X-Avatar: Expressive Human Avatars

Kaiyue Shen

Master Thesis
April 2022

Supervisors:
Chen Guo
Dr. Jie Song
Prof. Dr. Otmar Hilliges

ETH zürich

Advanced Interactive Technologies

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 in both data domains both quantitatively and qualitatively on the animation task. To facilitate future research on expressive avatars we contribute a new dataset, called X-Humans, containing 244 sequences of high-quality textured scans from 21 participants, totalling 37,005 data frames.

Contents

List of Figures

List of Tables

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.

Parametric body models such as the SMPL family [Loper et al. 2015, Romero et al. 2017, Pavlakos et al. 2019] 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. [Chen et al. 2021] 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 [Chen et al. 2021] only models the major body bones, other works have focused on creating implicit models of the face [Zheng et al. 2022, Gafni et al. 2021], the hands [Corona et al. 2022], or how to model humans that appear in garments [Dong et al. 2022] and how to additionally capture appearance [Saito et al. 2021, Tiwari et al. 2021]. 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 ex-

pressions, and clothing. To this end we adopt the full-body pose space of SMPL-X [Pavlakos et al. 2019]. This causes two key challenges for learning X-Avatars from data: (i) the significantly increased number of involved articulated body parts (9 used by [Chen et al. 2021] 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 [Chen et al. 2021] 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 [Saito et al. 2021], SNARF [Chen et al. 2021]) on the GRAB dataset [Taheri et al. 2020, Brahmbhatt et al. 2019] on the animation task of minimally clothed humans.

Second, we contribute a novel dataset consisting of 244 sequences of 21 clothed participants recorded in a high-quality volumetric capture stage [Collet et al. 2015]. 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 be learned from either 3D scans or (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 244 sequences, of high-quality textured scans showing 21 participants with varied body and hand movements, and facial expressions, totalling 37,005 frames.

Data, models and SMPL[-X] registrations, will be released for scientific purposes.

Related Work

2.1. 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 [Loper et al. 2015, Anguelov et al. 2005, Osman et al. 2020], the hands [Romero et al. 2017], or the face [Blanz and Vetter 1999, Li et al. 2017], while others treat the human more holistically like we do in this work, *e.g*. [Pavlakos et al. 2019, Xu et al. 2020, Zanfir et al. 2020, Osman et al. 2022, Joo et al. 2018]. 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 [Feng et al. 2021, Sun et al. 2021, Kocabas et al. 2021, Li et al. 2021, Choutas et al. 2020, Song et al. 2020, Kolotouros et al. 2019, Yuan et al. 2022, Sun et al. 2022, Kanazawa et al. 2018], RGB-D fitting [Yu et al. 2018, Chen et al. 2016, Bogo et al. 2015], fitting to body-worn sensor data [von Marcard et al. 2018, Huang et al. 2018, Yi et al. 2022], or 3D hand pose estimation [Hasson et al. 2019, Boukhayma et al. 2019] 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 [Alldieck et al. 2018, Alldieck et al. 2019a], also dubbed SMPL+D, pose-dependent 3D offsets [Ma et al. 2020], by modelling 3D garments and draping them over the SMPL mesh [Corona et al. 2021, Bhatnagar et al. 2019] or via local small surface patches [Ma et al. 2021]. Explicit models have seen a trend towards unification to model human expressiveness, *e.g*. SMPL-X [Pavlakos et al. 2019] and Adam [Joo et al. 2018]. X-Avatar shares this goal, but for implicit models.

2.2. 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 [Park et al. 2019], density [Mildenhall et al. 2020], or occupancy [Mescheder et al. 2019] given a point in space. To extend this idea to articulated shapes like the human body, NASA [Deng et al. 2020] used per body-part occupancy networks [Mescheder et al. 2019]. This per-part formulation creates artifacts, especially for unseen poses, which works such as [Chen et al. 2021, Mihajlovic et al. 2021, Mihajlovic et al. 2022] improve. SNARF [Chen et al. 2021] does so via a forward warping field which is compatible with the SMPL [Loper et al. 2015] skeleton, learns pose-independent skinning and generalizes well to unseen poses and people in clothing. Other works [Saito et al. 2021, Tiwari et al. 2021] model appearance and are learned from scans. [Xiu et al. 2022] creates avatars from video by relying on normals extracted from a 3D body model fitted to images and [Dong et al. 2022] does so from RGB-D video.

Moving beyond bodies, other work has investigated implicit models for faces [Zheng et al. 2022, Yenamandra et al. 2021, Ramon et al. 2021, Gafni et al. 2021] and hands [Corona et al. 2022]. 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 [Chen et al. 2021] driven by SMPL-X [Pavlakos et al. 2019]. 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 [Saito et al. 2021], 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 [Saito et al. 2021] for the hands and faces. Furthermore, in contrast to [Saito et al. 2021, Tiwari et al. 2021, Chen et al. 2021], X-Avatars can be fit to 3D scans *and* RGB-D videos.

2.3. Human Datasets

Publicly available datasets that show the full range of human expressiveness and contain clothed and textured ground-truth are rare. GRAB [Taheri et al. 2020], a subset of AMASS [Mahmood et al. 2019], contains minimally clothed SMPL-X registrations. BUFF [Zhang et al. 2017] and CAPE [Ma et al. 2020] do not model appearance and facial expressions. The CMU Panoptic Studio [Joo et al. 2015] dataset was used to fit Adam [Joo et al. 2018] which does model hands and faces, but is neither textured nor clothed. Also, [Joo et al. 2015] does not contain scans. To study X-Avatars on real clothed humans, we thus contribute our own dataset, X-Humans which contains 37,005 frames of high-quality, texturized scans of real clothed humans with corresponding SMPL[-X] registrations.

SMPL-X Registration

From our Volumetric Capture Stage, we get high-quality 3D scans and RGB images of 53 camera views. We fit an SMPL-X model [Pavlakos et al. 2019] to each scan. The gender-specific model $M(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\psi})$ is parameterized by the whole body pose $\boldsymbol{\theta}$, body shape $\boldsymbol{\beta}$, 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 $\boldsymbol{\theta}_b$. Before capture, our subjects indicate their gender on a questionnaire, so we subsequently use the corresponding gender-specific SMPL-X model for the fitting.

Our SMPL-X registration pipeline has three steps: 1. 2D landmarks detection (Sec. 3.1); 2. 3D landmarks generation (Sec. 3.2); 3. multi-stage fitting (Sec. 3.3).

3.1. 2D Landmarks Detection

As shown in the left part of Fig. 3.1, in the 2D landmarks detection stage, we first render the 3D scans with known camera parameters to get the corresponding binary human mask. Then we predict a tight bounding box from the mask and use it to crop out the human part from the RGB images. We resize the crop to fit the aspect-ratio of images expected by OpenPose [Cao et al. 2019, Simon et al. 2017]. We then feed the cropped and resized image into OpenPose to get 2D full body landmarks including the body keypoints, hand keypoints and facial landmarks. The cropping improves the resolution of the human body and the following resizing operation makes the image ratio more similar to OpenPose training images, so as to improve the detection results.

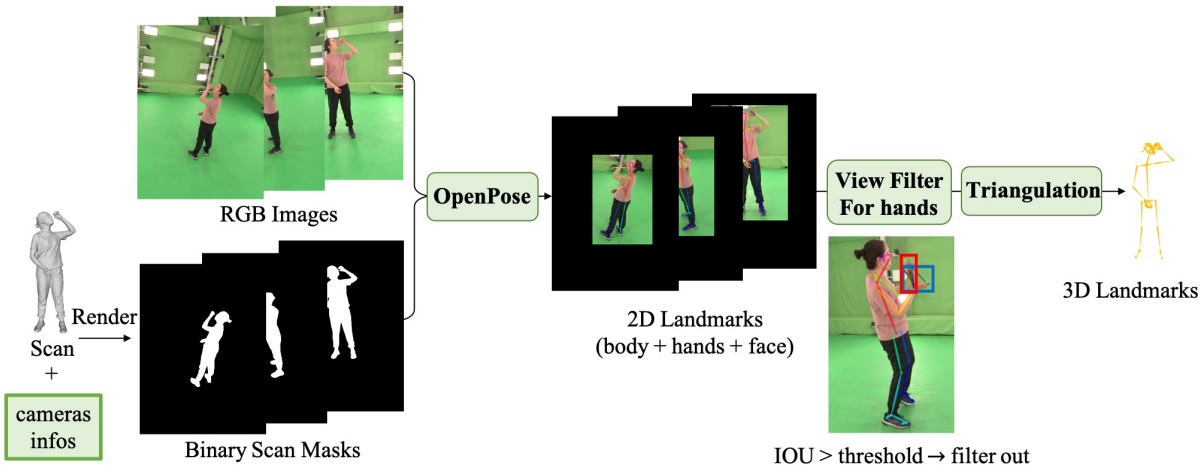

Figure 3.1.: Full-body Landmarks Generation Overview. *It consists of two steps: (a) 2D landmarks detection using OpenPose on cropped images; (b) 3D landmarks generation via triangulation. We filter out views which show a large overlap of 2D hand landmarks.*

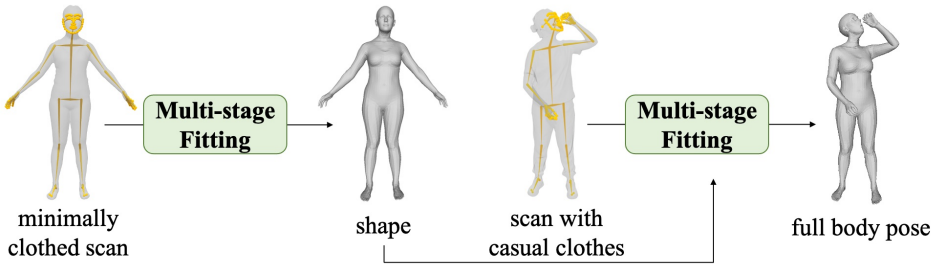

Figure 3.2.: Registration Overview. *For each subject, we first obtain the shape by running the multi-stage fitting pipeline (c.f. Tab. 3.1) once on minimally clothed scans (5 frames). Then we fix the shape and run the pipeline once more on scans where subjects wear casual clothing to get full body pose and facial expressions. By disentangling the optimization of ground-truth shape and ground-truth pose, we can dissolve the shape ambiguity caused by loose clothing.*

3.2. 3D Landmarks Generation

As shown in the right part of Fig. 3.1, in the 3D landmarks generation stage, we first pass the detected 2D landmarks through a view filter for hands and then use triangulation to get 3D full body landmarks. To be more specific, for each camera view, we first use the hand keypoints to estimate the tight bounding box for each hand, and compute the Intersection over Union (IoU) of the two bounding boxes. If the IoU is larger than the given threshold, we will set the confidence of all hand keypoints in this view to be 0, which means these 2D hand keypoints are ignored during the computation of the 3D hand keypoints. The reason behind this filter is that it is very likely for OpenPose to return bad predictions when there exists strong occlusions. The wrong 2D detection may further result in poor 3D landmarks, which is why we simply filter them out.

ID	Description	Optimized parameters	Losses
1	optimize pose	$\boldsymbol{\theta}_g$, $\boldsymbol{\theta}_b$	$\mathcal{L}_J$, $\mathcal{L}_{\boldsymbol{\theta}_b}$
2	optimize pose, shape	$\boldsymbol{\theta}_g$, $\boldsymbol{\theta}_b$, $\boldsymbol{\beta}$	$\mathcal{L}_J$, $\mathcal{L}_S$, $\mathcal{L}_{\boldsymbol{\theta}_b}$, $\mathcal{L}_{\boldsymbol{\beta}}$
3	refine pose, shape	$\boldsymbol{\theta}_g$, $\boldsymbol{\theta}_b$, $\boldsymbol{\beta}$	$\mathcal{L}_J$, $\mathcal{L}_S$, $\mathcal{L}_{\boldsymbol{\theta}_b}$, $\mathcal{L}_{\boldsymbol{\beta}}$, $\mathcal{L}_{reg}$
4	refine body pose	$\boldsymbol{\theta}_b$	$\mathcal{L}_J$, $\mathcal{L}_S$, $\mathcal{L}_{\boldsymbol{\theta}_b}$, $\mathcal{L}_{reg}$
5	refine hands	$\boldsymbol{\theta}_h$	$\mathcal{L}_{J_h}$, $\mathcal{L}_{\boldsymbol{\theta}_h}$
6	refine face	$\boldsymbol{\theta}_f$, $\boldsymbol{\psi}$	$\mathcal{L}_{J_f}$, $\mathcal{L}_{\boldsymbol{\theta}_f}$, $\mathcal{L}_{\boldsymbol{\psi}}$

Table 3.1.: Details of Multi-Stage fitting pipeline. *We propose a coarse-to-fine fitting pipeline where we first optimize for global pose $\boldsymbol{\theta}_g$, body pose $\boldsymbol{\theta}_b$ and shape $\boldsymbol{\beta}$ (stage 1-2). We then refine those parameters in stages 3-4, before moving the smaller parts of the body,* i.e. *hands $\boldsymbol{\theta}_h$, face $\boldsymbol{\theta}_f$ and facial expressions $\boldsymbol{\psi}$. $\mathcal{L}_J, \mathcal{L}_{J_h}, \mathcal{L}_{J_f}$ are data terms that penalize differences between body, hands, and face 3D landmarks. $\mathcal{L}_S$ minimizes the point-to-point distance between the scan and SMPL-X vertices. $\mathcal{L}_{reg}$ is the interpenetration loss that encourages the SMPL-X to be inside the scan, following [Alldieck et al. 2019b]. $\mathcal{L}_{\boldsymbol{\theta}_b}$, $\mathcal{L}_{\boldsymbol{\theta}_h}$, $\mathcal{L}_{\boldsymbol{\theta}_f}$ penalize unrealistic bending of the torso, hands, and face joints (*c.f. *Supp. Mat). $\mathcal{L}_{\boldsymbol{\beta}}, \mathcal{L}_{\boldsymbol{\psi}}$ are L2 regularizers on the body shape and facial expressions. The shape $\boldsymbol{\beta}$ is only optimized for the minimally-clothed sequence (*c.f. *Fig. 3.2). Registration is more robust in this coarse-to-fine manner.*

3.3. Multi-Stage Fitting

For fitting, similar to [Patel et al. 2021, Alldieck et al. 2019b], we adopt a multi-stage pipeline, shown in more detail in Tab. 3.1. First we initiate the SMPL-X parameters with the sparse 3D landmarks obtained from multi-view RGB images as described in Sec. 3.2. Then we refine the body pose and shape with dense surface information coming from the scan meshes. Finally, we refine the hand pose and facial expressions with the 3D landmarks.

