HumanRef: Single Image to 3D Human Generation via Reference-Guided Diffusion

Jingbo Zhang
City University of Hong Kong
jbzhang6-c@my.cityu.edu.hk

Xiaoyu Li *
Tencent AI Lab
xliea@connect.ust.hk

Qi Zhang
Tencent AI Lab
nwpuqzhang@gmail.com

Yanpei Cao
Tencent AI Lab
caoyanpei@gmail.com

Ying Shan
Tencent AI Lab
yingsshan@tencent.com

Jing Liao *
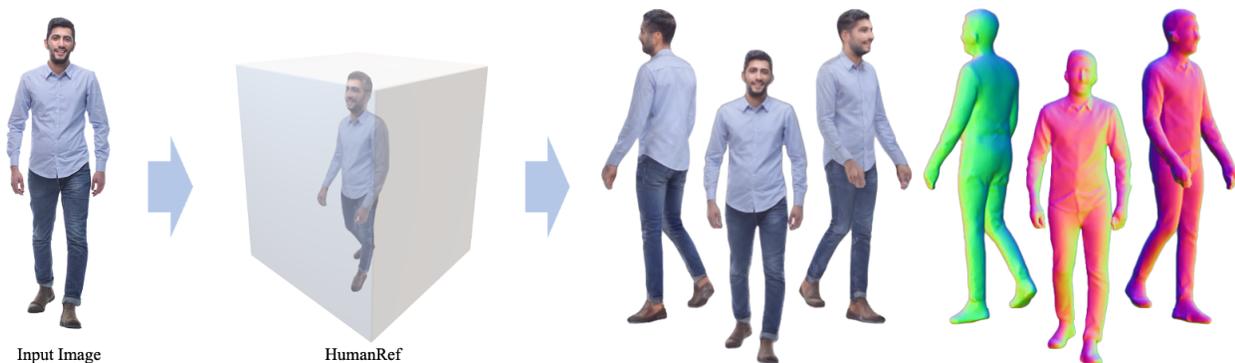City University of Hong Kong
jingliao@cityu.edu.hk

Figure 1. We propose HumanRef, a reference-guided 3D human generation framework. Our HumanRef is capable of generating 3D clothed human with realistic, view-consistent texture and geometry from a single image input.

Abstract

Generating a 3D human model from a single reference image is challenging because it requires inferring textures and geometries in invisible views while maintaining consistency with the reference image. Previous methods utilizing 3D generative models are limited by the availability of 3D training data. Optimization-based methods that lift text-to-image diffusion models to 3D generation often fail to preserve the texture details of the reference image, resulting in inconsistent appearances in different views. In this paper, we propose HumanRef, a 3D human generation framework from a single-view input. To ensure the generated 3D model is photorealistic and consistent with the input image, HumanRef introduces a novel method called reference-guided score distillation sampling (Ref-SDS), which effectively incorporates image guidance into the generation process. Furthermore, we introduce region-aware attention to Ref-SDS, ensuring accurate correspondence between different body regions. Experimental results demonstrate that HumanRef outperforms state-of-the-art methods in generating 3D clothed humans with fine geometry, photorealistic textures, and view-consistent appearances. Code and model are available at https://eckertzhang.github.io/HumanRef.github.io/.

1. Introduction

Clothed human reconstruction from single or multi-view images has received significant attention in the fields of computer vision and graphics due to its potential applications in virtual reality, movie industry, and immersive games [22, 71]. Unlike reconstruction using videos [1–3] or multi-view images [5, 6, 9, 54, 59], which allows for inferring 3D information from multi-view inputs, reconstructing a 3D human from a single view input is considerably more challenging. This task not only requires the reconstructed 3D clothed human to exhibit consistency in geometry and texture with the input view but also involves generating plausible geometry and texture that are not directly visible in the input. Therefore, compared to reconstruction, this task faces even greater challenges in human generation.

One approach to generating 3D human models from sin-

*Corresponding authors

gle images is by training a 3D generative model using 3D scanned human datasets. However, methods in this category [60, 61] tend to be more successful in generating geometry rather than texture, as textures are too diverse to be learned from limited 3D data. For example, ICON [61] and ECON [60] primarily estimate geometry, while PaMIR [72] and PIFu [49] tend to produce blurred textures when applied to in-the-wild data. Another approach to address the lack of 3D training data is to lift a 2D model pretrained on a large dataset to 3D. DreamFusion [42] is a pioneering work in this direction, leveraging the prior knowledge in a pretrained text-to-image diffusion model to supervise the optimization of 3D objects using a score distillation sampling (SDS) loss. Some follow-up works have extended this framework from text-to-3D generation to reference-guided 3D generation. This is achieved by extracting textual guidance from the reference image, either through textual inversion (e.g., RealFusion [35] and NeRDi [12]) or detailed text description using an image captioning model (e.g., Make-It-3D [53]). TeCH [22] combines both textual inversion and image captioning with 3D human priors to apply the SDS framework to 3D human generation from a single image.

Despite its success in 3D human generation, TeCH is limited by the SDS loss for two reasons. First, SDS is guided by texts, but text prompts or embeddings extracted from images can only represent the global semantic information of the reference images, which cannot capture the lower-level features necessary to provide detailed textures. As a result, TeCH's generated textures in invisible views often exhibit inconsistency with the reference image. Secondly, given the same text prompt, the diffusion model can generate images with a large diversity, making it difficult for the SDS loss optimization to converge. To address this issue, the SDS loss employs a large class-free guidance (CFG) scale in each diffusion denoising step, aiming to enhance the text relevance and reduce the diversity of image generation. However, while this approach ensures some stability in the 3D generation process, it also leads to generated texture over-saturation and over-smoothing, which is a well-known problem associated with the SDS loss. Although some efforts have been made to alleviate this problem, such as the variational score distillation (VSD) proposed by ProlificDreamer [58], these methods are primarily designed for text-to-3D generation and do not consider the reference image and human priors. Consequently, they are less optimal when applied to human generation from a single image.

