DiffHuman: Probabilistic Photorealistic 3D Reconstruction of Humans

Akash Sengupta[1,2*] Thiemo Alldieck[1] Nikos Kolotouros[1] Enric Corona[1]
Andrei Zanfir[1] Cristian Sminchisescu[1]
[1]Google Research [2]University of Cambridge

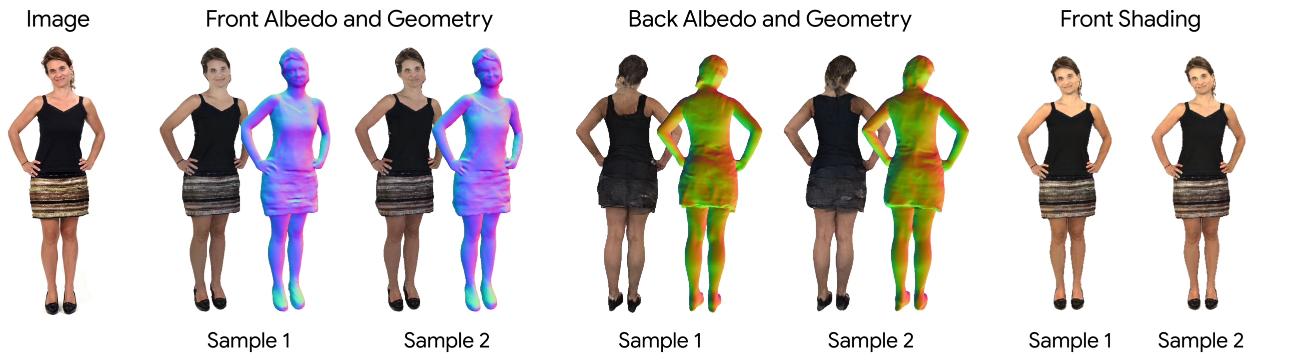

Figure 1. DiffHuman predicts a probability distribution over 3D human reconstructions conditioned on a single monocular RGB image. This enables us to sample multiple plausible, diverse and input-consistent reconstructions during inference. Samples from DiffHuman demonstrate a high level of geometric and colour-wise detail, particularly in unseen and uncertain regions of the human body surface.

Abstract

We present DiffHuman, a probabilistic method for photorealistic 3D human reconstruction from a single RGB image. Despite the ill-posed nature of this problem, most methods are deterministic and output a single solution, often resulting in a lack of geometric detail and blurriness in unseen or uncertain regions. In contrast, DiffHuman predicts a ***probability distribution*** *over 3D reconstructions conditioned on an input 2D image, which allows us to sample multiple detailed 3D avatars that are consistent with the image. DiffHuman is implemented as a conditional diffusion model that denoises pixel-aligned 2D observations of an underlying 3D shape representation. During inference, we may sample 3D avatars by iteratively denoising 2D renders of the predicted 3D representation. Furthermore, we introduce a generator neural network that approximates rendering with considerably reduced runtime (55× speed up), resulting in a novel dual-branch diffusion framework. Our experiments show that DiffHuman can produce diverse and detailed reconstructions for the parts of the person that are unseen or uncertain in the input image, while remaining competitive with the state-of-the-art when reconstructing visible surfaces.*

1. Introduction

Photorealistic 3D reconstruction of humans from a single image is a central problem for a wide range of applications. Avatar creation for virtual and mixed reality, games, movie production, or fitness and health applications all benefit from reliable and easy-to-use 3D human reconstruction. However, monocular 3D reconstruction is ill-posed: depth-ambiguities, (self)-occlusion, and unobserved body parts make it infeasible to reconstruct the true, veridical 3D shape and appearance of the subject. In fact, there exist an infinite number of 3D scenes that could have produced a given image; although not all of them would represent plausible human and clothing geometry, and realistic interplay between physical albedo and lighting. Yet, existing methods [5, 10, 50, 63] still treat the problem as a one-to-one mapping and return just *one* plausible 3D solution. Simply assuming that this single solution is correct can lead to failures in downstream usages of the 3D reconstruction. Moreover, deterministic methods often produce detail-less reconstructions of unobserved or uncertain surface regions, *e.g*. the back of a person. This is a well-known effect of applying deterministic training losses to ill-posed learning problems [8, 11, 37], which causes predictions to fall back towards the mean of the underlying training distribution when faced with ambiguity. The mean may not have high probability in a multi-modal distribution and often represents blurry and over-smooth 3D reconstructions.

In this work, we overcome the shortcomings of deterministic methods by predicting a *distribution* over possible 3D human reconstructions. Our method **DiffHuman** uses a single input image to condition a denoising diffusion model [18], which generates pixel-aligned front and back observa-

*Work done as an intern at Google Research.

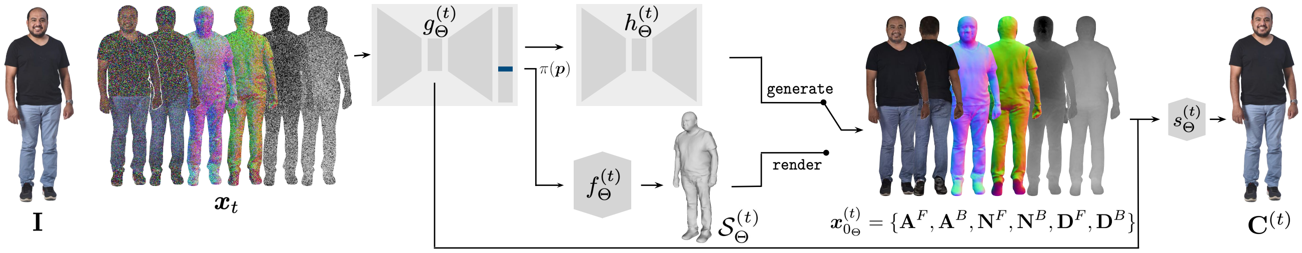

Figure 2. **Method overview**. We use a diffusion probabilistic model [18] to predict a distribution over plausible 3D reconstructions conditioned on a single RGB image. During training, we predict noise-dependent pixel-aligned features $g_\Theta^{(t)}(\boldsymbol{x}_t, \mathbf{I})$ given a noisy observation set $\boldsymbol{x}_t$ consisting of front/back albedo, depth and normal renders, and an RGB image $\mathbf{I}$. These features condition an SDF $f_\Theta^{(t)}$, which is dependent on both $\boldsymbol{x}_t$ and $\mathbf{I}$. $f_\Theta^{(t)}$ and $g_\Theta^{(t)}$ are neural networks that define an implicit surface $\mathcal{S}_\Theta^{(t)}(\boldsymbol{x}_\mathbf{t}, \mathbf{I})$. Then, we obtain an estimate of the denoised observation set $\boldsymbol{x}_{0_\Theta}^{(t)}$ by rendering $\mathcal{S}_\Theta^{(t)}$. We may additionally produce a shaded image $\mathbf{C}^{(t)}$ by applying a pixel-wise noise-dependent shading network $s_\Theta^{(t)}$. During inference, we can sample trajectories over observation sets $\boldsymbol{x}_{0:T} \sim p_\Theta(\boldsymbol{x}_{0:T}|\mathbf{I})$ by computing and rendering $\mathcal{S}_\Theta^{(t)}(\boldsymbol{x}_\mathbf{t}, \mathbf{I})$ in each denoising step. Our final 3D samples $\mathcal{S} \sim p_\Theta(\mathcal{S}|\mathbf{I})$ are obtained as the final reconstruction $\mathcal{S} = \mathcal{S}_\Theta^{(1)}(\boldsymbol{x}_1, \mathbf{I})$. To mitigate the computational cost of rendering an implicit surface in every step, we train a "generator" network $h_\Theta^{(t)}$ that imitates rendering by directly mapping $g_\Theta^{(t)}(\boldsymbol{x}_t, \mathbf{I})$ to $\boldsymbol{x}_{0_\Theta}^{(t)}$. During inference, we denoise using $h_\Theta^{(t)}$ and only explicitly compute the 3D reconstruction in the last step.

tions of the underlying 3D human. To enable full 3D reconstruction, we take inspiration from recent work [57, 59] and integrate rendering of an intermediate implicit 3D representation into the model's denoising step. This allows us, at test time, to reconstruct 3D meshes from a signed distance and colour field defined by this same intermediate representation. However, diffusion-via-rendering is notoriously slow. Thus, we develop a hybrid solution which replaces the expensive implicit surface rendering with a single forward pass through an additional generator network, resulting in a $55\times$ speed up at test time. Our probabilistic approach enables us to sample multiple input-consistent reconstructions and visualise prediction uncertainty, while significantly improving the quality of unseen surfaces. In summary, our contributions are:

- We present a probabilistic diffusion model for photorealistic 3D human reconstruction that predicts a distribution of plausible reconstructions conditioned on an input image.
- We propose a novel dual-branch framework that utilises an image generation network, alleviating the need for expensive implicit surface rendering at every denoising step.
- We show that our model produces 3D reconstructions with greater levels of geometric detail and colour sharpness in uncertain regions than the current state-of-the-art.

