SiTH: Single-view Textured Human Reconstruction with Image-Conditioned Diffusion

Hsuan-I Ho Jie Song Otmar Hilliges

Department of Computer Science, ETH Zürich

https://ait.ethz.ch/sith

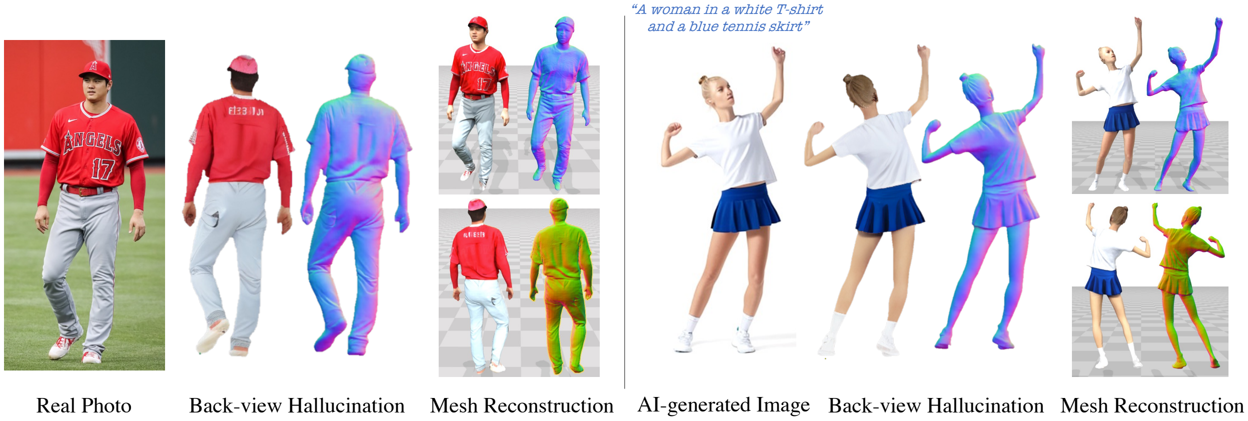

Figure 1. **Single-view textured human reconstruction**. SiTH is a novel pipeline for creating high-quality and fully textured 3D human meshes from single images. We first hallucinate back-view appearances through an image-conditioned diffusion model, followed by the reconstruction of full-body textured meshes using both the front and back-view images. Our pipeline enables the creation of lifelike and diverse 3D humans from unseen photos (*left*) and AI-generated images (*right*).

Abstract

A long-standing goal of 3D human reconstruction is to create lifelike and fully detailed 3D humans from single-view images. The main challenge lies in inferring unknown body shapes, appearances, and clothing details in areas not visible in the images. To address this, we propose SiTH, a novel pipeline that uniquely integrates an image-conditioned diffusion model into a 3D mesh reconstruction workflow. At the core of our method lies the decomposition of the challenging single-view reconstruction problem into generative hallucination and reconstruction subproblems. For the former, we employ a powerful generative diffusion model to hallucinate unseen back-view appearance based on the input images. For the latter, we leverage skinned body meshes as guidance to recover full-body texture meshes from the input and back-view images. SiTH requires as few as 500 3D human scans for training while maintaining its generality and robustness to diverse images. Extensive evaluations on two 3D human benchmarks, including our newly created one, highlighted our method's superior accuracy and perceptual quality in 3D textured human reconstruction.

1. Introduction

With the growing popularity of 3D and virtual reality applications, there has been increasing interest in creating realistic 3D human models. In general, crafting 3D humans is labor-intensive, time-consuming, and requires collaboration from highly skilled professionals. To bring lifelike 3D humans to reality and to support both expert and amateur creators in this task, it is essential to enable users to create textured 3D humans from simple 2D images or photos.

Reconstructing a fully textured human mesh from a single-view image presents an ill-posed problem with two major challenges. Firstly, the appearance information required for generating texture in unobserved regions is missing. Secondly, 3D information for mesh reconstruction, such as depth, surface, and body pose, becomes ambiguous in a 2D image. Previous efforts [4, 60, 79] attempted to tackle these challenges in a data-driven manner, focusing on training neural networks with image-mesh pairs. However, these approaches struggle with images featuring unseen appearances or poses, due to limited 3D human training data. More recent studies [61, 73, 79] introduced additional 3D reasoning modules to enhance robustness against unseen poses. Yet, generating realistic and full-body textures from unseen appearances still remains an unsolved problem.

To address the above challenges, we propose SiTH, a novel pipeline that integrates an image-conditioned diffusion model to reconstruct lifelike 3D textured humans from monocular images. At the core of our approach is the decomposition of the challenging single-view problem into two subproblems: generative back-view hallucination and mesh reconstruction. This decomposition enables us to exploit the generative capability of pretrained diffusion models to guide full-body mesh and texture reconstruction. The workflow is depicted in Fig. 1. Given a front-view image, the first stage involves hallucinating a perceptually consistent back-view image using image-conditioned diffusion. The second stage reconstructs full-body mesh and texture, utilizing both the front and back-view images as guidance.

More specifically, we employ the generative capabilities of pretrained diffusion models (e.g. Stable Diffusion [57]) to infer unobserved back-view appearances for full-body 3D reconstruction. The primary challenge in ensuring the realism of 3D meshes lies in generating images that depict spatially aligned body shapes and perceptually consistent appearances with the input images. While diffusion models demonstrate impressive generative abilities with text conditioning, they are limited in producing desired back-view images using the frontal images as *image conditions*. To overcome this, we adapt the network architecture to enable conditioning on frontal images and introduce additional trainable components following ControlNet [77] to provide pose and mask control. To fully tailor this model to our task while retaining its original generative power, we carefully fine-tune the diffusion model using multi-view images rendered from 3D human scans. Complementing this generative model, we develop a mesh reconstruction module to recover full-body textured mesh from front and back-view images. We follow prior work in handling 3D ambiguity through normal [61] and skinned body [73, 79] guidance. It is worth noting that the models for both subproblems are trained using the same public THuman2.0 [75] dataset, which consists of as few as 500 scans.

To advance research in single-view human reconstruction, we created a new benchmark based on the high-quality CustomHumans [22] dataset and conducted comprehensive evaluations against state-of-the-art methods. Compared to existing end-to-end methods [4, 60, 79], our two-stage pipeline can recover full-body textured meshes, including back-view details, and demonstrates robustness to unseen images. In contrast to time-intensive diffusion-based optimization methods [24, 35, 41], our pipeline efficiently produces high-quality textured meshes in under two minutes. Moreover, we explored applications combining text-guided diffusion models, showing SiTH's versatility in 3D human creation. Our contributions are summarized as follows:

- We introduce SiTH, a single-view human reconstruction pipeline capable of producing high-quality, fully textured 3D human meshes within two minutes.
- Through decomposing the single-view reconstruction task, SiTH can be efficiently trained with public 3D human scans and is more robust to unseen images.
- We establish a new benchmark featuring more diverse subjects for evaluating textured human reconstruction.