In this paper, we present a novel framework called HumanRef for 3D clothed human generation from a reference image. Our approach utilizes a hash-encoded signed distance field (SDF) network for 3D representation and optimizes the SDF parameters from coarse to fine. We incorporate human geometry constraints and, most importantly, introduce a novel Reference-Guided Score Distillation Sam-
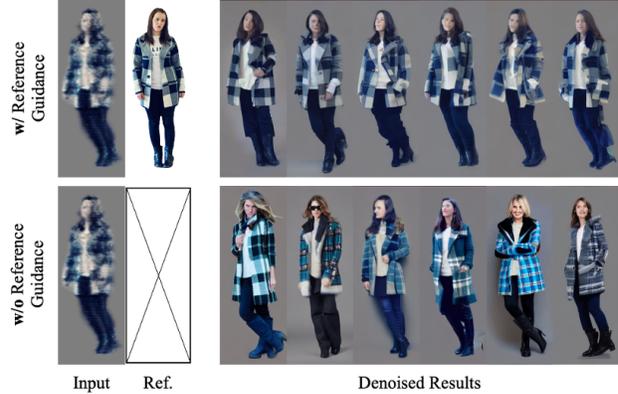

Figure 2. Showcasing the denoising process on initial input with and without reference image guidance. Starting with a coarse novel-view rendering image, we employ multiple rounds of random denoising using both reference-guided and text-guided diffusion models, highlighting the impact of reference guidance.

pling (Ref-SDS) loss in the optimization. Unlike the vanilla SDS loss, which is guided solely by text prompts, our Ref-SDS loss fully exploits the guidance of the reference image. We inject the reference image into the diffusion model to calculate the attentions between features of the reference and generated images at each denoising step. This process guides the diffusion model in generating results that better preserve the visual appearances of the reference image, as demonstrated in Figure 2. Consequently, our method is superior in generating view-consistent results matching the reference image. Moreover, by distilling less diverse images from the diffusion model, our Ref-SDS loss converges more easily, enabling us to reduce the CFG scale and generate more photo-realistic textures. To further incorporate human priors, we introduce Region-Aware Attentions for Ref-SDS. We employ human parsing to segment images into body regions, enhancing the attentions of corresponding regions (e.g., head to head) and suppressing the attention of non-corresponding regions when calculating the attentions between features of the reference and generated images. This refinement improves the precision of image guidance in our Ref-SDS. Thanks to these design choices, our HumanRef framework can generate 3D clothed humans with fine geometry, photorealistic textures, and view-consistent appearances from a single reference image.

To sum up, our contributions are three-fold as below:

- We propose HumanRef, a coarse-to-fine optimization framework, for 3D clothed human generation from a single image. It unifies the optimization process of geometry and appearance in a single SDF representation and does not introduce additional 3D representation and optimization stages.
- We propose a novel Reference-Guided Score Distillation Sampling (Ref-SDS) method for 3D generation. Ref-SDS injects image-level guidance into the denoising process of

a pretrained diffusion model, resulting in the production of more photorealistic and view-consistent 3D results.

- We introduce region-aware attention to Ref-SDS for 3D human generation, enhancing the precision of image guidance by ensuring accurate correspondence between different body regions.

2. Related Work

2.1. 3D Clothed Human Reconstruction

Unlike human pose and shape estimation methods [7, 16, 17, 27, 38, 40] that leverage a parametric body model for naked body reconstruction, clothed human reconstruction focuses on 3D humans with clothes, entailing more intricate details. This task has been explored in video [1–3] or multi-view settings [5, 6, 54, 59] for additional reconstruction constraints. However, the hardware requirements for additional inputs limit practical usage. Consequently, efforts have been made to recover 3D clothed humans from a single image [4, 11, 23, 30, 49, 50, 60, 61, 72]. Notably, PIFu [49] digitizes detailed clothed humans by inferring 3D geometry and texture from a single image, while PI-FuHD [50] introducing a coarse-to-fine framework for high-resolution geometry reconstruction. PaMIR [72] combines the parametric body model with a deep implicit function, while ICON [61] and ECON [60] recover fine geometry by inferring detailed clothed human normals. PHORHUM [4] and S3F [11] estimate albedo and shading information for relighting during reconstruction. Despite these advancements, such methods struggle to achieve clear and realistic textures, particularly for unseen areas in the input image. In this work, we leverage diffusion prior to synthesize high-quality, consistent textures for these invisible areas.

2.2. Diffusion Models

Diffusion models [21, 51] are latent-variable generative models that have garnered significant attention due to their impressive generation results. It consists of a forward process that slowly removes structure from data by adding noise and a reverse process or generative model that slowly adds structure from noise. To improve the performance of diffusion models, denoising diffusion implicit models [52] propose to use non-Markovian diffusion processes to reduce the generation steps and [13] proposes classifier guidance to improve the sample quality using a classifier to trade off diversity for fidelity. While [20] introduces classifier-free guidance by mixing the score estimates of a conditional diffusion model and a jointly trained unconditional diffusion model. Benefited from the scalability of the diffusion models and large-scale aligned image-text datasets, text-to-image has made great progress such as Glide [37], DALL-E 2 [46], Imagen [48] and StableDiffusion [47]. These pretrained diffusion models have been used as a diffusion prior

to promote the development of many other tasks like image or 3D editing and generation [8, 24, 66, 73].

2.3. 3D Generation Using 2D Diffusion

With the reduced dependence on 3D data, the recent development of applying pretrained 2D text-to-image diffusion models for 3D generation has significantly progressed after the pioneer works DreamFusion [42] and SJC [55]. The key technique is the Score Distillation Sampling (SDS) method proposed in DreamFusion which enables to use 2D diffusion models with score functions to optimize a 3D representation. Subsequently, numerous works [10, 31, 36, 58, 69, 73] have improved text input generation results, while another research line [12, 35, 43, 53, 62] focuses on 3D object reconstruction with a reference image. Compared to text-to-3D methods, image-to-3D requires generation results to closely resemble reference images. RealFusion [35] and NeRDi [12] extract text embedding from input images to provide additional visual cues to diffusion models, while Make-It-3D [53] uses an image captioning model for detailed text descriptions. However, both methods lack lower-level image features for detailed textures. The concurrent work, TeCH [22], uses both detailed text descriptions and text embedding for SDS-based 3D human generation from a single image but struggles with texture inconsistency. In contrast, we propose Ref-SDS to incorporate region-aware image guidance into the attention network of diffusion models, enabling more precise control over detailed textures.

