Disentangled Clothed Avatar Generation from Text Descriptions

Jionghao Wang[1,2]†, Yuan Liu[3]†, Zhiyang Dou[3,4],
Zhengming Yu[1], Yongqing Liang[1], Cheng Lin[3], Rong Xie[2],
Li Song[2]‡, Xin Li[1], and Wenping Wang[1]‡

[1] Texas A&M University
[2] Shanghai Jiao Tong University
[3] The University of Hong Kong
[4] TransGP

Fig. 1: Our method generates high-quality separated human body and clothes meshes from text prompts. Kinematics or simulation motions can drive the disentangled avatar representations to achieve photorealistic animations.

Abstract. In this paper, we introduce a novel text-to-avatar generation method that separately generates the human body and the clothes and allows high-quality animation on the generated avatar. While recent advancements in text-to-avat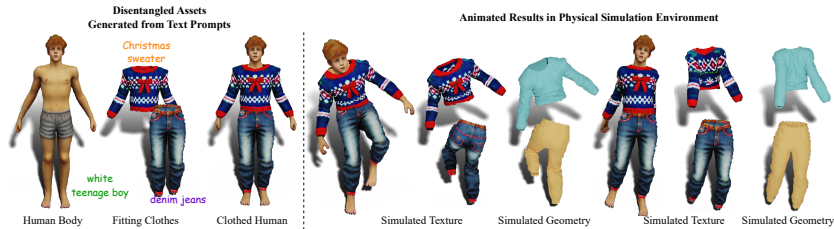ar generation have yielded diverse human avatars from text prompts, these methods typically combine all elements—clothes, hair, and body—into a single 3D representation. Such an entangled approach poses challenges for downstream tasks like editing or animation. To overcome these limitations, we propose a novel disentangled 3D avatar representation named Sequentially Offset-SMPL (SO-SMPL), building upon the SMPL model. SO-SMPL represents the human body and clothes with two separate meshes but associates them with offsets to ensure the physical alignment between the body and the clothes. Then, we design a Score Distillation Sampling (SDS)-based distillation framework to generate the proposed SO-SMPL representation from text prompts. Our approach not only achieves higher texture and geometry quality and better semantic alignment with text prompts, but also significantly improves the visual quality of character animation, virtual try-on, and avatar editing. Project page: this link.

Keywords: avatar and clothes generation · score distillation sampling

---

† Co-first authors.

‡ Corresponding authors.

1   Introduction

Human avatars play a significant role in visual storytelling, conveying narratives through actions, expressions, and interactions with other elements in the animated world, thereby creating a dynamic and engaging experience for the audience. Avatars can be designed to represent diverse characters, including different genders, races, ethnicities, and abilities, promoting inclusivity and challenging stereotypes in the media. However, creating high-quality avatars is costly and labor-intensive, requiring skilled 3D modelers. To achieve photorealistic, animatable, and customizable avatars, it is imperative to employ a disentangled representation consisting of distinct components, such as body and clothing elements. This approach enhances customization and reusability, allowing seamless alterations to clothing while maintaining the same underlying body structure. Additionally, separate components contribute to heightened realism in animation, as different parts exhibit distinct motion characteristics. For example, human body movements can be animated using forward kinematics, while cloth motion can be generated through physical simulation.

Despite evident benefits, most existing research [39, 97] focuses on reconstructing a singular avatar mesh from videos or images [71, 105], neglecting the potential for disentangled representations. Only a few studies, like DELTA [26] and SCARF [27], explore disentangled avatar reconstruction from videos but do not facilitate the arbitrary creation of such representations. Thus, a comprehensive method for creating disentangled human avatars remains an unresolved challenge. Extending text-to-avatar pipelines to generate disentangled avatars involves not only independently generating body and clothing components but also ensuring their precise alignment for visually appealing results. To address this, we introduce Sequentially-Offset-SMPL (SO-SMPL), based on SMPL-X [58] for characterizing coarse human shapes. SO-SMPL adds trainable offsets and appearance features to represent the unclothed human body, inspired by TADA [51], and further refines these to learn clothing specifics. This sequential approach ensures perfect alignment and separation of body and clothing meshes.

Leveraging the power of stable diffusion with SDS losses [69], we introduce a novel approach to generate our SO-SMPL representation from text prompts. Our method achieves disentangled avatar generation by first creating the unclothed human body and subsequently distilling the clothing components based on the underlying body structure. Experiments show that our method produces high-quality human avatars with better geometry details and appearances than baseline methods [35, 51]. The disentangled nature of our SO-SMPL representation unlocks strong customization capabilities, allowing for the seamless integration of various clothing options on the same human body. Furthermore, the avatars generated by our state-of-the-art method exhibit exceptionally realistic animation results, achieved by simulating distinct motion characteristics for both clothing and human body components. Moreover, by generating another layer of offsets, we can generate multiple-layered clothes, like a suit on a T-shirt, and each layer has a separate mesh.

2   Related work

Distilling diffusion for 3D tasks. Many recent works have utilized diffusion models to generate 3D objects or scenes. Some works [6, 43, 45, 56, 62, 70, 89] trained 3D diffusion models to directly generate 3D representations such as point clouds, meshes, radiance fields [61], SDFs [88], UDFs [106], DMTets [78] or tri-planes [11]. Others designed diffusion models to generate 3D-aware 2D images or textures [8, 12, 29, 53, 55, 57, 60, 74, 80, 87]. Both lines of works rely on 3D data [20, 21, 25] or multi-view images [73, 100, 104] to train.

With the success of 2D text-to-image diffusion models [22, 32, 75], recent works [5, 13, 16, 40, 52, 69, 76, 77, 84, 86, 92, 93, 95, 102, 116] propose to distill 3D representations from 2D diffusion models to circumvent the lack of 3D data. Some of them designed different 3D representations [13, 52, 84] for generation, while others explored the guidance schemes of diffusion models [69, 86, 92].

3D Human Avatar generation. Previous works on 3D avatar generation [7, 15, 28, 34, 64, 65, 110, 111] could generate diverse human geometries or textured avatars, but the generation process could not be controlled by text descriptions. With the help of large language models like CLIP [72] and text-to-image diffusion models [32, 75], recent works [9, 38, 46, 101, 114] were able to generate high-fidelity avatars with prompts or with both image and prompt [39, 68, 94]. Other works have attempted to generate human avatars by painting the given human meshes with generated textures [82, 103]. However, the generated avatars could not be animated and utilized in CG software. [4, 35, 41, 47, 51, 54, 99, 109] could produce animatable avatars, but they ignored garment-human interactions, leading to unsatisfactory animation quality.

Disentangled representations. Reconstructing the human body and clothes as separate geometries has long been studied in the field of computer vision [19, 26, 27, 42], but generating distinct body and clothes representations has barely been explored. Previous works have made attempts to generate 3D avatars as a combination of separate components [34] or layers [98], but they did not explicitly disentangle human body and accessories, making an individual component or layer not physically meaningful. A more recent work [36] trained a diffusion-based human generation model and distinctly represented the human body and different types of clothes as separate layers, but it did not have semantic controllability and suffers from over-smoothing textures. Another effort [91] proposed to use an SDS-based pipeline to generate a NeRF-based clothing layer separated from the human body, but could not be utilized in physical simulations due to its implicit representation. A concurrent work TECA [107] explored disentangling hair and head ornament from text and images while our work focuses on decomposing the avatar into a human body and clothes. Another recent work AvatarFusion [37] also separates clothes from human bodies. However, AvatarFusion adopts a post-processing strategy based on the outputs of AvatarClip [35], leading to low-quality results. Our method is natively designed for disentangled generation, which produces much better quality.