2. Related Work

Single-view human mesh reconstruction. Reconstructing 3D humans from monocular inputs [14, 17, 26, 28, 29, 62, 67, 71] has gained more popularity in research. In this context, we focus on methods that recover 3D human shapes, garments, and textures from a single image. As a seminal work, Saito *et al*. [60] first proposed a data-driven method with pixel-aligned features and neural fields [72]. Its follow-up work PIFuHD [61] further improved this framework with high-res normal guidance. Later approaches extended this framework with additional human body priors. For instance, PaMIR [79] and ICON [73] utilized skinned body models [42, 51] to guide 3D reconstruction. ARCH [25], ARCH++ [21], and CAR [39] transformed global coordinates into the canonical coordinates to allow for reposing. PHOHRUM [4] and S3F [12] further disentangled shading and albedo to enable relighting. Another line of work replaced the neural representations with conventional Poisson surface reconstruction [33, 34]. ECON [74] and 2K2K [18] trained normal and depth predictors to generate front and back 2.5D point clouds. The human mesh is obtained by fusing these point clouds with body priors and 3D heuristics. However, none of these methods produce realistic full-body texture and geometry in the unobserved regions. Our pipeline addresses this problem by incorporating a generative diffusion model into the 3D human reconstruction workflow.

3D generation with 2D diffusion models. Diffusion models [52, 56, 57, 59] trained with large collections of images have demonstrated unprecedented capability in creating 3D objects from text prompts. Most prior work [11, 19, 40, 47, 53, 68] followed an optimization workflow to update 3D representations (e.g. NeRF [48], SDF tetrahedron [63]) via neural rendering [66] and a score distillation sampling (SDS) [53] loss. While some methods [3, 10, 24, 27, 64, 76] applied this workflow to human bodies, they cannot produce accurate human bodies and appearances due to the ambiguity of text-conditioning. More recent work [41, 54] also tried to extend this workflow with more accurate image-conditioning. However, we show that they struggle to recover human clothing details and require a long optimization time. Most related to our work is Chupa [35], which also decomposes its pipeline into two stages. Note that Chupa is an optimization-based approach that relies on texts

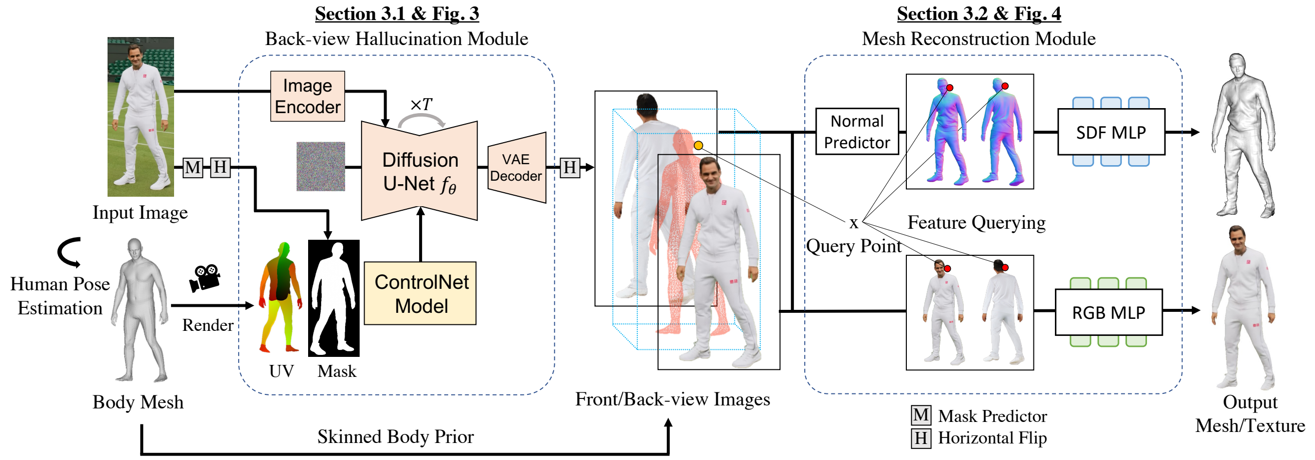

Figure 2. **Method overview**. SiTH is a two-stage pipeline composed of back-view hallucination and mesh reconstruction. The back-view hallucination module samples perceptually consistent back-view images through an iterative denoising process conditioned on the input image, UV map, and silhouette mask (Sec. 3.1). Based on the input and generated back-view images, the mesh reconstruction module recovers a full-body mesh and textures leveraging a skinned body prior as guidance (Sec. 3.2). Note that both modules in the pipeline can be trained with the same public 3D human dataset and generalize unseen images.

and cannot model colors. We address these issues by introducing an image-conditioning strategy and model. Most importantly, our method swiftly reconstructs full-texture human meshes without any optimization process.

Diffusion models adaptation. Foundation models [8, 13, 20, 36] trained on large-scale datasets have been shown to be adaptable to various downstream tasks. Following this trend, pretrained diffusion models [52, 56, 57, 59] have become common backbones for generative modeling. For instance, they can be customized by finetuning with a small collection of images [23, 37, 58]. ControlNet [77] introduced additional trainable plugins to enable image conditioning such as body skeletons. While these strategies have been widely adopted, none of them directly fit our objective. More relevant to our task is DreamPose [31], which utilizes DensePose [16] images as conditions to repose input images. However, it cannot handle out-of-distribution images due to overfitting. Similarly, Zero-1-to-3 [41] finetunes a diffusion model with multi-view images to allow for viewpoint control. However, we show that viewpoint conditioning is not sufficient for generating consistent human bodies. Our model addresses this issue by providing accurate body pose and mask conditions for back-view hallucination.

3. Methodology

Method overview. Given an input image of a human body and estimated SMPL-X [51] parameters, SiTH produces a full-body textured mesh. This mesh not only captures the observed appearances but also recovers geometric and textural details in unseen regions, such as clothing wrinkles on the back. The pipeline is composed of two modules and is summarized in Fig. 2. In the first stage, we hallucinate unobserved appearances leveraging the generative power of an image-conditioned diffusion model (Sec. 3.1). In the second stage, we reconstruct a full-body textured mesh given the input front-view image and the generated back-view image as guidance (Sec. 3.2). Notably, both modules are efficiently trained with 500 textured human scans in THuman2.0 [75].

