,190$ faces).

1. Automatic Registration

Key to scalability in model learning is automating the registration process. Previous work relied on laborious manual annotation processes which makes data processing laborious and possibly error prone. Here we focus on large-scale automation, which can be viewed as an additional contribution of this work. We start by registering our GHUM/GHUML models to unstructured data collected for the full body, as well as to closeup face and hand point cloud data, available in our GHS3D dataset as well as from Caesar. We first register a full body rest shape to all the neutral scans (including closeup scans) of a subject (section 1.1), and then deform and articulate the rest shape as conformly as possible (ACAP) [8] in order to match the data (§ 1.2).

The registration process is formulated as optimization of

$$\min_{\mathbf{A},\mathbf{b},\boldsymbol{\theta}} E_d(\mathbf{X}^*(\bar{\mathbf{X}},\boldsymbol{\theta}),\mathbf{Y},\mathbf{L}) + E_s(\bar{\mathbf{X}}(\mathbf{A},\mathbf{b})), \qquad (1)$$

where E_d and E_s are the data and shape prior terms respectively and each vertex i has 12 local affine transformation DOFs $[\mathbf{A}_i \in \mathbb{R}^9, \mathbf{b}_i \in \mathbb{R}^3]$ that deform the template mesh vertex $\mathbf{x}_i^0$ into $\bar{\mathbf{x}}_i = \mathbf{A}_i\mathbf{x}_i^0 + \mathbf{b}_i$; $\mathbf{X}^*$ is a posed mesh of $\bar{\mathbf{X}}$ with $\boldsymbol{\theta}$. During the registration phase, the articulation uses the artist-defined skinning where artifacts are compensated by deforming the rest shape.

The data term E_d consists of point-to-plane residuals evaluated on each scanning (data) point via iterative closest point (ICP) and point-to-point residuals evaluated on a set of automatically labelled landmarks (section 1.3). Specifically, we have

$$E_d(\mathbf{X}^*,\mathbf{Y},\mathbf{L}) = \frac{1}{2}\sum_{i=1}^{P} \|\mathbf{n}_i\mathbf{n}_i^\top(\mathbf{B}_i(\mathbf{X}^*) - \mathbf{Y}_i)\|^2 + \frac{1}{2}\sum_{i=1}^{Q} \|\mathbf{D}_i(\mathbf{X}^*) - \mathbf{L}_i\|^2, \qquad (2)$$

where $\mathbf{B}_i$ are barycentric coordinates for the closest vertices to scan points $\mathbf{Y}_i$ on the current deformed mesh $\mathbf{X}^*$, and $\mathbf{n}_i\mathbf{n}_i^\top$ accounts for the point-to-plane weighting with vertex normal $\mathbf{n}_i$. Our Q landmarks are either defined on surface and/or body joints. When $\mathbf{L}_i$ is on the surface, $\mathbf{D}_i$ is the same as $\mathbf{B}_i$; and otherwise

$$\mathbf{D}_i(\mathbf{X}^*) = \mathbf{T}_i(\boldsymbol{\theta},\mathbf{c})\mathbf{T}_i(\bar{\boldsymbol{\theta}},\mathbf{c})^{-1}\mathbf{c}_i, \qquad (3)$$

