imGHUM: Implicit Generative Models of 3D Human Shape and Articulated Pose

Thiemo Alldieck* Hongyi Xu* Cristian Sminchisescu

Google Research

{alldieck,hongyixu,sminchisescu}@google.com

Abstract

We present **imGHUM***, the first holistic generative model of 3D human shape and articulated pose, represented as a signed distance function. In contrast to prior work, we model the full human body implicitly as a function zero-level-set and without the use of an explicit template mesh. We propose a novel network architecture and a learning paradigm, which make it possible to learn a detailed implicit generative model of human pose, shape, and semantics, on par with state-of-the-art mesh-based models. Our model features desired detail for human models, such as articulated pose including hand motion and facial expressions, a broad spectrum of shape variations, and can be queried at arbitrary resolutions and spatial locations. Additionally, our model has attached spatial semantics making it straightforward to establish correspondences between different shape instances, thus enabling applications that are difficult to tackle using classical implicit representations. In extensive experiments, we demonstrate the model accuracy and its applicability to current research problems.*

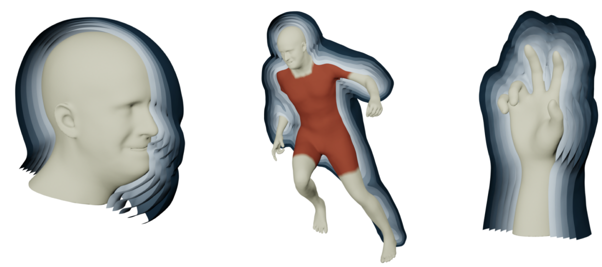

Figure 1. **imGHUM** is the first parametric full human body model represented as an implicit signed distance function. **imGHUM** successfully models broad variations in pose, shape, and facial expressions. The level sets of **imGHUM** are shown in blue-scale.

1. Introduction

Mathematical models of the human body have been proven effective in a broad variety of tasks. In the last decades models of varying degrees of realism have been successfully deployed e.g. for 3D human motion analysis [46], 3D human pose and shape reconstruction [24, 52], personal avatar creation [3, 54], medical diagnosis and treatment [16], or image synthesis and video editing [53, 21]. Modern statistical body models are typically learnt from large collections of 3D scans of real people, which are used to capture the body shape variations among the human population. Dynamic scans, when available, can be used to further model how different poses affect the deformation of the muscles and the soft-tissue of the human body.

The recently released GHUM model [49] follows this methodology by describing the human body, its shape variation, articulated pose including fingers, and facial expressions as a moderate resolution mesh based on a low-dimensional, partly interpretable parameterization. In the deep learning literature GHUM and similar models [27, 23] are typically used as fixed function layers. This means that the model is parameterized with the output of a neural network or some other non-linear function, and the resulting mesh is used to compute the final function value. While this approach works well for several tasks, including, more recently, 3D reconstruction, the question of how to best represent complex 3D deformable and articulated structures is open. Recent work dealing with the 3D visual reconstruction of general objects aimed to represent the output not as meshes but as implicit functions [28, 32, 7, 29]. Such approaches thus describe surfaces by the zero-level-set (decision boundary) of a function over points in 3D-space. This has clear benefits as the output is neither constrained by a template mesh topology, nor is it discretized and thus of fixed spatial resolution.

In this work, we investigate the possibility to learn a data-driven statistical body model as an implicit function. Given the maturity of state of the art explicit human models, it is crucial that an equivalent implicit representation maintains their key, attractive properties – representing comparable variation in shape and pose and similar level of detail. This is challenging since recently-proposed implicit function networks tend to produce overly smooth shapes and fail for articulated humans [8]. We propose a novel network architecture and a learning paradigm that enable, for the first time, constructing detailed generative models of human pose, shape, and semantics, represented as Signed Distance Functions (SDFs) (see fig. 1). Our multi-part architecture focuses on difficult to model body components like

* The first two authors contributed equally.

generative pose	generative shape	generative hands	gen. expression	interpolation	signed distances	semantics	continuous rep.	
✓	✓	✓	✓	✓	✗	✓	✗	GHUM [49]
✗	✗	✗	✗	✗	✗	✗	✓	IF-Net [8]
✗	✗	✗	✗	✓	✓	✗	✓	IGR [14]
✓	✗	✗	✗	✓	✗	✗	✓	NASA [11]
✓	✓	✓	✓	✓	✓	✓	✓	**imGHUM**

Table 1. Comparison of different approaches to model human bodies. GHUM is meshed-based and thus discretized. IGR only allows for shape interpolation. NASA lacks generative capabilities for shape, hands, and facial expressions and only returns occupancy values. Only imGHUM combines all favorable properties.

hands and faces. Moreover, imGHUM models its neighborhood through distance values, enabling e.g. collision tests. Our model is not bound to a specific resolution and thus can be easily queried at arbitrary locations. Being template-free further paves the way to our ultimate goal to fairly represent diversity of mankind, including disabilities which may not be always well covered by a generic template of standard topology. Finally, in contrast to recent implicit function networks, our model additionally carries on the explicit semantics of mesh-based models. Specifically, our implicit function also returns correspondences to a canonical representation near and on its zero-level-set, enabling e.g. texturing or body part labeling. This holistic approach is novel and significantly more difficult to produce, as can be noted in prior work which could only demonstrate individual properties, *c.f.* tab. 1. Our contribution – and the key to success – stems from the novel combination of adequate, generative latent representations, network architectures with fine grained encoding, implicit losses with attached semantics, and the consistent aggregation of multi-part components. Besides extensive evaluation of 3D deformable and articulated modeling capabilities, we also demonstrate surface completion using imGHUM and give an outlook to modeling varying topologies. Our models are available for research [1].

1.1. Related Work

We review developments in 3D human body modeling, variants of implicit function networks, and applications of implicit function networks for 3D human reconstruction.

Human Body Models. Parametric human body models based on geometric primitives have been proposed early on [48] and successfully applied e.g. for human reconstruction from video data [36, 46, 45]. SCAPE [35] was one of the first realistic large scale data-driven human body models. Later variants inspired by blend skinning [17] modeled correlations between body shape and pose [15], as well as soft-tissue dynamics [37]. SMPL variants [27, 23, 33, 31] are also popular parametric body models, with linear shape spaces, compatible with standard graphics pipelines and offering good full-body representation functionality. GHUM is a recent parametric model [49] that represents the full body model using deep non-linear models – VAEs for shape and normalizing flows for pose, respectively – with various trainable parameters, learned end-to-end. In this work, we rely on GHUM to build our novel implicit model. Specifically, besides the static and dynamic 3D human scans in our dataset, we also rely on GHUM (1) to represent the latent pose and shape state of our implicit model, (2) to generate supervised training data in the form of latent pose and shape codes with associated 3D point clouds, sampled from the underlying, posed, GHUM mesh.

