Implicit and Parametric Avatar Pose and Shape Estimation From a Single Frontal Image of a Clothed Human

Fares Mallek
École de technologie supérieure
Montreal, Quebec, Canada
fares.mallek.1@ens.etsmtl.ca

Carlos Vázquez
École de technologie supérieure
Montreal, Quebec, Canada
carlos.vazquez@etsmtl.ca

Eric Paquette
École de technologie supérieure
Montreal, Quebec, Canada
eric.paquette@etsmtl.ca

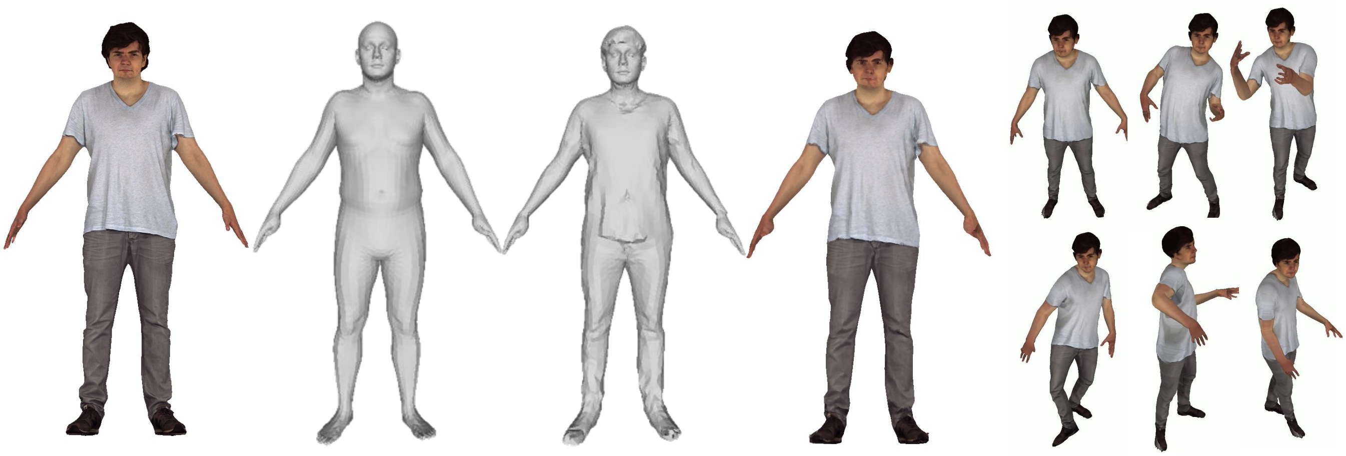


Figure 1: Illustration of our single-image reconstruction approach. From left to right: Input RGB image, SMPL-X after fitting, SMPL-X+D, rendered avatar, and avatar rendered in multiple poses.

Abstract

In this paper, we tackle the challenge of three-dimensional estimation of expressive, animatable, and textured human avatars from a single frontal image. Leveraging a Skinned Multi-Person Linear (SMPL) parametric body, we adjust the model parameters to faithfully reflect the shape and pose of the individual, relying on the mesh generated by a Pixel-aligned Implicit Function (PIFu) model. To robustly infer the SMPL parameters, we deploy a multi-step optimization process. Initially, we recover the position of 2D joints using an existing pose estimation tool. Subsequently, we utilize the 3D PIFu mesh together with the 2D pose to estimate the 3D position of joints. In the subsequent step, we adapt the body's parametric model to the 3D joints through rigid alignment, optimizing for global translation and rotation. This step provides a robust initialization for further refinement of shape and pose parameters. The next step involves optimizing the pose and the first component of the SMPL shape parameters while imposing constraints to enhance model robustness. We then refine the SMPL model pose and shape parameters by adding two new registration loss terms to the optimization cost function: a point-to-surface distance and a Chamfer distance. Finally, we introduce a refinement process utilizing a deformation vector field applied to the SMPL mesh, enabling more faithful modeling of tight to loose clothing geometry. A notable advantage of our approach is the ability to generate detailed avatars with fewer vertices compared to previous research, enhancing computational efficiency while maintaining high fidelity. To complete our model, we design a texture extraction and completion approach. Our entirely automated approach was evaluated against recognized benchmarks, X-Avatar and PeopleSnapshot, showcasing competitive performance against state-of-the-art methods. This approach contributes to advancing 3D modeling techniques, particularly in the realms of interactive applications, animation, and video games. We made our code available to the community: https://github.com/ETS-BodyModeling/ImplicitParametricAvatar.

CCS Concepts

• **Computing methodologies** → *Mesh geometry models*; *Regularization*; **Reconstruction**; **Shape inference**; **Procedural animation**.

Keywords

Human avatars, 3D modeling, parametric models, animation, textures, deep neural networks, computer vision.

MIG '24, November 21–23, 2024, Arlington, VA

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

ACM Reference Format:
Fares Mallek, Carlos Vázquez, and Eric Paquette. 2024. Implicit and Parametric Avatar Pose and Shape Estimation From a Single Frontal Image of a Clothed Human. In *Proceedings of The 16th ACM SIGGRAPH Conference on Motion, Interaction and Games Proceedings (MIG '24).* ACM, New York, NY, USA, 11 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Photo-realistic avatars has the potential to revolutionize fields ranging from XR to healthcare, and most notably the entertainment industry, by greatly enhancing the user experience while interacting with virtual humans. Despite significant recent advancements, the task of crafting realistic human avatars still presents significant challenges. Traditional methods [Ajanohoun et al. 2021; Alldieck et al. 2018b] rely on extensive input data such as multiple views, video sequences or depth information, underscoring the need for more efficient and accessible techniques. Progress in the field of 3D human modeling, while notable, encounters major challenges, particularly in faithfully reproducing the human morphology. The complexity of this task is exacerbated when modeling from a single image, a constraint that offers a promising path towards more accessible and practical applications. Deep learning-based methods [Feng et al. 2021; Zhang et al. 2021] for predicting parametric body models produce compact and animatable surfaces, but face difficulties in accurately capturing details such as clothing nuances and textures, essential aspects for creating realistic avatars. The Pixel-aligned Implicit Function (PIFu) based methods [Alldieck et al. 2022; Saito et al. 2019, 2020; Xiu et al. 2022] mark a significant advancement and are capable of reconstructing a 3D model with high resolution from a single image. However, these methods encounter difficulties in generating a compact mesh that accurately reconstructs all body parts, such as the hands and the head. Due to their representation by small pixel regions in the image space, recreating these parts proves particularly complex. This difficulty is exacerbated by the use of the marching-cubes algorithm to generate a mesh. Furthermore, the outputs of the PIFu-based methods are not directly animatable, their meshes are not compact, and they lack focusing in hard to represent areas (face, hands, and feet). The ICON method [Xiu et al. 2022] stands out for its use of a parametric human model. It optimizes the parameters to adjust to the rendering of the silhouette and normals. However, although directly animatable, this method can remove the fine details of the mesh due to the used parametric body mesh normals, which tend to be smooth, and lacks specific clothing details. The PHORHUM method [Alldieck et al. 2022], focusing on predicting the illumination to reconstruct albedo colors, encounters limits in color fidelity, thus diverging from realism. Moreover, PHORHUM, trained on perspective images, does not perform well across a wide range of camera configurations. In conclusion, while body shape modeling methods exist, they might not be as effective in texture reconstruction or animation. Combining these three aspects – modeling, animation, and texture reconstruction – from a single image remains a major challenge.