2. Related Work

We give an overview of related work on photorealistic and probabilistic 3D human reconstruction.

Photorealistic 3D human reconstruction. Methods for human reconstruction in 3D can be broadly categorised into three classes: mesh-based, implicit, and NeRF-based.

Several methods [9, 14, 25, 27, 28, 42, 44, 51, 58, 69, 70, 72] attempt to reconstruct humans in 3D by leveraging parametric body models [6, 34, 65]. However these approaches only reconstruct the body geometry under clothing and do not predict texture. Others focus on learning deformations on top of parametric models to model hair and clothing [1–4, 30, 62, 75]. These methods are generally fast to render, but have the drawback of working with low resolution meshes that cannot capture fine geometric and texture details, and more importantly cannot handle garments with topologies that deviate from the body, such as skirts or dresses.

Implicit methods model 3D surface geometry as the level-set of a signed-distance function [43] or occupancy field [38]. They are able to model surfaces of arbitrary topology, which makes them suitable for representing clothed humans. PIFu [49] and PIFuHD [50] predict occupancy and colour fields directly from an input image using pixel-aligned features. Geo-PIFu [16] and PaMIR [74] use a combination of pixel-aligned features and sampled features from a voxel grid to mitigate depth ambiguity issues. PHORHUM [5] replaces the occupancy field with a signed distance function and decouples albedo and shading. ARCH [22], ARCH++ [17], and S3F [10] leverage a human body prior and reconstruct animatable avatars. ICON [63] only predicts surface geometry, using a normal refinement procedure that alternates between normal prediction and body pose refinement. ECON [64] independently reconstructs front and back surfaces, which are then fused using a body model prior. TECH [21] is a concurrent optimisation-based method that uses guidance from a text-to-image diffusion model to reconstruct invisible surfaces. DiffuStereo [55] reconstructs detailed 3D human geometry using a multi-view stereo setup, whereas POSE-Fusion [32] uses a single RGB-D camera. D-IF [67] models uncertainty in occupancy field predictions based on the distance of a point from the surface. In contrast, our method learns a distribution over plausible implicit surfaces for a given image from which we can sample at test time.

Another line of work for photorealistic 3D human reconstruction uses Neural Radiance Fields (NeRFs) [40] as the underlying representation. However, these often require multi-view setups or long videos to train [23, 24, 33, 61, 66]. Towards the task of monocular reconstruction, SHERF [19] and ELICIT [20] learn animatable NeRFs from a single image. They are both driven by an underlying body model.

Probabilistic 3D human reconstruction. Several methods estimate distributions over 3D poses conditioned on an input image. For example, [31] uses mixture density networks to estimate a distribution over 3D keypoints conditioned on observed 2D keypoint locations. [60] replaces mixture density networks with normalising flows. More recent methods, such as [13] and [54], employ diffusion models for learning a distribution over 3D poses. 3D Multibodies [7] predicts a categorical distribution over SMPL [34] parameter hypotheses conditioned on an input image, while ProHMR [29] utilises conditional normalising flows to this end. Sengupta *et al*. [52] output hierarchical matrix-Fisher distributions that exploit the SMPL kinematic tree. HuManiFlow [53] predicts normalising flow distributions over ancestor-conditioned joint rotations, which respect the structure of the 3D rotation group $SO(3)$. All these methods, however, predict distributions over sparse 3D landmarks, joint rotations or body model parameters. In contrast, our method learns a much more expressive distribution over detailed surfaces corresponding to clothed human geometry and appearance.

3D diffusion models. The success of diffusion models for 2D image synthesis [46, 48] has motivated a few methods that apply these to 3D generation. A pertinent challenge in this task is the choice of an appropriate 3D representation. [36, 71] implement diffusion models for 3D point cloud generation. DiffRF [41] learns a diffusion model for generating volumetric radiance fields, but denoising 3D voxel grids is computationally expensive. HyperDiffusion [12] presents a method for 3D shape generation that performs diffusion in the weight space of occupancy networks. However, this requires offline fitting of an occupancy field to every training example. [57, 59] integrate rendering of an intermediate 3D representation into the denoising step of a 2D diffusion model, which enables 3D sampling during inference. Our method is similar, but we mitigate the cost of diffusion-via-rendering using a 2D generator neural network.

3. Method

This section details our method for predicting diffusion-based distributions over implicit surfaces representing human geometry and appearance. We begin with an overview of denoising diffusion models and implicit surfaces.

3.1. Background

Denoising Diffusion Probabilistic Models. (DDPMs) [18] are generative models that learn to sample from a target data distribution $q(\boldsymbol{x}_0)$ via a learned iterative denoising process. A forward Markov chain $q(\boldsymbol{x}_{0:T})$ progressively adds Gaussian noise to data samples $\boldsymbol{x}_0 \sim q(\boldsymbol{x}_0)$ such that

$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{1-\beta_t}\boldsymbol{x}_{t-1}, \beta_t \boldsymbol{I}), \qquad (1)$$

where $\beta_t \in (0, 1)$ represents the noise variance at a given timestep $t \in \{1, \dots, T\}$. The distribution $q(\boldsymbol{x}_t|\boldsymbol{x}_0)$ can be derived in closed form from Eq. (1). For sufficiently large T, the marginal $q(\boldsymbol{x}_T)$ approaches a standard normal distribution. A DDPM approximates the reverse Markov chain, iteratively transforming samples from a latent distribution $p(\boldsymbol{x}_T) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ onto the data manifold by following

$$p_\theta(\boldsymbol{x}_{0:T}) = p(\boldsymbol{x}_T) \prod_{t=1}^{T} p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t). \qquad (2)$$

The reverse transition kernels are defined as

$$p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) = \mathcal{N}(\boldsymbol{x}_{t-1}; \boldsymbol{\mu}_\theta^{(t)}(\boldsymbol{x}_t), \boldsymbol{\Sigma}_\theta^{(t)}(\boldsymbol{x}_t)), \qquad (3)$$

where the distribution parameters are typically predicted by a time-dependent neural network with weights θ. This network is trained to maximise a variational lower bound (VLB) on the log-likelihood $\mathbb{E}_{q(\boldsymbol{x}_0)}[\log p_\theta(\boldsymbol{x}_0)]$. The form of the VLB loss depends on the parameterisation used to predict $\boldsymbol{\mu}_\theta^{(t)}(\boldsymbol{x}_t)$. For our purposes, we train a neural network $\hat{\boldsymbol{x}}_{0_\theta}^{(t)}(\boldsymbol{x}_t)$ to estimate the "clean" sample $\boldsymbol{x}_0$ given the noisy sample $\boldsymbol{x}_t$. This results in a denoising objective of the form

$$\mathcal{L}_{\text{VLB}} = \mathbb{E}_{t, \boldsymbol{x}_0, \boldsymbol{x}_t|\boldsymbol{x}_0}\left[\|\boldsymbol{x}_0 - \hat{\boldsymbol{x}}_{0_\theta}^{(t)}(\boldsymbol{x}_t)\|_2^2\right], \qquad (4)$$

which corresponds to a weighted [18] version of the VLB.

