Supplementary Material
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

In this supplementary document, we provide additional materials to supplement our main paper. Please also refer to the supplementary video.

Contents

1. SMPL-X Registration Pipeline

From our Volumetric Capture Stage we get high-quality 3D scans and RGB images of 53 camera views. We fit an SMPL-X model [18] 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 model for the fitting.

Our SMPL-X registration pipeline has three steps: 1. 2D landmarks detection (Sec. 1.1); 2. 3D landmarks generation (Sec. 1.2); 3. multi-stage fitting (Sec. 1.3).

1.1. 2D Landmarks Detection

As shown in the left part of Fig. 10, 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 [5, 21]. 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.

1.2. 3D Landmarks Generation

As shown in the right part of Fig. 10, 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

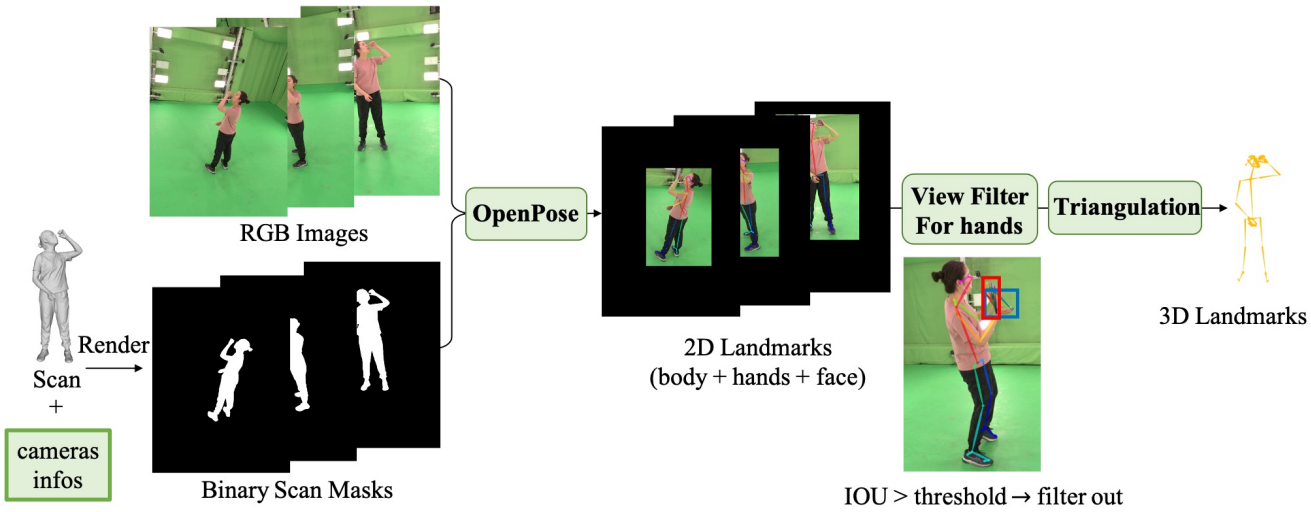

Figure 10. **Full-body landmarks generation overview.** It consists of two steps: (a) 2D landmarks detection using OpenPose on cropped images; (b) 3D landmarks generation via triangulation. We filter out views which show a large overlap of 2D hand landmarks.

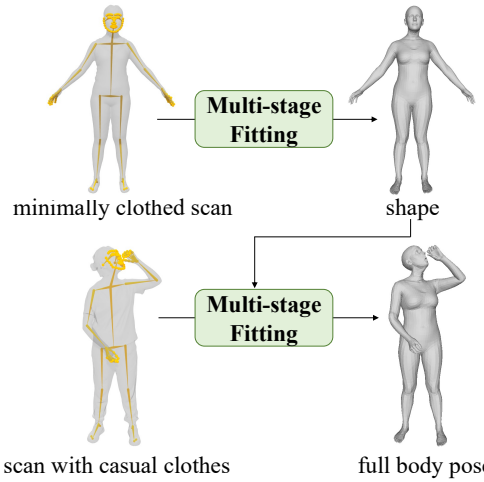

Figure 11. **Registration overview.** For each subject, we first obtain the shape by running the multi-stage fitting pipeline (*cf*. Tab. 5) once on minimally clothed scans (5 frames). Then we fix the shape and run the pipeline once more on scans where subjects wear casual clothing to get the full body 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.

detection may further result in poor 3D landmarks, which is why we simply filter them out.