3   Preliminaries

Score distillation sampling [69] is proposed to optimize a 3D representation through a frozen pre-trained large 2D diffusion model. Specifically, to optimize a given 3D representation with parameters θ using a diffusion model with parameters ϕ, SDS loss is computed as:

$$\nabla_\theta \mathcal{L}_{SDS} = E_{t,\epsilon} \left[(\hat{\epsilon}_\phi(x_t; \mathbf{y}, t) - \epsilon) \frac{\partial x}{\partial \theta} \right] \tag{1}$$

where t is the time step in diffusion model, $x = g(\theta)$ is the rendered image using a differentiable renderer g, $x_t = x + \epsilon$ is the noise version of x, ϵ is a sampled gaussian noise, $\hat{\epsilon}_\phi(x_t; \mathbf{y}, t)$ is the predicted noise of the diffusion model, and $\mathbf{y}$ is the input text prompt.

SMPL-X [67] is an expressive parametric human model that could produce a human mesh with a fixed topology. SMPL-X is conditioned by a given input shape parameter β, pose parameter θ, and expression parameters ψ, and follows a vertex-based linear blend skinning(LBS) based on skeleton $J(\beta)$ and skinning weights W for pose transforming and animation. Specifically, SMPL-X computes a posed human mesh $\mathbf{K}(\beta, \theta, \psi)$ as:

$$\begin{aligned} \mathbf{K}(\beta, \theta, \psi) &= \mathcal{W}(\mathbf{T}(\beta, \theta, \psi), J(\beta), \theta, W) \\ \mathbf{T}(\beta, \theta, \psi) &= T + B_s(\beta) + B_e(\psi) + B_p(\theta) \end{aligned} \tag{2}$$

where an T-pose body mesh $\mathbf{T}(\beta, \theta, \psi)$ is first calculated as the combination of a mean template T, shape, expression and pose blend shapes $\{B_s, B_e, B_p\}$ then warped to target pose θ with LBS operation $\mathcal{W}$.

4   Methodology

Fig. 2 is an overview of our disentangled avatar generation pipeline. Given text descriptions of the human avatar, our pipeline produces an animatable and high-quality unclothed avatar in the first stage. In the second stage, we create clothes on the target avatar. In the following, we introduce our disentangled representation, called SO-SMPL, of the human body and the clothes.

4.1   SO-SMPL Representation

Human body geometry. As shown in Fig. 3, we represent the human body mesh with a *densified* SMPL-X parametric model introduced in Eq. 2 and we add learnable vertex-wise offsets $\mathbf{O}_h$ to represent details on the body, similar to recent works [51]. We fix the pose of the human body mesh to an "A" pose θ and use a default expression parameter ψ during the text-to-avatar generation process so that θ and ψ are omitted for simplicity. Specifically, the human body geometry $\mathbf{T}_h$ is

$$\mathbf{T}_h(\beta, \mathbf{O}_h) = \mathbf{T}(\beta) + \mathbf{O}_h, \tag{3}$$
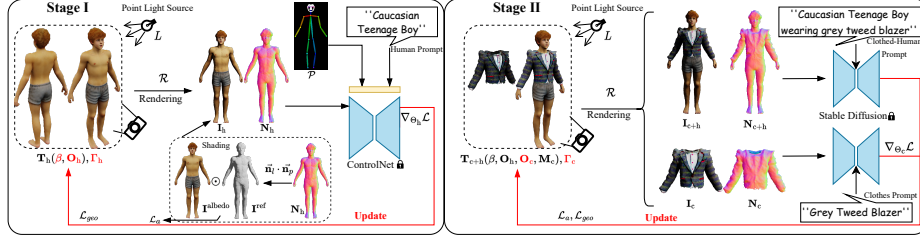
Fig. 2: An overview of our generation pipeline. Our pipeline has two stages. In Stage I, we generate a base human body model by optimizing its shape parameter β, vertex offset $\mathbf{O}_h$ and albedo texture Γ_h. A ControlNet is utilized to compute a Score Distillation Sampling(SDS) Loss conditioned on an "A"-pose map $\mathcal{P}$ for the rendered RGB image $\mathbf{I}_h$ and normal map $\mathbf{N}_h$. In Stage II, we freeze the human body model and optimize the clothes parameters $\mathbf{O}_c$ along with the albedo texture MLP Γ_c. The rendered RGB images and normal maps of both the clothed human and the clothes are used in computing the SDS losses. In both stages, we utilized a simple Phong shading model to render images from our SO-SMPL representations.

where $\mathbf{O}_h$ is trainable vertex-wise offset and β is trainable shape parameter.
Clothes geometry. Clothes are represented by additional vertex-wise masks $\mathbf{M}_c$ and offsets $\mathbf{O}_c$ on $\mathbf{T}_h$. The clothed human mesh $\mathbf{T}_{c+h}$ are

$$\mathbf{T}_{c+h}(\beta, \mathbf{O}_h, \mathbf{O}_c, \mathbf{M}_c) = \mathbf{T}_h(\beta, \mathbf{O}_h) + \mathbf{O}_c \odot \mathbf{M}_c, \tag{4}$$

where $\odot$ denotes the Hadamard product between the vertex-wise offsets and masks. We can also get the clothes geometry $\mathbf{T}_c$ only by masking on the clothed human mesh

$$\mathbf{T}_c(\beta, \mathbf{O}_h, \mathbf{O}_c, \mathbf{M}_c) = \mathbf{T}_{c+h}(\beta, \mathbf{O}_h, \mathbf{O}_c, \mathbf{M}_c) \odot \mathbf{M}_c. \tag{5}$$

To enable the generation of diverse clothes types, we use the body-part segments of SMPL-X and pre-define vertex-level mask templates of 6 different garment types, namely long shirts, short shirts, long pants, short pants, vests, and overalls. These templates are used to initialize clothes mask $\mathbf{M}_c$. We can also support skirts generation using the skirts template from BCNet [42].
Discussion. Previous works [37, 107] tried to represent clothes as a volume-based representation, e.g. NeRF [61], which is hard to be converted to a mesh. Our clothes representation is naturally a thin mesh representation, which is preferable for most graphics software and physical simulators [1, 2]. Meanwhile, by representing the clothes as a displacement surface that "grows" upon the base human, we ensure the alignment between the human body and clothes.