Implicit Function Networks (IFNs) have been proposed recently [28, 32, 7, 29]. Instead of representing shapes as meshes, voxels, or point clouds, IFNs learn a shape space as a function of a low-dimensional global shape code and a 3D point. The function either classifies the point as inside/outside [28, 7] (occupancy networks), or returns its distance to the closest surface [32] (distance functions). The global shape is then defined by the decision boundary or the zero-level-set of this function.

Despite advantages over mesh- and voxel-based representations in tasks like e.g. 3D shape reconstruction from partial views or given incomplete data, initial work has limitations. First, while the models can reliably encode rigid axis-aligned shape prototypes, they often fail for more complex shapes. Second, the reconstructions are often overly smooth, hence they lack detail. Different approaches have been presented to address these. Part-based models [13, 22, 12] assemble a global shape from smaller local models. Some methods do not rely on a global shape code but on features computed from convolving with an input observation [8, 10, 34, 9]. Others address such limitations by changing the learning methodology: tailored network initialization [4] and point sampling strategies [50], or second-order losses [14, 44] have been proposed towards this end. We found the latter to be extremely useful and rely on similar losses in this work.

IFNs for Human Reconstruction. Recently implicit functions have been explored to reconstruct humans. Huang et al. [18] learn an occupancy network that conditions on image features in a multi-view camera setup. Saito et al. [41] use features from a single image and an estimated normal image [42] together with depth values along camera rays as conditioning variables. ARCH [19] combines implicit function reconstruction and explicit mesh-based human models to represent dressed people. Karunratanakul et al. [25] propose to use SDFs to learn human grasps and augment their SDFs output with sparse regional labels. Similarly to us, Deng et al. [11] represent a pose-able human subject as a number of binary occupancy functions modeled in a kinematic structure. In contrast to our work, this framework is restricted to a single person and the body is only coarsely approximated, lacking facial features and hand detail. Also

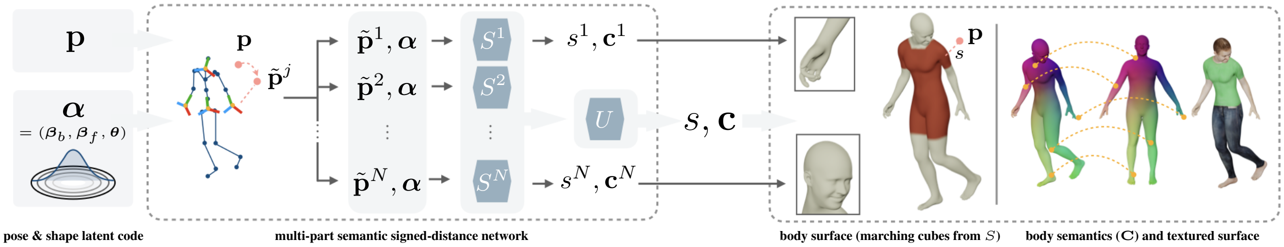

Figure 2. Overview of imGHUM. We compute the signed distance $s = S(\mathbf{p}, \boldsymbol{\alpha})$ and the semantics $\mathbf{c} = \mathbf{C}(\mathbf{p}, \boldsymbol{\alpha})$ of a spatial point $\mathbf{p}$ to the surface of an articulated human shape defined by the generative latent code $\boldsymbol{\alpha}$. Using an explicit skeleton, we transform the point $\mathbf{p}$ into the normalized coordinate frames as $\{\tilde{\mathbf{p}}^j\}$ for $N = 4$ sub-part networks, modeling body, hands, and head. Each sub-model $\{S^j\}$ represents a semantic signed-distance function. The sub-models are finally combined consistently using an MLP U to compute the outputs s and $\mathbf{c}$ for the full body. Our multi-part pipeline builds a full body model as well as sub-part models for head and hands, jointly, in a consistent training loop. On the right, we visualize the zero-level-set body surface extracted with marching cubes and the implicit correspondences to a canonical instance given by the output semantics. The semantics allows e.g. for surface coloring or texturing.

related, SCANimate [43] builds personalized avatars from multiple scans of a single person. Concurrent to our work, LEAP [30] learns an occupancy model of human shape and pose also without hand poses, expressions, or semantics. In this work we aim for a full implicit body model, featuring a large range of body shapes corresponding to diverse humans and poses, with detailed hands, and facial expressions.

2. Methodology

In this section, we describe our models and the losses used for training. We introduce two variants: a single-part model that encodes the whole human in a single network and a multi-part model. The latter constructs the full body from the output superposition of four body part networks.

Background. We rely on neural networks and implicit functions to generate 3D human shapes and articulated poses. Given a latent representation $\boldsymbol{\alpha}$ of the human shape and pose, together with an underlying probability distribution, we model the posed body as the zero iso-surface decision boundaries of Signed Distance Functions (SDFs) given by deep feed-forward neural networks. A signed distance $S(\mathbf{p}, \boldsymbol{\alpha}) \in \mathbf{R}$ is a continuous function which, given an arbitrary spatial point $\mathbf{p} \in \mathbf{R}^3$, outputs the shortest distance to the surface defined by $\boldsymbol{\alpha}$, where the sign indicates the inside (negative) or outside (positive) side w.r.t. the surface. The posed human body surface is implicitly given by $S(\cdot, \boldsymbol{\alpha}) = 0$.

GHUM [49] represents the human model as an articulated mesh $\mathbf{X}(\boldsymbol{\alpha})$. GHUM has a minimally-parameterized skeleton with $J = 63$ joints (124 Euler angle DOFs), and skinning deformations, explicitly sensitive to the pose kinematics $\boldsymbol{\theta} \in \mathbf{R}^{124}$. A kinematic prior based on normalizing flows defines the distribution of valid poses [52]. Each kinematic pose $\boldsymbol{\theta}$ represents a set of joint transformations $\mathbf{T}(\boldsymbol{\theta}, \mathbf{j}) \in \mathbf{R}^{J \times 3 \times 4}$ from the neutral to a posed state, where $\mathbf{j} \in \mathbf{R}^{J \times 3}$ are the joint centers that are dependent on the neutral body shape. The statistical body shapes are modeled using a nonlinear embedding $\boldsymbol{\beta}_b \in \mathbf{R}^{16}$. In addition to skeleton articulation, a nonlinear latent code $\boldsymbol{\beta}_f \in \mathbf{R}^{20}$ drives facial expressions. The implicit model we design here shares the same probabilistic latent representation as GHUM, $\boldsymbol{\alpha} = (\boldsymbol{\beta}_b, \boldsymbol{\beta}_f, \boldsymbol{\theta})$, but in contrast to computing an articulated mesh, we estimate a signed distance value $s = S(\mathbf{p}, \boldsymbol{\alpha})$ for each arbitrary spatial point $\mathbf{p}$.