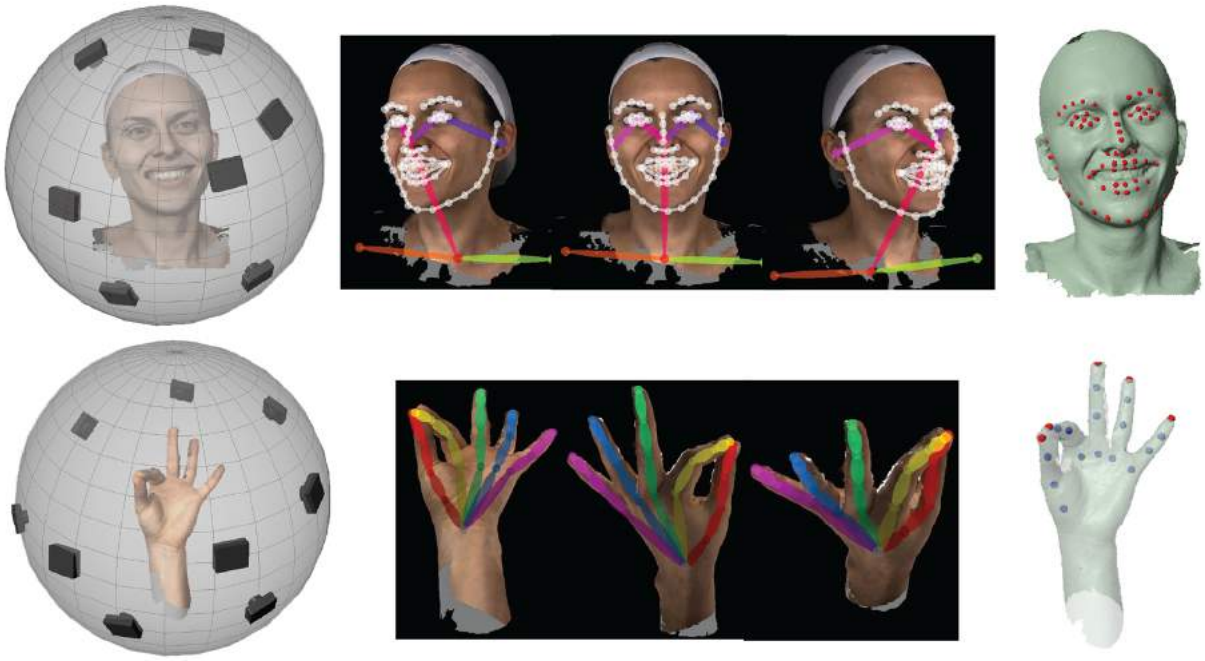

Figure 2. Automatic landmark detection. Left: we render images of the reconstructed 3d colored point cloud, with camera moving along a spiral trajectory on a sphere centered at the scan. Middle: we detect 2D landmarks (skeleton or contour structures highlighted) in the rendered images. Right: the 3D triangulated landmarks lifted to 3D point clouds (red: surface landmarks; blue: body joint landmarks).

where $\mathbf{c}_i = (\bar{\mathbf{C}}\bar{\mathbf{X}})_i$ is the center position of the corresponding joint at rest and it should get close to $\mathbf{L}_i$ when transformed with the pose $\boldsymbol{\theta}$.

We use an ACAP energy as our shape model E_s

$$E_s(\bar{\mathbf{X}}(\mathbf{A},\mathbf{b})) = \sum_{i=1}^{V}\Big(\sum_{j\in \mathrm{N}(i)} \|\mathbf{A}_i - \mathbf{A}_j\|^2 \tag{4}$$

$$+ \sum_{j\in \mathrm{N}(i)} \|\mathbf{A}_i(\bar{\mathbf{X}}_i - \bar{\mathbf{X}}_j) - (\bar{\mathbf{X}}_i - \bar{\mathbf{X}}_j)\|^2 \tag{5}$$

$$+ \sum_{j,k=1}^{3} \left[(\mathbf{A}_{i,j}^\top \mathbf{A}_{i,k})^2 + (\mathbf{A}_{i,j}^\top \mathbf{A}_{i,j} - \mathbf{A}_{i,k}^\top \mathbf{A}_{i,k})^2)\right] \tag{6}$$

where $\mathbf{A}_{i,j}$ is the jth column of $\mathbf{A}_i$, and $k = (j+1)\%3$. Eq. (6) encourages the rest shape deformation from template to be as conformal as possible whereas (4),(5) regularizes the linear transformations to be continuous and consistent between all pairs of neighboring vertices. Please refer to [8] for more details.