3.1. Back-view Hallucination

Preliminaries. Given an input front-view image $I^F \in \mathbb{R}^{H\times W\times 3}$, our goal is to infer a back-view image $I^B \in \mathbb{R}^{H\times W\times 3}$ which depicts unobserved body appearances. This task is under-constrained since there are multiple possible solutions to the same input images. Taking this perspective into account, we leverage a **latent diffusion model (LDM)** [57] to learn a conditional distribution of back-view images given a front-view image. First, a VAE autoencoder, consisting of an encoder $\mathcal{E}$ and a decoder $\mathcal{D}$, is pretrained on a corpus of 2D natural images through image reconstruction, i.e. $\tilde{I} = \mathcal{D}(\mathcal{E}(I))$. Afterwards, an LDM learns to produce a latent code z within the VAE latent distribution $z = \mathcal{E}(I)$ from randomly sampled noise. To sample an image, a latent code $\tilde{z}$ is obtained by iteratively denoising Gaussian noise. The final image is reconstructed through the decoder, i.e., $\tilde{I} = \mathcal{D}(\tilde{z})$.

Image-conditioned diffusion model. Simply applying the LDM architecture to our task is not sufficient since our goal is to learn a conditional distribution of back-view images given an input conditional image. To this end, we make several adaptations to allow for image-conditioning as shown in Fig. 3. First, we utilize the pretrained CLIP [55] image encoder and VAE encoder $\mathcal{E}$ to extract image features from the front-view image (i.e., I^F). These image features are used for conditioning the LDM, ensuring the output image shares a consistent appearance with the input image. Second, we follow the idea of ControlNet [77] and propose to use a UV map ($I^B_{UV} \in \mathbb{R}^{H\times W\times 3}$) and a silhou-

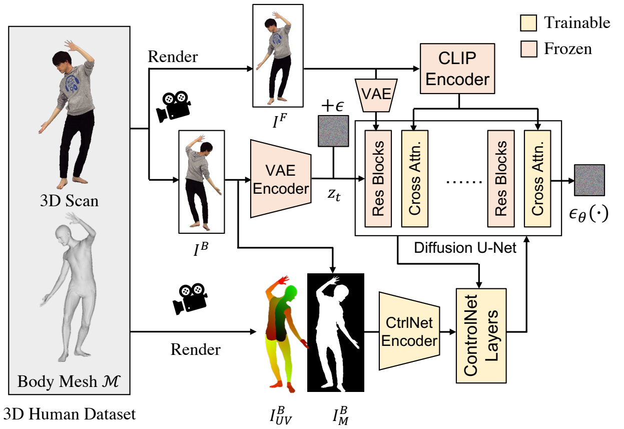

Figure 3. **Training of back-view hallucination module**. We employ a pretrained LDM and ControlNet architecture to enable image conditioning. To train our model, we render training pairs of conditional images I^F and ground-truth images I^B from 3D human scans. Given a noisy image latent z_t, the model predicts added noise ϵ given the conditional image I^F, UV map I^B_{UV}, and mask I^B_M as conditions. We train the ControlNet model and cross-attention layers while keeping other parameters frozen.

ette mask ($I^B_M \in \mathbb{R}^{H\times W}$) from the back view as additional conditions. These conditional signals provide additional information that ensures the output image has a similar body shape and pose to the conditional input image.

Learning hallucination from pretraining. Another challenge in training an image-conditioned LDM is data. Training the model from scratch is infeasible due to the requirement of a large number of paired images rendered from 3D textured human scans. Inspired by the concept of learning from large-scale pretraining [13, 20], we build our image-conditioned LDM on top of a pretrained diffusion U-Net [57]. We utilize the finetuning strategy [37, 77] to optimize cross-attention layers and ControlNet parameters while keeping most of the other parameters frozen (see Fig. 3). The design and training strategy of our image-conditioned diffusion model enables hallucinating plausible back-view images that are *cosistent* with the frontal inputs.

Training and inference. To generate pairwise training images from 3D human scans, we sample camera view angles and use orthographic projection to render RGBA images from 3D scans and UV maps from their SMPL-X fits. Given a pair of images rendered by a frontal and its corresponding back camera, the first image serves as the conditional input I^F while the other one is the ground-truth image I^B. During training, the ground-truth latent code $z_0 = \mathcal{E}(I^B)$ is perturbed by the diffusion process in t time steps, resulting in a noisy latent z_t. The image-conditoned LDM model ϵ_θ aims to predict the added noise ϵ given the noisy latent z_t, the time step $t \sim [0, 1000]$, the conditional image I^F, the silhouette mask I^B_M, and the UV map I^B_{UV} (See Fig. 3). The objective function for fine-tuning can be represented as:

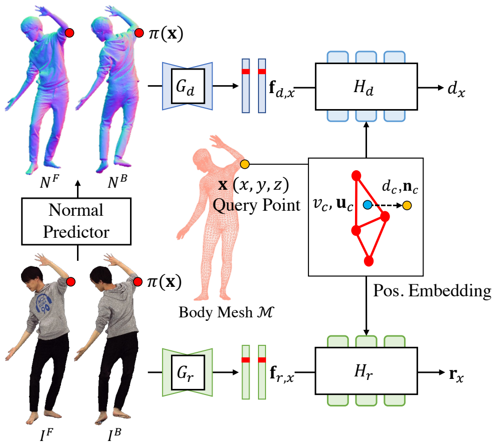

Figure 4. **Mesh reconstruction module**. Given front and back-view images (I^F, I^B) we predict their normal images (N^F, N^B) through a learned normal predictor. A 3D point $\mathbf{x}$ is projected onto these images for querying pixel-aligned features ($\mathbf{f}_{d,x}, \mathbf{f}_{r,x}$). To leverage human body mesh as guidance, we embed the point $\mathbf{x}$ into the local UV coordinates $\mathbf{u}_c$, vector $\mathbf{n_c}$, distance d_c, and visibility v_c. Finally, two decoders (H_d, H_r) predict SDF and RGB values at $\mathbf{x}$ given the positional embedding and pixel-aligned features.

$$\min_\theta \mathbb{E}_{z\sim\mathcal{E}(I),t,\epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I})} \left\|\epsilon - \epsilon_\theta(z_t, t, I^F, I^B_{UV}, I^B_M)\right\|_2^2 . \tag{1}$$

At test time, we obtain I^B_{UV}, I^B_M from an off-the-shelf pose predictor [9] and segmentation model [36]. To infer a back-view image, we sample a latent $\tilde{z}_0$ by performing the iterative denoising process starting from a Gaussian noise $z_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The back-view image can obtained by:

$$\tilde{I}^B = \mathcal{D}(\tilde{z}_0) = \mathcal{D}(f_\theta(z_T, I^F, I^B_{UV}, I^B_M)), \tag{2}$$

where f_θ is a function representing the iterative denoising process of our image-conditioned LDM (See Fig. 2 left).