The registration pipeline must deal with shape ambiguity caused by loose clothing. Instead of using the time-consuming skin-cloth segmentation as is the practice in [Patel et al. 2021], we disentangle the optimization of ground-truth shape and ground-truth poses as shown in Fig. 3.2. This is based on the assumption that the same person wearing different clothes still shares the same body shape. We first run the multi-stage fitting pipeline on minimally clothed scans, which is a short 5-frame sequence where the participants wear tight-fitting clothes. Then, for a regular sequence where participants wear their casual clothes, we initiate the multi-stage fitting pipeline with the previously learned shape parameters that then remains fixed during the optimization of full body pose, hand pose and facial expressions.

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. 4.1 and Fig. 4.2. 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 end (*c.f.* Sec. 4.4).

4.1. Recap: SMPL-X Unified Human Body Model

Our goal is to create fully controllable human avatars. We use the parameter space of SMPL-X [Pavlakos et al. 2019], 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, $|\boldsymbol{\theta}_g| = 3$, $|\boldsymbol{\theta}_b| = 63$, $|\boldsymbol{\theta}_h| = 90$, $|\boldsymbol{\theta}_f| = 9$, $|\boldsymbol{\beta}| = 10$, $|\boldsymbol{\psi}| = 10$, $N = 10,475$.

4.2. 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

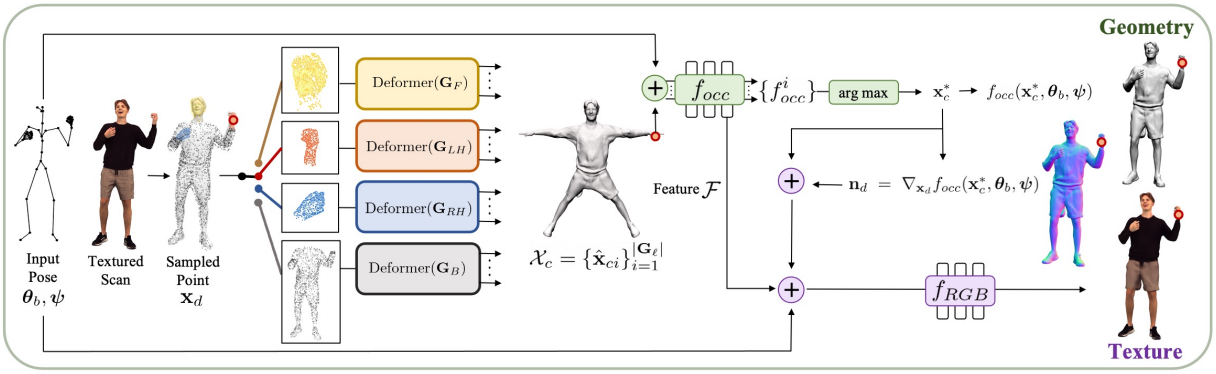

Figure 4.1.: Method Overview. *Given a posed scan with an SMPL-X registration, we first adaptively sample points $\mathbf{x}_d$ in deformed space per body part ℓ (face F, left hand LH, right hand RH, body B). A part-specific deformer network finds the corresponding candidate points $\hat{\mathbf{x}}_{ci}$ (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 $\mathbf{B}_\ell$ (cf. Fig. 4.2). 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 $\nabla_{\mathbf{x}_d} f_{occ}(\mathbf{x}^*_c, \boldsymbol{\theta}_b, \boldsymbol{\psi})$.*

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.

4.2.1. 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 [Niemeyer et al. 2019] are applied to the input. Hence, the shape model f_{occ} is denoted by:

$$f_{\text{occ}} : \mathbb{R}^3 \times \mathbb{R}^{|\boldsymbol{\theta}_b|} \times \mathbb{R}^{|\boldsymbol{\psi}|} \rightarrow [0, 1]. \tag{4.1}$$

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

$$\mathcal{S} = \{\, \mathbf{x}_c \mid f_{\text{occ}}(\mathbf{x}_c, \boldsymbol{\theta}_b, \boldsymbol{\psi}) = 0.5 \,\}. \tag{4.2}$$

4.2.2. Deformation

To model skeletal deformation, we follow previous work [Chen et al. 2021, Li et al. 2022, Dong et al. 2022, Zheng et al. 2022] and represent the skinning weight field in the canonical space by an MLP:

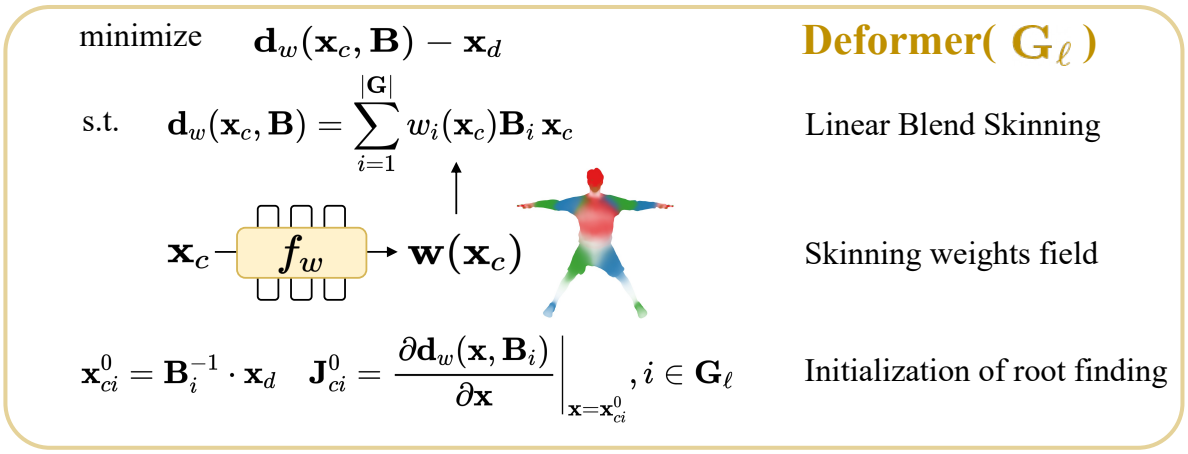

Figure 4.2.: Part-specific Deformer. *Each deformer shown in Fig. 4.1 is initialized with the bone transformations belonging to a specific part* $\mathbf{G}_\ell, \ell \in \{F, LH, RH, B\}$*, but shares the parameters of* f_w*.*

$$f_w : \mathbb{R}^3 \rightarrow \mathbb{R}^{n_b} \times \mathbb{R}^{n_h} \times \mathbb{R}^{n_f}, \tag{4.3}$$

where n_b, n_h, n_f denotes the number of body, finger, and face bones respectively. Similar to [Chen et al. 2021], 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:

$$\mathbf{x}_d = \mathbf{d}_w(\mathbf{x}_c, \mathbf{B}) = \sum_{i=1}^{|\mathbf{G}|} w_i(\mathbf{x}_c)\mathbf{B}_i\,\mathbf{x}_c. \tag{4.4}$$

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 [Chen et al. 2021] and employ correspondence search. We use Broyden's method [Broyden 1965] 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.

4.2.3. 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 (*c.f.* Tab. 5.1). 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 $\{\mathbf{x}_{ci}^0\}$ and Jacobian matrices $\{\mathbf{J}_{ci}^0\}$ as:

$$\mathbf{x}_{ci}^0 = \mathbf{B}_i^{-1} \cdot \mathbf{x}_d, \quad \mathbf{J}_{ci}^0 = \left.\frac{\partial \mathbf{d}_w(\mathbf{x}, \mathbf{B}_i)}{\partial \mathbf{x}}\right|_{\mathbf{x}=\mathbf{x}_{ci}^0}, i \in \mathbf{G}_\ell. \tag{4.5}$$

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 $\mathcal{X}_c = \{\hat{\mathbf{x}}_{ci}\}_{i=1}^{|\mathbf{G}_\ell|}$:

$$o(\mathbf{x}_d, \boldsymbol{\theta}_b, \boldsymbol{\psi}) = \max_{\hat{\mathbf{x}}_c \in \mathcal{X}_c} \{f_{\text{occ}}(\hat{\mathbf{x}}_c, \boldsymbol{\theta}_b, \boldsymbol{\psi})\}. \tag{4.6}$$

The correspondence in canonical space is given by:

$$\mathbf{x}_c^* = \arg\max_{\hat{\mathbf{x}}_c \in \mathcal{X}_c} \{f_{\text{occ}}(\hat{\mathbf{x}}_c, \boldsymbol{\theta}_b, \boldsymbol{\psi})\}. \tag{4.7}$$

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. 4.1. However, note that all deformers share the same skinning weight network f_w as highlighted in Fig. 4.2. The only difference between them is how the iterative root finding is initialized.

4.2.4. 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. 5.1, Fig. 5.2). Hence, we further propose a part-aware sampling strategy, to over-sample points per area for small body parts. The detailed part-aware sampling pipeline is shown in Fig. 4.3. Since the official document provides the FLAME and MANO vertex indices of the SMPL-X body and SMPL-X has the fixed topology, for every SMPL-X mesh, the part label for each vertex on the mesh is known. Assuming part labels $\mathcal{P} = \{F, LH, RH, B\}$, for each vertex $\mathbf{p}_i$ on the 3D scan mesh, its part label can be computed by finding the label of its closest SMPL-X vertex $\mathbf{v}_i$. We store the pre-computed 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.

4.2.5. LBS Regularization

To further account for the lower resolution and smaller scale of the 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 [Zheng et al. 2022]. Our ablations show that this greatly increases the quality of the results (*c.f.* Sec. 5.2).

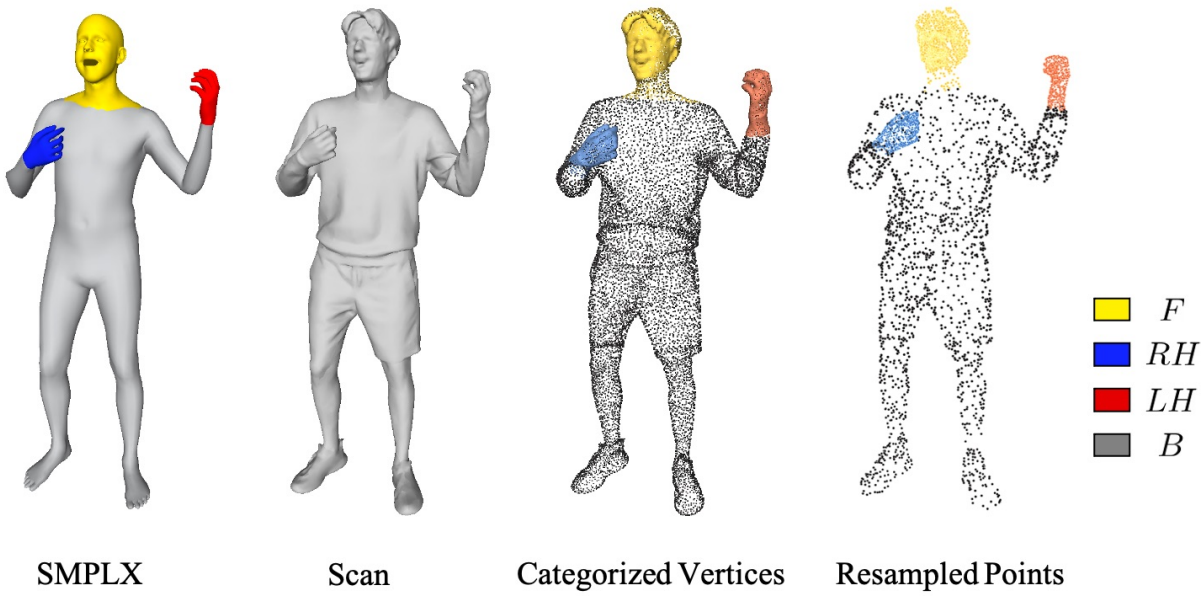

Figure 4.3.: Part-Aware Sampling Pipeline. *From left to right: 1. Categorized SMPL-X mesh (Known), 2. Corresponding scan mesh, 3. Categorized scan mesh vertices (by finding NN on SMPL-X mesh), 4. Points sampled from categorized vertices by applying a sampling rate specific to part label (used for training).*

4.2.6. Texture

Similar to [Saito et al. 2021, Tiwari et al. 2021] 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 [Chen et al. 2022, Zheng et al. 2022]. Following [Zheng et al. 2022], 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:

$$f_{\text{RGB}} : \mathbb{R}^3 \times \mathbb{R}^3 \times \mathbb{R}^{512} \times \mathbb{R}^{|\boldsymbol{\theta}_b|} \times \mathbb{R}^{|\boldsymbol{\psi}|} \rightarrow \mathbb{R}^3. \tag{4.8}$$

We apply positional encoding to all inputs to obtain better high-frequency details following best practices [Mildenhall et al. 2020].

