

TADA! Text to Animatable Digital Avatars

Tingting Liao^{1*}, Hongwei Yi^{2*}, Yuliang Xiu², Jiaxiang Tang³, Yangyi Huang⁴
Justus Thies², Michael J. Black²

¹Mohamed bin Zayed University of Artificial Intelligence ²Max Planck Institute for Intelligent Systems

³Peking University ⁴State 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

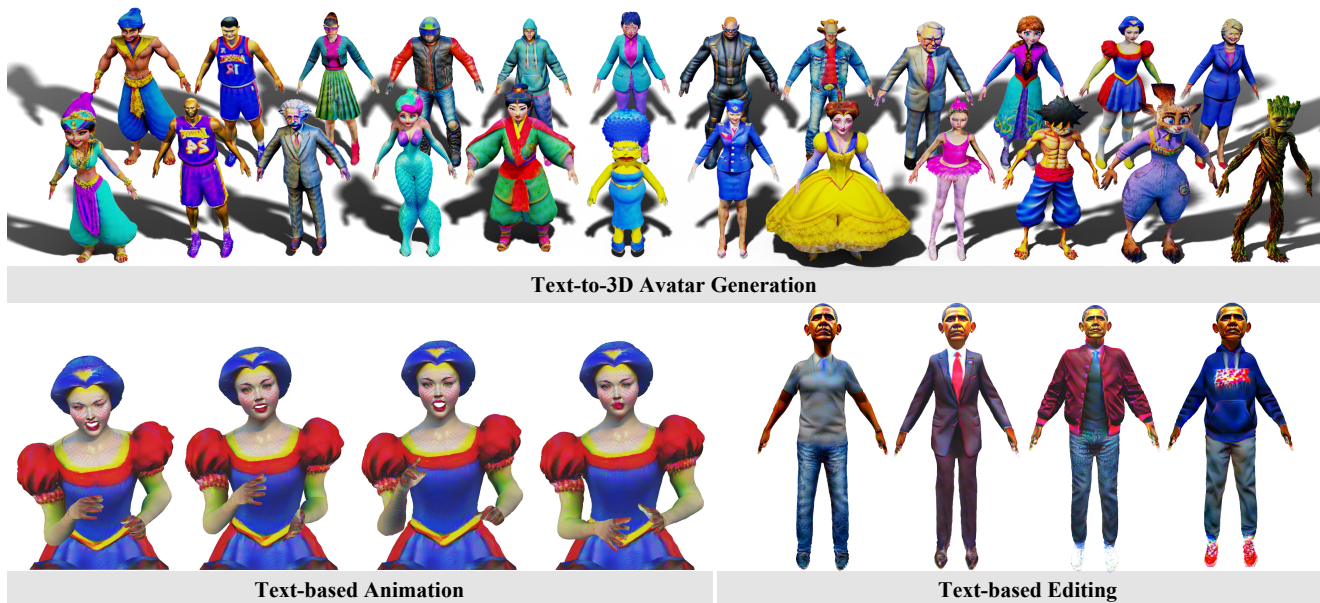


Figure 1. With only text descriptions as input, TADA generates high-fidelity 3D avatars with lifelike texture and detailed geometry, including high-resolution faces. Accurate alignment of texture and geometry, together with an underlying SMPL-X representation, enables expressive animation. TADA also 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 at 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 SMPL-X with 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 SDS optimization process. We further drive the face of character with multiple expressions during optimization, ensuring that its semantics remain consistent with the original SMPL-X model, for realistic animation with semantic alignment. 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 allows for text-guided editing. The code will be public for research purposes at tada.is.tue.mpg.de.

*denotes equal contribution.

1. Introduction

Digital avatars are a foundation for applications in augmented and virtual reality, immersive telepresence [27, 28, 49, 68, 78], virtual try-on [51, 52, 79], and video games [15, 77, 80]. Creating high-quality and expressive 3D avatars is challenging since the geometry and appearance of the character has to be modelled under a variety of different poses. Traditional pipelines used in the entertainment industry often use sophisticated multi-view capture studios [3, 22, 32] to create complex 3D models. Manual processes like cleaning and rigging the scans make creating an animatable character time-consuming and expensive. While there is recent progress on automatic learning-based body reconstruction from single image [20, 28, 29, 48, 49, 67, 68, 70], or sparse images [54], such methods are limited to real humans, fail on fictional characters, and are hard to edit and control. Thanks to the rapid progress on Large Language Models [4, 42] and Diffusion Models [17, 57–59, 66], recent work has shown that text-to-image models [43, 45] can be combined with differentiable neural 3D scene representations such as DeepSDF [38], NeRF [35] and DMTET [55] to generate realistic 3D models solely from textual descriptions. However, these methods have many limitations. The generated objects or characters are often rigid and lack of animation [7, 14, 30, 34, 44], they have difficulty in producing high-quality realistic 3D avatars in terms of geometry and texture [18], or the characters are incompatible with traditional CG workflows (NeRF based [5, 25, 40]).