4.2 SO-SMPL Rendering

On both the human body mesh and the clothes mesh, we model appearances with a Phong shading model where the meshes are characterized by their albedo and the lighting is simply a combination of a point light and an ambient light.
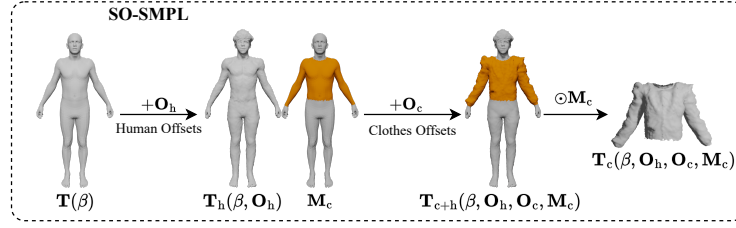
Fig. 3: Illustration of our SO-SMPL representation. Two vertex-wise offsets, namely the human offset $\mathbf{O}_h$ and the clothes offset $\mathbf{O}_c$ are sequentially added in order to the SMPL-X body mesh $\mathbf{T}(\beta)$ to obtain the human body mesh $\mathbf{T}_h(\beta, \mathbf{O}_h)$ and clothed human mesh $\mathbf{T}_{c+h}(\beta, \mathbf{O}_h, \mathbf{O}_c, \mathbf{M}_c)$, where a vertex mask $\mathbf{M}_c$ is calculated to determine the clothing region. Finally, we mask the clothed human mesh with $\mathbf{M}_c$ and obtain the separated clothes mesh $\mathbf{T}_c(\beta, \mathbf{O}_h, \mathbf{O}_c, \mathbf{M}_c)$.

Albedo. We use an MLP Γ_i to represent the albedo of the mesh. The albedo ρ of a 3D point $\mathbf{x}$ on the target mesh is computed by

$$\rho = \Gamma_i(\gamma(\mathbf{x}); \xi_i), \tag{6}$$

where $i \in \{h, c\}$ means clothes or human body, γ is positional encoding of the 3D position $\mathbf{x}$, and ξ_i is the parameter of the MLP.

Shading model. Given a randomly-sampled point light with a position of $\mathbf{v}$ and an intensity of $\mathbf{l}_d$, along with the ambient light intensity $\mathbf{l}_a$, the shading color $\mathbf{c}$ of a surface point $\mathbf{p}$ is computed by

$$\mathbf{c} = \rho_{\mathbf{p}}(\mathbf{l}_a + \max(0, \mathbf{n}_l \cdot \mathbf{n}_{\mathbf{p}})\mathbf{l}_d) \tag{7}$$

where $\mathbf{n}_{\mathbf{p}}$ is the normal direction of the point $\mathbf{p}$, $\mathbf{n}_l = (\mathbf{v} - \mathbf{p})/\|\mathbf{v} - \mathbf{p}\|_2$ is the light direction towards the surface vertice, and $\rho_{\mathbf{p}}$ is the albeodo on this point.

w/ baked-in artifacts   w/o baked-in artifacts

Fig. 4: An intuitive illustration of the impact of the baked-in artifacts in learned albedo. On the left, the generated pant suffers from severe baked-in wrinkles in its texture, resulting in non-photorealistic wrinkles and shadows in animation. On the right side, our shader prevents the shadows from baking into the albedo, hence significantly improving the visual quality of animations.

Discussion. Incorporating this shading model greatly reduces the chance that the shadows are baked into the albedo of the generated clothes, which improves

the consistency between geometry and textures, as shown in Fig. 4. This shading model associates the rendering colors with the generated normals so optimizing rendering colors also changes the normal directions, which reduces the chances of getting baked-in shadows or wrinkles.

Rendering clothes & human body. From any given camera viewpoint, we can render the images from our SO-SMPL representation using rasterization. We first determine the corresponding 3D point and its normal from $\mathbf{T}_{h+c}$ on every pixel. Then, the shading colors are computed using Eq. 7, resulting in the rendered image $\mathbf{I}_{h+c}$. In a similar fashion, we render separate images for the body ($\mathbf{I}_h$) and the clothes ($\mathbf{I}_c$), respectively. Concurrently, we also obtain the corresponding normal maps $\mathbf{N}_{h+c}$, $\mathbf{N}_h$, and $\mathbf{N}_c$. More details are included in the supplementary material.

4.3 SO-SMPL Generation

To learn the geometry and texture representations of the SO-SMPL model, we utilize SDS losses on both the rendered RGB images and the normal maps. Our approach involves a two-stage optimization process: In Stage I, we focus on generating an unclothed human body model. Following this, in Stage II, we proceed to generate the clothing, which is attached to the human body model established in Stage I.

Stage I: Human body generation. For the human body generation, we adopt an OpenPose [10]-based ControlNet [112] with SDS [69] losses to optimize a SO-SMPL representation. An "A"-pose joint map $\mathcal{P}$ rendered from 3D skeleton joints is utilized as the condition for ControlNet to compute SDS losses. Given an input rendered image $\mathbf{I} \in \{\mathbf{I}_h, \mathbf{N}_h\}$, the gradient is computed by

$$\nabla_{\Theta_h} \mathcal{L} = E_{t,\epsilon} \left[(\hat{\epsilon}_\phi(\mathbf{I}_t; \mathbf{y}_h, \mathcal{P}, t) - \epsilon) \frac{\partial \mathbf{I}}{\partial \Theta_h} \right], \tag{8}$$

where $\mathbf{I}_t = \mathbf{I} + \epsilon$ is a noisy version of the input image, $\Theta_h = \{\beta, \mathbf{O}_h, \xi_h\}$ is the trainable parameters during human body generation, $\{\beta, \mathbf{O}_h\}$ are geometry parameters, ξ_h is the texture parameters, $\hat{\epsilon}_\phi(\mathbf{I}_t; \mathbf{y}_h, \mathcal{P}, t)$ is the predicted noise given text embedding $\mathbf{y}_h$, noise step t and the pose map $\mathcal{P}$. To ensure our generated human body mesh is unclothed, we add specific descriptions to the human prompt such as "wearing tight shorts". Additionally, to further guide the body mesh and texture away from depicting clothing, we employ negative prompts like "loose clothes, accessories".

Stage II: Clothes generation. Similar to human body generation, we also utilize SDS losses to update clothing parameters $\Theta_c = \{\mathbf{O}_c, \xi_c\}$. We calculate SDS gradients through RGB and normal images of both the clothed human and the separated clothes. Given a rendered image $\mathbf{I} \in \{\mathbf{I}_c, \mathbf{N}_c, \mathbf{I}_{c+h}, \mathbf{N}_{c+h}\}$, and its corresponding text prompt embedding $\mathbf{y}_c$, the SDS loss is termed as:

$$\nabla_{\Theta_c} \mathcal{L} = E_{t,\epsilon} \left[(\hat{\epsilon}_\phi(\mathbf{I}_t; \mathbf{y}_c, t) - \epsilon) \frac{\partial \mathbf{I}}{\partial \Theta_c} \right]$$

Fig. 5: A gallery of our generated clothed avatars. Our method can generate avatars with varied ethnicities, genders, and clothes.

Albedo smoothness constraint. To avoid geometries (e.g. wrinkles and bumps) and lighting (shadows and highlights) being baked into the albedo of the model, we employ an albedo smoothness loss following [17, 113] to reduce these effects. With a small perturbation δ of the input position $\mathbf{x}$, we encourage the estimated albedo from texture MLP $\Gamma_i \in \{\Gamma_h, \Gamma_c\}$ to be consistent, i.e.

$$\mathcal{L}_a = \|\Gamma_i(\mathbf{x}) - \Gamma_i(\mathbf{x} + \delta)\|_2 \tag{9}$$

Geometry constraints. In both the human body and clothing generation, we utilize geometry constraints to ensure that the mesh is smooth and does not deviate too far from the original SMPL-X mesh. We employ a Laplacian smoothness term [44] $\mathcal{L}_s$ to minimize the norm of graph Laplacian. We also add an offset regularizing term $\mathcal{L}_o = \|\mathbf{O}_i\|_2$ for the body/clothes offset $\mathbf{O}_i \in \{\mathbf{O}_h, \mathbf{O}_c\}$. Thus, the geometry regularization loss is $\mathcal{L}_{\text{geo}} = \mathcal{L}_s + \mathcal{L}_o$. More implementation details such as camera configurations are included in the supplementary material.