2.1. Models and Training

Given a collection of full-body human meshes $\mathbf{Y}$, together with the corresponding GHUM encodings $\boldsymbol{\alpha} = (\boldsymbol{\beta}_b, \boldsymbol{\beta}_f, \boldsymbol{\theta})$, our goal is to learn a MLP-based SDF representation $S(\mathbf{p}, \boldsymbol{\alpha})$ so that it approximates the shortest signed distance to $\mathbf{Y}$ for any query point $\mathbf{p}$. Note that $\mathbf{Y}$ could be arbitrary meshes, such as raw human scans, mesh registrations, or samples drawn from the GHUM latent space. The zero iso-surface $S(\cdot, \boldsymbol{\alpha}) = 0$ is sought to preserve all geometric detail in $\mathbf{Y}$, including body shapes and poses, hand articulation, and facial expressions.

Single-part Network. We formulate one global neural network that decodes $S(\mathbf{p}, \boldsymbol{\alpha})$ for a given latent code $\boldsymbol{\alpha}$ and a spatial point $\mathbf{p}$. Instead of pre-computing the continuous SDFs from point samples as in DeepSDF [32], we train a MLP network $S(\mathbf{p}, \boldsymbol{\alpha}; \omega)$ with weights ω, similar in spirit to IGR [14], to output a solution to the Eikonal equation

$$\|\nabla_{\mathbf{p}} S(\mathbf{p}, \boldsymbol{\alpha}; \omega)\| = 1, \tag{1}$$

where S is a signed distance function that vanishes at the surface $\mathbf{Y}$ with gradients equal to surface normals. Mathematically, we formulate our total loss as a weighted combination of

$$L_o(\omega) = \frac{1}{|O|} \sum_{i \in O} (|S(\mathbf{p}_i, \boldsymbol{\alpha})| + \|\nabla_{\mathbf{p}_i} S(\mathbf{p}_i, \boldsymbol{\alpha}) - \mathbf{n}_i\|) \tag{2}$$

$$L_e(\omega) = \frac{1}{|F|} \sum_{i \in F} (\|\nabla_{\mathbf{p}_i} S(\mathbf{p}_i, \boldsymbol{\alpha})\| - 1)^2 \tag{3}$$

$$L_l(\omega) = \frac{1}{|F|} \sum_{i \in F} \mathrm{BCE}(l_i, \phi(kS(\mathbf{p}_i, \boldsymbol{\alpha}))), \tag{4}$$

where ϕ is the sigmoid function, O are surface samples from $\mathbf{Y}$ with normals $\mathbf{n}$, and F are off surface samples with inside/outside labels l, consisting of both uniformly sampled points within a bounding box and sampled points near the surface. The first term L_o encourages the surface samples to be on the zero-level-set and the SDF gradient to be equal to the given surface normals $\mathbf{n}_i$. The Eikonal loss L_e is derived from (1) where the SDF is differentiable everywhere with gradient norm 1. We obtain the SDF gradient $\nabla_{\mathbf{p}_i} S(\mathbf{p}_i, \boldsymbol{\alpha})$ analytically via network back-propagation. In practice, we also find it useful to include a binary cross-entropy error (BCE) loss L_l for off-the-surface samples, where k controls the sharpness of the decision boundary. We use $k = 10$ in our experiments. Our training losses only require surface samples with normals and inside/outside labels for off-surface samples. Those are much easier and faster to obtain than pre-computing ground truth SDF values.

Recent work suggests that standard coordinate-based MLP networks encounter difficulties in learning high-frequency functions, a phenomenon referred to as *spectral bias* [39, 47]. To address this limitation, inspired by [47], we therefore encode our samples using the basic Fourier mapping $\mathbf{e}_i = [\sin(2\pi\tilde{\mathbf{p}}_i), \cos(2\pi\tilde{\mathbf{p}}_i)]^\top$, where we first unpose the samples with the root rigid transformation $\mathbf{T}_0^{-1}$ and normalize them into $[0, 1]^3$ with a shared bounding box $\mathbf{B} = [\mathbf{b}_{min}, \mathbf{b}_{max}]$, as

$$\tilde{\mathbf{p}}_i = \frac{\mathbf{T}_0^{-1}(\boldsymbol{\theta}, \mathbf{j})[\mathbf{p}_i, 1]^\top - \mathbf{b}_{min}}{\mathbf{b}_{max} - \mathbf{b}_{min}}. \tag{5}$$

Note that our SDF is defined w.r.t. the original meshes $\mathbf{Y}$ and therefore we do not unpose and scale the sample normals. Also, the loss gradients are derived w.r.t. $\mathbf{p}_i$.

Multi-part Network. Our single-part network represents well the global geometric features for various human body shapes and kinematic poses. However, despite its spatial encoding, the network still has difficulties capturing facial expressions and articulated hand poses, where the SDF has local high-frequency variations. To augment geometric detail on face and hands regions, we therefore propose a multi-part network that decomposes the human body into $N = 4$ local regions, i.e. the head, left and right hand, and the remaining body, respectively. This significantly reduces spectral frequency variations within each local region allowing the specialized single-part networks to capture local geometric detail. A consistent full-body SDF $S(\mathbf{p}, \boldsymbol{\alpha})$ is composed from the local single-part SDF network outputs $s^j = S^j(\mathbf{p}, \boldsymbol{\alpha}), j \in \{1, \dots, N\}$.

We follow the training protocol described in §2.1 for each local sub-part network with surface and off-surface samples within a bounding box $\mathbf{B}^j$ defined for each part. Note that we use the neck and wrist joints as the the root transformation for the head and hands respectively. In GHUM, the joint centers $\mathbf{j}$ are obtained as a function given the neutral body shapes $\bar{\mathbf{X}}(\boldsymbol{\beta}_b)$. However, $\bar{\mathbf{X}}$ is not explicitly presented in our implicit representation. Therefore, we build a nonlinear joint regressor from $\boldsymbol{\beta}_b$ to $\mathbf{j}$, which is trained, supervised, using GHUM's latent space sampling.