During inference, $\hat{\boldsymbol{x}}_{0_\theta}^{(t)}(\boldsymbol{x}_t)$ is used to compute $\boldsymbol{\mu}_\theta^{(t)}(\boldsymbol{x}_t)$. Following [18], we set $\boldsymbol{\Sigma}_\theta^{(t)}(\boldsymbol{x}_t) = \sigma_t^2 \boldsymbol{I}$ to time-dependent constants that depend on the hyperparameters β_t. Given $\boldsymbol{\mu}_\theta$ and $\boldsymbol{\Sigma}_\theta$, samples $\boldsymbol{x}_0 \sim p_\theta(\boldsymbol{x}_0)$ are obtained using ancestral sampling, as in Eq. (2). DDPMs can be easily extended to sample from conditional distributions $p_\theta(\boldsymbol{x}_0|\boldsymbol{y})$, by passing a conditioning variable $\boldsymbol{y}$ to the denoising neural network such that "clean" sample estimates are given by $\hat{\boldsymbol{x}}_{0_\theta}^{(t)}(\boldsymbol{x}_t, \boldsymbol{y})$.

Neural Implicit Surfaces. A surface $\mathcal{S}$ in $\mathbb{R}^3$ can be implicitly defined as the zero-level-set or decision boundary of a function. Given an RGB image $\mathbf{I}$ of a subject, an estimate of the surface geometry of the subject may be obtained using an image-conditioned signed distance function (SDF). This can be represented by a coordinate-based neural network f_Θ, which outputs a signed distance value $d_{\boldsymbol{p}}$ and unshaded albedo colour $\boldsymbol{a}_{\boldsymbol{p}}$ given a query point $\boldsymbol{p} \in \mathbb{R}^3$. Hereafter, we use Θ to denote the set of all learnable parameters. The neural implicit surface corresponding to f_Θ is defined as

$$\mathcal{S}_\Theta(\mathbf{I}) = \left\{\boldsymbol{p} \in \mathbb{R}^3 | f_\Theta\left(\boldsymbol{p}; g_\Theta\left(\mathbf{I}\right)\right) = (0, \boldsymbol{a}_{\boldsymbol{p}})\right\}, \qquad (5)$$

where g_Θ is a feature extractor CNN that is used to condition f_Θ on the image $\mathbf{I}$ with pixel-aligned features $g_\Theta(\mathbf{I})$, following [5, 49]. The feature vector associated with $\boldsymbol{p}$, which we denote as $\boldsymbol{g}_{\boldsymbol{p}}$, is obtained by projecting $\boldsymbol{p}$ onto the image plane and bilinearly interpolating $g_\Theta(\mathbf{I})$ at this pixel location. The distance value $d_{\boldsymbol{p}}$ and albedo colour $\boldsymbol{a}_{\boldsymbol{p}}$ at $\boldsymbol{p}$ are concretely obtained as $(d_{\boldsymbol{p}}, \boldsymbol{a}_{\boldsymbol{p}}) = f_\Theta(\boldsymbol{p}, \boldsymbol{g}_{\boldsymbol{p}})$.

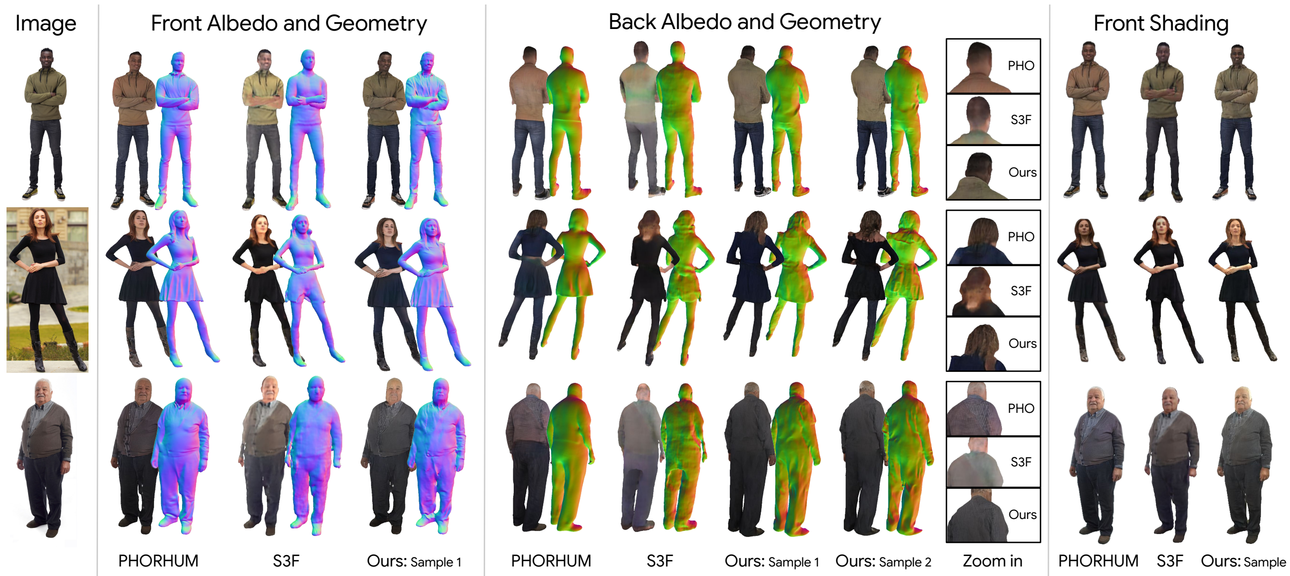

Figure 3. **Qualitative comparison against deterministic monocular 3D human reconstruction methods [5, 10] that predict geometry, surface albedo and shaded colour.** PHORHUM [5] (retrained on our dataset) outputs good front predictions, but exhibits over-smooth, flat geometry and blurry colours on the back. S3F [10] yields more detailed geometry, but colours are still often blurry. Moreover, both these methods occasionally paste the front colour predictions onto the back incorrectly (see row 3). Our method outputs *multiple* diverse samples, with a greater level of geometric detail and colour sharpness in uncertain regions, that are consistent with the input image after shading.

To decouple unshaded albedo colour and illumination-dependent shading, an additional neural network s_Θ [5] may be used to estimate a shading coefficient $\boldsymbol{s_p}$ at each surface point $\boldsymbol{p}$. This is obtained using

$$\boldsymbol{s_p} = s_\Theta\left(\boldsymbol{n_p}, \boldsymbol{l}(\mathbf{I})\right), \tag{6}$$

where $\boldsymbol{n_p} = \nabla_{\boldsymbol{p}} d_{\boldsymbol{p}}$ is the surface normal at $\boldsymbol{p}$ and $\boldsymbol{l}(\mathbf{I})$ is a scene illumination code estimated from the input image. The latter may be computed using the bottleneck of $g_\Theta(\mathbf{I})$, as in [5]. Then, the shaded colour at a point $\boldsymbol{p}$ is given by $\boldsymbol{c_p} = \boldsymbol{a_p} \odot \boldsymbol{s_p}$, where $\odot$ denotes element-wise multiplication.

Given an image-conditioned SDF f_Θ, various methods exist to extract and/or render the corresponding surface $\mathcal{S}$. An explicit mesh approximation of $\mathcal{S}$ is typically generated by running Marching Cubes [35] in a densely sampled 3D bounding box [5, 49, 50]. Standard graphics pipelines can be used to render various properties of the mesh, such as surface albedo, shaded colour, surface normal or depth maps. In addition, $\mathcal{S}$ may be directly rendered using sphere tracing [15], without generating an explicit mesh. Sphere tracing can be formulated as a differentiable operation [68], enabling the use of 2D rendering losses during training.

3.2. Implicit Surface Diffusion via Rendering

Our aim is to estimate the 3D surface geometry and appearance $\mathcal{S}$ of a human subject given a single RGB image $\mathbf{I}$. This is an ill-posed problem, since multiple 3D reconstructions can plausibly explain a 2D input image, *e.g.* due to occlusion or depth ambiguity. Thus, we seek to predict a *probability distribution* over 3D geometry and appearance conditioned on the RGB image, $p_\Theta(\mathcal{S}|\mathbf{I})$. Our method, **DiffHuman**, implements $p_\Theta(\mathcal{S}|\mathbf{I})$ using the framework of DDPMs [18], and enables us to sample multiple plausible 3D reconstructions.

