HumanNorm: Learning Normal Diffusion Model for High-quality and Realistic 3D Human Generation

Anonymous CVPR submission

Paper ID 9881

Figure 1. Taking text descriptions as input, HumanNorm has the capability to generate 3D human models with superior geometric quality and realistic textures. The 3D human models produced by HumanNorm can be exported as human meshes and texture maps, making them suitable for downstream applications.

Abstract

Recent text-to-3D methods employing diffusion models have made significant advancements in 3D human generation. However, these approaches face challenges due to the limitations of text-to-image diffusion models, which lack an understanding of 3D structures. Consequently, these methods struggle to achieve high-quality human generation, resulting in smooth geometry and cartoon-like appearances. In this paper, we propose HumanNorm, a novel approach for high-quality and realistic 3D human generation. The main idea is to enhance the model's 2D perception of 3D geometry by learning a normal-adapted diffusion model and a normal-aligned diffusion model. The normal-adapted diffusion model can generate high-fidelity normal maps corresponding to user prompts with view-dependent and body-aware text. The normal-aligned diffusion model learns to generate color images aligned with the normal maps, thereby transforming physical geometry details into realistic appearance. Leveraging the proposed normal diffusion model, we devise a progressive geometry generation strategy and a multi-step Score Distillation Sampling (SDS) loss to enhance the performance of 3D human generation. Comprehensive experiments substantiate HumanNorm's ability to generate 3D humans with intricate geometry and realistic appearances. HumanNorm outperforms existing text-to-3D methods in both geometry and texture quality.

1. Introduction

Large-scale generative models have achieved significant breakthroughs in diverse domains, including motion [41], audio [1, 26], and 2D image generation [25, 30, 31, 33, 34]. However, the pursuit of high-quality 3D content generation [5, 28, 37, 39] following the success of 2D generation poses a novel and meaningful challenge. Within the broader scope of 3D content creation, 3D human generation [10, 17, 18] holds particular significance. It plays a pivotal role in applications such as AR/VR, holographic communication, and the metaverse.

To achieve 3D content generation, a straightforward approach is to train generative models like GANs or diffusion models to generate 3D representations [2, 4, 12, 43]. However, these approaches face challenges due to the scarcity of current 3D datasets, resulting in restricted diversity and suboptimal generalization. To overcome these challenges, recent methods [19, 21, 28] adopt a 2D-guided approach to achieve 3D generation. Their core framework builds upon pre-trained text-to-image diffusion models and distills 3D contents from 2D generated images through Score Distillation Sampling (SDS) loss [28]. Leveraging the image generation priors learned from large-scale datasets, this framework enables more diverse 3D generation. However, current text-to-image diffusion models primarily emphasize the generation of natural RGB images, which results in a limited perception of 3D geometry structure and view direction. This limitation can result in Janus (multi-faced) artifacts and smooth geometry. Moreover, the texture of the 3D contents generated by existing methods is sometimes not based on geometry, which can result in fake 3D details, particularly in wrinkles and hair. Although some 3D human generation methods [3, 17, 18] introduce human body models such as SMPL [20] for animation and enhancing the quality of body details, they fail to address these fundamental limitations. Their results still suffer from sub-optimal geometry, fake 3D details and over-saturated texture.

In this paper, we present HumanNorm, a novel approach for generating high-quality and realistic 3D human models. The core idea is introducing a normal diffusion model to enhance the perception of 2D diffusion model for 3D geometry. HumanNorm is divided into two components: geometry generation and texture generation. For the geometry generation, we train a *normal-adapted diffusion model* using multi-view normal maps rendered from 3D human scans and prompts with view-dependent and body-aware text. Compared with text-to-image diffusion models, the normal-adapted diffusion model filters out the influence of texture and can generate high-fidelity surface normal maps according to prompts. This ensures the generation of 3D geometric details and avoids Janus artifacts. Since normal maps lack depth information, we also learn a depth-adapted diffusion model to further enhance the perception of 3D geometry. The 2D results generated by these diffusion models are presented in Fig. 2. The geometry is generated using both normal and depth SDS losses, which are based on our normal-adapted and depth-adapted diffusion models. Furthermore, a progressive strategy is designed to reduce geometric noise and enhance geometry quality.

As previously discussed, the core challenges for texture generation are fake 3D details and over-saturated appearances, as illustrated in Fig. 3. To avoid fake 3D details, we learn a *normal-aligned diffusion model* from normal-image pairs. This model efficiently integrates human geometric information into the texture generation process by taking normal maps as conditions. It accounts for elements such as shading caused by geometric folds and aligns the generated texture with surface normal. To tackle the over-saturated appearances, we introduce a multi-step SDS loss based on our normal-aligned diffusion model for texture generation. The loss recovers images with multiple diffusion steps, ensuring a more natural appearance of the generated texture.

The 3D models generated by HumanNorm are presented in Fig. 1. The key contributions of this paper are:

1. We propose a method for detailed human geometry generation by introducing a normal-adapted diffusion model that can generate normal maps from prompts with view-dependent and body-aware text.
2. We propose a method for geometry-based texture generation by learning a normal-aligned diffusion model, which transforms physical geometry details into realistic appearances.
3. We introduce the multi-step SDS loss to mitigate over-saturated texture and a progressive strategy for enhancing stability in geometry generation.

2. Related work

Our study is primarily centered on the realm of text-to-3D, with a specific emphasis on text-to-3D human generation. Here, we revisit some recent work related to our method.