In order to fuse the local SDFs into a consistent full-body SDF, while at the same time preserving local detail, we merge the last hidden layers of the local networks using an additional light-weight MLP U. To train the combined network, a sample point $\mathbf{p}_i$, defined for the full body, is transformed into the N local coordinate frames using $\mathbf{T}_0^j$ and then passed to the single-part local networks, see fig. 2. The union SDF MLP then aggregates the shortest distance to the full body among the local distances. We apply our losses to the union full-body SDF as well, to ensure that the output for full body satisfies the SDF property (1). Our multi-part pipeline produces sub-part models and a full-body one, trained jointly and leveraging data correlations among different body components.

Our spatial point encoding $\mathbf{e}_i$ requires all samples $\mathbf{p}$ to be inside the bounding box $\mathbf{B}$, which otherwise might result in periodic SDFs due to sinusoidal encoding. However, a point sampled from the full body is likely to be outside of a sub-part's local bounding box $\mathbf{B}^j$. Instead of clipping or projecting to the bounding box, we augment our encoding of sample $\mathbf{p}_i$ for sub-part networks S^j as $\mathbf{e}_i^j = [\sin(2\pi\tilde{\mathbf{p}}_i^j), \cos(2\pi\tilde{\mathbf{p}}_i^j), \tanh(\pi(\tilde{\mathbf{p}}_i^j - 0.5))]^\top$, where the last value indicates the relative spatial location of the sample w.r.t. the bounding box. If a point $\mathbf{p}_i$ is outside the bounding box $\mathbf{B}^j$, the union SDF MLP will learn to ignore $S^j(\mathbf{p}_i^j, \boldsymbol{\alpha})$ for the final union output.

Implicit Semantics. In contrast to explicit models like GHUM, implicit functions do not naturally come with point correspondences between different shape instances. However, many applications, such as pose tracking, texture mapping, semantic segmentation, surface landmarks, or clothing modeling, largely benefit from such correspondences. Given an arbitrary spatial point, on or near the surface $\mathbf{Y}$, i.e, $|S(\mathbf{p}_i, \boldsymbol{\alpha})| < \sigma$, we are therefore interested to interpret its *semantics*. We define the semantics as a 3D implicit function $\mathbf{C}(\mathbf{p}, \boldsymbol{\alpha}) \in \mathbf{R}^3$. Given a query point $\mathbf{p}_i$, it returns a correspondence point on a canonical GHUM mesh $\mathbf{X}(\boldsymbol{\alpha}_0)$ as

$$\mathbf{C}(\mathbf{p}_i, \boldsymbol{\alpha}) = \mathbf{w}_i \mathbf{v}_f(\boldsymbol{\alpha}_0) = \mathbf{c}_i, \quad \mathbf{p}_i^* = \mathbf{w}_i \mathbf{v}_f(\boldsymbol{\alpha}) \tag{6}$$

where $\mathbf{p}_i^*$ is the closest point of $\mathbf{p}_i$ in the GHUM mesh $\mathbf{X}(\boldsymbol{\alpha})$ with f the nearest face and $\mathbf{w}$ the barycentric weights of the vertex coordinates $\mathbf{v}_f$. In contrast to alternative semantic encodings, such as 2D texture coordinates, our semantic function $\mathbf{C}(\mathbf{p}, \boldsymbol{\alpha})$ is smooth in the spatial domain without distortion and boundary discontinuities, which favors the learning process, *c.f.* [5].

By definition, implicit SDFs return the shortest distance to the underlying implicit surface for a spatial point whereas implicit semantics associate the query point to its closest surface neighbor. Hence, we consider implicit semantics as highly correlated to SDF learning. We co-train both tasks with our augmented multi-part network (§2.1) computing both $S(\mathbf{p}, \boldsymbol{\alpha})$ and $\mathbf{C}(\mathbf{p}, \boldsymbol{\alpha})$. Semantics are trained fully supervised, using an L_1 loss for a collection of training sample points near and on the surface $\mathbf{Y}$. Due to the correlation between tasks, our network is able to predict both signed distance and semantics, without expanding its capacity.

Using trained implicit semantics, we can e.g. apply textures to arbitrary iso-surfaces at level set $|z| \leq \sigma$, reconstructed from our implicit SDF. During inference, an iso-surface mesh $S(\cdot, \boldsymbol{\alpha}) = z$ can be extracted using Marching Cubes [26]. Then for every generated vertex $\tilde{\mathbf{v}}_i$ we query its semantics $\mathbf{C}(\tilde{\mathbf{v}}_i, \boldsymbol{\alpha})$. The queried correspondence point $\mathbf{C}(\tilde{\mathbf{v}}_i, \boldsymbol{\alpha})$ might not be exactly on the canonical surface and therefore we project it onto $\mathbf{X}(\boldsymbol{\alpha}_0)$. Now, we can interpolate the UV texture coordinates and assign them to $\tilde{\mathbf{v}}_i$. Similarly, we can also assign segmentation labels or define on- or near-surface landmarks. In fig. 2 (right) we show an imGHUM reconstruction textured and with a binary 'clothing' segmentation. We use the latter throughout the paper demonstrating that our semantics allow the transfer of segmentation labels to different iso-surface reconstructions. Please refer to §3.3 for more applications of our implicit semantics e.g. landmarks or clothed human reconstruction.

Architecture. For the single-part network we use a similar feed-forward architecture as DeepSDF [32] or IGR [14] with eight 512-dimensional fully-connected layers. To enable higher-order derivatives, we use Swish nonlinear activation [40] instead of ReLU. IGR originally proposed SoftPlus, however, we found Swish superior (see tab. 3). The multi-part network is composed out of one 8-layer 256-dimensional MLP for the body and three 4-layer 256-dimensional MLPs for hands and head. Each sub-network has a skip connection to the middle layer. The last hidden layers of sub-networks are aggregated in a 128-dimensional fully-connected layer with Swish nonlinear activation, before the final network output. The final model features 2.49 million parameters and performs 4.99 million FLOPs per point query.

Dataset. Our training data consists of a collection of full-body human meshes $\mathbf{Y}$ together with the corresponding GHUM latent code $\boldsymbol{\alpha}$, where $\mathbf{X}(\boldsymbol{\alpha})$ best approximates $\mathbf{Y}$. For each mesh, we perform Poisson disk sampling on the surface and obtain $|O| = 32$K surface samples, together with their surface normals. In addition, within a predefined $2.2 \times 2.8 \times 2.2\,\mathrm{m}^3$ bounding box centered at the origin, we sample $|F|/2 = 16$K points uniformly. Another 16K samples are generated by randomly displacing surface sample points with isotropic normal noise with $\sigma = 0.05$m. All off-surface samples are associated with inside/outside labels, computed by casting randomized rays and checking parity. We also label semantics for on and near surface samples, which are drawn with random face indices and barycentric weights of the GHUM mesh and randomly displaced for near-surface samples. With the corresponding face and barycentric weights, semantic labels are generated using (6) in a light-weight computation with no need for projection or nearest neighbor search. Each mesh Y is then decomposed into $N = 4$ parts and we generate the same number of training samples per body part (we use $\sigma = 0.02$m for surface samples near the hands).

