2024 International Conference on 3D Vision (3DV) | 979-8-3503-6245-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/3DV62453.2024.00150

TADA! Text to Animatable Digital Avatars

Tingting Liao[1*], Hongwei Yi[2*], Yuliang Xiu[2], Jiaxiang Tang[3], Yangyi Huang[4]
Justus Thies[2], Michael J. Black[2]
[1]Mohamed bin Zayed University of Artificial Intelligence [2]Max Planck Institute for Intelligent Systems
[3]Peking University [4]State Key Lab of CAD & CG, Zhejiang University
tingting.liao@mbzuai.ac.ae, tjx@pku.edu.cn, huangyangyi@zju.edu.cn
{hongwei.yi, yuliang.xiu, justus.thies, black}@tuebingen.mpg.de

Figure 1. With only text descriptions as input, TADA generates high-fidelity 3D avatars with lifelike texture and detailed geometry, including high-resolution faces. Accurate alignment of texture and geometry, together with an underlying SMPL-X representation, enables expressive animation. TADA also supports applications such as virtual try-on and personalized editing using text.

Abstract

We introduce TADA, a simple-yet-effective approach that takes textual descriptions and produces expressive 3D avatars with high-quality geometry and lifelike textures, that can be animated and rendered with traditional graphics pipelines. Existing text-based character generation methods are limited in terms of geometry and texture quality, and cannot be realistically animated due to the misalignment between the geometry and the texture, particularly in the face region. To address these limitations, TADA leverages the synergy of a 2D diffusion model and a parametric body model. Specifically, we derive a high-resolution upsampled version of SMPL-X with a displacement layer and a texture map, and use hierarchical rendering with score distillation sampling (SDS) to create high-quality, detailed, holistic 3D avatars from text. To ensure alignment between the geometry and texture, we render normals and RGB images of the generated character and exploit their latent embeddings during the SDS optimization process. We further drive the character's face with multiple expressions during optimization, ensuring that its semantics remain consistent with the original SMPL-X model. Both qualitative and quantitative evaluations show that TADA significantly surpasses existing approaches. TADA enables large-scale creation of digital characters ready for animation and rendering, while also enabling text-guided editing. The code is public for research purposes at tada.is.tue.mpg.de.

[*]denotes equal contribution.

979-8-3503-6245-9/24/$31.00 ©2024 IEEE
DOI 10.1109/3DV62453.2024.00150

1. Introduction

Digital avatars are a foundation for AR/VR applications, such as immersive telepresence [27, 28, 51, 72, 80], virtual try-on [53, 54, 82], and video games [15, 79, 83]. Creating high-quality and expressive 3D avatars can be challenging due to the need to model both the geometry and appearance under various poses. Capture of high-quality 3D human scans [3, 22, 33], in particular, requires expensive scanning equipment, such as 4D scanners or light stages, and significant artist effort for cleaning, repairing and rigging the captured data. While there is recent progress on automatic learning-based body reconstruction from a single image [19, 20, 28, 29, 50, 51, 71, 72, 74], or sparse set of images [56], such methods are limited to real humans, fail on fictional characters, and struggle with editing.

Due to rapid progress in Large Language Models [4, 44] and Diffusion Vision Models [17, 60, 61, 62, 70], text-to-image models [45, 47] can be combined with differentiable rendering and 3D neural representations (*e.g.*, DeepSDF [39], NeRF [36], DMTet [57]) to generate realistic 3D shapes solely from textual descriptions. However, these methods have certain limitations when it comes to 3D humans. The 3D humans generated are rigid one-piece shapes without a rig for animation [7, 14, 30, 35, 46]. The generated geometry and texture are of poor quality, inconsistently aligned (see Fig. 2 (a,b)) [7, 25, 41, 46], and lack fine details [18]. And the final characters are incompatible with traditional graphics workflows (*e.g.*, NeRF representations [5, 25, 41]).

Here, we address these limitations with TADA, illustrated in Fig. 1. We have **three goals** for the generated 3D humans: (i) animatable and compatible with graphics engines (*i.e.*, LBS-based animation). (ii) diverse clothing and textures. (iii) semantically consistent with SMPL-X so that it can be correctly animated; that is, the body parts and vertices of the output avatars correspond to the same body parts and vertices on the SMPL-X mesh.

To achieve these goals, we build on SMPL-X [40], which is a commonly used parametric body model with minimal clothing, and then apply Score Distillation Sampling (SDS) [41] on the additional displacement layer of SMPL-X and UV texture map, for rich clothing deformations and textures. Specifically, we make **three key contributions**. (i) First, we devise a hierarchical optimization on a subdivided version of SMPL-X [40] with additional learnable displacements and a texture map. By optimizing the entire body hierarchically with varying focal lengths, we can achieve high-quality details, especially in the facial region. (ii) Secondly, to address the inconsistent alignment issue, TADA blends the RGB and normal maps in the latent space before computing the SDS loss. This approach helps to generate high-quality 3D avatars with well-aligned texture, see Fig. 2 (a,b). (iii) Third, to enforce the semantic consistency with SMPL-X, we introduce animations throughout the optimization process. Specifically, we deform the generated character in each optimization step by sampling predefined SMPL-X body poses and facial expressions. This ensures that our generated characters can be animated coherently, as depicted in Fig. 2 (c). In particular, once optimization is finished, our generated characters can be animated with any novel set of SMPL-X parameters. Especially, combined with existing text-to-motion generation methods [55, 63] or text-to-audio-to-motion methods [64, 75], we can animate the generated characters to interact with 3D scenes or communicate with other characters. This paves the way towards creating virtual 3D worlds with animatable digital avatars fully from text.

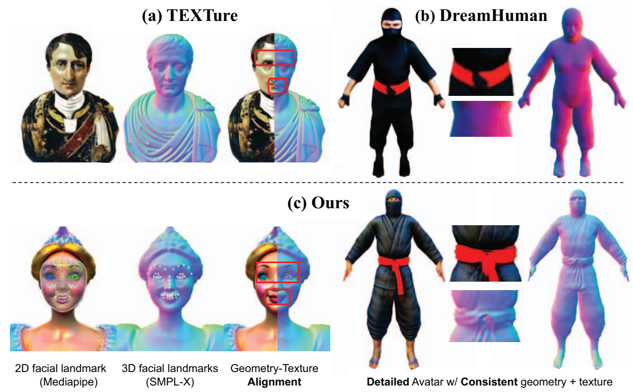

Figure 2. Compared with other existing methods [25, 41], our method can generate high-quality 3D avatars with well-aligned geometry and texture that are consistent with SMPL-X, enabling animation and rendering with existing graphics pipelines.