5   Experiments

5.1   Generated Avatars & Clothes

As demonstrated in Fig. 5, our method generates diverse avatars and various types of clothes with detailed textures. Besides, since the clothes are generated on a base human body, they fit with the intended human naturally.

5.2   Quality Comparisons

We conduct comparisons on three aspects, static clothed-avatar quality, clothes quality, and animation quality. For the generation quality of clothed avatars,

Fig. 6: Quality comparison of the generated clothed avatars. We compare results from TEXture [74], AvatarCLIP [35], TADA! [51] and our method with the same prompts.

we compare with state-of-the-art text-to-3D avatar generation methods TEXture [74], AvatarCLIP [35] and TADA! [51]. As for clothes generation, we conduct comparisons with two of the best existing text-to-3D methods ProlificDreamer [92] and TextMesh [84]. We then compare animation results in the physical simulation environment [2] with the animatable meshes generated by TEXture [74] and TADA! [51].

Clothed avatar quality. A visual comparison of the generated avatars is presented in Fig. 6. TEXture [74] takes a fixed SMPL-X mesh as input and suffers from texture inconsistency under different views. AvatarCLIP [35] often produces geometry artifacts and poor facial details. TADA! [51] tends to generate unnatural geometric artifacts and unaligned outfits, e.g. a hull on the African-American woman's belly and the white teenage girl's yellow pants that contradicts the "beige legging" described by the prompt. This is because they use a simple text-to-image diffusion model and lack sufficient regularization on geometry. In comparison, using a pose-controlled diffusion model and proper regularization, our method can produce not only the most realistic human with detailed textures but also high-quality clothes that are semantically aligned with texts.

Clothes quality. We also evaluate the quality of the separated clothes. As demonstrated in Fig. 7, neither TextMesh [84] or ProlificDreamer [92] can generate meaningful clothes shapes or textures. Besides, they cannot control the position and the size of the generated clothes, and thus cannot be fitted to an

avatar. In comparison, our method can produce clothes fitting to a specific avatar and also with high-quality textures and shapes.
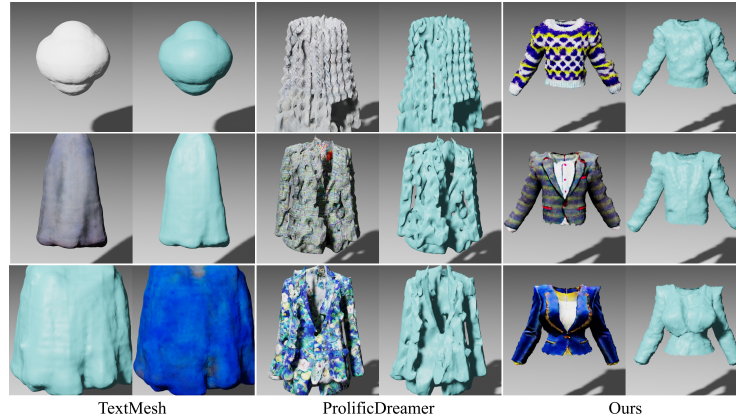

TextMesh   ProlificDreamer   Ours

Fig. 7: Comparisons of clothes generation quality. Garments in each row are generated using the same prompts.

Animation quality. Our generated human body mesh is built upon the SMPL-X model and thus allows us to use SMPL-X parameters to animate the body. On the other hand, the separately generated clothes can be used in a simulation environment [1, 2] to simulate their motions. We used motion sequences from AMASS [59] and AIST++ [50] to animate the generated avatars and simulate the clothes animations in the MarvelousDesigner [2]. Note that the recent progress in human motion synthesis [3, 18, 23, 83, 85, 115] or human motion capture from visual signals [14, 24, 49, 79, 90, 108] can also be employed to drive the generated character. The animation results of TADA!, TEXTure, and our method are shown in Fig. 8. For TADA! and TEXTure, the clothes are entangled on the generated avatar, which leads to inconsistent wrinkles and shadows in animation. Our method disentangles clothes and human bodies, enabling realistic clothes-body interactions and visual effects.

Methods	AvatarCLIP [35]	TEXture [74]	TADA [51]	Ours
FID ↓	166.0	132.8	90.2	**84.9**
CLIP-FID ↓	34.9	42.8	27.4	**26.2**
CLIP Score ↑	19.2	23.6	24.9	**28.4**

Table 1: Quantitative comparisons of generated "A"-pose avatar quality in FID, CLIP-FID, and CLIP score.

Fig. 8: Qualitative comparisons of animation results. Our model enables the generation of disentangled human body and clothes meshes, which thus enables more photorealistic animations.

Favourite (%, ↑)	TEXture [74]	TADA [51]	Ours
Texture & Geometry Quality	8.1	8.4	**83.4**
Animation Realism	3.4	7.0	**89.7**
Prompt Alignment	5.5	7.7	**86.8**

Table 2: User study from 26 users in terms of visual quality, animation quality, and prompt alignment preference. The number means the percentage of users favoring a specific method.

5.3   Quantitative Comparisons

Accurate evaluation of text-to-3D generation is a challenging task, which is rarely discussed in existing works [96]. Here, we use 3 metrics to evaluate generation fidelity: FID [31, 66], CLIP-FID [48] and CLIP Score [30]. CLIP Score is the CLIP feature distance between the input prompts and the rendered images of the generated avatars. For FID and CLIP-FID, we measure the distance between the rendered images from our generated 3D avatars, and images generated from Stable Diffusion [75] using the same prompts. We evaluate all methods using 30 prompts and render 300 views for each prompt, resulting in 9000 images. As shown in Table 1, our method achieves the best results in all 3 metrics compared with previous methods, demonstrating the high visual quality and textual alignment of our pipeline.

5.4   User Study

We also conduct a comprehensive user study to evaluate both the generation quality and the animation capability of our pipeline. We design a questionnaire and present 16 full-body animations made by TEXture [74], TADA! [51], and our method. The survey is conducted among 26 researchers and artists from either academia or industry. The participants are asked to pick the one with the best semantic consistency, visual quality, and animation quality. Note that we ask participants to exclude animation quality when evaluating texture& geometry quality. As shown in Table 2, our model is the most preferred one in all aspects of generation quality.

5.5   Ablation Study

We first evaluate the effects of our albedo smoothness constraints and 2D-pose ControlNet guidance. In Fig. 9, we compare the rendered albedo and RGB images w/ or w/o our albedo smoothness constraint $\mathcal{L}_a$. As can be seen from the visual results, the albedo generated without $\mathcal{L}_a$ shows baked-in shadows and wrinkles, which leads to unsatisfactory visual artifacts in animations. In comparison, using $\mathcal{L}_a$ largely alleviates the problem in the albedo while still enabling the correct rendering of the shadows and wrinkles on RGB images.