3. Method

We introduce HumanRef, a unified coarse-to-fine optimization framework for 3D clothed human generation from a single image, as shown in Fig. 3. Given an input image, we initially extract its text caption, SMPL-X body [41], front and back normal maps, and silhouette using estimators. A neural SDF network, initialized with the estimated SMPL-X body, is then employed for optimization-based generation. To maintain appearance and pose consistency, we use the input image, silhouette, and normal maps as optimization constraints. For invisible regions, we introduce Ref-SDS, a method that distills realistic textures from a pretrained diffusion model, yielding sharp, realistic 3D clothed humans that align with the input image.

3.1. Image Preprocessing

Creating a 3D human from a single image is inherently ill-posed, requiring effective regularization to constrain the outcomes. To tackle this challenge, we extract multiple annotations for optimization, serving as valuable constraints to enhance the accuracy of the results.

Image Segmentation. Given an input image I, we first employ Mask-RCNN [19] for background segmentation, deducing the human silhouette S_I. Then, we use a human
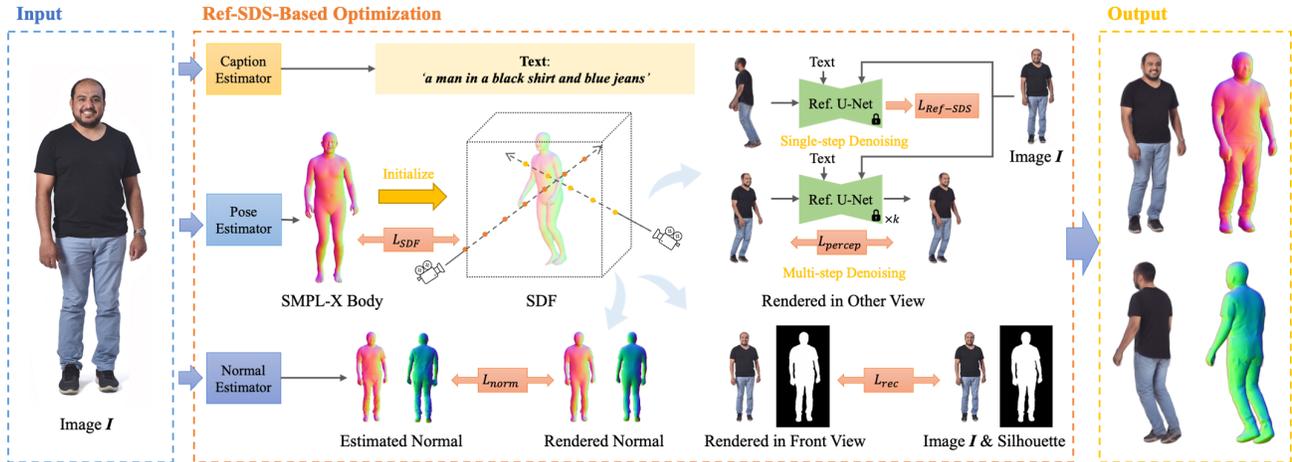
Figure 3. Overview of our proposed HumanRef for 3D clothed human generation from a single input image.

parsing method [29] to infer attribute masks for regions such as the head, coat, pants, etc. This enables us to divide the target human into four regions: head, upper body, lower body, and feet, each represented by their respective masks $\{M_j\}$. These regional masks will be used to instruct subsequent region-aware reference-guided generation.

Image Captioning. Similar to previous SDS-based 3D generation methods, our method is also based on a pretrained text-to-image diffusion model where text input is required. Thus, we adopt an off-the-shelf image captioning method [28] to estimate a rough text description.

SMPL-X Fitting. To capture the human pose and coarse shape, we employ a human pose estimation algorithm [41] to infer the corresponding SMPL-X body mesh from the image. We then convert this body mesh into an SDF representation and use it to initialize our SDF Network.

Normal Estimation. We employ the normal estimator of ECON [60] to predict the front and back normal maps from the image I. Such normal maps will provide effective geometric priors for our optimization.

3.2. Hash-Encoded SDF Representation

We adopt the hash-encoded SDF network [56] as the 3D representation of the human, with one sub-network f_s to predict the signed distance value s of a spatial query point $\mathbf{x}$ and another sub-network f_c to predict its color $\mathbf{c}$:

$$s(\mathbf{x}) = f_s(\mathbf{x}, h(\mathbf{x})), \quad \mathbf{c}(\mathbf{x}) = f_c(\mathbf{x}, h(\mathbf{x})), \quad (1)$$

where $h(\mathbf{x})$ is the feature queried from the hash grids with multi-level resolutions. Furthermore, we employ a cumulative distribution function [64] to model the density σ from the predicted s:

$$\sigma(\mathbf{x}) = \frac{\alpha}{2}\left(1 + sign(s(\mathbf{x}))\exp\left(-\alpha|s(\mathbf{x})|\right)\right), \quad (2)$$

where α is a learnable parameter. $sign(s)$ indicates the sign of the distance value s. Then, we can calculate the rendered

pixel color $\mathbf{C}$, normal $\mathbf{N}$, and silhouette S via the volume rendering method:

$$\Psi(\mathbf{r}) = \int_{t_n}^{t_f} T(t)\sigma(\mathbf{r}(t))\psi(\mathbf{r}(t))\,dt, \quad (3)$$

where Ψ indicates one of $\mathbf{C}$, $\mathbf{N}$, and S. $\psi(\mathbf{r}(t))$ is $\mathbf{c}(\mathbf{r}(t))$, $\mathbf{n}(\mathbf{r}(t))$, and 1 for $\mathbf{C}$, $\mathbf{N}$, and S, respectively. $\mathbf{n}(\mathbf{x}) = \nabla_{\mathbf{x}}s(\mathbf{x})$ indicates the predicted normal of query point $\mathbf{x}$. $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$ represents the coordinate of the sampled point on the camera ray emitted from the pixel center $\mathbf{o}$ with the direction $\mathbf{d}$. t_n and t_f are near and far bounds of the ray, respectively. $T(t) = \exp\left(-\int_{t_n}^{t}\sigma(\mathbf{r}(\tau))\,d\tau\right)$ is the accumulated transmittance along the ray. Note that, unlike previous volume rendering methods, we adopt orthographic rendering instead of perspective rendering. That is why $\mathbf{o}$ indicates the pixel center instead of the camera center.