Here we address these limitations with TADA, illustrated in Fig. 13. Since our goal is animatable avatars that are compatible with existing rendering engines, we build upon the SMPL-X body model [39]. SMPL-X, however, only represents a realistic, minimally clothed human body shape. Our goal is to create diverse avatars with a wider variety of body shapes and textures. Like recent work that generates avatars from text, we use Score Distillation Sampling (SDS) [40] but do so in several novel ways. Specifically, we make three key contributions. (i) First, we devise a hierarchical optimization of a hybrid mesh representation which is based on a subdivided version of SMPL-X [39] with additional learnable displacements and a texture map. To produce high-quality details, especially on the face region, we perform hierarchical optimization over hierarchically rendered images with different focal lengths, where the entire body, or only specific parts, are visible. (ii) Second, existing text-to-3D object methods [7, 40] suffer from inconsistent alignment between the reconstructed geometry and texture (see Fig. 2), as they evaluate the texture and geometry individually. This makes animation of the resulting avatars infeasible. (iii) Third, we want the generated character

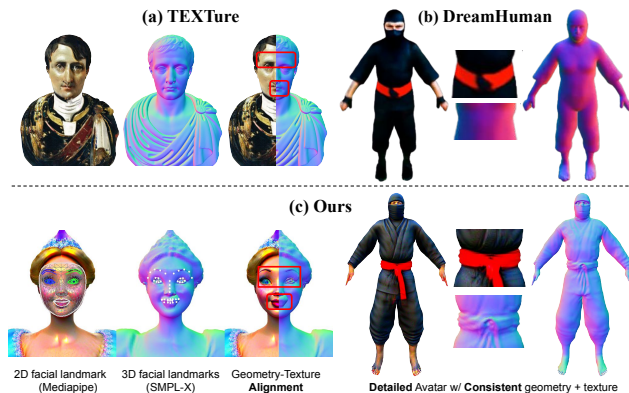


Figure 2. Compared with other existing methods [25, 40], our method can generate high-quality 3D avatars with well-aligned geometry and texture that is consistent with SMPL-X, enabling animation and rendering with existing graphics pipelines.

to be semantically consistent with SMPL-X so that it can be easily animated; that is, body parts and vertices on output avatars correspond to the same body parts and vertices on SMPL-X. To that end, 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 accurately and coherently, as depicted in Fig. 2 (c). Notably, 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 [53, 60] or text-to-audio-to-motion methods [61, 72], we can animate the generated characters to interact with the scenes or communicate with others. This paves the way towards creating virtual 3D worlds with animatable digital avatars fully from text.

In summary, with TADA we propose a user-friendly tool for avatar creation and editing, that can be solely controlled by the textual input and is 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 realistic iconic celebrities, customized humans, and cartoon characters. We validate our contributions with ablation studies, show qualitative comparisons to the state of the art, and conduct a user study that quantifies the performance of our method on the task of high-quality 3D avatar generation.

2. Related Work

Recently, there has been rapid progress on extending text-to-2D-image generation methods [11, 23, 46] to text-to-3D-content generation [33, 40, 65]. Here, we

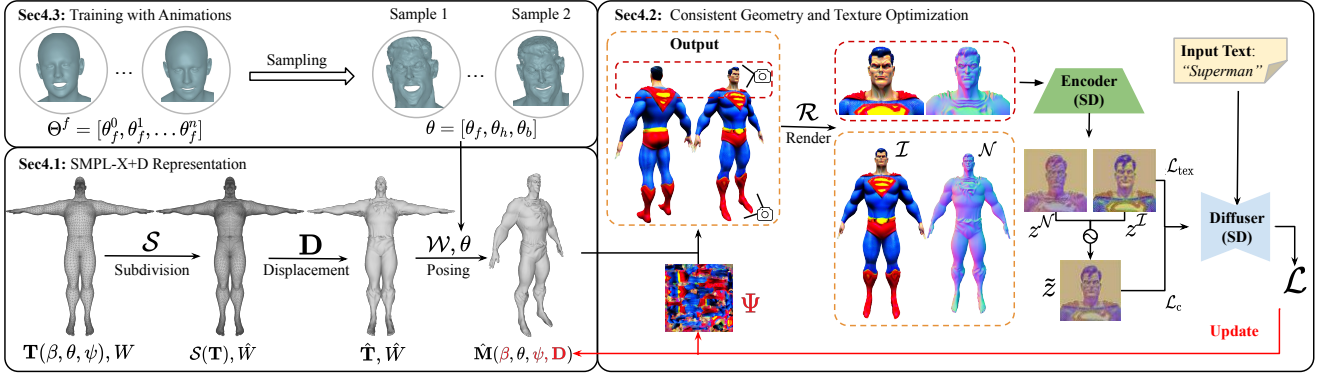


Figure 3. **Overview.** Initialized by a SMPL-X body $\mathbf{T}(\beta, \theta, \psi)$ with skinning weights W , we subdivide the body to obtain a denser mesh $\mathbf{S}(\mathbf{T})$ and add personalized displacements \mathbf{D} to it. The personalized mesh $\hat{\mathbf{T}}$ is transformed into the 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.

discuss the most relevant text-to-3D-content generation methods, while focusing primarily on text-to-3D-avatar generation (both human and anime characters).

Text-to-3D-Content Generation. The successes of Text-to-Image (T2I) generative modeling [9, 45, 47] have sparked a surge of interest in the field of text-to-3D generation [8, 50, 62, 69]. Despite progress, effectively describing and controlling 3D properties of an object using language, while ensuring coherence in the three-dimensional space, remains a challenge. One line of work [21, 34, 36] utilizes 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 the 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 circumvent the training of a 3D generative model and the problem of missing datasets, recent publications [7, 30, 40, 44] make significant strides by leveraging the power of score distillation sampling (SDS) [45] derived from 2D text-to-image diffusion models to create content from textual descriptions by optimizing a 3D representation. TEXTure [44] takes a mesh as input and only optimizes the texture map based on a given text prompt. In contrast, DreamFusion [40] optimizes a Neural Radiance Field (NeRF) [35] to represent the 3D content in terms of a density and radiance field. It can generate 3D models of arbitrary (fictional) objects. 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. All these methods focus on general, static, object/scene generation; they output is not animation-ready, which is necessary for 3D character creation.