4.3. Training Process

4.3.1. Objective Function

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

$$\mathcal{L} = \mathcal{L}_{occ} + \mathcal{L}_{RGB} + \mathcal{L}_{reg}. \tag{4.9}$$

$\mathcal{L}_{occ}$ 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:

$$\begin{aligned}\mathcal{L}_{occ} =& \lambda_{BCE}\mathcal{L}_{BCE} + \lambda_n\mathcal{L}_n \\ =& \lambda_{BCE} \sum_{\mathbf{x}_d \in \mathcal{P}_{\text{off}}} CE(o(\mathbf{x}_d, \boldsymbol{\theta}_b, \boldsymbol{\psi}), o^{GT}(\mathbf{x}_d)) \\ &+ \lambda_n \sum_{\mathbf{x}_d \in \mathcal{P}_{\text{on}}} \left\| \mathbf{n}_d - \mathbf{n}^{GT}(\mathbf{x}_d) \right\|_2 ,\end{aligned} \tag{4.10}$$

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 [Dong et al. 2022]. $\mathcal{L}_{RGB}$ supervises the point color:

$$\mathcal{L}_{RGB} = \lambda_{RGB} \sum_{\mathbf{x}_d \in \mathcal{P}_{\text{on}}} \left\| c(\mathbf{x}_c, \mathbf{n}_d, \mathcal{F}, \boldsymbol{\theta}_b, \boldsymbol{\psi}) - c^{GT}(\mathbf{x}_d) \right\|_1 . \tag{4.11}$$

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 $\mathcal{L}_{\text{surf}}$:

$$\begin{aligned}\mathcal{L}_{reg} =& \lambda_{\text{bone}}\mathcal{L}_{\text{bone}} + \lambda_{\text{joint}}\mathcal{L}_{\text{joint}} + \lambda_{\text{surf}}\mathcal{L}_{\text{surf}} \\ =& \lambda_{\text{bone}} \sum_{\mathbf{x}_c \in \mathcal{P}^c_{\text{bone}}} CE(f_{\text{occ}}(\mathbf{x}_c, \boldsymbol{\theta}_b, \boldsymbol{\psi}), 1) \\ &+ \lambda_{\text{joint}} \sum_{\mathbf{x}_c \in \mathcal{P}^c_{\text{joint}}} \sum_{i \in \mathcal{N}(i)} (w_i(\mathbf{x}_c) - 0.5)^2 \\ &+ \lambda_{\text{surf}} \sum_{\mathbf{x}_c \in \mathcal{P}^c_{\text{surf}}} \sum_{i \in \mathbf{G} \setminus \mathbf{G}_B} (w_i(\mathbf{x}_c) - w_i^{GT}(\mathbf{x}_c))^2,\end{aligned} \tag{4.12}$$

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 Chapter 3. $\mathcal{P}^c_{\text{bone}}$, $\mathcal{P}^c_{\text{joint}}$, $\mathcal{P}^c_{\text{surf}}$ 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 [Chen et al. 2021]. We add the last term to regularize the LBS weights for fingers and face which have low resolution and are more difficult to learn.

4.3.2. Model Initialization

To speed up the training process, we pre-train the geometry network f_{occ} and skinning network f_w with male and female SMPL-X meshes from AMASS [Mahmood et al. 2019].

4.4. Adaptation from 3D Scans to RGB-D Video

To enable learning X-Avatars from RGB-D videos, we make the following modifications to the scan-based version:

- We add a **data pre-processing** step, in which we generate colored point clouds from the RGB-D images with known camera parameters and estimate normals with points from the local neighborhood. See Supp. Mat for more details about the generation and usage of RGB-D data.
- In the **geometry** module, we replace the occupancy field f_{occ} with a signed distance field (SDF) f_{sdf} simply by removing the softmax activation function in the last layer. The reason is that without the surface from the scan, we cannot calculate the ground-truth occupancy, but we know all points lie on the surface so the ground-truth SDF naturally equals to zero.
- In the **deformation** module, we modify the pooling operation from maximum to minimum since the definition of inside and outside for occupancy and SDF are opposite.
- In the **objective function**, compared to Eq. (12) in the main paper, we replace the BCE loss $\mathcal{L}_{BCE}$ with an L1 loss $\mathcal{L}_1$, remove the bone occupancy loss $\mathcal{L}_{\text{bone}}$, and add an Eikonal loss $\mathcal{L}_{\text{eik}}$ following [Dong et al. 2022, Gropp et al. 2020]. The new objective function thus becomes:

$$\begin{aligned}
\mathcal{L} =& \lambda_1 \mathcal{L}_1 + \lambda_n \mathcal{L}_n + \lambda_{RGB} \mathcal{L}_{RGB} \\
&+ \lambda_{\text{joint}} \mathcal{L}_{\text{joint}} + \lambda_{\text{surf}} \mathcal{L}_{\text{surf}} + \lambda_{\text{eik}} \mathcal{L}_{\text{eik}} \\
\mathcal{L}_1 =& \sum_{\mathbf{x}_d \in \mathcal{P}_{\text{on}}} \| o(\mathbf{x}_d, \boldsymbol{\theta}_b, \boldsymbol{\psi}) \|_1 \\
\mathcal{L}_{\text{eik}} =& \sum_{\mathbf{x}_d \in \mathcal{P}_{\text{off}}} (\| f_{\text{sdf}}(\mathbf{x}_c, \boldsymbol{\theta}_b, \boldsymbol{\psi}) \| - 1)^2
\end{aligned} \tag{4.13}$$

Results

We first introduce the datasets and metrics that we use for our experiments in Sec. 5.1. Sec. 5.2 ablates all important design choices. In Sec. 5.3 we briefly describe the state-of-the-art methods to which we compare our method. Finally we show and discuss the results in Sec. 5.4-5.6. We focus on the challenging animation task, but for completeness, we also report the reconstruction results.

5.1. Datasets

5.1.1. GRAB

We use the GRAB [Taheri et al. 2020] subset of AMASS [Mahmood et al. 2019] 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 one for validation. This results in 9,756 frames for training and 235 test frames.

5.1.2. 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 [Collet et al. 2015]. We call the resulting dataset X-Humans. It consists of 21 subjects (12 males, 9 females) with various clothing types and hair styles. It is the first 3D

textured clothed human dataset that contains a large variation of body poses, hand gestures and facial expressions. As illustrated in Fig. 5.1, our participants not only do kicks, dancing, weight lifting and other kind of sports that involve large body movements, but also perform more fine-level finger movements, such as playing instruments, using tools, or counting. Along with different poses, people show corresponding emotions like laughing, frowning and screaming. A statistical analysis of the motion types can be found in Supp. Mat. For each subject, we split the motion sequences into a training and test set. In total, there are 30,278 poses for training and 6,727 test poses. The detailed statistics of our X-Humans dataset is provided in Supp. Mat. 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 can be found in Chapter 3. To provide SMPL registrations we convert our SMPL-X fits using the official transfer code [Pavlakos et al. 2019].

The collection and publication of X-Humans has been reviewed and approved by an internal ethics committee. All subjects have participated voluntarily, signed a consent form and have received monetary compensation for their time required to complete the capture.

Figure 5.1.: X-Humans Gallery. *With our high-quality, multi-view volumetric capture stage [Collet et al. 2015], we provide X-Humans, which consists of 21 subjects with various clothing types, colors, hair styles, genders and ages. It is the first dataset of 3D textured clothed human scans with a large variation of body pose, hand gestures and facial expressions. Ground truth SMPL[-X] registrations are also provided.*

5.1.3. 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. More details about the generation and usage of RGB-D version of X-Humans can be found in Supp. Mat.

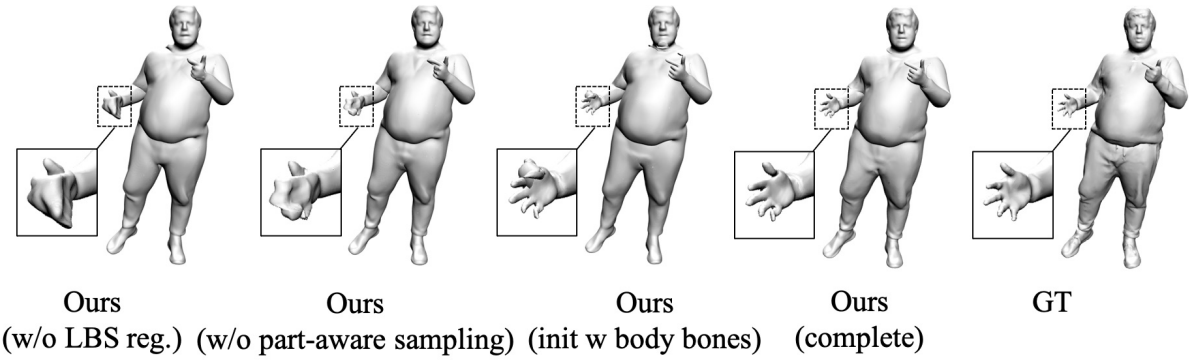

Figure 5.2.: Effect of our design decisions on the resulting geometry. *Notice how all baselines struggle to recover accurate hand geometry.*

5.1.4. Metrics

We evaluate the geometric accuracy via volumetric IoU, Chamfer distance (CD) (mm) and normal consistency (NC) metrics, following the practice in PINA [Dong et al. 2022]. Because these metrics are dominated by large surface areas, we always report the metrics for the entire body (*All*) and the hands separately (*Hands*).

5.2. Ablation Study

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 5.1.: 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. 5.2 and Fig. 5.4.*

5.2.1. 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 [Chen et al. 2021]. 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 (*c.f.* Fig. 5.2 and Tab. 5.1, A1). The final model is 3 times faster than A2 (0.7 iterations per second vs. 0.25), yet it still retains high fidelity and even outperforms A2 by a small margin (*c.f.* Tab. 5.1, A2 and Fig. 5.3 for qualitative results). Thus we conclude that part-based initialization of the deformer is an efficient way to find accurate correspondences.

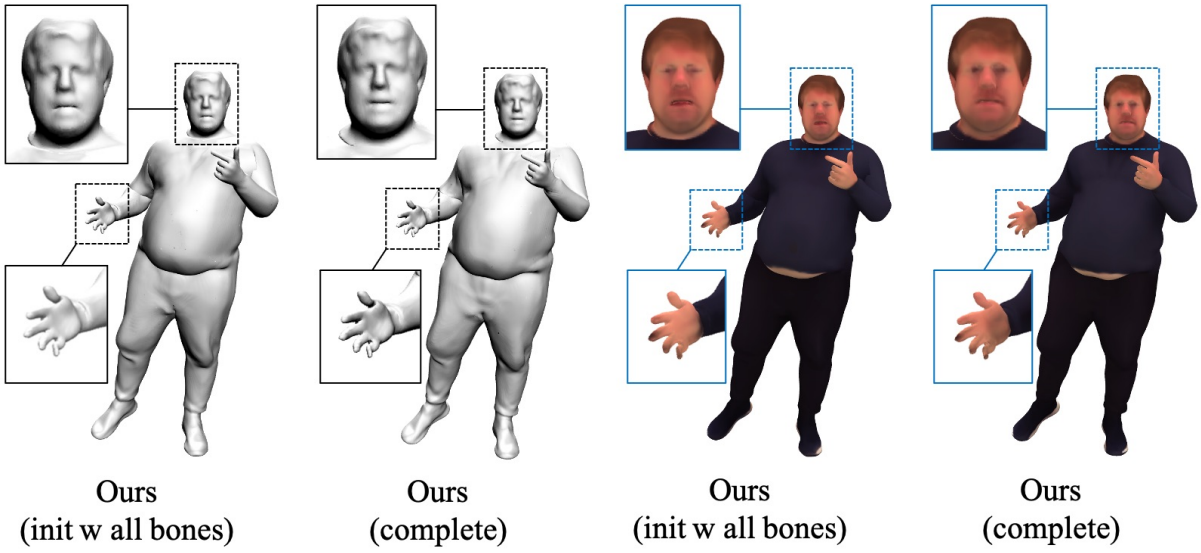

Figure 5.3.: Effect of our part-aware initialization strategy. *Our final model gives similar predictions on hands, face geometry and texture but with 3 times the speed of the baseline.*

5.2.2. 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 (*c.f.* second column of Fig. 5.2 and Tab. 5.1, A3) and b) it improves texture details in the eye and mouth region (*c.f.* Fig. 5.4).

5.2.3. LBS Weights Regularization for Hands and Face

The first column in Fig. 5.2 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 (*c.f.* Tab. 5.1, A4).

5.2.4. Texture-Conditioning

To increase the quality of the learned texture, we condition our texture field f_{RGB} on both high-level geometry features $\mathcal{F}$ and low-level normals $\mathbf{n}_d$ derived from the deformed space, following

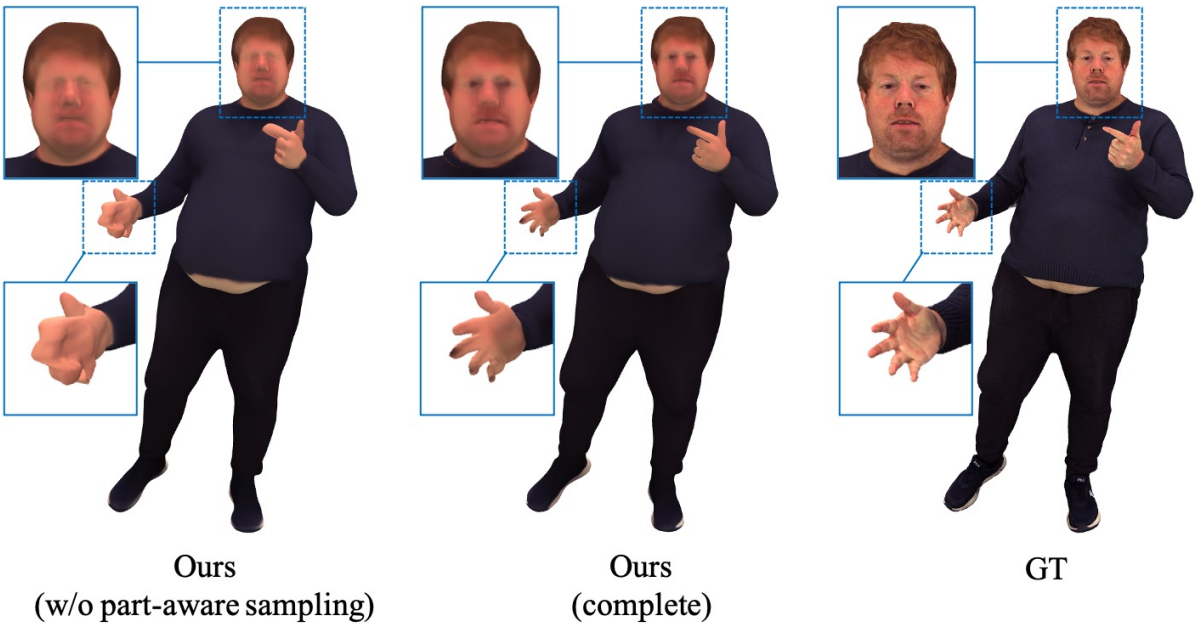

Figure 5.4.: Effect of our part-aware sampling strategy *on the hand geometry and texture prediction of the face.*

[Zheng et al. 2022, Chen et al. 2022]. Though we are not the first to do this, we still carry out the ablation study on texture field conditions for completeness.