Text-to-3D content generation. Early methods, such as CLIP-Forge [35], DreamFields [14], and CLIP-Mesh [23], combine a pre-trained CLIP [29] model with 3D representations, and generate 3D content under the supervision of CLIP loss. DreamFusion [28] introduces the SDS loss and generates NeRF [22] under the supervision of a text-to-image diffusion model. Following this, Magic3D [19] proposes a two-stage method that employs both NeRF and mesh for high-resolution 3D content generation. Latent-NeRF [21] optimizes NeRF in the latent space using a latent diffusion model to avoided the burden of encoding images. TEXTure [32] introduces a method for texture generation, transfer, and editing. Fantasia3D [5] decomposes the generation process into geometry and texture generation to enhance the performance of 3D generation. To address the over-saturation issue, ProlificDreamer [44] proposes a Variational Score Distillation (VSD) loss to produce high-quality NeRF. IT3D [6] introduces GAN loss and leverages generated 2D images to enhance the quality of 3D contents. MVDream [37] proposes a multi-view diffusion model to generate consistent multi-views for 3D generation. DreamGaussian [40] uses 3D Gaussian splatting [16] to accelerate the generation process. However, these methods are un-

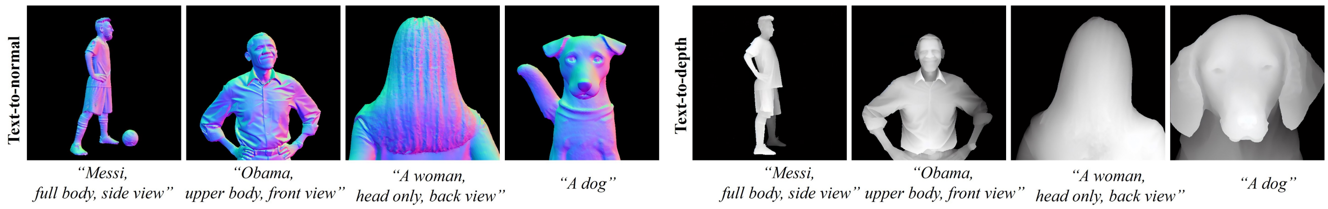


Figure 2. **2D results by normal-adapted and depth-adapted diffusion models.** The view-dependent texts like "front view" are utilized to control the view direction. The body-aware texts like "upper body" are employed to control which body part is generated.

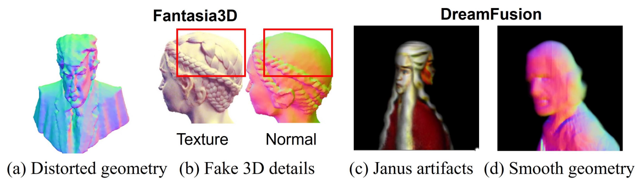


Figure 3. **Problems of existing methods.**

able to generate high-quality 3D humans, leading to Janus artifacts and unreasonable body proportions. Our method addresses these issues by introducing normal-adapted diffusion model that can generate normal maps from prompts with view-dependent and body-aware text.

Text-to-3D human generation. Recently, EVA3D [11], LSV-GAN [46], GETAvatar [50], Get3DHuman [45] introduce GAN-based frameworks to directly generate 3D representations for 3D human generation. AvatarCLIP [10] integrates SMPL and Neus [42] to create 3D humans, leveraging CLIP for a supervision. DreamAvatar [3] and AvatarCraft [15] utilize the pose and shape of the parametric SMPL model as a prior, guiding the generation of humans. DreamWaltz [13] creates 3D humans using a parametric human body prior, incorporating 3D-consistent occlusion-aware SDS and 3D-aware skeleton conditioning. DreamHuman [17] generates animatable 3D humans by introducing a pose-conditioned NeRF that is learned using imGHUM. AvatarBooth [47] uses dual fine-tuned diffusion models separately for the human face and body, enabling the creation of personalized humans from casually captured face or body images. The most recent model, AvatarVerse [48], trains a ControlNet with DensePose [7] as conditions to enhance the view consistency of 3D human generation. TADA [18] derives SMPL-X [27] with a displacement layer and a texture map, using hierarchical rendering with SDS loss to produce 3D humans. While these methods reduce Janus artifacts and unreasonable body shapes by introducing human body models, they still produce 3D humans with fake 3D details, over-saturation and smooth geometry. Moreover, the introduction of SMPL presents challenges for these methods in generating 3D humans with intricate clothing such as puffy skirts and hats. Our method addresses these issues by learning normal diffusion model and introducing multi-step SDS loss, thereby enhancing the both geometry and texture quality of 3D humans.

3. Preliminary

3.1. Diffusion-guided 3D Generation Framework

When provided with text y as the generation target, the core of the diffusion-guided 3D generation framework aims to align the images $\mathbf{x}_0$ rendered from the 3D representation θ with the generated image distribution $p(\mathbf{x}_0|y)$ of the 2D diffusion model. Specifically, during the 3D generation process, the rendered images $\mathbf{x}_0$ are obtained by randomly sampling cameras $\mathbf{c}$ and rendering through a differentiable rendering function $g(\theta, \mathbf{c})$. Suppose the rendered images from various angles are distributed as $q^\theta(\mathbf{x}_0|y) = \int q^\theta(\mathbf{x}_0|y, \mathbf{c})p(\mathbf{c})d\mathbf{c}$, the optimization objective of diffusion-guided 3D generation framework can be represented as follows:

$$\min_\theta D_{KL}(q^\theta(\mathbf{x}_0|y) \parallel p(\mathbf{x}_0|y)). \tag{1}$$

Directly optimizing this objective is highly challenging, and recent methods have proposed losses such as SDS [28] and VSD [44] to solve it. To further enhance the quality of geometry, Fantasia3D [5] proposes to disentangle the geometry θ_g and appearance θ_c in the 3D representation θ. In the geometry stage, it aligns $q^{\theta_g}(\mathbf{z}_0^n|y)$, the distribution of the rendered normal maps $\mathbf{z}_0^n$, with the natural image distribution $p(\mathbf{x}_0|y)$:

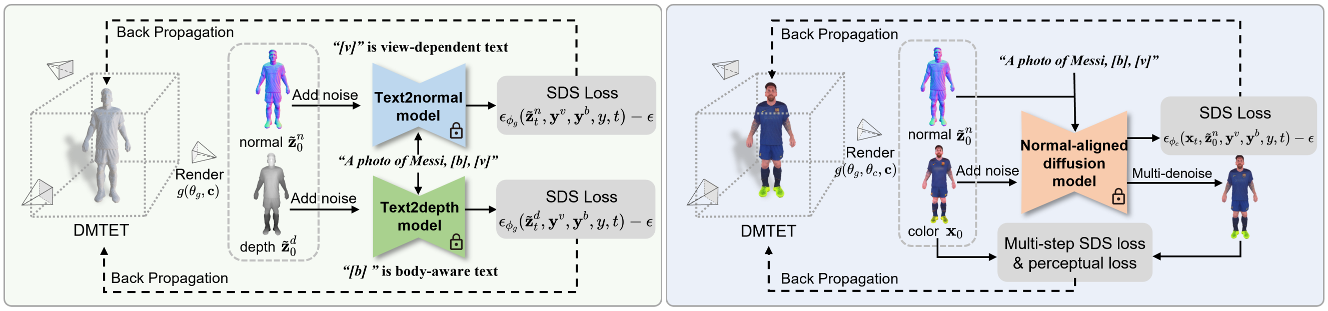
$$\min_{\theta_g} D_{KL}(q^{\theta_g}(\mathbf{z}_0^n|y) \parallel p(\mathbf{x}_0|y)). \tag{2}$$

In the texture stage, the texture of 3D objects is optimized through Eq. (1).

3.2. Bottleneck of Diffusion-guided 3D Generation

The bottleneck of the diffusion-guided 3D generation lies in the T2I (text-to-image) diffusion model, which confines itself to parameterize the probability distribution of natural RGB images, denoted as $p(\mathbf{x}_0|y)$. Therefore, current T2I diffusion model lacks the understanding of both view direction and geometry. Consequently, 3D generation directly guided by the T2I diffusion model (Eq. (1)) leads to Janus artifacts and low-quality geometry as shown in Fig. 3 (c-d). Although Fantasia3D disentangles geometry and texture, it still encounters issues originating from the T2I diffusion model in both geometry and texture stages. In the geometry

CVPR
#9881



(a) Geometry Generation (b) Texture Generation

Figure 4. **Overview of HumanNorm.** Our method is designed for high-quality and realistic 3D human generation from given prompts. The whole framework consists of geometry and texture generation. We first propose the normal-adapted and depth-adapted diffusion model for the geometry generation. These two models can guide the rendered normal and depth maps to approach the learned distribution of high-fidelity normal and depth maps through the SDS loss, thereby achieving high-quality geometry generation. In terms of texture generation, we introduce the normal-aligned diffusion model. The normal-aligned diffusion model leverages normal maps as guiding cues to ensure the alignment of the generated texture with geometry. We first exclusively employ the SDS loss and then incorporate the multi-step SDS and perceptual loss to achieve realistic texture generation.

stage, directly aligning the rendered normal maps distribution $q^{\theta_g}(\mathbf{z}_0^n|y)$ with the natural images distribution $p(\mathbf{x}_0|y)$ is inappropriate since normal maps significantly differ from RGB images. This alignment results in geometry distortions and artifacts, as depicted in Fig. 3 (a). In the texture stage, minimizing the divergence between the appearance distribution $q^{\theta_c}(\mathbf{x}_0|y)$ and the natural image distribution $p(\mathbf{x}_0|y)$ may lead to fake 3D details due to the absence of geometric guidance, as presented in Fig. 3 (b).

4. Method

We propose HumanNorm to achieve high-quality and realistic 3D human generation. The whole generation framework has a geometry stage and a texture stage, as shown in Fig. 4. In this section, we first introduce our normal diffusion model, which consists of a normal-adapted diffusion model and a normal-aligned diffusion model (Sec. 4.1). Then in the geometry stage, based on the normal-adapted diffusion model, we utilize the DMTET [36] as the 3D representation and propose a progressive generation strategy to achieve high-quality geometry generation (Sec. 4.2). In texture stage, building upon the normal-aligned diffusion model, we propose the multi-step SDS loss for high-fidelity and realistic appearance generation (Sec. 4.3).

4.1. Normal Diffusion Model

In the pursuit of generating a high-quality and realistic 3D human from a given text target y, the first challenge lies in achieving precise geometry generation. This entails aligning the distributions of rendered normal maps $q^{\theta_g}(\mathbf{z}_0^n|\mathbf{c}, y)$ from multiple viewpoints $\mathbf{c}$ with an ideal normal maps distribution $\hat{p}(\mathbf{z}_0^n|\mathbf{c}, y)$. The next challenge is to generate the realistic texture θ_c while ensuring its coherence with the established geometry θ_g. Therefore, minimizing the divergence between the distribution of rendered images $q^{\theta_c}(\mathbf{x}_0|\mathbf{c}, y)$ and an ideal geometry-aligned images distribution $\hat{p}(\mathbf{x}_0|\mathbf{c}, \theta_g, y)$ becomes essential. The ideal optimization objective is formulated as follows:

$$\begin{aligned}\min_{\theta_g,\theta_c} & \underbrace{D_{KL}(q^{\theta_g}(\mathbf{z}_0^n|\mathbf{c}, y) \parallel \hat{p}(\mathbf{z}_0^n|\mathbf{c}, y))}_{geometry\ generation\ objective} \\ & + \underbrace{D_{KL}(q^{\theta_c}(\mathbf{x}_0|\mathbf{c}, y) \parallel \hat{p}(\mathbf{x}_0|\mathbf{c}, \theta_g, y))}_{texture\ generation\ objective}.\end{aligned} \tag{3}$$

