X-Avatar: Expressive Human Avatars

Kaiyue Shen*[1] Chen Guo*[1] Manuel Kaufmann[1] Juan Jose Zarate[1]
Julien Valentin[2] Jie Song†[1] Otmar Hilliges [1]
[1]ETH Zürich [2]Microsoft

Figure 1. We propose **X-Avatar**, an animatible implicit human avatar model capable of capturing human body pose, hand pose, facial expressions, and appearance. X-Avatar can be created from input 3D scans (top row) or RGB-D images (bottom row) and displays high-quality geometry as well as appearance under animation. X-Avatar captures facial expressions and hand gestures (right), making it the first implicit human avatar model to capture the richness of the human state in a unified model.

Abstract

We present X-Avatar, a novel avatar model that captures the full expressiveness of digital humans to bring about life-like experiences in telepresence, AR/VR and beyond. Our method models bodies, hands, facial expressions and appearance in a holistic fashion and can be learned from either full 3D scans or RGB-D data. To achieve this, we propose a part-aware learned forward skinning module that can be driven by the parameter space of SMPL-X, allowing for expressive animation of X-Avatars. To efficiently learn the neural shape and deformation fields, we propose novel part-aware sampling and initialization strategies. This leads to higher fidelity results, especially for smaller body parts while maintaining efficient training despite increased number of articulated bones. To capture the appearance of the avatar with high-frequency details, we extend the geometry and deformation fields with a texture network that is conditioned on pose, facial expression, geometry and the normals of the deformed surface. We show experimentally that our method outperforms strong baselines both quantitatively and qualitatively on the animation task. To facilitate future research on expressive avatars we contribute a new dataset, called X-Humans, containing 233 sequences of high-quality textured scans from 20 participants, totalling 35,500 data frames. Project page: https://ait.ethz.ch/X-Avatar.

1. Introduction

A significant part of human communication is non-verbal in which body pose, appearance, facial expressions, and hand gestures play an important role. Hence, it is clear that the quest towards immersive, life-like remote telepresence and other experiences in AR/VR, will require methods to capture the richness of human expressiveness in its entirety. Yet, it is not clear how to achieve this. Non-verbal communication involves an intricate interplay of several articulated body parts at different scales, which makes it difficult to capture and model algorithmically.

*These authors contributed equally to this work
†Corresponding author

Parametric body models such as the SMPL family [35, 47, 49] have been instrumental in advancing the state-of-the-art in modelling of digital humans in computer vision and graphics. However, they rely on mesh-based representations and are limited to fixed topologies and in resolution of the 3D mesh. These models are focused on minimally clothed bodies and do not model garments or hair. Hence, it is difficult to capture the full appearance of humans.

Neural implicit representations hold the potential to overcome these limitations. Chen et al. [13] introduced a method to articulate human avatars that are represented by continuous implicit functions combined with learned forward skinning. This approach has been shown to generalize to arbitrary poses. While SNARF [13] only models the major body bones, other works have focused on creating implicit models of the face [22, 67], the hands [17], or how to model humans that appear in garments [20] and how to additionally capture appearance [50, 55]. Although neural implicit avatars hold great promise, to date no model exists that holistically captures the body and all the parts that are important for human expressiveness jointly.

In this work, we introduce X-Avatar, an animatable, implicit human avatar model that captures the shape, appearance and deformations of complete humans and their hand poses, facial expressions, and clothing. To this end we adopt the full-body pose space of SMPL-X [47]. This causes two key challenges for learning X-Avatars from data: (i) the significantly increased number of involved articulated body parts (9 used by [13] vs. 45 when including hands and face) and (ii) the different scales at which they appear in observations. The hands and the face are much smaller in size compared to the torso, arms and legs, yet they exhibit similarly or even more complex articulations.

X-Avatar consists of a shape network that models the geometry in canonical space and a deformation network to establish correspondences between canonical and deformed space via learned linear blend skinning (LBS). The parameters of the shape and deformation fields must be learned only from posed observations. SNARF [13] solves this via iterative correspondence search. This optimization problem is initialized by transforming a large number of candidate points via the bone transformations. Directly adopting SNARF and initializing root-finding with only the body bones leads to poor results for the hands and face. Hence, to account for the articulation of these smaller body parts, their bone transformations must also be considered. However, correspondence search scales poorly with the number of bones, so naïvely adding them makes training slow. Therefore, we introduce a *part-aware initialization* strategy which is almost 3 times faster than the naïve version while outperforming it quantitatively. Furthermore, to counteract the imbalance in scale between the body, hands, and face, we propose a *part-aware sampling* strategy, which increases the sampling rate for smaller body parts. This significantly improves the fidelity of the final result. To model the appearance of X-Avatars, we extend the shape and deformation fields with an additional appearance network, conditioned on pose, facial expression, geometry and the normals in deformed space. All three neural fields are trained jointly.

X-Avatar can learn personalized avatars for multiple people and from multiple input modalities. To demonstrate this, we perform several experiments. First, we compare our method to its most related work (SCANimate [50], SNARF [13]) on the GRAB dataset [9, 54] on the animation task of minimally clothed humans. Second, we contribute a novel dataset consisting of 233 sequences of 20 clothed participants recorded in a high-quality volumetric capture stage [16]. The dataset consists of subjects that perform diverse body and hand poses (*e.g.*, counting, pointing, dancing) and facial expressions (*e.g.*, laughing, screaming, frowning). On this dataset we show that X-Avatar can learn from 3D scans and (synthesized) RGB-D data. Our experiments show that X-Avatar outperforms strong baselines both in quantitative and qualitative measures in terms of animation quality. In summary, we contribute:

- X-Avatar, the first expressive implicit human avatar model that captures body pose, hand pose, facial expressions and appearance in a holistic fashion.
- Part-aware initialization and sampling strategies, which together improve the quality of the results and keep training efficient.
- X-Humans, a new dataset consisting of 233 sequences, of high-quality textured scans showing 20 participants with varied body and hand movements, and facial expressions, totalling 35,500 frames.

2. Related Work

Explicit Human Models Explicit models use a triangulated 3D mesh to represent the underlying shape and are controlled by a lower-dimensional set of parameters. Some models focus on capturing a specific part of the human, *e.g.*, the body [3, 35, 44], the hands [49], or the face [6, 34], while others treat the human more holistically like we do in this work, *e.g.* [28, 45, 47, 58, 63]. Explicit models are popular because the 3D mesh neatly fits into existing computer graphics pipelines and because the low-dimensional parameter space lends itself well for learning. Only naturally have such models thus been applied to tasks such as RGB-based pose estimation [15, 21, 29–32, 51–53, 62, 65, 66], RGB-D fitting [7, 14, 61], fitting to body-worn sensor data [26, 56, 60], or 3D hand pose estimation [8, 25] with a resounding success. Because the SMPL family does not natively model clothing, researchers have investigated ways to extend it, *e.g.* via fixed additive 3D offsets [1, 2, 23], also dubbed SMPL+D, pose-dependent 3D offsets [37], by modelling 3D garments and draping them over the SMPL

mesh [5, 18] or via local small surface patches [36]. Explicit models have seen a trend towards unification to model human expressiveness, *e.g*. SMPL-X [47] and Adam [28]. X-Avatar shares this goal, but for implicit models.