1.3. Multi-Stage Fitting

For fitting, similar to [1, 17], we adopt a multi-stage pipeline, shown in more detail in Tab. 5. First we initiate the SMPL-X parameters with the sparse 3D landmarks obtained from multi-view RGB images as described in Sec. 1.2. 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 [17], we disentangle the optimization of ground-truth shape and ground-truth poses as shown in Fig. 11. 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.

2. X-Avatar: Implementation Details

2.1. Network Architecture

We implement our models in PyTorch [16]. Fig. 12 illustrates the network architectures for the geometry-, texture-, and deformation-networks. We use geometric initialization [2] 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 [15] 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

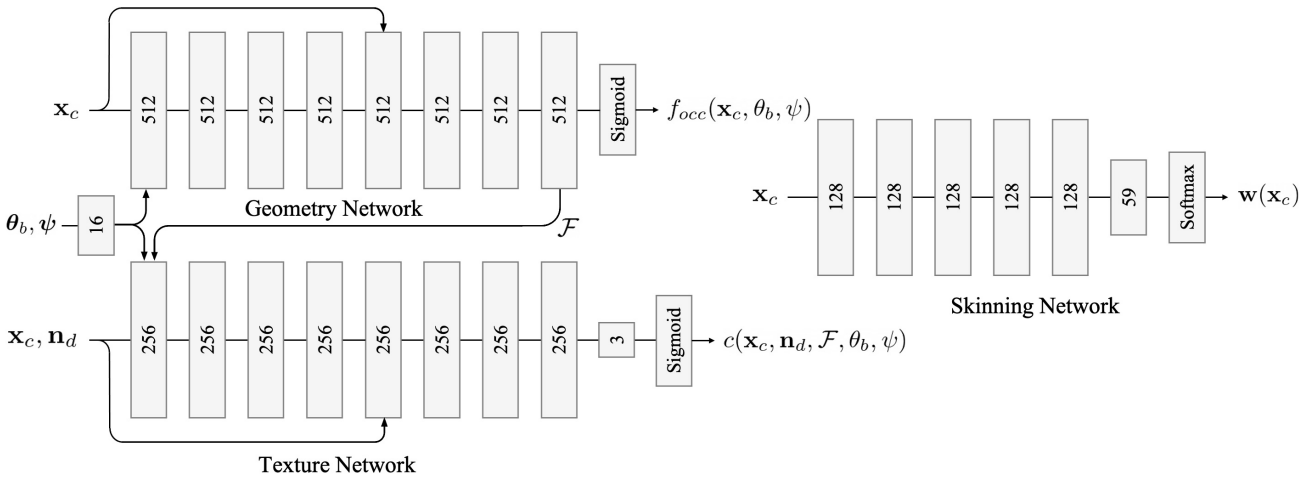

Figure 12. **Network Architecture.** Each block represents a linear layer with its output dimension specified in the inset, followed by a weight normalization layer [20] and a Softplus [11] activation layer.

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 5. **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 [1]. $\mathcal{L}_{\boldsymbol{\theta}_b}, \mathcal{L}_{\boldsymbol{\theta}_h}, \mathcal{L}_{\boldsymbol{\theta}_f}$ penalize unrealistic bending of the torso, hands, and face joints. $\mathcal{L}_{\boldsymbol{\beta}}, \mathcal{L}_{\boldsymbol{\psi}}$ are L2 regularizers on the body shape and facial expressions. The shape $\boldsymbol{\beta}$ is only optimized for the minimally-clothed sequence (*cf*. Fig. 11). Registration is more robust in this coarse-to-fine manner.

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. We ablate on the design choices for the texture conditions in Sec. 5.2 of this Supp. Mat.

2.2. 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 [14].

2.3. Part Label Assignment

We use a hard assignment and every point is assigned to only one part. The assignment works as follows: We first compute the part label of each SMPL-X vertex. Because SMPL-X has a fixed topology and it is a priory known which vertex IDs originally belonged to the face or hands, this is a simple look-up that only needs to be done once. Then, for each vertex on the 3D scan, its part label is determined by the pre-computed label of its closest SMPL-X vertex.

2.4. Correspondence Search

Following SNARF [8], we use Broyden's method [4] 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}.

2.5. Canonical Pose

Following [8, 22], 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.

2.6. 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.
- 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 [10, 12]. The new objective function thus becomes:

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

2.7. Losses

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

2.8. Training Details

We train our networks using the the Adam optimizer [13] 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.