We represent $\mathcal{S}$ as an implicit surface $\mathcal{S}_\Theta(\mathbf{I})$ defined by the corresponding neural networks f_Θ and g_Θ, as detailed in Eq. (5). If we wish to directly apply a DDPM to learn $p_\Theta(\mathcal{S}|\mathbf{I})$, we need to define forward and reverse diffusion processes over $\mathcal{S}$. This requires a suitable representation of $\mathcal{S}$ that can be noised and denoised. In the framework of implicit surfaces, the neural network f_Θ is fixed and acts as a decoder for the pixel-aligned features $g_\Theta(\mathbf{I})$. The latter could be an adequate choice for a representation; however, they are unknown before the network is trained. Specifically, we do not have access to ground-truth pixel aligned features $g_\Theta(\mathbf{I})$ for a given $\mathbf{I}$ *a priori*, and thus cannot add noise to or denoise them directly.

Instead, we model a distribution over image-based, pixel-aligned *observations* of $\mathcal{S}$ that cover the true $\mathcal{S}$ well. Specifically, we consider three types of observations of the front and back of $\mathcal{S}$ with respect to a fixed camera π: (i) unshaded albedo colour images $\mathbf{A}^F$ and $\mathbf{A}^B$, (ii) surface normal images $\mathbf{N}^F$ and $\mathbf{N}^B$, and (iii) depth maps $\mathbf{D}^F$ and $\mathbf{D}^B$. These are concatenated together to form an observation set

$$\boldsymbol{x}_0 = \{\mathbf{A}^F, \mathbf{A}^B, \mathbf{N}^F, \mathbf{N}^B, \mathbf{D}^F, \mathbf{D}^B\}. \tag{7}$$

In practice, this observation set is represented as an array $\boldsymbol{x}_0 \in [-1, 1]^{H \times W \times C}$. Given a surface $\mathcal{S}$, the corresponding $\boldsymbol{x}_0 = \texttt{render}(\mathcal{S}, \pi)$ can be obtained via rendering. Since $\boldsymbol{x}_0$ is effectively a multichannel image, a conventional image-based DDPM may be directly applied to learn $p_\Theta(\boldsymbol{x}_0|\mathbf{I})$. This would involve training a neural network $\hat{\boldsymbol{x}}^{(t)}_{0_\Theta}(\boldsymbol{x}_t, \mathbf{I})$ to estimate the "clean" observation set $\boldsymbol{x}_0$ by denoising $\boldsymbol{x}_t$, as detailed in Sec. 3.1. We modify this denoising step by incorporating a neural implicit surface as an intermediate 3D reconstruction and obtain the denoised estimate via rendering of this surface. This enables us to sample from a learned distribution over 3D surfaces during inference.

Specifically, given a noisy observation set $\boldsymbol{x}_t$ and conditioning image $\mathbf{I}$, we first compute noise-dependent pixel-aligned features $g_\Theta^{(t)}(\boldsymbol{x}_t, \mathbf{I})$. These are used to condition an SDF $f_\Theta^{(t)}(\boldsymbol{p}; g_\Theta^{(t)}(\boldsymbol{x}_t, \mathbf{I}))$, which is dependent on both $\boldsymbol{x}_t$ and $\mathbf{I}$. The networks $f_\Theta^{(t)}$ and $g_\Theta^{(t)}$ define a neural implicit surface $\mathcal{S}_\Theta^{(t)}(\boldsymbol{x}_\mathbf{t}, \mathbf{I})$, which is given by adding noise-dependence to Eq. (5). Then, we obtain an estimate of the denoised observation set with

$$\hat{\boldsymbol{x}}_{0_\Theta}^{(t)}(\boldsymbol{x}_t, \mathbf{I}) = \mathtt{render}\left(\mathcal{S}_\Theta^{(t)}(\boldsymbol{x}_t, \mathbf{I}), \pi\right). \tag{8}$$

Furthermore, we can obtain a shaded image $\mathbf{C}^{(t)}$ by applying a timestep-dependent shading network $s_\Theta^{(t)}$ to the front albedo and normal predictions that comprise $\hat{\boldsymbol{x}}_{0_\Theta}^{(t)}$:

$$\mathbf{C}^{(t)} = \mathbf{A}^F \odot s_\Theta^{(t)}(\mathbf{N}^F, \boldsymbol{l}(\mathbf{I})). \tag{9}$$

$f_\Theta^{(t)}$ and $g_\Theta^{(t)}$ are trained by minimising the following DDPM loss in each training iteration:

$$\mathcal{L}_{\text{VLB}}^{\mathtt{render}} = \left\|\boldsymbol{x}_0 - \mathtt{render}\left(\mathcal{S}_\Theta^{(t)}(\boldsymbol{x}_t, \mathbf{I}), \pi\right)\right\|_2^2, \tag{10}$$

which follows from Eq. (4). Images and corresponding clean observation sets are sampled from a target data distribution $\boldsymbol{x}_0, \mathbf{I} \sim q(\boldsymbol{x}_0, \mathbf{I})$ and a timestep is sampled from $t \sim \mathcal{U}(\{1, ..., T\})$. Noised observations are sampled from $\boldsymbol{x}_t \sim q(\boldsymbol{x}_t|\boldsymbol{x}_0)$. Moreover, we can supervise on $\mathbf{C}^{(t)}$ to ensure that all 3D samples from our predicted distribution are consistent with the conditioning image after rendering and shading. Please refer to the Suppl. Mat. for details.

Once $f_\Theta^{(t)}$ and $g_\Theta^{(t)}$ are trained, we can ancestrally sample reverse process trajectories over observation sets $\boldsymbol{x}_{0:T} \sim p_\Theta(\boldsymbol{x}_{0:T}|\mathbf{I})$, by computing and rendering $\mathcal{S}_\Theta^{(t)}(\boldsymbol{x}_\mathbf{t}, \mathbf{I})$ in each denoising step. Notably, 3D samples $\mathcal{S} \sim p_\Theta(\mathcal{S}|\mathbf{I})$ are given by the final reconstruction $\mathcal{S} = \mathcal{S}_\Theta^{(1)}(\boldsymbol{x}_1, \mathbf{I})$.

The above formulation of diffusion via rendering is similar to [57, 59]. These approaches implement diffusion over multiple images of an underlying 3D scene from different views, and incorporate rendering of an intermediate volumetric 3D representation [39] into the reverse process. Our method considers various pixel-aligned observations of a 3D human from the same view and reconstructs intermediate implicit surfaces during denoising. We also incorporate probabilistic scene illumination estimation via Eqn. 9. Nonetheless, all these approaches involve rendering of a neural 3D representation in every single denoising step, which is computationally expensive during inference. Consequently, Sec. 3.3 introduces a hybrid diffusion model that integrates both rendering and learned generation in the denoising process, enabling 3D sampling at considerably reduced runtime.

3.3. Hybrid Implicit Surface Diffusion

The diffusion-via-rendering formulation introduced in Sec. 3.2 involves rendering an implicit surface in every denoising step. This is memory- and time-intensive, both when the surface is directly rendered using sphere tracing, and also when an explicit mesh is extracted with Marching Cubes and subsequently rasterised. Sphere tracing requires K successive evaluations of $f_\Theta^{(t)}$ *per pixel*, where K equals the number of tracing steps until a surface is found or the ray is terminated; $K \approx 30$ in our experiments resulting in 7.9M function evaluations for an image of 512×512px. Marching Cubes, on the other hand, requires one $f_\Theta^{(t)}$ evaluation *per 3D grid point*. This can be accelerated via octree sampling, but still requires $> 10^5$ function evaluations for a mesh of medium spatial resolution.