1.1. Full Body Shape Registration

To estimate a full body shape for a subject, we register the neutral body pose but also leverage shape detail for the head and hands if closeup scans are available. As the neutral pose defined by the template character will not necessarily remained aligned to the neutral scan during optimization, we will need to also co-optimize the body pose. To fuse the registered mesh for neutral body, head and hand of a subject, we adjust (1) as

$$\min_{\mathbf{A},\mathbf{b},\{\boldsymbol{\theta}^d\}} \sum_{d=b,f,h} E_d(\mathbf{X}^{*d}(\bar{\mathbf{X}},\boldsymbol{\theta}^d),\mathbf{Y}^d,\mathbf{L}^d) + E_s(\bar{\mathbf{X}}(\mathbf{A},\mathbf{b})), \tag{7}$$

where we estimate a single rest shape $\bar{\mathbf{X}}$ for the neutral poses. Solving (7) alone could lead to undesirable solutions that match targets by deforming the rest shape instead of articulating it by means of the kinematic constraints of the skeleton. To eliminate the nullspace, we regularize the problem by biasing the bone directions of the rest pose to align with the template skeleton (*i.e.* the 'A' pose) in the world coordinate as

$$E_o(\mathbf{c}(\bar{\mathbf{X}})) = \frac{1}{J}\sum_{i=1}^{J}(\|(\mathbf{c}_i - \mathbf{c}_{p(i)})\mathbf{t}_{i,a}\|^2 + \|(\mathbf{c}_i - \mathbf{c}_{p(i)})\mathbf{t}_{i,b}\|^2) \tag{8}$$

where $\mathbf{t}_{i,a}, \mathbf{t}_{i,b}$ are the two unit vectors orthogonal to the bone direction of the template character estimated from joint i to its parent $p(i)$. Hand and head closeup scans do not have point to constraint their complementary body (mesh) regions, hence we add a weak L2-regularizer on the poses $\boldsymbol{\theta}$ so that the rest would be close to neutral for those regions.

Additionally, our solver is augmented with a set of hard constraints. First, to remove the ambiguity due to rigid translation, we enforce the root joint to stay at the same location as the template root joint (origin of the world coordinate in our case). We also enforce symmetry for the

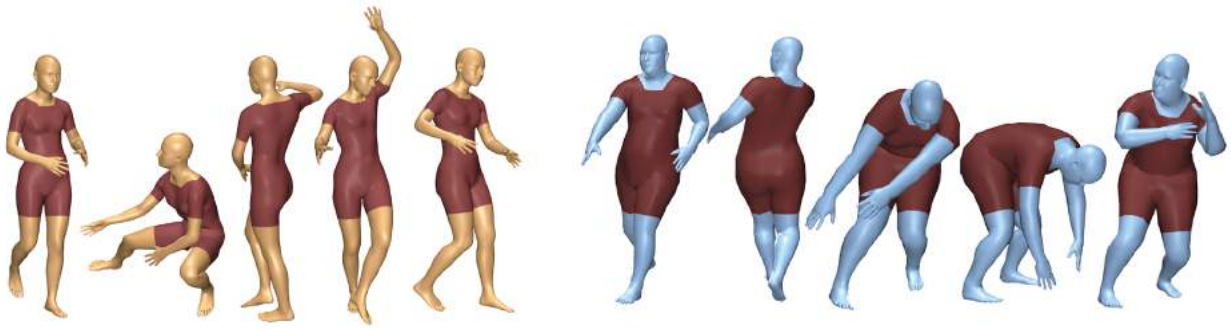

Figure 3. Retargeted CMU and Human3.6M motion sequence for GHUM and GHUML. Notice good posing without skinning artefacts for subjects with both low and high BMI.

rest-shape joint locations via $\mathbf{c}_i = \mathbf{S}(\mathbf{c}_{m(i)})$, where $m(i)$ is the mirrored joint for i and $\mathbf{S}$ is an operator to mirror the location. The hard constraints are enforced by means of an augmented Lagrangian method. Results of this process are shown in fig. 3 in the main paper.