We use two types of human meshes for our imGHUM training. We first randomly sample 75K poses from H36M and the CMU mocap dataset, with Gaussian sampled body shapes, expressions and hand poses from the GHUM latent priors, where $\mathbf{Y}$ are the posed GHUM meshes. In addition, we collect 35K human scans, on which we perform As-Conformal-As-Possible (ACAP) registrations [51] with the GHUM topology and fit GHUM parameters as well. Our human scans include the CAESAR dataset, full body pose scans, as well as close-up head and hand scans. Due to the noise and incompleteness in some of the raw scans we use the registrations for training. We fine-tune imGHUM – initially trained on GHUM sampling – using the registration dataset. In this way, imGHUM can capture geometric detail not well represented by GHUM (see tab. 2).

3. Experiments

We evaluate imGHUM qualitatively and quantitatively in multiple experiments. First, we compare imGHUM with its explicit counterpart GHUM (§3.1). Then, we perform an extensive baseline and ablation study, demonstrating the effect of imGHUM's architecture and training scheme (§3.2). We also build a model to compare to the recent single-subject occupancy model NASA. Finally, we show the performance of imGHUM on three representative applications demonstrating its usefulness and versatility (§3.3).

We report three different metrics. Bi-directional Chamfer-L_2 distance measures the accuracy and completeness of the surface (lower is better). Normal Consistency (NC) evaluates estimated surface normals (higher is better). Volumetric Intersection over Union (IoU) compares the reconstructed volume with the ground truth shape (higher is better). The latter can only be reported for watertight shapes. Please note that metrics not always correlate with the perceived quality of the reconstructions. We therefore additionally include qualitative side-by-side comparisons.

For visualization and numerical evaluation we extract meshes from imGHUM using Marching Cubes [26]. To this end, we approximate the bounding box of the surface though probing and then run Marching Cubes with a resolution of 256^3 within the bounding box. Hereby, the signed distances support acceleration using Octree sampling: we use the highest grid density only near the surface and sam-

Figure 3. Bodies generated and reconstructed using imGHUM. *Left*: imGHUM with Gaussian sampling of the shape, expression and pose latent space. *Middle*: Reconstructed motion sequence from the CMU mocap dataset [2] (fixed body shape). *Right*: Body shape and facial expressions latent code interpolation (fixed pose). See supplementary material for more examples.

ple far less frequently away from it. However, we note that for most applications, such as human reconstruction and collision detection, Marching Cubes are not needed, except only once for the final mesh visualization.

3.1. Representational Power

In fig. 3, we show reconstructions of a motion capture sequence applied to imGHUM. Our model captures well the articulated full-body motion, with consistent body shape for various poses. By sharing the latent priors with GHUM, imGHUM supports realistic body shape and pose generation (fig. 3, left) as well as smooth interpolation within the shape and expression latent spaces (fig. 3, right). Our model generalizes well to novel body shapes, expressions, and poses, and has interpretable and decoupled latent representations.

In tab. 2, we compare the representation power of imGHUM with the explicit GHUM on our registration testset. imGHUM better captures present detail as numerically demonstrated. An imGHUM model trained only using GHUM samples captures the body deformation due to articulation less well, indicating that GHUM is a useful surrogate to 'synthetically' bootstrap the training of the implicit network, but that real data is important as well.

Limitations of imGHUM are sometimes apparent for very extreme pose configurations that have not been covered in the training set, such as anthropometrically invalid poses that are impossible for a human, e.g. resulting in self-intersection or by bending joints beyond their anatomical range of motion. imGHUM produces plausible results for inputs not too far from expected configurations, but the results occasionally feature some defects e.g. distorted or incomplete geometry or inaccurate semantics, see fig. 8 for examples.

3.2. Baseline Experiments

In the next section, we compare imGHUM to various baselines inspired by recent work. The first is an auto-encoder, where the encoder side is PointNet++ [38] and the decoder is our single-part network. The idea is to let the network find the best representation instead of precomputing a low dimensional representation. In practice this means that latent codes are not interpretable. Further,

Model	IoU ↑	Chamfer $\times 10^{-3}$ ↓	NC ↑
imGHUM ‡	0.900	0.071	0.977
GHUM	0.913	0.055	0.983
imGHUM	**0.932**	**0.040**	**0.984**

Table 2. GHUM comparisons on registration dataset. imGHUM marked with ‡ is trained only based on GHUM sampling data.

Model	IoU ↑	Chamfer $\times 10^{-3}$ ↓	NC ↑
Autoencoder	0.831	2.457	0.923
Single-part †	0.957	0.085	0.983
Single-part ⊕	0.958	0.070	0.983
Single-part	0.965	0.052	0.986
Single-part deeper †	0.961	0.070	0.984
Single-part deeper	0.967	0.058	0.986
imGHUM †	0.955	0.095	0.984
imGHUM w/o L_l (4)	0.966	0.051	0.988
imGHUM	**0.969**	**0.036**	**0.989**

Table 3. Numerical comparison with baselines. Models marked with † don't use the Fourier input mapping. ⊕ marks Softplus activation as in [14].

Model	IoU ↑ Head / Hands	Ch. $\times 10^{-3}$ ↓ Head / Hands	NC ↑ Head / Hands
Single-part	0.967 / 0.818	0.010 / 0.201	0.937 / 0.790
Single-part deep.	0.968 / 0.832	0.011 / 0.271	0.938 / 0.811
imGHUM	**0.976 / 0.929**	**0.007 / 0.031**	**0.944 / 0.934**

Table 4. Unidirectional metrics (GT to generated mesh) for critical body parts. Our multi-part architecture significantly improves the head and hand reconstruction accuracy.

we experiment with our single part network without Fourier input mapping, largely following the training scheme proposed by IGR [14]. We also use input mapping and finally trained a deeper single-part network variant (10 layers) having roughly the same number of variables as imGHUM.

In tab. 3 we report the metrics for different variants on our test set containing 1 000 GHUM samples. In fig. 4, we show a side-by-side comparison. The Fourier input mapping consistently improves results for all variants. We have also tried higher-dimensional Fourier features but empirically found the basic encoding to work best in our setting. The auto-encoder produces large artifacts especially in the hand region. Similar problems, large blobs or missing pieces, can be observed in results from single-part variants, especially for the hands and, less severe, also for the facial region. These problems, however, are not well captured by globally evaluating the whole shape. To this end, we evaluate imGHUM and our single-part models specifically for these critical regions, see tab. 4. Only imGHUM consistently produces high-quality results also for hands and the face, supporting the proposed architecture choices.

