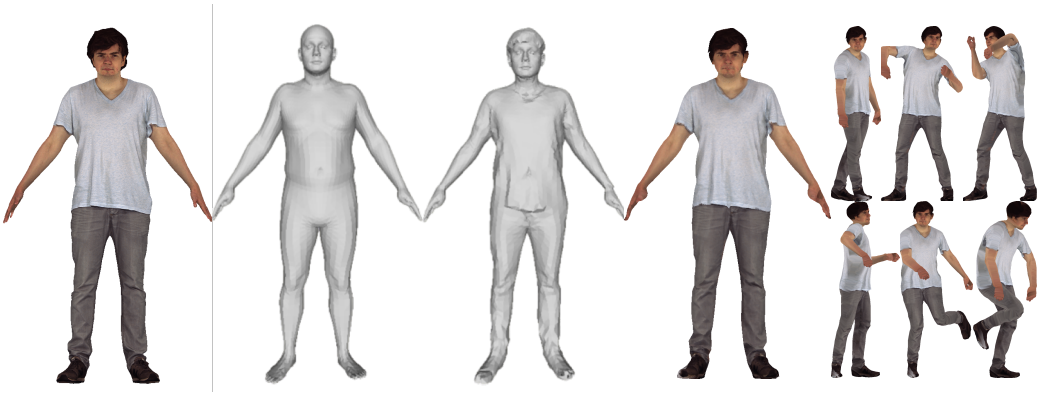
Graphical Abstract

Parametric Model Fitting
for Textured and Animatable 3D Avatar
From a Single Frontal Image of a Clothed Human

Fares Mallek, Carlos Vázquez, Eric Paquette



Highlights

Parametric Model Fitting
for Textured and Animatable 3D Avatar
From a Single Frontal Image of a Clothed Human

Fares Mallek, Carlos Vázquez, Eric Paquette

- Introduction of a deformation vector to model the details from the PIFuHD mesh
- Multi-step optimization process to fit SMPL-X to humans wearing clothing
- Design of an easily animatable new SMPL-X mesh topology with shoe-like feet
- Addition of a subdivision of the SMPL-X mesh to improve the representation of details while computing the deformation vectors
- Approach to resolve silhouette and back of the head texture artifacts

Parametric Model Fitting for Textured and Animatable 3D Avatar From a Single Frontal Image of a Clothed Human

Fares Mallek[a,*], Carlos Vázquez[a], Eric Paquette[a]

[a]*Software and IT Engineering Department, École de technologie supérieure, 1100 Notre-Dame Street West Montréal, H3C 1K3, Québec, Canada*

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 model, 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. As most other works, we optimize based on images of people wearing shoes, resulting in artifacts in the toes region of SMPL. We thus introduce a new shoe-like mesh topology which greatly improves the quality of the reconstructed feet. 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. We also demonstrate how to gain even more details, while maintaining the advantages of SMPL. 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 will make our code and our improved SMPL mesh topology available to the community: https://github.com/ETS-BodyModeling/ImplicitParametricAvatar.

Keywords: Human avatar, Reconstruction, SMPL-X, Optimization, 3D modeling, Parametric model, Animation, Textures, Computer vision

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 [1, 2] 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 [3, 4] 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 [5, 6, 7, 8] mark a significant advancement and are capable of reconstructing a 3D model with high resolution

*Corresponding author
Email address: fares.mallek.1@ens.etsmtl.ca (Fares Mallek)

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 [8] 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 [5], 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. The method of Mallek *et al.* [9] reconstructs an animatable SMPL-X avatar with a good texture, but its optimization of the feet region introduces visible and annoying artifacts. Moreover, the geometric details of the clothing are limited by the lower resolution of the SMPL-X mesh. 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 the method of Mallek *et al.* [9]. Figure 1 represents our 3D human body reconstruction pipeline, which relies on the Pixel-aligned Implicit Function for high-resolution 3D Human Digitization (PIFuHD) [7] as well as on OpenPose [10] 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) [11] 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 anima-

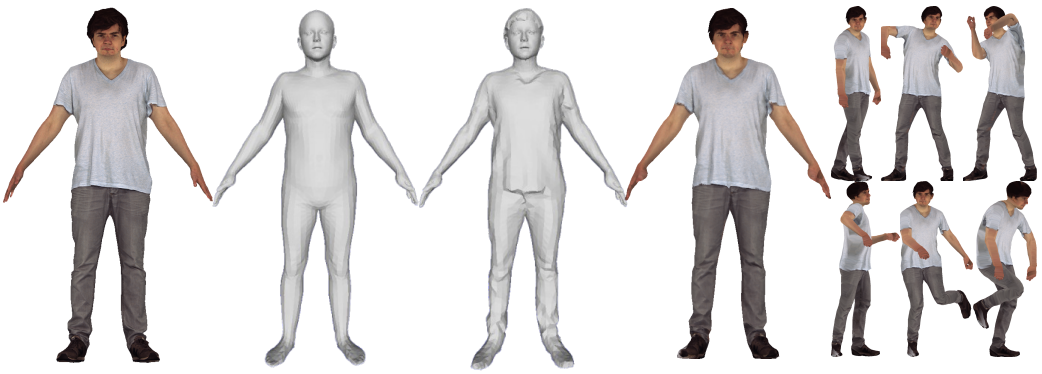


Figure 1: Illustration of our single-image reconstruction approach. From left to right: Input RGB image, SMPL-X after fitting, SMPL-X+D, rendered avatar, and avatar rendered in multiple poses.

tion provided by the SMPL-X model. We also demonstrate how to further increase the fine details while preserving the advantages of SMPL-X. 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. The design of an easily animatable new SMPL-X mesh topology, appropriate for images of people wearing shoes;
4. 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 [10, 12, 13]. It conceptualizes the human skeleton as a hierarchical tree structure, incorporating articulated joints. The second strategy utilizes parametric models [11, 14, 15], allowing for separate optimization of body shape and posture. The Skinned Multi-Person Linear (SMPL) model [15] 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 [11] 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 [2, 16, 17], video sequences [18, 19, 20], or depth information [21, 22, 23, 24]. 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 [1, 2, 9, 11, 25, 26]. Most of the related work optimizes SMPL-X based on images of people wearing shoes or socks. For instance, the DINAR [27] method, as well as the PeopleSnapshot [1] and X-Avatar [28] datasets, consist of only people wearing shoes or socks. In other papers (PIFu [6], PIFuHD [7], ICON [8], and PHORHUM [5]) and datasets (Renderpeople [29] and THuman [30]), the proportion of images corresponding to barefoot people is small (less than 3 %). While improving the reconstruction of feet for