In summary, TADA is a user-friendly tool for avatar creation and editing. And it can be solely controlled by textual input, and the generated avatars are fully compatible with traditional graphics pipelines. The output model is graphics-ready because the underlying model is SMPL-X with displacements and a texture map. Our method can generate iconic celebrities, customized humans, and cartoon characters. We validate our contributions through ablation studies, provide qualitative comparisons with the state of the art, and conduct a perceptual study to quantify the performance of our method.

2. Related Work

Recently, there has been rapid progress on extending text-to-2D-image generation methods [11, 23, 48] to text-to-3D-content generation [34, 41, 68]. Here, we discuss the most relevant text-to-3D-content generation methods, while focusing primarily on text-to-3D-avatar generation (both human and anime characters).

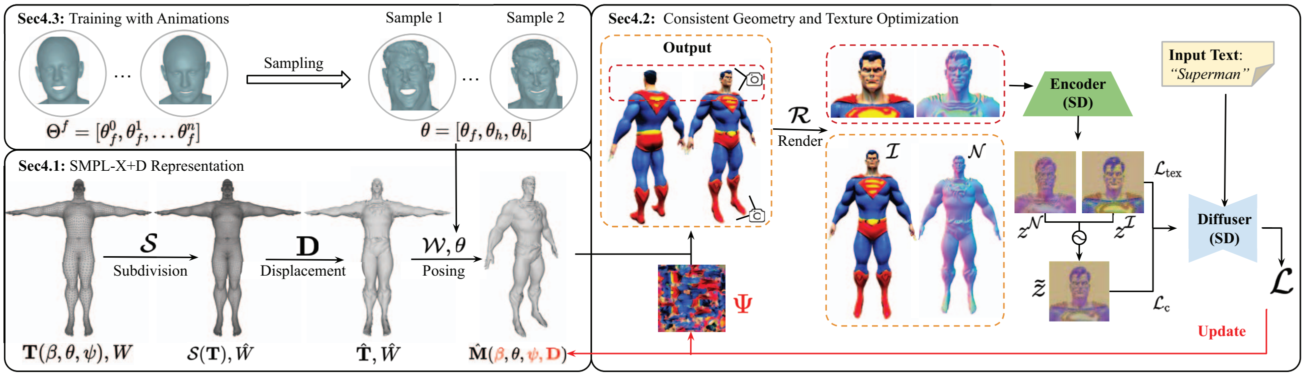

Figure 3. **Overview**. Initialized by a SMPL-X body $\mathbf{T}(\beta, \theta, \psi)$ with skinning weights W, we subdivide the body to obtain a denser mesh $\mathbf{S}(\mathbf{T})$ and add personalized displacements $\mathbf{D}$ to it. The personalized mesh $\hat{\mathbf{T}}$ is transformed into the posed space denoted as $\hat{\mathbf{M}}$ using randomly sampled expressions and poses from an animation database. In each optimization step, the expressions and poses are changed and $\hat{\mathbf{M}}$ is rendered under a novel view. Based on the rendered RGB $\mathbf{I}$ and normal image $\mathbf{N}$, the geometry and texture of the mesh are simultaneously optimized by a Score Distillation Sampling (SDS) loss.

Text-to-3D-Content Generation. The success of Text-to-Image (T2I) generative modeling [9, 47, 49] has sparked a surge of interest in the field of text-to-3D generation [8, 52, 65, 73]. Despite progress, effectively describing and controlling 3D properties of an object using language, while ensuring coherence in the 3D space, remains a challenge. One line of work [21, 35, 37] uses CLIP-space similarities to guide shape and texture optimization. However, these methods often fail to generate convincing and realistic 2D renderings. CLIP-based optimization can be combined with a generative appearance model to improve quality, as shown in CLIPFace [2]. However, this requires learning a GAN-model for the 3D appearance, which is challenging for full-body avatars that can vary from real humans to cartoon characters.

To avoid training a 3D generative model, and circumvent the issue of limited 3D training data, recent publications take advantage of the power of the 2D diffusion model [47] to create 3D content from texts. TEXTure [46] takes a mesh as input and optimizes the texture map based on a given text prompt. DreamFusion [41] introduces score distillation sampling (SDS), and uses it to optimize the 3D NeRF [36] to represent the 3D content of arbitrary (fictional) objects in terms of a density and radiance field. However, it faces challenges due to slow optimization of NeRF and low-resolution image space supervision, resulting in long processing times and low-quality 3D models. To overcome these limitations, Magic3D [30] introduces a two-stage optimization framework, using NeRF in the first stage and a textured mesh in the second stage. Fantasia3D [7] extends this to generate 3D meshes by disentangling geometry and texture, and optimizes them separately. These methods focus on general, static, object/scene generation; the output is not animation-ready, which is necessary for 3D character creation. In addition to the aforementioned "optimization via multiview SDS" approach, recent attention has been drawn to the "reconstruction via direct view-conditional generation" [6, 31, 32, 42, 69, 81].

Text-to-3D-Avatar Generation. Several methods generate 3D head avatars from text [13, 16, 67, 77, 78]. In contrast, we focus on generating full-body characters, including the detailed face. AvatarCLIP [18] leverages NeuS [66] and the SMPL-X model with a CLIP-guide loss to facilitate avatar generation. Similarly, DreamAvatar [5] uses the shape parameters of SMPL as a prior to learn a NeRF-based color field. DreamHuman [25] leverages imGHUM [1] as a prior, which represents a signed distance field conditioned on pose and shape parameters, to learn a NeRF of the human. However, the NeRF representation remains problematic due to its relatively low geometry and appearance quality, and it is incompatible with traditional graphics workflows, especially for animation. In the domain of explicit representations, Text2Mesh [35] and Chupa [24] employ vertex displacement on a predefined mesh template. However, the inherent limitation of fixed topology poses challenges in accurately generating diverse character shapes. TeCH [20] uses DMTet [58] as the 3D representation, and considers image-based reconstruction as a conditional generation task, taking conditions from both the input image and the derived descriptions, to create lifelike avatars, which are pixel-aligned with the single image. In contrast, our approach jointly optimizes the shape, expression, and displacement. Thus, the generated characters exhibit superior quality, can be easily animated with SMPL-X motions, and integrate seamlessly into existing rendering and animation workflows.