To mitigate this computational overhead, we note that we are ultimately only interested in 3D reconstruction samples obtained at the end of the denoising process $\mathcal{S} = \mathcal{S}_\Theta^{(1)}(\boldsymbol{x}_1, \mathbf{I})$. Thus, explicitly computing and rendering $\mathcal{S}_\Theta^{(t)}$ in every denoising step during inference is wasteful. Instead, we introduce an additional "generator" neural network $h_\Theta^{(t)}$ that is trained to imitate the rendering of an implicit surface conditioned on pixel-aligned features. Concretely, $h_\Theta^{(t)}$ directly maps noise-dependent pixel-aligned features $g_\Theta^{(t)}(\boldsymbol{x}_t, \mathbf{I})$ to an estimate of the denoised observation set

$$\bar{\boldsymbol{x}}_{0_\Theta}^{(t)}(\boldsymbol{x}_t, \mathbf{I}) = h_\Theta^{(t)}\left(g_\Theta^{(t)}(\boldsymbol{x}_t, \mathbf{I})\right), \tag{11}$$

where $\bar{\boldsymbol{x}}_{0_\Theta}^{(t)}$ should approximate $\hat{\boldsymbol{x}}_{0_\Theta}^{(t)}$ obtained via rendering (Eq. (8)). $h_\Theta^{(t)}$ is trained with the following objective:

$$\mathcal{L}_{\text{VLB}}^{\mathtt{generate}} = \left\|\boldsymbol{x}_0 - h_\Theta^{(t)}\left(g_\Theta^{(t)}(\boldsymbol{x}_t, \mathbf{I})\right)\right\|_2^2. \tag{12}$$

During inference, we sample reverse process trajectories $\boldsymbol{x}_{1:T} \sim p_\Theta(\boldsymbol{x}_{1:T}|\mathbf{I})$ using generated denoised estimates $\bar{\boldsymbol{x}}_{0_\Theta}^{(t)}$ instead of rendering. We only explicitly compute 3D reconstruction samples $\mathcal{S} = \mathcal{S}_\Theta^{(1)}(\boldsymbol{x}_1, \mathbf{I})$ at the end of the reverse process, using the final noisy observations $\boldsymbol{x}_1$. Marching Cubes is simply applied once to extract the final mesh. In general, a forward pass through the neural network $h_\Theta^{(t)}$ is computationally cheaper than explicit rendering via sphere tracing or Marching Cubes and rasterisation. The computational cost savings sum together over the reverse process, which may involve hundreds of denoising steps.

Generating denoised estimates with the composition of neural networks $h_\Theta^{(t)} \circ g_\Theta^{(t)}$ is reminiscent of the standard denoising network architecture used in conventional DDPMs. However, we simultaneously apply both $\mathcal{L}_{\text{VLB}}^{\mathtt{render}}$ and $\mathcal{L}_{\text{VLB}}^{\mathtt{generate}}$ during training, resulting in a hybrid diffusion framework that combines rendering and generation. This ensures that $g_\Theta^{(t)}(\boldsymbol{x}_t, \mathbf{I})$ continue to be features that validly

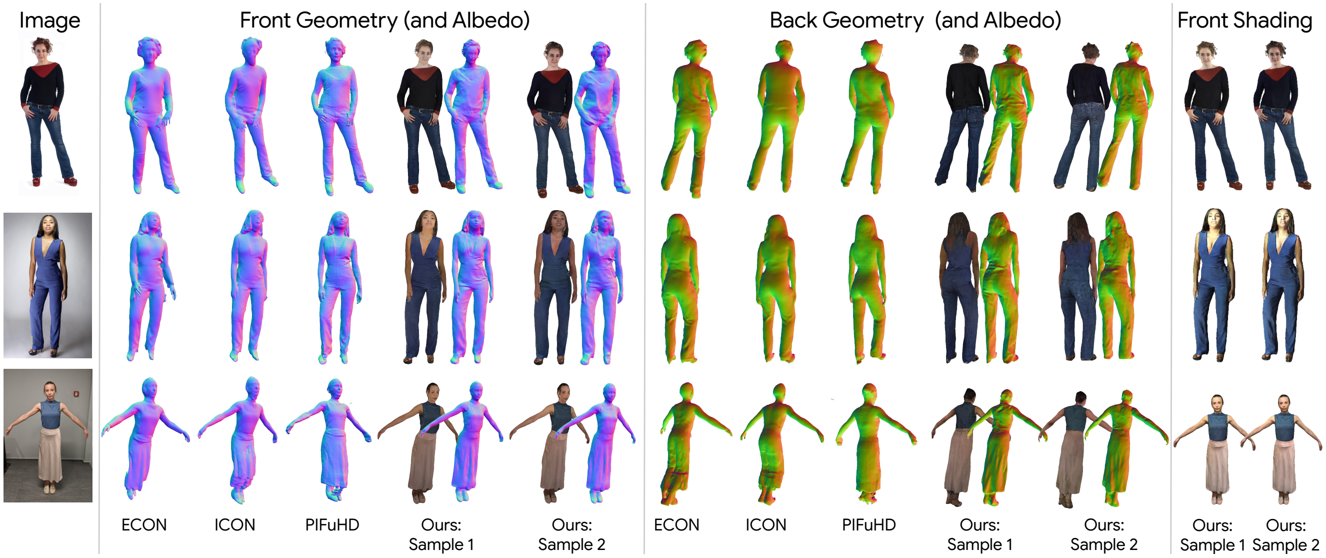

Figure 4. **Qualitative comparison against deterministic monocular 3D human reconstruction methods that predict only surface geometry: PIFuHD [50], ICON [63] and ECON [64].** Samples from our method generally exhibit greater geometric detail in uncertain regions, while maintaining a high level of consistency with the input image in shaded renders. Moreover, deterministic methods often fall back towards the mean of the training data distribution when faced with ambiguous and challenging inputs [8, 11, 37]; *e.g.* predicting trousers from the back instead of a long skirt in row 3. This can be mitigated by learning to predict a distribution over reconstructions instead.

condition an SDF and $h_\Theta^{(t)}$ learns to decode these features into the observation sets corresponding to the SDF.

3.4. Implementation Details

Our networks are trained with a synthetic training dataset, consisting of HDRI-based illuminated renders of real body scans from [45] and our own captured data. We use ~ 5.9K scans of ~ 1.1K identities to render ~ 450K training examples, each consisting of a 512×512px image (with masked-out background) and an observation set $\boldsymbol{x}_0$. The back-views in $\boldsymbol{x}_0$ are created by inverting the z-buffer, and thus rendering the scans back-to-front. [45] provides ground-truth albedo textures, which we use to generate $\mathbf{A}^F$ and $\mathbf{A}^B$ in $\boldsymbol{x}_0$. Our scans approximately capture albedo using even ambient lighting. However, this is not perfect and causes our model to sometimes yield shading artefacts in albedo predictions.

In addition to $\mathcal{L}^*_{\text{VLB}}$, we use 3D losses on $d_{\boldsymbol{p}}$, $\boldsymbol{a}_{\boldsymbol{p}}$ and $\boldsymbol{n}_{\boldsymbol{p}}$ to improve training stability (see Suppl. Mat. for details). During training, we render 32×32px patches using differentiable ray tracing [68] to form $\hat{\boldsymbol{x}}_{0_\Theta}^{(t)}$ for $\mathcal{L}^{\texttt{render}}_{\text{VLB}}$. We apply $\mathcal{L}^{\texttt{generate}}_{\text{VLB}}$ on the full resolution generation $\bar{\boldsymbol{x}}_{0_\Theta}^{(t)}$. g_Θ and h_Θ are U-Nets [47] with 13 encoder-decoder layers each and skip connections. Both networks double the filter size in each encoder layer, starting from 64 up to 512 for g_Θ and up to 128 for h_Θ. g_Θ outputs a pixel-aligned feature map in $\mathbb{R}^{512 \times 512 \times 256}$. f_Θ and s_Θ are MLPs, following [5].

At test time, we perform 100 DDIM [56] denoising steps. At each step, we can choose to denoise via `render` or `generate` and we ablate different strategies in Sec. 4.1. For faster inference, we use Marching Cubes and rasterisation, instead of sphere tracing, in `render`. If we `render` at higher noise (large t), we run Marching Cubes on a 256^3 grid. For small t, we use 512^3. The final denoising step always has to be a `render` step, since `generate` does not produce 3D geometry. However, we do not perform a full step of `render` at $t = 1$, but only reconstruct $\mathcal{S}$ and omit rasterisation of $\boldsymbol{x}_0$.

4. Experiments

This section quantitatively compares DiffHuman with the state-of-the-art photorealistic human reconstruction methods, and visually demonstrates the quality of 3D samples conditioned on internet images. Furthermore, we experimentally ablate a number of crucial design choices. Please see the Suppl. Mat. for additional results and experiments.