barefoot images is another interesting problem, we propose to improve the reconstruction for images of people wearing shoes and socks. With its detailed toes, the SMPL-X model cannot properly fit a shoe shape. Through their deformation vector, Mallek *et al.* [9] deform the toes toward the shape of shoes, but given the mesh topology of the toes, their optimization process often generates artifacts in the feet and toes region. Alternatively, the parametric model's parameters can be directly regressed via a Deep Neural Network (DNN) model [3, 4, 31, 32]. 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 [5, 6, 7, 8, 33]. The primary objective of these PIFu-based methods lies in obtaining an abundance of details, encompassing hair, and clothing. PaMIR [34] uses a DNN-based method to generate an implicit field with features extracted from the input image. ECON [35] generates front and back normals from the input image, which are passed to a DNN-based method to reconstruct the front and back meshes which are then aligned and completed. HumanRef [36], SiTH [37], TeCH [38] and DiffHuman [39] generate front and back normals from the input image, and use DNN-based methods to generate a distance field from which a mesh is later extracted. A significant drawback of the methods presented in this paragraph lies in the inaccurate modeling of small geometric details such as hands and face. These methods often produce lower-quality results in the hands and face 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, which contains many more triangles than a parametric body mesh. Additionally, it is difficult to animate the mesh, and the animation often needs to resort to advanced DNN techniques [40]. 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 [41] leverages DensePose [42] 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. DINAR [27] 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 [11], 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. To get rid of the concerns related to the SMPL-X model, some methods [6, 34] extract a fine-detailed mesh from the input image before generating the textures. Nevertheless, these methods produce blurry and low quality textures. Another group of methods [33, 36, 37, 38, 39] uses diffusion models to generate the textures, but does not use the SMPL-X model, thus resolving some of the concerns faced by DINAR. While the resulting textures are interesting, the avatar is hard to animate with off-the-shelf software.

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, some struggle in reconstructing proper shoe-like feet geometry, 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. Building upon the work of Mallek *et al.* [9], 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, hands, and feet, 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. Our pipeline begins with the extraction of the target mesh,

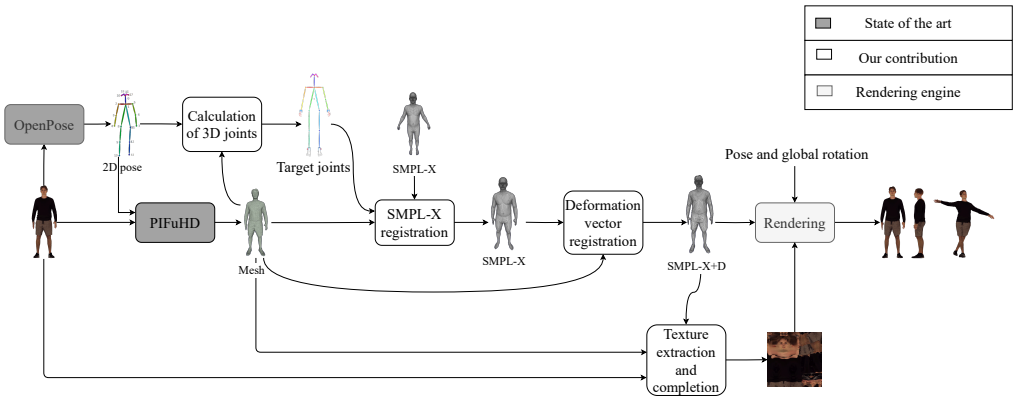


Figure 2: Illustration of the reconstruction and texturing of the SMPL-X+D mesh from a single image, along with rendering results in various poses and viewpoints.

utilizing PIFuHD [7], coupled with the acquisition of 2D pose estimations via OpenPose [10]. We then compute three-dimensional joints. We optimize a global alignment of the SMPL-X model [11], optimizing its translation and rotation parameters, and then further refine 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 $\psi \in \mathbb{R}^{10}$. This mesh has a fixed topology with a constant number of vertices and faces:

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

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:

$$M_{\text{PIFuHD}} = \{V_{\text{PIFuHD}}, F_{\text{PIFuHD}}\} \subseteq \mathbb{R}^{n_2 \times 3}, \mathbb{N}^{m_2 \times 3}. \tag{2}$$

3.2. Pose Estimation

Utilizing OpenPose [10], we extract 2D skeletal data, represented as blue points in Figure 3, which correspond to joints within the image. We project

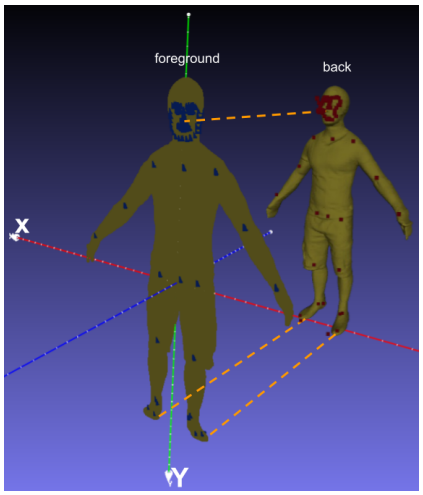


Figure 3: Orthographic projection and 3D pose estimation approach. The back shows the M_{PIFuHD} mesh, while the foreground shows the orthographic projection, M'_p, of this mesh onto the XY plane. The blue points illustrate the 2D joint estimates obtained through OpenPose. The red points correspond to 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.

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 simple process is not

overly sensitive to the precision of the 2D pose estimation algorithm and allow us to obtain a robust initialization to the 3D location of joints without requiring complex constraints normally used in 3D lifting of 2D poses [43]. Furthermore, our approach can take advantage of future pose detectors, as long as they are compatible with the SMPL-X joints.

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, allowing 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:

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

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:

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

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 between two meshes M_A, which will correspond to M_{PIFuHD}, and M_B, which will correspond to $M_{\text{SMPL-X}}$ in this section. We introduce mesh M_B here as in Section 3.3.3 we will use the same loss with the SMPL-X plus deformation vector mesh. The loss computes the distance between vertex pairs projected onto the normal vector of the vertex from mesh M_B. Our loss favors adjustment of the M_B vertices locally and perpendicular to the M_B surface, thus reducing lateral sliding:

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

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

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

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):

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

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 [1, 19], 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:

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

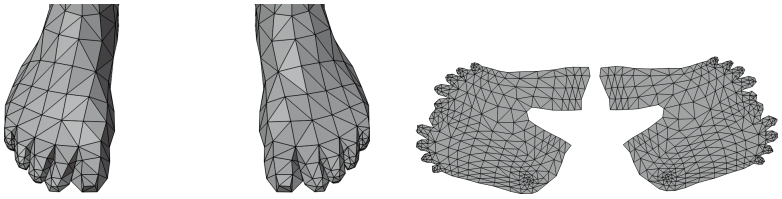
where the two losses are simply added together. This deformation vector optimization greatly improves the clothing representation, capturing the wrinkles and later helping with the texture extraction. Our optimization strategy

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.