3.3. Reference-Guided Optimization

In our unified reference-guided generation framework, we employ a coarse-to-fine optimization strategy. Initially, we optimize at a low rendering resolution (64×64) for quick convergence to the target space. Subsequently, we progressively increase the rendering resolution up to 512×512, enabling the model to refine geometry and texture. Unlike previous SDS- or VSD-based methods [42, 45, 58], we propose a modified Ref-SDS method to introduce reference guidance for the 3D generation (Sec. 3.3.1), and region-aware attention for precise local-region guidance in Ref-SDS (3.3.2). Besides, we also employ some constraints to align our generated 3D human with the input I (Sec. 3.3.3).

3.3.1 Reference-Guided Score Distillation Sampling

Inspired by [8], we modify the vanilla SDS by injecting region-aware image guidance into the attention network of
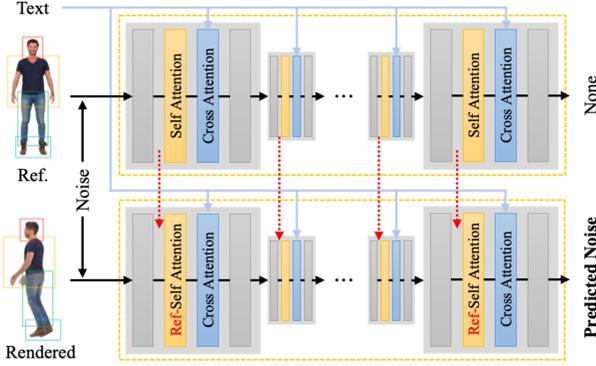
Figure 4. The framework of our reference U-Net.

Figure 5. Example of Region-Aware Attention.

U-Net during the single-step denoising process. In the implementation of vanilla SDS, a noisy rendered image, as well as random time step t and text prompt, are fed into the diffusion U-Net to predict corresponding noise. Then, the weighted noise residual between predicted and added noises is regarded as the SDS loss that backpropagates to the 3D representation through the rendered image. In fact, as shown in the upper branch of Fig. 4, the noisy image will undergo several transformer blocks and up/downsampling processes in the denoising U-Net. In each transformer block, the latent feature derived from the input is first subjected to self-attention operation and then cross-attention operation with text and time embedding, thereby realizing the text-guided denoising process. Note that we do not draw the data flow of time embedding in the figure because we do not make any changes to this part. To achieve the reference image guidance, as shown in Fig. 4, we first perform the diffusion and denoising process normally on the reference image and save the latent features p_i^{ref} before each self-attention operation. Subsequently, we feed the noisy rendered image to the same U-Net to perform denoising. Instead of undergoing the self-attention and cross-attention, we concatenate the latent features p_i^{tar} before self-attention with the corresponding features p_i^{ref} derived from the denoising process of the reference image, and perform ref-self-attention between the original features p_i^{tar} and the concatenated features $p_i^{tar} \cup p_i^{ref}$. In this way, the information from the reference image is transferred to the denoising process of the rendered image, thereby our denoising U-Net achieves reference image guidance. Please refer to the supplementary material for more details.

3.3.2 Region-Aware Attention

The mutual process in Sec. 3.3.1 only provides global attention to the denoising process of the rendered image. To achieve region-aware attention, we further introduce additional attention masks into the above process. In fact, we have inferred the regional masks $\{M_j\}$ of reference image
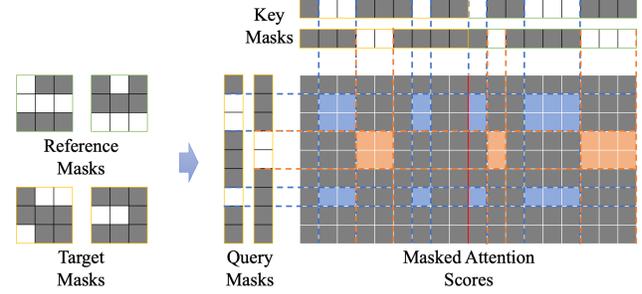
in preprocessing. To obtain the regional masks $\{\bar{M}_j\}$ of the rendered image, we first divide the 3D space into different regions in the vertical direction according to the boundaries of region masks $\{M_j\}$, and then divide the rendered image into corresponding regions according to the projection of the space division in the rendering view. Please refer to the supplementary material for more details. Subsequently, we feed these regional masks into the denoising process for region-aware attention.

As shown in Fig. 5, we implement the region-aware attention on the attention score matrix calculated from *query* vector and *key* vector. Assuming a 3×3 latent feature and its two attention masks, to implement ref-self-attention, the target feature and its masks are flattened into 1×9 vectors. After passing a linear network, the feature vector is then used as the *query* vector to calculate the attention score matrix together with the *key* vector formed by concatenating the target and reference features. Correspondingly, we use the flattened masks and the concatenated masks as the *query* and *key* masks to infer the local regions with same semantics, as shown in the blue and orange regions of Fig. 5. After determining the local attention regions, we multiply attention scores outside these regions by a coefficient $\gamma < 1$, and then normalize the whole score matrix, thereby improving the network's attention on these determined local regions. In practice, we set γ as 0.3.

After that, we now implement the denoising process with region-aware reference image guidance. Continuing from vanilla SDS [42], we also use the weighted noise residual between predicted and added noises as our Ref-SDS for 3D generation. Thanks to the image-level guidance, 3D generation based on our Ref-SDS is robust, thus we could set the CFG scale in the common level as the image generation. In this way, our Ref-SDS supports the generation of realistic textures that are consistent with the input reference image.