Implicit Human Models Explicit body models are limited by their fixed mesh topology and resolution, and thus the expressive power required to model clothing and appearance necessitates extending these models beyond their original design. In contrast, using implicit functions to represent 3D geometry grants more flexibility. With implicit models, the shape is defined by neural fields, typically parameterized by MLPs that predict signed distance fields [46], density [42], or occupancy [39] given a point in space. To extend this idea to articulated shapes like the human body, NASA [19] used per body-part occupancy networks [39]. This per-part formulation creates artifacts, especially for unseen poses, which works such as [13, 40, 41] improve. SNARF [13] does so via a forward warping field which is compatible with the SMPL [35] skeleton, learns pose-independent skinning and generalizes well to unseen poses and people in clothing. Other works [50, 55] model appearance and are learned from scans. [24, 57] create avatars from RGB video and [20] does so from RGB-D video.

Moving beyond bodies, other work has investigated implicit models for faces [22, 48, 59, 67] and hands [17]. Yet, an implicit model that incorporates body, hands, face, and clothing in a single model is missing. X-Avatar fills this gap. We do so by adopting neural forward skinning [13] driven by SMPL-X [47]. This seemingly simple change necessitates non-trivial improvements to the correspondence search as otherwise the iterative root finding is too slow and leads to poor results which we show empirically. We propose to do so by introducing part-aware initialization and sampling strategies, which are incorporated into a single model. Similar to [50], we obtain color with an MLP that is fed with canonical points and conditioned on the predicted geometry. Thanks to the part-aware sampling strategy, our method produces higher quality results than [50] for the hands and faces. Furthermore, in contrast to [13, 50, 55], X-Avatars can be fit to 3D scans *and* RGB-D videos.

Human Datasets Publicly available datasets that show the full range of human expressiveness and contain clothed and textured ground-truth are rare. GRAB [54], a subset of AMASS [38], contains minimally clothed SMPL-X registrations. BUFF [64] and CAPE [37] do not model detailed hand gestures and facial expressions. The CMU Panoptic Studio [27] dataset was used to fit Adam [28] which does model hands and faces, but is neither textured nor clothed. Also, [27] does not contain 4D scans. To study X-Avatars on real clothed humans, we thus contribute our own dataset, X-Humans which contains 35,500 frames of high-quality, textured scans of real clothed humans with corresponding SMPL[-X] registrations.

3. Method

We introduce X-Avatar, a method for the modeling of implicit human avatars with full body control including body movements, hand gestures, and facial expressions. For an overview, please refer to Fig. 2 and Fig. 3. Our model can be learned from two types of inputs, *i.e*., 3D posed scans and RGB-D images. We first recap the SMPL-X full body model. Then we describe the X-Avatar formulation, training scheme, and our part-aware initialization and sampling strategies. For simplicity, we discuss the scan-based version without loss of generality and list the differences to depth-based acquisition in the Supp. Mat.

3.1. Recap: SMPL-X Unified Human Body Model

Our goal is to create fully controllable human avatars. We use the parameter space of SMPL-X [47], which itself extends SMPL to include fully articulated hands and an expressive face. SMPL-X is defined by a function $M(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\psi}) : \mathbb{R}^{|\boldsymbol{\theta}|} \times \mathbb{R}^{|\boldsymbol{\beta}|} \times \mathbb{R}^{|\boldsymbol{\psi}|} \rightarrow \mathbb{R}^{3N}$, parameterized by the shape $\boldsymbol{\beta}$, whole body pose $\boldsymbol{\theta}$ and facial expressions $\boldsymbol{\psi}$. The pose can be further divided into the global pose $\boldsymbol{\theta}_g$, head pose $\boldsymbol{\theta}_f$, articulated hand poses $\boldsymbol{\theta}_h$, and remaining body poses $\boldsymbol{\theta}_b$. Here, $|\boldsymbol{\theta}_g| = 3$, $|\boldsymbol{\theta}_b| = 63$, $|\boldsymbol{\theta}_h| = 90$, $|\boldsymbol{\theta}_f| = 9$, $|\boldsymbol{\beta}| = 10$, $|\boldsymbol{\psi}| = 10$, $N = 10,475$.

3.2. Implicit Neural Avatar Representation

To deal with the varying topology of clothed humans and to achieve higher geometric resolution and increased fidelity of overall appearance, X-Avatar proposes a human model defined by articulated neural implicit surfaces. We define three neural fields: one to model the geometry via an implicit occupancy network, one to model deformation via learned forward linear blend skinning (LBS) with continuous skinning weights, and one to model appearance as an RGB color value.

Geometry We model the geometry of the human avatar in the canonical space with an MLP that predicts the occupancy value f_{occ} for any 3D point $\mathbf{x}_c$ in this space. To capture local non-rigid deformations such as facial or garment wrinkles, we condition the geometry network on the body pose $\boldsymbol{\theta}_b$ and facial expression coefficients $\boldsymbol{\psi}$. We found empirically that high-frequency details are preserved better if positional encodings [43] are applied to the input. Hence, the shape model f_{occ} is denoted by:

$$f_{\text{occ}} : \mathbb{R}^3 \times \mathbb{R}^{|\boldsymbol{\theta}_b|} \times \mathbb{R}^{|\boldsymbol{\psi}|} \rightarrow [0, 1]. \tag{1}$$

The canonical shape is defined as the 0.5 level set of f_{occ}:

$$\mathcal{S} = \{\, \mathbf{x}_c \mid f_{\text{occ}}(\mathbf{x}_c, \boldsymbol{\theta}_b, \boldsymbol{\psi}) = 0.5 \,\}. \tag{2}$$