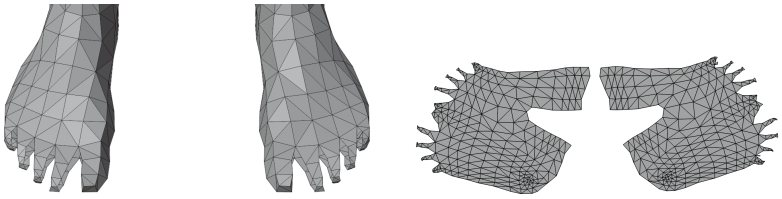
The high-resolution mesh of M_{PIFuHD} results in significant computational time and memory usage during the optimization. Our experiments demonstrated that subsampling M_{PIFuHD} to match the vertex count of the $M_{\text{SMPL-X}}$ mesh, significantly reduces computation time while having a negligible impact on the resulting quality. To achieve a reduction in the number of vertices V_{PIFuHD}, we employed a farthest point sampling method [44]. 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.

Like those in most related work, our reconstructions rely on images of people wearing shoes or socks. The SMPL-X model, with its detailed geometry, including individually articulated toes, is ill-designed for optimization toward a shoe geometry. We thus introduce a modification to the SMPL-X mesh topology in the foot region. We replaced the toe details with a closed surface resembling shoes (Figure 4). We manually selected the faces corresponding to the inner sides of the toes and removed them from both the 3D model and the UV map (Figure 4(b)). We then added new faces to close the 3D mesh and added corresponding faces to the UV map (essential for texture extraction, Section 3.4). Note that, in this process, we rely only on the original SMPL-X vertices. Vertices which were located on the sides of the toes are now unused (and ignored in all optimization steps). As we rely only on the original SMPL-X vertices, we preserve the ability to use the original skinning-based animation without any change. Only the number of faces and vertices is slightly less. By adopting this approach, the modified model retains the general shape of the feet while easing the optimization process. We will release our proposed shoe-like SMPL-X mesh topology to the community.

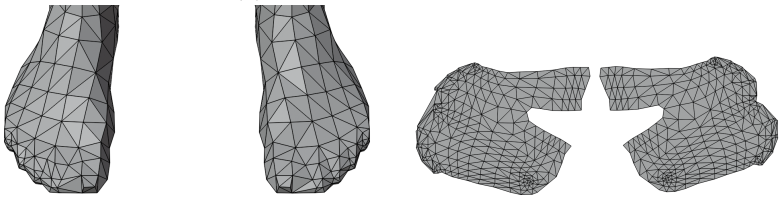
Even though SMPL-X can model most clothing details found in PIFuHD, it sometimes fails to recover finer clothing wrinkles and discontinuities. To



(a) Original SMPL-X



(b) SMPL-X after faces removal



(c) New SMPL-X mesh topology

Figure 4: Topology modification of the SMPL-X model. (a) Original SMPL-X feet mesh with corresponding UV map. (b) Mesh resulting from the toe inner side removal. (c) New shoe-like topology.

overcome this, we create a new mesh, SMPL-X$_{\times 4}$, which is a finer version of SMPL-X (1-to-4 subdivision), as seen in Figure 5, and adapted some aspects of our framework. First, we compute skinning weights for the new vertices and also we subdivide the UV-map mesh. The added vertices are considered in the optimization steps. The point-to-surface loss function (Equations 7 and 8) now uses a regular Euclidean distance between the corresponding vertices $v \in M_{\text{SMPL-X}_{\times 4}}$ and $p \in M_{\text{PIFuHD}}$:

$$\text{dist}(p, v) = \|v - p\|_2. \tag{12}$$

This change is justified by the fact that the resolution of the subdivided mesh is sufficiently high, eliminating the need for projecting vertices with respect to the normal vector. Furthermore, because of the difference in number of vertices, the point to surface loss ($\mathcal{L}_{\text{P2S}}$, Equation 7) behaves in a slightly different way. The increased number of vertices sometimes pulls the vertices of $M_{\text{SMPL-X}_{\times 4}}$ in regions of M_{PIFuHD} showcasing a noisy surface or erroneous protrusions. To avoid pulling the surface too far at each optimization iteration of Equation 7, each optimization iteration ignores mesh M_A (M_{PIFuHD}) vertices for which the distance of Equation 12 is further than 2 cm. Through the successive optimization iterations, the surface deforms more locally and gradually. As such, the surface is smoother. Also, the tuning of loss parameters has been adjusted to better regulate the influence of each component of the loss functions on the overall model training. Specifically, λ_1 was adjusted to 2 to increase its regulative impact, whereas λ_2 was increased to 1e5 and λ_4 to 1e8 (λ_3 remained unchanged).

In our computational framework, the Adam optimizer [45] is consistently utilized across all stages. We conducted a parameter sweep to select good learning rates for each step of our approach (See Tables 5-8 of the Appendix for details). The selected learning rates are as follows: 10^{-3} for the rigid transformation optimization (Section 3.3.1, Equation 3), 10^{-4} for the pose optimization (Section 3.3.1, Equation 6), 10^{-2} for the shape optimization (Section 3.3.2), and 10^{-4} for the deformation vector optimization (Section 3.3.3).

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

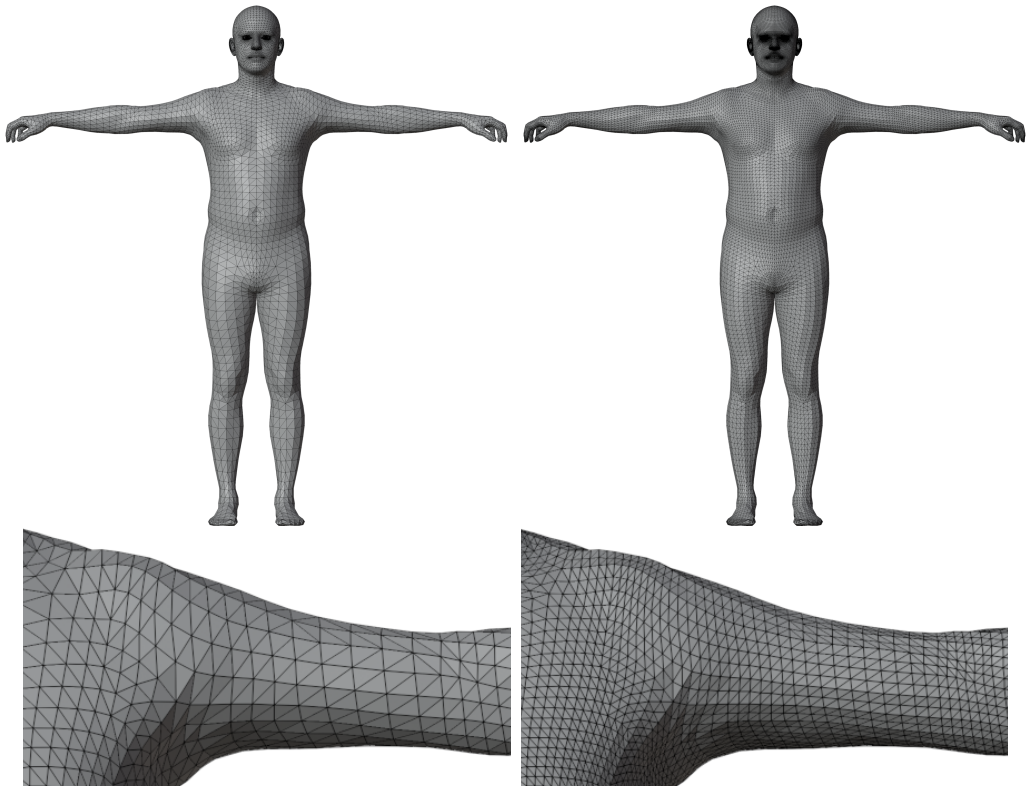


Figure 5: Left: Original SMPL-X mesh. Right: SMPL-$X_{\times 4}$ (1-to-4 subdivision). The bottom images zoom in on the shoulder and arm regions.