To verify the importance of geometry features $\mathcal{F}$ and normals $\mathbf{n}_d$, we separately remove them from the condition, and compare the qualitative results with the full version as shown in Fig. 5.5. With both normals and features as conditions, our complete version produces sharper contours of the mouth, eyes, and more details like white teeth and shadows on pants than the other two baselines.

5.3. Baselines

5.3.1. Scan-based Methods

We compare our 3D scan-based method variant on both GRAB and X-Humans to SMPLX+D, SCANimate and SNARF baselines. We adapt SMPLX+D from SMPL+D introduced in [Bhatnagar et al. 2020a]. This baseline uses an explicit body model, SMPL-X, and models clothing with additive vertex offsets. For each subject, we optimize the offset between the scans and SMPL-X meshes, average over the training set to get the template offset. We then add the offset on all testing SMPL-X meshes to model the clothed human. To compare with SCANimate and SNARF, we use publicly available code. Since these two methods require SMPL registration, we use the official model parameter transfer code provided in [Pavlakos et al. 2019] to convert SMPL-X parameters to SMPL parameters.

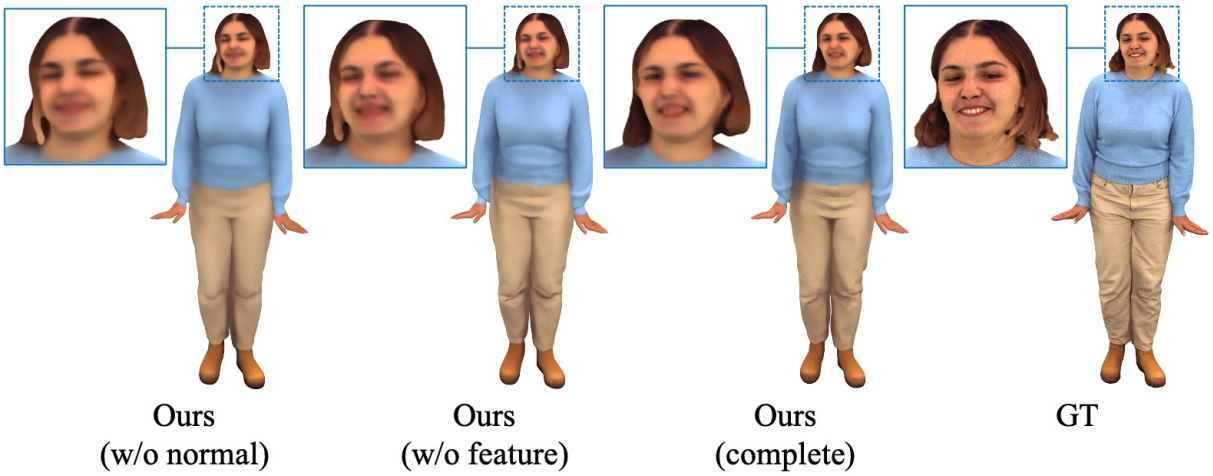

Figure 5.5.: Effect of our design decisions on the resulting texture. *The one with both normals and geometry features as conditons produces the sharpest details around the mouth, eyes, and clothes.*

5.3.2. RGB-D Video-based Methods

We compare our RGB-D method variant on the X-Humans (RGB-D) dataset to PINA [Dong et al. 2022], 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.

5.4. Results on GRAB Dataset

Tab. 5.2 summarizes results on the GRAB dataset. Overall, our method beats all baselines, especially for the hands, where the margin is large. Fig. 5.6 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 [Saito et al. 2021]	2.32	6.25	54.86	55.18	0.970	0.804	0.938	0.656
SNARF [Chen et al. 2021]	1.11	3.75	29.98	29.72	0.980	0.851	0.975	0.724
Ours	**0.88**	**0.75**	**16.76**	**4.87**	**0.984**	**0.962**	**0.994**	**0.901**

Table 5.2.: Quantitative results on GRAB dataset. *Our method outperforms all baselines, especially for the hand part (*c.f. *Fig. 5.6).*

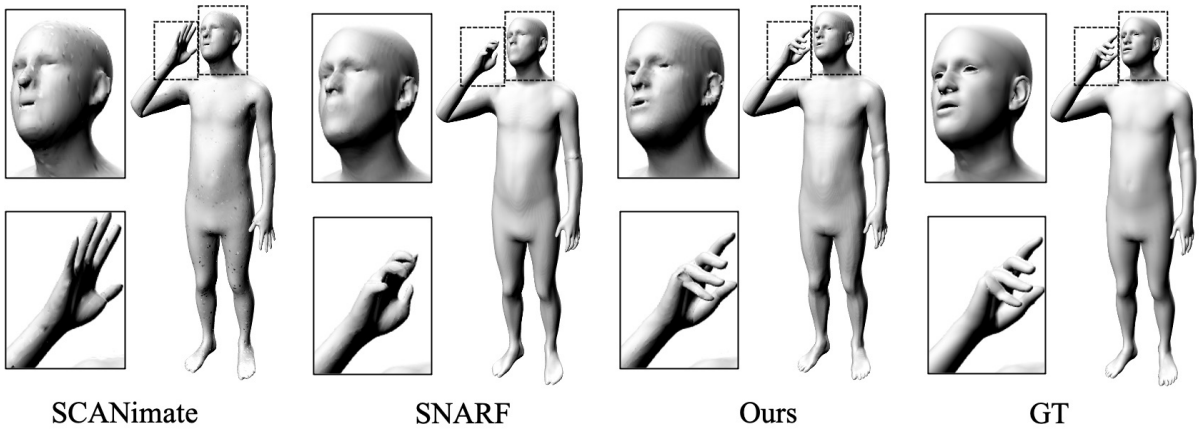

***Figure 5.6.: Qualitative results on GRAB dataset**. Our method recovers hand articulation and facial expression most accurately.*

5.5. Results on X-Humans (Scans)

5.5.1. Animation

Tab. 5.3 shows that our method also outperforms the baselines on X-Humans. Fig. 5.7 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. 5.11, Fig. 5.9, Fig. 5.12, Fig. 5.13, Fig. 5.14 show more animation results, with learned avatars driven by motions from either the monocular video or our X-Humans dataset.

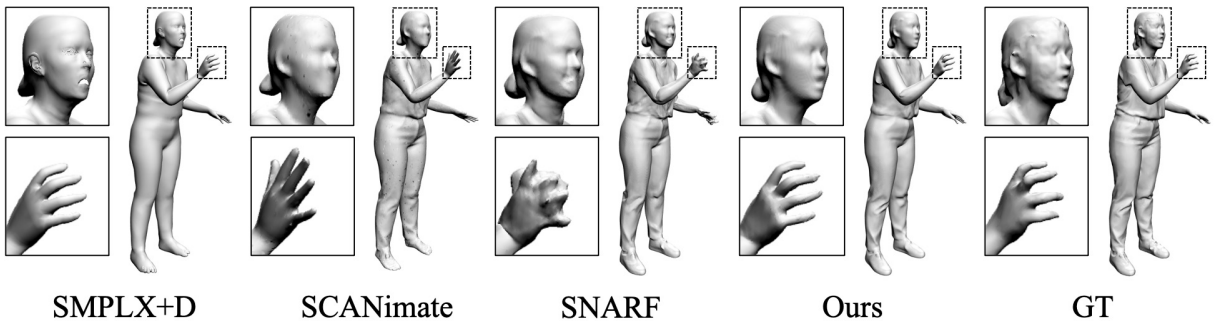

***Figure 5.7.: Qualitative animation comparison on X-Humans (Scans).** SMPLX+D fails to model face and garment details. SCANimate and SNARF generate poor hands (static or incomplete). Our method produces the most plausible face and hands, and keeps the clothing details comparable to strong baselines.*

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 [Saito et al. 2021]	6.54	9.78	59.71	48.32	0.925	0.726	0.919	0.557
SNARF [Chen et al. 2021]	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 5.3.: Quantitative animation results on X-Humans (Scans). *We beat all baselines both for the entire body (All) and hands only (Hands).*

5.5.2. Reconstruction

Tab. 5.4 summarizes reconstruction results on X-Humans with all scan-based methods. SNARF has the best score among all methods. However, notice that all numbers are reported on the training set, which means they only reflect the over-fitting capabilities. Combining the findings in the animation task, the hand learned by SNARF seems to overfit drastically on the training set. When given an unseen pose, it tends produce a shape that barely looks like a human hand. Though our method is not the best on the reconstruction task, on one hand, its performance does not differ too much from its best competitor, and on the other hand, it demonstrates stronger generalization ability to unseen hand poses as demonstrated in the animation task.

Method	CD↓		CD-MAX ↓		NC ↑		IoU ↑	
	All	Hands	All	Hands	All	Hands	All	Hands
SMPLX+D	6.45	5.18	49.71	20.72	0.918	0.792	0.953	0.754
SCANimate [Saito et al. 2021]	6.38	10.42	61.8	50.85	0.928	0.729	0.904	0.540
SNARF [Chen et al. 2021]	**2.55**	**2.29**	**43.03**	**15.4**	0.955	**0.925**	0.974	**0.792**
Ours	2.66	4.78	43.53	22.18	**0.957**	0.810	**0.980**	0.790

Table 5.4.: Quantitative reconstruction results on X-Humans (Scans). *Metrics show that SNARF can fit well on the training set but mostly due to the over-fitting (especially for the face and hands), which is verified in the animation task.*

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

5.6.1. Animation

Tab. 5.5 shows our model's performance compared to PINA [Dong et al. 2022]. Our method outperforms PINA on all metrics. Fig. 5.8 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

Method	CD↓		CD-MAX ↓		NC ↑		IoU ↑	
	All	Hands	All	Hands	All	Hands	All	Hands
PINA [Dong et al. 2022]	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 5.5.: Quantitative animation results on X-Humans (RGB-D). *Our method outperforms PINA in all metrics. Improvements are more pronounced for hands (*c.f. *Fig. 5.8 for visual comparison).*

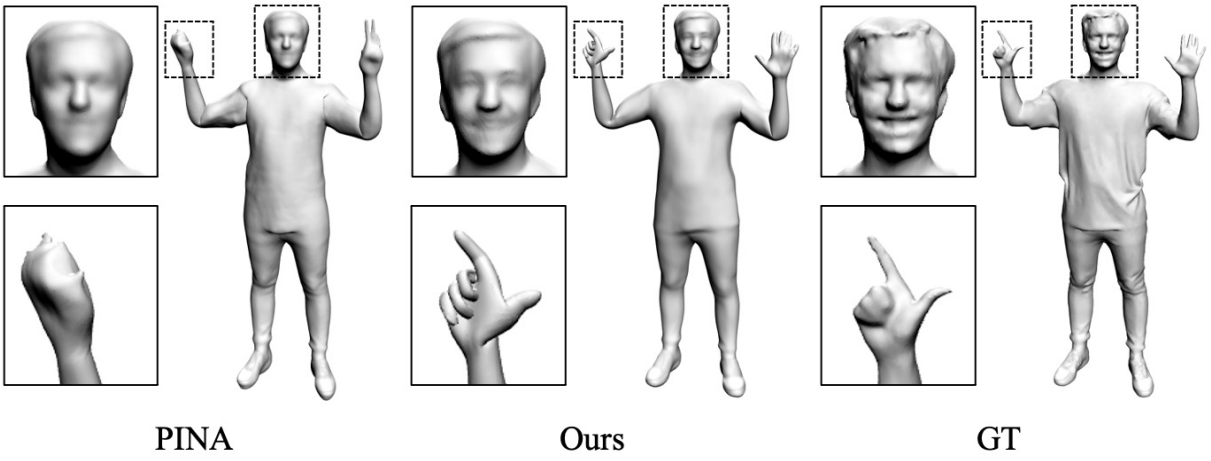

Figure 5.8.: X-Avatars created from RGB-D input compared to PINA. *Notice how we obtain better hand and face geometry.*

network is conditioned on facial expression and b) better hand poses because we initialize the root finding with hand bone transformations.

5.6.2. Reconstruction

Method	CD↓		CD-MAX ↓		NC ↑		IoU ↑	
	All	Hands	All	Hands	All	Hands	All	Hands
PINA [Dong et al. 2022]	4.78	7.37	64.01	42.18	0.935	0.816	0.926	0.614
Ours	**4.48**	**4.85**	**49.75**	**21.62**	**0.943**	**0.819**	**0.952**	**0.785**

Table 5.6.: Quantitative reconstruction results on X-Humans (RGB-D). *Our method beats PINA, showing our new expressive human representation can better fuse partial observations and learn body, hands and face entirely.*

In Tab. 5.6, we lists comparisons on RGB-D data with PINA [Dong et al. 2022]. Different from training with scans where we have the complete mesh information, when we learn from RGB-D images, for each frame, we only have partial observations from certain view points. Therefore, the reconstruction results measures the capability of fusing partial observations from different

view points into an implicit surface representation. Our method outperforms PINA especially for the hand region, which means our new human representation can better model body, hands and face as an entirety.

Figure 5.9.: More animation results on X-Humans (Scans).