3.2. Human Mesh Reconstruction

After obtaining the back-view image, our goal is to construct a full-body human mesh and its textures using the input and back-view image as guidance. We follow the literature [60, 61] to model this task with a data-driven method. Given pairwise training data (i.e., front/back-view images and 3D scans), we learn a data-driven model that maps these images to a 3D representation (e.g., a signed distance field (SDF)). We define this mapping as below:

$$\begin{aligned} \mathbf{\Phi} : \mathbb{R}^{H\times W\times 3} \times \mathbb{R}^{H\times W\times 3} \times \mathbb{R}^3 &\rightarrow \mathbb{R} \times \mathbb{R}^3 \\ (I^F, I^B, \mathbf{x}) &\mapsto d_x, \mathbf{r}_x, \end{aligned} \tag{3}$$

where $\mathbf{x}$ is the 3D coordinate of a query point, and $d_x, \mathbf{r}_x$ denote the signed distance and RGB color value at point $\mathbf{x}$. The network components we used for learning the mapping function are depicted in Fig. 4.

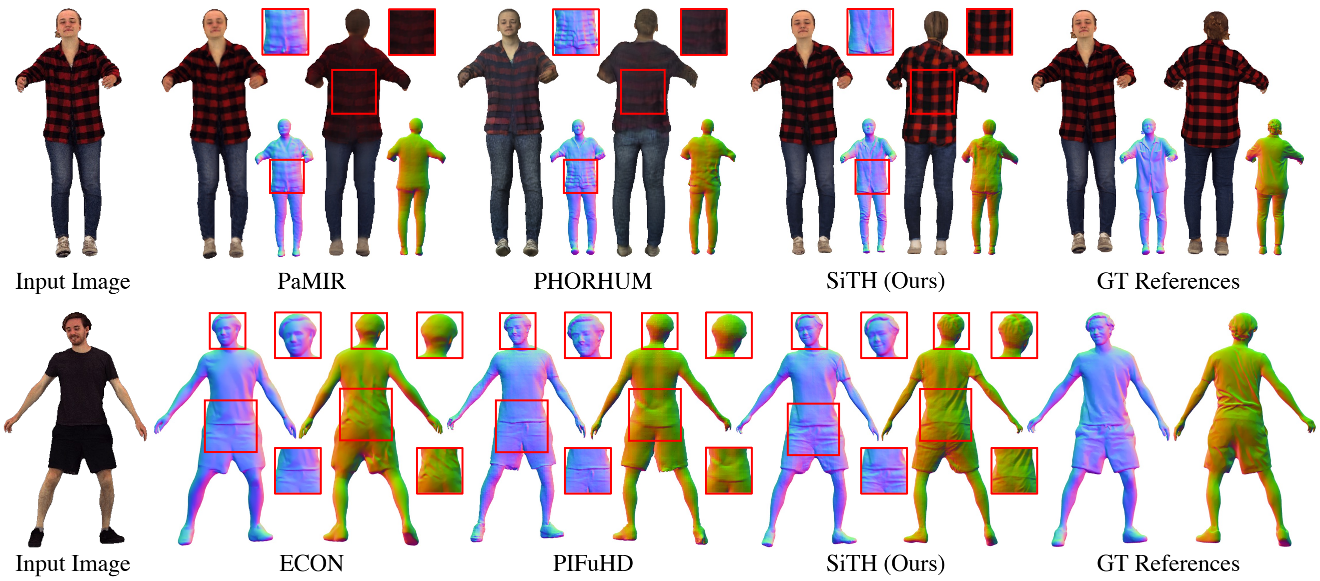

Figure 5. **Qualitative comparison on CustomHumans**. *Top*: Results of methods generating mesh and texture. *Bottom*: Results of methods generating mesh only. Note that single-view reconstruction is not possible to replicate exact back-view texture and geometry. Our method generates realistic texture and clothing wrinkles perceptually close to the real scans while other baselines only produce smooth colors and surfaces in the back regions. Best viewed in color and zoom in.

Local feature querying. To learn a generic mapping function that is robust to unseen images, it is important that the model is conditioned solely on local image information with respect to the position of $\mathbf{x}$. Therefore, we employ the idea of pixel-aligned feature querying [60, 61] and separate our model into two branches, i.e., color and geometry. Our model contains a normal predictor that converts the RGB image pair (I^F, I^B) into normal maps (N^F, N^B). Two image feature encoders G_d, G_r then extract color and geometry feature maps $(\mathbf{f}_d, \mathbf{f}_r) \in \mathbb{R}^{H' \times W' \times D}$ from the images and normal maps respectively (for simplicity we describe the process for a single image and leave out the superscripts, but both front and back images are treated the same). Finally, we project the query point $\mathbf{x}$ onto the image coordinate (Fig. 4 red points) to retrieve the local features $(\mathbf{f}_{d,x}, \mathbf{f}_{r,x}) \in \mathbb{R}^D$:

$$\begin{aligned} \mathbf{f}_{d,x} &= \mathcal{B}(\mathbf{f}_d, \pi(\mathbf{x})) = \mathcal{B}(G_d(N), \pi(\mathbf{x})), \\ \mathbf{f}_{r,x} &= \mathcal{B}(\mathbf{f}_r, \pi(\mathbf{x})) = \mathcal{B}(G_r(I), \pi(\mathbf{x})), \end{aligned} \tag{4}$$

where $\mathcal{B}$ is a local feature querying operation using bilinear interpolation and $\pi(\cdot)$ denotes orthographic projection.

Local positional embedding with skinned body prior. As mentioned in Sec. 1, a major difficulty in mesh reconstruction is 3D ambiguity where a model has to infer unknown depth information between the front and back images. To address this issue, we follow prior work [22, 73, 79] leveraging a skinned body mesh [51] for guiding the reconstruction task. This body mesh is regarded as an anchor that provides an approximate 3D shape of the human body.