3. Preliminaries

SMPL-X [40] is an animatable parametric 3D body model that consists of the human body, face and hands. It has $N = 10,475$ vertices and $K = 54$ joints. Given the shape β, pose θ (including body joints pose θ_b, jaw pose θ_f and finger pose θ_h) and expression ψ parameters, SMPL-X models the human body as $\mathbf{M}(\beta, \theta, \psi)$:

$$\begin{aligned} \mathbf{M}(\beta,\theta,\psi) &= \mathcal{W}(\mathbf{T}(\beta,\theta,\psi), J(\beta), \theta, W) \\ \mathbf{T}(\beta,\theta,\psi) &= T + B_s(\beta) + B_e(\psi) + B_p(\theta), \end{aligned} \tag{1}$$

where T is a mean shape template, B_s, B_e and B_p are shape, expression and pose blend shapes, respectively. $\mathcal{W}$ is the linear blend-skinning function transforming $\mathbf{T}(\beta,\theta,\psi)$ to the target pose θ, with the skeleton joints $J(\beta)$ and skinning weights $W \in \mathbb{R}^{N\times K}$.

Score Distillation Sampling (SDS). DreamFusion proposes SDS [41] to utilize a pre-trained 2D diffusion model to optimize the parameters η of a 3D model, given a text y as input. Given the diffusion model ϕ with the noise prediction network $\hat{\epsilon}_\phi(x_t; y, t)$, SDS optimizes parameters η by directly minimizing the injected noise ϵ added to the rendered images $x = g(\eta)$ and the predicted noise:

$$\nabla_\eta \mathcal{L}_{SDS}(\phi, x) = E_{t,\epsilon}\left[w(t)(\hat{\epsilon}_\phi(x_t; y, t) - \epsilon)\frac{\partial x}{\partial \theta}\right], \tag{2}$$

where $g(\eta)$ denotes the differentiable rendering of the 3D model parameterized by η, x_t is the noised image, and $w(t)$ is a weighting function that depends on the noise level t.

4. Method

Given an input text prompt, TADA aims to generate a high-fidelity animatable full-body avatar. As illustrated in Fig. 3, our method initializes the 3D avatar with upsampled SMPL-X, which is parameterized with shape, pose, and expression parameters. Based on it, learnable displacements are incorporated, resulting in a "clothed" avatar with increased density (Sec. 4.1). Then, we optimize the 3D character with consistent geometry and texture using SDS losses that considers both the rendered normal and RGB images in the latent space (Sec. 4.2). To encourage semantic consistency with the SMPL-X, we sample different gestures and expressions during training (Sec. 4.3). This enables the future animation using the SMPL-X pose and expression space.

4.1. SMPL-X+D Representation

TADA adopts an SMPL-X+D to model animatable clothed avatars. The learnable displacement (D) accounts for personalized details that are independent of pose, shape, and expression. To generate a high-quality character with a detailed face, we apply a partial mesh subdivision on the original SMPL-X model, which is adapted as (Eq. (1)):

$$\begin{aligned} \hat{\mathbf{M}}(\beta,\theta,\psi,\mathbf{D}) &= \mathcal{W}(\hat{\mathbf{T}}(\beta,\theta,\psi,\mathbf{D}), J(\beta), \theta, \hat{W}) \\ \hat{\mathbf{T}}(\beta,\theta,\psi,\mathbf{D}) &= \mathcal{S}(\mathbf{T}(\beta,\theta,\psi)) + \mathbf{D}, \end{aligned} \tag{3}$$

where $\mathcal{S} : \mathbb{R}^{N\times 3} \to \mathbb{R}^{N_s\times 3}$ is the mesh subdivision operation, $\mathbf{D} \in \mathbb{R}^{N_s\times 3}$, $\hat{W} \in \mathbb{R}^{N_s\times J}$ and N_s are the vertex displacement, skinning weights and vertex number of the subdivided body, respectively. Besides the displacement $\mathbf{D}$, the parameters β, θ, ψ are also learnable. This helps to generate characters with various shapes, such as human-like or anime characters with largely deformed body shapes, like exaggerated proportions, elongated limbs, large eyes, *etc*.

Partial Mesh Subdivision. The vertices on the surface of the SMPL-X body are irregularly distributed, *i.e*., around 4,000 vertices for the head with the remaining 6,000 for the body. The sparsity of vertices on the body surface results in less detailed deformations there. Simply increasing the mesh density by subdividing the whole body mesh leads to noisy results, especially, in the face area during geometry optimization. To address this issue, we employ an adaptive upsampling technique on the triangles and interpolate their skinning weights within areas of low mesh density, such as the body region and the back of the head. This process yields a more refined mesh with uniformly distributed vertices and smoother skinning weights.

4.2. Consistent Geometry and Texture Learning

To generate animatable characters, we need to ensure the consistency between the geometry and the texture. Therefore, we propose to blend the SDS loss of the rendered normal and RGB images to achieve a well-aligned geometry and texture. Given a mesh $\hat{\mathbf{M}}$ parameterized by $\mathbf{D}$, β and ψ and albedo Ψ, we render its normal image $\mathcal{N}$ and colored image $\mathcal{I}$ using a differentiable render [26], denoted as $\mathcal{R}$:

$$\mathcal{N} = \mathcal{R}(\hat{\mathbf{M}}, \pi), \quad \mathcal{I} = \mathcal{R}(\Psi, \hat{\mathbf{M}}, \pi) \tag{4}$$

where π are the camera parameters. In each iteration, the camera is randomly positioned in one of two perspectives: a full-body view or a zoom-in head view. The head zoom-in allows us to reconstruct a detailed face region.

Texture SDS Objective. Given a text prompt, the texture generation is guided by a pretrained Stable Diffusion (SD) model [47], denoted as ϕ, which measures the similarity between the rendered image and the provided text prompt within the added and predicted noise space:

$$\nabla_\Psi \mathcal{L}_{\text{tex}}(\phi, \mathcal{I}) = \mathbb{E}_{t,\epsilon}\left[w(t)(\hat{\epsilon}_\phi(z_t^{\mathcal{I}}; y, t) - \epsilon)\frac{\partial \mathcal{I}}{\partial \Psi}\frac{\partial z^{\mathcal{I}}}{\partial \mathcal{I}}\right], \tag{5}$$

Figure 4. TADA enables holistic animation over the face, body, and hands. We show animation examples of the avatars "*Lionel Messi*" and "*Mabel Pines in Gravity Falls*" with full-body motions generated by TalkSHOW [75] (w/ TTS [*]) and PriorMDM [55], respectively.

where $z^{\mathcal{I}}$ is the latent feature of $\mathcal{I}$, encoded by image encoder (SD), $\hat{\epsilon}_\phi(z_t^{\mathcal{I}}; y, t)$ is the predicted noise given text embedding y and noise level t, ϵ is the pre-computed noise.