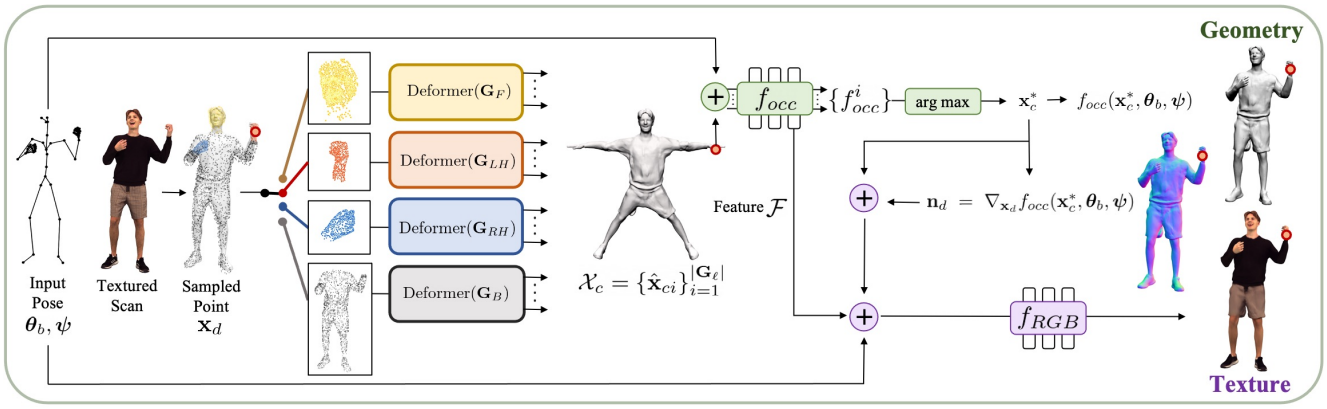

Figure 2. **Method Overview.** Given a posed scan with an SMPL-X registration, we first adaptively sample points $\mathbf{x}_d$ in deformed space per body part ℓ (face F, left hand LH, right hand RH, body B). A part-specific deformer network finds the corresponding candidate points $\hat{\mathbf{x}}_{ci}$ (for $1 \leqslant i \leqslant |\mathbf{G}_\ell|$) in canonical space via iterative root finding. The deformers share the parameters of the skinning network, but each deformer is initialized with only the bone transformations $\mathbf{G}_\ell$ (cf. Fig. 3). The final shape is obtained via an occupancy network f_{occ}. We further model appearance via a texture network that takes as input the body pose $\boldsymbol{\theta}_b$, facial expression $\boldsymbol{\psi}$, the last layer $\mathcal{F}$ of f_{occ}, the canonical point $\mathbf{x}^*_c$ and the normals $\mathbf{n}_d$ in deformed space. The normals correspond to the gradient $\nabla_{\mathbf{x}_d} f_{occ}(\mathbf{x}^*_c, \boldsymbol{\theta}_b, \boldsymbol{\psi})$.

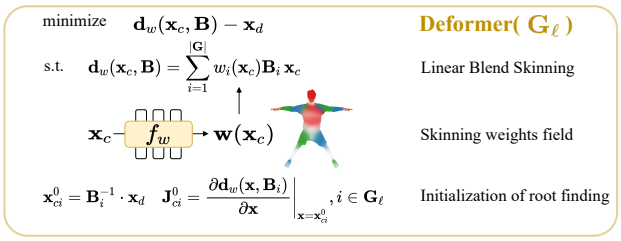

Figure 3. **Part-specific Deformer.** Each deformer shown in Fig. 2 is initialized with the bone transformations belonging to a specific part $\mathbf{G}_\ell, \ell \in \{F, LH, RH, B\}$, but shares the parameters of f_w.

Deformation To model skeletal deformation, we follow previous work [13, 20, 33, 67] and represent the skinning weight field in the canonical space by an MLP:

$$f_w : \mathbb{R}^3 \rightarrow \mathbb{R}^{n_b} \times \mathbb{R}^{n_h} \times \mathbb{R}^{n_f}, \tag{3}$$

where n_b, n_h, n_f denotes the number of body, finger, and face bones respectively. Similar to [13], we assume a set of bones $\mathbf{G}$ and require the weights $\mathbf{w} \in \mathbb{R}^{|\mathbf{G}|}$ to fulfill $w_i \geq 0$ and $\sum_i w_i = 1$. With the learned deformation field $\mathbf{w}$ and given bone transformations $\mathbf{B} = \{\mathbf{B}_1, ..., \mathbf{B}_{|\mathbf{G}|}\}$, for each point $\mathbf{x}_c$ in the canonical space, its deformed counterpart is then uniquely determined:

$$\mathbf{x}_d = \mathbf{d}_w(\mathbf{x}_c, \mathbf{B}) = \sum_{i=1}^{|\mathbf{G}|} w_i(\mathbf{x}_c)\mathbf{B}_i\,\mathbf{x}_c. \tag{4}$$

Note that the canonical shape is a-priori unknown and learned during training. Since the relationship between deformed and canonical points is only implicitly defined, we follow [13] and employ correspondence search. We use Broyden's method [10] to find canonical correspondences $\mathbf{x}_c$ for each deformed query point $\mathbf{x}_d$ iteratively as the roots of $\mathbf{d}_w(\mathbf{x}_c, \mathbf{B}) - \mathbf{x}_d = 0$. In cases of self-contact, multiple valid solutions exist. Therefore the optimization is initialized multiple times by transforming deformed points $\mathbf{x}_d$ rigidly to the canonical space with each bone transformation. Finally, the set of valid correspondences $\mathcal{X}_c$ is determined via analysis of the local convergence.

Part-Aware Initialization At the core of our method lies the problem of jointly learning the non-linear deformations introduced by body poses *and* dexterous hand articulation *and* facial expressions. The above method to attain multiple correspondences scales poorly with the number of bones. Therefore, naïvely adding finger and face bones of SMPL-X to the initialization procedure, causes prohibitively slow training. Yet our ablations show that these are required for good animation quality (*cf*. Tab. 1). Hence, we propose a part-aware initialization strategy, in which we first separate all SMPL-X bones $\mathbf{G}$ into four groups $\mathbf{G}_B$, $\mathbf{G}_{LH}$, $\mathbf{G}_{RH}$, $\mathbf{G}_F$. For a given deformed point with part label ℓ, we then initialize the states $\{\mathbf{x}^0_{ci}\}$ and Jacobian matrices $\{\mathbf{J}^0_{ci}\}$ as:

$$\mathbf{x}^0_{ci} = \mathbf{B}_i^{-1} \cdot \mathbf{x}_d, \quad \mathbf{J}^0_{ci} = \frac{\partial \mathbf{d}_w(\mathbf{x}, \mathbf{B}_i)}{\partial \mathbf{x}}\Big|_{\mathbf{x}=\mathbf{x}^0_{ci}}, i \in \mathbf{G}_\ell. \tag{5}$$

We explain how we obtain the label ℓ for each point further below. The final occupancy prediction is determined via the maximum over all valid candidates $\mathcal{X}_c = \{\hat{\mathbf{x}}_{ci}\}_{i=1}^{|\mathbf{G}_\ell|}$:

$$o(\mathbf{x}_d, \boldsymbol{\theta}_b, \boldsymbol{\psi}) = \max_{\hat{\mathbf{x}}_c \in \mathcal{X}_c} \{f_{\text{occ}}(\hat{\mathbf{x}}_c, \boldsymbol{\theta}_b, \boldsymbol{\psi})\}. \tag{6}$$

The correspondence in canonical space is given by:

$$\mathbf{x}_c^* = \underset{\hat{\mathbf{x}}_c \in \mathcal{X}_c}{\arg\max} \{ f_{\text{occ}}(\hat{\mathbf{x}}_c, \boldsymbol{\theta}_b, \boldsymbol{\psi}) \}. \tag{7}$$

This part-aware initialization is based on the observation that a point close to a certain body part is likely to be mostly affected by the bones in that part. This scheme effectively creates four deformer networks, as shown in Fig. 2. However, note that all deformers share the same skinning weight network f_w as highlighted in Fig. 3. The only difference between them is how the iterative root finding is initialized.