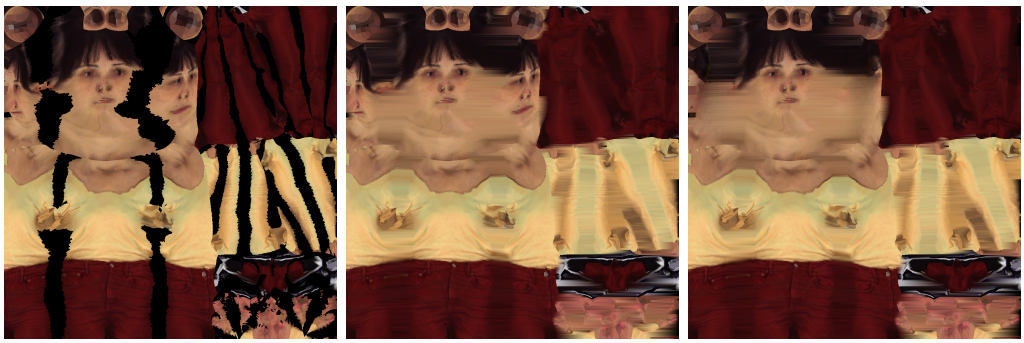
the closest triangle and convert the texel's position to barycentric coordinates within this triangle of the SMPL-X+D mesh. From the corresponding 3D position, we fetch the color from the nearest PIFuHD mesh vertex.

Colors at the silhouette of the PIFuHD mesh exhibit color leakage from the background as can be seeing in Figure 6. To identify these wrong silhou-

(a) (b) (c)

Figure 6: Silhouette color leakage. From left to right - the input image, the PIFuHD mesh, and the resulting texture extracted from PIFuHD.

ette 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 [46] 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 7 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 [47]. 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 7 (c).



(a) (b) (c)

Figure 7: From left to right: Texture extracted from PIFu-HD, 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 [28] 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 [1] 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 [6], PIFuHD [7], ICON [8], PHORHUM [5] and DINAR [27]. This comparison is based on a set of specific metrics. Intersection over Union (IoU) [48] 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) [49] evaluates the similarity between two sets of vertices, with lower values denoting closer matches. Normal Consistency (NC) [50] 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) [51] and Peak Signal-to-Noise Ratio (PSNR) [52] 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) [53] metric evaluates perceptual similarity between images, focusing on high-level visual features significant for human perception, where closer matches yield lower scores.

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. "Ours subdiv" corresponds to using the SMPL-$X_{\times 4}$ mesh.

	3D Metrics		Rendered Normals			Rendered RGB Images				Nbr
Method	CD ↓	NC ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	IoU ↑	vertices ↓
PIFu [6]	1.16	0.808	0.835	0.142	18.54	0.832	0.144	19.90	0.971	50,000
PIFuHD [7]	**0.76**	**0.823**	0.857	**0.089**	**21.62**	**0.912**	0.093	21.55	**0.984**	170,000
PHORHUM [5]	2.48	0.75	0.782	0.216	13.96	0.76	0.192	13.67	0.890	100,000
ICON [8]	2.98	0.721	0.833	0.125	18.48	0.805	0.143	17.89	0.947	48,000
Mallek et al. [9]	0.91	0.803	0.869	0.127	20.75	0.896	0.075	**23.23**	0.974	**10,475**
Ours	0.91	0.805	0.870	0.125	20.82	0.900	0.073	**23.23**	0.976	**10,475**
Ours subdiv	0.84	0.807	**0.876**	0.108	20.74	0.911	**0.066**	23.03	0.979	41,738

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 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 PHORHUM, being specifically trained on perspective data, has a weaker performance on our orthographic projection setting. To reduce the misalignment between the source and the reprojected images, we have applied minor translation and scale adjustments before computing the quality metrics to allow for a fairer comparison. 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. Finally, we can note that our approach, with and without subdivision, outperforms the method of Mallek et al. [9] with the only exception of PSNR on RGB images. When considering our approach without subdivision, the numerical differences to the method of Mallek et al. [9] are smaller, mainly because our contribution in the foot modeling results in relatively few pixels in the overall image. Note that our optimization approach being non-deterministic, the optimized avatars slightly differ every time the optimization is computed. The results in Table 1 for our approach correspond to the median value over 10 optimizations.

We evaluated our approach alongside DINAR on the PeopleSnapshot dataset, and the results are shown in Table 2. This dataset consists of real-world perspective images, which correspond to the training environment of DINAR. Additionally, since DINAR's rendered outputs do not perfectly align with the input images, we further applied cropping and scaling adjustments to ensure a fairer comparison. Despite these conditions, our method consistently achieves higher fidelity and segmentation quality, demonstrating robustness across both orthographic and perspective-based scenarios.

Table 2: Quantitative comparison on the PeopleSnapshot dataset using rendered RGB image metrics. Our method achieves higher fidelity and segmentation quality than DINAR.

	Rendered RGB Images			IoU
Method	SSIM ↑	LPIPS ↓	PSNR ↑	IoU ↑
DINAR [27]	0.947	0.070	26.88	0.871
Ours subdiv	**0.985**	**0.029**	**33.55**	**0.955**

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 8 and Figure 9, as well as a comparison on real images in Figure 10. Figure 8 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