However, as discussed in Sec. 3.1, the existing T2I (text-to-image) diffusion model is limited to parameterize the distribution of natural RGB images, denoted as $p(\mathbf{x}_0|y)$, which deviates significantly from the ideal distributions $\hat{p}(\mathbf{z}_0^n|\mathbf{c}, y)$ and $\hat{p}(\mathbf{x}_0|\mathbf{c}, \theta_g, y)$. To bridge this gap, we propose the incorporation of normal maps, representing the 2D perception of human geometry, into the T2I diffusion model to approximate $\hat{p}(\mathbf{z}_0^n|\mathbf{c}, y)$ and $\hat{p}(\mathbf{x}_0|\mathbf{c}, \theta_g, y)$. For the geometry component, we propose to fine-tune the diffusion model, adapting it to generate the distribution of normal map $p(\mathbf{z}_0^n|y)$. In the context of texturing, we utilize normal maps $\mathbf{z}_0^n$ as conditions to guide the diffusion model $p(\mathbf{x}_0|\mathbf{z}_0^n, y)$ in generating normal-aligned images, which ensures that the generated texture aligns with the geometry. In addition, we further introduce view-dependent text $\mathbf{y}^v$ (*e.g. "front view"*) and body-aware text $\mathbf{y}^b$ (*e.g. "upper body"*), serving as an additional condition for the diffusion model. This strategy ensures that the generated images align with the view direction and enables body part generation, as depicted in Fig. 2. The final optimization objective is:

$$\begin{aligned}\min_{\theta_g,\theta_c} & D_{KL}(q^{\theta_g}(\mathbf{z}_0^n|\mathbf{c}, y) \parallel p(\mathbf{z}_0^n|\mathbf{y}^v, \mathbf{y}^b, y)) + \\ & D_{KL}(q^{\theta_c}(\mathbf{x}_0|\mathbf{c}, y) \parallel p(\mathbf{x}_0|\mathbf{z}_0^n, \mathbf{y}^v, \mathbf{y}^b, y)).\end{aligned} \tag{4}$$

Next, we will introduce our 3D human generation framework and construction of the normal-adapted diffusion model and normal-aligned diffusion model used to parameterize $p(\mathbf{z}_0^n|\mathbf{y}^v, \mathbf{y}^b, y)$ and $p(\mathbf{x}_0|\mathbf{z}_0^n, \mathbf{y}^v, \mathbf{y}^b, y)$ for geometry and texture generation.

4.2. Geometry Generation

4.2.1 Normal-adapted Diffusion Model

Constructing the normal-adapted diffusion model for high-quality geometry generation faces several challenges. First, existing 3D human datasets are scarce, leading to a limited number of normal maps for training. Therefore, we employ a fine-tuning strategy to adapt a text-to-image diffusion model into a text-to-normal diffusion model. Then we find the rendered normal maps undergo dramatic changes with variations in viewing angles, which results in potential overfitting or underfitting issues. To mitigate this effect and encourage the diffusion model to focus on perceiving the details of geometry, we transform the normal maps $\mathbf{z}_0^n$ from the world coordinate to camera coordinates by the rotation R of the camera parameters. The transformed normal maps $\tilde{\mathbf{z}}_0^n$ are used for training of the normal-adapted diffusion model. As discussed in Sec. 4.1, we add the view-dependent text $\mathbf{y}^v$ and body-aware text $\mathbf{y}^b$ as addition conditions. The fine-tuning process employs this optimization objective:

$$\min_{\phi_g} \mathbb{E}_{\mathbf{c},t,\epsilon} \left[\|\epsilon_{\phi_g}(\alpha_t \tilde{\mathbf{z}}_0^n + \sigma_t, \mathbf{y}^v, \mathbf{y}^b, y, t) - \epsilon\|_2^2 \right], \quad (5)$$

where $\mathbf{c}$ is a camera pose, t is a timestep, ϵ denotes noise and y is a prompt. σ_t and α_t are the parameters of the diffusion scheduler. $\epsilon_{\phi_g}(\cdot)$ is the normal-adapted diffusion model.

SDS loss [28] is widely employed in various diffusion-guided 3D generation frameworks. It translates the optimization objective in Eq. (1) into the optimization of the divergence between two distributions with diffusion noise, thereby achieving 3D generation. Our geometry is optimized by the normal SDS loss based on the trained normal-adapted diffusion model:

$$\nabla \mathcal{L}_{SDS}(\theta_g) = \mathbb{E}_{\mathbf{c},t,\epsilon} \left[\omega(t)(\epsilon_{\phi_g}(\tilde{\mathbf{z}}_t^n, \mathbf{y}^v, \mathbf{y}^b, y, t) - \epsilon) \frac{\partial g(\theta_g, \mathbf{c})}{\partial \theta_g} \right]. \quad (6)$$

where $\tilde{\mathbf{z}}_t^n$ corresponds to the rendered normal map $\tilde{\mathbf{z}}_t^0$ with the noise ϵ at timestep t. $\omega(t)$ is the parameters of the diffusion scheduler. $g(\theta_g, \mathbf{c})$ denotes render the normal map at camera pose $\mathbf{c}$ from geometry θ_g. In addition to normal SDS loss, we also fine-tune a depth-adapted diffusion model by simply changing normal maps to depth maps to calculate depth SDS loss. We found the depth SDS loss can reduce geometry distortion and artifacts in geometry generation, as shown in Fig. 8.

4.2.2 Progressive Geometry Generation

DMTET [36] is used as our 3D representation. To augment the robustness of 3D human generation, we initialize it with a neutral body mesh. We propose a progressive strategy including progressive positional encoding and progressive SDF loss to mitigate geometric noise and enhance the overall quality of geometry generation.