Our proposal offers a unique approach to generating a compact, animatable, expressive, and textured 3D avatar from a single frontal image in A-Pose, building upon existing methods [Ajanohoun et al. 2021; Alldieck et al. 2019, 2018b]. Figure 1 represents our 3D human body reconstruction pipeline, which relies on the Pixel-aligned Implicit Function for high-resolution 3D Human Digitization (PIFuHD) [Saito et al. 2020] as well as on OpenPose [Cao et al. 2019] to initialize the shape and pose of the avatar. We extract the 3D pose based on the 2D pose, and then fit the Skinned Multi-Person Linear eXpressive (SMPL-X) [Pavlakos et al. 2019] model to the target PIFuHD mesh. Compared to the PIFuHD mesh, SMPL-X is easy to animate and has a compact mesh. Conversely, the SMPL-X model does not allow to model the specific shape details found in the PIFuHD mesh. To overcome this, we then add a deformation vector field to the mesh and optimize it to model geometric details, such as the clothing geometry. This approach allows us to combine the detailed PIFuHD mesh with the compactness and ease of animation provided by the SMPL-X model. Next, we extract the texture and complete it using color interpolation and an image inpainting method. Our approach aims to offer a faithful representation of a wide range of human morphologies while facilitating the animation of the obtained avatar, thus widening its application potential in various contexts. Our main scientific and theoretical contributions are:

1. The introduction of a deformation vector field to model the details from the PIFuHD mesh onto the compact and easy to animate SMPL-X model;
2. A multi-step optimization process to adjust the SMPL-X model to fit humans wearing tight to loose clothing;
3. A novel approach for the generation and completion of textures resolving silhouette and back of the head artifacts.

With these contributions, our approach ensures realistic, fast, and stable animation of clothed avatars directly in off-the-shelf animation software.

2 Related Work

This section explores three elements of research regarding the reconstruction of 3D human body. We begin by exploring parametric models, then proceed to discuss 3D reconstruction, and conclude by analyzing texture extraction and completion methods.

2.1 Parametric Body Model

Two primary strategies stand out in 3D human body modeling. The first one is based on the kinematic skeleton, emphasizing an articulated structure that primarily focuses on joint movement without capturing body shape details. The skeleton model is widely utilized in 2D human pose estimation [Cao et al. 2019; Fischler and Elschlager 1973; Toshev and Szegedy 2014]. It conceptualizes the human skeleton as a hierarchical tree structure, incorporating articulated joints. The second strategy utilizes parametric models [Anguelov et al. 2005; Loper et al. 2015; Pavlakos et al. 2019], allowing for separate optimization of body shape and posture. The Skinned Multi-Person Linear (SMPL) model [Loper et al. 2015] utilizes a base shape and linear deformations to capture a variety of human shapes and poses. Its popularity in both industry and academia is attributed to its flexibility and its ability to seamlessly animate the avatar in off-the-shelf animation software. SMPL-X [Pavlakos et al. 2019] represents a significant evolution of the SMPL model, incorporating fully articulated hands and an expressive face, while still providing a compact mesh.

2.2 3D Reconstruction of the Human Body

Significant advancements have been made in the field of avatar creation. Some methods utilize multiple images [Ajanohoun et al. 2021; Mu et al. 2023; Xu et al. 2023], video sequences [Alldieck et al. 2018a; Guo et al. 2023; Jiang et al. 2023], or depth information [Lu et al. 2022; Song et al. 2023; Wang et al. 2021; Zheng et al. 2023]. While these methods are interesting when having access to more sophisticated capture setup, our research concentrates on the challenge of reconstructing avatars from a single image. Reconstructing 3D avatars from a single image typically revolves around two distinct strategies. The first strategy relies on the use of a parametric body model. A parametric model approximates the shape of the human body to be reconstructed and is characterized by a small set of parameters. These parameters define the shape and pose of the body. The estimation of a parametric model can be achieved through an optimization process of its parameters [Ajanohoun et al. 2021; Alldieck et al. 2018b; Bogo et al. 2016; Kolotouros et al. 2019b; Pavlakos et al. 2019]. Alternatively, the parametric model's parameters can be directly regressed via a Deep Neural Network (DNN) model [Feng et al. 2021; Kanazawa et al. 2017; Kolotouros et al. 2019a; Zhang et al. 2021]. DNN-based methods have recently shown promising results in reconstructing human meshes from a single image. These methods directly map raw pixels to model parameters, allowing for the production of parametric models in a feed-forward manner through neural networks.

The second strategy estimates morphology details in the form of an implicit function representation [Alldieck et al. 2022; Saito et al. 2019, 2020; Xiu et al. 2022]. The primary objective of these PIFu-based methods lies in obtaining an abundance of details, encompassing hair, and clothing. A significant drawback of the PIFu-based methods lies in the inaccurate modeling of small geometric details such as hands and face. These methods often produce lower-quality results in these areas due to the limited number of pixels compared to their complexity, resulting in inaccuracies or distortions. Another concern with these methods is the mesh: it has many more triangles than a parametric body mesh. Additionally, another major limitation is the model's inability to be animated without resorting to advanced DNN techniques [Saito et al. 2021]. Furthermore, the distribution and shape of the triangles provides lower quality animations compared to parametric body meshes.

2.3 Texture Extraction and Completion

Recent advancements in texture extraction and completion for 3D human body reconstruction from single images have shown promising developments. The Pose with Style method [AlBahar et al. 2021] leverages DensePose [Güler et al. 2018] to map the image space to the UV space of SMPL textures. It also enables the automatic synthesis of missing texture parts. While effective, this method struggles with preserving subject face details and accurately reproducing hands and clothing textures. On another front, DINAR [Svitov et al. 2023] introduced a method combining neural textures with the SMPL-X body model. DINAR achieved good quality and easily animatable avatars. It uses a diffusion model that enables realistic reconstruction of the texture in occluded regions, such as the back of a person from a frontal view. However, despite the realism of people wearing tight clothing, challenges arise from defects in the SMPL-X mesh generated by SMPLify-X [Pavlakos et al. 2019], essential for texture extraction. These defects, particularly noticeable in clothing regions, stem from the limitation of the SMPL-X model, designed solely for modeling human bodies and not clothing.

In conclusion, the SMPL-X parametric body model has several advantages (easy to optimize, compact mesh and animatable). Methods which reconstruct avatars with the SMPL-X body representation often lack details such as clothing, and many of them do not reconstruct the texture for the avatar. PIFu-based methods provide fine details, but are hard to animate, do not provide easy to use texture maps, and struggle to reconstruct fine details such as those found in the hands and the face. Finally, texture extraction and completion methods often struggle with hands and clothing. We propose a new approach to cope with all of the problems at once: recreating an easily animatable avatar, from a single image of human wearing tight or loose clothing. Our avatars benefit from fine details, good representation of the face and hands, a compact mesh, and textures.

3 Proposed Methodology

Our methodology (See Figure 2), designed as a multi-step pipeline, aims for detailed, animatable 3D reconstruction of a human subject from a single frontal image. The process initiates with the extraction of the target mesh, utilizing the PIFuHD method [Saito et al. 2020], coupled with the acquisition of 2D pose estimations via OpenPose [Cao et al. 2019]. Subsequently, this procedure progresses to the computation of three-dimensional joints. The process includes global alignment of the SMPL-X model [Pavlakos et al. 2019], optimizing its translation and rotation parameters, and further refining the model's pose and shape parameters. We introduce a deformation vector adjustment to overcome SMPL-X's clothing modeling limitations, followed by a specialized algorithm for texture extraction and completion based on the PIFuHD mesh colors. Finally, we can render the textured SMPL-X+D mesh in various poses and camera angles.