Part-Aware Sampling Because hands and faces are comparatively small, while still exhibiting complex deformations, we found that a uniform sampling strategy for points $\mathbf{x}_d$ leads to poor results (cf. Tab. 1, Fig. 4). Hence, we further propose a part-aware sampling strategy, to over-sample points per area for small body parts. Assuming part labels $\mathcal{P} = \{F, LH, RH, B\}$, for each point $\mathbf{p}_i$ in the 3D scan, we first find the closest SMPL-X vertex $\mathbf{v}_i$ and store its precomputed body part label $k_i \in \mathcal{P}$. Then, for each part $\ell \in \mathcal{P}$ we extract all points $\{\mathbf{p}_i \mid k_i = \ell\}$ and re-sample the resulting mesh with a sampling rate specific to part ℓ to obtain N_ℓ many deformed points $\{\mathbf{x}_{di}\}_{i=1}^{N_\ell}$ for training.

LBS regularization To further account for the lower resolution and smaller scale of face and hands, we regularize the LBS weights of these parts to be close to the weights given by SMPL-X. A similar strategy has also been used by [67]. Our ablations show that this greatly increases the quality of the results (*cf*. Sec. 4.2).

Texture Similar to [50, 55] we introduce a third neural texture field to predict RGB values in canonical space. Its output is the color value $c(\mathbf{x}_c, \mathbf{n}_d, \mathcal{F}, \boldsymbol{\theta}_b, \boldsymbol{\psi})$. This is, in addition to pose and facial expression, the color depends on the last layer $\mathcal{F}$ of the geometry network and the normals $\mathbf{n}_d$ in deformed space. This conditions the color prediction on the deformed geometry and local high-frequency details, which has been shown to be helpful [12, 67]. Following [67], the normals are obtained via $\mathbf{n}_d = \nabla_{\mathbf{x}_d} f_{occ}(\mathbf{x}_c^*, \boldsymbol{\theta}_b, \boldsymbol{\psi})$. Therefore, the texture model f_{RGB} is formulated as:

$$f_{\text{RGB}} : \mathbb{R}^3 \times \mathbb{R}^3 \times \mathbb{R}^{512} \times \mathbb{R}^{|\boldsymbol{\theta}_b|} \times \mathbb{R}^{|\boldsymbol{\psi}|} \rightarrow \mathbb{R}^3. \tag{8}$$

We apply positional encoding to all inputs to obtain better high-frequency details following best practices [42].

3.3. Training Process

Objective Function For each 3D scan, we minimize the following objective:

$$\mathcal{L} = \mathcal{L}_{occ} + \mathcal{L}_{RGB} + \mathcal{L}_{reg}. \tag{9}$$

$\mathcal{L}_{occ}$ supervises the geometry and consists of two losses: the binary cross entropy loss $\mathcal{L}_{BCE}$ between the predicted occupancy $o(\mathbf{x}_d, \boldsymbol{\theta}_b, \boldsymbol{\psi})$ and the ground-truth value $o^{GT}(\mathbf{x}_d)$, and an L2 loss $\mathcal{L}_n$ on the normals:

$$\begin{aligned} \mathcal{L}_{occ} =& \lambda_{BCE} \mathcal{L}_{BCE} + \lambda_n \mathcal{L}_n \\ =& \lambda_{BCE} \sum_{\mathbf{x}_d \in \mathcal{P}_{\text{off}}} CE(o(\mathbf{x}_d, \boldsymbol{\theta}_b, \boldsymbol{\psi}), o^{GT}(\mathbf{x}_d)) \\ &+ \lambda_n \sum_{\mathbf{x}_d \in \mathcal{P}_{\text{on}}} \left\| \mathbf{n}_d - \mathbf{n}^{GT}(\mathbf{x}_d) \right\|_2, \end{aligned} \tag{10}$$

where $\mathcal{P}_{\text{on}}$, $\mathcal{P}_{\text{off}}$ separately denote points on the scan surface and points within a thin shell surrounding the surface [20]. $\mathcal{L}_{RGB}$ supervises the point color:

$$\mathcal{L}_{RGB} = \lambda_{RGB} \sum_{\mathbf{x}_d \in \mathcal{P}_{\text{on}}} \left\| c(\mathbf{x}_c, \mathbf{n}_d, \mathcal{F}, \boldsymbol{\theta}_b, \boldsymbol{\psi}) - c^{GT}(\mathbf{x}_d) \right\|_1. \tag{11}$$

Finally, $\mathcal{L}_{reg}$ represents the regularization term, consisting of the bone occupancy loss $\mathcal{L}_{\text{bone}}$, joint LBS weights loss $\mathcal{L}_{\text{joint}}$ and surface LBS weights loss $\mathcal{L}_{\text{surf}}$:

$$\begin{aligned} \mathcal{L}_{reg} =& \lambda_{\text{bone}} \mathcal{L}_{\text{bone}} + \lambda_{\text{joint}} \mathcal{L}_{\text{joint}} + \lambda_{\text{surf}} \mathcal{L}_{\text{surf}} \\ =& \lambda_{\text{bone}} \sum_{\mathbf{x}_c \in \mathcal{P}_{\text{bone}}^c} CE(f_{\text{occ}}(\mathbf{x}_c, \boldsymbol{\theta}_b, \boldsymbol{\psi}), 1) \\ &+ \lambda_{\text{joint}} \sum_{\mathbf{x}_c \in \mathcal{P}_{\text{joint}}^c} \sum_{i \in \mathcal{N}(i)} (w_i(\mathbf{x}_c) - 0.5)^2 \\ &+ \lambda_{\text{surf}} \sum_{\mathbf{x}_c \in \mathcal{P}_{\text{surf}}^c} \sum_{i \in \mathbf{G} \setminus \mathbf{G}_B} (w_i(\mathbf{x}_c) - w_i^{GT}(\mathbf{x}_c))^2, \end{aligned} \tag{12}$$

where $\mathcal{N}(i)$ are the neighboring bones of joint i and w_i^{GT} are the skinning weights taken from SMPL-X. $\mathcal{L}_{reg}$ makes use of the supervision from registered SMPL-X meshes. For more details on the registration, please refer to the Supp. Mat. $\mathcal{P}_{\text{bone}}^c$, $\mathcal{P}_{\text{joint}}^c$, $\mathcal{P}_{\text{surf}}^c$ refer to points sampled on the SMPL-X bones, the SMPL-X joints and from the SMPL-X mesh surface respectively. The first two terms follow the definition of [13]. We add the last term to regularize the LBS weights for fingers and face which have low resolution and are more difficult to learn.