Positional encoding [22, 24] maps each component of input vectors to a higher-dimensional space, thereby enhancing the 3D representation's ability to capture high-frequency details. However, we found that the high frequency of positional encoding can also lead to noisy surface. This is due to the DMTET prioritizing coarse geometry during the initial optimization stage, resulting in the failure to translate high-frequency input into geometric details. To solve this, we employ a mask to suppress high-frequency components of positional encoding for SDF function in DMTET during the initial stage. This allows the network to focus on low-frequency components of geometry and improving the training stability in the beginning. As training progresses, we gradually reduce the mask for high-frequency components. Thereby enhancing the details such as clothes wrinkle.

In addition, the progressive SDF loss is introduced to further improve the quality of geometry generation. We first record the SDF functions of DMTET before reducing the high-frequency mask, denoted as $\mathbf{s}(x)$. Then as training progresses, we add the SDF loss to mitigate strange geometry deformations:

$$\mathcal{L}_{SDF}(\theta_g) = \sum_{x \in P} \|\tilde{\mathbf{s}}_{\theta_g}(x) - \mathbf{s}(x)\|_2^2, \quad (7)$$

where $\tilde{\mathbf{s}}_{\theta_g}(x)$ is the SDF function in DMTET and P is the set of random sampling points. This strategy can effectively avoid unreasonable body proportions.

4.3. Texture Generation

4.3.1 Normal-aligned Diffusion Model

In texture generation, we fix the geometry parameters θ_g and introduce the normal-aligned diffusion model as guidance. The normal-aligned diffusion model can translate physical geometry details into realistic appearance and ensure the generated texture is aligned with the geometry. Specifically, we employ the strategy of ControlNet [49] to incorporate transformed normal maps $\tilde{\mathbf{z}}_0^n$ as the guided condition of the T2I diffusion model. The training objective of the normal-aligned diffusion model is as follows:

$$\min_{\phi_c} \mathbb{E}_{\mathbf{c},t,\epsilon} \left[\|\epsilon_{\phi_c}(\alpha_t \mathbf{x}_0 + \sigma_t, \tilde{\mathbf{z}}_0^n, \mathbf{y}^v, \mathbf{y}^b, y, t) - \epsilon\|_2^2 \right] \quad (8)$$

After training, we propose a multi-step SDS loss based on the normal-aligned diffusion model for photo-realistic texture generation.

Figure 5. **Examples of 3D humans generated by HumanNorm.** A single view and the corresponding normal map are rendered for visualization. **See supplementary for video results**.

4.3.2 Multi-step SDS Loss

We generate texture in two stages. In the initial stage, we employ the vanilla SDS loss of the normal-aligned diffusion model ϵ_{ϕ_c} for texture generation:

$$\nabla \mathcal{L}_{SDS}(\theta_c) = \mathbb{E}_{\mathbf{c},t,\epsilon}\left[\omega(t)(\epsilon_{\phi_c}(\mathbf{x}_t, \tilde{\mathbf{z}}_0^n, \mathbf{y}^v, \mathbf{y}^b, y, t) - \epsilon)\frac{\partial g(\theta_c, \mathbf{c})}{\partial \theta_c}\right]. \tag{9}$$

While SDS loss can lead to over-saturated styles and appear less natural as shown in Fig. 7 (c), it efficiently optimizes a reasonable texture as an initial value. We subsequently refine the texture through multi-step SDS and perceptual loss. Different from SDS loss, multi-step SDS loss needs multiple diffusion steps to recover the distribution of RGB images, which promotes stability during optimization and avoids getting trapped in local optima. As a result, the generated images appear more natural. To further prevent over-saturation effects, the perceptual loss is also applied to keep the natural style of the rendering images consistent with the images generated by the normal-aligned diffusion model. The loss is defined as:

$$\nabla \mathcal{L}_{MSDS}(\theta_c) \approx \mathbb{E}_{\mathbf{c},t,\epsilon}\left[\omega(t)(h(\mathbf{x}_t, \tilde{\mathbf{z}}_0^n, \mathbf{y}^v, \mathbf{y}^b, y, t) - \mathbf{x}_0)\frac{\partial g(\theta_c, \mathbf{c})}{\partial \theta}\right] + \lambda_p \mathbb{E}_{\mathbf{c},t,\epsilon} \left[\left(V(h(\mathbf{x}_t, \tilde{\mathbf{z}}_0^n, \mathbf{y}^v, \mathbf{y}^b, y, t)) - V(\mathbf{x}_0)\right)\frac{\partial V(\mathbf{x}_0)}{\partial \mathbf{x}_0}\frac{\partial g(\theta_c, \mathbf{c})}{\partial \theta_c}\right], \tag{10}$$

where V is the first k layers of the VGG network [38]. $h(\mathbf{x}_t, \tilde{\mathbf{z}}_0^n, \mathbf{y}^v, \mathbf{y}^b, y, t)$ denotes the multi-step image generation function of the normal-aligned diffusion model. λ_p is the weights of perceptual loss.

5. Experiment

5.1. Implementation Details

For each prompt, our method needs 15K iterations for geometry generation and 10K iterations for texture generation. The entire generation process takes about 2 hours on a single NVIDIA RTX 3090 GPU with 24 GB memory. The final rendered images and videos have a resolution of 1024×1024. Additional details, including dataset, training settings, and more, can be found in our supplementary.

5.2. Qualitative Evaluation

The examples of 3D humans generated by HumanNorm is shown in Fig. 5. Furthermore, we present qualitative comparisons with text-to-3D content methods including DreamFusion [28], LatentNeRF [21], TEXTure [32], and Fantasia3D [5], as well as text-to-3D human methods including DreamHuman [17] and TADA [18].