3.1 Mesh Definitions

Meshes are denoted by M, defined as a set $\{V, F\}$, where V represents the vertices and F represents the triangular faces. The SMPL-X model takes as input a translation $\mathcal{T} \in \mathbb{R}^3$, a global rotation $\mathcal{G} \in \mathbb{R}^3$, pose parameters for the body and hands $\theta = \{\theta_b, \theta_h\} \in \{\mathbb{R}^{23\times3}, \mathbb{R}^{30\times3}\}$, shape parameters for the body $\beta \in \mathbb{R}^{300}$, as well as facial expression parameters $\psi \in \mathbb{R}^{300}$. This mesh has a fixed topology with a constant number of vertices and faces:

$$M_{\text{SMPL-X}}(\mathcal{T}, \mathcal{G}, \theta, \beta, \psi) = \{V_{\text{SMPL-X}}, F_{\text{SMPL-X}}\} \subset \mathbb{R}^{n_1\times3}, \mathbb{N}^{m_1\times3}, \tag{1}$$

where $n_1 = 10475$ is the number of vertices and $m_1 = 20908$ is the number of faces. The PIFuHD mesh exhibits a variable topology, adapting its number of vertices n_2 and faces m_2 to the level of detail captured from the input image:

$$M_{\text{PIFuHD}} = \{V_{\text{PIFuHD}}, F_{\text{PIFuHD}}\} \subseteq \mathbb{R}^{n_2\times3}, \mathbb{N}^{m_2\times3}. \tag{2}$$

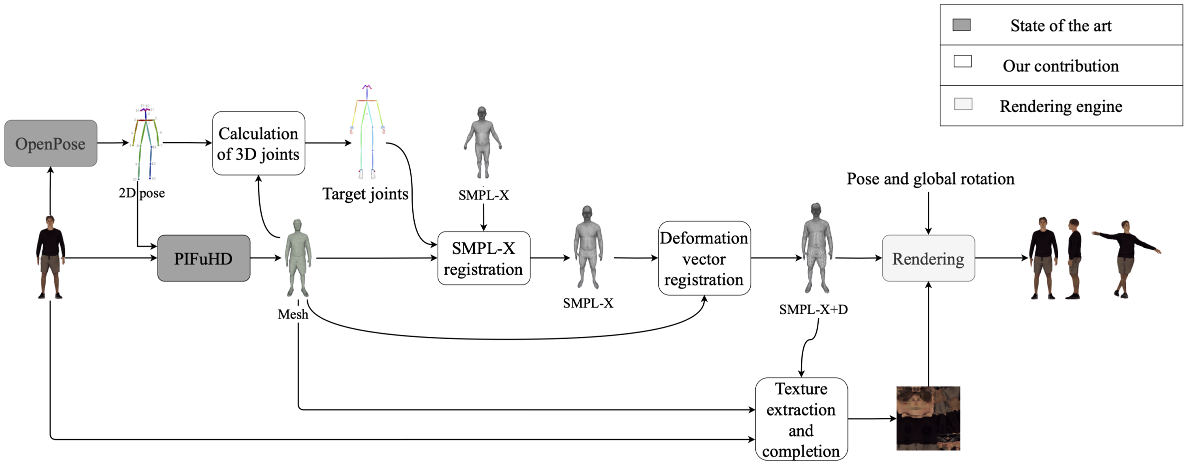


Figure 2: Illustration of the reconstruction and texturing of the SMPL-X+D mesh from a single image, along with rendering results in various poses and viewpoints.

3.2 Pose Estimation

Utilizing OpenPose [Cao et al. 2019], we extract 2D skeletal data, represented as blue points in Figure 3, which correspond to joints within the image. We project the PIFuHD mesh onto the image plane to generate the projected mesh vertices $M'_p = \{(x, y, 0) \mid (x, y, z) \in V_{\text{PIFuHD}}\}$. The 2D joints and projected vertices are now in the same reference frame. We select $k = 20$ points from M'_p closest to each OpenPose-detected joint J_i, employing a K-means algorithm to split the corresponding vertices from M_{PIFuHD} into two distinct sets, $\mathcal{F}_i$ and $\mathcal{B}_i$, laying respectively onto the front and back surfaces of the 3D mesh. We then average the centroids of these sets for each joint, thus achieving the 3D joint estimation $J_{\text{target}}(i)$. For facial keypoints, a similar technique is adopted, but this time, only the center point of the front set is used to lift each keypoint to 3D. Note that this process is not overly sensitive to the precision of the 2D pose estimation algorithm. Furthermore, our approach can take advantage of future pose detectors, as long as they are compatible with the SMPL-X joints.

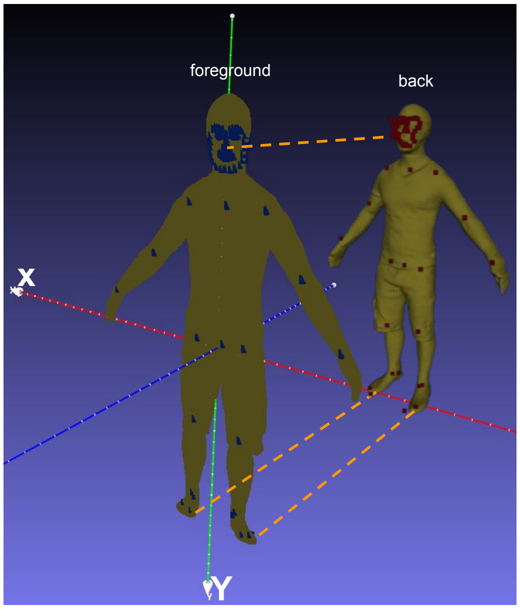


Figure 3: Orthographic projection and 3D pose estimation approach. The back shows the M_{PIFuHD} mesh, while the foreground shows the orthographic projection, M'_p, of this mesh onto the XY plane. The blue points illustrate the 2D joint estimates obtained through OpenPose. The red points correspond to these blue points lifted to the front and back surfaces of the M_{PIFuHD} mesh. While the joints for the hands are processed in the same way, they are not shown here because the density of points was not appropriate for the visualization.

3.3 Multi-Step Registration Approach

Our methodology emphasizes a sequential optimization for the SMPL-X model parameters, further refined by a deformation vector applied to the resultant $M_{\text{SMPL-X}}$ mesh, aiming for convergence with the target M_{PIFuHD} mesh. This process involves minimizing specific cost functions at successive stages.

Our pose optimization concentrates on body θ_b and hand θ_h joint parameters. The joints of the jaw and eyes in the SMPL-X model are not adjusted due to their minimal impact on the avatar's overall appearance. The optimization is carried out within a differentiable framework, relying on a cost function derived from the output mesh $M_{\text{SMPL-X}}(\mathcal{T}, \mathcal{G}, \theta, \beta, \psi)$ and the joint positions $J_{\text{SMPL-X}}(\mathcal{T}, \mathcal{G}, \theta, \beta, \psi)$, where $\mathcal{T}$ and $\mathcal{G}$ represent global translation and rotation, respectively, and θ, β, and ψ denote pose, shape, and facial expression parameters.

3.3.1 *Pose Optimization.*

In the initial stage, we set the SMPL-X model parameters $\mathcal{G}$, β, and ψ to zero, and establish a neutral "A" pose for θ. The initial translation $\mathcal{T} = T_0$ is estimated from the difference in the bounding box centers of M_{PIFuHD} and $M_{\text{SMPL-X}}$. Note that PIFuHD and SMPL-X are by default of similar sizes, corresponding to human proportions, allows for their alignment without the need for scaling. Subsequently, we refine subsets of our parameters trough a sequence of optimization stages, each using specific optimization criteria. We begin by refining $\mathcal{T}$ and $\mathcal{G}$, aiming to minimize a joint discrepancy cost function:

$$\underset{\mathcal{T},\mathcal{G}}{\operatorname{argmin}} \left(\mathcal{L}_{\text{joints}}\right), \tag{3}$$

where $\mathcal{L}_{\text{joints}}$ measures the squared L_2 norm of the difference between the SMPL-X joints and $J_{\text{target}}(i)$ joints extracted from M_{PIFuHD}.

Next, we address potential local minima leading to non-human poses by introducing a soft constraint on hand, idx_h, and body, idx_b, joints:

$$\mathcal{L}_{\text{sc}} = \sum_{i \in \text{idx}_h} (\max(0, a - \theta_i) + \max(0, \theta_i - b)) + \sum_{k \in \text{idx}_b} \alpha_k \|\theta_k\|_2^2, \tag{4}$$

where $a = -0.8$ rad, and $b = 0.5$ rad (values are not symmetric because of the SMPL-X hand rest pose) and α_k are weighting coefficients:

$$\alpha_k = \begin{cases} 10 & \text{if } k \in \{2, 5, 8, 9, 10, 11, 12, 13, 14\} \\ 1 & \text{Otherwise} \end{cases}. \tag{5}$$

The range of values for k corresponds to selected joints in the head, shoulders, torso and feet regions. A higher weight on these prevents the reconstructed body from incorrectly leaning forward/backward.

