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.

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 [48], 3D human pose and shape reconstruction [25, 54], personal avatar creation [3, 56], medical diagnosis and treatment [17], or image synthesis and video editing [55, 22]. 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 [51] 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 [29, 24] 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 [30, 34, 7, 31]. 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
and offering good full-body representation functionality. Shape spaces, compatible with standard graphics pipelines are also popular parametric body models, with linear soft-tissue dynamics [39]. SMPL variants [29, 24, 35, 33] correlate between body shape and pose [16], as well as later variants inspired by blend skinning [18] modeled from video data [38, 48, 47]. SCAPE [37] was one of the properties, c.f prior work which could only demonstrate individual properties, near and on its zero-level-set, enabling e.g. texturing or body part labeling. This holistic approach is novel and returns correspondences to a canonical representation near and on its zero-level-set, enabling e.g. collision tests. 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, 23, 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, 36, 9]. Others address such limitations by changing the learning methodology: tailored network initialization [4] and point sampling strategies [52], or second-order losses [14, 46] have been proposed towards this end. We found the latter to be extremely useful and rely on similar losses in this work.

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.

|                | generative pose | generative shape | generative hands | interpolation | signed distances | semantics | continuous icp |
|----------------|-----------------|------------------|------------------|---------------|------------------|----------|----------------|
| GHUM [51]      | ✓               | ✓                | ✗                | ✗             | ✗                | ✗        | ✗              |
| IF-Net [8]     | ✗               | x                | x                | ✓             | ✗                | ✗        | ✗              |
| IGR [6]        | ✗               | ✗                | ✗                | ✓             | ✗                | ✗        | ✗              |
| NASA [11]      | ✗               | ✗                | ✗                | ✓             | ✗                | ✗        | ✗              |
| imGHUM         | ✓               | ✓                | ✓                | ✗             | ✗                | ✗        | ✗              |

GHUM is a recent parametric model [51] 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 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 [30, 34, 7, 31]. 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 [30, 7] (occupancy networks), or returns its distance to the closest surface [34] (distance functions). The global shape is then defined by the decision boundary or the zero-level-set of this function.

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 [50] and successfully applied e.g. for human reconstruction from video data [38, 48, 47]. SCAPE [37] was one of the first realistic large scale data-driven human body models. Later variants inspired by blend skinning [18] modeled correlations between body shape and pose [16], as well as soft-tissue dynamics [39]. SMPL variants [29, 24, 35, 33] are also popular parametric body models, with linear shape spaces, compatible with standard graphics pipelines and offering good full-body representation functionality. 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.

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.

|                | generative pose | generative shape | generative hands | interpolation | signed distances | semantics | continuous icp |
|----------------|-----------------|------------------|------------------|---------------|------------------|----------|----------------|
| GHUM [51]      | ✓               | ✓                | ✗                | ✗             | ✗                | ✗        | ✗              |
| IF-Net [8]     | ✗               | x                | x                | ✓             | ✗                | ✗        | ✗              |
| IGR [6]        | ✗               | ✗                | ✗                | ✓             | ✗                | ✗        | ✗              |
| NASA [11]      | ✗               | ✗                | ✗                | ✓             | ✗                | ✗        | ✗              |
| imGHUM         | ✓               | ✓                | ✓                | ✗             | ✗                | ✗        | ✗              |

Human Body Models. Parametric human body models based on geometric primitives have been proposed early on [50] and successfully applied e.g. for human reconstruction from video data [38, 48, 47]. SCAPE [37] was one of the first realistic large scale data-driven human body models. Later variants inspired by blend skinning [18] modeled correlations between body shape and pose [16], as well as soft-tissue dynamics [39]. SMPL variants [29, 24, 35, 33] 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 [51] 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 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 [30, 34, 7, 31]. 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 [30, 7] (occupancy networks), or returns its distance to the closest surface [34] (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, 23, 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, 36, 9]. Others address such limitations by changing the learning methodology: tailored network initialization [4] and point sampling strategies [52], or second-order losses [14, 46] have been proposed towards this end. We found the latter to be extremely useful and rely on similar losses in this work.
related, SCANimate \cite{45} builds personalized avatars from multiple scans of a single person. Concurrent to our work, LEAP \cite{32} 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 full body and poses, hand articulation, 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 $\alpha$ 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 $S(p, \alpha) \in \mathbb{R}$ is a continuous function which, given an arbitrary spatial point $p \in \mathbb{R}^3$, outputs the shortest distance to the surface defined by $\alpha$, 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 $S(\cdot, \alpha) = 0$.