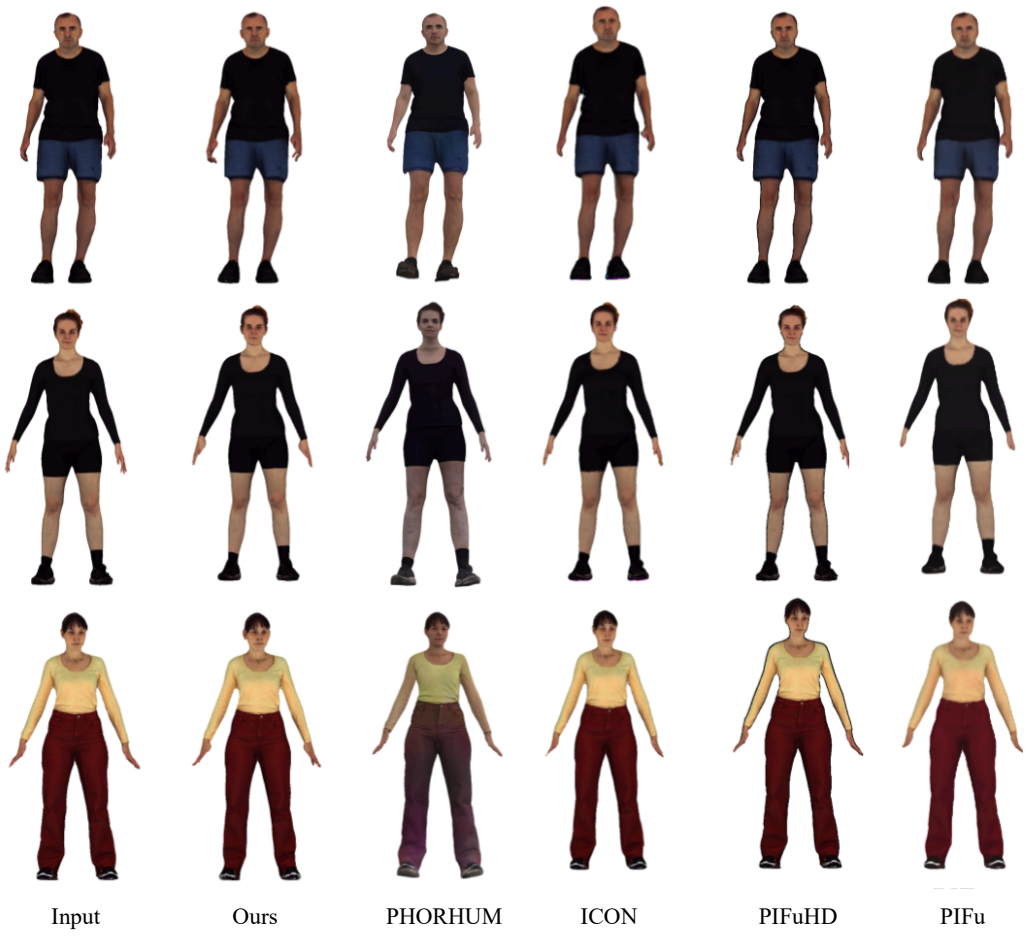


Figure 8: Qualitative evaluation of X-Avatar samples (same as input view).

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. Figure 11 illustrates a comparison of foot reconstruction, highlighting the differences between the SMPL-X foot topology used in the method of Mallek et al. [9] and our shoe-like topology. Figure 12 highlights the finer details on the clothes that are recovered when using the proposed subdivided mesh SMPL-$X_{\times 4}$. While this subdivision strategy is optional, Table 1 shows that it improves the quantitative metric results, increasing the advantage of our proposed approach compared to the method of Mallek et al. [9].

We then assess the performance of our approach in generating rendered images from new perspectives with the X-Avatar (Figure 9) and PeopleSnapshot (Figure 10) 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 13 presents three animations generated from the extensive AMASS dataset of human motions [54] showcasing the versatility of our approach.

Animation 1 (Figure 13a) features a series of dance poses. Animation 2 (Figure 13b) depicts an avatar executing gymnastic poses. Animation 3 (Figure 13c) 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 3 allows us to isolate and understand the impact of each component on the overall performance. The

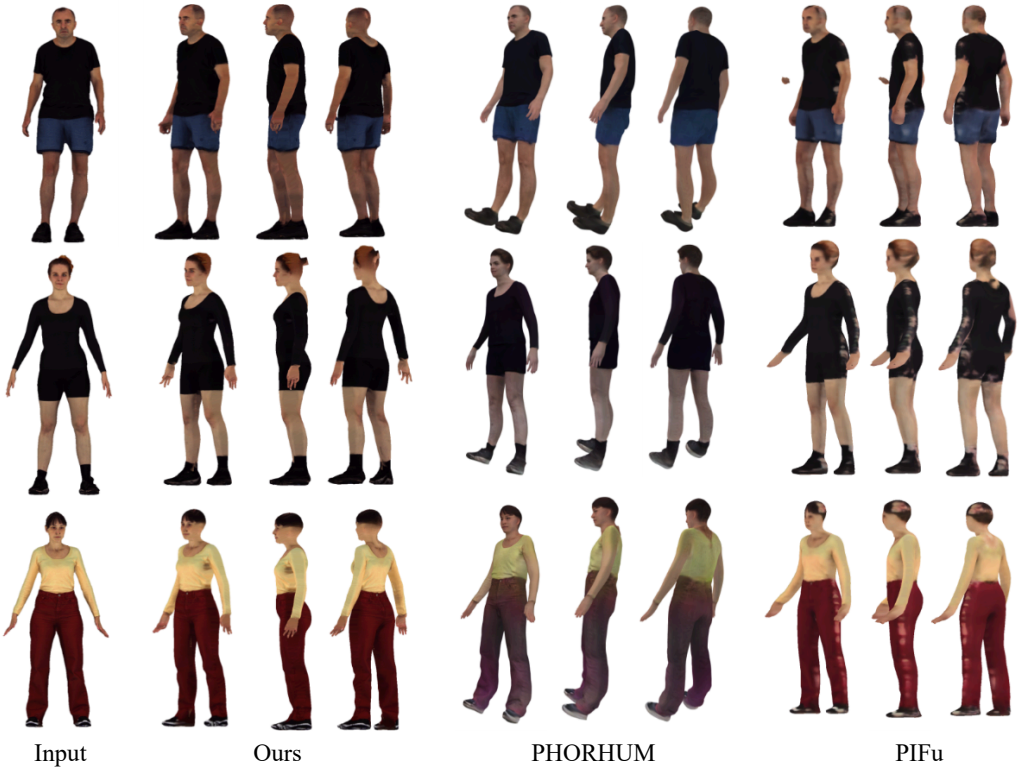


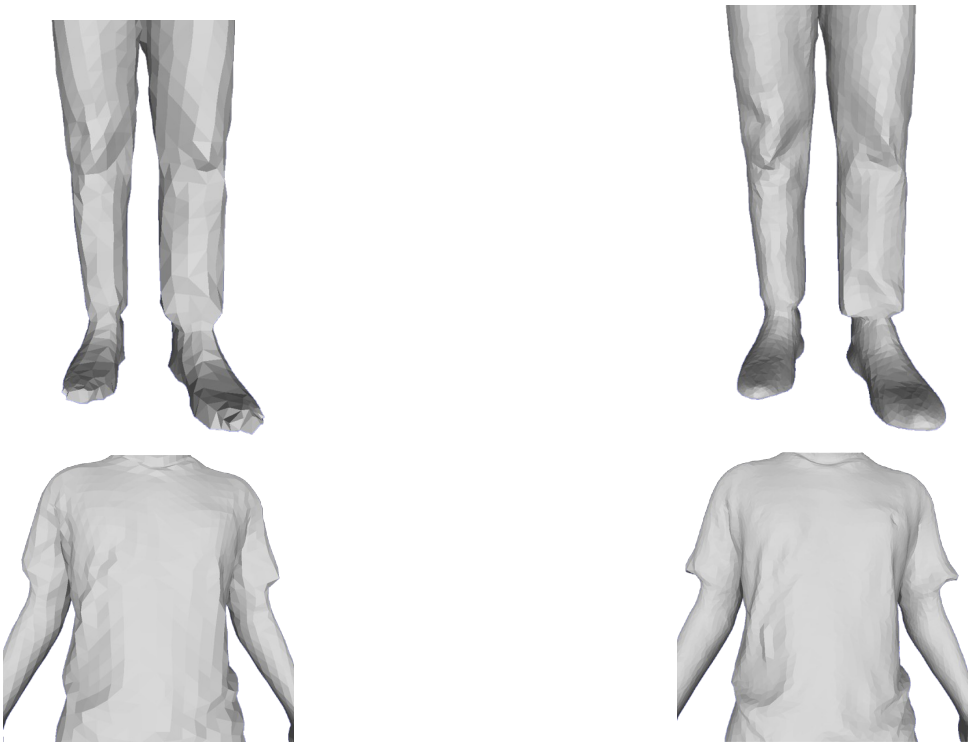
Figure 9: Qualitative evaluation of X-Avatar samples across varied perspectives, distinct from the initial view