4. Experiments

We first introduce the datasets and metrics that we use for our experiments in Sec. 4.1. Sec. 4.2 ablates all important design choices. In Sec. 4.3 we briefly describe the state-of-the-art methods to which we compare our method. Finally we show and discuss the results in Sec. 4.4-4.6. We focus on the challenging animation task, hence all the comparisons are conducted on entirely unseen poses. For completeness, we also report reconstruction results in the Supp. Mat.

ID	Method	CD↓		CD-MAX ↓		NC ↑		IoU ↑	
		All	Hands	All	Hands	All	Hands	All	Hands
A1	Ours (init w body bones)	5.42	5.05	57.54	25.10	0.940	0.824	0.964	0.812
A2	Ours (init w all bones)	4.55	4.35	44.86	20.71	0.945	0.845	**0.974**	0.811
A3	Ours (w/o part-aware sampling)	4.68	4.81	47.51	20.88	0.947	0.840	0.972	0.810
A4	Ours (w/o LBS reg.)	4.98	7.27	57.11	43.38	0.940	0.797	0.968	0.768
A	Ours (complete)	**4.46**	**4.15**	**44.36**	**20.61**	**0.948**	**0.853**	0.973	**0.829**

Table 1. **Ablation experiments for our major design choices.** We compute the metrics on the entire body (*All*) and separately on the hands (*Hands*) to better highlight the differences for the hands. All results are computed on a subset of X-Humans (Scans). Our final model (A) only marginally outperforms A2, but is roughly 3 times faster to train. For qualitative comparisons please refer to Fig. 4 and Fig. 5.

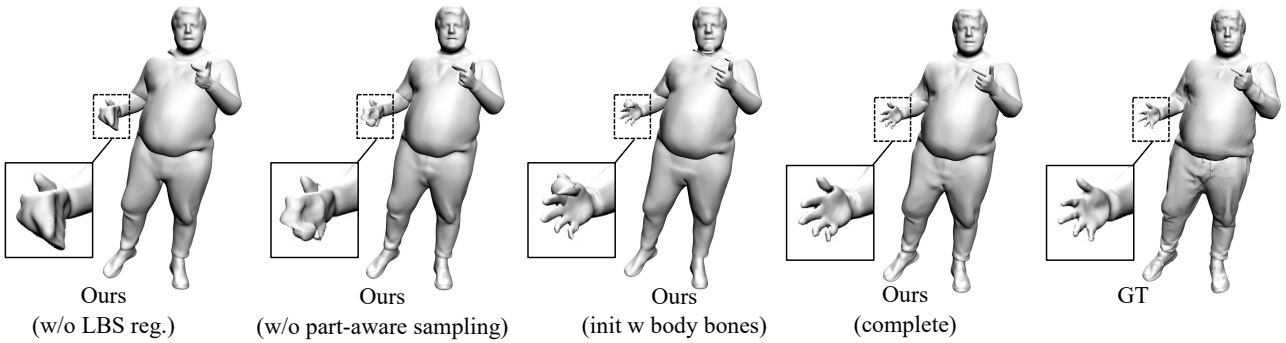

Figure 4. **Effect of our design decisions on the resulting geometry.** Notice how all baselines struggle to recover accurate hand geometry.

4.1. Datasets

GRAB [54] We use the GRAB subset of AMASS [38] for training and evaluate our model on SMPL-X meshes of minimally clothed humans. GRAB contains a diverse set of hand poses and facial expressions with several subjects. We pick the subject with the most pose variation and randomly select 9 sequences for training and three for validation. This results in 9,756 frames for training and 1,272 test frames.

X-Humans (Scans) Currently, there exists no publicly available dataset containing textured 3D clothed scans of humans with a large variation of body poses, hand gestures and facial expressions. Therefore, we captured our own dataset, for which we leveraged a high-quality, multi-view volumetric capture stage [16]. We call the resulting dataset X-Humans. It consists of 20 subjects (11 males, 9 females) with various clothing types and hair style. The collection of this dataset has been approved by an internal ethics committee. For each subject, we split the motion sequences into a training and test set. In total, there are 29,036 poses for training and 6,439 test poses. X-Humans also contains ground-truth SMPL-X parameters, obtained via a custom SMPL-X registration pipeline specifically designed to deal with low-resolution body parts. More details on the registration process and contents of X-Humans are in Supp. Mat.

X-Humans (RGB-D) We take the textured and posed scans from X-Humans and render them to obtain corresponding synthetic RGB-D images. For every time step, we render exactly one RGB-D image from a virtual camera, while the camera gradually rotates around the participant during the duration of the sequence. This is, the RGB-D version of X-Humans contains the same amount of frames as the scan version in both the training and test set.

Metrics We evaluate the geometric accuracy via volumetric IoU, Chamfer distance (CD) (mm) and normal consistency (NC) metrics, following the practice in PINA [20]. Because these metrics are dominated by large surface areas, we always report the metrics for the entire body (*All*) and the hands separately (*Hands*).

4.2. Ablation Study

Part-Aware Initialization The part-aware initialization for correspondence search is critical to accelerate training and to find good correspondences in small body parts. To verify this, we compare with two variations adapted from SNARF [13]. First, (A1) initiates the optimization states only via the body's bone transformations, while (A2) initializes using all bones (body, hands, face). **Results:** A1 suffers from strong artifacts for hands and the jaw (*cf*. Fig. 4 and Tab. 1, A1). The final model is 3 times faster than A2 (0.7 iterations per second vs. 0.25), yet it still retains high fidelity

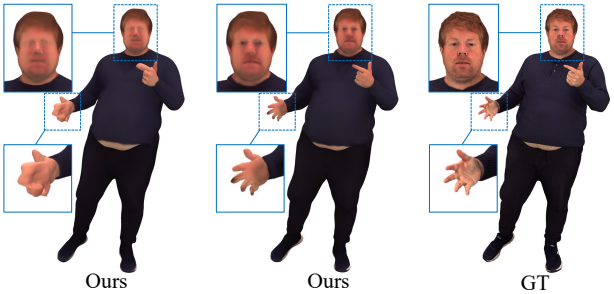

Figure 5. **Effect of our part-aware sampling strategy** on the hand geometry and texture prediction of the face.

and even outperforms A2 by a small margin (*cf*. Tab. 1, A2 and the Supp. Mat. for qualitative results). Thus we conclude that part-based initialization of the deformer is an efficient way to find accurate correspondences.

Part-Aware Sampling To verify the importance of part-aware sampling, we compare our model to a uniform sampling baseline (A3). **Results:** This component has two effects: a) it strongly improves the hand shape (*cf*. second column of Fig. 4 and Tab. 1, A3) and b) it improves texture details in the eye and mouth region (*cf*. Fig. 5).

LBS Weights Regularization for Hands and Face The first column in Fig. 4 shows that without regularizing the learned LBS weights with the SMPL-X weights, the learned hand shape is poor. This is further substantiated by a 75% increase in Chamfer distance for the hand region, compared to our final method (*cf*. Tab. 1, A4).