GHUM \cite{51} represents the human model as an articulated mesh $X(\alpha)$. GHUM has a minimally-parameterized skeleton with $J = 63$ joints (124 Euler angle DOFs), and skinning deformations, explicitly sensitive to the pose kinematics $\theta \in \mathbb{R}^{124}$. A kinematic prior based on normalizing flows defines the distribution of valid poses \cite{54}. Each kinematic pose $\theta$ represents a set of joint transformations $T(\theta, j) \in \mathbb{R}^{J \times 3 \times 4}$ from the neutral to a posed state, where $j \in \mathbb{R}^{J \times 3}$ are the joint centers that are dependent on the neutral body shape. The statistical body shapes are modeled using a nonlinear embedding $\beta_k \in \mathbb{R}^{16}$. In addition to skeleton articulation, a nonlinear latent code $\beta_f \in \mathbb{R}^{20}$ drives facial expressions. The implicit model we design here shares the same probabilistic latent representation as GHUM, $\alpha = (\beta_b, \beta_f, \theta)$, but in contrast to computing an articulated mesh, we estimate a signed distance value $s = S(p, \alpha)$ for each arbitrary spatial point $p$.

2.1. Models and Training

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

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

$$||\nabla_p S(p, \alpha; \omega)|| = 1, \quad (1)$$

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

$$L_\alpha(\omega) = \frac{1}{|O|} \sum_{i \in O} \langle |S(p_i, \alpha)| + ||\nabla_p S(p_i, \alpha) - n_i|| \rangle \quad (2)$$

$$L_\omega(\omega) = \frac{1}{|F|} \sum_{i \in F} (||\nabla_p S(p_i, \alpha)|| - 1)^2 \quad (3)$$

$$L_l(\omega) = \frac{1}{|F|} \sum_{i \in F} \text{BCE}(l_i, \phi(kS(p_i, \alpha))) \quad (4)$$

Figure 2. Overview of imGHUM. We compute the signed distance $s = S(p, \alpha)$ and the semantics $c = C(p, \alpha)$ of a spatial point $p$ to the surface of an articulated human shape defined by the generative latent code $\alpha$. Using an explicit skeleton, we transform the point $p$ into the normalized coordinate frames as $\{\tilde{p}^i\}$ 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 $c$ 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.
where $\phi$ is the sigmoid function, $O$ are surface samples from $Y$ with normals $n_i$, and $F$ are off-surface samples with inside/outside labels $l_i$, 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 $n_i$. The Eikonal loss $L_e$ is derived from (1) where the SDF is differentiable everywhere with gradient norm 1. We obtain the SDF gradient $\nabla_{p_i} S(p_i, \alpha)$ analytically via network back-propagation. In practice, we also find it useful to include a binary cross-entropy error (BCE) loss $L_1$ for off-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 [41, 49]. To address this limitation, inspired by [49], frequency functions, a phenomenon referred to as spectral MLP networks encounter difficulties in learning high-frequency functions, a phenomenon referred to as spectral.