Text-to-3D-Avatar Generation. Several methods generate 3D head avatars from text [13, 16, 64, 75, 76]. In contrast, we focus on generating full-body characters including the detailed face. AvatarCLIP [18] leverages NeuS [63] and the SMPL-X model with a CLIP-guide loss to facilitate the generation of avatars. Similarly, DreamAvatar [5] utilizes the shape parameters from 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 not compatible with traditional graphics workflows, especially for animation. In the domain of explicit representations, Text2Mesh [34] and Chupa [24] employ vertex displacement on a predefined mesh template. Nonetheless, the inherent limitation of fixed topology poses challenges in accurately generating diverse character shapes. 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 seamlessly integrate into existing CG rendering and animation workflows.

3. Preliminaries

SMPL-X [39] is an animatable parametric 3D body model that consists of the human body, hands and face. 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 $\mathbf{M}(\beta, \theta, \psi)$:

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

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 [40] has been proposed in DreamFusion 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:

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

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)):

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

where $\mathcal{S} : \mathbb{R}^{N \times 3} \rightarrow \mathbb{R}^{N_s \times 3}$ 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 vertices number of the subdivided body, respectively. Note that besides the displacement \mathbf{D} , the parameters β, θ, ψ are also learnable. This helps to generate various kinds of characters with various shapes, such as human-like characters and anime characters with large deformed body shapes, like exaggerated proportions, elongated limbs, large eyes, and *etc.*

Partial Mesh Subdivision. The vertices on the surface of the SMPL-X body are irregularly distributed, i.e., around 4,000 vertices are located on the head with the remaining 6,000 on 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 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} :

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

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 [45], denoted as ϕ , which measures the similarity between the rendered image and the provided text prompt within the added and predicted noise space:

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

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.

Text→Audio→Expressive Motion (TTS+TalkSHOW)

“...that everyone can do, to combat the kidney shortage...”



Text→Full-body Motion (priorMDM)

“A person makes a long leap forward”

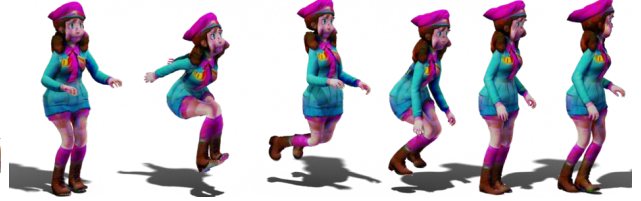


Figure 4. TADA enables holistic animation over the face, body and hands. We show animation examples of the avatars “Lionel Messi” and “Mabel Pines in Gravity Falls” using expressions and body poses from TalkSHOW [72] (with TTS [61]) and priorMDM [53], respectively.

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.

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

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}_{tex} and the geometry consistency loss \mathcal{L}_c , where λ_{tex} and λ_c are the corresponding loss weights:

$$\mathcal{L} = \lambda_{\text{tex}} \mathcal{L}_{\text{tex}} + \lambda_c \mathcal{L}_c, \quad (7)$$

Based on Eq. (7), the geometry and texture are optimized jointly. We employ a progressive optimization strategy for the rendered color image \mathcal{I} in the Eq. (5). Initially, this image is generated at a low resolution (32×32), which gradually increases 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 approach ensures both texture-text consistency and geometry-texture alignment, preventing misalignment that could result 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 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 [72]. The final optimization process minimizes the following objective:

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

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 generated by TADA using only with text as input. In the first case, we convert text to audio [61] and then use TalkSHOW [72] create expressive SMPL-X animations of the upper body, face and hands. In the second case we use priorMDM [53] to convert text into SMPL [31] animations, which we convert to SMPL-X [39]. Thanks semantic correspondence with SMPL-X, the characters are easily animated with natural movements of the fully body and face. This consistency with SMPL-X means that avatars generated by TADA can be animated using any of the recent text to animation methods that output SMPL-X.