Comparison with text-to-3D content methods. As illustrated in Fig. 6, the results produced by text-to-3D content methods present some challenges. The proportions of the generated 3D humans tend to be distorted, and the texture appears to be over-saturated and noisy. DreamFusion struggles to generate full-body humans, often missing the feet,

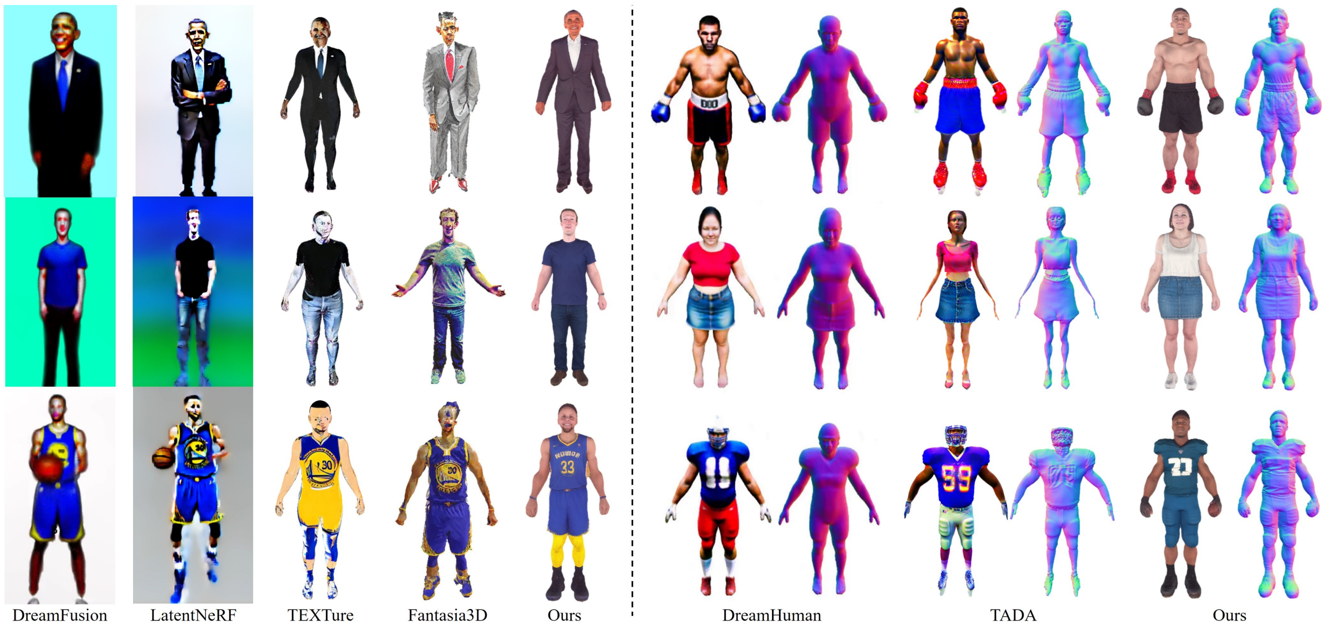


Figure 6. **Comparisons with text-to-3D content methods and text-to-3D human methods.** The results of DreamFusion are generated by unofficial code. The results of DreamHuman are taken from its original paper and project page.

Method	FID ↓	CLIP Score ↑
DreamFusion	145.2	28.65
LatentNeRF	152.6	27.42
TEXTure	142.8	27.08
Fantasia3D	120.6	28.47
DreamHuman	111.3	30.15
TADA	120.0	30.65
HumanNorm (Ours)	92.5	31.70

Table 1. **Quantitative comparisons with text-to-3D content and text-to-3D human methods.**

even given a prompt like “the full body of...”. In contrast, our method delivers superior results with more accurate geometry and realistic textures.

Comparison with text-to-3D human methods. As shown in Fig. 6, text-to-3D human methods yield outcomes with enhanced geometry due to the integration of SMPL-X and imGHUM human body models. In contrast, HumanNorm can create 3D humans with a higher level of geometric detail, such as wrinkles in clothing and distinct facial features. Furthermore, text-to-3D human methods also encounter issues with over-saturation, while our method can generate more lifelike appearances thanks to the multi-step SDS loss.

5.3. Quantitative Evaluation

Evaluating the quality of generated 3D models quantitatively can be challenging. However, we attempt to assess HumanNorm using two specific metrics. Firstly, we compute the Fréchet Inception Distance (FID) [9], a measure that compares the distribution of two image datasets. In our case, we calculate the FID between the views rendered from the generated 3D humans and the images produced by Stable Diffusion V1.5 [33]. In total, 30 prompts are used and 120 images are rendered or generated for each prompt. Secondly, we utilize the CLIP score [8] to measure the compatibility between the prompts with the rendered views of 3D humans. The results are detailed in Tab. 1. As can be observed, HumanNorm achieves a lower FID score. This suggests that the views rendered from our 3D humans are more closely aligned with the high-quality 2D images generated by the stable diffusion model. Furthermore, the superior CLIP score of HumanNorm indicates our enhanced capability to generate humans that are more accurately aligned with the prompts. Finally, we also conduct a user study to evaluate HumanNorm. The details of this study are provided in our supplementary.

5.4. Ablation Studies

Effectiveness of normal-adapted and depth-adapted diffusion models. In Fig. 7 (a), we show the geometry generated by a text-to-image diffusion model instead of our normal-adapted and depth-adapted diffusion models. One can see that the method struggles to generate facial geometry, and holes appear on ears. Additionally, the results display smoother clothing wrinkles. The experiment demonstrates that our normal-adapted and depth-adapted diffusion models are beneficial in generating high-quality geometry.

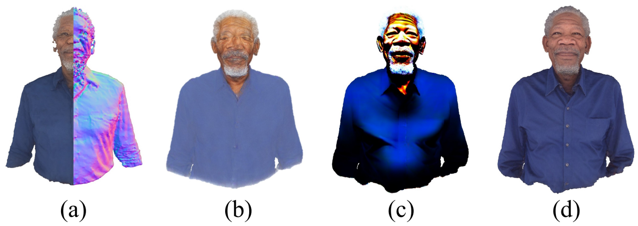


Figure 7. **Ablation studies.** (a) Without normal-adapted and depth-adapted diffusion. (b) Without normal-aligned diffusion model. (c) Without multi-step SDS loss. (d) The full method.