Geometry Consistency SDS Objective. Similarly, rendered normal images can be used for the diffusion model as shape encoding to facilitate the geometry synthesis. However, this approach may encounter challenges in ensuring perfect consistency between geometry and texture. To address this issue, we compute the SDS loss on the interpolation between normal and color image latents.

$$\nabla_\gamma \mathcal{L}_{\mathrm{c}}(\phi, x) = \mathbb{E}_{t,\epsilon}\left[w(t)(\hat{\epsilon}_\phi(\tilde{z}_t; y, t) - \epsilon)\frac{\partial \mathcal{N}}{\partial \gamma}\frac{\partial z}{\partial \mathcal{N}}\right], \tag{6}$$

where $\gamma = \{\beta, \psi, \mathbf{D}\}$ are the geometry related parameters, $\tilde{z} = \alpha z^{\mathcal{I}} + (1-\alpha)z^{\mathcal{N}}$ denotes the resulting interpolated latent code, while $z^{\mathcal{I}}$ and $z^{\mathcal{N}}$ represent the latent codes corresponding to the RGB and normal image, respectively.

Overall Optimization Objective. The learning objectives can be formulated as a combination of the texture SDS objective $\mathcal{L}_{\mathrm{tex}}$ and the geometry consistency loss $\mathcal{L}_{\mathrm{c}}$, where λ_{tex} and λ_{c} are the corresponding loss weights:

$$\mathcal{L} = \lambda_{\mathrm{tex}}\mathcal{L}_{\mathrm{tex}} + \lambda_{\mathrm{c}}\mathcal{L}_{\mathrm{c}}, \tag{7}$$

Based on Eq. (7), the geometry and texture are optimized jointly. We employ a progressive optimization strategy for the rendered color $\mathcal{I}$ in the Eq. (5). Initially, this image is generated at a low resolution (32×32), which is gradually increased during the optimization process, ultimately reaching 512×512 resolution. In contrast, both the rendered normal image $\mathcal{N}$ and color image $\mathcal{I}$ in the Eq. (6) remain 512×512 resolution throughout the entire procedure. Additionally, we detach the gradients of $z^{\mathcal{I}}$ in Eq. (6), allowing only geometric updates, while optimizing textures using the texture SDS loss. This improves texture-text consistency and geometry-texture alignment, preventing misalignment that results in unrealistic animation.

4.3. Training with Animations

To ensure plausible animations, particularly for the face region, it is essential to maintain semantic correspondence with the SMPL-X model. However, during optimization, certain parts may undergo changes and do not align perfectly with the original ones (*e.g.* the mouth may be mapped to the chin area or become distorted). If not addressed, animated results will have severe artifacts as the wrong parts will be deformed with the SMPL-X model. To tackle this problem, we optimize the avatar using various animations (see Fig. 3). In particular, we find that using different jaw poses during training helps produce well-aligned faces. We found that animating the SMPL-X expression parameters, made little visible difference. We suspect that these would become relevant with an even higher-resolution face mesh. Specifically, during optimization, we randomly sample one jaw pose in each iteration from an expression gallery Θ, i.e., a motion sequence from TalkSHOW [75]. The final optimization process minimizes the following objective:

$$\min_{\beta,\psi,\mathbf{D},\Psi} \mathbb{E}_{\theta\in\Theta}\left[\mathcal{L}(\phi, x(\beta, \theta, \psi, \mathbf{D}, \Psi))\right]. \tag{8}$$

5. Experiments

We first demonstrate our expressive, holistic, animation of the avatars, then evaluate their quality, and the consistency between texture and geometry. Finally, ablation studies are conducted to analyze the effectiveness of each component.

5.1. Expressive Holistic Body Animation

One crucial feature that distinguishes our method from others is that TADA enables natural full-body animations over the face, body and hands. Figure 4 illustrates the animation of characters from TADA with only text input. On the one hand, we can convert text to audio with TTS[*] and then use any Audio2Motion method, such as TalkSHOW

[*] https://play.ht

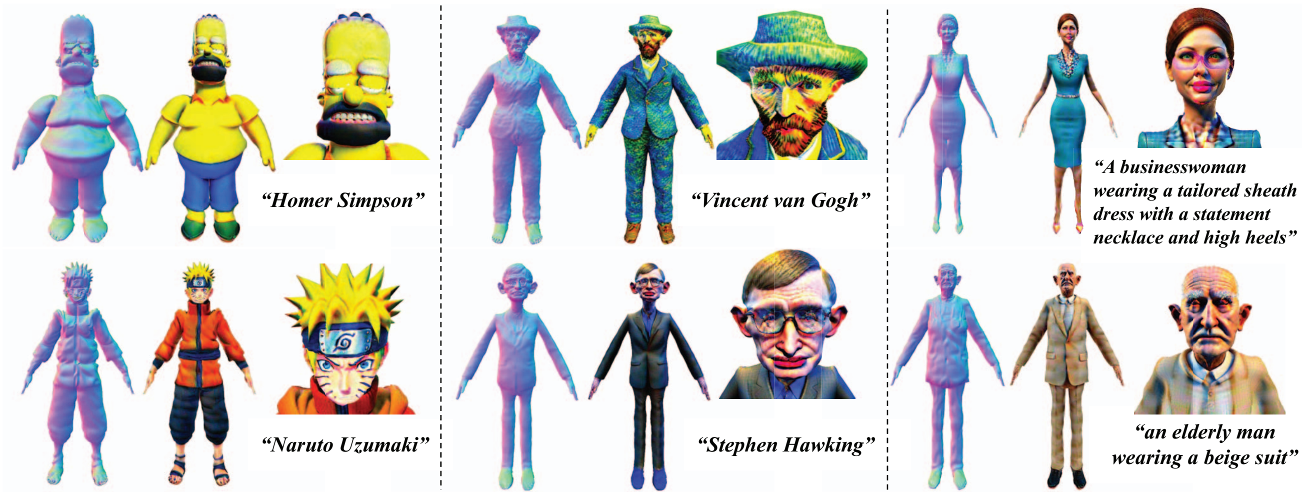

Figure 5. **Diverse range of avatar generation**. TADA can generate a broad spectrum of characters, which includes iconic figures, celebrities, and customized avatars based on textual descriptions.