We now optimize for θ and β_0 with:

$$\underset{\theta,\beta_0}{\operatorname{argmin}} \left(\lambda_{joints}\mathcal{L}_{\text{joints}} + \lambda_{sc}\mathcal{L}_{\text{sc}}\right), \tag{6}$$

where $\lambda_{\text{joints}} = 2$, $\lambda_{\text{sc}} = 1$, and β_0 corresponds to the first component of the SMPL-X shape parameters and can be seen as mostly controlling the scale of the body.

3.3.2 Shape Optimization. Our shape optimization framework is built upon two principal cost functions: a Chamfer loss ($\mathcal{L}_{\text{chamfer}}$) and a bidirectional point-to-surface loss ($\mathcal{L}_{\text{P2S}}$), chosen to refine the SMPL-X model's alignment with the PIFuHD mesh. The Chamfer loss quantifies the proximity between SMPL-X and PIFuHD vertices. Our point-to-surface loss selects the closest pairs of vertices from PIFuHD and SMPL-X, and computes the distance between these pairs projected onto the normal vector of the SMPL-X vertex. As such, our loss favors adjustment of the SMPL-X vertices locally and perpendicular to the SMPL-X surface, thus reducing lateral sliding:

$$\begin{aligned} \mathcal{L}_{\text{P2S}} = &\frac{1}{|M_{\text{PIFuHD}}|} \sum_{p \in M_{\text{PIFuHD}}} \text{dist}(p, \tilde{v}) + \\ &\frac{1}{|M_{\text{SMPL-X}}|} \sum_{v \in M_{\text{SMPL-X}}} \text{dist}(\tilde{p}, v), \end{aligned} \tag{7}$$

where $\tilde{v} = \underset{v \in M_{\text{SMPL-X}}}{\operatorname{argmin}} \|p - v\|_2^2$ and $\tilde{p} = \underset{p \in M_{\text{PIFuHD}}}{\operatorname{argmin}} \|p - v\|_2^2$. The distance $\text{dist}(p, v)$ is expressed as:

$$\text{dist}(p, v) = \frac{|\vec{n}_v \cdot (v - p)|}{\|\vec{n}_v\|_2}, \tag{8}$$

where $\vec{n}_v$ denotes the normal at vertex v, obtained by the normalized average of the normals of the faces adjacent to v.

Our optimization function at this stage fine-tunes the SMPL-X model parameters ($\mathcal{T}$, $\mathcal{G}$, θ, β, ψ):

$$\underset{\mathcal{T},\mathcal{G},\theta,\beta,\psi}{\operatorname{argmin}} \left(\lambda_{ch}\mathcal{L}_{\text{chamfer}} + \lambda_{P2S}\mathcal{L}_{\text{P2S}} + \lambda_{joints}\mathcal{L}_{\text{joints}} + \lambda_{sc}\mathcal{L}_{\text{sc}}\right), \tag{9}$$

with weighting coefficients $\lambda_{\text{ch}} = 10$, $\lambda_{\text{P2S}} = 1$, $\lambda_{\text{joints}} = 1000$, and $\lambda_{\text{sc}} = 1$.

3.3.3 Deformation Vector Optimization. To address the SMPL model's limitations in representing clothing, we add per-vertex deformation vectors. Inspired by previous work [Alldieck et al. 2018a,b], but adapted to our single-image context, this method allows for more precise clothing representation. We optimize deformation vectors $\mathcal{D} \in \mathbb{R}^{n_1 \times 3}$ to adjust to the clothing geometry on the SMPL-X mesh, aiming to minimize the same point-to-surface loss between the adjusted mesh and the PIFuHD target. To ensure stability and realistic mesh deformation, we incorporate a regularization term $\mathcal{L}_{\text{reg}}$, combining Laplacian smoothing, normal consistency and an L_2 norm on the deformation vector:

$$\mathcal{L}_{\text{reg}} = \lambda_1\mathcal{L}_{\text{Laplacian}} + \lambda_2\mathcal{L}_{\text{normals}} + \lambda_3\|\mathcal{D}\|_2^2 + \lambda_4\|\mathcal{D}_{\text{idx}_{f\&h}}\|_2^2, \tag{10}$$

where λ_1=10 and λ_2=10. We set a different weighting on the deformation vector loss $\mathcal{D}_{\text{idx}_{f\&h}}$ for the face and hands ($\lambda_4 = 10^4$) compared to the deformation vector loss $\mathcal{D}$ for the other parts of the body ($\lambda_3 = 1$). The hands and face are not always correctly reconstructed by PIFuHD and it is best in these regions to favor the SMPL-X shape by penalizing large deformation vectors. At this stage, our optimization equation is thus formulated as:

$$\underset{\mathcal{D}}{\operatorname{argmin}} \left(\mathcal{L}_{\text{P2S}} + \mathcal{L}_{\text{reg}}\right), \tag{11}$$

where the two losses are simply added together. This deformation vector optimization greatly improves the clothing representation, capturing the wrinkles and later helping with the texture extraction. Our optimization strategy effectively integrates local adjustments within a broader global framework through the parameterization of the SMPL-X model. This approach ensures that any local changes, such as those between specific points and vertices, are seamlessly incorporated into the overall structure of the SMPL-X model. Additionally, we enhance the fidelity of these adjustments by employing Laplacian and normal consistency losses. These losses are crucial as they maintain the mesh smoothness and continuity, ensuring that local optimizations do not compromise the global integrity and realistic appearance of the model. Thus, our method achieves a balance between refining detailed features and preserving surface smoothness.

The high-resolution mesh of M_{PIFuHD} results in significant computational time and memory usage during the optimization. Our experiments demonstrated that subsampling M_{PIFuHD} to match the vertex count of the $M_{\text{SMPL-X}}$ mesh, significantly reduces computation time while having a negligible impact on the resulting quality. To achieve a reduction in the number of vertices V_{PIFuHD}, we employed a farthest point sampling method [Ge et al. 2018]. Note that we do not coarsen the mesh; we only subsample the vertices as the polygons of PIFuHD are not needed in our loss functions.

In our computational framework, the Adam optimizer [Kingma and Ba 2014] is consistently utilized across all stages. The selection of learning rates is tailored to each specific aspect of the optimization process: 10^{-3} for rigid transformations, 10^{-4} for pose adjustments, 10^{-2} for shape optimization, and 10^{-4} for the refinement of deformation vectors.

3.4 Texture Extraction and Completion

Now that the geometry is adjusted, we extract the color information for the avatar from the PIFuHD mesh. Employing a blend of interpolation techniques followed by a texture inpainting technique ensures a faithful texture representation. For each texel center in the UV map of SMPL-X, we identify the closest triangle and convert the texel's position to barycentric coordinates within this triangle of the SMPL-X+D mesh. From the corresponding 3D position, we fetch the color from the nearest PIFuHD mesh vertex.

Colors at the silhouette of the PIFuHD mesh exhibit color leakage from the background as can be seeing in Figure 4. To identify these wrong silhouette texel colors, we extract the colors from the original image, and from an image with a different uniform background color. This second image is generated by detecting the background in the original image using the Rembg tool [Gatis 2023] and replacing it with a uniform color. Texels exhibiting differences in colors correspond to silhouette texels and should be synthesized. Horizontal linear interpolation is used to fill these silhouette texels from the left and right "valid" texel colors. Figure 5 illustrates this process. Another challenge in the extracted texture lies in the fact that the PIFuHD method employs a naive symmetry to assign colors to the back of the avatar. This negatively impacts occluded parts in the region at the back of the head. To address this issue, we employ the LaMa image inpainting method [Suvorov et al. 2021]. This method requires an input image and a mask specifying the area to be inpainted. In our case, we manually crafted a static mask targeting the back of the head. This mask remains unchanged and applied to all reconstructions, regardless of variations in the input images. This approach is justified by the fact that in the UV space of SMPL-X, the posterior region of the head is always at the same position. The use of this method allows for a more realistic back of the head, as illustrated in Figure 5 (c).