4.3. Baselines

Scan-based methods We compare our 3D scan-based method variant on both GRAB and X-Humans to SMPLX+D, SCANimate and SNARF baselines. All methods learn avatars in a personalized fashion, the same as ours. We adapt SMPLX+D from SMPL+D introduced in [4]. This baseline uses an explicit body model, SMPL-X, and models clothing with additive vertex offsets. To compare with SCANimate and SNARF, we use publicly available code. For details on the baselines, we refer to the Supp. Mat.

RGB-D Video-based methods We compare our RGB-D method variant on the X-Humans (RGB-D) dataset to PINA [20], a SMPL-based implicit human avatar method learned from RGB-D inputs. We assume that the ground truth pose and shape are known. For a fair comparison, we do not optimize these parameters in PINA.

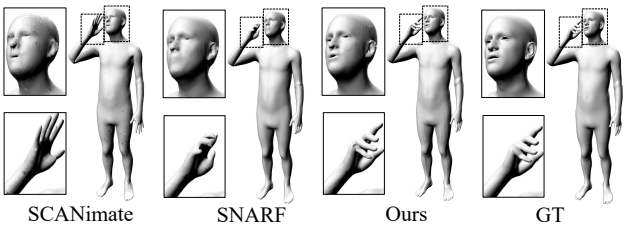

Figure 6. **Qualitative results on GRAB dataset**. Our method recovers hand articulation and facial expression most accurately.

4.4. Results on GRAB Dataset

Tab. 2 summarizes results on the GRAB dataset. Overall, our method beats all baselines, especially for the hands, where the margin is large. Fig. 6 visually shows that the quality of the hands and face learned by our method is much higher: SCANimate learns a mean hand and SNARF generalizes badly to the unseen poses. Since GRAB meshes are minimally clothed, we omit SMPLX+D from comparison.

Method	CD↓		CD-MAX ↓		NC ↑		IoU ↑	
	All	Hands	All	Hands	All	Hands	All	Hands
SCANimate [50]	2.60	8.39	54.75	54.22	0.967	0.760	0.941	0.569
SNARF [13]	1.37	5.13	33.86	33.51	0.977	0.818	0.967	0.739
Ours	**0.94**	**0.79**	**21.43**	**4.79**	**0.985**	**0.957**	**0.991**	**0.895**

Table 2. **Quantitative results on GRAB dataset.** Our method outperforms all baselines, especially for the hand part (*cf*. Fig. 6).

4.5. Results on X-Humans (Scans)

Tab. 3 shows that our method also outperforms the baselines on X-Humans. Fig. 7 qualitatively shows differences. SMPLX+D, limited by its fixed topology and low resolution, cannot model details like hair and wrinkles in clothing. SCANimate and SNARF are SMPL-driven, so they either learn a static or incomplete hand. Our method balances the different body parts so that hands are well-structured, but also the details on the face and body are maintained. Fig. 1 and Fig. 9 show more animation results.

Method	CD↓		CD-MAX ↓		NC ↑		IoU ↑	
	All	Hands	All	Hands	All	Hands	All	Hands
SMPLX+D	5.75	5.19	48.41	23.48	0.921	0.790	0.957	0.774
SCANimate [50]	6.54	9.78	59.71	48.32	0.925	0.726	0.919	0.557
SNARF [13]	5.05	7.23	55.06	37.15	0.934	0.788	0.937	0.608
Ours	**4.43**	**5.14**	**47.56**	**22.15**	**0.939**	**0.793**	**0.965**	**0.776**

Table 3. **Quantitative results on X-Humans (Scans).** We beat all baselines both for the entire body (*All*) and hands only (*Hands*).

Method	CD↓		CD-MAX ↓		NC ↑		IoU ↑	
	All	Hands	All	Hands	All	Hands	All	Hands
PINA [20]	5.41	9.51	66.05	48.07	0.928	0.771	0.910	0.566
Ours	**5.33**	**5.27**	**51.73**	**22.86**	**0.936**	**0.797**	**0.947**	**0.768**

Table 4. **Quantitative results on X-Humans (RGB-D).** Our method outperforms PINA in all metrics. Improvements are more pronounced for hands (*cf*. Fig. 8 for visual comparison).

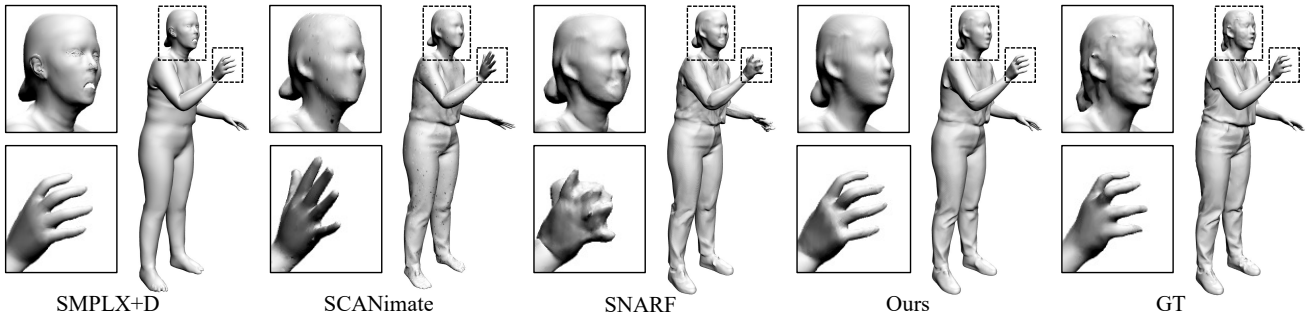

Figure 7. **Qualitative animation comparison on X-Humans (Scans).** SMPLX+D fails to model face and garment details. SCANimate and SNARF generate poor hands (static or incomplete). Our method produces the most plausible face and hands, and keeps the clothing details comparable to strong baselines.

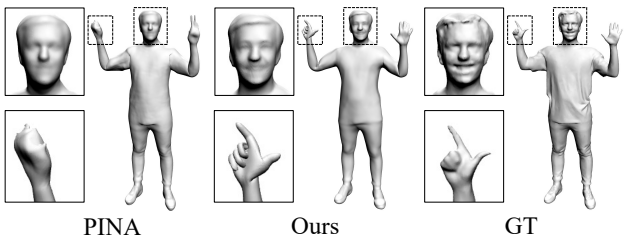

Figure 8. **X-Avatars created from RGB-D input compared to PINA.** Notice how we obtain better hand and face geometry.

4.6. Results on X-Humans (RGB-D)

Tab. 4 shows our model's performance compared to PINA [20]. Our method outperforms PINA on all metrics. Fig. 8 further qualitatively shows that without utilizing the hand and face information in the modelling process, the face and hands produced by PINA are not consistent with the input pose. Our model generates a) more realistic faces as the shape network is conditioned on facial expression and b) better hand poses because we initialize the root finding with hand bone transformations.