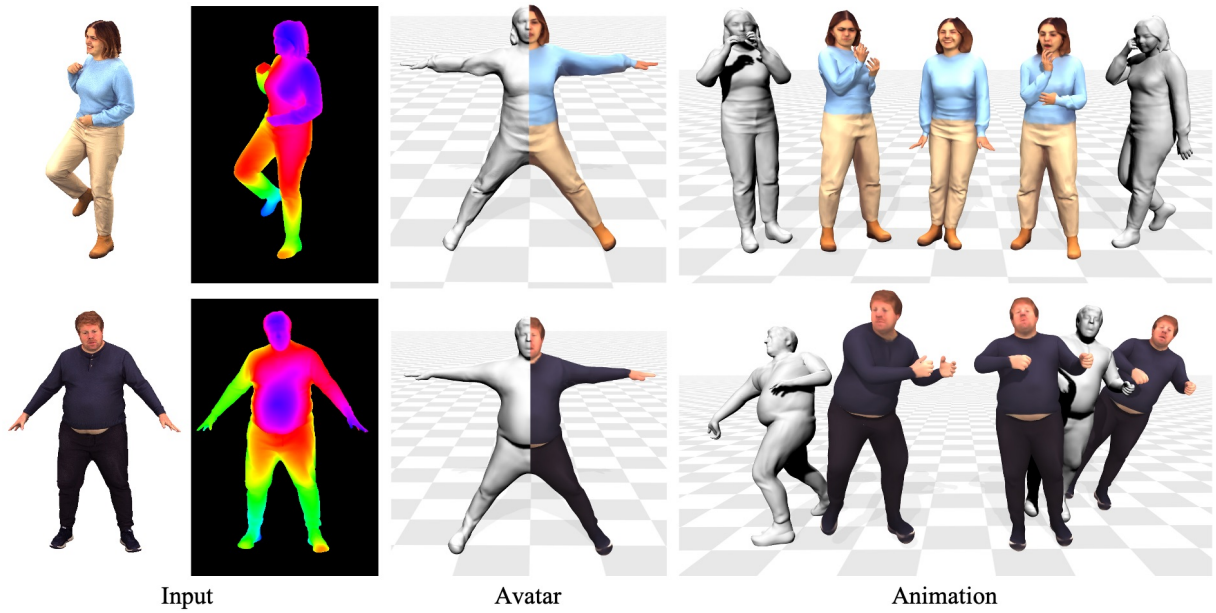

Figure 5.10.: More animation results on X-Humans (RGB-D).

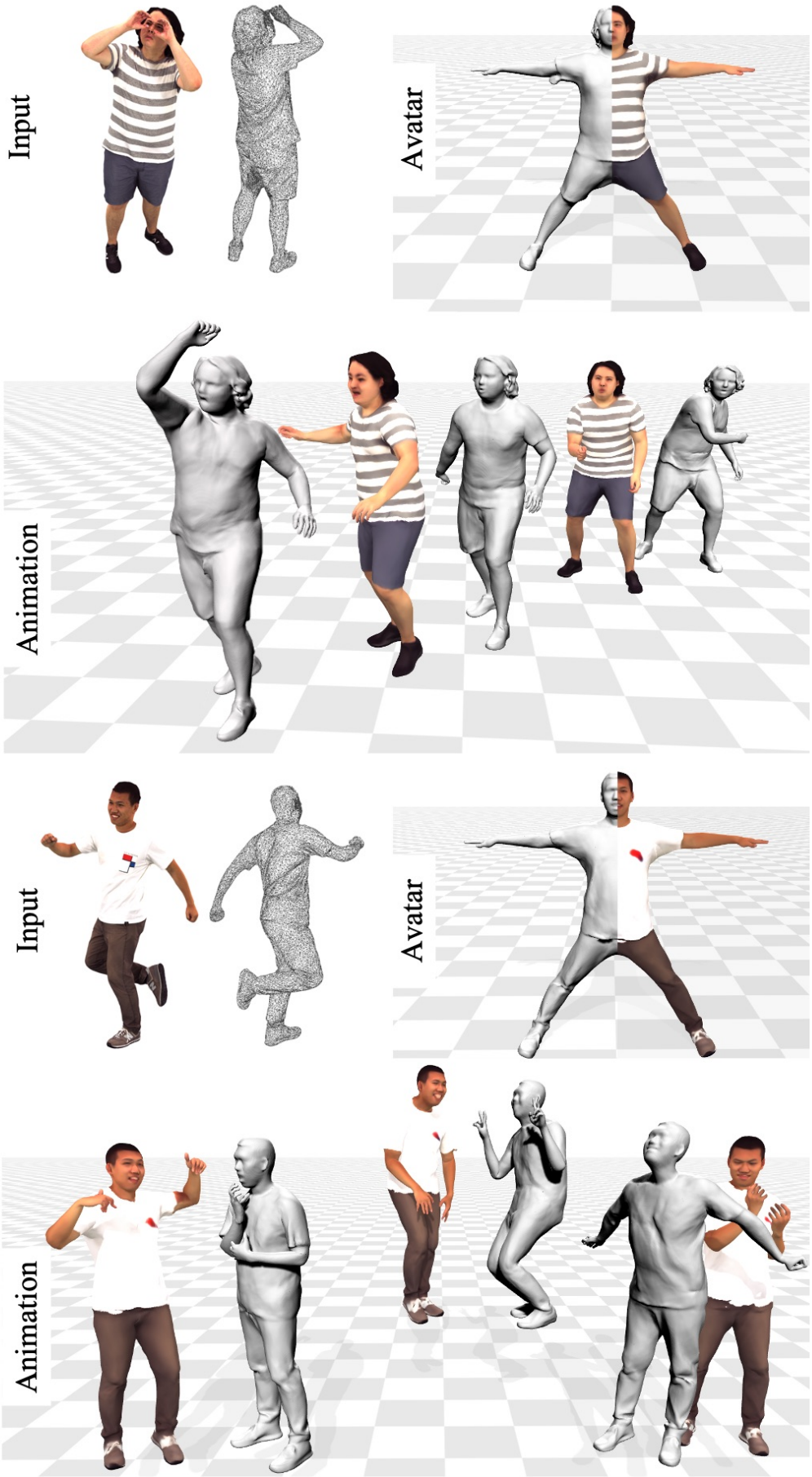

Figure 5.11.: Animation demonstration on X-Humans (Scans). *Our method can handle relatively complex clothing patterns, hairstyles, and varied facial expressions, hand, and body poses.*

Figure 5.12.: Demonstration of multiple X-Avatars driven by motions from YouTube video (Tennis).

Figure 5.13.: Demonstration of multiple X-Avatars driven by motions from YouTube video (Ballet Dance).

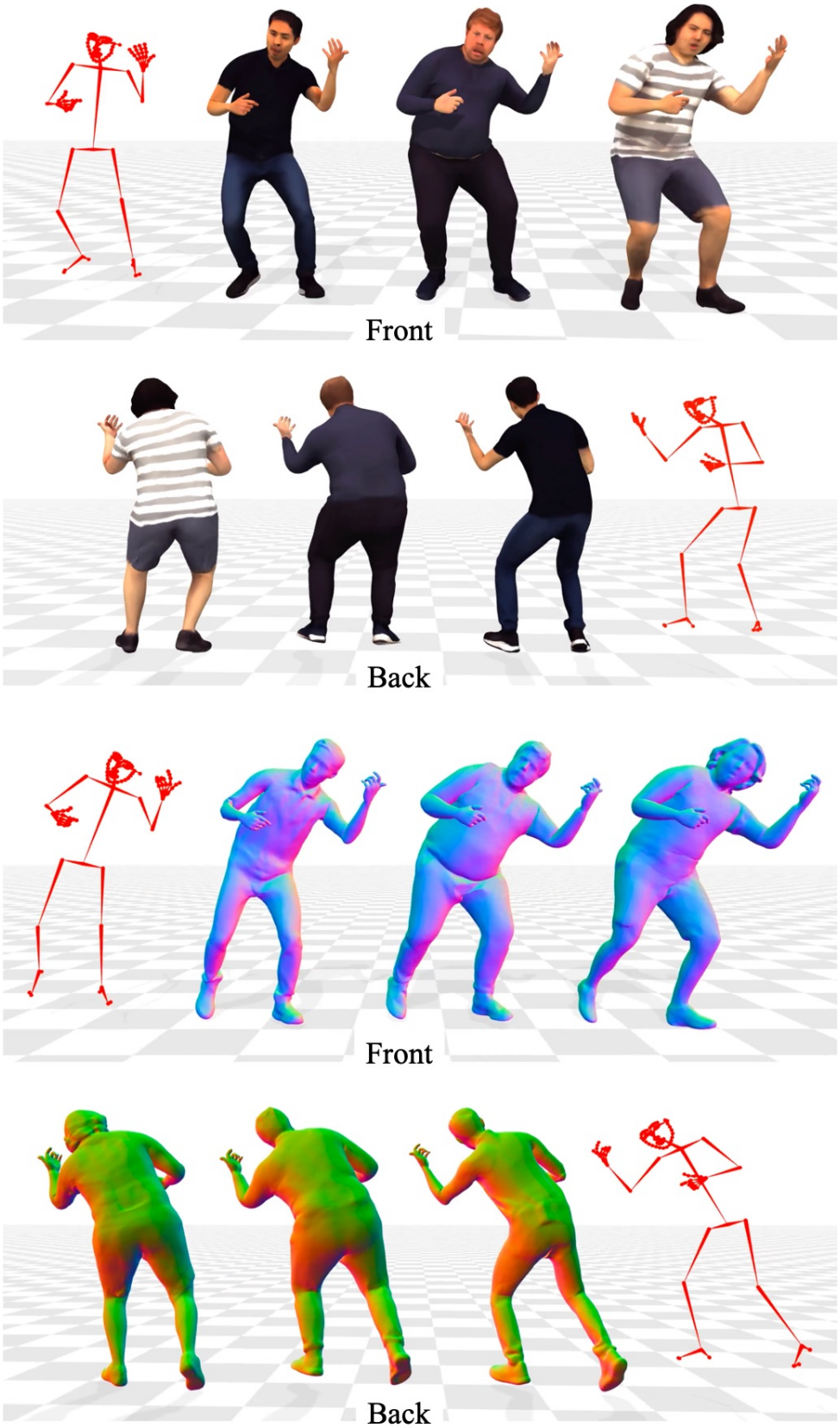

Figure 5.14.: Demonstration of multiple X-Avatars driven by motions from X-Humans.

Figure 5.15.: Demonstration of an X-Avatar driven by motions from other subjects in X-Humans.

Conclusion and Outlook

6.1. 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.

6.2. Conclusion

We propose X-Avatar, the first expressive implicit human avatar model that captures body pose, hand pose, facial expressions and appearance in a holistic fashion. We have demonstrated our method's expressive power, the benefit of our proposed part-aware initialization and sampling strategy, 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.

Appendix

A.1. SMPL-X Registration Details

A.1.1. Range of Joint Rotation

In the third step of SMPL-X registration pipeline, i.e., multi-stage fitting (*c.f*. Sec. 3.3), we introduce some priors $\mathcal{L}_{\theta_b}$, $\mathcal{L}_{\theta_h}$, $\mathcal{L}_{\theta_f}$ to penalize unrealistic bending of the torso, hands, and face joints. To be more specific, we manually set the limitation of rotation along each axis for every joint, which are detailed in Tab. A.1 (*torso*), Tab. A.2 (*hand*) and Tab. A.3 (*face*). All range of movement are set based on human limits.

A.1.2. Comparison with Other Registration Pipelines

We visually compare the fitting results of our SMPL(-X) registration pipeline with that of other registration pipelines on our captured 3D scans. The SMPL fitting baseline is EasyMocap [eas 2021], which only uses multi-view images as the input. Our SMPL fitting pipeline, besides the multi-view images, uses extra 3D scan information. The SMPL-X fitting baseline is MPI_MeshRegistration [Bhatnagar et al. 2020a, Bhatnagar et al. 2020b], which uses both multi-view images and 3D scans as inputs, the same setting as ours. Fig. A.1 shows that both our SMPL and SMPL-X registration pipelines achieve better alignment than their strong competitors. At the same time, our fitted SMPL-X model are more well-aligned with the ground truth than the SMPL model in terms of the hands and the face.

Description	Joint ID in *body pose*	Axis	Range
left, right knees	3, 4	x	[0°, 180°]
left, right collars	12, 13	x	[-60°, 60°]
left, right shoulders	15, 16	x	[-60°, 60°]
left, right elbows	17, 18	x	[-60°, 60°]
left elbow	17	y	[-180°, 0°]
right elbow	18	y	[0°, 180°]
left, right wrists	19, 20	x	[-90°, 90°]
left, right ankles	6, 7	z	0°
left, right feet	9, 10	x, y, z	0°
other torso joints	/	/	[-180°, 180°]

Table A.1.: Detailed range of movement of SMPL-X torso joints. *We manually set the limitation of rotation for every torso joint based on our observation of the human body. In this way, we can avoid weird body poses that violate human limits.*

Description	Joint ID in *hand pose*	Axis	Range
index, middle, ring 1, 2, 3	0, 1, 2, 3, 4, 5, 9, 10, 11	x	[-15°, 15°]
pinky 1, 2, 3	6, 7, 8	x	[-60°, 20°]
index, middle, pinky, ring 2, 3	1, 2, 4, 5, 7, 8, 10, 11	y	[-10°, 10°]
index, middle, pinky, ring 1	0, 3, 6, 9	y	[-20°, 20°]
index, middle, pinky, ring 1, 2, 3	[0, 11]	z	[-90°, 20°]
thumb 1	12	x	[-60°, 180°]
thumb 1	12	y	[-60°, 120°]
thumb 2, 3	13, 14	x	[-60°, 120°]
thumb 2	13	y	[-60°, 180°]
thumb 3	14	y	[-20°, 90°]
thumb 1, 2, 3	12, 13, 14	z	[-90°, 60°]

Table A.2.: Detailed range of movement of SMPL-X (Left/Right) joints. *Similar to Tab. A.1, we manually set the limitation of rotation for every hand joint based on our observation of the human hand. Since the left hand and right hand are symmetric, we only list out the range for the left hand for simplicity. The only difference is that for the right hand, the range of rotation along the y axis and z axis is the opposite of its left counterpart.*

Description	Axis	Range
left, right eye	x, y, z	[-180°, 180°]
jaw	x	[-180°, 180°]
jaw	y, z	0°

Table A.3.: Detailed range of movement of SMPL-X face joints. *Similar to Tab. A.1, we manually set the limitation of rotation for every face joint based on our observation of the human face.*