Given an arbitrary spatial point, on or near the surface $Y$, i.e., $|S(p_i, \alpha)| < \sigma$, we are therefore interested to interpret its semantics. We define the semantics as a 3D implicit function $C(p, \alpha) \in \mathbb{R}^3$. Given a query point $p_i$, it returns a correspondence point on a canonical GHUM mesh $X(\alpha_0)$ as

$$C(p_i, \alpha) = w_i \cdot v_f(\alpha_0) = e_i, \quad p_i^* = w_i \cdot v_f(\alpha)$$

where $p_i^*$ is the closest point of $p_i$ in the GHUM mesh $X(\alpha)$ with $f$ the nearest face and $w$ the barycentric weights of the vertex coordinates $v_f$. In contrast to alternative semantic encodings, such as 2D texture coordinates, our semantic function $C(p, \alpha)$ 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 \( S(p, \alpha) \) and \( C(p, \alpha) \). Semantics are trained fully-supervised, using an \( L_1 \) loss for a collection of training sample points near and on the surface \( Y \). 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 \( S(\cdot, \alpha) = z \) can be extracted using Marching Cubes [28]. Then for every generated vertex \( \tilde{v}_i \) we query its semantics \( C(\tilde{v}_i, \alpha) \). The queried correspondence point \( C(\tilde{v}_i, \alpha) \) might not be exactly on the canonical surface and therefore we project it onto \( X(\alpha_0) \). Now, we can interpolate the UV texture coordinates and assign them to \( \tilde{v}_i \). 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 [34] or IGR [14] with eight 512-dimensional fully-connected layers. To enable higher-order derivatives, we use Swish nonlinear activation [32] 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 \( Y \) together with the corresponding GHUM latent code \( \alpha \), where \( X(\alpha) \) best approximates \( Y \). For each mesh, we perform Poisson disk sampling on the surface and obtain \(|O| = 32K \) surface samples, together with their surface normals. In addition, within a predefined \( 2.2 \times 2.8 \times 2.2 \) m\(^3\) bounding box centered at the origin, we sample \(|F|/2 = 16K \) points uniformly. Another 16K samples are generated by randomly displacing surface sample points with isotropic normal noise with \( \sigma = 0.05m \). 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 \( \sigma = 0.02m \) 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 \( Y \) are the posed GHUM meshes. In addition, we collect 35K human scans, on which we perform As-Conformal-As-Possible (ACAP) registrations [53] 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 [28]. 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 test-set. 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++ [40] and the decoder is our single-part network. The idea is to let the network find the best representation instead of pre-computing a low dimensional representation. In practice this means that latent codes are not interpretable. Further, 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

| 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 deep ‡  | 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 $L_i$ (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 ↑ | Ch. $\times 10^{-3}$ ↓ | NC ↑ |
|---------------------|-------|------------------------|------|
| 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.
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 [21] 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× fewer layers than our full multi-subject model, each with half-dimensionality, using the same dataset. Even though GHUM-based NASA has 3× 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 structure 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 \( \{ \hat{v} \} \in \mathbb{R}^{3n} \) along with oriented normals \( \{ \hat{n} \} \in \mathbb{R}^{3n} \), 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 \( \alpha = (\beta_v, \beta_f, \theta) \) such that all vertices \( \hat{v} \) are close to the implicit surface \( S(\cdot, \alpha) = 0 \). Moreover, we enforce gradients at \( \hat{v} \) to be close to normals \( \hat{n} \), and generated off-surface samples to have distances with the expected signs. In addition, we sample near surface points with a small distance \( \eta \) along surface normals, and enforce \( S(\hat{v} \pm \eta \hat{n}, \alpha) = \pm \eta \), as in [34]. 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.](image332x510 to 521x595)

![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).](image332x646 to 521x720)
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 \( \alpha \) 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 \( L_1 \) 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 \( (\downarrow) \) 0.156 \( \times \) \( 10^{-3} \) (ours) vs. 0.315 \( \times \) \( 10^{-3} \) (theirs) and NC \( (\uparrow) \) 0.954 (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 parameterization 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.

**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].
Supplementary Material

In this supplementary material, we detail our implementation and used architectures, show further results, discuss further ablations, and give more details on our experiments and comparisons. We also demonstrate how imGHUM can be used for differentiable rendering.

A. Implementation Details

In the following, we detail the implementation of imGHUM. We specify the used hyper-parameters and the architectures used in the ablation experiments. Finally, we give running times for imGHUM mesh extraction via Marching Cubes [28].