Figure 10: Qualitative evaluation of PeopleSnapshot samples

(a) SMPL-X foot topology (b) Our foot topology

Figure 11: Comparison of foot reconstruction.

last row (Ours) shows that our full pipeline has the best and second best values for five out of nine measures, demonstrating that it outperforms most of the other configurations. Rows labelled "w/o P2S" in Table 3 and the column labelled "w/o P2S" in Figure 14 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 3 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 14 "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 14 is quite degraded compared to our full pipeline. According to Table 3, 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 14), with flipped and intersecting triangles on the body and hands.



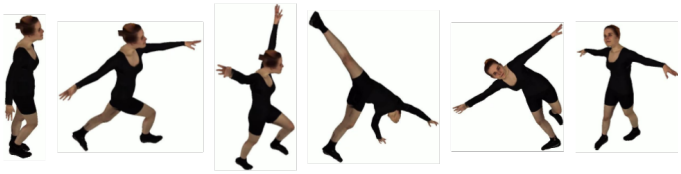
(a) w/o subdiv (SMPL-X) (b) with subdiv (SMPL-X$_{\times 4}$)

Figure 12: Comparison without and with mesh subdivision.

Table 3: Comparison with respect to the ablated components. Best results highlighted in **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	0.910	**0.803**	**0.869**	0.127	20.75	0.896	**0.075**	23.23	0.974

(a) Animation 1



(b) Animation 2

(c) Animation 3

Figure 13: Presentation of three rendered animations featuring three subjects in diverse body poses and expressions

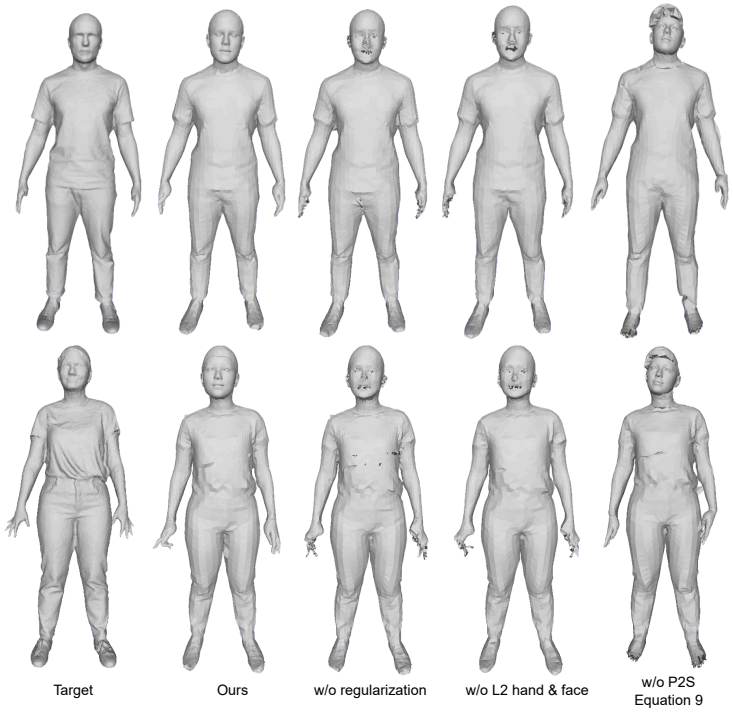


Figure 14: Qualitative ablation

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. Figure 15 presents the relative computation times of a representative example. We can see that with the regular SMPL-X mesh, most of the time goes toward texture extraction, while with the subdivided SMPL-X mesh, most of the computation time goes toward computing the deformation vectors. We used a computer with 2 cores at 2.2 GHz, 24 GB of memory and an NVidia L4 GPU.

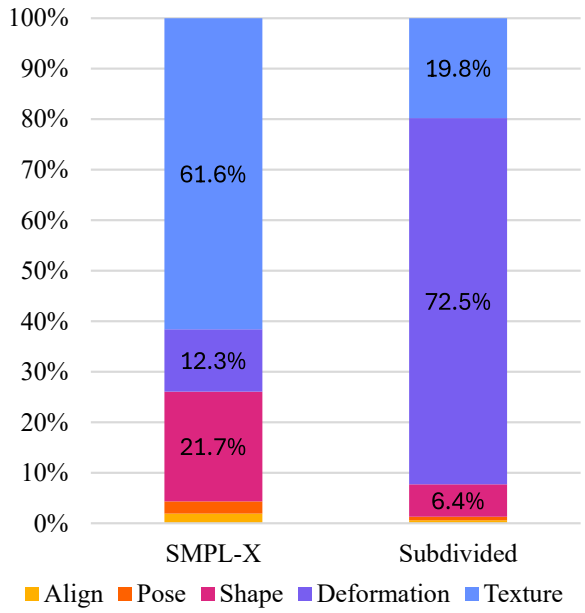


Figure 15: Relative computation times for the various steps of our approach, Align (Equation 3), Pose (Section 3.3.1), Shape (Section 3.3.2), Deformation (Section 3.3.3), and Texture (Section 3.4), as well as for the original vs. subdivided SMPL-X mesh.

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 4 highlights the distinctions between our approach and other methods.