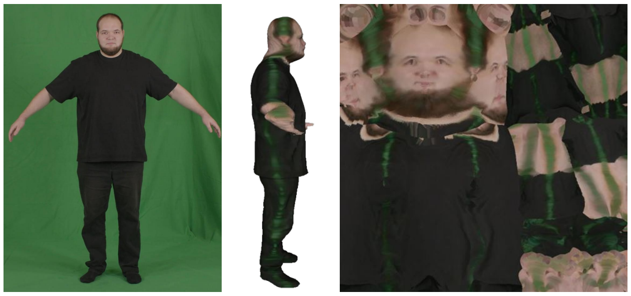

(a) (b) (c)

Figure 4: Silhouette color leakage. From left to right - the input image, the PIFuHD mesh, and the resulting texture extracted from PIFuHD.

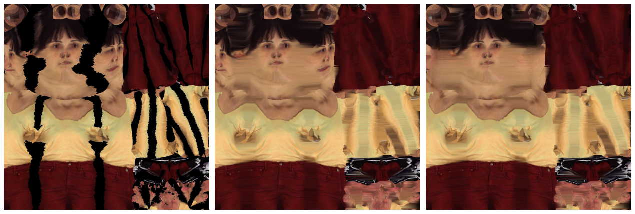

(a) (b) (c)

Figure 5: From left to right: Texture extracted from PIFuHD, texture after linear interpolation in the silhouette areas, texture following the application of the LaMa inpainting method on the back of the head.

4 Results

In this section, we evaluate our 3D reconstruction approach using two open-access datasets. The X-Avatar dataset [Shen et al. 2023] features 20 subjects from scanned real bodies, with synthetically generated images using PyTorch3D. It presents a good diversity across body shapes, poses, and demographics. PeopleSnapshot [Alldieck et al. 2018b] captures 12 subjects in A-pose through perspective RGB video from a camera 2 meters away. For testing, we used the video's first frame showing the subject's frontal view. Note that these two datasets do not overlap with PIFuHD training dataset.

4.1 Quantitative Evaluation

We benchmarked our results against those achieved by PIFu [Saito et al. 2019], PIFuHD [Saito et al. 2020], and ICON [Xiu et al. 2022]. This comparison is based on a set of specific metrics. Intersection over Union (IoU) [Harouni and Baghmaleki 2018] measures segmentation accuracy by calculating the ratio of overlap between the predicted and actual silhouettes, where a higher score indicates better performance. Chamfer Distance (CD) [Barrow et al. 1977] evaluates the similarity between two sets of vertices, with lower values denoting closer matches. Normal Consistency (NC) [Mescheder et al. 2019] assesses the agreement of surface normals between the reconstructed model and the reference, aiming for a score close to one for an ideal match. The Structural Similarity Index (SSIM) [Wang et al. 2004] and Peak Signal-to-Noise Ratio (PSNR) [Horé and Ziou 2010] gauge image quality, considering aspects like texture, luminance, and contrast, with higher values indicating superior image reconstruction. Finally, the Learned Perceptual Image Patch Similarity (LPIPS) [Zhang et al. 2018] metric evaluates perceptual similarity between images, focusing on high-level visual features significant for human perception, where closer matches yield lower scores.

Table 1 presents comparative results based on the X-Avatar dataset. Our approach exhibits robust and competitive performance across various metrics, affirming its efficacy for single-view 3D reconstruction. While slightly outperformed in some cases, the differences are minor. The slight performance decrement is partly attributed to the use of a parametric body model, which, despite offering substantial flexibility, may struggle to capture small body or clothing details. Our results do not exhibit a pronounced advantage in metrics such as LPIPS, PSNR for Rendered Normals, and SSIM for

Table 1: Numerical comparisons of single-view 3D reconstructions on the X-Avatar dataset. Best results are highlighted in bold green and second-best in amber.

	3D Metrics		Rendered Normals			Rendered RGB Images				
Method	CD ↓	NC ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	IoU ↑	Nbr vertices ↓
PIFu [Saito et al. 2019]	1.16	0.808	0.835	0.142	18.54	0.832	0.144	19.90	0.971	50000
PIFuHD [Saito et al. 2020]	**0.76**	**0.823**	0.857	**0.089**	**21.62**	**0.912**	0.093	21.55	**0.984**	170000
ICON [Xiu et al. 2022]	2.98	0.721	0.833	0.125	18.48	0.805	0.143	17.89	0.947	48000
Ours	0.91	0.803	**0.869**	0.127	20.75	0.896	**0.075**	**23.23**	0.974	**10475**

Rendered RGB Images primarily due to the underlying structure of our model. Our reconstruction relies on a parametric model which utilizes less than six percent of the vertices of the PifuHD model. This reduction in vertex density inherently limits our model's capacity to capture extremely fine geometric details, such as hair, and to precisely converge to the complex geometries exemplified by PifuHD. Note that, for the comparison found in Table 1, we excluded PHORHUM due to its training specificity on perspective data and DINAR for its focus on neural network applications rather than geometric reconstruction of clothed bodies. This ensures a fair and relevant comparison across methods grounded in orthographic rendering. Finally, in Table 1, ICON performs worse than PIFuHD in terms of Chamfer distance. In the ICON paper, the experiments use difficult poses, effectively highlighting how ICON is significantly better than PIFuHD in that context. In contrast, our experiments were conducted with frontal images and relatively simple poses, a setting in which PIFuHD outperforms ICON, which explains the apparent discrepancy in Chamfer distance between our study and that reported in the ICON paper.

4.2 Qualitative Evaluation

Quantitative evaluations do not always align with human perception. Therefore, we present qualitative results of our approach alongside the methods of PIFu, PIFuHD, ICON, and PHORHUM on synthetic images in Figure 6 and Figure 7, as well as a comparison on real images in Figure 8. Figure 6 focuses on comparing input images to rendered images from identical viewpoints. Our rendered images closely mirror the source images. Conversely, PHORHUM reveals deficiencies in color restitution, attributed to their unreliable attempt at estimating scene lighting for albedo color reconstruction. Alternative methods, including ICON, PIFu, and PIFuHD, exhibit performances comparable to ours, with the lower resolution of ICON and PIFu resulting in a slight loss of sharpness in the rendered images.

We then assess the performance of our approach in generating rendered images from new perspectives with the X-Avatar (Figure 7) and PeopleSnapshot (Figure 8) datasets. Our approach excels in estimating shape, pose, and colors, outperforming PIFu and PHORHUM. PHORHUM, in particular, exhibits anomalies in color and pose estimation, while PIFu struggles with color completion issues, especially near the silhouette of the body. Furthermore, our approach benefits from the use of a parametric model, enabling the generation of more natural and realistic face and hand shapes.

Concluding this evaluation, it is crucial to highlight a distinctive advantage of our approach: the ability to easily animate the reconstructed 3D avatars using linear blend skinning. This feature starkly contrasts with other methods that do not facilitate such direct animation. Illustrating the animation capability of the proposed approach, Figure 9 presents three animations generated from the extensive AMASS dataset of human motions [Mahmood et al. 2019] showcasing the versatility of our approach.