Hyper-parameters. We train imGHUM with a batch-size of 32, each of which contains 32 instances of \( \alpha \) paired with 512 on surface, 256 near surface, and 256 uniform samples for each instance. Our loss is composed as

\[
L = \lambda_{o_1} L_{o_1} + \lambda_{o_2} L_{o_2} + \lambda_{e} L_{e} + \lambda_{l} L_{l} + \lambda_{s} L_{s},
\]

where \( L_{o_1} \) refers to the first part of \( L_{o} \) (distance) and \( L_{o_2} \) the second part (gradient direction), respectively, and \( L_{e} \) refers to the semantics loss. We choose \( \lambda_{o_1} = 1 \), \( \lambda_{o_2} = 1 \), \( \lambda_{e} = 0.1 \), \( \lambda_{l} = 0.1 \), and \( \lambda_{s} = 0.5 \). Empirically we found that linearly increasing \( \lambda_{o_1} \) to 50 over 100K iterations leads to perceptually better results. We train imGHUM until convergence using the Adam optimizer [27] with a learning rate of \( 0.2 \times 10^{-3} \) exponentially decaying by a factor of 0.9 over 100K iterations.

Architectures. The following architectures have been used for the baseline experiments: The single-part network has been used as described in the main paper totaling in 2.01M parameters. The deeper single-part network uses 10 instead of 8 layers, resulting in 2.53M parameters. The autoencoder is composed from a PointNet++ [40] encoder and our single-part decoder with a total number of parameters of 3.91M. The encoder consists of three PointNet++ set abstraction modules and two 512-dimensional fully-connected layers with ReLU activation.

Running Times. We extract meshes from imGHUM using Octree sampling. Reconstructing a mesh in its bounding box and with a maximum grid resolution of \( 256^3 \) takes on average 1.08s using a NVIDIA Tesla V100. Hereby, the network query time sums up to 0.44s and Marching Cubes [28] (on CPU) takes 0.34s. The rest of the time is used by identifying the bounding box through probing (0.17s), Octree logic (0.05s), and transforming the samples to the part reference frames (0.07s). We query imGHUM in batches with a maximum batch-size of 64\(^3\) samples, where one full batch takes on average 0.13s to compute. The resulting meshes feature approximately 100K vertices and 200K facets. Note that imGHUM allows creating meshes in arbitrary resolutions and can be queried and also rendered (c.f. §D.3) without generating an explicit mesh. For reference, we show imGHUM mesh reconstructions in different resolutions in fig. 9.

B. Results

In this supplemental material we show additional results for our application experiments (fig. 11, 12, 14, 15). Additionally, fig. 10 displays a large number of imGHUM instances with great variety in poses, shapes, hand poses, and facial expressions sampled from imGHUM’s generative latent space. This demonstrates once more that imGHUM’s level of detail, expressiveness and generative power is on par with state-of-the-art mesh-based models. Moreover, imGHUM can additionally be queried at arbitrary resolutions and spatial locations and models not only the surface, but also the space around the person.

C. Ablations

In this section, we report further results of our dataset ablation experiment and results of an additional ablation study on joint rotation parameterization.

Dataset. In the main paper we have shown that imGHUM benefits from being trained on both samples of GHUM and additionally on As-Conformal-As-Possible (ACAP) registrations of a corpus of human scans. While training only on scan data can represent the distribution of the scans well (tab. 5), it does not generalize sufficiently to poses that are not covered in this limited training set, as we show in tab. 6.

In fig. 13, we qualitatively show the effect of fine-tuning with scan data. Please note the increased level of detail in the faces and the enhanced soft-tissue deformation.

Rotation Representations. In tab. 7, we report metrics for imGHUM using different rotation representations for joint rotations \( \theta \). We have experimented with Euler angles, basic \( \sin \), \( \cos \) Fourier mapping [54], and the recently
D. Applications

In the following, we explain the losses used in our triangle set surface reconstruction experiment, detail the residual model of the dressed and inclusive modeling experiment, and finally introduce another application namely pose estimation from silhouettes using differentiable rendering.