Next, we compare imGHUM to the recent single-subject multi-pose implicit human occupancy model NASA [11]. With a fixed body shape, we generate 22 500 random GHUM full-body training poses and 2 500 testing poses from Human3.6M [20] and the CMU mocap dataset [2], including head and hand poses. Using the original point sampling strategy in NASA, we have trained the network until convergence, based on the original source code. Please see the supplementary material for details on how we adapted NASA for the GHUM skeleton. For comparison, we have trained an imGHUM architecture with $2\times$ fewer layers than our full multi-subject model, each with half-dimensionality, using the same dataset. Even though GHUM-based NASA has $3\times$ more parameters, our smaller-size single-subject imGHUM still performs significantly better in representing both the global shape and local detail (see hand reconstructions in fig. 5). In contrast to NASA, which computes binary occupancy, imGHUM returns more informative signed distance values which produce smooth decision boundaries and preserve the detailed geometry much better. Further key differences to NASA are our considerably simpler architecture that requires far less computation to produce a reconstruction, our semantics, and the carefully chosen learning model (i.e. Fourier encoding, second-order losses) that pays particular attention to surface detail. Moreover, imGHUM additionally models body shape, fingers, and facial expressions using generative latent codes (tab. 1).

3.3. Applications

We apply imGHUM to three key tasks: body surface reconstruction, partial point cloud completion, and dressed and inclusive human reconstruction.

Triangle Set Surface Reconstruction. Given a triangle set ('soup') with n vertices $\{\hat{\mathbf{v}}\} \in \mathbf{R}^{3n}$ along with oriented normals $\{\hat{\mathbf{n}}\} \in \mathbf{R}^{3n}$, we deploy our parametric implicit SDF for surface reconstruction with semantics. This task is necessary for triangle soups produced by 3D scanners. To extract the surface from an incomplete scan, we apply a BFGS optimizer to fit $\boldsymbol{\alpha} = (\boldsymbol{\beta}_b, \boldsymbol{\beta}_f, \boldsymbol{\theta})$ such that all vertices $\hat{\mathbf{v}}$ are close to the implicit surface $S(\cdot, \boldsymbol{\alpha}) = 0$. Moreover, we enforce gradients at $\hat{\mathbf{v}}$ to be close to normals $\hat{\mathbf{n}}$, and generated off-surface samples to have distances with the expected signs. In addition, we sample near surface points with a small distance η along surface normals, and enforce $S(\hat{\mathbf{v}} \pm \eta\hat{\mathbf{n}}, \boldsymbol{\alpha}) = \pm\eta$, as in [32]. Note that all these operations can be easily implemented and are fully differential due to imGHUM being a SDF. When 3D landmarks are available on the target surface, e.g. as triangulated from

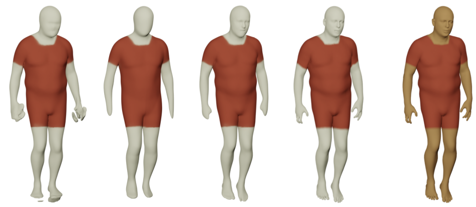

Figure 4. Qualitative comparison with baseline experiments. From left to right: autoencoder, single-part model without and with Fourier input mapping, our multi-part imGHUM, ground-truth GHUM. We use our semantics network to color baseline results.

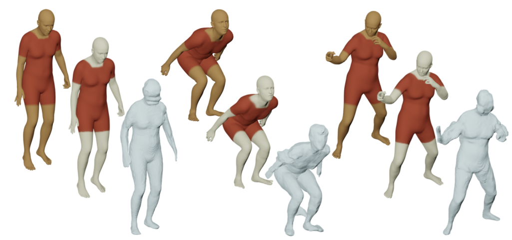

Figure 5. Comparison with NASA [11] on our single-subject multi-pose dataset. Top to bottom: GT, single-subject imGHUM, and NASA reconstructions. imGHUM better captures global and local geometry, despite using a significantly smaller network version in this experiment. Also numerically our results are superior: IoU (↑) 0.962 (ours) vs. 0.839 (theirs), Ch. (↓) 0.068×10^{-3} (ours) vs. 3.53×10^{-3} (theirs), NC (↑) 0.985 (ours) vs. 0.903 (theirs).

2D detected landmarks of raw scanner images, we additionally augment the optimization with landmark losses based on the imGHUM semantics. Please see the supplementary material for details of the losses.

For reference, we also show results on IF-Net [8], a recent method for implicit surface extraction, completion, and voxel super-resolution. We trained IF-Net with the same pose and shape variation as used for imGHUM – presumably much more variation than the 2 183 scans in the original paper. In both training and testing we generate 15K random samples from the observed shape and pass them through IF-Net for surface reconstruction. Note that IF-Net is using less information compared to our method, but is also solving an easier task as it is not computing a global and semantically meaningful shape code. An entirely fair comparison is thus not possible. However, we believe that by comparing with IF-Net, we show that imGHUM is adequate for this task. Fig. 6 qualitatively shows examples of both imGHUM fits and of IF-Net inference results for 150 human scans containing 20 subjects. Our model not only fits well to the volume of the scans but also reconstructs the facial expressions and hand poses. Using landmarks and ICP losses, one could also fit GHUM to the triangle sets. However, our fully differential imGHUM losses show superior performance over ICP-based GHUM fitting (Chamfer (↓) 0.77×10^{-3}, NC (↑) 0.921).

Partial Point Cloud Completion. Another relevant task for many applications is shape completion. Here we show

Figure 6. *Left:* Triangle set surface reconstruction (input scan, imGHUM fit, and IF-Net inference from left to right). Numerically, imGHUM fits are better than IF-Net with Chamfer distance (↓) 0.156×10^{-3} (ours) vs. 0.844×10^{-3} (IF-Net), and NC (↑) 0.954 (ours) vs. 0.914 (IF-Net). *Right:* Partial point cloud completion (input point cloud, imGHUM fit, IF-Net, and ground truth scan).

surface reconstruction and completion from partial point clouds as recorded e.g. using a depth sensor. We synthesize depth maps from A-posed scans of 10 subjects from the Faust dataset [6] using the intrinsics and the resolution of a Kinect V2 sensor. To complete the partial view, we search for the $\boldsymbol{\alpha}$ such that all points from the depth point cloud are close to imGHUM's zero-level-set. We sample additional points along surface normals (estimated from depth image gradients) and enforce estimated distances by imGHUM to be close to true distances. We also sample points in front of the depth cloud and around it and enforce their L_l label loss. Finally, we also supervise the estimated normals. We do not rely on landmarks or other semantics in this experiment.