3.3.3 Loss Functions

During the optimization, we render our SDF network in front, back, and other random views. In the front view, we construct a reconstruction loss formed as $L_{rec} = L_{rgb} + L_{IoU}$. Here, L_{rgb} is a L_2-form loss calculated between

the rendered and input color images. L_{IoU} indicates the intersection-over-union (IoU) loss [67] between the rendered silhouette S and the silhouette S_I of input image:

$$L_{IoU} = 1 - \frac{\|S \otimes S_I\|_1}{\|S \oplus S_I - S \otimes S_I\|_1}, \quad (4)$$

where $\otimes$ and $\oplus$ indicate element-wise product and sum operator, respectively. Furthermore, we calculate a L_2-form normal loss L_{norm} between the rendered and estimated normal maps in the front and back views. For views other than the front view, we implement the proposed Ref-SDS loss $L_{Ref-SDS}$ to generate realistic texture and geometry. To further enhance the texture details, we implement a multi-step denoising process based on our region-aware reference-guided U-Net, and calculate the perceptual loss [25] L_{percep} between the rendered image and the enhanced image. To constrain the pose of the generated human, we additionally introduce a L_1-form SDF loss L_{SDF} between the predicted signed distance value s and that queried from SMPL-X body mesh. Besides, we also adopt the normal smooth loss L_{smooth} as the previous generation methods [32, 53] during the optimization process.

In summary, we train the entire network using the following objective function:

$$\begin{aligned} L = &\lambda_1 L_{rec} + \lambda_2 L_{norm} + \lambda_3 L_{Ref-SDS} \\ &+ \lambda_4 L_{percep} + \lambda_5 L_{SDF} + \lambda_6 L_{smooth}, \end{aligned} \quad (5)$$

where $\{\lambda_1, \ldots, \lambda_6\}$ are the weights used to balance different loss terms. In practice, we empirically set the weights as $10000, 100, 0.001, 20, 100, 5$, respectively.

3.4. Implementation Details

We implement our HumanRef with the ThreeStudio [18] framework in Pytorch [39] on a A100 GPU. For optimization, we adopt the Adam [26] optimizer with default hyper-parameters and a learning rate of 0.001 for all learnable parameters. To guarantee the quality of the generated model, we uniformly sample random views in the elevation range $[-20, 20]$ and azimuth range $[-180, 180]$ outside the front and back views. In addition, to facilitate the model to distinguish generated foreground objects and background, we follow [53] to set a random background color for the rendering results in each optimization step. To perform region-aware Ref-SDS generation and multi-step denoising, we adopt the stable diffusion model in version 1.5 [47] and the fast diffusion sampling scheduler UniPC [70].

4. Experiments

4.1. Setup

Baseline Methods To evaluate the performance of our HumanRef on 3D clothed human generation, we compareit

Methods	PIFu	PaMIR	TeCH	**Ours**
LPIPS ↓	0.054	0.050	0.044	**0.032**
Contextual ↓	3.180	2.961	2.882	**1.969**
CLIP Score ↑	80.1%	81.4%	85.3%	**90.0%**

Table 1. Quantitative comparison for 3D clothed human generation on 2D human images.

against three state-of-the-art baseline methods: PIFu [49], PaMIR [72], and a concurrent work TeCH [22]. Here, PIFu and PaMIR are learning-based methods that acquire the ability to infer the geometry and texture of a 3D clothed human from input images after being training on a large number of scanned human datasets. TeCH is a multi-stage optimization algorithm based on the SDS method designed for 3D clothed human generation. Additionally, to further assess the quality of the generated geometry, we incorporate three additional reconstruction-related baseline methods (FOF [14], D-IF [63], and ECON [60]) to conduct evaluation experiments on 3D human datasets: *CAPE* [33] and *THuman2.0* [65].

Evaluation Metrics Following [53], we adopt LPIPS [68] between the input image and rendered image at the input view to evaluate the reconstruction quality. Besides, we use the contextual distance [34] and CLIP score [44] as the generation quality metrics to measure the texture similarity and semantic similarity between the input image and rendered images at novel views. For evaluation on geometric experiments, we adopt the L_2 Normal error between normal maps rendered from scans and generated results to evaluate the geometry quality. Meanwhile, we employ PSNR, SSIM [57], LPIPS [68], and contextual distance to measure the texture quality on colored images.

4.2. Comparisons

We evaluate our HumanRef and baseline methods for 3D clothed human generation on 50 diverse human images released by [4, 15, 60], as shown in Fig. 6. Additionally, we provide quantitative comparison results to assess the performance of different methods in terms of reconstruction and generation qualities, as shown in Tab. 1. Clearly, our method surpasses the baseline methods in both qualitative and quantitative comparisons.