D.1. Triangle Set Surface Reconstruction

We describe our triangle set surface reconstruction experiment in the main paper (§3.3) and show more examples here in fig. 11. Our imGHUM reconstructions are performed under a weighted combination of losses as

\[
\min_{\alpha} L_o(\alpha) + L^+_v(\alpha) + L^+_i(\alpha) \tag{8}
\]

\[
L_o(\alpha) = \frac{1}{n}\sum_{i} |S(\hat{v}_i, \alpha)| + \|\nabla \phi_S(\hat{v}_i, \alpha) - \hat{n}_i\| \tag{9}
\]

\[
L^+_v(\alpha) = \frac{1}{n}\sum_{i} (\phi(kS(\hat{v}_i + \gamma_i\hat{n}_i, \alpha)) - 1)^2 \tag{10}
\]

\[
L^+_i(\alpha) = \frac{1}{n}\sum_{i} (\phi(kS(\hat{v}_i - \gamma_i\hat{n}_i, \alpha)) )^2, \tag{11}
\]

where \(L_o\) is a surface sample loss (similar to eq. 2 in the main paper), and \(L^+_v, L^+_i\) are sign classification losses defined for points sampled along and opposite to the normals respectively (\(\gamma_i \in [0, 0.05]\) is a Gaussian sampled distance).

Enabled by the implicit semantics of imGHUM, we can additionally exploit landmark losses as,

\[
L_j(\alpha) = \frac{1}{|M_j|}\sum_{i \in M_j} \|T_i(\alpha)j_i - m_{j,i}\|^2 \tag{12}
\]

\[
L_s(\alpha) = \frac{1}{|M_s|}\sum_{i \in M_s} \|C(m_{s,i}, \alpha) - \bar{m}_{s,i}\|^2, \tag{13}
\]

where \(M_j = \{m_j\}, M_s = \{m_s\}\) are a collection of 3D landmarks defined over the joints and the surface, respectively. \(\bar{m}_s\) are the corresponding surface landmarks defined on the canonical mesh \(X(\alpha_0)\). \(L_j\) aligns the transformed joint centers with the joint landmarks. The surface landmarks loss \(L_s\) queries the semantics for the ground-truth surface landmarks \(m_s\) conditioned on \(\alpha\). The semantics describe the position of the landmarks \(m_s\) w.r.t. the canonical mesh, and thus should match their correspondences \(\bar{m}_s\).

Given a triangle set mesh, one could also fit GHUM with landmarks and ICP losses. However, we note that imGHUM is not only able to perform equivalently on the landmark losses to mesh-based representations, but also exploits more information of the triangle set with its differential losses (eq. (8)) compared to ICP. The process of finding the nearest point for ICP at each optimization iteration is non-differentiable and the accuracy of the nearest point correspondences are highly sensitive to the initialization. In contrast, our imGHUM losses are fully differential everywhere and also exploit additional information encoded in the surface normals and the sign labels. Numerical comparisons are reported in §3.3 of the main paper.

D.2. Dressed and Inclusive Human Modeling

In the following, we detail our dressed and inclusive modeling experiment from the main paper. We also show more results in fig. 14. In order to learn a personalized shape of a given scan, we augment imGHUM with an MLP \(\hat{S}\) consisting of four 256-dimensional layers. Each layer is followed by Swish nonlinear activation, and a skip connection is added to the middle layer. \(\hat{S}\) modulates the signed distance field of the body to match the scan. These distance residuals could come from clothing, hair, other ap-
Figure 10. Random imGHUM full-body and part instances sampled from imGHUM’s generative latent codes. On the right, we show textured examples. Texturing and binary coloring is enabled by imGHUM’s semantics.

Figure 11. More examples for the triangle set surface reconstruction experiment. Each pair shows the ground truth scan (left) and our reconstruction (right). Notice the reconstructed facial expressions and hand poses.