We show IF-Net [8] results for comparison. We trained IF-Net specifically for this task while we use the same imGHUM for all experiments. Our reconstructions are numerically better with Chamfer distance (↓) 0.103×10^{-3} (ours) vs. 0.315×10^{-3} (theirs) and NC (↑) 0.962 (ours) vs. 0.936 (theirs). Qualitatively, our results contain much more of the desirable reconstruction detail, especially for hands and faces, see fig. 6, right. Note, again, that IF-Net only reconstructs a surface while we recover the parametrization of a body model, a considerably harder task.

Dressed and Inclusive Human Modeling. imGHUM is template-free which is a valuable property for future developments. While this work deals primarily with the methodology of learning a generative implicit human model – in itself a complex and novel task – we also give an outlook for possible future directions. Building a detailed model of the human body shape including hair and clothing, or learning inclusive models could be such directions. However, currently the data needed for building such models does not exist at large enough scale. To demonstrate that imGHUM is a valuable building block for such models, we leverage it as an inner layer for personalized human models. Concretely, we augment imGHUM with a light-weight residual SDF network, conditioned on the output of imGHUM, both the signed distances and semantics. We estimate the residual model using the same learning scheme as for imGHUM, but limit training to a single scan. The final output models the human with layers, including the inner body shape represented with imGHUM and the personalization (hair, clothing, non-standard body topology) as residuals, *c.f.* fig. 7. This layered representation can be reposed by changing the parameterization of the underlying imGHUM. Hereby, the residual model acts as a fitted layer around imGHUM and deforms according to the distance and semantic field defined by imGHUM. Please see the supplementary material for more examples, a numerical evaluation, and implementation details.

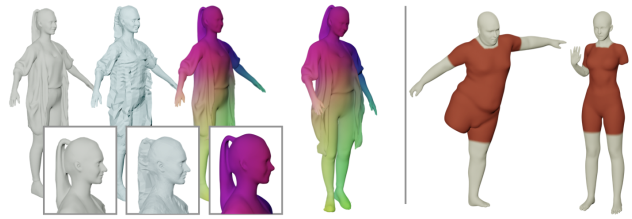

Figure 7. From left to right: scan, GHUM template mesh ACAP registration, imGHUM+residual fit (color-scale represents semantics), reposed imGHUM+residual, imGHUM+residual fits to people with limb differences. In contrast to the fitted template mesh, imGHUM+residual successfully models topologies different from the plain human body and captures more geometric detail.

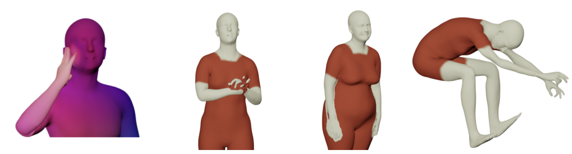

Figure 8. Failure modes. Interpenetration can lead to unwanted shapes and semantics (leaked hand semantics to the cheek). Extreme poses may produce deformed body parts (thin arms).

4. Discussion and Conclusion

We introduced imGHUM, the first 3D human body model, with controllable pose and shape, represented as an implicit signed distance function. imGHUM has comparable representation power to state-of-the-art mesh-based models and can represent significant variations in body pose, shape, and facial expressions, as well as underlying, precise, semantics. imGHUM has additional valuable properties, since its underlying implicit SDF represents not only the surface of the body but also its neighborhood, which e.g. enables collision tests with other objects or efficient distance losses. imGHUM can be used to build diverse, fair models of humans who may not match a standard template. This paves the way for transformative research and inclusive applications like modeling clothing, enabling immersive virtual apparel try-on, or free-viewpoint photorealistic visualization. Our models are available for research [1].

References

[1] https://github.com/google-research/google-research/tree/master/imghum. 2, 8

[2] CMU graphics lab motion capture database. 2009. http://mocap.cs.cmu.edu/. 6, 7

[3] Thiemo Alldieck, Marcus Magnor, Bharat Lal Bhatnagar, Christian Theobalt, and Gerard Pons-Moll. Learning to reconstruct people in clothing from a single RGB camera. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1175–1186. IEEE, 2019. 1

[4] Matan Atzmon and Yaron Lipman. Sal: Sign agnostic learning of shapes from raw data. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2565–2574, 2020. 2

[5] Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Loopreg: Self-supervised learning of implicit surface correspondences, pose and shape for 3d human mesh registration. In *Adv. Neural Inform. Process. Syst.*, 2020. 4

[6] Federica Bogo, Javier Romero, Matthew Loper, and Michael J. Black. FAUST: Dataset and evaluation for 3D mesh registration. In *IEEE Conf. Comput. Vis. Pattern Recog.* IEEE, 2014. 8

[7] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5939–5948, 2019. 1, 2

[8] Julian Chibane, Thiemo Alldieck, and Gerard Pons-Moll. Implicit functions in feature space for 3d shape reconstruction and completion. In *IEEE Conf. Comput. Vis. Pattern Recog.* IEEE, jun 2020. 1, 2, 7, 8

[9] Julian Chibane, Aymen Mir, and Gerard Pons-Moll. Neural unsigned distance fields for implicit function learning. In *Adv. Neural Inform. Process. Syst.*, December 2020. 2

[10] Julian Chibane and Gerard Pons-Moll. Implicit feature networks for texture completion from partial 3d data. In *Eur. Conf. Comput. Vis. Worksh.* Springer, August 2020. 2

[11] Boyang Deng, JP Lewis, Timothy Jeruzalski, Gerard Pons-Moll, Geoffrey Hinton, Mohammad Norouzi, and Andrea Tagliasacchi. Neural articulated shape approximation. In *Eur. Conf. Comput. Vis.* Springer, August 2020. 2, 7

[12] Kyle Genova, Forrester Cole, Avneesh Sud, Aaron Sarna, and Thomas Funkhouser. Local deep implicit functions for 3d shape. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4857–4866, 2020. 2

[13] Kyle Genova, Forrester Cole, Daniel Vlasic, Aaron Sarna, William T Freeman, and Thomas Funkhouser. Learning shape templates with structured implicit functions. In *Int. Conf. Comput. Vis.*, pages 7154–7164, 2019. 2

[14] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. In *Int. Conf. on Mach. Learn.*, pages 3569–3579. 2020. 2, 3, 5, 6

[15] Nils Hasler, Carsten Stoll, Martin Sunkel, Bodo Rosenhahn, and H-P Seidel. A statistical model of human pose and body shape. *Comput. Graph. Forum*, 28(2):337–346, 2009. 2