3. X-Humans: Dataset Details

With our high-quality, multi-view volumetric capture stage [9], we provide X-Humans, which consists of 20 subjects with various clothing types, hair styles, genders and ages. 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. 13, 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. For each subject, we split the motion sequences into a training and testing set. Overall, there are 233 sequences (35,475 frames), with 190 sequences (29,036 frames) for training and 43 sequences (6,439 frames) for testing. We also provide the ground truth SMPL[-X] registration and the way to obtain them is described in Sec. 1 of this Supp. Mat. To provide SMPL registrations we convert our SMPL-X fits using the official transfer code [18].

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.

4. Baselines: Implementation Details

For baseline SMPLX+D, we adapt the implementation from [3]. 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. For SCANimate [19] and SNARF [8], we use their public code. Since these two methods require SMPL registration, we use the official model parameter transfer code provided in [18] to convert SMPL-X parameters to SMPL parameters.

5. Supplementary Results

5.1. Metrics on Faces.

In the main paper, due to space constraints, we only include metrics for the entire body (*All*) and hands (*Hands*). We provide additional metrics for the face region (*Face*) in this Supp. Mat. In Tab. 6 and Tab. 7, we observe the same trend on the face as we did for the hands, namely that X-Avatar outperforms the baselines.

5.2. Additional Ablation Studies

Part-Aware Initialization In the main paper, we quantitatively show that compared with our part-aware initialization, the baseline that is initialized with all bones (body, hands, face) has comparable performance in terms of the geometry but with much lower computation efficiency. Fig. 15 shows further visual comparison on the hands and face. Both the shape and color look similarly for two methods, which is consistent with the quantitative analysis.

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 [7, 22]. Though we are not the first to do this, we still carry out the ablation study on texture field conditions for completeness.

Figure 13. **X-Humans gallery.** With our high-quality, multi-view volumetric capture stage [9], we provide X-Humans, which consists of 20 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.

Subject	Method	CD↓			CD-MAX ↓			NC ↑			IoU ↑		
		All	Hands	Face	All	Hands	Face	All	Hands	Face	All	Hands	Face
S1	SCANimate	2.60	8.39	2.43	54.75	54.22	19.35	0.967	0.760	0.957	0.941	0.569	0.898
	SNARF	1.37	5.13	1.28	33.86	33.51	13.07	0.977	0.818	0.966	0.967	0.739	0.937
	Ours	**0.94**	**0.79**	**0.85**	**21.43**	**4.79**	**11.50**	**0.985**	**0.957**	**0.971**	**0.991**	**0.895**	**0.943**
S5	SCANimate	3.31	6.77	8.88	44.01	43.54	27.57	0.969	0.776	0.919	0.933	0.590	0.732
	SNARF	3.04	6.93	3.12	44.30	44.08	17.25	0.972	0.768	0.949	0.936	0.586	0.890
	Ours	**0.96**	**0.73**	**1.01**	**19.55**	**3.56**	**12.34**	**0.984**	**0.960**	**0.967**	**0.992**	**0.884**	**0.947**

Table 6. More results on GRAB on the same subject reported in the main paper (S1, male) and a new subject (S5, female).

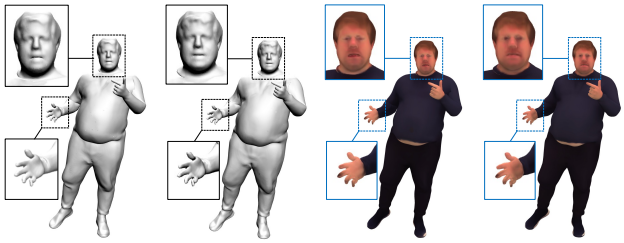

Figure 14. **Effect of our part-aware initialization strategy**. Our final model gives similar predictions on hands, face geometry and texture but with 3 times the speed of the baseline.

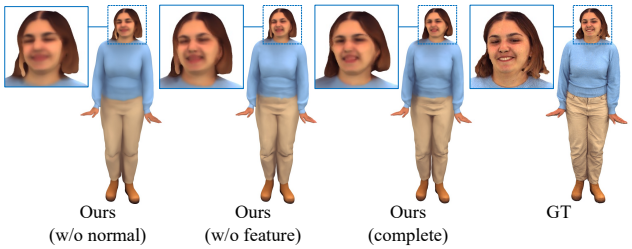

Figure 15. **Effect of our design decisions on the resulting texture.** The one with both normals and geometry features as conditons produces the sharpest details around the mouth, eyes, and clothes.