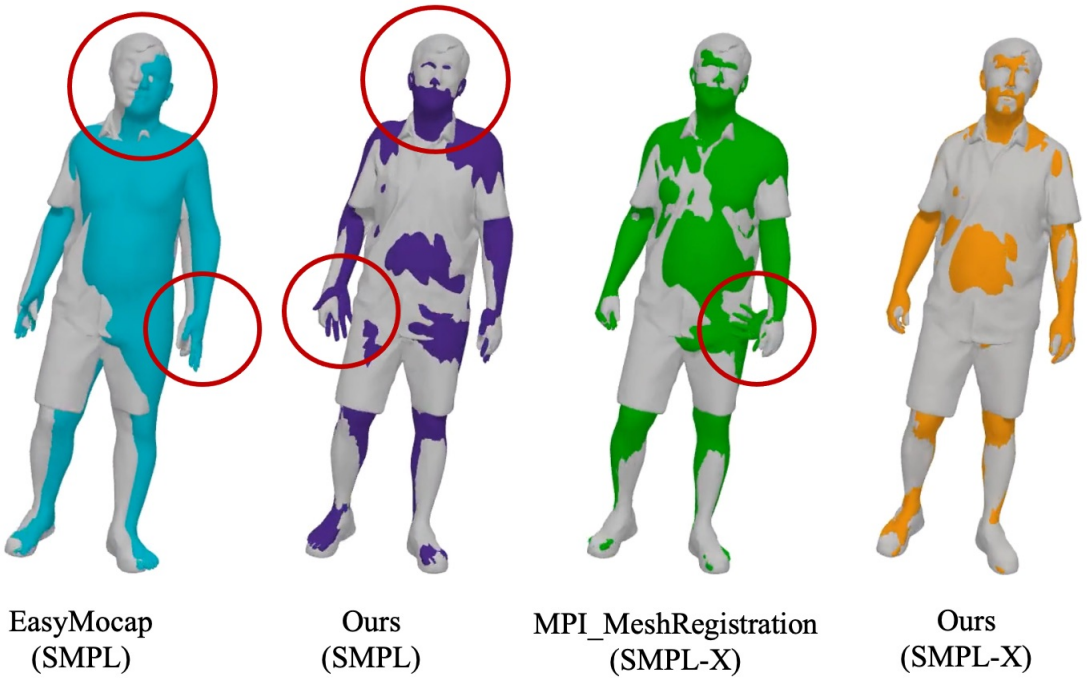

Figure A.1.: Qualitative comparison with other SMPL(-X) registration pipelines. *The grey meshes refer to the ground truth scans and the colored meshes refer to the SMPL(-X) model fitted with each method. Our SMPL-X fitting achieves the best alignment while our SMPL fitting and other methods all have badly aligned hands or faces as shown in the red circles.*

A.2. X-Humans: Dataset Details

A.2.1. Statistics

We list out all detailed information of X-Humans dataset in Tab. A.4 for reference, including the unique ID, the gender, the number of training and testing motion sequences, and the number of training and testing motion frames for each subject. As we can see, the genders are quite balanced, and each subject has quite enough frames for training an avatar.

We also compute the statistics in terms of the motion variety and visualize the result in Fig. A.2. The pie chart depicts the proportion of each motion type measured in the number of sequences. The chart is quite balanced and has lots of slices, showing the diversity of our X-Humans dataset.

A.2.2. Generation and Usage of RGB-D Version

The complete pipeline of the generation and usage of RGB-D data is shown in Fig. A.4.

ID	Gender	# Seq		# Frames	
		train	test	train	test
00001	female	8	2	1595	330
00002	female	10	2	1738	360
00003	male	9	2	1780	360
00004	male	10	2	1396	300
00016	male	9	2	1330	296
00017	male	9	2	1339	299
00018	male	9	2	1340	300
00019	female	7	2	1015	286
00020	male	10	3	1721	376
00021	female	6	1	1000	150
00022	male	9	2	1242	288
00024	male	12	3	1799	398
00025	female	9	2	1335	300
00027	female	10	2	1439	300
00028	male	10	3	1444	448
00034	male	10	2	1415	300
00035	male	10	2	1332	300
00036	female	10	3	1464	449
00039	female	10	2	1396	292
00041	male	12	2	1727	300
Total:	21 subjects	199	45	30278	6727

Table A.4.: Detailed statistics on X-Humans dataset. *Every row lists out the information per subject: unique ID, gender, number of training/testing sequences/frames.*

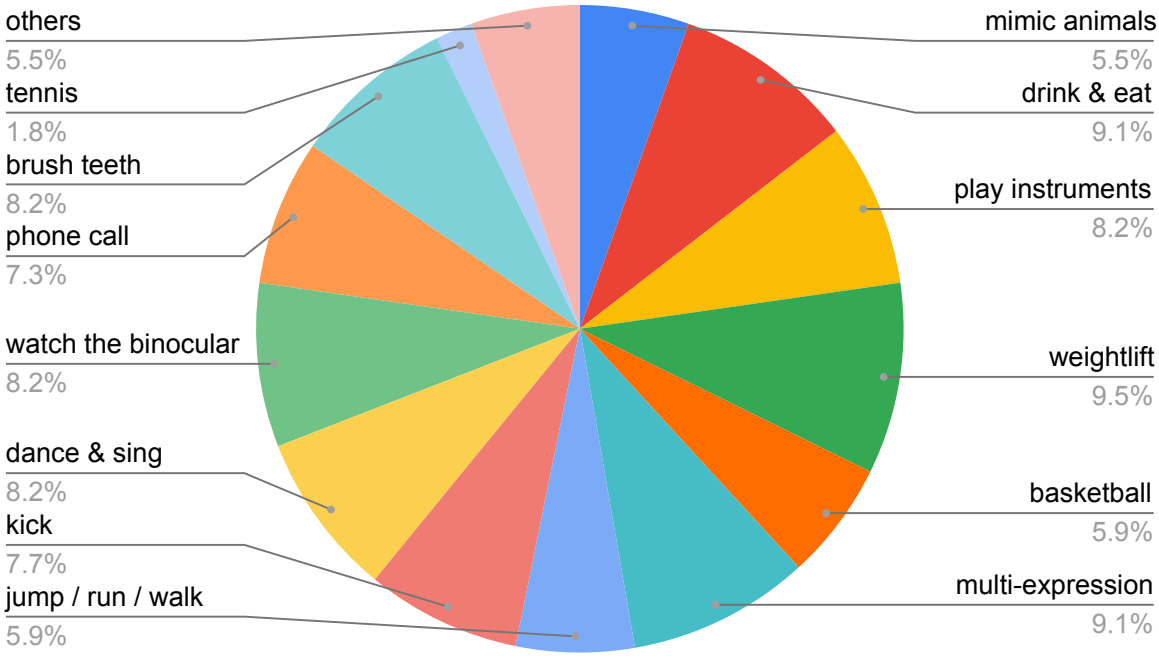

Figure A.2.: Statistics of Types of Motions in X-Humans Dataset. *The motion types have a great variety involving all scales of movement: body poses (e.g., dance, kick, jump), hand gestures (e.g., brush teeth, play instruments, watch the binocular) and facial expressions (e.g., multi-expression, drink, eat).*

Generation We use AITViewer [Kaufmann et al. 2022] to get the rendered RGB-D images from our captured textured and posed scans. Basically, we have an virtual camera moving around the target and render one RGB-D image at each frame. For the camera, we set the field of view (FoV) to be 30°, inverses of the width and height of a pixel to be 600, and the width and height of the output image to be 800 and 1200 pixels. We visualize the trajectory of the moving camera in Fig. A.3, which is a circle with a radius of 6 meters, centered on the scan center of the first frame and located in the horizontal plane. The virtual camera moves in a constant speed: one step per frame and the whole circle consists of 100 segments. Though the circle is predefined and fixed, it is guaranteed to surround our target during the whole sequence since our volumetric capture stage always requires our participant to stand in a roughly 3-meter circle. During the rendering, we save both RGB-D images and the corresponding camera intrinsic and extrinsic.

Usage During the rendering, we assume that the virtual RGB camera and depth camera lie in the same coordinate system. And we already know the intrinsic and extrinsic of the cameras. So we can easily use the basic knowledge in computer vision to compute the colored point cloud from RGB-D images. With the corresponding SMPL-X registration, we can classify the point cloud into different body parts and apply our part-aware sampling and initialization strategies. The remaining steps are all same as the scan-based version.

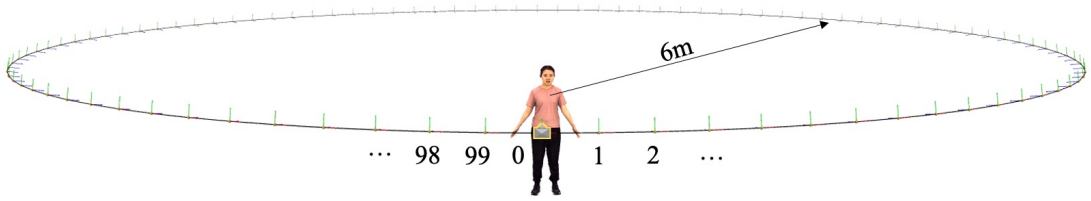

Figure A.3.: Trajectory of virtual camera during rendering. *The trajectory is a circle with a radius of 6 meters, centered on the scan center of the first frame and located in the horizontal plane. The circle is uniformly divided into 100 segments and the virtual camera moves one step per frame. So for each frame, we only have partial observation of the participant from a certain camera view.*

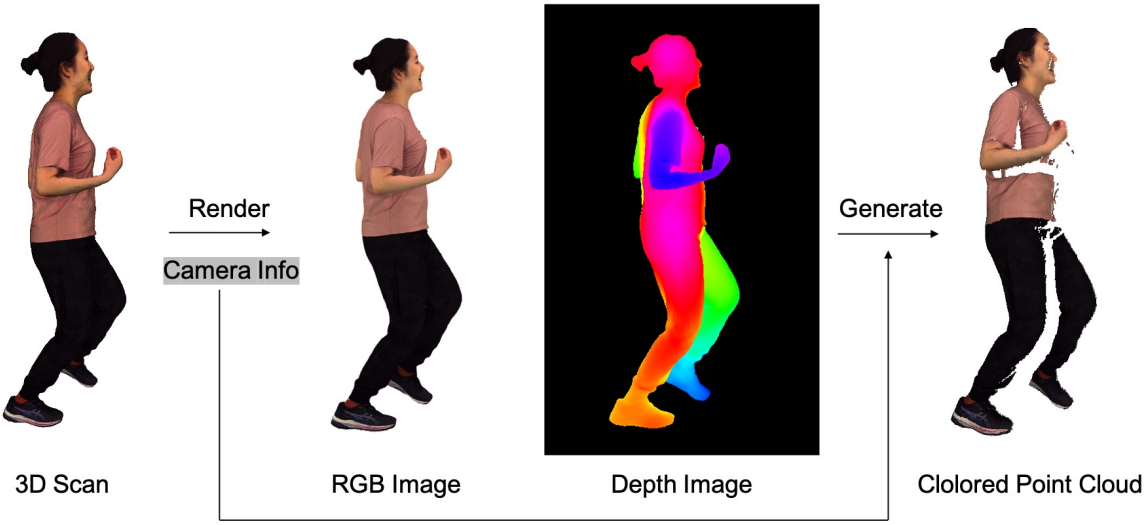

Figure A.4.: Colored Point Cloud Generation Pipeline. *The 3D scan is first rendered with a virtual camera to get the RGB and depth image. With the known camera information including the extrinsic and intrinsic matrix, we then generate the colored point cloud from the RGB and depth image.*

A.3. X-Avatar: Implementation Details

A.3.1. Network Architecture

We implement our models in PyTorch [Paszke et al. 2019]. Fig. A.5 illustrates the network architectures for the geometry-, texture-, and deformation-networks. We use geometric initialization [Atzmon and Lipman 2020] for the geometry network's weights and PyTorch's default initialization for the weights of the skinning network and texture network. For both the geometry network and texture network, we apply positional encoding [Mildenhall et al. 2021] with 4 frequency components on the input points to model high-frequency details, and condition the networks on pose $\boldsymbol{\theta}_b$ and facial expressions $\boldsymbol{\psi}$ to handle pose-dependent deformations. We additionally condition the texture network on the last layer feature $\mathcal{F}$ of the geometry network and the normal $\mathbf{n}_d$ in deformed space so that the texture network is aware of the underlying geometry.

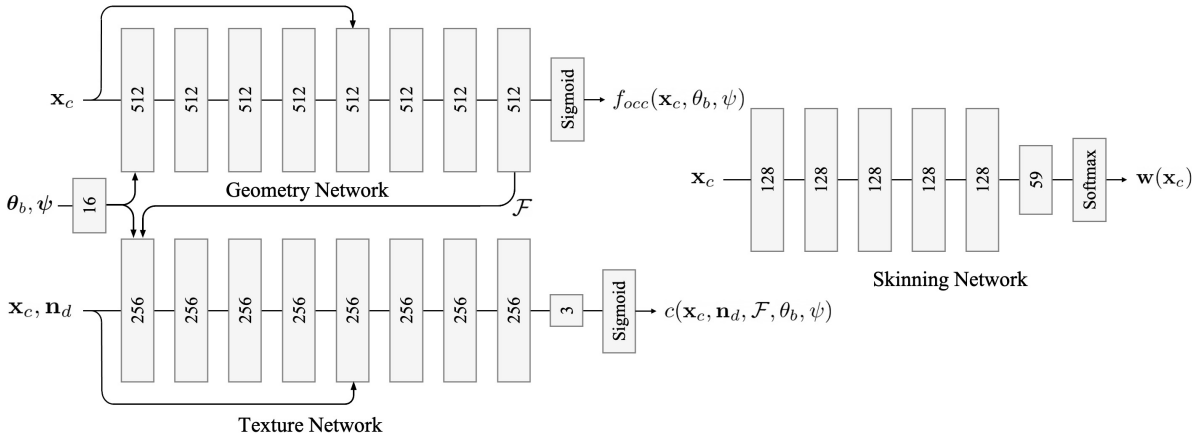

Figure A.5.: Network Architecture. *Each block represents a linear layer with its output dimension specified in the inset, followed by a weight normalization layer [Salimans et al. 2016] and a Softplus [Dugas et al. 2000] activation layer.*

A.3.2. Correspondence Search

Following SNARF [Chen et al. 2021], we use Broyden's method [Broyden 1965] for our correspondence search. We apply our part-aware strategy to the initialization stage. For each deformed point $\mathbf{x}_d$ with part label ℓ, we initialize the states by inversely transforming $\mathbf{x}_d$ with bone transformations of the corresponding bone group $\mathbf{G}_\ell$. Here, $\ell \in \{F, LH, RH, B\}$, $|\mathbf{G}_F| = 3$, $|\mathbf{G}_{LH}| = |\mathbf{G}_{RH}| = 16$, $|\mathbf{G}_B| = 9$. In the experiments, we set the maximum number of update steps to 50 and the convergence threshold to 10^{-5}.