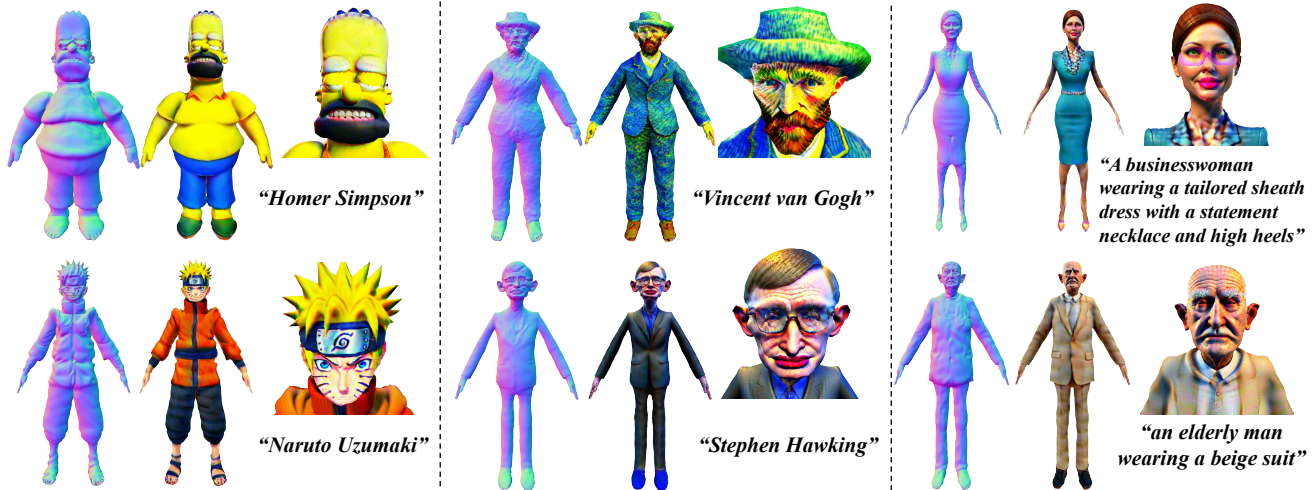


Figure 5. **Diverse Range of Avatar Generation.** TADA has the ability to generate a broad spectrum of characters, which includes iconic figures, celebrities, and customized avatars based on textual descriptions.



Figure 6. **Qualitative comparison.** The prompts (top \rightarrow down) are “Stormtrooper”, “Woody in Toy Story”, “Kristoff in Frozen”. Compared with baselines using: A) body mesh w/o displacement clothing layer (TEXTure [44], AvatarCLIP [18]), B) Neural fields (DreamAvatar [5]), C) DMTET (Fantasia3D [7]). TADA generates more high-quality characters in terms of both geometry and texture.

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 based on prompts generated by ChatGPT. This capability opens up numerous real-world applications, enabling users to effortlessly generate avatars



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 greater fidelity.

| Preference (% , \uparrow) | 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 indicates that TADA significantly outperforms other baselines in terms of geometry, texture, and consistency with the input prompt. Its superior performance is evident across all three key aspects.

with a wide range of shapes, appearances, and clothing styles.

5.3. Qualitative Comparison

We compare our method with existing approaches on the task of text-to-3D human avatar generation. We consider four state-of-the-art methods for full body avatar generation, where the original implementation is available: TEXTure [44], AvatarCLIP [18], DreamAvatar [5] and Fantasia3D [7]. For head-only avatar generation, we compare our approach with DreamFace [75] and HeadSculpt [16].

Full Body Avatar. Figure 6 provides a qualitative comparison of avatars generated by [5, 18, 44] and by TADA. In comparison to the baselines, TADA generates avatars with considerably more realistic textures. 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. Furthermore, TADA generates high-quality head avatars as shown in Fig. 7. We compare with DreamFace [75] and HeadSculpt [16], a shape sculpting 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. However, others have different limitations. DreamFace [75] avatars can look realistic but are strongly biased towards natural head shapes and cannot capture more varied facial details like mustaches or cartoon shapes. Like our results, the head avatars can be animated.

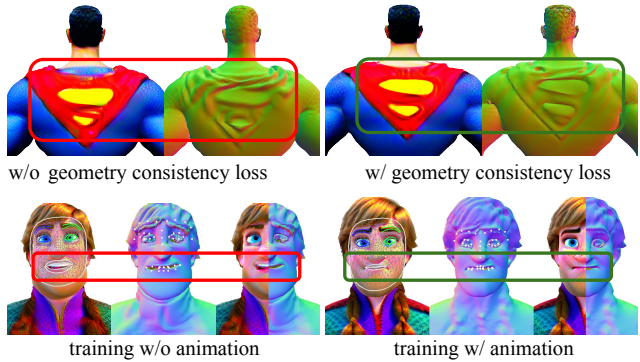


Figure 8. **Ablation study on 1) geometry consistency loss, and 2) training with animation.** The geometry consistency loss generates better well-aligned geometry and texture, while training with animation helps remain the semantic correspondences with the original SMPL-X, especially at the mouth region.

Meanwhile, HeadSculpt [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 a A/B user study with 17 CV students. We used a Google Survey Form to assess the (1) geometry quality, (2) texture quality, and (3) consistency with input prompts. We used ChatGPT to automatically generate a large set of character descriptions, including celebrities, characters in movies and anime, and general occupation character descriptions, select 27 of these at random, and generate the corresponding avatars; see Sup. Mat. for details. In A/B tests, the participants were asked to select the preferred reconstruction from randomly selected videos from the baselines (AvatarCLIP and DreamAvatar) and our method (see Tab. 1). The results show that our proposed method achieves considerable 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 animations in our method. 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 mouth region. These advancements enable us to effortlessly animate our high-resolution avatars, leveraging the pose and expression space of the SMPL-X model.



Figure 9. **Virtual Try-on.** We demonstrate five individuals: Abraham Lincoln, Donald Trump, Barack Obama, Mark Zuckerberg, and Bruce Li, each with two different outfits. The first one represents their typical dressing style, personalized according to their known preferences with their name as the only input. The other one is imagined by ChatGPT, complete with detailed descriptions of the attire.

6. Editing Applications

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



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

Virtual Try-on. TADA can be used for virtual try-on, i.e., we can ask ChatGPT [6] to design fashion outfits for a specific person as depicted in Fig. 9. The visual results indicate that our method can generate avatars with text-guided personalized textures while preserving the identity.

Texture Editing. Figure 10 shows examples of modifying outfit textures by changing the input text. This is particularly valuable for film or game character design, allowing easy alterations such as changing the color palette of a character. Designers can quickly visualize their desired aesthetic appeal and bring their creative vision to life.