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. 15. 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. Additional Results on GRAB Dataset

In the main paper, we show results on a male subject (S1) from GRAB dataset. We also evaluate our method on a female subject from GRAB (S5, 2,392 training and 766 testing frames). Tab. 6 shows that for S5 our method's improvement compared to the baselines is even larger than for S1.

Method	CD↓			CD-MAX ↓			NC ↑			IoU ↑		
	All	Hands	Face	All	Hands	Face	All	Hands	Face	All	Hands	Face
SMPLX+D	5.75	5.19	3.41	48.41	23.48	17.69	0.921	0.790	0.915	0.957	0.774	0.905
SCANimate	6.54	9.78	4.63	59.71	48.32	23.41	0.925	0.726	0.931	0.919	0.557	0.858
SNARF	5.05	7.23	2.98	55.06	37.15	18.39	0.934	0.788	0.936	0.937	0.608	0.914
Ours	**4.43**	**5.14**	**2.29**	**47.56**	**22.15**	**15.05**	**0.939**	**0.793**	**0.948**	**0.965**	**0.776**	**0.928**

Table 7. Separate results for the entire body, hands and face on X-Humans (Scans).

5.4. Quantitative Results for Reconstruction

In the main paper, we compare our method with other baselines on the animation task. Though it's not our main focus, we also provide the results on the reconstruction task.

Reconstruction results on X-Humans (Scans) Tab. 8 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 in 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 main paper.

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	6.38	10.42	61.8	50.85	0.928	0.729	0.904	0.540
SNARF	**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 8. **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.

Method	CD↓		CD-MAX ↓		NC ↑		IoU ↑	
	All	Hands	All	Hands	All	Hands	All	Hands
PINA	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 9. **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.

Reconstruction results on X-Humans (RGB-D) In Tab. 9, we lists comparisons on RGB-D data with PINA [10]. 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.

5.5. Qualitative Results on X-Humans

We show more qualitative animation results on X-Humans (from scans in Fig. 16, from RGB-D in Fig. 17) to demonstrate that our model can generalize to various people with different body shape, clothing types, patterns and hair styles. More results can be found in the video.

5.6. Robustness to Noisy SMPL-X

We investigate the robustness of our method with respect to the noisy estimation of the input SMPL-X parameters. Specifically, we re-train our model on the GRAB subject (S1) but induce random noise with a standard deviation of 2 degrees on the SMPL-X pose parameters. I.e., we compute noisy inputs to our method as $\boldsymbol{\theta}_{\text{noisy}} = \boldsymbol{\theta} + \mathcal{N}(\mathbf{0}, 2\pi/180)$. Note that we draw individual samples for each coordinate of $\boldsymbol{\theta}$. We do not perturb the global orientation and translation. This noise level leads to significant differences between the noisy SMPL-X inputs and the original inputs (1st row, Tab. 10). Although our original method struggles with this noisy input (2nd row, Tab. 10), it does not fail entirely. Please note however that our method does not explicitly optimize the SMPL-X pose parameters. If we optimize the pose parameters jointly, robustness to the input noise increases drastically (3rd row, Tab. 10).

Method	CD↓		CD-MAX ↓		NC ↑		IoU ↑	
	All	Hands	All	Hands	All	Hands	All	Hands
Noisy SMPL-X Input	9.60	23.83	70.80	60.79	0.912	0.634	0.897	0.376
Ours (w noise, w/o pose opt.)	5.02	18.62	74.60	69.41	0.934	0.585	0.883	0.335
Ours (w noise, w pose opt.)	3.85	1.93	29.31	8.37	0.960	0.916	0.958	0.862
Ours (w/o noise)	**0.94**	**0.79**	**21.43**	**4.79**	**0.985**	**0.957**	**0.991**	**0.895**

Table 10. Robustness to noisy SMPL-X on GRAB.

5.7. Speed

When we initialize our method with all bones (Method ID A2 in Tab. 1 of the main paper, L.651), we observe a speed of 3.85 seconds per iteration (s/it) during training.