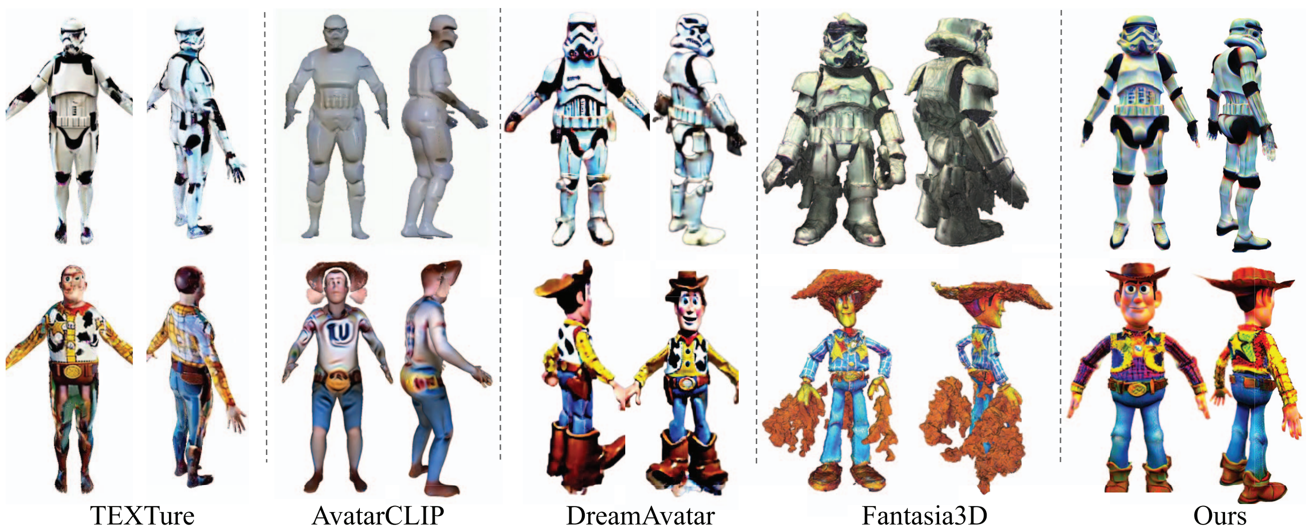

Figure 6. **Qualitative comparison**. The prompts (top $\rightarrow$ down) are *"Stormtrooper", "Woody in Toy Story"*. Compared with baselines using: A) body mesh w/o displacement clothing layer (TEXTure [46], AvatarCLIP [18]), B) Neural fields (DreamAvatar [5]), C) DMTET (Fantasia3D [7]). TADA generates better high-quality characters in terms of both geometry and texture.

[75], to create expressive motions of the upper body, face and hands. On the other hand, any Text2Motion method, such as PriorMDM [55], could also be used to convert text into motions. Here we all convert the motions to SMPL-X [40] formats for compatibility.

5.2. Diverse Range of Avatars

As shown in Fig. 5, TADA produces a wide variety of 3D avatars characterized by their high-quality geometry and realistic textures. These avatars contain fictional characters from animated films, real-life celebrities, and custom-made characters. This enables practical applications, enabling users to effortlessly generate avatars with a wide range of shapes, appearances, and clothing styles.

5.3. Qualitative Comparison

For full-body human generation task, we mainly compare with TEXTure [46], AvatarCLIP [18], DreamAvatar [5] and Fantasia3D [7]. For head-only avatar generation, we compare TADA with DreamFace [77] and HeadSculpt [16].

Full Body Avatar. As Fig. 6 shows, TADA generates avatars with higher quality than baselines. Also TADA produces a wide range of 3D body shapes (cf. TEXTure), without geometric artifacts (cf. AvatarCLIP, DreamAvatar and Fantasia3D), and with a semantically correct texture that is consistent with the geometry.

Head Avatar. As shown in Fig. 7, We compare TADA with DreamFace [77] and HeadSculpt [16], a shape sculpting

Figure 7. **Comparison of head generator**. While DreamFace excels in generating CG-compatible facial assets, it struggles with shapes that deviate significantly from the norm, such as accessories like hats. HeadSculpt often produces noisy artifacts in its output. In contrast, TADA generates a broader range of detailed shapes and appearances with better fidelity.

method specifically designed for head avatar generation. Note that TADA creates visually appealing head avatars with consistent and well-aligned geometry as well as high-fidelity textures. DreamFace [77] avatars can look realistic but are strongly biased towards natural head shapes and cannot capture more varied facial details like mustaches or cartoon shapes. Meanwhile, HeadScuplt [16] generates noisy geometry and texture, making the output less useful for downstream tasks like animation.

5.4. Quantitative Evaluation.

To quantitatively evaluate TADA, we conducted an A/B user study with 17 participants to assess the (1) geometry quality, (2) texture quality, and (3) consistency with input prompts. We used ChatGPT (see Sup.Mat.) to automatically generate 105 character descriptions, including celebrities, movie and anime characters, and characters from various occupations, select 27 of these at random, and generate the corresponding avatars; In A/B tests, the participants were asked to select the preferred reconstruction from randomly selected videos from the baselines and our method (see Tab. 1). The results show that our proposed method achieves considerably higher preference over the baseline methods over all three metrics.

5.5. Ablation Study

We conducted ablation studies to evaluate the effects of the geometry consistency loss and the optimization with animation scheme. The results shown in Fig. 8 demonstrate the effectiveness of these components. The consistency loss improves the alignment between the geometry and texture on the backside of the "Superman" while training with animations improves the face geometry by enforcing the semantic correspondence with SMPL-X, particularly at the mouth. These advancements enable us to effortlessly animate our high-resolution avatars, leveraging the pose and expression space of the SMPL-X model.

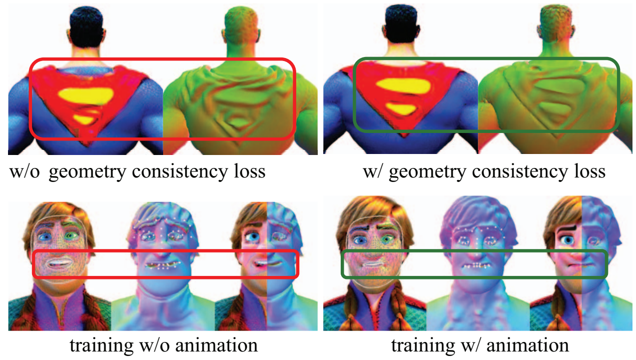

Figure 8. **Ablation study on 1) geometry consistency loss, and 2) training with animation**. The geometry consistency loss generates better well-aligned geometry and texture, while training with animation helps remain the semantic correspondences with the original SMPL-X, especially at the mouth region.