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 any additional effort. Fig. 11 gives an example of face editing on four individuals. This is also applicable to body or clothing transferring. In addition to geometry or texture transferring, TADA can also be utilized for local shape sculpting through user-friendly prompts as input. This feature is particularly helpful for artists in designing customized avatars.

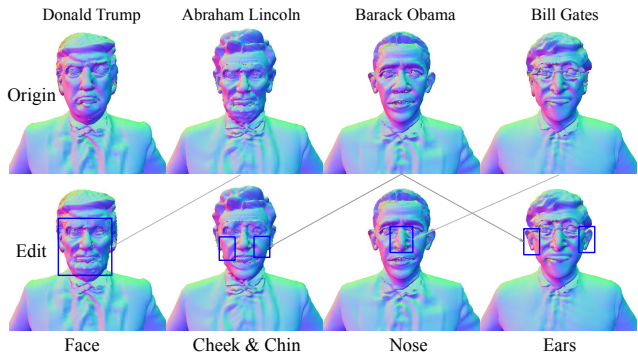


Figure 11. **Local Shape Editing.** We demonstrate an example of face swapping across four different celebrities.

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, e.g. generated indoor rooms from MIME [71], thus enabling photo-realistic rendering with human-scene interactions. This can benefit from using 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 [37, 41, 74] can be beneficial to provide more detailed control over a character’s face or clothing. And the *compositional generation* of separate haircut [56],

accessories [12], and decoupled outfits [10] could also be a valuable exploration direction.

Social Impact. As the technique progresses, it raises concerns about deep-fake 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 Zhen Liu and Weiyang Liu for their fruitful discussion, Haofan Wang and Xu Tang for their technical support, and Benjamin Pelkofer for IT support. Hongwei Yi is supported in part by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039B. Yuliang Xiu is funded by the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No.860768 (CLIFE). Jiaxiang Tang is supported by National Natural Science Foundation of China (Grant Nos: 61632003, 61375022, 61403005). Yangyi Huang is supported by the National Nature Science Foundation of China (Grant Nos: 62273302, 62036009, 61936006).

Disclosure. MJB has received research gift funds from Adobe, Intel, Nvidia, Meta/Facebook, and Amazon. MJB has financial interests in Amazon, Datagen Technologies, and Meshcapade GmbH. While MJB is a consultant for Meshcapade, his research in this project was performed solely at, and funded solely by, the Max Planck Society.

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)*, pages 1877–1901, 2020. 2
- [5] Yunkang 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, 7, 13, 14, 16
- [6] chatgpt. <https://chat.openai.com/>, 2022. 8
- [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, 7, 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. 9
- [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. 9
- [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. *arXiv preprint arXiv:2306.03038*, 2023. 3, 7, 13
- [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)*, 41(4):1–19, 2022. 2, 3, 6, 7, 13, 16
- [19] Yukun Huang, Jianan Wang, Ailing Zeng, He Cao, Xianbiao Qi, Yukai Shi, Zheng-Jun Zha, and Lei Zhang. Dreamwaltz: Make a scene with complex 3d animatable avatars. *arXiv preprint arXiv:2305.12529*, 2023. 13, 14
- [20] 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
- [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, Thimeo Alldieck, Andrei Zanfir, Eduard Gabriel Bazavan, Mihai Fieraru, and Cristian Sminchisescu. Dreamhuman: Animatable 3d avatars from text. *arXiv preprint:2306.09329*, 2023. 2, 3, 13, 17
- [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] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, 2015. 5
- [32] 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
- [33] Gal Metzger, 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
- [34] 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
- [35] 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
- [36] 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
- [37] 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
- [38] 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
- [39] 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, 3, 5
- [40] 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, 13, 14

- [41] 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. *arXiv preprint:2306.07280*, 2023. 8
- [42] 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
- [43] 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
- [44] Elad Richardson, Gal Metzer, Yuval Alaluf, Raja Giryes, and Daniel Cohen-Or. Texture: Text-guided texturing of 3d shapes. *arXiv preprint:2302.01721*, 2023. 2, 3, 6, 7
- [45] 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)*, pages 10684–10695, 2022. 2, 3, 4
- [46] 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
- [47] 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
- [48] 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
- [49] 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
- [50] 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
- [51] 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
- [52] 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
- [53] 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
- [54] 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
- [55] 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
- [56] Vanessa Sklyarova, Jenya Chelishhev, 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
- [57] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 2
- [58] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*, 2021.
- [59] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2019. 2
- [60] 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
- [61] tts. <https://play.ht>, 2023. 2, 5
- [62] 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
- [63] 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
- [64] 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)*, pages 4563–4573, 2023. 3
- [65] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *arXiv preprint arXiv:2305.16213*, 2023. 2
- [66] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *International Conference on Machine Learning (ICML)*, pages 681–688, 2011. 2