Figure 16. **More Animation results on X-Humans (Scans).**

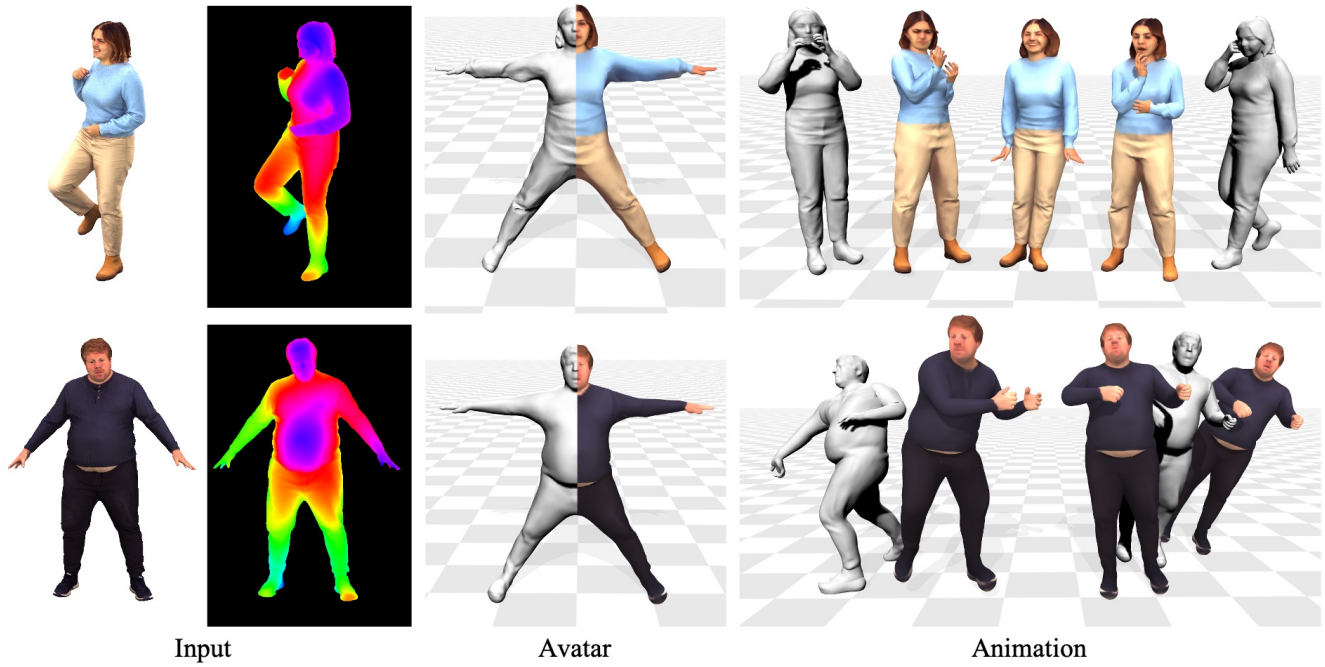

Figure 17. **More Animation results on X-Humans (RGB-D).**

Our final method takes 1.29 s/it (hence the reported speed-up of 3x). We train on a single RTX 6000 (24 GB) for 24 hours per sequence (see Sec. 2.7.). At inference time, rendering takes roughly 7 seconds per frame, which includes time for marching cubes and mesh rendering. Please note that very recently a faster version of SNARF has been proposed [6], which reports a speed-up of 150×. This is a drop-in replacement for our deformers and will accelerate our method significantly.

6. 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 counter-measures and detect malicious uses.

References

[1] Thiemo Alldieck, Gerard Pons-Moll, Christian Theobalt, and Marcus Magnor. Tex2shape: Detailed full human body geometry from a single image. In *IEEE International Conference on Computer Vision (ICCV)*. IEEE, oct 2019. 2, 3

[2] Matan Atzmon and Yaron Lipman. Sal: Sign agnostic learning of shapes from raw data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2565–2574, 2020. 2

[3] 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. 4

[4] Charles G Broyden. A class of methods for solving nonlinear simultaneous equations. *Mathematics of computation*, 19(92):577–593, 1965. 3

[5] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 1

[6] 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. 7

[7] 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. 4

[8] 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. 3, 4

[9] 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. 4, 5

[10] 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. 4, 6

[11] Charles Dugas, Yoshua Bengio, François Bélisle, Claude Nadeau, and René Garcia. Incorporating second-order functional knowledge for better option pricing. *Advances in neural information processing systems*, 13, 2000. 3

[12] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. In *Proceedings of Machine Learning and Systems 2020*, pages 3569–3579. 2020. 4

[13] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4

[14] 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

[15] 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. *Communications of the ACM*, 65(1):99–106, 2021. 2

[16] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 2

[17] Priyanka Patel, Chun-Hao P. Huang, Joachim Tesch, David T. Hoffmann, Shashank Tripathi, and Michael J. Black. AGORA: Avatars in geography optimized for regression analysis. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2021. 2

[18] 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. 1, 4

[19] 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. 4

[20] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016. 3

[21] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *CVPR*, 2017. 1

[22] 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. 3, 4