A.3.3. Canonical Pose

Following [Chen et al. 2021, Zheng et al. 2022], we set the roll value of left hip and right hip to $\pi/6$ and $-\pi/6$, and the pitch value of the jaw to 0.2. With this definition, the canonical shape is in a star-like pose with little self-contact and smooth boundaries, which makes the learning easier as MLPs tends to produce smooth outputs.

A.3.4. Loss Details

We set the weights of the losses to $\lambda_{BCE} = \lambda_1 = 1, \lambda_n = 1$ ($\lambda_n = 0.1$ for RGB-D), $\lambda_{RGB} = 1, \lambda_{\text{bone}} = 1, \lambda_{\text{joint}} = \lambda_{\text{surf}} = 10, \lambda_{\text{eik}} = 0.5$.

A.3.5. Training Details

We train our networks using the the Adam optimizer [Kingma and Ba 2014] with a learning rate $\eta = 10^{-3}$ ($\eta = 10^{-4}$ for RGB-D), and $\beta = (0.9, 0.999)$, without weight decay or learning rate decay. Training a model takes around 24h on a single Nvidia RTX 6000 GPU.

A.4. Supplementary Results

A.4.1. More Qualitative Comparison

We illustrate more examples in Fig. A.6, Fig. A.7 and Fig. A.8 to make the qualitative comparison on the animation task on GRAB dataset and our X-Humans dataset more convincing. On each dataset, for each subject, our method beats other baselines well in terms of the body details, hand and face fidelity, which is consistent with the statement and the quantitative numbers in the main paper.

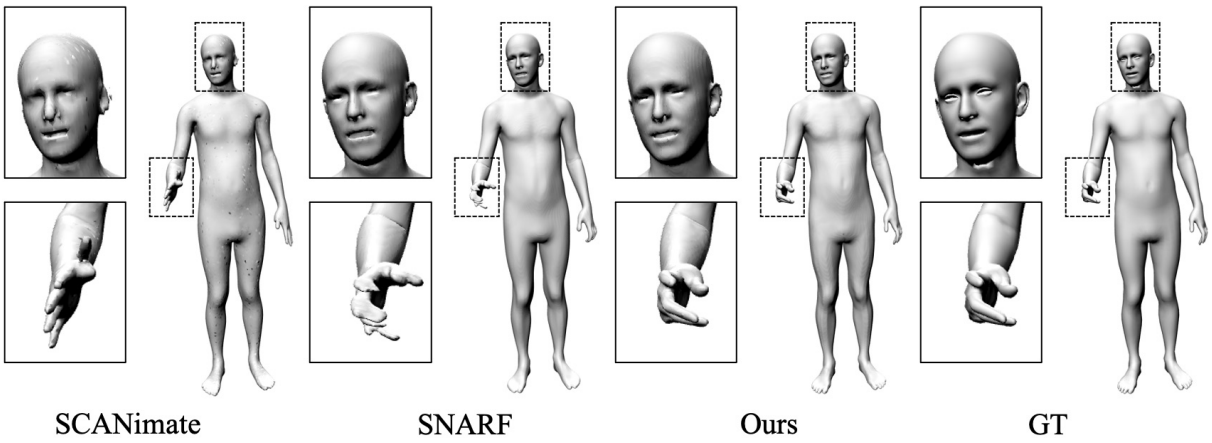

Figure A.6.: More qualitative comparison on GRAB dataset*. From left to right are results of SCANimate, SNARF, our method and ground truth. Our method recovers hand articulation and facial expression most accurately.*

A.5. Societal Impact Discussion

X-Avatar enables building fully animatable human avatars from either 3D scans or RGB-D video, which has great potential in immersive, life-like remote telepresence and other experiences in AR/VR. The method presented here is intended for uses that are beneficial to society, *e.g*. by bringing people closer together in mixed reality who are otherwise large distances apart in the real world. However, we unfortunately cannot rule out that the technology might be abused for nefarious purposes. Because our method can animate personalized avatars with poses and facial expressions that are completely unseen, the biggest concern is that it might be misused to generate deep fakes. Although there is still a way to go to achieve a level of quality that is indistinguishable from real footage, the rapid progress of recent years in related fields, such as image generation, may have fore-shadowed a similar trend in the modelling of 3D human avatars. We believe that open-sourcing such research is vital to build a general knowledge about how such models can be created - this understanding will in turn help to build countermeasures and detect malicious uses.

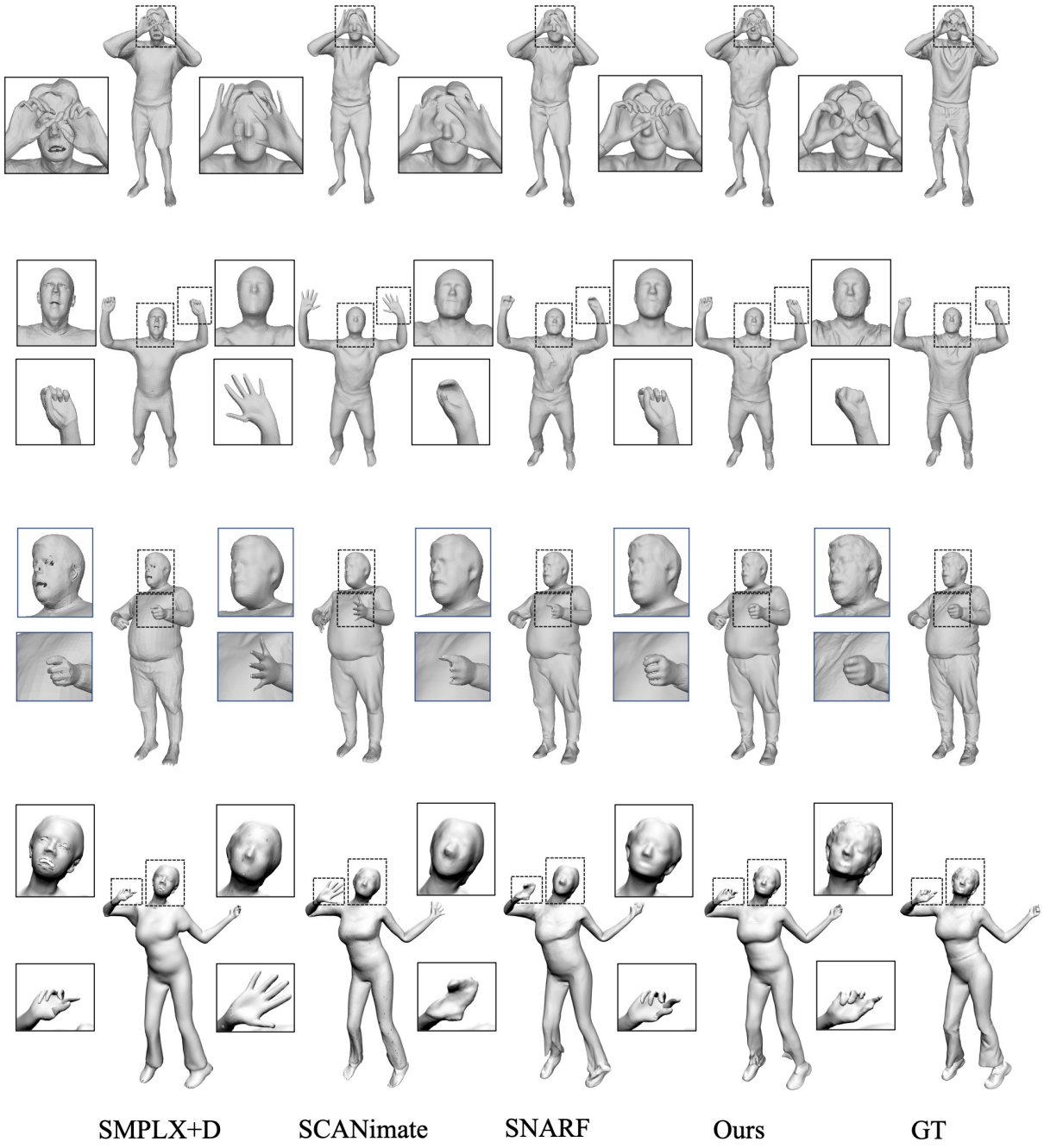

Figure A.7.: More qualitative comparison on X-Humans (Scans). *From left to right are results of SMPLX+D, SCANimate, SNARF, our method and ground truth. SMPLX+D fails to model face and garment details. SCANimate and SNARF generate poor hands (static or incomplete). Our method produces the most plausible face and hands, and keeps the clothing details comparable to strong baselines.*

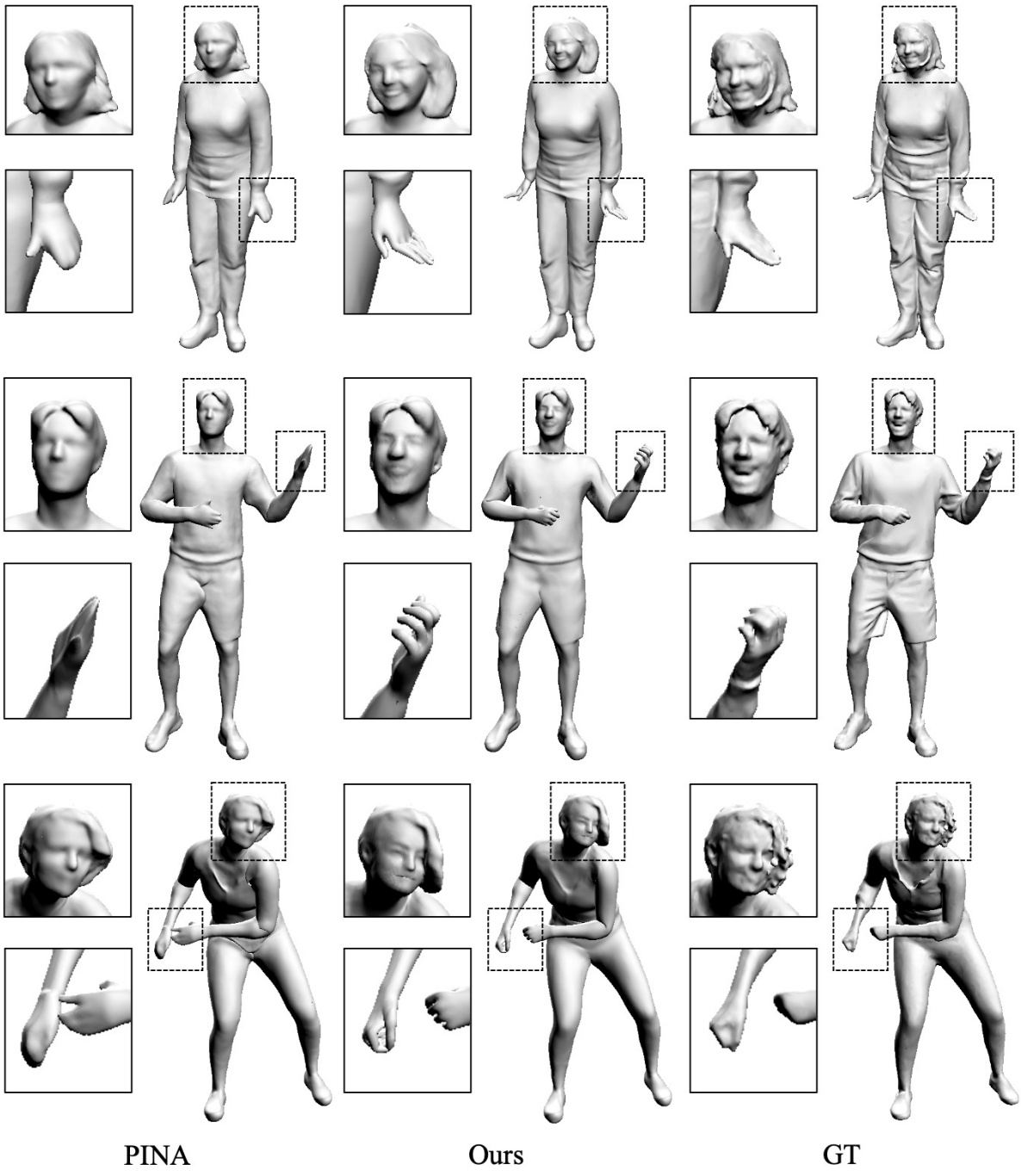

Figure A.8.: More qualitative comparison on X-Humans (RGB-D). *From left to right are results of PINA, our method and ground truth. Notice how we obtain better hand and face geometry compared to PINA.*

Bibliography

ALLDIECK, T., MAGNOR, M., XU, W., THEOBALT, C., AND PONS-MOLL, G. 2018. Video based reconstruction of 3d people models. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8387–8397.

ALLDIECK, T., MAGNOR, M., BHATNAGAR, B. L., THEOBALT, C., AND PONS-MOLL, G. 2019. Learning to reconstruct people in clothing from a single rgb camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1175–1186.

ALLDIECK, T., PONS-MOLL, G., THEOBALT, C., AND MAGNOR, M. 2019. Tex2shape: Detailed full human body geometry from a single image. In *IEEE International Conference on Computer Vision (ICCV)*, IEEE.

ANGUELOV, D., SRINIVASAN, P., KOLLER, D., THRUN, S., RODGERS, J., AND DAVIS, J. 2005. Scape: Shape completion and animation of people. *ACM Trans. Graph. 24*, 3 (jul), 408–416.