To exploit this body prior, we devise a local positional embedding function that transforms the query point $\mathbf{x}$ into the local body mesh coordinate system. We look for the closest point $\mathbf{x}_c^*$ on the body mesh (Fig. 4 blue point), i.e.,

$$\mathbf{x}_c^* = \arg\min_{\mathbf{x}_c} \|\mathbf{x} - \mathbf{x}_c\|_2, \tag{5}$$

where $\mathbf{x}_c$ are points on the skinned body mesh $\mathcal{M}$. Our positional embedding $\mathbf{p}$ constitutes four elements: a signed distance value d_c between $\mathbf{x}_c^*$ and $\mathbf{x}$, a vector $\mathbf{n}_c = (\mathbf{x} - \mathbf{x}_c^*)$, the UV coordinates $\mathbf{u}_c \in [0, 1]^2$ of the point $\mathbf{x}_c^*$, and a visibility label $v_c \in \{1, -1, 0\}$ that indicates whether $\mathbf{x}_c^*$ is visible in the front/back image or neither. Finally, two separate MLPs H_d, H_r take the positional embedding $\mathbf{p} = [d_c, \mathbf{n}_c, \mathbf{u}_c, v_c]$ and the local texture/geometry features $(\mathbf{f}_{d,x}, \mathbf{f}_{r,x})$ as inputs to predict the final SDF and RGB values at point $\mathbf{x}$:

$$\begin{aligned} d_x &= H_d(\mathbf{f}_{d,x}^F, \mathbf{f}_{d,x}^B, \mathbf{p}), \\ \mathbf{r}_x &= H_r(\mathbf{f}_{r,x}^F, \mathbf{f}_{r,x}^B, \mathbf{p}). \end{aligned} \tag{6}$$

Training and inference. We used the same 3D dataset described in Sec. 3.1 to render training image pairs (I^F, I^B) from the 3D textured scans. For each training scan, query points $\mathbf{x}$ are sampled within a 3-dimensional cube $[-1, 1]^3$. For each point, we compute the ground-truth signed distance values d to the scan surface, closest texture RGB values r, and surface normal $\mathbf{n}$. Finally, we jointly optimized the normal predictors, the image encoders, and the MLPs in

	CAPE [45]				CustomHuman [22]			
Method	CD: P-to-S / S-to-P (cm)↓	NC↑	f-Score↑	LPIPS: F $(\times 10^{-2})\downarrow$	CD: P-to-S / S-to-P (cm)↓	NC↑	f-Score↑	LPIPS: F / B $(\times 10^{-2})\downarrow$
PIFu [60]	2.368 / 3.763	0.778	33.842	2.720	2.209 / 2.582	0.805	34.881	6.073 / 8.496
PIFuHD [61]	2.401 / 3.522	0.772	35.706	-	2.107 / 2.228	0.804	**39.076**	-
PaMIR [79]	2.190 / 2.806	0.804	36.725	2.085	2.181 / 2.507	0.813	35.847	4.646 / 7.152
2K2K [18]	2.478 / 3.683	0.782	28.700	-	2.488 / 3.292	0.796	30.186	-
FOF [15]	2.196 / 4.040	0.777	34.227	-	2.079 / 2.644	0.808	36.013	-
ICON [73]	2.516 / 3.079	0.786	29.630	-	2.256 / 2.795	0.791	30.437	-
ECON [74]	2.475 / 2.970	0.788	30.488	-	2.483 / 2.680	0.797	30.894	-
SiTH (Ours)	**1.899 / 2.261**	**0.816**	**37.763**	**1.977**	**1.871 / 2.045**	**0.826**	37.029	**3.929 / 6.803**

Table 1. **Single-view human reconstruction benchmarks**. We report Chamfer distance (CD), normal consistency (NC), and f-score between ground truth and predicted meshes. To evaluate texture reconstruction quality, we compute LPIPS between the image rendering of GT and generated textures. **The best** and the second best methods are highlighted in bold and underlined respectively. Note that gray color denotes models trained on more commercial 3D human scans while the others are trained on the public THuman2.0 dataset.

both branches with the following reconstruction losses:

$$\mathcal{L}_d = \|d - d_x\|_1 + \lambda_n(1 - \mathbf{n} \cdot \nabla_{\mathbf{x}} d_x), \tag{7}$$

$$\mathcal{L}_r = \|\mathbf{r} - \mathbf{r}_x\|_1. \tag{8}$$

Note that $\nabla_{\mathbf{x}}$ indicates numerical finite differences for computing local normals at point $\mathbf{x}$ and λ_n is a hyperparameter.

During inference, we use the input image I^F and the back-view image $\tilde{I}^B$ obtained from Sec. 3.1 to reconstruct 3D mesh and textures. First, we align both images with the estimated body mesh $\mathcal{M}$ to ensure that image features can be properly queried around the 3D anchor. We adopt a similar strategy of SMPLify [7] to optimize the scale and the offset of the body mesh with silhouette and 2D joint errors. Finally, we perform the marching cube algorithm [43] by querying SDF and RGB values within a dense voxel grid via Eq. (3) (see Fig. 2 right).

4. Experiments

4.1. Experimental Setup

Dataset. Previous work relied on training data from commercial datasets such as RenderPeople [1]. While these datasets offer high-quality textured meshes, they also limit reproducibility due to limited accessibility. For fair comparisons, we follow ICON [73] by training our method on the public 3D dataset THuman2.0 [75] and using the CAPE [45] dataset for evaluation. However, we observed potential biases in the evaluation due to the low-res ground-truth meshes and image rendering defects in the CAPE dataset (for a detailed discussion, please refer to Supp-Sec. 6). Consequently, we further create a new benchmark that evaluates the baselines on a higher-quality 3D human dataset CustomHumans [22]. In the following, we provide a summary of the datasets used in our experiments:

- **THuman2.0** [75] contains approximately 500 scans of humans wearing 150 different garments in various poses. We use these 3D scans as the training data.
- **CAPE** [45] contains 15 subjects in 8 types of tight outfits. The test set, provided by ICON, consists of 100 meshes. We use CAPE for the quantitative evaluation (Sec. 4.2).
- **CustomHumans** [22] contains 600 higher-quality scans of 80 subjects in 120 different garments and varied poses. We selected 60 subjects for all quantitative experiments, user studies, and ablation studies. (Sec. 4.2 - Sec. 4.4)

Evaluation protocol. We follow the evaluation protocol in OccNet [46] and ICON [73] to compute 3D metrics Chamfer distance **(CD)**, normal consistency **(NC)**, and **f-Score** [65] on the generated meshes. To evaluate reconstructed mesh texture, we report **LPIPS** [78] of front and back texture rendering. In user studies, 30 participants rank the meshes obtained by four different methods. We report the **average ranking** ranging from 1 (best) to 4 (worst).

4.2. Single-view Human Reconstruction

Benchmark evaluation. We compared SiTH with state-of-the-art single-view human reconstruction methods, including **PIFu** [60], **PIFuHD** [61], **PaMIR** [79], **FOF** [15], **ICON** [73], **PHORHUM** [4], **2K2K** [18], and **ECON** [74] on CAPE and CustomHumans. Note that PHORHUM is only used for qualitative comparison since a different camera system is used, leading to the misalignment with ground-truth meshes. We visualize the generated mesh texture and normals in Fig. 5. Existing methods produce over-smoothed texture and normals, particularly in the back. Our method not only generates photorealistic and perceptually consistent appearances in unobserved regions but also recovers underlying geometric details like clothing wrinkles.