Test dataset and metrics. We use the test set of [5] for numerical evaluation, and report both 3D metrics and image-based (pixel-aligned) metrics.

The 3D metrics consist of bi-directional Chamfer distance $\times 10^{-3}$ (Ch. $\downarrow$), Normal Consistency (NC $\uparrow$), and Volumetric Intersection over Union (IoU $\uparrow$). Iterative Closest Points is used to first align 3D predictions with the ground-truth.

3D metrics are sensitive to the assumed camera model. In contrast, image-based metrics partially ignore errors due to an incorrect camera, instead focusing on surface structure. This is better correlated with perceived quality. Image-based metrics are computed by rendering 3D predictions with each model's assumed camera, and comparing the resulting images against ground-truth 2D renders. Specifically, we report Structural Similarity Index (SSIM $\uparrow$), Learned Perceptual Image Patch Similarity (LPIPS $\downarrow$) [73], and Peak Signal-to-Noise Ratio (PSNR $\uparrow$) for albedo and shaded colour renders. For normal renders, we report the angular error in degrees

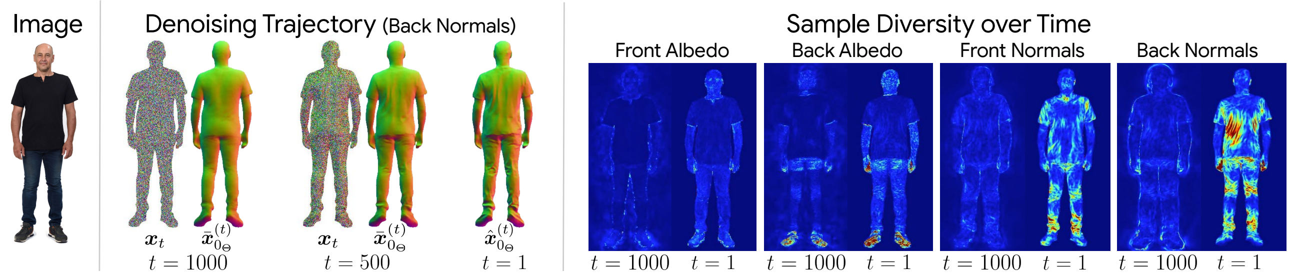

Figure 5. **Visualisation of the reverse process.** The denoising trajectory shows noisy samples $\boldsymbol{x}_t$ and generated clean predictions $\bar{\boldsymbol{x}}_{0_\Theta}^{(t)}$ at each timestep. Clean predictions are initially very simple, akin to many deterministic approaches, and become detailed over time. The heatmaps show sample diversity, computed as the per-pixel variance of the observations in $\bar{\boldsymbol{x}}_{0_\Theta}^{(t)}$ over 10 samples. Diversity is low at the start of the denoising process ($t = 1000$), but increases gradually as the samples diverge. Back diversity is, intuitively, greater than the front.

Render frequency	Runtime s / sample	3D Ch. ↓	Alb. F / B PSNR ↑	Nor. F / B Ang. ↓	Sha. F PSNR ↑
Per step	496	**0.99**	22.92 / 20.95	21.59 / 22.88	26.57
Per 10 steps	86	1.12	23.24 / 21.07	**19.05 / 22.46**	27.08
Per 25 steps	34	1.16	**23.26 / 21.06**	19.11 / 22.52	**27.09**
Final step	9	1.16	23.26 / 21.05	19.12 / 22.55	**27.09**

Table 1. **Ablation of hybrid implicit surface diffusion.** $N = 5$ samples are obtained using 100 DDIM [56] steps. We periodically `render` every 1, 10 or 25 steps, or only in the final step. All other denoising steps use `generate`. While per-step `render` performs best on 3D metrics, predominantly using `generate` results in better colour. The **best** and second best results are marked.

(Ang. ↓) and LPIPS. SSIM and LPIPS evaluate structure rather than pixel-to-pixel errors. The latter can be misleadingly low for over-smooth reconstructions; SSIM and LPIPS, in our experience, better capture the perceived quality.

For our method, we report metrics using the best-of-$N \in \{1, 5, 10\}$ reconstructions, following [7, 29, 53]. Specifically, we obtain N different 3D samples for each test image, and aggregate metrics using the numerical best reconstruction. The Suppl. Mat. additionally reports mean metrics, and discusses the use of best-of-N vs. mean metrics for this task.

Baselines. We compare with a large number of recent approaches to monocular 3D human reconstruction. Only PHORHUM [5], S3F [10], and PIFu [49] reconstruct surface color, but the latter does not decompose albedo and shading. ARCH [22] and ARCH++ [17] do not reconstruct true surface details but use normal mapping to enhance the visual fidelity of results. For fairness, we evaluate the estimated normals instead of true surface normals for these methods. We also compare against a version of PHORHUM retrained with our larger dataset, resulting in a strong baseline method.

4.1. Ablation Studies

Since we retrained PHORHUM [5] on our larger synthetic dataset, we consider it as an ablation of our main design choice: probabilistically modelling the reconstruction process using a diffusion model that predicts distributions over 3D human reconstructions. Even though the retrained PHORHUM model turned out to be a very strong baseline, DiffHuman is able to produce reconstructions with higher visual fidelity and better numerical performance, especially for unseen regions. Better performance for unseen regions can be explained by our probabilistic approach being less prone to averaging effects caused by the inherent aleatoric [26] uncertainty in an ill-posed problem.

Additionally, we ablate our hybrid diffusion strategy: the generator network $h_\Theta^{(t)}$ and denoising via `generate` instead of `render`. In Tab. 1 we compare the use of `render` in every denoising step, every 10, every 25, and only at the final step (to extract the final mesh with Marching Cubes). We use `generate` for all other steps. The performances of all variants are comparable, suggesting that $h_\Theta^{(t)}$ has learned to imitate `render` sufficiently well. Nonetheless, Ch. is slightly better when using `render` in every step, while colour metrics are improved with lower `render` frequency. This is expected: `render` actually reconstructs 3D geometry whereas `generate` only synthesises observations. Errors in this approximate synthesis operation may accumulate over the course of the denoising process. On the other hand, `generate` may also "fix" inconsistent colour extracted from the signed-distance and colour field. Crucially, the use of `generate` results in an up to $55\times$ speed up at comparable quality. We use the "final step" strategy for all remaining experiments. The Suppl. Mat. shows that 3D samples obtained via per-step and final step rendering are visually similar.

4.2. Reconstruction Accuracy

Tab. 2 and Tab. 3 compare recent methods in terms of image-based metrics and 3D metrics respectively. DiffHuman yields improved performance for image-based back metrics, especially with growing number of samples N, while front metrics are stable for all N. This is because the back is unobserved – more samples means a higher chance of finding the correct reconstruction – while the front is visible and thus variation is lower. We do not split observed and unobserved parts for 3D metrics, but also find a global improvement with higher N. Our retrained PHORHUM also performs well, and often slightly better than $N = 1$. This is again expected, as DiffHuman is conducting a much harder task: while PHORHUM outputs a single solution, DiffHuman models the distribution over possible 3D reconstructions.