5. Conclusion

Limitations X-Avatar struggles to model loose clothing that is far away from the body (*e.g.* skirts). Furthermore, generalization capability beyond a single person is still limited, *i.e.* we train one model for each subject. The inference speed ($\approx$ 7 seconds per frame) is yet not optimal but a faster version of SNARF [11] is a drop-in replacement for our deformers and will accelerate our method significantly.

Conclusion We propose X-Avatar, the first expressive implicit human avatar model that captures body/hand poses, facial expressions and appearance holistically. We have demonstrated our method's expressive power and the capability of creating it from multiple input modalities with the aid of our newly introduced X-Humans dataset. We believe that our method along with X-Humans will promote further scientific research in creating expressive digital avatars.

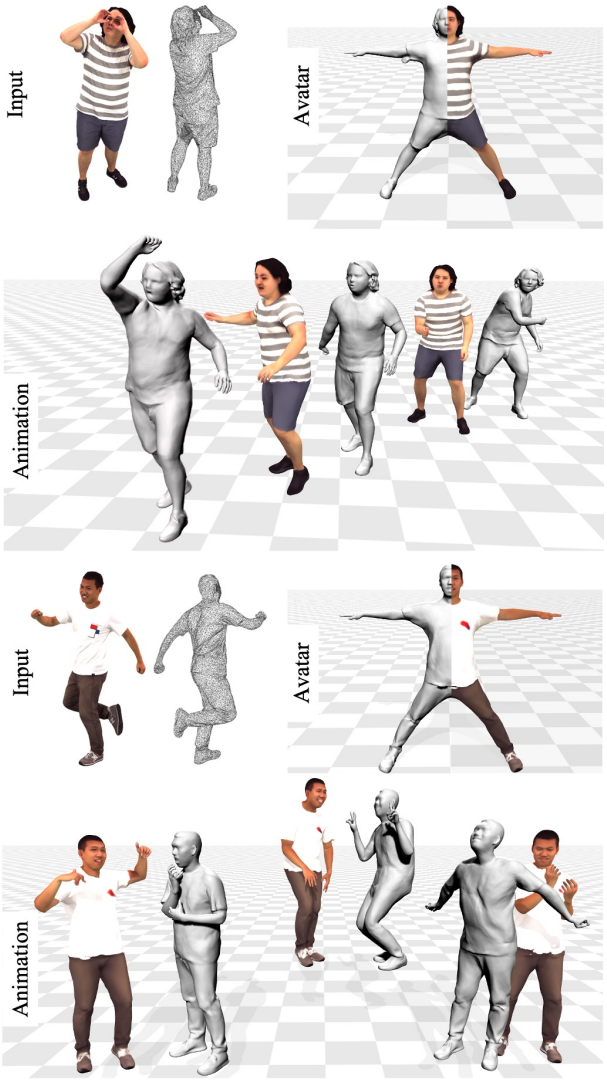

Figure 9. **Animation demonstration on X-Humans (Scans).** Our method can handle relatively complex clothing patterns, hair styles, and varied facial expressions, hand, and body poses.

References

[1] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1175–1186, 2019. 2

[2] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3d people models. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8387–8397, 2018. 2

[3] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. Scape: Shape completion and animation of people. *ACM Trans. Graph.*, 24(3):408–416, jul 2005. 2

[4] Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Combining implicit function learning and parametric models for 3d human reconstruction. In *European Conference on Computer Vision (ECCV)*. Springer, aug 2020. 7

[5] Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. Multi-garment net: Learning to dress 3d people from images. In *IEEE International Conference on Computer Vision (ICCV)*. IEEE, Oct 2019. 3

[6] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '99, page 187–194, USA, 1999. ACM Press/Addison-Wesley Publishing Co. 2

[7] Federica Bogo, Michael J. Black, Matthew Loper, and Javier Romero. Detailed full-body reconstructions of moving people from monocular rgb-d sequences. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2300–2308, 2015. 2

[8] Adnane Boukhayma, Rodrigo de Bem, and Philip HS Torr. 3d hand shape and pose from images in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10843–10852, 2019. 2

[9] Samarth Brahmbhatt, Cusuh Ham, Charles C. Kemp, and James Hays. ContactDB: Analyzing and predicting grasp contact via thermal imaging. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[10] Charles G Broyden. A class of methods for solving nonlinear simultaneous equations. *Mathematics of computation*, 19(92):577–593, 1965. 4

[11] Xu Chen, Tianjian Jiang, Jie Song, Max Rietmann, Andreas Geiger, Michael J. Black, and Otmar Hilliges. Fast-snarf: A fast deformer for articulated neural fields, 2022. 8

[12] Xu Chen, Tianjian Jiang, Jie Song, Jinlong Yang, Michael J. Black, Andreas Geiger, and Otmar Hilliges. gdna: Towards generative detailed neural avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20427–20437, June 2022. 5

[13] Xu Chen, Yufeng Zheng, Michael J Black, Otmar Hilliges, and Andreas Geiger. Snarf: Differentiable forward skinning for animating non-rigid neural implicit shapes. In *International Conference on Computer Vision (ICCV)*, 2021. 2, 3, 4, 5, 6, 7

[14] Yin Chen, Z. Cheng, Chao Lai, Ralph Robert Martin, and Gang Dang. Realtime reconstruction of an animating human body from a single depth camera. *IEEE Transactions on Visualization and Computer Graphics*, 22:2000–2011, 2016. 2

[15] Vasileios Choutas, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. Monocular expressive body regression through body-driven attention. In *European Conference on Computer Vision (ECCV)*, 2020. 2

[16] Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese, Hugues Hoppe, Adam Kirk, and Steve Sullivan. High-quality streamable free-viewpoint video. *ACM Trans. Graph.*, 34(4), jul 2015. 2, 6

[17] Enric Corona, Tomas Hodan, Minh Vo, Francesc Moreno-Noguer, Chris Sweeney, Richard Newcombe, and Lingni Ma. Lisa: Learning implicit shape and appearance of hands. In *CVPR*, 2022. 2, 3

[18] Enric Corona, Albert Pumarola, Guillem Alenyà, Gerard Pons-Moll, and Francesc Moreno-Noguer. Smplicit: Topology-aware generative model for clothed people. In *CVPR*, 2021. 3

[19] Boyang Deng, JP Lewis, Timothy Jeruzalski, Gerard Pons-Moll, Geoffrey Hinton, Mohammad Norouzi, and Andrea Tagliasacchi. Neural articulated shape approximation. In *The European Conference on Computer Vision (ECCV)*. Springer, August 2020. 3

[20] Zijian Dong, Chen Guo, Jie Song, Xu Chen, Andreas Geiger, and Otmar Hilliges. Pina: Learning a personalized implicit neural avatar from a single rgb-d video sequence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20470–20480, June 2022. 2, 3, 4, 5, 6, 7, 8

[21] Yao Feng, Vasileios Choutas, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. Collaborative regression of expressive bodies using moderation. In *International Conference on 3D Vision (3DV)*, 2021. 2

[22] Guy Gafni, Justus Thies, Michael Zollhöfer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8649–8658, June 2021. 2, 3

[23] Chen Guo, Xu Chen, Jie Song, and Otmar Hilliges. Human performance capture from monocular video in the wild. In *2021 International Conference on 3D Vision (3DV)*, pages 889–898. IEEE, 2021. 2

[24] Chen Guo, Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. Vid2avatar: 3d avatar reconstruction from videos in the wild via self-supervised scene decomposition. *arXiv*, 2023. 3

[25] Yana Hasson, Gül Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11799–11808, 2019. 2

[26] Yinghao Huang, Manuel Kaufmann, Emre Aksan, Michael J. Black, Otmar Hilliges, and Gerard Pons-Moll. Deep inertial poser: Learning to reconstruct human pose from sparse inertial measurements in real time. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 37:185:1–185:15, Nov. 2018.