In Fig. 9, we validate the effectiveness of OpenPose-based ControlNet as guidance during our human body generation. As can be observed from the results, generated avatars without ControlNet have over-smoothed textures and geometry. In contrast, our utilization of ControlNet produces significantly finer details in both geometry and texture.

w/o $\mathcal{L}_a$ w/ $\mathcal{L}_a$ w/o $\mathcal{L}_a$ w/ $\mathcal{L}_a$ w/o ControlNet   w/ ControlNet   w/o ControlNet   w/ ControlNet

Fig. 9: Ablation study on our shading model (left) and ControlNet as guidance (right).

In Fig. 10 we demonstrate the necessity of using both guidances from clothes renders, $\{\mathbf{I}_c, \mathbf{N}_c\}$, and clothed human renders, $\{\mathbf{I}_{c+h}, \mathbf{N}_{c+h}\}$. Applying $\{\mathbf{I}_{c+h}, \mathbf{N}_{c+h}\}$ without clothes leads to the incorrect generation of skin colors on the clothes meshes, as seen on the left of Fig. 10. On the other hand, with only clothes guidance, the garments are prone to learn texture and semantics incoherent with the human body, such as the back collar on the right of Fig. 10.

5.6   Applications

Complex garments generation. Our SO-SMPL design can also be generalized to more complex garment types. For instance, by adding additional layers of

w/o $\mathbf{I}_c, \mathbf{N}_c$ w/ $\mathbf{I}_c, \mathbf{N}_c$ w/o $\mathbf{I}_{c+h}, \mathbf{N}_{c+h}$ w/ $\mathbf{I}_{c+h}, \mathbf{N}_{c+h}$

Fig. 10: Ablation study on guidance from clothes renders $\{\mathbf{I}_c, \mathbf{N}_c\}$ and clothed human renders $\{\mathbf{I}_{c+h}, \mathbf{N}_{c+h}\}$.

offsets on top of the first clothing layer, we can generate more complex multi-layer garments, as shown in Fig. 11. Besides, leveraging BCNet [42], a parametric skirts model conditioned on SMPL parameters, our pipeline can be extended to generating clothes that do not follow SMPL-X's topological structure such as skirts and dresses, as depicted in Fig. 11. More details can be found in the supplementary material.

"denim skirts"   "Scottish kilt"   "navy pleated skirts"   "coral skirts with floral details"   "layered green skirts"   "turquoise skirts with tropical print"   "yellow A-line skirt"

"burgundy shirt"   "charcoal blazer on burgundy shirt"   "deep navy vest on burgundy shirt"

Fig. 11: Complex garments generation. Our pipeline could be generalized to generate more complex garments like skirts and multi-layered garments.

Virtual try-on. Our disentangled human-clothes representation inherently allows us to change the outfit of a certain avatar, or put the same clothes on different avatars. As depicted in Fig. 12, while previous methods like TADA! [51] also enable virtual try-on, they struggle to maintain consistency in clothing and human identity during the try-on process. As can be seen from the left half of Fig. 12, when changing clothes for the white teenage boy with TADA!, the facial details and body shape are changed; on the other hand, the colors and patterns of the clothes also changed when putting the same clothes on different avatars with TADA!. On the contrary, our disentangled representation supports fixing the human identity and clothes details when switching outfits.
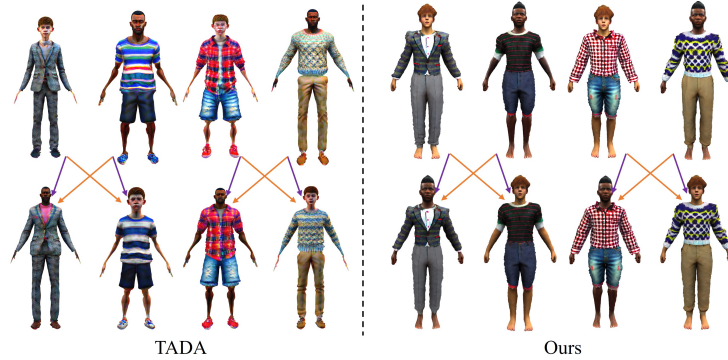
Fig. 12: Comparisons in virtual try-on applications. The orange and purple arrows represent compositions of the human body and the clothes, respectively. Changing clothes in TADA! [51] often leads to undesirable changes in human identity or clothes while changing clothes in our method is more photorealistic and keeps the same human identity and clothes.

6   Limitations&Conlusion

Limitations. Despite promising results, our pipeline still has limitations. Since images generated by diffusion models are of final shadings, it is difficult to get fully shadow-free material decomposition even with our shading model and albedo smoothness constraint. Besides, our results still suffer from cartoonish colors and over-smoothing textures, which is a common problem for distillation-based frameworks. Another limitation is the over-saturation of colors in our distillation-based framework. Although distillation techniques [92] have aimed to address this issue, it remains a challenge. Our offset-based representation also can not handle avatars with long hair, which requires advanced 3D representations [81]. What's more, our generated clothes do not follow any sewing patterns and are without any physical attributes such as stretch, bending, and friction. Therefore, generating garments with sewing patterns is a promising direction for exploration. We leave these potential tasks to future works.

Conclusion. In this paper, we have presented an innovative approach for generating human avatars through our Sequentially Offset-SMPL (SO-SMPL) representation from textual descriptions. To the best of our knowledge, our pipeline is the first to produce avatars in a disentangled manner: first generate a human body mesh then a clothes mesh on top of it. The 3D avatars and clothes generated by our pipeline exhibit remarkable diversity and high fidelity in both texture and geometric detailing. They could also be easily utilized in CG software for animation and simulation, thereby opening up exciting possibilities in fields such as virtual reality, gaming, and digital fashion. Experiments demonstrate that our pipeline outperforms existing text-to-3D generation methods in texture&geometry quality, alignment with text descriptions, and animation quality.

Acknowledgements

References

1. Clo3d, 2023. `https://www.clo3d.com/`
2. Marvelous designer, 2023. `https://www.marvelousdesigner.com`
3. Alexanderson, S., Nagy, R., Beskow, J., Henter, G.E.: Listen, denoise, action! audio-driven motion synthesis with diffusion models. ACM Transactions on Graphics (TOG) **42**(4), 1–20 (2023)
4. Anonymous: Avatarstudio: High-fidelity and animatable 3d avatar creation from text (2023)
5. Armandpour, M., Zheng, H., Sadeghian, A., Sadeghian, A., Zhou, M.: Re-imagine the negative prompt algorithm: Transform 2d diffusion into 3d, alleviate janus problem and beyond. arXiv preprint arXiv:2304.04968 (2023)
6. Bautista, M.A., Guo, P., Abnar, S., Talbott, W., Toshev, A., Chen, Z., Dinh, L., Zhai, S., Goh, H., Ulbricht, D., et al.: Gaudi: A neural architect for immersive 3d scene generation. Advances in Neural Information Processing Systems **35**, 25102–25116 (2022)
7. Bergman, A.W., Kellnhofer, P., Wang, Y., Chan, E.R., Lindell, D.B., Wetzstein, G.: Generative neural articulated radiance fields. In: NeurIPS (2022)
8. Cao, T., Kreis, K., Fidler, S., Sharp, N., Yin, K.: Texfusion: Synthesizing 3d textures with text-guided image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4169–4181 (2023)
9. Cao, Y., Cao, Y.P., Han, K., Shan, Y., Wong, K.Y.K.: Dreamavatar: Text-and-shape guided 3d human avatar generation via diffusion models. arXiv preprint arXiv:2304.00916 (2023)
10. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7291–7299 (2017)
11. Chan, E.R., Lin, C.Z., Chan, M.A., Nagano, K., Pan, B., De Mello, S., Gallo, O., Guibas, L.J., Tremblay, J., Khamis, S., et al.: Efficient geometry-aware 3d generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16123–16133 (2022)
12. Chen, D.Z., Siddiqui, Y., Lee, H.Y., Tulyakov, S., Nießner, M.: Text2tex: Text-driven texture synthesis via diffusion models. arXiv preprint arXiv:2303.11396 (2023)
13. Chen, R., Chen, Y., Jiao, N., Jia, K.: Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. arXiv preprint arXiv:2303.13873 (2023)