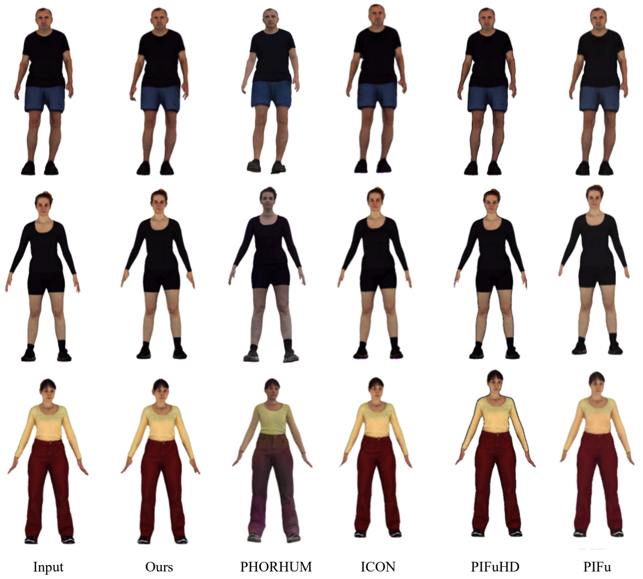


Figure 6: Qualitative evaluation of X-Avatar samples (same as input view).

Animation 1 (Figure 9a) features a series of dance poses. Animation 2 (Figure 9b) depicts an avatar executing gymnastic poses. Animation 3 (Figure 9c) demonstrates the capacity of our approach to capture and reproduce a range of facial expressions and hand movements.

4.3 Ablation Study

In this section, we present an ablation study on the multiple steps and optimizations of our model, focusing on geometric and color reconstruction using the X-Avatar dataset. We conduct a series of tests where individual components are removed from our pipeline. Table 2 allows us to isolate and understand the impact of each component on the overall performance. The last row (Ours) shows that our full pipeline has the best and second best values for five

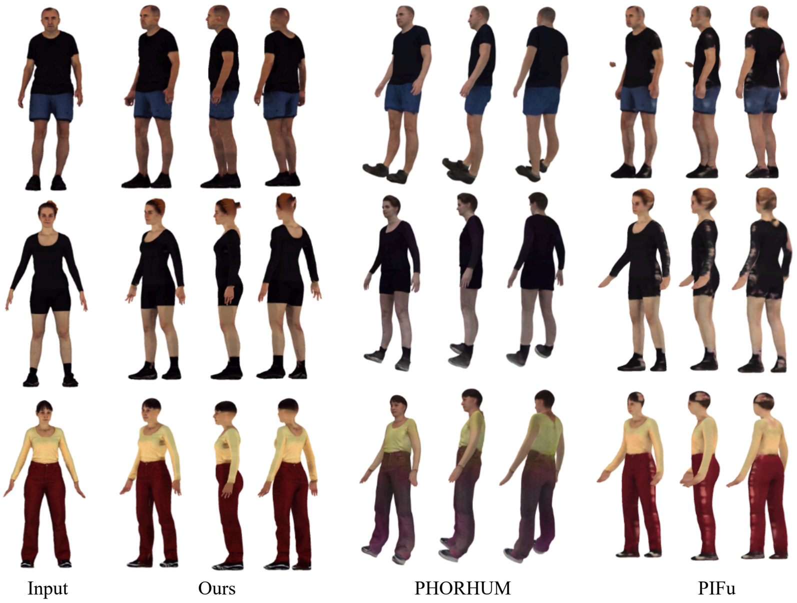


Figure 7: Qualitative evaluation of X-Avatar samples across varied perspectives, distinct from the initial view

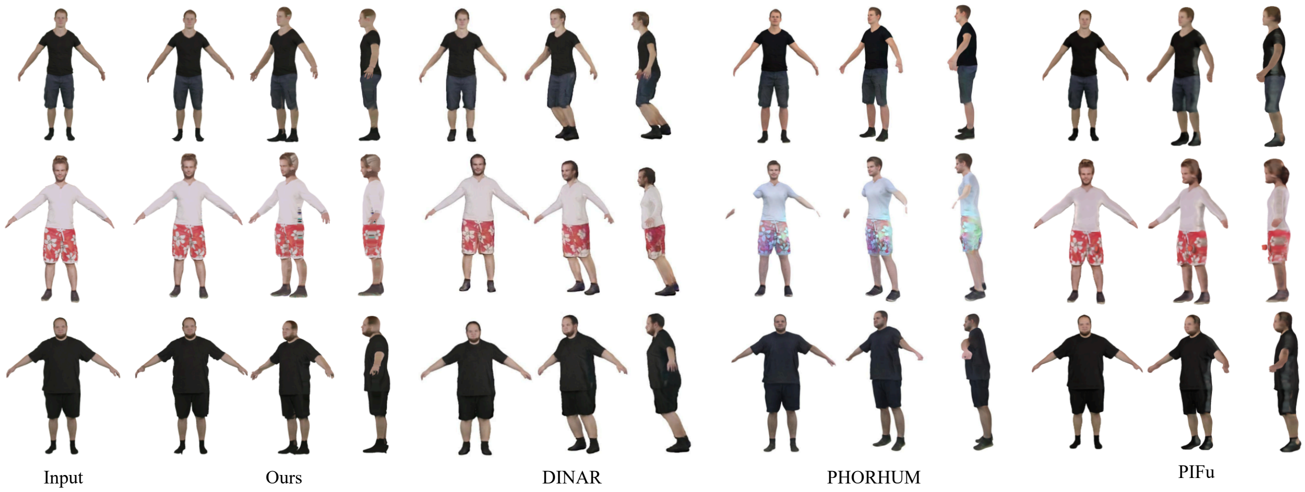
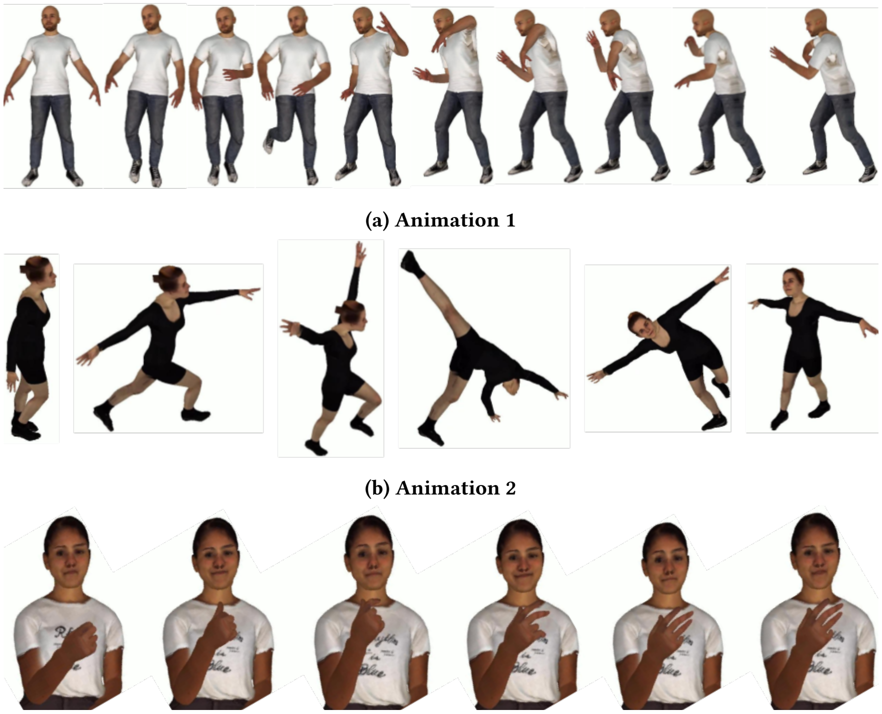


Figure 8: Qualitative evaluation of PeopleSnapshot samples

out of nine measures, demonstrating that it outperforms most of the other configurations. Rows labelled "w/o P2S" in Table 2 and the column labelled "w/o P2S" in Figure 10 illustrate the critical role of the point-to-surface loss in Equation 9 and 11, collecting the worst quantitative metric values. Rows "w/o L_2 norm hand & face" and "w/o L_2 norm body, hand & face" in Table 2 show that the quantitative measures are better without the L_2 norm, but the qualitative results are much worse as can be seen in Figure 10 "w/o L_2 norm hand & face" (similar qualitative problems occur for "w/o L_2 norm body, hand & face"). The removal of the L_2 norm for the hand and face parts in our model increases flexibility in the deformation process, allowing for a better coverage of these areas when projected in image space. However, one can see that the reconstruction of the hands in column "w/o L_2 norm hand & face" of Figure 10 is quite degraded compared to our full pipeline. According to Table 2, Equation 11 performs better in terms of Chamfer distance when ignoring the regularization term, but again we can see that the qualitative result is worse than the full pipeline (column "w/o regularization" in Figure 10), with flipped and intersecting triangles on the body and hands.