1.2. Posed Scan Registration

After estimating the full body shape with neutral-pose scans, we can register all the point cloud data individually by minimizing (1). At this stage, we still allow the rest shape to deform such that facial expressions that cannot be explained by d.o.f.s associated to facial articulation can be registered, and skinning artifacts can be compensated for. However, we add a L2 regularizer to the rest shape in order to be close to $\bar{\mathbf{X}}$ estimated in §1.1. This ensures the deformed rest shape is at neutral pose, and corresponds to the same subject for partial or noisy point clouds.

Note that estimates $\boldsymbol{\theta}, \bar{\mathbf{X}}$ resulting from our articulated ACAP optimization (1) are not the input fed to our training pipeline. After the registration is completed for all scans, we use our pose predictor (section 3.3 in the main paper) to re-estimate the pose $\boldsymbol{\theta}$ using $\mathbf{X}^*$ as targets and with $\bar{\mathbf{X}}$ estimated at neutral poses, as rest shape.

1.3. Automatic Landmark Triangulation

For registration, we automatically detect both surface landmarks and joint landmarks on point clouds. These are critical in order to provide long range model(mesh)-to-point cloud correspondence constraints, which, in addition to dense ICP losses, would make the registration scalable and robust.

Given a point cloud reconstruction with texture, obtained using our 3d capture system, we first render a set of images corresponding to different camera observation viewpoints. The camera is placed on a sphere centered at the scan, oriented to the scan, and moves along a spiral trajectory, to observe and render the scan from different viewpoints. We sample 3 different radii for the sphere for views of different level of detail. We then run automatic detectors to identify 2D landmarks $\mathbf{h}$ in rendered images. Once such estimates are available, we can triangulate to obtain the 3D landmarks $\mathbf{L}_i$ via

$$\min_{\mathbf{L}} \sum_{j=1}^{G} \sum_{i=1}^{Q} w_{i,j} \|\mathbf{P}_j \mathbf{L}_i - \mathbf{h}_{i,j}\|^2, \tag{9}$$

where G is the number of rendered images, $\mathbf{P}_j$ is the camera projection operator for rendered image j and $w_{i,j}$ are 2D landmark detection confidences. We illustrate our automatic landmark detection and triangulation process in fig. 2.

2. Kinematic Prior

In this section we briefly describe the construction of our kinematics prior which consists of two components: mapping existing motion capture data to GHUM/GHUML's skeleton representation and building a prior/regularizer based on that data, in our case using normalizing flow.

2.1. Motion Retargetting

In order to build kinematic priors that capture the statistics of human body pose and motion, we need to retarget our model to existing motion capture datasets. We use marker data (from Human3.6M[4] and CMU [1]) to fit GHUM and obtain valid 3d joint angle configurations. We manually annotate the markers on GHUM/GHUML so that positions correspond to the particular marker sets of CMU and Human3.6M. We process each motion sequence independently, and assume one shape parameters per sequence, and one pose per frame. The objective function is set as the MSE between the 3D mocap markers and the model predictions (N.B. those would require both a kinematic and a shape configuration). We start by initializing the shape parameters by an average of initial fits, and rely on coordinate-descent, by alternating the following steps, for all frames simultaneously, to convergence

1. Update the pose of the model $\boldsymbol{\theta}$
2. Update the shape of the model $\boldsymbol{\beta}^b$, using the latent body VAE representation

Figure 4. Sampling *both the shape and the kinematic parameters* $\boldsymbol{\alpha} = (\boldsymbol{\theta}, \boldsymbol{\beta})$ of the model, *for all body components*, including facial expressions and articulation of the hands.