Preference (%, ↑)	AvatarCLIP	DreamAvatar
Geometry Quality	94.45	87.77
Texture Quality	94.74	82.67
Consistency with Input Prompt	95.00	81.52

Table 1. **User Study**. User preference results indicate that TADA severely outperforms other baselines across all three key aspects.

6. Editing Applications

TADA facilitates several applications, such as virtual try-on, text-guided texture editing, and local geometry transferring.

Outfit & Texture Editing. TADA enables the text-guided outfit editing while preserving identity, as depicted in Fig. 9. Figure 10 shows examples of modifying the textures of the outfit by changing the input text. This is particularly valuable for film or game character design. Designers can efficiently visualize their desired aesthetic appeal and bring their creative vision to fruition.

Local Shape Editing. Thanks to the body-part segments of SMPL-X, our method supports direct local body and face swapping between two avatars without additional effort. Fig. 11 gives an example of facial editing on four individuals. This is also applicable to body or clothing transferring. This feature is particularly helpful for artists in designing customized avatars.

7. Discussion

While TADA shows promising results, it still has several limitations. Additionally, further investigation is needed to assess any potential negative social impact.

Limitations & Future works. One aspect that requires improvement is the *relighting capabilities* in different environments, enabling photo-realistic rendering with human-scene interactions. This can benefit from using

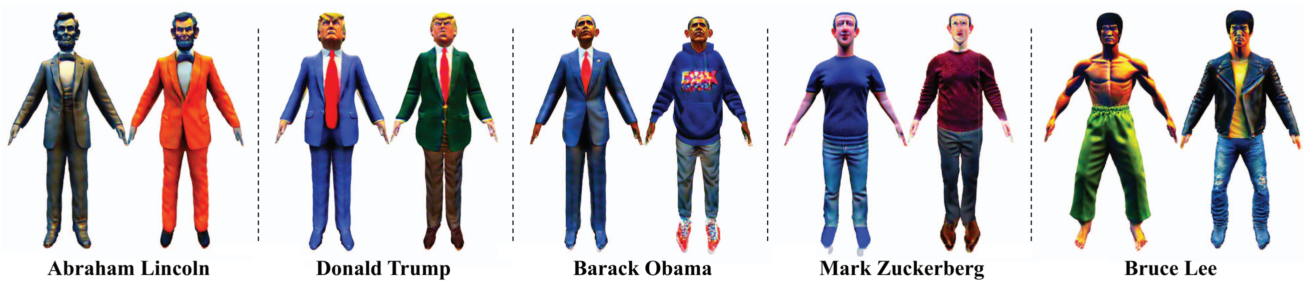

Figure 9. **Outfit Editing**. We demonstrate five individuals: Abraham Lincoln, Donald Trump, Barack Obama, Mark Zuckerberg, and Bruce Li, each with two different outfits. The first image represents their usual style of clothing, while the other is based on a new text description (created via ChatGPT), complete with detailed attire descriptions.

Figure 10. **Text-guided texture editing**. TADA possesses the ability to modify the color of clothing via changing texts.

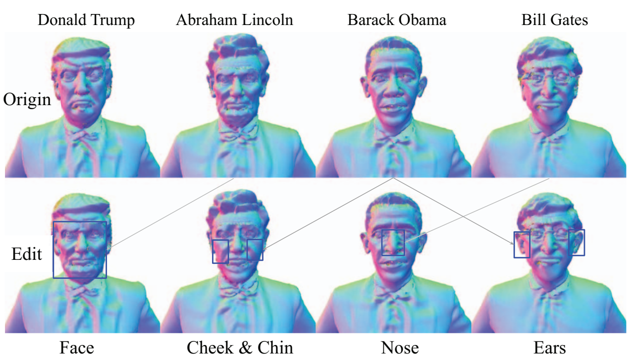

Figure 11. **Local Shape Editing**. We demonstrate an example of face swapping across four different celebrities.

BRDF, by separating the texture into separate components (*i.e.*, material, albedo, and lighting) like Fantasia3D [7]. Furthermore, TADA can generate avatars with diverse body shapes, some of which may deviate largely from the base SMPL-X model. In such cases, using the original skinning weights may lead to unrealistic animations. Therefore, exploring the joint learning of *adaptable skinning weights* specifically tailored to text input could be a promising direction. Textual descriptions alone may not fully capture the nuanced and intricate aspects of a character's appearance. Combining existing controllable text-to-image models [38, 43, 76] can be beneficial to provide more detailed control over a character's face or clothing. And the *compositional generation* of separate haircut [59], accessories [12], and decoupled outfits [10] could also be a valuable exploration direction.

Social Impact. As the technique progresses, it raises concerns about DeepFake and intellectual property (IP) when we generate iconic characters. Regulations should be established to address these issues alongside the benefits in the entertainment industry. Additionally, it is crucial to prioritize gender and cultural diversity. For instance, if the term "police officer" consistently generates a male instead of considering both genders, it implies potential gender bias. Ensuring inclusivity and avoiding stereotypes are essential in mitigating any adverse social impact.

8. Conclusion

We introduce TADA, a simple yet effective method for generating high-quality and animatable 3D textured avatars solely from text input. These avatars cover a wide range of individuals, including celebrities and customized characters. They seamlessly integrate into existing CG pipelines, catering to various industries like fashion and entertainment. The key contributions include: 1) utilizing a subdivided version of SMPL-X with learned displacement layer and UV texture, 2) employing hierarchical optimization with adaptive focal lengths, 3) enforcing geometry-texture alignment through geometric consistency loss, and 4) training with animation to keep semantic correspondence with SMPL-X. We validate these components through ablation studies and demonstrate the superiority of TADA over other SOTAs with both qualitative and quantitative results.

Acknowledgments. Thanks Z. Liu and W. Liu for their fruitful discussion, H. Wang and X. Tang for technical support, and Benjamin Pelkofer for IT support. H. Yi is supported in part by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039B. Y. Xiu is funded by Marie Skłodowska-Curie grant agreement No.860768 (CLIPE). J. Tang and Y. Huang are supported by NSFC (Grant Nos: 61632003, 61375022, 61403005, 62273302, 62036009, 61936006).