Figure 12. Remaining partial point cloud completion results. Left to right: input point cloud, imGHUM fit, and ground truth scan.
D.3. Differentiable Rendering

A benefit of imGHUM’s SDF representation is the potential for rendering using sphere tracing [15]. During ray tracing the surface is located by stepping from the camera along a ray until a surface is passed. In sphere tracing the save step length is given by the current minimal distance to any point on the surface, i.e., the SDF value at the current location. For inexact SDFs, one can take a damped step to reduce the likelihood of over-shooting. Using this technique we can render among other things: imGHUM depth maps, normal maps, and semantics. Hereby, each pixel contains the last queried value of its corresponding camera ray. In the following, we compute differentiable binary silhouettes via sphere tracing and fit imGHUM to images using a silhouette alignment loss.

We implement differentiable approximate sphere tracing by taking a fixed number of steps. Concretely, we step $T = 15$ save steps into the SDF in the direction of each camera ray. At each final point $p_T$ of each camera ray, we query the signed distance value and generate the binarized pixel as:

$$b = \frac{1}{\eta S(p_T, \alpha)^2 + 1},$$

with $\eta = 5000$ in our experiment. $b$ is differentiable w.r.t. $\alpha$ and thus can be used in optimization losses. We formulate a standard silhouette overlap loss and a sparse 2D joint landmark loss and use both to fit imGHUM to image evidence. Fig. 16 shows results of fitting imGHUM to image silhouettes.

E. Details on Compared Methods

As reported in the main paper, we change NASA [11] in contrast to their original version. Firstly, we train NASA based on the GHUM skeleton containing 63 parts. Originally, NASA was trained on SMPL containing only 24 parts. Another difference is the topology of GHUM. In contrast to SMPL, GHUM features an oral cavity that is also represented in our training data. Summarizing, we deploy NASA for a higher-dimensional model and thus a harder task. For a fair comparison, we therefore use a larger and deeper architecture with eight 64-dimensional fully-connected layers for each part instead of the original four 40-dimensional layers. The new architecture features 1.92M parameters (original version has 0.38M) and has shown significantly better representation power. In contrast, we use a much smaller imGHUM architecture in this experiment. imGHUM has been originally designed to also explain shape variation and facial expressions. Since this experiment only features variation in pose, we can use a much smaller version. We use $2 \times$ fewer layers in each part, each with half-dimensionality, resulting in only 0.64M parameters. This smaller-size imGHUM still performs significantly better than NASA in our experiments.

We have trained IF-Net [8] based on their original source code. Specifically, we use IF-Net for point clouds with $128^3$ resolution featuring 2.6M parameters. We also follow their sampling and resizing strategy, such that the input point cloud always has a maximum side length of one unit. Finally, we train IF-Net task-specific (for full and partial point clouds), while we use the same imGHUM in all our comparisons.
Figure 13. Added detail after fine-tuning on the registration dataset. We show imGHUM reconstructions before fine-tuning (left) and after fine-tuning (right) qualitatively and using error heat-maps (red means $\geq 2$cm). Please pay attention to the faces, body shapes, and soft-tissue deformations (digital zoom in recommended).

Figure 14. Dressed human modeling. From left to right: Scan, GHUM ACAP mesh registration, imGHUM+residual fit, reposed or reshaped imGHUM+residual. imGHUM+residual accurately explains all detail present in the input scan. GHUM ACAP mesh registrations have difficulties with complicated and layered structures. By changing the parameterization of the underlying imGHUM, we can repose and reshape the personalized models. The color-scale represents imGHUM semantics and thus correspondences between different instances.

Figure 15. Inclusive human modeling. Left: imGHUM+residual can explain body shapes that do not match the standard template. Right: GHUM ACAP mesh registrations fail to explain these body shapes. For reference, we show ground truth scans in small. Missing limbs are deformed but still present.

Figure 16. Visual 3D reconstruction of imGHUM using differentiable silhouette and landmark losses. Left to right: image, observed silhouette, estimated silhouette, imGHUM reconstruction. By using a silhouette loss, we are able to accurately reconstruct body shapes.