Contrary to TeCH and ours, PIFu and PaMIR employ 3D generators trained on scanned human datasets to predict human body geometry and texture. Their performance thus is limited by the training data and model design, struggling to infer detailed textures and fine geometry from a single image, particularly in areas invisible to the input. The first and third rows in Fig. 6 illustrate the disparity between front and back views, with the latter appearing blurrier and less detailed. As a result, they receive lower evaluation scores shown in Tab. 1. TeCH, however, can generate de-
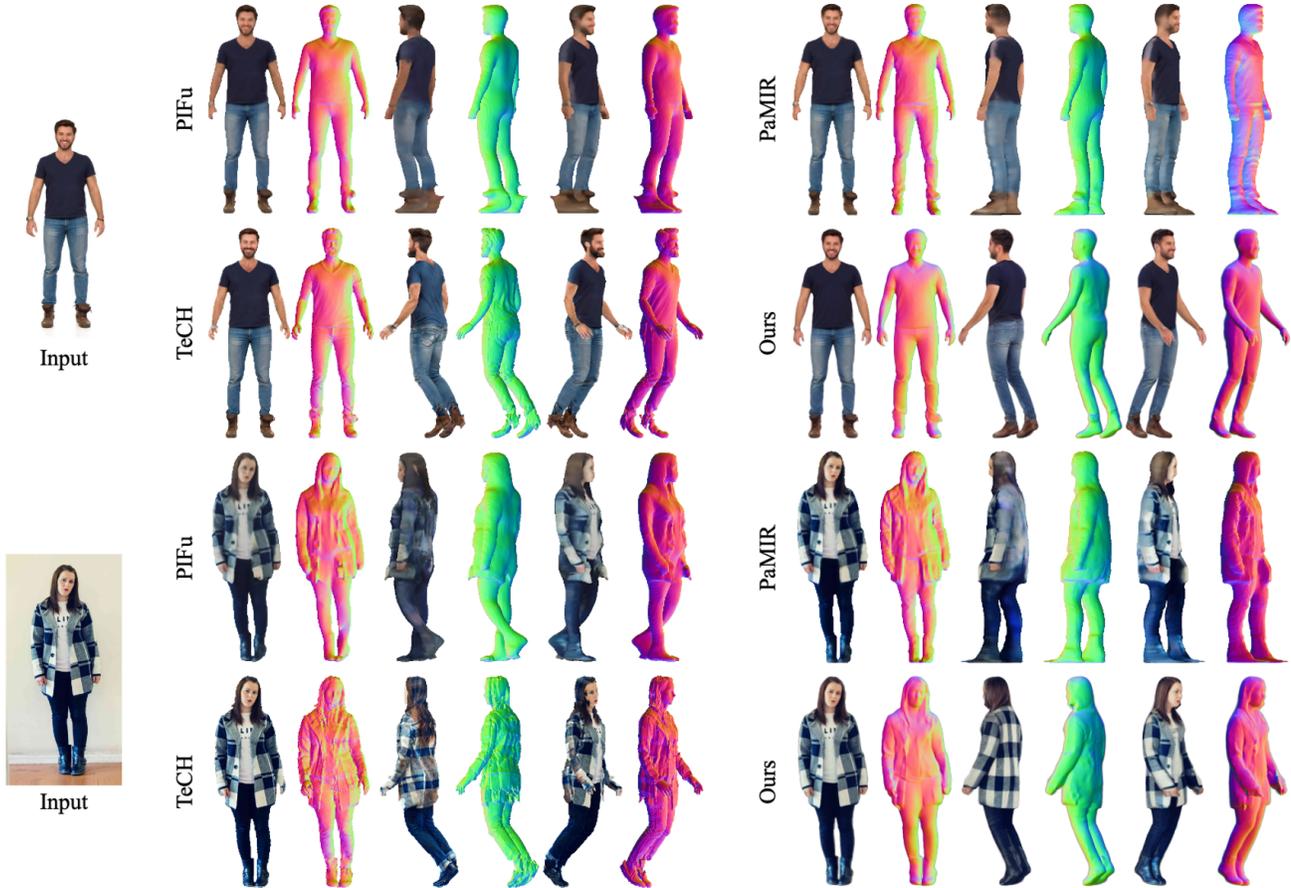
Figure 6. Qualitative comparison for 3D clothed human generation on 2D human images.

Methods	CAPE					THuman2.0				
	Normal↓	PSNR↑	SSIM↑	LPIPS↓	Contextual↓	Normal↓	PSNR↑	SSIM↑	LPIPS↓	Contextual↓
PIFu	0.0630	26.2845	0.9226	0.0910	3.1387	0.0816	22.3781	0.8913	0.1152	3.0899
PaMIR	0.0563	25.4044	0.9156	0.0942	3.1063	0.0682	22.0976	0.8870	0.1213	3.1452
TeCH	**0.0414**	27.4912	0.9396	0.0749	2.8076	0.0653	24.8015	0.9144	0.0877	2.8926
FOF	0.0582	-	-	-	-	0.0742	-	-	-	-
D-IF	0.0483	-	-	-	-	0.0682	-	-	-	-
ECON	0.0422	-	-	-	-	**0.0629**	-	-	-	-
Ours	0.0423	**28.1453**	**0.9489**	**0.0613**	**2.1603**	0.0648	**25.6963**	**0.9274**	**0.0770**	**2.2167**

Table 2. Quantitative comparison on 3D human datasets *CAPE* and *THuman2.0*. Compared with baseline methods, our HumanRef outperforms all baselines in texture evaluations and achieves comparable geometry quality to the human shape reconstruction methods.

tailed textures in unseen areas due to SDS-based text-guided optimization and pretrained diffusion model priors. To mitigate over-saturation in SDS-based generation and enhance realism, TeCH employs strategies to minimize SDS denoising diversity, including precise text inference via a question-answering algorithm, textural invention, and diffusion model fine-tuning. Nonetheless, TeCH struggles to achieve texture consistency with the input image due to the lack of image-level guidance in SDS denoising. In the *Man* example of Fig. 6, TeCH generates a realistic blue T-shirt

texture on the back, but there is still a noticeable difference from the dark blue texture in the input image. In contrast, our method, HumanRef, generates realistic and detail-rich textures that are more consistent with the input image. For geometry generation, although lacking a specialized design like TeCH, our method produces reasonable and complete 3D human structures, while TeCH suffers from issues like geometric breakage and stretching in the generated human feet. Moreover, in the *Woman* example of Fig. 6, our method successfully generates realistic view-consistent tex-
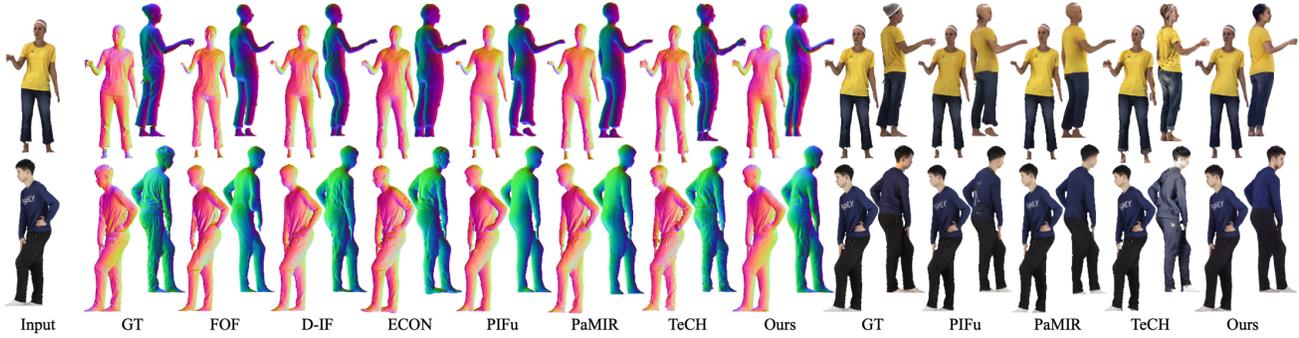
Figure 7. Qualitative comparison on 3D human datasets *CAPE* and *THuman2.0*.