The quantitative results are summarized in Tab. 1. It's worth noting that most methods are trained with commercial datasets (gray color in Tab. 1), while the others are trained on the public THuman2.0 dataset. To evaluate the methods leveraging a skinned body prior (i.e., PaMIR, ICON, ECON, FOF, and SiTH), we use the same pose alignment procedure in their original implementations for a fair

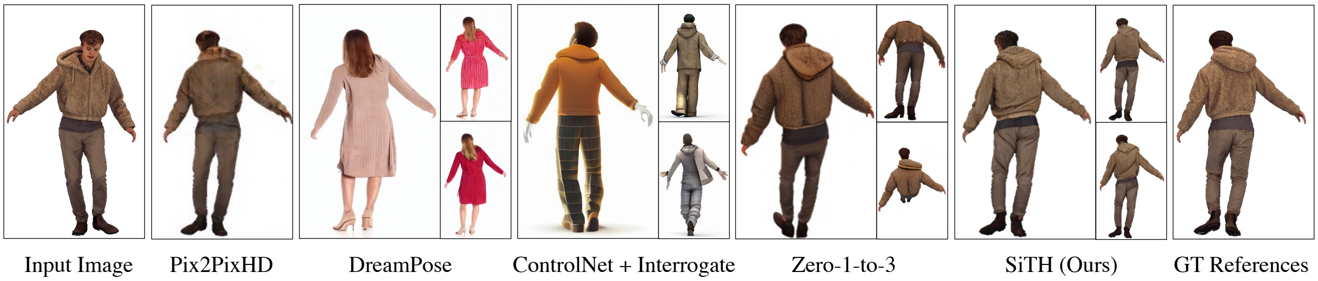

Figure 6. **Qualitative comparison of back-view hallucination**. We visualize back-view images generated by the baseline methods. Note that the three different images are sampled from different random seeds. Our results are perceptually close to the ground-truth image in terms of appearances and poses. Moreover, our method also preserves generative stochasticity for handling tiny wrinkle changes.

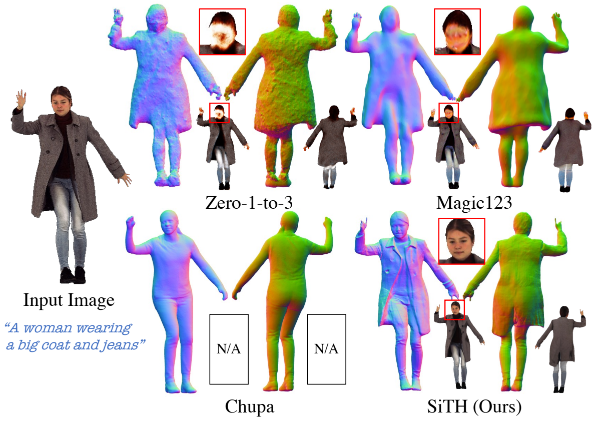

Figure 7. **Comparison with optimization-based methods**. Compared to methods that utilize diffusion models for optimization, our result is closer to the input image and contains local geometric details. Note that Chupa is not conditioned on the image, and fails to generate correct clothing from the text prompts.

comparison. Results in Tab. 1 show that the method using a body prior (PaMIR) outperformed the end-to-end method (PIFuHD) on tight clothing and challenging poses in CAPE. However, it falls short in handling diverse outfits in CustomHumans. Moreover, the methods trained on commercial datasets achieve better performance than those trained with public data (ICON, ECON). Notably, our method is robust across both benchmarks, achieving performance comparable to the methods trained on high-quality commercial data.

Compared with optimization-based methods. We compared SiTH with methods that use pretrained diffusion models and a score distillation sampling loss [53] to optimize 3D meshes. In the case of **Zero-1-to-3** [41], we used the input image to optimize an instant NGP [49] radiance field, and for **Magic-123** [54], we provided additional text prompts to optimize an SDF tetrahedron [63]. From Fig. 7, we see that while both methods can handle full-body textures, they struggle with reasoning the underlying geometry and clothing details. It is worth noting that Zero-1-to-3 and Magic-123 require 10 minutes and 6 hours in optimization, respectively, while our method takes under 2 minutes to generate a textured mesh with a marching cube of 512^3 resolution.

	ICON	ECON	PIFuHD	Ours
Front Geometry	3.127	2.720	2.607	**1.547**
Back Geometry	3.193	2.513	3.093	**1.200**
Similarity	3.093	2.660	2.780	**1.456**

	PIFu	PaMIR	PHOHRUM	Ours
Front Texture	3.067	2.153	3.450	**1.327**
Back Texture	3.450	2.355	3.140	**1.054**
Similarity	3.307	2.192	3.416	**1.093**

	Chupa	Ours
User Preference	36.0%	**64.0%**

Table 2. **User study results**. *Top*: 30 users are asked to rank the quality of surface normal images from best (1) to worst (4). We report the average ranking of each method. *Middle*: Similar to the first task, users are asked to rank the quality of RGB textures. *Bottom*: We ask users to choose the mesh with a better quality.

Additionally, more similar to our method is **Chupa** [35], which generates front/back-view normals for mesh reconstruction. Note that Chupa is not conditioned on images and does not generate texture. Instead, we provided body poses and text prompts generated by an image-to-text interrogator [38] as their conditional inputs. From Fig. 7, it's clear that text-conditioning is less accurate than image-conditioning, and the method struggles to generate unseen clothing styles such as coats. By contrast, our method can reconstruct correct clothing geometry and texture from unseen images. We present more discussions and comparisons with optimization-based methods in Supp-Sec. 9.1.

User study. The above metrics may not fully capture the quality of 3D meshes in terms of realism and local details. To address this, we conducted a user study to compare the texture and geometry quality among various baselines. We invited 30 users to rank the front/back-view texture and normal renderings of 3D meshes generated by four different methods. Additionally, we asked the users to assess the similarity between the input images and the generated meshes. The results (Tab. 2) support our claim that existing methods struggle to efficiently generate desirable back-view texture/geometry from single-view images. Our method,

Method	CD (cm)↓	NC↑	f-Score↑
W/o Body Mesh	2.471	0.801	33.244
W/o Hallucination	1.960	**0.840**	36.677
Full Pipeline	**1.958**	0.826	**37.029**
W/ GT Body Mesh	1.172	0.891	58.858
W/ GT Body and I^B	1.059	0.914	63.356

Table 3. **Ablation study on CustomHumans**. We ablate the hallucination module and the skinned body mesh in our pipeline. Please refer to our discussion in Sec. 4.4.

which leverages the generative capability of diffusion models, consistently outperforms each baseline. It also produces more preferred front-view textures and geometries, as evidenced by higher user rankings. We also conducted a user study with Chupa (in Tab. 2 *bottom*) which also indicates more users prefer the 3D meshes generated by our method.