Disclosure. https://files.is.tue.mpg.de/black/CoI_3DV_2024.txt

References

[1] Thiemo Alldieck, Hongyi Xu, and Cristian Sminchisescu. imghum: Implicit generative models of 3d human shape and articulated pose. In *International Conference on Computer Vision (ICCV)*, pages 5461–5470, 2021. 3

[2] Shivangi Aneja, Justus Thies, Angela Dai, and Matthias Nießner. Clipface: Text-guided editing of textured 3d morphable models. In *SIGGRAPH '23 Conference Proceedings*, 2023. 3

[3] Bernd Bickel, Mario Botsch, Roland Angst, Wojciech Matusik, Miguel Otaduy, Hanspeter Pfister, and Markus Gross. Multi-scale capture of facial geometry and motion. *Transactions on Graphics (TOG)*, 26(3):33–es, 2007. 2

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020. 2

[5] 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:2304.00916*, 2023. 2, 3, 6

[6] Eric R. Chan, Koki Nagano, Matthew A. Chan, Alexander W. Bergman, Jeong Joon Park, Axel Levy, Miika Aittala, Shalini De Mello, Tero Karras, and Gordon Wetzstein. GeNVS: Generative novel view synthesis with 3D-aware diffusion models. In *International Conference on Computer Vision (ICCV)*, 2023. 3

[7] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3D: Disentangling Geometry and Appearance for High-quality Text-to-3D Content Creation. In *International Conference on Computer Vision (ICCV)*, 2023. 2, 3, 6, 8

[8] Yongwei Chen, Rui Chen, Jiabao Lei, Yabin Zhang, and Kui Jia. TANGO: Text-driven Photorealistic and Robust 3D Stylization via Lighting Decomposition. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2022. 3

[9] dalle2. `https://openai.com/dall-e-2`, 2022. 3

[10] Yao Feng, Jinlong Yang, Marc Pollefeys, Michael J. Black, and Timo Bolkart. Capturing and Animation of Body and Clothing from Monocular Video. In *SIGGRAPH Asia 2022 Conference Papers*, 2022. 8

[11] Jianglin Fu, Shikai Li, Yuming Jiang, Kwan-Yee Lin, Chen Qian, Chen-Change Loy, Wayne Wu, and Ziwei Liu. Stylegan-human: A data-centric odyssey of human generation. *European Conference on Computer Vision (ECCV)*, 2022. 2

[12] Daiheng Gao, Yuliang Xiu, Kailin Li, Lixin Yang, Feng Wang, Peng Zhang, Bang Zhang, Cewu Lu, and Ping Tan. DART: Articulated Hand Model with Diverse Accessories and Rich Textures. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 8

[13] William Gao, Noam Aigerman, Groueix Thibault, Vladimir Kim, and Rana Hanocka. Textdeformer: Geometry manipulation using text guidance. In *ACM Transactions on Graphics (SIGGRAPH)*, 2023. 3

[14] William Gao, Noam Aigerman, Groueix Thibault, Vladimir Kim, and Rana Hanocka. Textdeformer: Geometry manipulation using text guidance. In *ACM Transactions on Graphics (SIGGRAPH)*, 2023. 2

[15] Marc Habermann, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. Deepcap: Monocular human performance capture using weak supervision. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2020. 2

[16] Xiao Han, Yukang Cao, Kai Han, Xiatian Zhu, Jiankang Deng, Yi-Zhe Song, Tao Xiang, and Kwan-Yee K. Wong. Headsculpt: Crafting 3d head avatars with text. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023. 3, 6, 7

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Conference on Neural Information Processing Systems (NeurIPS)*, pages 6840–6851, 2020. 2

[18] Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, and Ziwei Liu. AvatarCLIP: Zero-Shot Text-Driven Generation and Animation of 3D Avatars. *Transactions on Graphics (TOG)*, 2022. 2, 3, 6

[19] Yangyi Huang, Hongwei Yi, Weiyang Liu, Haofan Wang, Boxi Wu, Wenxiao Wang, Binbin Lin, Debing Zhang, and Deng Cai. One-shot implicit animatable avatars with model-based priors. In *International Conference on Computer Vision (ICCV)*, 2023. 2

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

[21] Ajay Jain, Ben Mildenhall, Jonathan T. Barron, Pieter Abbeel, and Ben Poole. Zero-shot text-guided object generation with dream fields. In *Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[22] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *Computer Vision and Pattern Recognition (CVPR)*, pages 3334–3342, 2015. 2

[23] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[24] Byungjun Kim, Patrick Kwon, Kwangho Lee, Myunggi Lee, Sookwan Han, Daesik Kim, and Hanbyul Joo. Chupa: Carving 3D Clothed Humans from Skinned Shape Priors using 2D Diffusion Probabilistic Models. In *International Conference on Computer Vision (ICCV)*, 2023. 3

[25] Nikos Kolotouros, Thiemo Alldieck, Andrei Zanfir, Eduard Gabriel Bazavan, Mihai Fieraru, and Cristian Sminchisescu. Dreamhuman: Animatable 3d avatars from text. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023. 2, 3

[26] Samuli Laine, Janne Hellsten, Tero Karras, Yeongho Seol, Jaakko Lehtinen, and Timo Aila. Modular primitives for high-performance differentiable rendering. *Transactions on Graphics (TOG)*, 39(6), 2020. 4

[27] Ruilong Li, Kyle Olszewski, Yuliang Xiu, Shunsuke Saito, Zeng Huang, and Hao Li. Volumetric human teleportation. In *ACM SIGGRAPH 2020 Real-Time Live*, 2020. 2

[28] Ruilong Li, Yuliang Xiu, Shunsuke Saito, Zeng Huang, Kyle Olszewski, and Hao Li. Monocular real-time volumetric performance capture. In *European Conference on Computer Vision (ECCV)*, pages 49–67. Springer, 2020. 2

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

[30] 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 *Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 3

[31] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund T, Zexiang Xu, and Hao Su. One-2-3-45: Any Single Image to 3D Mesh in 45 Seconds without Per-Shape Optimization. *arXiv preprint*, 2023. 3

[32] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot One Image to 3D Object. In *International Conference on Computer Vision (ICCV)*, 2023. 3

[33] Wan-Chun Ma, Andrew Jones, Jen-Yuan Chiang, Tim Hawkins, Sune Frederiksen, Pieter Peers, Marko Vukovic, Ming Ouhyoung, and Paul Debevec. Facial performance synthesis using deformation-driven polynomial displacement maps. *Transactions on Graphics (TOG)*, 27 (5):1–10, 2008. 2

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

[35] Oscar Michel, Roi Bar-On, Richard Liu, Sagie Benaim, and Rana Hanocka. Text2Mesh: Text-Driven Neural Stylization for Meshes. In *Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 3