3. Update model's marker positions w.r.t. to the mesh surface, *i.e*. markers need to lay on the triangulated mesh

The result consists of 2.2 million retargeted marker frames from H3.6M and 2.8 million from CMU. In fig. 3, we pose our models (GHUM and GHUML) using motion capture sequences retargetted from H36M [4] and CMU. We observe natural posing and body shape deformation of subjects with very different body shapes.

2.2. Normalizing Flow Kinematic Prior

To build a kinematic prior for the hands and the rest of the body $\boldsymbol{\theta}$, we use normalizing flow representations. Please see our accompanying paper [9] for details. A normalizing flow[7, 2, 3, 5] is a sequence of invertible transformations applied to an original input distribution. The end-result is a warped space with a potentially simple and tractable density function, e.g. $\mathbf{t} \sim \mathcal{N}(0; \mathbf{I})$. We consider $\boldsymbol{\theta} \sim p^*(\boldsymbol{\theta})$ sampled from an unknown distribution. One way to learn it is to use a dataset $\mathcal{D}$ and maximize data log-likelihood with respect to a parametric model $p_{\boldsymbol{\phi}}(\boldsymbol{\theta})$. In practice we build two normalizing flow models, one for the hands and one for the rest of the body except the face. For the hands we used $2,400$ poses collected from 40 people, whereas for the rest of the body (excluding the face) we use $500,000$ frames representing a factor of 10 sub-sampling of 2.8 million frames from CMU and 2.2 million frames from Human3.6M. For the other skeleton variables of the face, e.g. jaw or eyeballs, we used anatomical angle limit constraints available with our skeleton structure. Most image-based 3d reconstructions we show rely, for pose, on optimization in the latent kinematic space but for joints that we didn't have data for, e.g. the jaw, we optimize in the ambient (original joint angle) space. This takes advantage of both worlds – data constraints for those skeleton component where we do have motion capture, and anatomical joint angle limit constraints for those joints where we do not.

Architecture The normalizing flow priors for hands and the rest of the body except face (called here the body in a slight abuse of the term) have the same architecture, but data dimensionality is different. The body prior is based on 22 joints (48 Euler angles), each with a variable amount of sine and cosine of its angular DoFs. For the body, we use a FC96-PReLU-FC96-PreLU-FC96-PreLU-FC96-PReLU-FC96 architecture. Here, 'FC96' denotes a fully-connected layer with 96 neurons, while PreLU is the parametric ReLU activation function. In total, this network has $\approx 46,500$ trainable parameters. For the hands body prior, the network architecture has the same structure, except that each fully-connected layer will have 50 neurons (this corresponds to the sine and cosine of the 25 Euler angles for the 15 joints that articulate the hand), with $\approx 12,750$ trainable parameters.

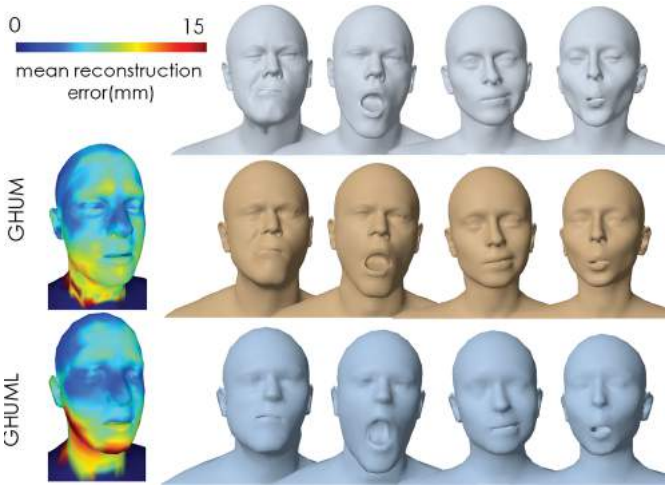

Figure 5. Evaluation of head reconstruction with facial expressions. Left: Per-vertex Euclidean distance error to the registration for GHUM and GHUML. Right: (top to bottom, registrations, GHUM and GHUML). Please note the head and neck motion besides the facial expressions.