Table 4: Comparison according to several criteria

Method	Single Image Input	Shape Variability	Animation	Expression	Textured	Compact Representation
SMPLify-X	✓	✗	✓	✓	✗	✓
Video Avatar	✗	✓	✓	✗	✓	✓
PIFuHD	✓	✓	✗	✗	✗	✗
PHORHUM	✓	✓	✗	✗	✓	✗
Ours	✓	✓	✓	✓	✓	✓

Our approach, while using a mesh with fewer vertices compared to PIFu, PIFuHD, ICON, and PHORHUM ($\approx$ 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 extends the work of Mallek *et al.* [9]. We are 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. Most often, avatars are reconstructed from images of people wearing shoes or socks. Thus, we modified the SMPL-X mesh topology to reflect that. Our

modification maintains the same ease of use and animation of SMPL-X as we kept the exact same vertices and only changed the mesh topology. Furthermore, we demonstrate how to adapt our approach to a finer resolution SMPL-X mesh. We also showed that this subdivision strategy improves the quantitative metrics. 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 generation requires further refinement for enhanced fidelity. Investigating the use of diffusion-based models [33, 36, 37, 38, 39] has the potential to better capture the back side of the avatar. While our approach is successful regarding certain types of loose clothing, it does not yet support very loose garments, like skirts. Investigating other methods [34, 35, 37, 38, 39] which successfully support loose garments could help in rethinking our use of the SMPL-X mesh to allow for different garment topologies while preserving the ability to easily animate the resulting avatar. 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 generative 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.

Appendix

Tables 5-8 showcase the parameter sweep we conducted to select the best learning rates for the different steps of our approach.

Table 5: Parameter sweep with respect to the learning rate for the rigid optimization phase (Section 3.3.1, Equation 3). Best results are highlighted in bold **green** and second-best in amber.

	3D Metrics		Rendered Normals			Rendered RGB Images			IoU ↑
Learning rate	CD ↓	NC ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	
1e-2	**0.828**	0.806	0.877	0.110	20.68	**0.911**	0.068	22.88	**0.978**
1e-3	0.850	0.806	**0.878**	**0.109**	**20.76**	**0.911**	0.068	**22.97**	**0.978**
1e-4	0.850	**0.808**	0.877	0.110	20.66	0.910	**0.066**	22.77	**0.978**

Table 6: Parameter sweep with respect to the learning rate for the pose optimization phase (Section 3.3.1, Equation 6). Best results are highlighted in bold **green** and second-best in amber.

	3D Metrics		Rendered Normals			Rendered RGB Images			IoU ↑
Learning rate	CD ↓	NC ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	
1e-3	**0.850**	**0.807**	0.877	**0.109**	20.71	**0.911**	**0.067**	22.87	**0.978**
1e-4	**0.850**	0.806	**0.878**	**0.109**	**20.76**	**0.911**	0.068	**22.97**	**0.978**
1e-5	**0.850**	**0.807**	0.877	0.110	20.65	0.910	0.068	22.83	**0.978**

Table 7: Parameter sweep with respect to the learning rate for the shape optimization phase (Section 3.3.2). Best results are highlighted in bold **green** and second-best in amber.

	3D Metrics		Rendered Normals			Rendered RGB Images			IoU ↑
Learning rate	CD ↓	NC ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	
1e-1	0.87	0.800	0.871	0.117	20.37	0.905	0.073	22.74	0.973
1e-2	**0.85**	0.806	**0.878**	**0.109**	**20.76**	**0.911**	0.068	**22.97**	**0.978**
1e-3	0.88	0.798	0.871	0.116	20.35	0.905	0.073	22.75	0.973

References

[1] T. Alldieck, M. Magnor, W. Xu, C. Theobalt, G. Pons-Moll, Video based reconstruction of 3D people models, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8387–8397.

Table 8: Parameter sweep with respect to the learning rate for the deformation vector optimization phase (Section 3.3.3). Best results are highlighted in bold **green** and second-best in amber.

	3D Metrics		Rendered Normals			Rendered RGB Images			IoU ↑
Learning rate	CD ↓	NC ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	
1e-3	0.83	0.794	0.856	0.143	18.42	0.897	0.092	19.23	0.966
1e-4	**0.85**	**0.806**	**0.878**	**0.109**	**20.76**	**0.911**	0.068	**22.97**	**0.978**
1e-5	**0.85**	0.805	0.876	0.111	20.66	0.90	0.069	22.90	0.977

[2] J. Ajanohoun, E. Paquette, C. Vázquez, Multi-view human model fitting using bone orientation constraint and joints triangulation, in: 2021 IEEE International Conference on Image Processing (ICIP), 2021, pp. 1094–1098.

[3] Y. Feng, V. Choutas, T. Bolkart, D. Tzionas, M. J. Black, Collaborative regression of expressive bodies using moderation, in: 2021 International Conference on 3D Vision (3DV), IEEE, 2021, pp. 792–804.

[4] H. Zhang, Y. Tian, X. Zhou, W. Ouyang, Y. Liu, L. Wang, Z. Sun, PyMAF: 3D human pose and shape regression with pyramidal mesh alignment feedback loop, IEEE/CVF International Conference on Computer Vision (ICCV) (2021) 11426–11436.

[5] T. Alldieck, M. Zanfir, C. Sminchisescu, Photorealistic monocular 3D reconstruction of humans wearing clothing, IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022) 1496–1505.

[6] S. Saito, Z. Huang, R. Natsume, S. Morishima, A. Kanazawa, H. Li, PIFu: Pixel-aligned implicit function for high-resolution clothed human digitization, in: IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2304–2314.

[7] S. Saito, T. Simon, J. Saragih, H. Joo, PIFuHD: Multi-level pixel-aligned implicit function for high-resolution 3D human digitization, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 84–93.

[8] Y. Xiu, J. Yang, D. Tzionas, M. J. Black, ICON: Implicit clothed humans obtained from normals, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2022, pp. 13286–13296.

[9] F. Mallek, C. Vázquez, E. Paquette, Implicit and parametric avatar pose and shape estimation from a single frontal image of a clothed human, in: Proceedings of the 17th ACM SIGGRAPH Conference on Motion, Interaction and Games (MIG), 2024, pp. 1–11.

[10] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, Y. A. Sheikh, OpenPose: Realtime multi-person 2D pose estimation using part affinity fields, IEEE Transactions on Pattern Analysis and Machine Intelligence (2019) 14.

[11] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. Osman, D. Tzionas, M. J. Black, Expressive body capture: 3D hands, face, and body from a single image, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10975–10985.

[12] M. A. Fischler, R. A. Elschlager, The representation and matching of pictorial structures, IEEE Transactions on computers 100 (1) (1973) 67–92.

[13] A. Toshev, C. Szegedy, Deeppose: Human pose estimation via deep neural networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 1653–1660.

[14] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, J. Davis, SCAPE: Shape completion and animation of people, ACM Trans. Graph. 24 (3) (2005) 408–416.

[15] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, M. J. Black, SMPL: A skinned multi-person linear model, ACM Trans. Graph. 34 (6) (oct 2015).

[16] B. Xu, J. Zhang, K.-Y. Lin, C. Qian, Y. He, 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), 2023, pp. 17878–17888.

[17] J. Mu, S. Sang, N. Vasconcelos, X. Wang, ActorsNeRF: Animatable few-shot human rendering with generalizable NeRFs, in: 2023 IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 18345–18355.

[18] T. Jiang, X. Chen, J. Song, O. Hilliges, InstantAvatar: Learning avatars from monocular video in 60 seconds, in: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16922–16932.