- [67] 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
- [68] 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
- [69] 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
- [70] 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
- [71] Hongwei Yi, Chun-Hao P. Huang, Shashank Tripathi, Lea Hering, Justus Thies, and Michael J. Black. MIME: Human-aware 3D scene generation. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 8
- [72] 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
- [73] Huichao Zhang, Bowen Chen, Hao Yang, Liao Qu, Xu Wang, Li Chen, Chao Long, Feida Zhu, Kang Du, and Min Zheng. Avatarverse: High-quality & stable 3d avatar creation from text and pose. *arXiv preprint:2308.03610*, 2023. 13, 14, 15
- [74] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 8
- [75] Longwen Zhang, Qiwei Qiu, Hongyang Lin, Qixuan Zhang, Cheng Shi, Wei Yang, Ye Shi, Sibe Yang, Lan Xu, and Jingyi Yu. Dreamface: Progressive generation of animatable 3d faces under text guidance. *arXiv preprint arXiv:2304.03117*, 2023. 3, 7
- [76] 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)*, pages 21013–21023, 2023. 3
- [77] 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
- [78] 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
- [79] 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
- [80] Luyang Zhu, Konstantinos Rematas, Brian Curless, Steve Seitz, and Ira Kemelmacher-Shlizerman. Reconstructing nba players. In *European Conference on Computer Vision (ECCV)*, 2020. 2

Appendices

A. Additional Qualitative Comparisons

We provide additional qualitative comparisons with existing baselines in three categories: head avatar generation, full-body iconic and customized avatar generation.

Head Avatar Generation. In Fig. 12, we provide examples generated by HeadSculpt [16] and our method TADA. Unlike HeadSculpt often generates noisy geometry and suffers from inconsistency between texture and geometry, TADA could produce high-quality character geometries with well-aligned textures.

Full-body Iconic Avatar Generation. We conduct a comprehensive comparison of the full-body iconic avatar generation with existing methods [5, 18, 19, 40, 73] in Figs. 13 to 15. NeRF-based methods [5, 19, 40, 73] tends to generate low-quality geometric fields, which are not compatible with existing traditional CG workflows, such as rasterization and animation. The mesh-based method AvatarClip [18] tends to generate minimal clothed geometry, and low-quality texture with severe artifacts (see Fig. 15). In contrast, TADA has the capability to produce a wide range of characters with superior geometries and well-aligned textures. These outputs seamlessly integrate with conventional graphics workflows, making them readily suitable for animation and rendering.

Full-body Customized Avatars Generation. To explore the generalization of our method, we compare TADA with DreamHuman [25] on the task of full-body customized avatars generation. In Fig. 16, we can see that TADA generates high-quality characters with more consistent geometry and texture than DreamHuman.

B. Implementation details

We select camera positions (r, θ, ϕ) in a spherical coordinate system, where $r, \theta,$ and ϕ denote the radius, elevation, and azimuth angle, respectively. During the optimization, the virtual cameras are positioned as: 1) **full-body mode**: camera with full-body field of view (FOV), and 2) **head-mode**: zoom-in camera focusing the head. The head-mode camera is sampled with a probability of 30%, and full-body model with 70%. We sample θ_b values within the range of $[60^\circ, 90^\circ]$ under full-body mode. Conversely, for the head view θ_h , we opt for values from the range of $[75^\circ, 85^\circ]$. Additionally, we sample ϕ within the intervals of $[-180^\circ, 180^\circ]$ for the full body, and $[-30^\circ, 30^\circ]$ for the head. In each iteration, the camera radius is adjusted according to the body height and the head size.

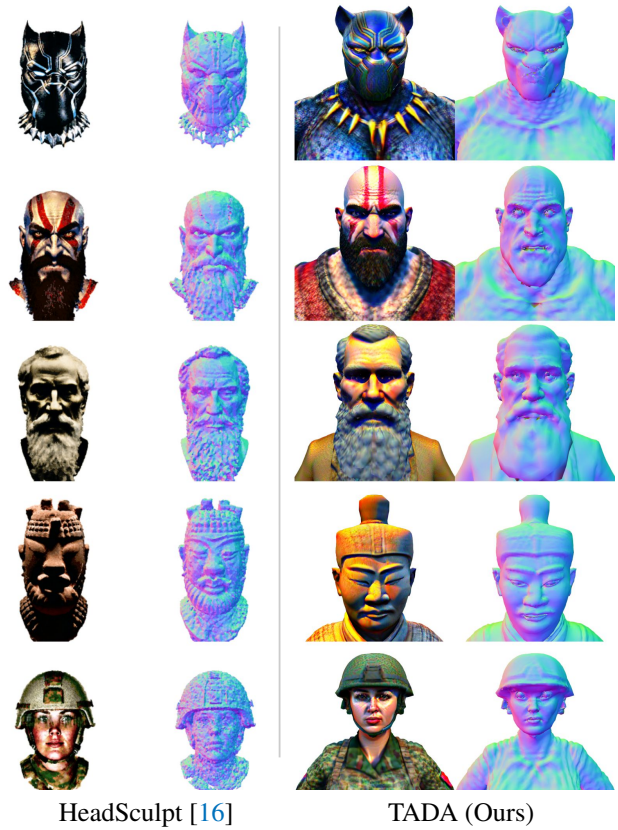


Figure 12. Qualitative comparison of our method TADA with HeadSculpt [16] on the task of head avatar generation.



Figure 13. **Qualitative comparison on full-body avatar generation of icons.** The prompts (top → down) are “*Spiderman*”, “*Joker*”, “*Stormtrooper*”. Compared with the baselines [5, 19, 40, 73], TADA generates a higher quality in terms of both geometry and texture.