3. Model Evaluation

Templates. In fig. 1, we visualize our quad-mesh templates for GHUM and GHUML in wireframe. For both templates, the mesh vertices are regularly distributed. They share a same skeleton, an atomically parameterized, with 63 joints. Anatomical joint angle limits are available and used during different optimization processes, e.g. during registration, end-to-end model learning and for 3D reconstruction from images.

Training Parameters. To train our pipeline end-to-end, we use batch size 64 and a learning rate of $1e^{-5}$. Besides the latent layer, the encoder and decoder of our VAEs take 3 PReLU layers with a size of 128 (for body shape VAE, the first of the encoder and last layer of decoder has the equal size to the latent code). PSD regressor has 1 ReLU layer with 32, followed with a linear regressor with strong L1 sparsity regularizer. We train $3,200$ epoches and we update poses with our pose estimator after each 600 epochs.

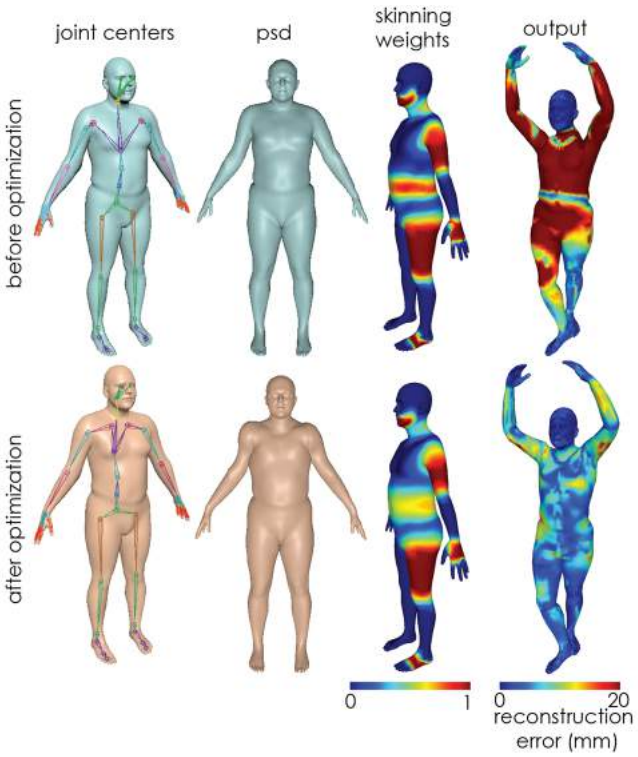

Figure 6. Effect of optimized skinning parameters. For clarity, we only visualize skinning weights for body parts associated to joints like the lower spine, neck, jaw, upper-arm and the thigh.

Optimization. Our end-to-end training significantly improves skinning (see fig. 6). Compared with the initial guess of the joint centers $\bar{\mathbf{C}}\bar{\mathbf{X}}$, we observe that optimization moves the clavicle joints much closer and lifts up the thigh, neck and head joints. Pose space deformation adds smooth bulges to the shoulder in order to compensate volume loss when humans raise their arms. The PSD model additionally learns from data to perform soft tissue deformation and produces more natural skinning (fig. 7). The optimized skinning weights are smoother overall after learning, and the skinned model has significantly lower reconstruction error to the registered mesh.

Head Reconstruction. In addition to our body and hand reconstructions shown in fig. 1, 9 of the main paper, we here also illustrate hand reconstruction results evaluated on the test set of GHS3D, *c.f*. fig. 5. Both GHUM and GHUML are able to accurately capture the details in the registration, as a result of optimizing the skinning parameters in an end-to-end training pipeline.