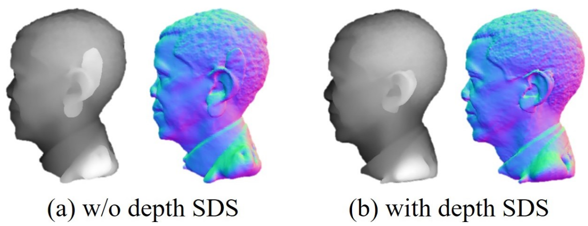


Figure 8. **Importance of depth SDS.**

Effectiveness of normal-aligned diffusion model. In Fig. 7 (b), we experiment with the removal of the normal-aligned diffusion model, opting instead for a text-to-image diffusion model for texture generation. The resulting texture, as can be observed, is somewhat blurry and fails to accurately display geometric details. This is because the text-to-image diffusion model struggle to align the generated texture with geometry. However, using the normal-aligned diffusion model, our method manages to overcome these limitations. It achieves more precise and intricate details, leading to a significant enhancement for the appearance of the 3D humans.

Effectiveness of multi-step SDS loss. In Fig. 7 (c), we present the result generated when only the SDS loss is used in the texture generation. The generated model is noticeably over-saturated. However, as shown in Fig. 7 (d), the texture generated through multi-step SDS loss exhibits a more realistic and natural color, which underscores the effectiveness of the multi-step SDS loss.

Effectiveness of depth SDS. Since normal maps lack depth information, optimizing geometry by only calculating normal SDS loss may lead to failed geometry in some regions. As shown in Fig. 8 (a), the ear exhibits artifacts when only using normal SDS loss. This is because the normal of the artifacts is similar to the normal of the head, making it non-salient for the normal diffusion model. In contrast, we can clearly see the artifacts in the depth map. In Fig. 8 (b), it's evident that the artifacts are reduced when adding the additional depth SDS loss based on our depth-adapted diffusion model, which demonstrates the effectiveness of introducing depth SDS.

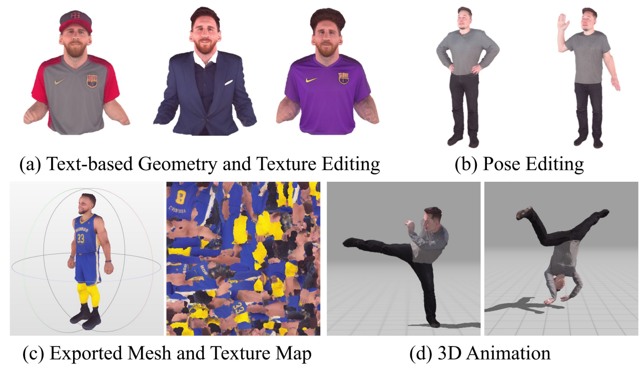


Figure 9. **Applications of HumanNorm.**

5.5. Applications

Text-based Editing. HumanNorm offers the capability to edit both the texture and geometry of the generated 3D humans by adjusting the input prompt. As demonstrated in Fig. 9 (a), we modify the color and style of Messi's clothing, as well as his hairstyle.

Pose Editing. HumanNorm also provides the ability to edit the pose of generated 3D humans by adjusting the pose of the mesh used for initialization and modifying the prompts. The results of pose editing are displayed in Fig. 9 (b).

3D Animation. HumanNorm enables the creation of lifelike human mesh featuring about 400K distinct faces and intricate 2K-resolution texture map. Based on the high-quality models, we can animate them using full-body motion sequences. Results are presented in Fig. 9 (c-d)

6. Conclusion

We presented HumanNorm, a novel method for high-quality and realistic 3D human generation. By learning the normal diffusion model, we improved the capabilities of 2D diffusion models for 3D human generation. Utilizing the trained normal diffusion model, we introduced a diffusion-guided 3D generation framework. Additionally, we devised the progressive strategy for robust geometry generation and the multi-step SDS loss to address the over-saturation problem. We demonstrated that HumanNorm can generate 3D humans with intricate geometric details and realistic appearances, outperforming existing methods.

Limitations and future work. HumanNorm primarily focuses on addressing the geometric and textural challenges present in existing methods. As a result, 3D humans generated by HumanNorm necessitate a rigged human skeleton for 3D animation. In our future work, we plan to incorporate SMPL-X to directly animate 3D humans and improve the quality of body details such as fingers. Additionally, our generated texture may exhibit undesired shading. To address this, we are considering the use of Physically-Based Rendering (PBR) for material estimation and relighting.

References

[1] Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, et al. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*, 2023. 1

[2] Sizhe An, Hongyi Xu, Yichun Shi, Guoxian Song, Umit Ogras, and Linjie Luo. Panohead: Geometry-aware 3d full-head synthesis in 360°. *CVPR*, 2023. 2

[3] 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, 3

[4] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *CVPR*, pages 16123–16133, 2022. 2

[5] 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. 1, 2, 3, 6

[6] Yiwen Chen, Chi Zhang, Xiaofeng Yang, Zhongang Cai, Gang Yu, Lei Yang, and Guosheng Lin. It3d: Improved text-to-3d generation with explicit view synthesis. *arXiv preprint arXiv:2308.11473*, 2023. 2

[7] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *CVPR*, pages 7297–7306, 2018. 3

[8] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, 2021. 7

[9] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, pages 6626–6637, 2017. 7

[10] Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, and Ziwei Liu. Avatarclip: Zero-shot text-driven generation and animation of 3d avatars. *ACM TOG*, 41(4), 2022. 2, 3

[11] Fangzhou Hong, Zhaoxi Chen, LAN Yushi, Liang Pan, and Ziwei Liu. Eva3d: Compositional 3d human generation from 2d image collections. In *ICLR*, 2023. 3