14. Chen, X., Su, Z., Yang, L., Cheng, P., Xu, L., Fu, B., Yu, G.: Learning variational motion prior for video-based motion capture. arXiv preprint arXiv:2210.15134 (2022)
15. Chen, X., Jiang, T., Song, J., Yang, J., Black, M.J., Geiger, A., Hilliges, O.: gdna: Towards generative detailed neural avatars. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20427–20437 (2022)
16. Chen, Y., Zhang, C., Yang, X., Cai, Z., Yu, G., Yang, L., Lin, G.: It3d: Improved text-to-3d generation with explicit view synthesis. arXiv preprint arXiv:2308.11473 (2023)
17. Chen, Z., Ding, C., Guo, J., Wang, D., Li, Y., Xiao, X., Wu, W., Song, L.: L-tracing: Fast light visibility estimation on neural surfaces by sphere tracing. In: Proceedings of the European Conference on Computer Vision (ECCV) (2022)
18. Cong, P., Dou, Z.W., Ren, Y., Yin, W., Cheng, K., Sun, Y., Long, X., Zhu, X., Ma, Y.: Laserhuman: Language-guided scene-aware human motion generation in free environment. arXiv preprint arXiv:2403.13307 (2024)
19. Corona, E., Pumarola, A., Alenya, G., Pons-Moll, G., Moreno-Noguer, F.: Smplicit: Topology-aware generative model for clothed people. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11875–11885 (2021)
20. Deitke, M., Liu, R., Wallingford, M., Ngo, H., Michel, O., Kusupati, A., Fan, A., Laforte, C., Voleti, V., Gadre, S.Y., et al.: Objaverse-xl: A universe of 10m+ 3d objects. arXiv preprint arXiv:2307.05663 (2023)
21. Deitke, M., Schwenk, D., Salvador, J., Weihs, L., Michel, O., VanderBilt, E., Schmidt, L., Ehsani, K., Kembhavi, A., Farhadi, A.: Objaverse: A universe of annotated 3d objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13142–13153 (2023)
22. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)
23. Dou, Z., Chen, X., Fan, Q., Komura, T., Wang, W.: C·ase: Learning conditional adversarial skill embeddings for physics-based characters. arXiv preprint arXiv:2309.11351 (2023)
24. Dou, Z., Wu, Q., Lin, C., Cao, Z., Wu, Q., Wan, W., Komura, T., Wang, W.: Tore: Token reduction for efficient human mesh recovery with transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15143–15155 (2023)
25. Downs, L., Francis, A., Koenig, N., Kinman, B., Hickman, R., Reymann, K., McHugh, T.B., Vanhoucke, V.: Google scanned objects: A high-quality dataset of 3d scanned household items. In: 2022 International Conference on Robotics and Automation (ICRA). pp. 2553–2560. IEEE (2022)
26. Feng, Y., Liu, W., Bolkart, T., Yang, J., Pollefeys, M., Black, M.J.: Learning disentangled avatars with hybrid 3d representations. arXiv preprint arXiv:2309.06441 (2023)
27. Feng, Y., Yang, J., Pollefeys, M., Black, M.J., Bolkart, T.: Capturing and animation of body and clothing from monocular video. In: SIGGRAPH Asia 2022 Conference Papers. pp. 1–9 (2022)
28. Grigorev, A., Iskakov, K., Ianina, A., Bashirov, R., Zakharkin, I., Vakhitov, A., Lempitsky, V.: Stylepeople: A generative model of fullbody human avatars. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5151–5160 (2021)

29. Guo, Y., Zuo, X., Dai, P., Lu, J., Wu, X., Cheng, L., Yan, Y., Xu, S., Wu, X.: Decorate3d: Text-driven high-quality texture generation for mesh decoration in the wild. In: Thirty-seventh Conference on Neural Information Processing Systems (2023)
30. Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning. arXiv preprint arXiv:2104.08718 (2021)
31. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
32. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
33. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
34. Hong, F., Chen, Z., Lan, Y., Pan, L., Liu, Z.: Eva3d: Compositional 3d human generation from 2d image collections. arXiv preprint arXiv:2210.04888 (2022)
35. Hong, F., Zhang, M., Pan, L., Cai, Z., Yang, L., Liu, Z.: Avatarclip: Zero-shot text-driven generation and animation of 3d avatars. arXiv preprint arXiv:2205.08535 (2022)
36. Hu, S., Hong, F., Hu, T., Pan, L., Mei, H., Xiao, W., Yang, L., Liu, Z.: Humanliff: Layer-wise 3d human generation with diffusion model. arXiv preprint arXiv:2308.09712 (2023)
37. Huang, S., Yang, Z., Li, L., Yang, Y., Jia, J.: Avatarfusion: Zero-shot generation of clothing-decoupled 3d avatars using 2d diffusion. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 5734–5745 (2023)
38. Huang, X., Shao, R., Zhang, Q., Zhang, H., Feng, Y., Liu, Y., Wang, Q.: Humannorm: Learning normal diffusion model for high-quality and realistic 3d human generation. arXiv preprint arXiv:2310.01406 (2023)
39. Huang, Y., Yi, H., Xiu, Y., Liao, T., Tang, J., Cai, D., Thies, J.: Tech: Text-guided reconstruction of lifelike clothed humans. arXiv preprint arXiv:2308.08545 (2023)
40. Huang, Y., Wang, J., Shi, Y., Qi, X., Zha, Z.J., Zhang, L.: Dreamtime: An improved optimization strategy for text-to-3d content creation. arXiv preprint arXiv:2306.12422 (2023)
41. Huang, Y., Wang, J., Zeng, A., Cao, H., Qi, X., Shi, Y., Zha, Z.J., Zhang, L.: Dreamwaltz: Make a scene with complex 3d animatable avatars. arXiv preprint arXiv:2305.12529 (2023)
42. Jiang, B., Zhang, J., Hong, Y., Luo, J., Liu, L., Bao, H.: Bcnet: Learning body and cloth shape from a single image. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16. pp. 18–35. Springer (2020)
43. Jun, H., Nichol, A.: Shap-e: Generating conditional 3d implicit functions. arXiv preprint arXiv:2305.02463 (2023)
44. Kanazawa, A., Tulsiani, S., Efros, A.A., Malik, J.: Learning category-specific mesh reconstruction from image collections. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 371–386 (2018)
45. Karnewar, A., Vedaldi, A., Novotny, D., Mitra, N.J.: Holodiffusion: Training a 3d diffusion model using 2d images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18423–18433 (2023)
46. Kim, B., Kwon, P., Lee, K., Lee, M., Han, S., Kim, D., Joo, H.: Chupa: Carving 3d clothed humans from skinned shape priors using 2d diffusion probabilistic models. arXiv preprint arXiv:2305.11870 (2023)