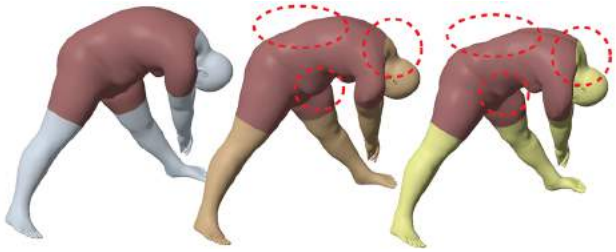

Figure 7. Soft tissue deformation learnt from pose space deformation. From left to right, registration, GHUM with and without pose-space deformation. Notice better modeling for the former.

Sampling. In fig. 4, we sample our body shape VAE, face expression VAE and kinematic priors all together for the full body. With normal distribution of sampling, our VAEs produces natural body shape and face expressions, and with our model, it shows natural deformations with sampled body and hand poses from our kinematics prior. In fig. 8, we linearly interpolate a few samples in the latent space of our VAEs and it has demonstrated smooth and natural transitions.

3D Pose and Shape from Monocular Images. While the main objective of this paper is to present an end-to-end statistical full human body model construction, and not 3D reconstruction from images (which is a broad topic in itself), we provide illustration of how the model can be used to understand the full body pose and shape, including facial expression and hand gestures. In fig. 11, we illustrate the reconstruction of the pose and shape parameters of GHUM from monocular images taken from MSCoco[6]. We optimize by initializing at the 0 mean latent space pose and kinematic configuration $\boldsymbol{\alpha} = \mathbf{0}$ and use both a body joint re-projection error and a semantic body part alignment loss[9] which relies on a deep neural network to provide, as for registration, body joint detections and the semantic segmentation of body parts. Please refer to the *Supplementary Video* for additional 3D reconstruction

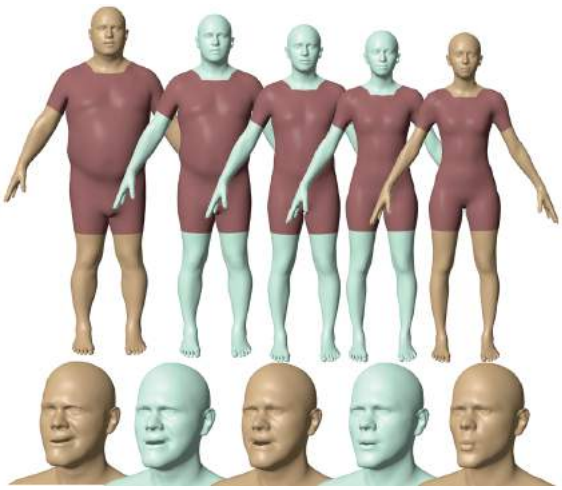

Figure 8. Decoded body and face expression with interpolated latent code. Brown: sample from datasets; green: interpolations.

results. Note that hand reconstructions are challenging on images with low-resolution since the 2d keypoint detections are not always accurate. We show reconstructions on higher-resolution images in fig. 9, where it is clear that body, hand poses and facial expressions are reconstructed, simultaneously, with good accuracy. We also demonstrate close-up facial reconstruction in fig. 10

Figure 9. Image reconstruction with full-body poses and facial expressions.

Figure 10. Facial expressions reconstruction from close-up head images.

Figure 11. 3D reconstructions of GHUML's pose and shape from monocular images in MSCoco.

References

[1] CMU graphics lab motion capture database. 2009. http://mocap.cs.cmu.edu/.

[2] Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014.

[3] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

[4] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3d human sensing in natural environments. *PAMI*, 2014.

[5] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *NeurIPS*, 2018.

[6] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014.

[7] Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. *arXiv preprint arXiv:1505.05770*, 2015.

[8] Yusuke Yoshiyasu, Wan-Chun Ma, Eiichi Yoshida, and Fumio Kanehiro. As-conformal-as-possible surface registration. *Computer Graphics Forum*, 2014.

[9] Andrei Zanfir, Eduard Gabriel Bazavan, Hongyi Xu, Bill Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Weakly supervised 3d human pose and shape reconstruction with normalizing flows. *arXiv preprint arXiv:2003.10350*, 2020.