4.3. Generative Capability

Image quality comparison. Our hallucination module is a unique and essential component that generates spatially aligned human images to guide 3D mesh reconstruction. Given that our focus is on back-view hallucination, we compare the quality of generated images with the relevant generative methods in Fig. 6. We trained a baseline **Pix2PixHD** [69] model, which produced smooth and blurry results on unseen images due to overfitting to 500 subjects. Another method closely related to ours is **DreamPose** [31], which conditions the model with DensePose images and finetunes the diffusion model with paired data. However, their model failed to handle unseen images, in contrast to our approach. While **Zero-1-to-3** [41] can generalize to unseen images, their method faces challenges in generating consistent body poses given the same back-view camera. Moreover, we designed another baseline that provides **ControlNet** [77] for corresponding text prompts using an image-to-text interrogator [38]. However, without proper image conditioning and fine-tuning, such a method cannot generate images that faithfully match the input appearances. Our method not only addresses these issues but also handles stochastic appearances (e.g., tiny differences in wrinkles) from different random seeds. We report 2D generative evaluation metrics and more results in Supp-Sec. 8.1.

4.4. Ablation Study

We conducted controlled experiments to validate the effectiveness of our proposed modules. As shown in Fig. 8, the skinned body mesh is a crucial component for 3D human reconstruction. Without this body mesh as a 3D anchor, the output mesh contains an incorrect body shape due to the depth ambiguity issue. Conversely, removing the hallucination module has minimal impact on 3D reconstruction metrics, though it slightly degrades normal consistency. However, the overall quality in both texture and geometry is incomparable with our full model (see Fig. 8 *right*). This is consistent with our findings in user studies, indicating that 3D metrics may not accurately reflect the perceptual quality of 3D meshes. Finally, we tested two additional variants, leveraging ground-truth body meshes and real back-view images in our full pipeline, representing the upper bound of our method. As shown in Tab. 3 *bottom*, this additional information notably improves the 3D metrics. These results highlight the persistent challenges in the single-view reconstruction problem, including pose ambiguity and the stochastic nature of clothing geometry. For more experiments on our design choices, please refer to Supp-Sec. 8.5.

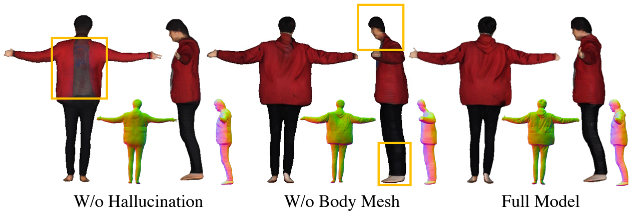

Figure 8. **Ablation study**. We visualize back and side-view rendering of the reconstructed meshes. Our full model produced a correct body shape and more realistic clothing geometry.

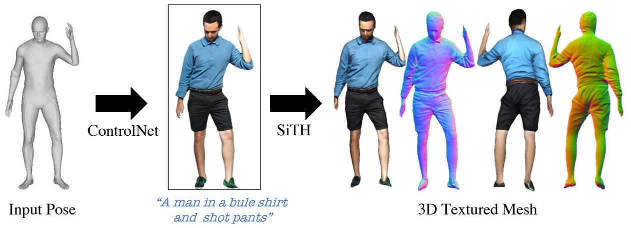

Figure 9. **Applications**. Our pipeline can be incorporated with text-to-image diffusion models for 3D human creation.

4.5. Applications

Inheriting the generative capability of LDM, SiTH is robust to diverse inputs, such as out-of-distribution or AI-generated images. We demonstrate a unique solution to link photo-realistic AI photos and high-fidelity 3D humans. In Fig. 9, we introduce a 3D creation workflow integrating powerful text-to-image generative models. Given a body pose, we generate a front-view image using Stable Diffusion and ControlNet using text prompts. SiTH then creates a full-body textured human from the AI-generated image.

5. Conclusion

We propose an innovative pipeline designed to create fully textured 3D humans from single-view images. Our approach seamlessly integrates an image-conditioned diffusion model into the existing data-driven 3D reconstruction workflow. Leveraging the generative capabilities of the diffusion model, our method efficiently produces lifelike 3D humans from a diverse range of unseen images in under two minutes. We expect our work will advance the application of generative AI in 3D human creation.

Acknowledgements. This work was partially supported by the Swiss SERI Consolidation Grant ”AI-PERCEIVE”.

References

[1] Renderpeople, https://renderpeople.com/. 6

[2] Stable diffusion image variations. huggingface.co/lambdalabs/stable-diffusion-image-conditioned. 1

[3] Badour AlBahar, Shunsuke Saito, Hung-Yu Tseng, Changil Kim, Johannes Kopf, and Jia-Bin Huang. Single-image 3d human digitization with shape-guided diffusion. In *SIGGRAPH Asia*, 2023. 2, 7

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

[5] Paul J Besl and Neil D McKay. Method for registration of 3-d shapes. In *Sensor fusion IV: control paradigms and data structures*, pages 586–606. Spie, 1992. 4

[6] Mikolaj Binkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD gans. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018. 2

[7] 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)*. Springer International Publishing, 2016. 6

[8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:1877–1901, 2020. 3

[9] Zhongang Cai, Wanqi Yin, Ailing Zeng, Chen Wei, Qingping Sun, Yanjun Wang, Hui En Pang, Haiyi Mei, Mingyuan Zhang, Lei Zhang, Chen Change Loy, Lei Yang, and Ziwei Liu. Smpler-x: Scaling up expressive human pose and shape estimation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 4

[10] Yukang Cao, Yan-Pei Cao, Kai Han, Ying Shan, and Kwan-Yee K Wong. Dreamavatar: Text-and-shape guided 3d human avatar generation via diffusion models. *arXiv preprint arXiv:2304.00916*, 2023. 2

[11] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2023. 2

[12] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3, 4

[14] Zijian Dong, Chen Guo, Jie Song, Xu Chen, Andreas Geiger, and Otmar Hilliges. Pina: Learning a personalized implicit neural avatar from a single rgb-d video sequence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[15] Qiao Feng, Yebin Liu, Yu-Kun Lai, Jingyu Yang, and Kun Li. Fof: Learning fourier occupancy field for monocular real-time human reconstruction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 6, 5

[16] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[17] Chen Guo, Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. Vid2avatar: 3d avatar reconstruction from videos in the wild via self-supervised scene decomposition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[18] Sang-Hun Han, Min-Gyu Park, Ju Hong Yoon, Ju-Mi Kang, Young-Jae Park, and Hae-Gon Jeon. High-fidelity 3d human digitization from single 2k resolution images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 6, 4