ATZMON, M., AND LIPMAN, Y. 2020. Sal: Sign agnostic learning of shapes from raw data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2565–2574.

BHATNAGAR, B. L., TIWARI, G., THEOBALT, C., AND PONS-MOLL, G. 2019. Multi-garment net: Learning to dress 3d people from images. In *IEEE International Conference on Computer Vision (ICCV)*, IEEE.

BHATNAGAR, B. L., SMINCHISESCU, C., THEOBALT, C., AND PONS-MOLL, G. 2020. Combining implicit function learning and parametric models for 3d human reconstruction. In *European Conference on Computer Vision (ECCV)*, Springer.

BHATNAGAR, B. L., SMINCHISESCU, C., THEOBALT, C., AND PONS-MOLL, G. 2020. Loopreg: Self-supervised learning of implicit surface correspondences, pose and shape for 3d human mesh registration. In *Advances in Neural Information Processing Systems (NeurIPS)*.

BLANZ, V., AND VETTER, T. 1999. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, ACM Press/Addison-Wesley Publishing Co., USA, SIGGRAPH '99, 187–194.

BOGO, F., BLACK, M. J., LOPER, M., AND ROMERO, J. 2015. Detailed full-body reconstructions of moving people from monocular rgb-d sequences. *2015 IEEE International Conference on Computer Vision (ICCV)*, 2300–2308.

BOUKHAYMA, A., BEM, R. D., AND TORR, P. H. 2019. 3d hand shape and pose from images in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 10843–10852.

BRAHMBHATT, S., HAM, C., KEMP, C. C., AND HAYS, J. 2019. ContactDB: Analyzing and predicting grasp contact via thermal imaging. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

BROYDEN, C. G. 1965. A class of methods for solving nonlinear simultaneous equations. *Mathematics of computation 19*, 92, 577–593.

CAO, Z., HIDALGO MARTINEZ, G., SIMON, T., WEI, S., AND SHEIKH, Y. A. 2019. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

CHEN, Y., CHENG, Z., LAI, C., MARTIN, R. R., AND DANG, G. 2016. Realtime reconstruction of an animating human body from a single depth camera. *IEEE Transactions on Visualization and Computer Graphics 22*, 2000–2011.

CHEN, X., ZHENG, Y., BLACK, M. J., HILLIGES, O., AND GEIGER, A. 2021. Snarf: Differentiable forward skinning for animating non-rigid neural implicit shapes. In *International Conference on Computer Vision (ICCV)*.

CHEN, X., JIANG, T., SONG, J., YANG, J., BLACK, M. J., GEIGER, A., AND HILLIGES, O. 2022. gdna: Towards generative detailed neural avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 20427–20437.

CHOUTAS, V., PAVLAKOS, G., BOLKART, T., TZIONAS, D., AND BLACK, M. J. 2020. Monocular expressive body regression through body-driven attention. In *European Conference on Computer Vision (ECCV)*.

COLLET, A., CHUANG, M., SWEENEY, P., GILLETT, D., EVSEEV, D., CALABRESE, D., HOPPE, H., KIRK, A., AND SULLIVAN, S. 2015. High-quality streamable free-viewpoint video. *ACM Trans. Graph. 34*, 4 (jul).

CORONA, E., PUMAROLA, A., ALENYÀ, G., PONS-MOLL, G., AND MORENO-NOGUER, F. 2021. Smplicit: Topology-aware generative model for clothed people. In *CVPR*.

CORONA, E., HODAN, T., VO, M., MORENO-NOGUER, F., SWEENEY, C., NEWCOMBE, R., AND MA, L. 2022. Lisa: Learning implicit shape and appearance of hands. In *CVPR*.

DENG, B., LEWIS, J., JERUZALSKI, T., PONS-MOLL, G., HINTON, G., NOROUZI, M., AND TAGLIASACCHI, A. 2020. Neural articulated shape approximation. In *The European Conference on Computer Vision (ECCV)*, Springer.

DONG, Z., GUO, C., SONG, J., CHEN, X., GEIGER, A., AND HILLIGES, O. 2022. 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)*, 20470–20480.

DUGAS, C., BENGIO, Y., BÉLISLE, F., NADEAU, C., AND GARCIA, R. 2000. Incorporating second-order functional knowledge for better option pricing. *Advances in neural information processing systems 13*.

2021. Easymocap - make human motion capture easier. Github.

FENG, Y., CHOUTAS, V., BOLKART, T., TZIONAS, D., AND BLACK, M. J. 2021. Collaborative regression of expressive bodies using moderation. In *International Conference on 3D Vision (3DV)*.

GAFNI, G., THIES, J., ZOLLHÖFER, M., AND NIESSNER, M. 2021. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8649–8658.

GROPP, A., YARIV, L., HAIM, N., ATZMON, M., AND LIPMAN, Y. 2020. Implicit geometric regularization for learning shapes. In *Proceedings of Machine Learning and Systems 2020*. 3569–3579.

HASSON, Y., VAROL, G., TZIONAS, D., KALEVATYKH, I., BLACK, M. J., LAPTEV, I., AND SCHMID, C. 2019. Learning joint reconstruction of hands and manipulated objects. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11799–11808.

HUANG, Y., KAUFMANN, M., AKSAN, E., BLACK, M. J., HILLIGES, O., AND PONS-MOLL, G. 2018. Deep inertial poser: Learning to reconstruct human pose from sparse inertial measurements in real time. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia) 37* (Nov.), 185:1–185:15.

JOO, H., LIU, H., TAN, L., GUI, L., NABBE, B., MATTHEWS, I., KANADE, T., NOBUHARA, S., AND SHEIKH, Y. 2015. Panoptic studio: A massively multiview system for social motion capture. In *2015 IEEE International Conference on Computer Vision (ICCV)*, 3334–3342.

JOO, H., SIMON, T., AND SHEIKH, Y. 2018. 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)*.

KANAZAWA, A., BLACK, M. J., JACOBS, D. W., AND MALIK, J. 2018. End-to-end recovery of human shape and pose. In *Computer Vision and Pattern Recognition (CVPR)*.

KAUFMANN, M., VECHEV, V., AND MYLONOPOULOS, D., 2022. AITViewer, 7.

KINGMA, D. P., AND BA, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

KOCABAS, M., HUANG, C.-H. P., HILLIGES, O., AND BLACK, M. J. 2021. PARE: Part attention regressor for 3D human body estimation. In *Proc. International Conference on Computer Vision (ICCV)*, 11127–11137.

KOLOTOUROS, N., PAVLAKOS, G., BLACK, M. J., AND DANIILIDIS, K. 2019. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *ICCV*.

LI, T., BOLKART, T., BLACK, M. J., LI, H., AND ROMERO, J. 2017. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia) 36*, 6.

LI, J., XU, C., CHEN, Z., BIAN, S., YANG, L., AND LU, C. 2021. 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*, 3383–3393.

LI, R., TANKE, J., VO, M., ZOLLHOFER, M., GALL, J., KANAZAWA, A., AND LASSNER, C. 2022. Tava: Template-free animatable volumetric actors.

LOPER, M., MAHMOOD, N., ROMERO, J., PONS-MOLL, G., AND BLACK, M. J. 2015. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia) 34*, 6 (Oct.), 248:1–248:16.

MA, Q., YANG, J., RANJAN, A., PUJADES, S., PONS-MOLL, G., TANG, S., AND BLACK,

M. J. 2020. Learning to dress 3d people in generative clothing. In *Computer Vision and Pattern Recognition (CVPR)*.

MA, Q., SAITO, S., YANG, J., TANG, S., AND BLACK, M. J. 2021. SCALE: Modeling clothed humans with a surface codec of articulated local elements. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 16082–16093.

MAHMOOD, N., GHORBANI, N., TROJE, N. F., PONS-MOLL, G., AND BLACK, M. J. 2019. AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision*, 5442–5451.

MESCHEDER, L., OECHSLE, M., NIEMEYER, M., NOWOZIN, S., AND GEIGER, A. 2019. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4460–4470.

MIHAJLOVIC, M., ZHANG, Y., BLACK, M. J., AND TANG, S. 2021. LEAP: Learning articulated occupancy of people. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.

MIHAJLOVIC, M., SAITO, S., BANSAL, A., ZOLLHOEFER, M., AND TANG, S. 2022. COAP: Compositional articulated occupancy of people. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.

MILDENHALL, B., SRINIVASAN, P. P., TANCIK, M., BARRON, J. T., RAMAMOORTHI, R., AND NG, R. 2020. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*.

MILDENHALL, B., SRINIVASAN, P. P., TANCIK, M., BARRON, J. T., RAMAMOORTHI, R., AND NG, R. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM 65*, 1, 99–106.

NIEMEYER, M., MESCHEDER, L., OECHSLE, M., AND GEIGER, A. 2019. Occupancy flow: 4d reconstruction by learning particle dynamics. In *Proceedings of the IEEE/CVF international conference on computer vision*, 5379–5389.

OSMAN, A. A. A., BOLKART, T., AND BLACK, M. J. 2020. STAR: A sparse trained articulated human body regressor. In *European Conference on Computer Vision (ECCV)*, 598–613.

OSMAN, A. A. A., BOLKART, T., TZIONAS, D., AND BLACK, M. J. 2022. SUPR: A sparse unified part-based human body model. In *European Conference on Computer Vision (ECCV)*.

PARK, J. J., FLORENCE, P., STRAUB, J., NEWCOMBE, R., AND LOVEGROVE, S. 2019. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 165–174.

PASZKE, A., GROSS, S., MASSA, F., LERER, A., BRADBURY, J., CHANAN, G., KILLEEN, T., LIN, Z., GIMELSHEIN, N., ANTIGA, L., ET AL. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems 32*.

PATEL, P., HUANG, C.-H. P., TESCH, J., HOFFMANN, D. T., TRIPATHI, S., AND BLACK, M. J. 2021. AGORA: Avatars in geography optimized for regression analysis. In *Proceed-*

ings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR).

PAVLAKOS, G., CHOUTAS, V., GHORBANI, N., BOLKART, T., OSMAN, A. A. A., TZIONAS, D., AND BLACK, M. J. 2019. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.

RAMON, E., TRIGINER, G., ESCUR, J., PUMAROLA, A., GARCIA, J., GIRO-I NIETO, X., AND MORENO-NOGUER, F. 2021. H3d-net: Few-shot high-fidelity 3d head reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5620–5629.

ROMERO, J., TZIONAS, D., AND BLACK, M. J. 2017. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia) 36*, 6 (Nov.), 245:1–245:17.

SAITO, S., YANG, J., MA, Q., AND BLACK, M. J. 2021. SCANimate: Weakly supervised learning of skinned clothed avatar networks. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*.

SALIMANS, T., GOODFELLOW, I., ZAREMBA, W., CHEUNG, V., RADFORD, A., AND CHEN, X. 2016. Improved techniques for training gans. *Advances in neural information processing systems 29*.

SIMON, T., JOO, H., MATTHEWS, I., AND SHEIKH, Y. 2017. Hand keypoint detection in single images using multiview bootstrapping. In *CVPR*.

SONG, J., CHEN, X., AND HILLIGES, O. 2020. Human body model fitting by learned gradient descent.

SUN, Y., BAO, Q., LIU, W., FU, Y., MICHAEL J., B., AND MEI, T. 2021. Monocular, one-stage, regression of multiple 3d people. In *ICCV*.

SUN, Y., LIU, W., BAO, Q., FU, Y., MEI, T., AND BLACK, M. J. 2022. Putting people in their place: Monocular regression of 3d people in depth. In *CVPR*.

TAHERI, O., GHORBANI, N., BLACK, M. J., AND TZIONAS, D. 2020. GRAB: A dataset of whole-body human grasping of objects. In *European Conference on Computer Vision (ECCV)*.

TIWARI, G., SARAFIANOS, N., TUNG, T., AND PONS-MOLL, G. 2021. Neural-gif: Neural generalized implicit functions for animating people in clothing. In *International Conference on Computer Vision (ICCV)*.

VON MARCARD, T., HENSCHEL, R., BLACK, M., ROSENHAHN, B., AND PONS-MOLL, G. 2018. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *European Conference on Computer Vision (ECCV)*.

XIU, Y., YANG, J., TZIONAS, D., AND BLACK, M. J. 2022. ICON: Implicit Clothed humans Obtained from Normals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13296–13306.

XU, H., BAZAVAN, E. G., ZANFIR, A., FREEMAN, W. T., SUKTHANKAR, R., AND SMINCHISESCU, C. 2020. Ghum & ghuml: Generative 3d human shape and articulated pose

models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6184–6193.

YENAMANDRA, T., TEWARI, A., BERNARD, F., SEIDEL, H., ELGHARIB, M., CREMERS, D., AND THEOBALT, C. 2021. i3dmm: Deep implicit 3d morphable model of human heads. In *Proceedings of the IEEE / CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

YI, X., ZHOU, Y., HABERMANN, M., SHIMADA, S., GOLYANIK, V., THEOBALT, C., AND XU, F. 2022. 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)*.

YU, T., ZHENG, Z., GUO, K., ZHAO, J., DAI, Q., LI, H., PONS-MOLL, G., AND LIU, Y. 2018. 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.

YUAN, Y., IQBAL, U., MOLCHANOV, P., KITANI, K., AND KAUTZ, J. 2022. Glamr: Global occlusion-aware human mesh recovery with dynamic cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

ZANFIR, A., BAZAVAN, E. G., XU, H., FREEMAN, W. T., SUKTHANKAR, R., AND SMINCHISESCU, C. 2020. Weakly supervised 3d human pose and shape reconstruction with normalizing flows. In *Computer Vision – ECCV 2020*, 465–481.

ZHANG, C., PUJADES, S., BLACK, M. J., AND PONS-MOLL, G. 2017. Detailed, accurate, human shape estimation from clothed 3d scan sequences. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

ZHENG, Y., ABREVAYA, V. F., BÜHLER, M. C., CHEN, X., BLACK, M. J., AND HILLIGES, O. 2022. I M Avatar: Implicit morphable head avatars from videos. In *Computer Vision and Pattern Recognition (CVPR)*.