47. Kolotouros, N., Alldieck, T., Zanfir, A., Bazavan, E.G., Fieraru, M., Sminchisescu, C.: Dreamhuman: Animatable 3d avatars from text. arXiv preprint arXiv:2306.09329 (2023)
48. Kynkäänniemi, T., Karras, T., Aittala, M., Aila, T., Lehtinen, J.: The role of imagenet classes in fr\'echet inception distance. arXiv preprint arXiv:2203.06026 (2022)
49. Li, J., Xu, C., Chen, Z., Bian, S., Yang, L., Lu, C.: 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. pp. 3383–3393 (2021)
50. Li, R., Yang, S., Ross, D.A., Kanazawa, A.: Learn to dance with aist++: Music conditioned 3d dance generation (2021)
51. Liao, T., Yi, H., Xiu, Y., Tang, J., Huang, Y., Thies, J., Black, M.J.: TADA! Text to Animatable Digital Avatars. In: International Conference on 3D Vision (3DV) (2024)
52. Lin, C.H., Gao, J., Tang, L., Takikawa, T., Zeng, X., Huang, X., Kreis, K., Fidler, S., Liu, M.Y., Lin, T.Y.: Magic3d: High-resolution text-to-3d content creation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 300–309 (2023)
53. Liu, R., Wu, R., Van Hoorick, B., Tokmakov, P., Zakharov, S., Vondrick, C.: Zero-1-to-3: Zero-shot one image to 3d object. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9298–9309 (2023)
54. Liu, X., Zhan, X., Tang, J., Shan, Y., Zeng, G., Lin, D., Liu, X., Liu, Z.: Humangaussian: Text-driven 3d human generation with gaussian splatting. arXiv preprint arXiv:2311.17061 (2023)
55. Liu, Y., Lin, C., Zeng, Z., Long, X., Liu, L., Komura, T., Wang, W.: Syncdreamer: Generating multiview-consistent images from a single-view image. arXiv preprint arXiv:2309.03453 (2023)
56. Liu, Z., Feng, Y., Black, M.J., Nowrouzezahrai, D., Paull, L., Liu, W.: Meshdiffusion: Score-based generative 3d mesh modeling. arXiv preprint arXiv:2303.08133 (2023)
57. Long, X., Guo, Y.C., Lin, C., Liu, Y., Dou, Z., Liu, L., Ma, Y., Zhang, S.H., Habermann, M., Theobalt, C., et al.: Wonder3d: Single image to 3d using cross-domain diffusion. arXiv preprint arXiv:2310.15008 (2023)
58. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. ACM Trans. Graphics (Proc. SIGGRAPH Asia) **34**(6), 248:1–248:16 (Oct 2015)
59. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: AMASS: Archive of motion capture as surface shapes. In: International Conference on Computer Vision. pp. 5442–5451 (Oct 2019)
60. Michel, O., Bar-On, R., Liu, R., Benaim, S., Hanocka, R.: Text2mesh: Text-driven neural stylization for meshes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13492–13502 (2022)
61. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)
62. Nichol, A., Jun, H., Dhariwal, P., Mishkin, P., Chen, M.: Point-e: A system for generating 3d point clouds from complex prompts. arXiv preprint arXiv:2212.08751 (2022)

63. Nicki Skafte Detlefsen, Jiri Borovec, Justus Schock, Ananya Harsh, Teddy Koker, Luca Di Liello, Daniel Stancl, Changsheng Quan, Maxim Grechkin, William Falcon: TorchMetrics - Measuring Reproducibility in PyTorch (Feb 2022). https://doi.org/10.21105/joss.04101, https://github.com/Lightning-AI/torchmetrics

64. Noguchi, A., Sun, X., Lin, S., Harada, T.: Neural articulated radiance field. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5762–5772 (2021)

65. Noguchi, A., Sun, X., Lin, S., Harada, T.: Unsupervised learning of efficient geometry-aware neural articulated representations. arXiv:2204.08839 (2022)

66. Parmar, G., Zhang, R., Zhu, J.Y.: On aliased resizing and surprising subtleties in gan evaluation. In: CVPR (2022)

67. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A., Tzionas, D., Black, M.J.: 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. pp. 10975–10985 (2019)

68. Peng, H.Y., Zhang, J.P., Guo, M.H., Cao, Y.P., Hu, S.M.: Charactergen: Efficient 3d character generation from single images with multi-view pose canonicalization. arXiv preprint arXiv:2402.17214 (2024)

69. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. arXiv (2022)

70. Qian, G., Mai, J., Hamdi, A., Ren, J., Siarohin, A., Li, B., Lee, H.Y., Skorokhodov, I., Wonka, P., Tulyakov, S., et al.: Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. arXiv preprint arXiv:2306.17843 (2023)

71. Qian, Z., Wang, S., Mihajlovic, M., Geiger, A., Tang, S.: 3dgs-avatar: Animatable avatars via deformable 3d gaussian splatting. arXiv preprint arXiv:2312.09228 (2023)

72. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)

73. Reizenstein, J., Shapovalov, R., Henzler, P., Sbordone, L., Labatut, P., Novotny, D.: Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10901–10911 (2021)

74. Richardson, E., Metzer, G., Alaluf, Y., Giryes, R., Cohen-Or, D.: Texture: Text-guided texturing of 3d shapes. arXiv preprint arXiv:2302.01721 (2023)

75. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models (2021)

76. Seo, H., Kim, H., Kim, G., Chun, S.Y.: Ditto-nerf: Diffusion-based iterative text to omni-directional 3d model. arXiv preprint arXiv:2304.02827 (2023)

77. Seo, J., Jang, W., Kwak, M.S., Ko, J., Kim, H., Kim, J., Kim, J.H., Lee, J., Kim, S.: Let 2d diffusion model know 3d-consistency for robust text-to-3d generation. arXiv preprint arXiv:2303.07937 (2023)

78. Shen, T., Gao, J., Yin, K., Liu, M.Y., Fidler, S.: Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. Advances in Neural Information Processing Systems **34**, 6087–6101 (2021)

79. Shi, M., Aberman, K., Aristidou, A., Komura, T., Lischinski, D., Cohen-Or, D., Chen, B.: Motionet: 3d human motion reconstruction from monocular video with skeleton consistency. ACM Transactions on Graphics (TOG) **40**(1), 1–15 (2020)