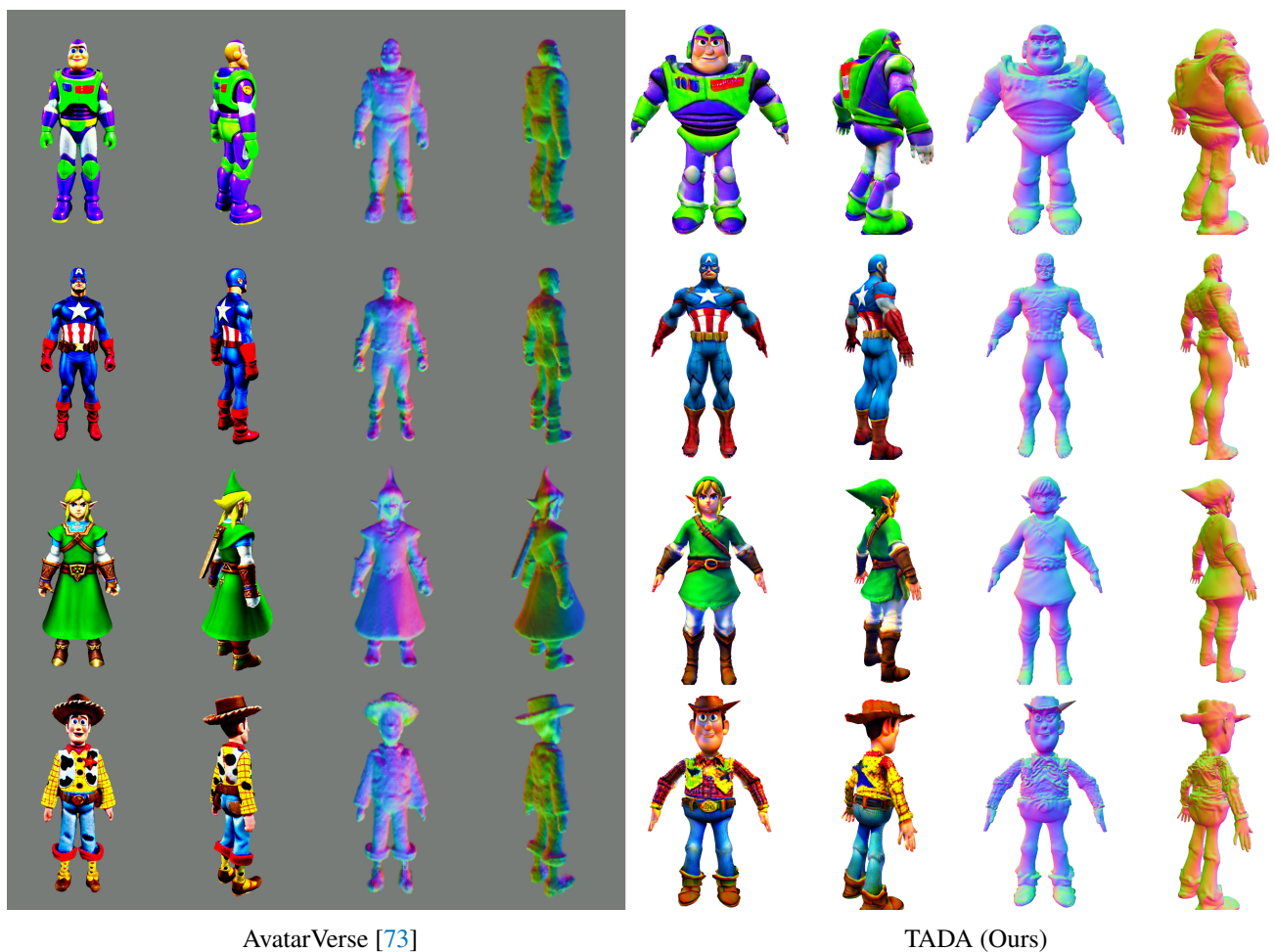


Figure 14. **Qualitative comparison of the full-body iconic avatar generation with AvatarVerse [73].** Left is from AvatarVerse. Right is ours. The generated avatars from AvatarVerse lack high-quality details in geometry, especially on the face region. In contrast, TADA generates high-quality meshes with well-aligned textures.



Figure 15. Qualitative comparison against DreamAvatar [5] and AvatarClip [18] on the full-body iconic avatar generation.



DreamHuman [25]

TADA (Ours)

Figure 16. Qualitative comparison with DreamHuman [25] on the full-body customized avatar generation.

C. Prompts used by ChatGPT

We generate characters of three groups: celebrity, characters in movies & anime (fictional characters), and general job descriptions. We also provide the prompts used to imagine outfits for Virtual Try-on application. In the following, we list the prompts used for generation.

Celebrities. We ask ChatGPT to output the names of superstars, scientists, businessmen and presidents. Here are the corresponding generated prompts:

Donald Trump
Abraham Lincoln
Barack Obama
Hillary Clinton
Yao Ming
Kobe Bryant
Messi
Bruce Lee
Steven Paul Jobs
Mark Elliot Zuckerberg
Joe Biden
Bill Gates
Warren Buffett
Elon Musk
Jeff Bezos
Jack Ma
Albert Einstein
Marie Curie
Stephen Hawking
Vincent van Gogh
Michelangelo
Wolfgang Amadeus Mozart
Ludwig van Beethoven
Michael Jackson
Kim Kardashian