[19] Ayaan Haque, Matthew Tancik, Alexei Efros, Aleksander Holynski, and Angjoo Kanazawa. Instruct-nerf2nerf: Editing 3d scenes with instructions. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2023. 2

[20] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16000–16009, 2022. 3, 4

[21] Tong He, Yuanlu Xu, Shunsuke Saito, Stefano Soatto, and Tony Tung. Arch++: Animation-ready clothed human reconstruction revisited. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021. 2

[22] Hsuan-I Ho, Lixin Xue, Jie Song, and Otmar Hilliges. Learning locally editable virtual humans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 5, 6, 1

[23] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022. 3

[24] Yangyi Huang, Hongwei Yi, Yuliang Xiu, Tingting Liao, Jiaxiang Tang, Deng Cai, and Justus Thies. TeCH: Text-guided Reconstruction of Lifelike Clothed Humans. In *International Conference on 3D Vision (3DV)*, 2024. 2, 7

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

[26] Boyi Jiang, Yang Hong, Hujun Bao, and Juyong Zhang. Selfrecon: Self reconstruction your digital avatar from monocular video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[27] Ruixiang Jiang, Can Wang, Jingbo Zhang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. Avatarcraft: Transforming text into neural human avatars with parameterized shape and pose control. *arXiv preprint arXiv:2303.17606*, 2023. 2

[28] Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. Instantavatar: Learning avatars from monocular video in 60 seconds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[29] 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. 2

[30] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 694–711. Springer, 2016. 2

[31] Johanna Karras, Aleksander Holynski, Ting-Chun Wang, and Ira Kemelmacher-Shlizerman. Dreampose: Fashion image-to-video synthesis via stable diffusion. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2023. 3, 8, 2

[32] Manuel Kaufmann, Velko Vechev, and Dario Mylonopoulos. aitviewer, 2022. 4

[33] Michael Kazhdan and Hugues Hoppe. Screened poisson surface reconstruction. *ACM Transactions on Graphics (TOG)*, 32(3):1–13, 2013. 2

[34] Michael Kazhdan, Matthew Bolitho, and Hugues Hoppe. Poisson surface reconstruction. In *Proceedings of the fourth Eurographics symposium on Geometry processing*, 2006. 2

[35] Byungjun Kim, Patrick Kwon, Kwangho Lee, Myunggi Lee, Sookwan Han, Daesik Kim, and Hanbyul Joo. Chupa: Carving 3d clothed humans from skinned shape priors using 2d diffusion probabilistic models. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2023. 2, 7

[36] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 3, 4

[37] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3, 4

[38] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning (ICML)*, 2022. 7, 8

[39] Tingting Liao, Xiaomei Zhang, Yuliang Xiu, Hongwei Yi, Xudong Liu, Guo-Jun Qi, Yong Zhang, Xuan Wang, Xiangyu Zhu, and Zhen Lei. High-Fidelity Clothed Avatar Reconstruction from a Single Image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[40] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[41] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2023. 2, 3, 7, 8, 1

[42] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics (TOG)*, 34(6):248:1–248:16, 2015. 2

[43] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm, 1998. 6

[44] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. Mediapipe: A framework for perceiving and processing reality. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, 2019. 2

[45] Qianli Ma, Jinlong Yang, Anurag Ranjan, Sergi Pujades, Gerard Pons-Moll, Siyu Tang, and Michael J. Black. Learning to Dress 3D People in Generative Clothing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 6, 1, 5

[46] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 6

[47] Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. Latent-nerf for shape-guided generation of 3d shapes and textures. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[48] 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. 2

[49] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (TOG)*, 41(4):102:1–102:15, 2022. 7

[50] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2016. 2

[51] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 3, 5

[52] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 2, 3

[53] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023. 2, 7

[54] Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov, and Bernard Ghanem. Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. *arXiv preprint arXiv:2306.17843*, 2023. 2, 7

[55] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021. 3

[56] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 2, 3

[57] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 3, 4

[58] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[59] 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. *Advances in Neural Information Processing Systems (NeurIPS)*, 35, 2022. 2, 3

[60] 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 International Conference on Computer Vision (ICCV)*, 2019. 1, 2, 4, 5, 6

[61] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 2, 4, 5, 6

[62] Kaiyue Shen, Chen Guo, Manuel Kaufmann, Juan Zarate, Julien Valentin, Jie Song, and Otmar Hilliges. X-avatar: Expressive human avatars. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[63] Tianchang Shen, Jun Gao, Kangxue Yin, Ming-Yu Liu, and Sanja Fidler. Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2, 7

[64] David Svitov, Dmitrii Gudkov, Renat Bashirov, and Victor Lempitsky. Dinar: Diffusion inpainting of neural textures for one-shot human avatars. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2023. 2

[65] Maxim Tatarchenko, Stephan R Richter, René Ranftl, Zhuwen Li, Vladlen Koltun, and Thomas Brox. What do single-view 3d reconstruction networks learn? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 6

[66] Ayush Tewari, Justus Thies, Ben Mildenhall, Pratul Srinivasan, Edgar Tretschk, W Yifan, Christoph Lassner, Vincent Sitzmann, Ricardo Martin-Brualla, Stephen Lombardi, et al. Advances in neural rendering. In *Annual Conference of the European Association for Computer Graphics (EUROGRAPHICS)*, pages 703–735. Wiley Online Library, 2022. 2

[67] Yating Tian, Hongwen Zhang, Yebin Liu, and Limin Wang. Recovering 3d human mesh from monocular images: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2023. 2

[68] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[69] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 8, 2

[70] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers*, 2003. 2

[71] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16210–16220, 2022. 2

[72] Yiheng Xie, Towaki Takikawa, Shunsuke Saito, Or Litany, Shiqin Yan, Numair Khan, Federico Tombari, James Tompkin, Vincent Sitzmann, and Srinath Sridhar. Neural fields in visual computing and beyond. In *Annual Conference of the European Association for Computer Graphics (EUROGRAPHICS)*, pages 641–676. Wiley Online Library, 2022. 2

[73] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J. Black. ICON: Implicit Clothed humans Obtained from Normals. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2, 5, 6, 4

[74] 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*

Conference on Computer Vision and Pattern Recognition (CVPR), 2023. 2, 6, 4, 5

[75] Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 3, 6, 1

[76] Huichao Zhang, Bowen Chen, Hao Yang, Liao Qu, Xu Wang, Li Chen, Chao Long, Feida Zhu, Kang Du, and Min Zheng. Avatarverse: High-quality stable 3d avatar creation from text and pose. *arXiv preprint arXiv:2308.03610*, 2023. 2

[77] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2023. 2, 3, 4, 8, 1

[78] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 6, 2

[79] Zerong Zheng, Tao Yu, Yebin Liu, and Qionghai Dai. Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2021. 1, 2, 5, 6