Figure 8. 3D cloth human generation based on SDS, VSD, and our Ref-SDS, with estimated text prompts ʼ*a woman in a striped top and jeans*ʼ and ʼ*a woman in black and white polka dot print skirt*ʼ.

ture and geometry for humans with complex clothing textures that TeCH fails to infer.

Additionally, we further compare our HumanRef with baseline methods on 3D human datasets *CAPE* and *THuman2.0* to evaluate the quality of our generated geometry. Specifically, we randomly select 40 scans from *CAPE* and 40 scans from *THuman2.0* encompassing various poses and clothing styles. As shown in Fig 7 and Tab. 2, our method outperforms all baselines in texture evaluations and achieves comparable geometry quality to the human shape reconstruction methods.

4.3. Ablation Studies

Ref-SDS. To illustrate the superiority of the proposed Ref-SDS in our framework, we conduct a comparative experiment by replacing it with exiting SDS [42] and VSD [58] methods. Here, SDS and VSD are both text-guided methods, with VSD being a modified version of SDS to address over-saturation. In Fig. 8, SDS generates textures with higher saturation and lacks realism, while VSD produces realistic textures that are not consistent with the input. In contrast, our Ref-SDS-based method generates realistic and

Figure 9. Ablation study on the region-aware attention.

view consistent textures thanks to the reference image guidance in the denoising process.

Region-Aware Attention. Additionally, we evaluate the impact of the region-aware attention in our Ref-SDS by comparing our full framework with a version without it. In Fig. 9, we observe that the implementation without region-aware attention may spontaneously focus on undesired features in the reference image during ref-self-attention, resulting in realistic but unreasonable textures in some cases. By incorporating region-aware attention, we effectively limit the attention area of the network and obtain more reasonable texture inference results.

5. Conclusion

In this paper, we propose HumanRef, a unified framework for generating 3D clothed humans from a single input image. Our approach addresses the challenge of realistic texture generation in the invisible areas by proposing a modified Ref-SDS method that fully exploits the guidance of the reference image during the denoising process. Additionally, we introduce region-aware attention into our Ref-SDS, enhancing the precision of image guidance. Overall, our HumanRef framework empowers the generation of 3D clothed humans with view-consistent realistic textures and reasonable geometry from a single reference image. For a detailed discussion on the limitations of our work, we kindly request you to refer to the supplementary material.

References

[1] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Detailed human avatars from monocular video. In *2018 International Conference on 3D Vision (3DV)*, pages 98–109. IEEE, 2018. 1, 3

[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 Conference on Computer Vision and Pattern Recognition*, pages 8387–8397, 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*, pages 1175–1186, 2019. 1, 3

[4] 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*, pages 1506–1515, 2022. 3, 6

[5] Alexandru O Balan, Leonid Sigal, Michael J Black, James E Davis, and Horst W Haussecker. Detailed human shape and pose from images. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007. 1, 3

[6] Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. Multi-garment net: Learning to dress 3d people from images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5420–5430, 2019. 1, 3

[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 *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 561–578. Springer, 2016. 3

[8] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. *arXiv preprint arXiv:2304.08465*, 2023. 3, 4

[9] Lu Chen, Sida Peng, and Xiaowei Zhou. Towards efficient and photorealistic 3d human reconstruction: a brief survey. *Visual Informatics*, 5(4):11–19, 2021. 1

[10] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. *arXiv preprint arXiv:2303.13873*, 2023. 3

[11] Enric Corona, Mihai Zanfir, Thiemo Alldieck, Eduard Gabriel Bazavan, Andrei Zanfir, and Cristian Sminchisescu. Structured 3d features for reconstructing controllable avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16954–16964, 2023. 3

[12] Congyue Deng, Chiyu Jiang, Charles R Qi, Xinchen Yan, Yin Zhou, Leonidas Guibas, Dragomir Anguelov, et al. Nerdi: Single-view nerf synthesis with language-guided diffusion as general image priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20637–20647, 2023. 2, 3

[13] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 3

[14] Qiao Feng, Yebin Liu, Yu-Kun Lai, Jingyu Yang, and Kun Li. Fof: Learning fourier occupancy field for monocular real-time human reconstruction. *Advances in Neural Information Processing Systems*, 35:7397–7409, 2022. 6

[15] Jianglin Fu, Shikai Li, Yuming Jiang, Kwan-Yee Lin, Chen Qian, Chen Change Loy, Wayne Wu, and Ziwei Liu. Stylegan-human: A data-centric odyssey of human generation. In *European Conference on Computer Vision*, pages 1–19. Springer, 2022. 6

[16] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4d: Reconstructing and tracking humans with transformers. *arXiv preprint arXiv:2305.20091*, 2023. 3

[17] Peng Guan, Alexander Weiss, Alexandru O Balan, and Michael J Black. Estimating human shape and pose from a single image. In *2009 IEEE 12th International Conference on Computer Vision*, pages 1381–1388. IEEE, 2009. 3

[18] Yuan-Chen Guo, Ying-Tian Liu, Ruizhi Shao, Christian Laforte, Vikram Voleti, Guan Luo, Chia-Hao Chen, Zi-Xin Zou, Chen Wang, Yan-Pei Cao, and Song-Hai Zhang. threestudio: A unified framework for 3d content generation. https://github.com/threestudio-project/threestudio, 2023. 6