Method	Albedo Front			Albedo Back			Normals Front		Normals Back		Shaded Front		
	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	Ang. ↓	LPIPS ↓	Ang. ↓	LPIPS ↓	SSIM ↑	LPIPS ↓	PSNR ↑
PIFu [49]	–	–	–	–	–	–	26.69	0.17	28.49	0.26	0.83	0.16	24.57
PIFuHD [50]	–	–	–	–	–	–	23.04	0.10	26.33	0.22	–	–	–
Geo-PiFU [16]	–	–	–	–	–	–	30.17	0.19	31.93	0.26	–	–	–
ARCH [22]	–	–	–	–	–	–	32.20	0.20	33.96	0.27	0.72	0.23	19.28
ARCH++ [17]	–	–	–	–	–	–	27.20	0.17	30.62	0.24	0.83	0.17	22.69
PHORHUM [5]	0.85	0.12	22.23	**0.76**	0.22	20.19	20.53	**0.11**	23.55	0.20	0.85	0.13	24.01
PaMIR [74]	–	–	–	–	–	–	22.88	0.14	27.26	0.23	–	–	–
ICON [63]	–	–	–	–	–	–	23.57	0.14	26.98	0.23	–	–	–
ECON [64]	–	–	–	–	–	–	22.27	0.15	26.98	0.23	–	–	–
D-IF [67]	–	–	–	–	–	–	24.52	0.15	27.84	0.23	–	–	–
S3F [10]	0.60	0.36	15.33	0.63	0.39	15.99	23.76	0.25	23.72	0.27	0.61	0.33	17.38
PHORHUM (retr.)	0.85	**0.11**	22.57	0.73	**0.21**	19.74	**18.41**	0.12	22.82	0.19	0.86	0.10	25.23
DiffHuman: $N = 1$	0.84	0.12	22.44	0.71	0.23	19.77	19.70	0.14	24.34	0.18	0.89	0.10	26.81
DiffHuman: $N = 5$	**0.86**	**0.11**	23.26	0.74	0.22	21.05	19.12	0.13	22.55	0.16	**0.90**	0.10	27.09
DiffHuman: $N = 10$	**0.86**	**0.11**	**23.47**	0.75	**0.21**	**21.24**	18.91	0.13	**22.34**	**0.15**	**0.90**	**0.09**	**27.15**

Table 2. **Quantitative comparison against other monocular 3D human reconstruction methods in terms of pixel-aligned metrics.** Since DiffHuman predicts a distribution over 3D reconstructions, we report metrics using the best of $N = 1, 5$ and 10 samples drawn for each test image. We `render` only in the final denoising step, and use `generate` otherwise. The **best** and second best results are marked.

Method	Ch. ↓	IoU ↑	NC ↑
PIFu [49]	3.21	0.61	0.77
PIFuHD [50]	4.54	0.62	0.78
Geo-PIFu [16]	4.98	0.54	0.72
ARCH [22] †	3.58	0.57	0.75
ARCH++ [17] †	3.48	0.59	0.77
PaMIR [74] †	2.88	0.61	0.77
PHORHUM [5]	1.29	**0.73**	0.85
ICON [63] †	2.44	0.62	0.78
ECON[64] †	3.48	0.61	0.76
D-IF [67] †	2.97	0.58	0.78
S3F [10] †	2.35	0.63	0.80
PHORHUM (retrained)	1.10	**0.73**	**0.87**
DiffHuman: $N = 1$	1.98	0.69	0.83
DiffHuman: $N = 5$	1.16	0.72	0.86
DiffHuman: $N = 10$	**1.09**	**0.73**	0.86

Table 3. **Quantitative comparison against other monocular 3D human reconstruction methods in terms of 3D metrics.** Since DiffHuman predicts a distribution over 3D reconstructions, we report metrics using the best of $N = 1, 5$ and 10 samples drawn for each test image. We `render` only in the final denoising step, and use `generate` otherwise. The **best** and second best results are marked. Methods marked with † use a parametric body model.

Furthermore, the 3D ground-truth is *but one* plausible reconstruction in a monocular setting. Our method is able to yield other 3D solutions that are input-consistent, but differ from the ground-truth resulting in worse metrics. Nevertheless, DiffHuman is competitive even for $N = 1$, and produces qualitatively better reconstructions, as discussed below.

4.3. Qualitative Results and Diversity

We show the qualitative performance of DiffHuman in Figs. 3 and 4 side-by-side with state-of-the-art approaches. All competing methods only return one solution, while DiffHuman allows us to sample multiple diverse results. We show two reconstructions per image. Consistent with the numerical results in Tab. 3, DiffHuman can shine the most when reconstructing unobserved back-sides. Despite performing well numerically, our retrained PHORHUM baseline does not produce good reconstructions of uncertain regions, with blurry colours and a lack of geometric detail. Methods that explicitly estimate a back normal map [50, 63, 64] produce over-smooth back reconstructions. In contrast, samples from DiffHuman exhibit fine wrinkles and details both for observed and unobserved regions, as shown by rows 1 and 2 in Fig. 4. PHORHUM and S3F [10] tend to simply clone front colours to the person's back-side – a reasonable approach for some but not all garments. *E.g.* in the 3rd row of Fig. 3, the subject's shirt is cloned onto the jacket in the back. In contrast, DiffHuman reliably colours unobserved regions without such artefacts. Moreover, we can obtain diverse reconstructions, shown by the different dresses and hairstyles in Fig. 1 and row 2 of Fig. 3. We visualise diversity over the denoising process in Fig. 5. The Suppl. Mat. contains additional qualitative results, as well as unconditional and edge-conditioned samples from DiffHuman. These exhibit greater diversity than image-conditioned samples, as RGB images are strong conditioning signals.

5. Conclusion

We presented DiffHuman, a probabilistic method for photorealistic 3D human reconstruction from a single RGB image. We build on top of recent advances in diffusion-based generative modelling and propose a novel pipeline for fast sampling of 3D human shapes. Our model is numerically competitive with the state-of-the-art, while improving the visual fidelity and the level of detail of unseen surfaces. Furthermore, we can sample multiple input-consistent but diverse 3D human reconstructions. Our novel hybrid implicit surface diffusion speeds up 3D sampling at test time compared with diffusion-via-rendering [57, 59], giving a general framework for computationally cheaper diffusion over implicit 3D representations. A limitation of our method is that it currently requires examples with known 3D geometry for training, which constrains the amount of data that can be used. In future work, we plan to overcome this by leveraging data with partial 2D and 2.5D supervision.

References

[1] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Detailed human avatars from monocular video. In *Proceedings of the Interntional Conference on 3D Vision (3DV)*, 2018. 2

[2] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3D people models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[3] Thiemo Alldieck, Marcus Magnor, Bharat Lal Bhatnagar, Christian Theobalt, and Gerard Pons-Moll. Learning to reconstruct people in clothing from a single rgb camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[4] Thiemo Alldieck, Gerard Pons-Moll, Christian Theobalt, and Marcus Magnor. Tex2shape: Detailed full human body geometry from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2293–2303, 2019. 2

[5] Thiemo Alldieck, Mihai Zanfir, and Cristian Sminchisescu. Photorealistic monocular 3D reconstruction of humans wearing clothing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2, 3, 4, 6, 7, 8

[6] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. Scape: shape completion and animation of people. *SIGGRAPH*, 2005. 2

[7] Benjamin Biggs, David Novotny, Sebastien Ehrhardt, Hanbyul Joo, Ben Graham, and Andrea Vedaldi. 3d multi-bodies: Fitting sets of plausible 3d human models to ambiguous image data. *Advances in Neural Information Processing Systems*, 33:20496–20507, 2020. 3, 7

[8] Christopher M. Bishop. Mixture Density Networks. Technical report, Aston University, 1994. 1, 6

[9] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. 2

[10] Enric Corona, Mihai Zanfir, Thiemo Alldieck, Eduard Gabriel Bazavan, Andrei Zanfir, and Cristian Sminchisescu. Structured 3d features for reconstructing relightable and animatable avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 2, 4, 7, 8

[11] Qiongjie Cui, Huaijiang Sun, and Fei Yang. Learning dynamic relationships for 3d human motion prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6519–6527, 2020. 1, 6

[12] Ziya Erkoç, Fangchang Ma, Qi Shan, Matthias Nießner, and Angela Dai. Hyperdiffusion: Generating implicit neural fields with weight-space diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 3

[13] Jia Gong, Lin Geng Foo, Zhipeng Fan, Qiuhong Ke, Hossein Rahmani, and Jun Liu. Diffpose: Toward more reliable 3d pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13041–13051, 2023. 3

[14] Riza Alp Guler and Iasonas Kokkinos. Holopose: Holistic 3D human reconstruction in-the-wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[15] John C Hart. Sphere tracing: A geometric method for the antialiased ray tracing of implicit surfaces. *The Visual Computer*, 12(10):527–545, 1996. 4

[16] Tong He, John Collomosse, Hailin Jin, and Stefano Soatto. Geo-pifu: Geometry and pixel aligned implicit functions for single-view human reconstruction. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2, 8

[17] Tong He, Yuanlu Xu, Shunsuke Saito, Stefano Soatto, and Tony Tung. Arch++: Animation-ready clothed human reconstruction revisited. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 7, 8