80. Shi, Y., Wang, P., Ye, J., Long, M., Li, K., Yang, X.: Mvdream: Multi-view diffusion for 3d generation. arXiv preprint arXiv:2308.16512 (2023)
81. Sklyarova, V., Zakharov, E., Hilliges, O., Black, M.J., Thies, J.: Haar: Text-conditioned generative model of 3d strand-based human hairstyles. ArXiv (Dec 2023)
82. Svitov, D., Gudkov, D., Bashirov, R., Lempitsky, V.: Dinar: Diffusion inpainting of neural textures for one-shot human avatars. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7062–7072 (2023)
83. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-Or, D., Bermano, A.H.: Human motion diffusion model. arXiv preprint arXiv:2312.02256 (2023)
84. Tsalicoglou, C., Manhardt, F., Tonioni, A., Niemeyer, M., Tombari, F.: Textmesh: Generation of realistic 3d meshes from text prompts. arXiv preprint arXiv:2304.12439 (2023)
85. Wan, W., Dou, Z., Komura, T., Wang, W., Jayaraman, D., Liu, L.: Tlcontrol: Trajectory and language control for human motion synthesis. arXiv preprint arXiv:2311.17135 (2023)
86. Wang, H., Du, X., Li, J., Yeh, R.A., Shakhnarovich, G.: Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12619–12629 (2023)
87. Wang, J., Chen, Z., Ling, J., Xie, R., Song, L.: 360-degree panorama generation from few unregistered nfov images. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 6811–6821 (2023)
88. Wang, P., Liu, L., Liu, Y., Theobalt, C., Komura, T., Wang, W.: Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. arXiv preprint arXiv:2106.10689 (2021)
89. Wang, T., Zhang, B., Zhang, T., Gu, S., Bao, J., Baltrusaitis, T., Shen, J., Chen, D., Wen, F., Chen, Q., et al.: Rodin: A generative model for sculpting 3d digital avatars using diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4563–4573 (2023)
90. Wang, W., Ge, Y., Mei, H., Cai, Z., Sun, Q., Wang, Y., Shen, C., Yang, L., Komura, T.: Zolly: Zoom focal length correctly for perspective-distorted human mesh reconstruction. arXiv preprint arXiv:2303.13796 (2023)
91. Wang, Y., Ma, J., Shao, R., Feng, Q., Lai, Y.K., Liu, Y., Li, K.: Humancoser: Layered 3d human generation via semantic-aware diffusion model. arXiv preprint arXiv:2312.05804 (2023)
92. Wang, Z., Lu, C., Wang, Y., Bao, F., Li, C., Su, H., Zhu, J.: Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. arXiv preprint arXiv:2305.16213 (2023)
93. Watson, D., Chan, W., Martin-Brualla, R., Ho, J., Tagliasacchi, A., Norouzi, M.: Novel view synthesis with diffusion models. arXiv preprint arXiv:2210.04628 (2022)
94. Weng, Z., Wang, Z., Yeung, S.: Zeroavatar: Zero-shot 3d avatar generation from a single image. arXiv preprint arXiv:2305.16411 (2023)
95. Wu, J., Gao, X., Liu, X., Shen, Z., Zhao, C., Feng, H., Liu, J., Ding, E.: Hd-fusion: Detailed text-to-3d generation leveraging multiple noise estimation. arXiv preprint arXiv:2307.16183 (2023)
96. Wu, T., Yang, G., Li, Z., Zhang, K., Liu, Z., Guibas, L., Lin, D., Wetzstein, G.: Gpt-4v(ision) is a human-aligned evaluator for text-to-3d generation. arXiv preprint arXiv:2401.04092 (2024)

97. Xiu, Y., Yang, J., Cao, X., Tzionas, D., Black, M.J.: Econ: Explicit clothed humans optimized via normal integration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 512–523 (2023)

98. Xu, Y., Yifan, W., Bergman, A.W., Chai, M., Zhou, B., Wetzstein, G.: Efficient 3d articulated human generation with layered surface volumes. arXiv preprint arXiv:2307.05462 (2023)

99. Xu, Y., Yang, Z., Yang, Y.: Seeavatar: Photorealistic text-to-3d avatar generation with constrained geometry and appearance. arXiv preprint arXiv:2312.08889 (2023)

100. Yang, Z., Cai, Z., Mei, H., Liu, S., Chen, Z., Xiao, W., Wei, Y., Qing, Z., Wei, C., Dai, B., Wu, W., Qian, C., Lin, D., Liu, Z., Yang, L.: Synbody: Synthetic dataset with layered human models for 3d human perception and modeling (2023)

101. Youwang, K., Oh, T.H.: Text-driven human avatar generation by neural reparameterized texture optimization

102. Yu, C., Zhou, Q., Li, J., Zhang, Z., Wang, Z., Wang, F.: Points-to-3d: Bridging the gap between sparse points and shape-controllable text-to-3d generation. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 6841–6850 (2023)

103. Yu, J., Zhu, H., Jiang, L., Loy, C.C., Cai, W., Wu, W.: Painthuman: Towards high-fidelity text-to-3d human texturing via denoised score distillation. arXiv preprint arXiv:2310.09458 (2023)

104. Yu, X., Xu, M., Zhang, Y., Liu, H., Ye, C., Wu, Y., Yan, Z., Zhu, C., Xiong, Z., Liang, T., et al.: Mvimgnet: A large-scale dataset of multi-view images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9150–9161 (2023)

105. Yu, Z., Cheng, W., Liu, X., Wu, W., Lin, K.Y.: Monohuman: Animatable human neural field from monocular video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16943–16953 (2023)

106. Yu, Z., Dou, Z., Long, X., Lin, C., Li, Z., Liu, Y., Müller, N., Komura, T., Habermann, M., Theobalt, C., et al.: Surf-d: High-quality surface generation for arbitrary topologies using diffusion models. arXiv preprint arXiv:2311.17050 (2023)

107. Zhang, H., Feng, Y., Kulits, P., Wen, Y., Thies, J., Black, M.J.: Text-guided generation and editing of compositional 3d avatars. arXiv preprint arXiv:2309.07125 (2023)

108. Zhang, H., Yuan, Y., Makoviychuk, V., Guo, Y., Fidler, S., Peng, X.B., Fatahalian, K.: Learning physically simulated tennis skills from broadcast videos. ACM Transactions On Graphics (TOG) **42**(4), 1–14 (2023)

109. Zhang, H., Chen, B., Yang, H., Qu, L., Wang, X., Chen, L., Long, C., Zhu, F., Du, K., Zheng, M.: Avatarverse: High-quality & stable 3d avatar creation from text and pose. arXiv preprint arXiv:2308.03610 (2023)

110. Zhang, J., Jiang, Z., Yang, D., Xu, H., Shi, Y., Song, G., Xu, Z., Wang, X., Feng, J.: Avatargen: a 3d generative model for animatable human avatars. In: European Conference on Computer Vision. pp. 668–685. Springer (2022)

111. Zhang, J., Sangineto, E., Tang, H., Siarohin, A., Zhong, Z., Sebe, N., Wang, W.: 3d-aware semantic-guided generative model for human synthesis. In: European Conference on Computer Vision. pp. 339–356. Springer (2022)

112. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847 (2023)

113. Zhang, X., Srinivasan, P.P., Deng, B., Debevec, P., Freeman, W.T., Barron, J.T.: Nerfactor: Neural factorization of shape and reflectance under an unknown illumination. ACM Transactions on Graphics (ToG) **40**(6), 1–18 (2021)
114. Zhang, X., Zhang, J., Chacko, R., Xu, H., Song, G., Yang, Y., Feng, J.: Getavatar: Generative textured meshes for animatable human avatars. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2273–2282 (2023)
115. Zhou, W., Dou, Z., Cao, Z., Liao, Z., Wang, J., Wang, W., Liu, Y., Komura, T., Wang, W., Liu, L.: Emdm: Efficient motion diffusion model for fast, high-quality motion generation. arXiv preprint arXiv:2309.11351 (2023)
116. Zhu, J., Zhuang, P.: Hifa: High-fidelity text-to-3d with advanced diffusion guidance. arXiv preprint arXiv:2305.18766 (2023)