[19] T. Alldieck, M. Magnor, W. Xu, C. Theobalt, G. Pons-Moll, Detailed human avatars from monocular video, in: 2018 International Conference on 3D Vision (3DV), IEEE, 2018, pp. 98–109.

[20] C. Guo, T. Jiang, X. Chen, J. Song, O. Hilliges, 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.

[21] H. Song, B. Yoon, W. Cho, W. Woo, RC-SMPL: Real-time cumulative smpl-based avatar body generation, in: 2023 IEEE International Symposium on Mixed and Augmented Reality (ISMAR), IEEE Computer Society, Los Alamitos, CA, USA, 2023, pp. 89–98.

[22] Y. Lu, H. Yu, W. Ni, L. Song, 3D real-time human reconstruction with a single RGBD camera, Applied Intelligence 53 (8) (2022) 8735–8745.

[23] R. Zheng, P. Li, H. Wang, T. Yu, Learning visibility field for detailed 3D human reconstruction and relighting, 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023) 216–226.

[24] S. Wang, M. Mihajlovic, Q. Ma, A. Geiger, S. Tang, Metaavatar: Learning animatable clothed human models from few depth images, in: Advances in Neural Information Processing Systems, 2021, pp. 2810–2822.

[25] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, M. J. Black, Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image, in: European Conference on Computer Vision (ECCV), Springer, 2016, pp. 561–578.

[26] N. Kolotouros, G. Pavlakos, M. J. Black, K. Daniilidis, 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.

[27] D. Svitov, D. Gudkov, R. Bashirov, V. Lempitsky, DINAR: Diffusion inpainting of neural textures for one-shot human avatars, in: 2023 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE, 2023, pp. 7039–7049.

[28] K. Shen, C. Guo, M. Kaufmann, J. J. Zarate, J. Valentin, J. Song, O. Hilliges, X-avatar: Expressive human avatars, IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023) 16911–16921.

[29] RenderPeople, `renderpeople.com` (2025).

[30] Z. Zheng, T. Yu, Y. Wei, Q. Dai, Y. Liu, DeepHuman: 3D Human Reconstruction From a Single Image, in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE, Seoul, Korea (South), 2019, pp. 7738–7748.

[31] A. Kanazawa, M. J. Black, D. W. Jacobs, J. Malik, End-to-end recovery of human shape and pose, IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2017) 7122–7131.

[32] N. Kolotouros, G. Pavlakos, K. Daniilidis, Convolutional mesh regression for single-image human shape reconstruction, IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 4496–4505.

[33] B. Albahar, S. Saito, H.-Y. Tseng, C. Kim, J. Kopf, J.-B. Huang, Single-Image 3D Human Digitization with Shape-guided Diffusion, in: SIGGRAPH Asia 2023 Conference Papers, SA '23, Association for Computing Machinery, New York, NY, USA, 2023, pp. 1–11.

[34] Z. Zheng, T. Yu, Y. Liu, Q. Dai, Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction, IEEE Transactions on Pattern Analysis and Machine Intelligence 44 (06) (2022) 3170–3184.

[35] Y. Xiu, J. Yang, X. Cao, D. Tzionas, M. J. Black, ECON: Explicit Clothed humans Optimized via Normal integration, in: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Vancouver, BC, Canada, 2023, pp. 512–523.

[36] J. Zhang, X. Li, Q. Zhang, Y. Cao, Y. Shan, J. Liao, HumanRef: Single Image to 3D Human Generation via Reference-Guided Diffusion, in: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Seattle, WA, USA, 2024, pp. 1844–1854.

[37] H.-I. Ho, J. Song, O. Hilliges, SiTH: Single-view Textured Human Reconstruction with Image-Conditioned Diffusion, in: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Seattle, WA, USA, 2024, pp. 538–549.

[38] Y. Huang, H. Yi, Y. Xiu, T. Liao, J. Tang, D. Cai, J. Thies, TeCH: Text-Guided Reconstruction of Lifelike Clothed Humans, in: 2024 International Conference on 3D Vision (3DV), 2024, pp. 1531–1542, iSSN: 2475-7888.

[39] A. Sengupta, T. A. Andrei Zanfir, C. Cris, N. Kolotouros, E. Corona, A. Zanfir, C. Sminchisescu, DiffHuman: Probabilistic Photorealistic 3D Reconstruction of Humans, in: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Seattle, WA, USA, 2024, pp. 1439–1449.

[40] S. Saito, J. Yang, Q. Ma, M. J. Black, Scanimate: Weakly supervised learning of skinned clothed avatar networks, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2886–2897.

[41] B. AlBahar, J. Lu, J. Yang, Z. Shu, E. Shechtman, J.-B. Huang, Pose with style: Detail-preserving pose-guided image synthesis with conditional stylegan, ACM Transactions on Graphics (TOG) 40 (6) (2021) 1–11.

[42] R. A. Güler, N. Neverova, I. Kokkinos, Densepose: Dense human pose estimation in the wild, 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018) 7297–7306.

[43] S. Mehraban, Y. Qin, B. Taati, Evaluating recent 2d human pose estimators for 2d-3d pose lifting, 2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition, FG 2024 (2024). doi:10.1109/FG59268.2024.10581948.

[44] L. Ge, Y. Cai, J. Weng, J. Yuan, Hand pointnet: 3D hand pose estimation using point sets, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8417–8426.

[45] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, CoRR abs/1412.6980 (2014).

[46] D. Gatis, Rembg: Rembg is a tool to remove images background, `https://github.com/danielgatis/rembg`, accessed: 2024-03 (2023).

[47] R. Suvorov, E. Logacheva, A. Mashikhin, A. Remizova, A. Ashukha, A. Silvestrov, N. Kong, H. Goka, K. Park, V. S. Lempitsky, Resolution-robust large mask inpainting with fourier convolutions, 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (2021) 3172–3182.

[48] M. Harouni, H. Y. Baghmaleki, Color image segmentation metrics, in: P. A. Laplante (Ed.), Encyclopedia of Image Processing, Vol. 95, CRC Press, 2018, pp. 10–21.

[49] H. G. Barrow, J. M. Tenenbaum, R. C. Bolles, H. C. Wolf, Parametric correspondence and chamfer matching: Two new techniques for image matching, in: Proceedings: Image Understanding Workshop, Science Applications, Inc, 1977, pp. 21–27.

[50] L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, A. Geiger, Occupancy networks: Learning 3D reconstruction in function space, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4455–4465. doi:10.1109/CVPR.2019.00459.

[51] Z. Wang, L. Lu, A. C. Bovik, Video quality assessment based on structural distortion measurement, Signal Processing: Image Communication 19 (2004) 121–132.

[52] A. Horé, D. Ziou, Image quality metrics: PSNR vs. SSIM, in: 2010 20th International Conference on Pattern Recognition, 2010, pp. 2366–2369.

[53] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, O. Wang, The unreasonable effectiveness of deep features as a perceptual metric, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 586–595.

[54] N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, M. J. Black, Amass: Archive of motion capture as surface shapes, in: IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 5442–5451.