2

[27] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3334–3342, 2015. 3

[28] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2, 3

[29] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[30] Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. PARE: Part attention regressor for 3D human body estimation. In *Proc. International Conference on Computer Vision (ICCV)*, pages 11127–11137, Oct. 2021.

[31] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *ICCV*, 2019.

[32] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3383–3393, 2021. 2

[33] Ruilong Li, Julian Tanke, Minh Vo, Michael Zollhofer, Jurgen Gall, Angjoo Kanazawa, and Christoph Lassner. Tava: Template-free animatable volumetric actors. 2022. 4

[34] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), 2017. 2

[35] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, Oct. 2015. 2, 3

[36] Qianli Ma, Shunsuke Saito, Jinlong Yang, Siyu Tang, and Michael J. Black. SCALE: Modeling clothed humans with a surface codec of articulated local elements. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 16082–16093, 2021. 3

[37] Qianli Ma, Jinlong Yang, Anurag Ranjan, Sergi Pujades, Gerard Pons-Moll, Siyu Tang, and Michael J. Black. Learning to dress 3d people in generative clothing. In *Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2, 3

[38] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision*, pages 5442–5451, Oct. 2019. 3, 6

[39] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4460–4470, 2019. 3

[40] Marko Mihajlovic, Shunsuke Saito, Aayush Bansal, Michael Zollhoefer, and Siyu Tang. COAP: Compositional articulated occupancy of people. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2022. 3

[41] Marko Mihajlovic, Yan Zhang, Michael J Black, and Siyu Tang. LEAP: Learning articulated occupancy of people. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2021. 3

[42] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 3, 5

[43] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Occupancy flow: 4d reconstruction by learning particle dynamics. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5379–5389, 2019. 3

[44] Ahmed A A Osman, Timo Bolkart, and Michael J. Black. STAR: A sparse trained articulated human body regressor. In *European Conference on Computer Vision (ECCV)*, pages 598–613, 2020. 2

[45] Ahmed A A Osman, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. SUPR: A sparse unified part-based human body model. In *European Conference on Computer Vision (ECCV)*, 2022. 2

[46] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 165–174, 2019. 3

[47] 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 *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 3

[48] Eduard Ramon, Gil Triginer, Janna Escur, Albert Pumarola, Jaime Garcia, Xavier Giro-i Nieto, and Francesc Moreno-Noguer. H3d-net: Few-shot high-fidelity 3d head reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5620–5629, 2021. 3

[49] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6):245:1–245:17, Nov. 2017. 2

[50] Shunsuke Saito, Jinlong Yang, Qianli Ma, and Michael J. Black. SCANimate: Weakly supervised learning of skinned clothed avatar networks. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2021. 2, 3, 5, 7

[51] Jie Song, Xu Chen, and Otmar Hilliges. Human body model fitting by learned gradient descent. 2020. 2

[52] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Black Michael J., and Tao Mei. Monocular, one-stage, regression of multiple 3d people. In *ICCV*, 2021.

[53] Yu Sun, Wu Liu, Qian Bao, Yili Fu, Tao Mei, and Michael J Black. Putting people in their place: Monocular regression of 3d people in depth. In *CVPR*, 2022. 2

[54] Omid Taheri, Nima Ghorbani, Michael J. Black, and Dim-

itrios Tzionas. GRAB: A dataset of whole-body human grasping of objects. In *European Conference on Computer Vision (ECCV)*, 2020. 2, 3, 6

[55] Garvita Tiwari, Nikolaos Sarafianos, Tony Tung, and Gerard Pons-Moll. Neural-gif: Neural generalized implicit functions for animating people in clothing. In *International Conference on Computer Vision (ICCV)*, October 2021. 2, 3, 5

[56] Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *European Conference on Computer Vision (ECCV)*, sep 2018. 2

[57] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J. Black. ICON: Implicit Clothed humans Obtained from Normals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13296–13306, June 2022. 3

[58] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6184–6193, 2020. 2

[59] T Yenamandra, A Tewari, F Bernard, HP Seidel, M Elgharib, D Cremers, and C Theobalt. i3dmm: Deep implicit 3d morphable model of human heads. In *Proceedings of the IEEE / CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021. 3

[60] Xinyu Yi, Yuxiao Zhou, Marc Habermann, Soshi Shimada, Vladislav Golyanik, Christian Theobalt, and Feng Xu. Physical inertial poser (pip): Physics-aware real-time human motion tracking from sparse inertial sensors. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022. 2

[61] Tao Yu, Zerong Zheng, Kaiwen Guo, Jianhui Zhao, Qionghai Dai, Hao Li, Gerard Pons-Moll, and Yebin Liu. Doublefusion: Real-time capture of human performances with inner body shapes from a single depth sensor. In *The IEEE International Conference on Computer Vision and Pattern Recognition(CVPR)*. IEEE, June 2018. 2

[62] Ye Yuan, Umar Iqbal, Pavlo Molchanov, Kris Kitani, and Jan Kautz. Glamr: Global occlusion-aware human mesh recovery with dynamic cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[63] 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 *Computer Vision – ECCV 2020*, pages 465–481, 2020. 2

[64] Chao Zhang, Sergi Pujades, Michael J. Black, and Gerard Pons-Moll. Detailed, accurate, human shape estimation from clothed 3d scan sequences. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 3

[65] Hongwen Zhang, Yating Tian, Yuxiang Zhang, Mengcheng Li, Liang An, Zhenan Sun, and Yebin Liu. Pymaf-x: Towards well-aligned full-body model regression from monocular images. *arXiv preprint arXiv:2207.06400*, 2022. 2

[66] Hongwen Zhang, Yating Tian, Xinchi Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun. Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop. In *Proceedings of the IEEE International Conference on Computer Vision*, 2021. 2

[67] Yufeng Zheng, Victoria Fernández Abrevaya, Marcel C. Bühler, Xu Chen, Michael J. Black, and Otmar Hilliges. I M Avatar: Implicit morphable head avatars from videos. In *Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 3, 4, 5