[36] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision (ECCV)*, 2020. 2, 3

[37] Nasir Mohammad Khalid, Tianhao Xie, Eugene Belilovsky, and Tiberiu Popa. Clip-mesh: Generating textured meshes from text using pretrained image-text models. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–8, 2022. 3

[38] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint:2302.08453*, 2023. 8

[39] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 165–174, 2019. 2

[40] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive Body Capture: 3D Hands, Face, and Body from a Single Image. In *Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 4, 6

[41] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. DreamFusion: Text-to-3d using 2d diffusion. In *International Conference on Learning Representations (ICLR)*, 2023. 2, 3, 4

[42] Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov, et al. Magic123: One Image to High-Quality 3D Object Generation Using Both 2D and 3D Diffusion Priors. *arXiv preprint:2306.17843*, 2023. 3

[43] Zeju Qiu, Weiyang Liu, Haiwen Feng, Yuxuan Xue, Yao Feng, Zhen Liu, Dan Zhang, Adrian Weller, and Bernhard Schölkopf. Controlling Text-to-Image Diffusion by Orthogonal Finetuning. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023. 8

[44] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020. 2

[45] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning (ICML)*, pages 8821–8831. PMLR, 2021. 2

[46] Elad Richardson, Gal Metzer, Yuval Alaluf, Raja Giryes, and Daniel Cohen-Or. Texture: Text-guided texturing of 3d shapes. In *International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, 2023. 2, 3, 6

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

[48] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[49] 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. In *Conference on Neural Information Processing Systems (NeurIPS)*, pages 36479–36494, 2022. 3

[50] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *International Conference on Computer Vision (ICCV)*, 2019. 2

[51] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *CVPR*, 2020. 2

[52] 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 *Computer Vision and Pattern Recognition (CVPR)*, pages 18603–18613, 2022. 3

[53] Igor Santesteban, Nils Thuerey, Miguel A Otaduy, and Dan Casas. Self-Supervised Collision Handling via Generative 3D Garment Models for Virtual Try-On. In *Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[54] Igor Santesteban, Miguel A. Otaduy, Nils Thuerey, and Dan Casas. ULNeF: Untangled layered neural fields for mix-and-match virtual try-on. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2022. 2

[55] Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit H Bermano. Human motion diffusion as a generative prior. *arXiv preprint arXiv:2303.01418*, 2023. 2, 5, 6

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

[57] Tianchang Shen, Jun Gao, Kangxue Yin, Ming-Yu Liu, and Sanja Fidler. Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. In *Conference on Neural Information Processing Systems (NeurIPS)*, pages 6087–6101, 2021. 2

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

[59] Vanessa Sklyarova, Jenya Chelishev, Andreea Dogaru, Igor Medvedev, Victor Lempitsky, and Egor Zakharov. Neural Haircut: Prior-Guided Strand-Based Hair Reconstruction. In *International Conference on Computer Vision (ICCV)*, 2023. 8

[60] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*. PMLR, 2015. 2

[61] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*, 2021. 2

[62] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2019. 2

[63] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *International Conference on Learning Representations (ICLR)*, 2023. 2

[64] tts. `https://play.ht`, 2023. 2

[65] Can Wang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. Clip-nerf: Text-and-image driven manipulation of neural radiance fields. In *Computer Vision and Pattern Recognition (CVPR)*, pages 3835–3844, 2022. 3

[66] 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. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2021. 3

[67] 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 *Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[68] 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. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023. 2

[69] Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel View Synthesis with Diffusion Models (3DiM). In *International Conference on Learning Representations (ICLR)*, 2023. 3

[70] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *International Conference on Machine Learning (ICML)*, 2011. 2

[71] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J. Black. ICON: Implicit Clothed humans Obtained from Normals. In *Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[72] Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J. Black. ECON: Explicit Clothed humans Optimized via Normal integration. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[73] Jiale Xu, Xintao Wang, Weihao Cheng, Yan-Pei Cao, Ying Shan, Xiaohu Qie, and Shenghua Gao. Dream3d: Zero-shot text-to-3d synthesis using 3d shape prior and text-to-image diffusion models. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[74] Xueting Yang, Yihao Luo, Yuliang Xiu, Wei Wang, Hao Xu, and Zhaoxin Fan. D-IF: Uncertainty-aware Human Digitization via Implicit Distribution Field. In *International Conference on Computer Vision (ICCV)*, 2023. 2

[75] Hongwei Yi, Hualin Liang, Yifei Liu, Qiong Cao, Yandong Wen, Timo Bolkart, Dacheng Tao, and Michael J Black. Generating holistic 3d human motion from speech. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 5, 6

[76] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *International Conference on Computer Vision (ICCV)*, 2023. 8

[77] Longwen Zhang, Qiwei Qiu, Hongyang Lin, Qixuan Zhang, Cheng Shi, Wei Yang, Ye Shi, Sibei Yang, Lan Xu, and Jingyi Yu. Dreamface: Progressive generation of animatable 3d faces under text guidance. In *International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, 2023. 3, 6, 7

[78] Rui Zhao, Wei Li, Zhipeng Hu, Lincheng Li, Zhengxia Zou, Zhenwei Shi, and Changjie Fan. Zero-shot text-to-parameter

translation for game character auto-creation. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[79] Yang Zheng, Ruizhi Shao, Yuxiang Zhang, Tao Yu, Zerong Zheng, Qionghai Dai, and Yebin Liu. Deepmulticap: Performance capture of multiple characters using sparse multiview cameras. In *International Conference on Computer Vision (ICCV)*, 2021. 2

[80] Zerong Zheng, Xiaochen Zhao, Hongwen Zhang, Boning Liu, and Yebin Liu. Avatarrex: Real-time expressive full-body avatars. *ACM Transactions on Graphics (TOG)*, 42(4), 2023. 2

[81] Zhizhuo Zhou and Shubham Tulsiani. Sparsefusion: Distilling view-conditioned diffusion for 3d reconstruction. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[82] Heming Zhu, Lingteng Qiu, Yuda Qiu, and Xiaoguang Han. Registering explicit to implicit: Towards high-fidelity garment mesh reconstruction from single images. In *Computer Vision and Pattern Recognition (CVPR)*, pages 3845–3854, 2022. 2

[83] Luyang Zhu, Konstantinos Rematas, Brian Curless, Steve Seitz, and Ira Kemelmacher-Shlizerman. Reconstructing nba players. In *European Conference on Computer Vision (ECCV)*, 2020. 2