[16] N. Hesse, S. Pujades, M.J. Black, M. Arens, U. Hofmann, and S. Schroeder. Learning and tracking the 3D body shape of freely moving infants from RGB-D sequences. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(10):2540–2551, 2020. 1

[17] David A Hirshberg, Matthew Loper, Eric Rachlin, and Michael J Black. Coregistration: Simultaneous alignment and modeling of articulated 3D shape. In *Eur. Conf. Comput. Vis.*, pages 242–255, 2012. 2

[18] Zeng Huang, Tianye Li, Weikai Chen, Yajie Zhao, Jun Xing, Chloe LeGendre, Linjie Luo, Chongyang Ma, and Hao Li. Deep volumetric video from very sparse multi-view performance capture. In *Eur. Conf. Comput. Vis.*, pages 336–354, 2018. 2

[19] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. Arch: Animatable reconstruction of clothed humans. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3093–3102, 2020. 2

[20] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2014. 7

[21] Arjun Jain, Thorsten Thormählen, Hans-Peter Seidel, and Christian Theobalt. Moviereshape: Tracking and reshaping of humans in videos. *ACM Trans. Graph.*, 29(5), 2010. 1

[22] Chiyu Jiang, Avneesh Sud, Ameesh Makadia, Jingwei Huang, Matthias Nießner, and Thomas Funkhouser. Local implicit grid representations for 3d scenes. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6001–6010, 2020. 2

[23] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3D deformation model for tracking faces, hands, and bodies. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 8320–8329. IEEE, 2018. 1, 2

[24] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. 1

[25] Korrawe Karunratanakul, Jinlong Yang, Yan Zhang, Michael J Black, Krikamol Muandet, and Siyu Tang. Grasping field: Learning implicit representations for human grasps. In *Int. Conf. on 3D Vis.*, pages 333–344. IEEE, 2020. 2

[26] Thomas Lewiner, Hélio Lopes, Antônio Wilson Vieira, and Geovan Tavares. Efficient implementation of marching cubes' cases with topological guarantees. *Journal of Graphics Tools*, 8(2):1–15, 2003. 5

[27] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graph.*, 34(6):248:1–248:16, 2015. 1, 2

[28] Lars M. Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4460–4470, 2019. 1, 2

[29] Mateusz Michalkiewicz, Jhony K Pontes, Dominic Jack, Mahsa Baktashmotlagh, and Anders Eriksson. Deep level sets: Implicit surface representations for 3D shape inference. *arXiv preprint arXiv:1901.06802*, 2019. 1, 2

[30] Marko Mihajlovic, Yan Zhang, Michael J Black, and Siyu Tang. LEAP: Learning articulated occupancy of people. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 3

[31] Ahmed A A Osman, Timo Bolkart, and Michael J. Black. STAR: A spare trained articulated human body regressor. In *Eur. Conf. Comput. Vis.*, 2020. 2

[32] Jeong Joon Park, Peter Florence, Julian Straub, Richard A. Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 165–174, 2019. 1, 2, 3, 5, 7

[33] 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 *IEEE Conf. Comput. Vis. Pattern Recog.* IEEE, 2019. 2

[34] Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. In *Eur. Conf. Comput. Vis.*, Cham, 2020. Springer International Publishing. 2

[35] Leonid Pishchulin, Stefanie Wuhrer, Thomas Helten, Christian Theobalt, and Bernt Schiele. Building statistical shape spaces for 3d human modeling. *Pattern Recognition*, 67:276–286, 2017. 2

[36] Ralf Plankers and Pascal Fua. Articulated soft objects for video-based body modeling. In *Int. Conf. Comput. Vis.*, pages 394–401. IEEE, 2001. 2

[37] Gerard Pons-Moll, Javier Romero, Naureen Mahmood, and Michael J Black. Dyna: a model of dynamic human shape in motion. *ACM Trans. Graph.*, 34:120, 2015. 2

[38] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Adv. Neural Inform. Process. Syst.*, pages 5099–5108, 2017. 6

[39] Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *Int. Conf. on Mach. Learn.*, pages 5301–5310. PMLR, 2019. 4

[40] Prajit Ramachandran, Barret Zoph, and Quoc V Le. Searching for activation functions. *arXiv preprint arXiv:1710.05941*, 2017. 5

[41] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Int. Conf. Comput. Vis.*, pages 2304–2314, 2019. 2

[42] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 2

[43] Shunsuke Saito, Jinlong Yang, Qianli Ma, and Michael J. Black. SCANimate: Weakly supervised learning of skinned clothed avatar networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 3

[44] Vincent Sitzmann, Julien N.P. Martel, Alexander W. Bergman, David B. Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. In *Adv. Neural Inform. Process. Syst.*, 2020. 2

[45] Cristian Sminchisescu, Atul Kanaujia, and Dimitris Metaxas. Learning joint top-down and bottom-up processes for 3D visual inference. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1743–1752. IEEE, 2006. 2

[46] Cristian Sminchisescu and Alexandru Telea. Human pose estimation from silhouettes. a consistent approach using distance level sets. In *10th Int. Conf. on Computer Graphics, Visualization and Computer Vision*, 2002. 1, 2

[47] Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T. Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *Adv. Neural Inform. Process. Syst.*, 2020. 4

[48] Daniel Thalmann, Jianhua Shen, and Eric Chauvineau. Fast realistic human body deformations for animation and VR applications. In *Proceedings of CG International'96*, pages 166–174, 1996. 2

[49] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. GHUM & GHUML: Generative 3d human shape and articulated pose models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6184–6193, 2020. 1, 2, 3

[50] Yifan Xu, Tianqi Fan, Yi Yuan, and Gurprit Singh. Ladybird: Quasi-monte carlo sampling for deep implicit field based 3d reconstruction with symmetry. In *Eur. Conf. Comput. Vis.*, 2020. 2

[51] Yusuke Yoshiyasu, Wan-Chun Ma, Eiichi Yoshida, and Fumio Kanehiro. As-conformal-as-possible surface registration. *Comput. Graph. Forum*, 2014. 5

[52] Andrei Zanfir, Eduard Gabriel Bazavan, Hongyi Xu, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Weakly supervised 3d human pose and shape reconstruction with normalizing flows. In *Eur. Conf. Comput. Vis.*, 2020. 1, 3

[53] Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Andrei Zanfir, and Cristian Sminchisescu. Human synthesis and scene compositing. In *AAAI*, pages 12749–12756, 2020. 1

[54] Tiancheng Zhi, Christoph Lassner, Tony Tung, Carsten Stoll, Srinivasa G. Narasimhan, and Minh Vo. TexMesh: Reconstructing detailed human texture and geometry from RGB-D video. In *Eur. Conf. Comput. Vis.* Springer International Publishing, 2020. 1