(a) Animation 1

(b) Animation 2

(c) Animation 3

Figure 9: Presentation of three rendered animations featuring three subjects in diverse body poses and expressions

Table 2: Comparison with respect to the ablated components. Best results highlighted in bold green, second-best in amber, worst in *red* italics.

	3D Metrics		Rendered Normals			Rendered RGB Images			
Method	CD ↓	NC ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	IoU ↑
Ours w/o sc Eq. 6	0.927	0.802	0.867	0.126	20.67	0.895	0.076	22.89	0.973
Ours w/o sc Eq. 9	0.910	0.801	0.867	0.127	20.79	0.896	0.076	23.29	0.973
Ours w/o P2S Eq. 9	*1.304*	*0.768*	*0.830*	0.169	*18.47*	0.864	0.111	19.41	*0.936*
Ours w/o Chamfer Eq. 9	0.916	0.801	0.867	0.126	20.62	0.896	0.076	22.79	0.972
Ours w/o regularization Eq. 10	**0.899**	0.795	0.865	0.130	20.57	0.896	0.080	22.87	0.972
Ours w/o Laplacian Eq. 10	0.920	0.801	0.866	0.126	20.65	0.895	0.076	22.82	0.973
Ours w/o normals Eq. 10	0.924	0.800	0.868	0.126	20.65	0.896	**0.075**	22.74	0.973
Ours w/o L_2 norm body Eq. 10	0.917	0.801	0.866	0.127	20.65	0.895	0.077	22.83	0.972
Ours w/o L_2 norm hand & face Eq. 10	0.903	0.801	0.869	0.126	**20.88**	**0.900**	**0.075**	23.61	**0.975**
Ours w/o L_2 norm body, hand & face Eq. 10	0.906	0.800	0.868	**0.125**	**20.88**	**0.900**	**0.075**	**23.64**	**0.975**
Ours w/o P2S Eq. 11	0.97	0.800	0.831	*0.174*	18.51	*0.837*	*0.126*	*18.90*	0.938
Ours	0.910	**0.803**	**0.869**	0.127	20.75	0.896	**0.075**	23.23	0.974

4.4 Discussion

The quantitative and qualitative evaluations confirm the ability of our approach to deliver high-quality 3D reconstruction. It validates not only the numerical accuracy of our approach but also its robustness and flexibility across varied visual and functional scenarios. Our approach is reasonably fast, requiring 2 to 4 minutes of computation to reconstruct the pose, shape, and texture of the results presented in this paper. We used a computer with 2 cores at 2.2 GHz, 24 GB of memory and an NVidia L4 GPU.

The conducted experiments confirmed fidelity of the resulting mesh. Notably, the incorporation of a Laplacian regularization loss significantly smoothed the mesh, reducing the irregularities and discontinuities seen in previous methods. Table 3 highlights the distinctions between our approach and other methods.

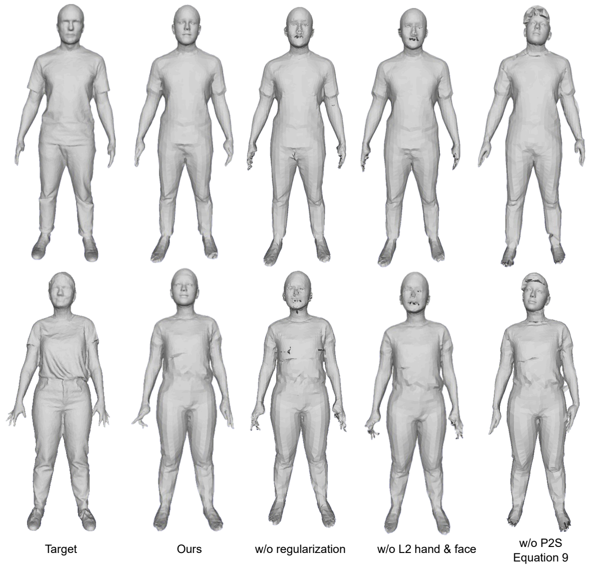


Figure 10: Qualitative ablation

Table 3: Comparison according to several criteria

	Single image	Shape variability	Animation	Expression	Textured	Compact
SMPLify-X	✓	✗	✓	✓	✗	✓
Video Avatar	✗	✓	✓	✗	✓	✓
PIFuHD	✓	✓	✗	✗	✗	✗
PHORHUM	✓	✓	✗	✗	✓	✗
Ours	✓	✓	✓	✓	✓	✓

Our approach, while using a mesh with fewer vertices compared to PIFu, PIFuHD, ICON, and PHORHUM (≈ 6% compared to PIFuHD), achieves levels of details that are comparable to implicit function-based methods, leading to fine-detailed avatars. Unlike the PIFu-based methods relying on deep learning models like SCANimate for animation, our approach uses the SMPL-X model, favouring robust, widely-used animation techniques like linear blend skinning. In terms of expressiveness, our approach, through the use of SMPL-X, allows for animations with a wider range of facial expressions and hand movements, surpassing other methods limited to body postures. Our texture process also outperforms others, providing avatars with rich and more detailed textures.

5 Conclusion

In this paper, we tackled the challenge of generating 3D human avatars from a single image. Our approach was driven by the objective to make these avatars realistic, animatable and expressive. By leveraging cutting-edge techniques such as PIFuHD, OpenPose, and the SMPL-X model, we have succeeded in producing 3D avatars that faithfully replicate the human morphology. We utilized PIFuHD to generate an accurate target 3D mesh and relied on OpenPose to estimate 2D joints that are subsequently lifted to 3D. We then fit an SMPL-X model to this target mesh by applying a sequence of optimization steps. We started with a rigid registration and then refined the shape and pose parameters. We introduced a final refinement process by applying a deformation vector to the SMPL-X mesh for a more faithful modeling of clothing geometry. Finally, we incorporated a phase of texture extraction and completion. We showed that our approach outperforms the related work when considering several evaluation criteria: reconstructs from a single image, uses a compact mesh, models humans wearing tight to loose clothing, produces a plausible reconstruction of hands and face, synthesizes a realistic texture, and allows easy animation of the avatars. None of the methods we have compared to could simultaneously achieve a good performance on all of these criteria.

Overall, the proposed approach represents a significant step toward achieving realistic and animatable human avatars, laying the groundwork for future improvements. While promising, our texture completion requires further refinement for enhanced fidelity. Integrating advanced deep learning techniques might better capture fine details and complex textures, particularly in challenging areas like hair and clothing folds. While our approach is successful regarding certain types of loose clothing, it does not yet support very loose garments, like skirts. Improvements to include such garments would require to rethink our use of the SMPL-X mesh to allow for different garment topologies. While PIFuHD works well for the global shape of the body, its reconstruction of the hands is sometimes poor, and our approach suffers from that. Investigating better methods for the reconstruction of hands could provide significant improvements in that sense.

Acknowledgments

This work was supported in part by grants from NSERC (# RGPIN-2021-04293 and # RGPIN-2019-05252) and MITACS (# IT19934).

Declaration of AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT-4 in order to improve the flow and language of the text. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

Jordy Ajanohoun, Eric Paquette, and Carlos Vázquez. 2021. Multi-View Human Model Fitting Using Bone Orientation Constraint and Joints Triangulation. In *2021 IEEE International Conference on Image Processing (ICIP)*. 1094–1098.