[12] Shoukang Hu, Fangzhou Hong, Tao Hu, Liang Pan, Haiyi Mei, Weiye Xiao, Lei Yang, and Ziwei Liu. Humanliff: Layer-wise 3d human generation with diffusion model. *arXiv preprint arXiv:2308.09712*, 2023. 2

[13] Yukun Huang, Jianan Wang, Ailing Zeng, He Cao, Xianbiao Qi, Yukai Shi, Zheng-Jun Zha, and Lei Zhang. Dreamwaltz: Make a scene with complex 3d animatable avatars. *arXiv preprint arXiv:2305.12529*, 2023. 3

[14] Ajay Jain, Ben Mildenhall, Jonathan T Barron, Pieter Abbeel, and Ben Poole. Zero-shot text-guided object generation with dream fields. In *CVPR*, pages 867–876, 2022. 2

[15] 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. 3

[16] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM TOG*, 42(4):1–14, 2023. 2

[17] Nikos Kolotouros, Thiemo Alldieck, Andrei Zanfir, Eduard Gabriel Bazavan, Mihai Fieraru, and Cristian Sminchisescu. Dreamhuman: Animatable 3d avatars from text. *arXiv preprint arXiv:2306.09329*, 2023. 2, 3, 6

[18] Tingting Liao, Hongwei Yi, Yuliang Xiu, Jiaxaing Tang, Yangyi Huang, Justus Thies, and Michael J Black. Tada! text to animatable digital avatars. In *3DV*, 2024. 2, 3, 6

[19] 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 *CVPR*, pages 300–309, 2023. 2

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

[21] Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. Latent-nerf for shape-guided generation of 3d shapes and textures. In *CVPR*, pages 12663–12673, 2023. 2, 6

[22] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 2, 5

[23] Nasir Mohammad Khalid, Tianhao Xie, Eugene Belilovsky, and Tiberiu Popa. Clip-mesh: Generating textured meshes from text using pretrained image-text models. In *ACM SIGGRAPH Asia Conference Proceedings*, pages 1–8, 2022. 2

[24] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM TOG*, 41(4):1–15, 2022. 5

[25] 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. 1

[26] Aaron Oord, Yazhe Li, Igor Babuschkin, Karen Simonyan, Oriol Vinyals, Koray Kavukcuoglu, George Driessche, Edward Lockhart, Luis Cobo, Florian Stimberg, et al. Parallel wavenet: Fast high-fidelity speech synthesis. In *ICML*, pages 3918–3926. PMLR, 2018. 1

[27] 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 *CVPR*, pages 10975–10985, 2019. 3

[28] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *ICLR*, 2023. 1, 2, 3, 5, 6

[29] 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 *ICML*, pages 8748–8763. PMLR, 2021. 2

[30] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, pages 8821–8831. PMLR, 2021. 1

[31] 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. 1

[32] Elad Richardson, Gal Metzer, Yuval Alaluf, Raja Giryes, and Daniel Cohen-Or. Texture: Text-guided texturing of 3d shapes. In *ACM SIGGRAPH Conference Proceedings*, 2023. 2, 6

[33] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 1, 7

[34] 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. *NeurIPS*, 35:36479–36494, 2022. 1

[35] Aditya Sanghi, Hang Chu, Joseph G Lambourne, Ye Wang, Chin-Yi Cheng, Marco Fumero, and Kamal Rahimi Malekshan. Clip-forge: Towards zero-shot text-to-shape generation. In *CVPR*, pages 18603–18613, 2022. 2

[36] Tianchang Shen, Jun Gao, Kangxue Yin, Ming-Yu Liu, and Sanja Fidler. Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. *NeurIPS*, 34: 6087–6101, 2021. 4, 5

[37] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023. 1, 2

[38] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *ICLR*, 2015. 6

[39] Jingxiang Sun, Bo Zhang, Ruizhi Shao, Lizhen Wang, Wen Liu, Zhenda Xie, and Yebin Liu. Dreamcraft3d: Hierarchical 3d generation with bootstrapped diffusion prior. *arXiv preprint arXiv:2310.16818*, 2023. 1

[40] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint arXiv:2309.16653*, 2023. 2

[41] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. In *ICLR*, 2023. 1

[42] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *NeurIPS*, 34:27171–27183, 2021. 3

[43] Tengfei Wang, Bo Zhang, Ting Zhang, Shuyang Gu, Jianmin Bao, Tadas Baltrusaitis, Jingjing Shen, Dong Chen, Fang Wen, Qifeng Chen, et al. Rodin: A generative model for sculpting 3d digital avatars using diffusion. In *CVPR*, pages 4563–4573, 2023. 2

[44] 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

[45] Zhangyang Xiong, Di Kang, Derong Jin, Weikai Chen, Linchao Bao, Shuguang Cui, and Xiaoguang Han. Get3dhuman: Lifting stylegan-human into a 3d generative model using pixel-aligned reconstruction priors. In *ICCV*, pages 9287–9297, 2023. 3

[46] Yinghao Xu, Wang Yifan, Alexander W Bergman, Menglei Chai, Bolei Zhou, and Gordon Wetzstein. Efficient 3d articulated human generation with layered surface volumes. *arXiv preprint arXiv:2307.05462*, 2023. 3

[47] Yifei Zeng, Yuanxun Lu, Xinya Ji, Yao Yao, Hao Zhu, and Xun Cao. Avatarbooth: High-quality and customizable 3d human avatar generation. *arXiv preprint arXiv:2306.09864*, 2023. 3

[48] 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. 3

[49] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023. 5

[50] Xuanmeng Zhang, Jianfeng Zhang, Rohan Chacko, Hongyi Xu, Guoxian Song, Yi Yang, and Jiashi Feng. Getavatar: Generative textured meshes for animatable human avatars. In *ICCV*, pages 2273–2282, 2023. 3