[19] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 3

[20] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 3

[21] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3

[22] Yangyi Huang, Hongwei Yi, Yuliang Xiu, Tingting Liao, Jiaxiang Tang, Deng Cai, and Justus Thies. Tech: Text-guided reconstruction of lifelike clothed humans. *arXiv preprint arXiv:2308.08545*, 2023. 1, 2, 3, 6

[23] 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*, pages 3093–3102, 2020. 3

[24] 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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14371–14382, 2023. 3

[25] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 694–711. Springer, 2016. 6

[26] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[27] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3383–3393, 2021. 3

[28] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 4

[29] Peike Li, Yunqiu Xu, Yunchao Wei, and Yi Yang. Self-correction for human parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):3260–3271, 2020. 4

[30] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 8662–8672, 2023. 3

[31] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 300–309, 2023. 3

[32] 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/CVF International Conference on Computer Vision*, pages 9298–9309, 2023. 6

[33] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 6469–6478, 2020. 6

[34] Roey Mechrez, Itamar Talmi, and Lihi Zelnik-Manor. The contextual loss for image transformation with non-aligned data. In *Proceedings of the European conference on computer vision (ECCV)*, pages 768–783, 2018. 6

[35] Luke Melas-Kyriazi, Iro Laina, Christian Rupprecht, and Andrea Vedaldi. Realfusion: 360deg reconstruction of any object from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8446–8455, 2023. 2, 3

[36] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 12663–12673, 2023. 3

[37] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 3

[38] 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 *2018 international conference on 3D vision (3DV)*, pages 484–494. IEEE, 2018. 3

[39] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 6

[40] 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. 3

[41] 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/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019. 3, 4

[42] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 2, 3, 4, 5, 8

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

[44] 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*, pages 8748–8763. PMLR, 2021. 6

[45] Amit Raj, Srinivas Kaza, Ben Poole, Michael Niemeyer, Nataniel Ruiz, Ben Mildenhall, Shiran Zada, Kfir Aberman, Michael Rubinstein, Jonathan Barron, et al. Dreambooth3d: Subject-driven text-to-3d generation. *arXiv preprint arXiv:2303.13508*, 2023. 4

[46] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 3

[47] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 3, 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. *Advances in Neural Information Processing Systems*, 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*, pages 2304–2314, 2019. 2, 3, 6

[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*, pages 84–93, 2020. 3

[51] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 3

[52] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 3

[53] Junshu Tang, Tengfei Wang, Bo Zhang, Ting Zhang, Ran Yi, Lizhuang Ma, and Dong Chen. Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior. *arXiv preprint arXiv:2303.14184*, 2023. 2, 3, 6

[54] Daniel Vlasic, Pieter Peers, Ilya Baran, Paul Debevec, Jovan Popović, Szymon Rusinkiewicz, and Wojciech Matusik. Dynamic shape capture using multi-view photometric stereo. In *ACM SIGGRAPH Asia 2009 papers*, pages 1–11. 2009. 1, 3

[55] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 12619–12629, 2023. 3

[56] Yiming Wang, Qin Han, Marc Habermann, Kostas Daniilidis, Christian Theobalt, and Lingjie Liu. Neus2: Fast learning of neural implicit surfaces for multi-view reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3295–3306, 2023. 4

[57] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 6

[58] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *arXiv preprint arXiv:2305.16213*, 2023. 2, 3, 4, 8

[59] Chenglei Wu, Kiran Varanasi, and Christian Theobalt. Full body performance capture under uncontrolled and varying illumination: A shading-based approach. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part IV 12*, pages 757–770. Springer, 2012. 1, 3

[60] Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J Black. Econ: Explicit clothed humans obtained from normals. *arXiv preprint arXiv:2212.07422*, 2022. 2, 3, 4, 6

[61] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J Black. Icon: Implicit clothed humans obtained from normals. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13286–13296. IEEE, 2022. 2, 3

[62] Dejia Xu, Yifan Jiang, Peihao Wang, Zhiwen Fan, Yi Wang, and Zhangyang Wang. Neurallift-360: Lifting an in-the-wild 2d photo to a 3d object with 360deg views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4479–4489, 2023. 3

[63] 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*, pages 9122–9132, 2023. 6

[64] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. *Advances in Neural Information Processing Systems*, 34:4805–4815, 2021. 4

[65] 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/CVF conference on computer vision and pattern recognition*, pages 5746–5756, 2021. 6

[66] Liang Yuan, Dingkun Yan, Suguru Saito, and Issei Fujishiro. Diffmat: Latent diffusion models for image-guided material generation. *Visual Informatics*, 2024. 3

[67] Jingbo Zhang, Ziyu Wan, and Jing Liao. Adaptive joint optimization for 3d reconstruction with differentiable rendering. *IEEE Transactions on Visualization and Computer Graphics*, 29(6):3039–3051, 2022. 6

[68] 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*, pages 586–595, 2018. 6

[69] Minda Zhao, Chaoyi Zhao, Xinyue Liang, Lincheng Li, Zeng Zhao, Zhipeng Hu, Changjie Fan, and Xin Yu. Efficientdreamer: High-fidelity and robust 3d creation via orthogonal-view diffusion prior. *arXiv preprint arXiv:2308.13223*, 2023. 3

[70] Wenliang Zhao, Lujia Bai, Yongming Rao, Jie Zhou, and Jiwen Lu. Unipc: A unified predictor-corrector framework for fast sampling of diffusion models. *arXiv preprint arXiv:2302.04867*, 2023. 6

[71] Yuheng Zhao, Jinjing Jiang, Yi Chen, Richen Liu, Yalong Yang, Xiangyang Xue, and Siming Chen. Metaverse: Perspectives from graphics, interactions and visualization. *Visual Informatics*, 6(1):56–67, 2022. 1

[72] 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*, 44(6): 3170–3184, 2021. 2, 3, 6

[73] Joseph Zhu and Peiye Zhuang. Hifa: High-fidelity text-to-3d with advanced diffusion guidance. *arXiv preprint arXiv:2305.18766*, 2023. 3