Badour AlBahar, Jingwan Lu, Jimei Yang, Zhixin Shu, Eli Shechtman, and Jia-Bin Huang. 2021. Pose with Style: Detail-Preserving Pose-Guided Image Synthesis with Conditional StyleGAN. *ACM Transactions on Graphics (TOG)* 40, 6 (2021), 1–11.

Thiemo Alldieck, Marcus Magnor, Bharat Lal Bhatnagar, Christian Theobalt, and Gerard Pons-Moll. 2019. Learning to Reconstruct People in Clothing from a Single RGB Camera. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 1175–1186.

Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. 2018a. Detailed human avatars from monocular video. In *2018 International Conference on 3D Vision (3DV)*. IEEE, 98–109.

Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. 2018b. Video based reconstruction of 3D people models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 8387–8397.

Thiemo Alldieck, Mihai Zanfir, and Cristian Sminchisescu. 2022. Photorealistic Monocular 3D Reconstruction of Humans Wearing Clothing. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), 1496–1505.

Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. 2005. SCAPE: Shape Completion and Animation of People. *ACM Trans. Graph.* 24, 3 (jul 2005), 408–416.

Harry G Barrow, Jay M Tenenbaum, Robert C Bolles, and Helen C Wolf. 1977. Parametric correspondence and chamfer matching: Two new techniques for image matching. In *Proceedings: Image Understanding Workshop*. Science Applications, Inc, 21–27.

Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. 2016. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *European Conference on Computer Vision (ECCV)*. Springer, 561–578.

Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. 2019. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019), 14.

Yao Feng, Vasileios Choutas, Timo Bolkart, Dimitrios Tzionas, and Michael J Black. 2021. Collaborative Regression of Expressive Bodies using Moderation. In *2021 International Conference on 3D Vision (3DV)*. IEEE, 792–804.

Martin A Fischler and Robert A Elschlager. 1973. The Representation and Matching of Pictorial Structures. *IEEE Transactions on computers* 100, 1 (1973), 67–92.

Daniel Gatis. 2023. Rembg: Rembg is a Tool to Remove Images Background. https://github.com/danielgatis/rembg. https://github.com/danielgatis/rembg Accessed: 2024-03.

Liuhao Ge, Yujun Cai, Junwu Weng, and Junsong Yuan. 2018. Hand PointNet: 3D Hand Pose Estimation using Point Sets. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8417–8426.

Riza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. 2018. DensePose: Dense Human Pose Estimation in the Wild. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), 7297–7306. https://api.semanticscholar.org/CorpusID:13637778

Chen Guo, Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. 2023. Vid2Avatar: 3D Avatar Reconstruction from Videos in the Wild via Self-supervised Scene Decomposition. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), 12858–12868. https://api.semanticscholar.org/CorpusID:257078760

Majid Harouni and Hadi Yazdani Baghmaleki. 2018. Color image segmentation metrics. In *Encyclopedia of Image Processing*, Phillip A. Laplante (Ed.). Vol. 95. CRC Press, 10–21.

Alain Horé and Djemel Ziou. 2010. Image Quality Metrics: PSNR vs. SSIM. In *2010 20th International Conference on Pattern Recognition*. 2366–2369.

Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. 2023. InstantAvatar: Learning Avatars from Monocular Video in 60 Seconds. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 16922–16932. https://doi.org/10.1109/CVPR52729.2023.01623

Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. 2017. End-to-End Recovery of Human Shape and Pose. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), 7122–7131.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *CoRR* abs/1412.6980 (2014). https://api.semanticscholar.org/CorpusID:6628106

Nikos Kolotouros, Georgios Pavlakos, Michael J. Black, and Kostas Daniilidis. 2019b. Learning to Reconstruct 3D Human Pose and Shape via Model-Fitting in the Loop. *IEEE/CVF International Conference on Computer Vision (ICCV)* (2019), 2252–2261.

Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. 2019a. Convolutional Mesh Regression for Single-Image Human Shape Reconstruction. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), 4496–4505.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. 2015. SMPL: A Skinned Multi-Person Linear Model. *ACM Trans. Graph.* 34, 6, Article 248 (oct 2015), 16 pages. https://doi.org/10.1145/2816795.2818013

Yang Lu, Han Yu, Wei Ni, and Liang Song. 2022. 3D real-time human reconstruction with a single RGBD camera. *Applied Intelligence* 53, 8 (aug 2022), 8735–8745. https://doi.org/10.1007/s10489-022-03969-4

Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. 2019. AMASS: Archive of Motion Capture as Surface Shapes. In *IEEE/CVF International Conference on Computer Vision (ICCV)*. 5442–5451.

Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. 2019. Occupancy Networks: Learning 3D Reconstruction in Function Space. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 4455–4465. https://doi.org/10.1109/CVPR.2019.00459

Jiteng Mu, Shen Sang, Nuno Vasconcelos, and Xiaolong Wang. 2023. ActorsNeRF: Animatable Few-shot Human Rendering with Generalizable NeRFs. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. 18345–18355. https://doi.org/10.1109/ICCV51070.2023.01686

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. 2019. Expressive Body Capture: 3D Hands, Face, and Body from a Single Image. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10975–10985.

Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. 2019. PIFu: Pixel-Aligned Implicit Function for High-Resolution Clothed Human Digitization. In *IEEE/CVF International Conference on Computer Vision (ICCV)*. 2304–2314.

Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. 2020. PIFuHD: Multi-Level Pixel-Aligned Implicit Function for High-Resolution 3D Human Digitization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 84–93.

Shunsuke Saito, Jinlong Yang, Qianli Ma, and Michael J Black. 2021. SCANimate: Weakly Supervised Learning of Skinned Clothed Avatar Networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2886–2897.

Ka Shen, Chen Guo, Manuel Kaufmann, Juan Jose Zarate, Julien Valentin, Jie Song, and Otmar Hilliges. 2023. X-Avatar: Expressive Human Avatars. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), 16911–16921.

H. Song, B. Yoon, W. Cho, and W. Woo. 2023. RC-SMPL: Real-time Cumulative SMPL-based Avatar Body Generation. In *2023 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. 89–98. https://doi.org/10.1109/ISMAR59233.2023.00023

Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor S. Lempitsky. 2021. Resolution-robust Large Mask Inpainting with Fourier Convolutions. *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* (2021), 3172–3182.

David Svitov, Dmitrii Gudkov, Renat Bashirov, and Victor Lempitsky. 2023. DINAR: Diffusion inpainting of neural textures for one-shot human avatars. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. 7039–7049.

Alexander Toshev and Christian Szegedy. 2014. DeepPose: Human Pose Estimation via Deep Neural Networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1653–1660.

Shaofei Wang, Marko Mihajlovic, Qianli Ma, Andreas Geiger, and Siyu Tang. 2021. MetaAvatar: Learning Animatable Clothed Human Models from Few Depth Images. In *Advances in Neural Information Processing Systems*.

Zhou Wang, Ligang Lu, and Alan C. Bovik. 2004. Video quality assessment based on structural distortion measurement. *Signal Processing: Image Communication* 19 (2 2004), 121–132. Issue 2.

Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J Black. 2022. ICON: Implicit Clothed humans Obtained from Normals. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 13286–13296.

Baixin Xu, Jiarui Zhang, Kwan-Yee Lin, Chen Qian, and Ying He. 2023. Deformable Model-Driven Neural Rendering for High-Fidelity 3D Reconstruction of Human Heads Under Low-View Settings. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. 17878–17888. https://doi.org/10.1109/ICCV51070.2023.01643

Hongwen Zhang, Yating Tian, Xinchi Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun. 2021. PyMAF: 3D Human Pose and Shape Regression with Pyramidal Mesh Alignment Feedback Loop. *IEEE/CVF International Conference on Computer Vision (ICCV)* (2021), 11426–11436.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 586–595.

Ruichen Zheng, Peng Li, Haoqian Wang, and Tao Yu. 2023. Learning Visibility Field for Detailed 3D Human Reconstruction and Relighting. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), 216–226. https://api.semanticscholar.org/CorpusID:258298156