Fictional Characters. We collect superheroes, Disney princesses, and characters in films such as Forzen, Aladdin, South Park, Simpson, Rick and Morty, *etc.*

superman
Deadpool
Batman
Ant-man
spiderman
Iron Man
Captain America
Woody in Toy Story
Buzz Lightyear in Toy Story
Elsa in Frozen
Anna in Frozen
Kristoff in Frozen
Aladdin in Aladdin
Jasmine in Aladdin
Mulan in Mulan
Olaf in Frozen
Jessie in Toy Story
Sun Wukong
groot
Moana in Moana
Judy Hopps in Zootopia
Goku in Dragon Ball series
Naruto Uzumaki in Naruto series
Luffy in One Piece
Kiki in Kiki's Delivery Service
San in Princess Mononoke
Eric Cartman in South Park: Bigger, Longer & Uncut
Dipper Pines in Gravity Falls
Mabel Pines in Gravity Falls
Rick Sanchez in Rick and Morty
Morty Smith in Rick and Morty
Stan Marsh in South Park: Bigger, Longer & Uncut
Grunkle Stan in Gravity Falls
Soos Ramirez in Gravity Falls
Kyle Broflowski in South Park: Bigger, Longer & Uncut
Kenny McCormick in South Park: Bigger, Longer & Uncut
Bojack Horseman in Bojack Horseman
Diane Nguyen in Bojack Horseman
Ariel in The Little Mermaid
Belle in Beauty and the Beast
Homer Simpson in The Simpsons
Marge Simpson in The Simpsons
Bart Simpson in The Simpsons
Lisa Simpson in The Simpsons
astronaut
Michelangelo style statue of an astronaut
A highly detailed stone bust of Theodoros Kolokotronis
Captain Jack Sparrow from Pirates of the Caribbean
Mobile suit Gundam
Harry Potter, adult
Optimus Prime, transformer
Ronald Weasley, adult
stormtrooper
Casino Dealer

Ancient Prince of India
Nick Fury
Captain Marvel
robot
Thor
Doctor Strange
Snow White
Cinderella
Crayon Shin-chan
Misae Nohara

General Job Descriptions. In addition, we ask GPT to generate general job descriptions such as policeman, doctor, chef and *etc.*

gardener
warrior
knight
witch
wizard
Joker
Clown
Chef
Nurse
Dentist
Doctor
Artist
Carpenter
Farmer
Athlete
Pilot
Firefighter
Police officer
astronaut

Imagined Outfits for Virtual Try-on. Here is an example, how we use ChatGPT to generate different outfit descriptions for Barack Obama.

We: please generate 50 prompts describing Obama wearing different clothes

ChatGPT:
Sure, here are 50 writing prompts that describe Barack Obama wearing different clothes:

Barack Obama wearing a sharp navy blue suit and a red tie.
Barack Obama wearing a casual grey t-shirt and jeans.
Barack Obama wearing a black leather jacket and aviator sunglasses.
Barack Obama wearing a traditional African dashiki.
Barack Obama wearing a classic American baseball cap and polo shirt.
Barack Obama wearing a white dress shirt and black dress pants.
Barack Obama wearing a bright yellow raincoat and rain boots.
Barack Obama wearing a navy blue hoodie and sneakers.
Barack Obama wearing a tan safari jacket and khaki pants.
Barack Obama wearing a red and black plaid flannel shirt and jeans.
Barack Obama wearing a crisp white dress shirt and a black suit.
Barack Obama wearing a light blue chambray shirt and khaki shorts.
Barack Obama wearing a navy blue blazer and chinos.
Barack Obama wearing a black and white houndstooth suit and tie.
Barack Obama wearing a green military-style jacket and combat boots.
Barack Obama wearing a purple dress shirt and grey suit.
Barack Obama wearing a denim shirt and khaki pants.
Barack Obama wearing a pink dress shirt and navy blue suit.
Barack Obama wearing a brown leather jacket and brown boots.
Barack Obama wearing a yellow polo shirt and white shorts.
Barack Obama wearing a grey sweatshirt and athletic pants.
Barack Obama wearing a black leather bomber jacket and jeans.
Barack Obama wearing a white t-shirt and denim jacket.
Barack Obama wearing a burgundy sweater and dress pants.
Barack Obama wearing a black and white checkered suit and tie.
Barack Obama wearing a green army jacket and cargo pants.
Barack Obama wearing a navy blue blazer and grey dress pants.
Barack Obama wearing a grey hoodie and black sweatpants.
Barack Obama wearing a yellow rain jacket and rain boots.
Barack Obama wearing a light blue dress shirt and navy blue suit.
Barack Obama wearing a black turtleneck and khaki pants.
Barack Obama wearing a red plaid shirt and jeans.
Barack Obama wearing a grey suit and a red tie.
Barack Obama wearing a white polo shirt and black shorts.
Barack Obama wearing a blue suit and a white dress shirt.
Barack Obama wearing a black t-shirt and black jeans.
Barack Obama wearing a purple dress shirt and black suit.
Barack Obama wearing a brown leather jacket and jeans.
Barack Obama wearing a white t-shirt and cargo pants.
Barack Obama wearing a grey suit and a blue tie.
Barack Obama wearing a black hoodie and black sweatpants.
Barack Obama wearing a green sweater and khaki pants.
Barack Obama wearing a navy blue suit and a white dress shirt.
Barack Obama wearing a red polo shirt and khaki shorts.
Barack Obama wearing a brown leather bomber jacket and jeans.
Barack Obama wearing a white dress shirt and black suit pants.
Barack Obama wearing a green camouflage jacket and cargo pants.
Barack Obama wearing a grey sweater and black dress pants.
Barack Obama wearing a blue plaid shirt and jeans.
Barack Obama wearing a black suit and a light blue dress shirt.