[18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *arXiv preprint arxiv:2006.11239*, 2020. 1, 2, 3, 4

[19] Shoukang Hu, Fangzhou Hong, Liang Pan, Haiyi Mei, Lei Yang, and Ziwei Liu. Sherf: Generalizable human nerf from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 3

[20] Yangyi Huang, Hongwei Yi, Weiyang Liu, Haofan Wang, Boxi Wu, Wenxiao Wang, Binbin Lin, Debing Zhang, and Deng Cai. One-shot implicit animatable avatars with model-based priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8974–8985, 2023. 3

[21] Yangyi Huang, Hongwei Yi, Yuliang Xiu, Tingting Liao, Jiaxiang Tang, Deng Cai, and Justus Thies. TeCH: Text-guided Reconstruction of Lifelike Clothed Humans. In *Proceedings of the International Conference on 3D Vision (3DV)*, 2024. 2

[22] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. Arch: Animatable reconstruction of clothed humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 7, 8

[23] Mustafa Işık, Martin Rünz, Markos Georgopoulos, Taras Khakhulin, Jonathan Starck, Lourdes Agapito, and Matthias Nießner. Humanrf: High-fidelity neural radiance fields for humans in motion. *ACM Transactions on Graphics (TOG)*, 42(4):1–12, 2023. 3

[24] Wei Jiang, Kwang Moo Yi, Golnoosh Samei, Oncel Tuzel, and Anurag Ranjan. Neuman: Neural human radiance field from a single video. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 3

[25] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7122–7131, 2018. 2

[26] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 7

[27] Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. PARE: Part attention regressor for 3D

human body estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2

[28] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2252–2261, 2019. 2

[29] Nikos Kolotouros, Georgios Pavlakos, Dinesh Jayaraman, and Kostas Daniilidis. Probabilistic modeling for human mesh recovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11605–11614, 2021. 3, 7

[30] Verica Lazova, Eldar Insafutdinov, and Gerard Pons-Moll. 360-degree textures of people in clothing from a single image. In *Proceedings of the International Conference on 3D Vision (3DV)*, pages 643–653. IEEE, 2019. 2

[31] Chen Li and Gim Hee Lee. Generating multiple hypotheses for 3d human pose estimation with mixture density network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9887–9895, 2019. 3

[32] Zhe Li, Tao Yu, Zerong Zheng, Kaiwen Guo, and Yebin Liu. Posefusion: Pose-guided selective fusion for single-view human volumetric capture. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14162–14172, 2021. 2

[33] Zhe Li, Zerong Zheng, Yuxiao Liu, Boyao Zhou, and Yebin Liu. Posevocab: Learning joint-structured pose embeddings for human avatar modeling. In *ACM SIGGRAPH Conference Proceedings*, 2023. 3

[34] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ToG*, 2015. 2, 3

[35] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3D surface construction algorithm. *SIGGRAPH*, 1987. 4

[36] Shitong Luo and Wei Hu. Diffusion probabilistic models for 3d point cloud generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2837–2845, 2021. 3

[37] Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. *arXiv preprint arXiv:1511.05440*, 2015. 1, 6

[38] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3D reconstruction in function space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[39] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 5

[40] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 2

[41] Norman Müller, Yawar Siddiqui, Lorenzo Porzi, Samuel Rota Bulò, Peter Kontschieder, and Matthias Nießner. Diffrf: Rendering-guided 3d radiance field diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4328–4338, 2023. 3

[42] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter Gehler, and Bernt Schiele. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *Proceedings of the International Conference on 3D Vision (3DV)*, 2018. 2

[43] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019. 2

[44] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3d human pose and shape from a single color image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 459–468, 2018. 2

[45] RenderPeople Dataset. Renderpeople dataset. https://renderpeople.com/. 6

[46] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10684–10695, 2022. 3

[47] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015. 6

[48] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Adv. Neural Inform. Process. Syst.*, 35: 36479–36494, 2022. 3

[49] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 2, 3, 4, 7, 8

[50] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3D human digitization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 2, 4, 6, 8

[51] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Synthetic training for accurate 3d human pose and shape estimation in the wild. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2020. 2

[52] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Hierarchical kinematic probability distributions for 3D human shape and pose estimation from images in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11219–11229, 2021. 3

[53] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Humaniflow: Ancestor-conditioned normalising flows on SO (3)

manifolds for human pose and shape distribution estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4779–4789, 2023. 3, 7

[54] Wenkang Shan, Zhenhua Liu, Xinfeng Zhang, Zhao Wang, Kai Han, Shanshe Wang, Siwei Ma, and Wen Gao. Diffusion-based 3d human pose estimation with multi-hypothesis aggregation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 3

[55] Ruizhi Shao, Zerong Zheng, Hongwen Zhang, Jingxiang Sun, and Yebin Liu. Diffustereo: High quality human reconstruction via diffusion-based stereo using sparse cameras. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 2

[56] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020. 6, 7

[57] Stanislaw Szymanowicz, Christian Rupprecht, and Andrea Vedaldi. Viewset diffusion: (0-)image-conditioned 3D generative models from 2D data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2, 3, 5, 8

[58] Vince J. K. Tan, Ignas Budvytis, and Roberto Cipolla. Indirect deep structured learning for 3D human shape and pose prediction. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2017. 2

[59] Ayush Tewari, Tianwei Yin, George Cazenavette, Semon Rezchikov, Joshua B. Tenenbaum, Frédo Durand, William T. Freeman, and Vincent Sitzmann. Diffusion with forward models: Solving stochastic inverse problems without direct supervision. *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 2, 3, 5, 8

[60] Tom Wehrbein, Marco Rudolph, Bodo Rosenhahn, and Bastian Wandt. Probabilistic monocular 3d human pose estimation with normalizing flows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11199–11208, 2021. 3

[61] Chung-Yi Weng, Brian Curless, Pratul P. Srinivasan, Jonathan T. Barron, and Ira Kemelmacher-Shlizerman. HumanNeRF: Free-viewpoint rendering of moving people from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16210–16220, 2022. 3

[62] Donglai Xiang, Fabian Prada, Chenglei Wu, and Jessica Hodgins. Monoclothcap: Towards temporally coherent clothing capture from monocular rgb video. In *Proceedings of the International Conference on 3D Vision (3DV)*, pages 322–332. IEEE, 2020. 2

[63] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J Black. Icon: Implicit clothed humans obtained from normals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2, 6, 8

[64] Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J. Black. ECON: Explicit Clothed humans Optimized via Normal integration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 6, 8

[65] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Ghum & ghuml: Generative 3D human shape and articulated pose models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[66] Hongyi Xu, Thiemo Alldieck, and Cristian Sminchisescu. H-nerf: Neural radiance fields for rendering and temporal reconstruction of humans in motion. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 3

[67] Xueting Yang, Yihao Luo, Yuliang Xiu, Wei Wang, Hao Xu, and Zhaoxin Fan. D-if: Uncertainty-aware human digitization via implicit distribution field. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9122–9132, 2023. 2, 8

[68] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 4, 6

[69] Andrei Zanfir, Eduard Gabriel Bazavan, Mihai Zanfir, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Neural descent for visual 3d human pose and shape. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14484–14493, 2021. 2

[70] Mihai Zanfir, Andrei Zanfir, Eduard Gabriel Bazavan, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Thundr: Transformer-based 3D human reconstruction with markers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[71] Xiaohui Zeng, Arash Vahdat, Francis Williams, Zan Gojcic, Or Litany, Sanja Fidler, and Karsten Kreis. Lion: Latent point diffusion models for 3d shape generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 3

[72] Hongwen Zhang, Jie Cao, Guo Lu, Wanli Ouyang, and Zhenan Sun. Danet: Decompose-and-aggregate network for 3D human shape and pose estimation. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 935–944, 2019. 2

[73] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. 6

[74] Zerong Zheng, Tao Yu, Yebin Liu, and Qionghai Dai. Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. *PAMI*, 2021. 2, 8

[75] Hao Zhu, Xinxin Zuo, Sen Wang, Xun Cao, and Ruigang Yang. Detailed human shape estimation from a single image by hierarchical mesh deformation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2