DreamAvatar: Text-and-Shape Guided 3D Human Avatar Generation via Diffusion Models

Yukang Cao[1*]   Yan-Pei Cao[2*]   Kai Han[1†]   Ying Shan[2]   Kwan-Yee K. Wong[1†]

[1]The University of Hong Kong   [2]ARC Lab, Tencent PCG

Abstract

*We present **DreamAvatar**, a text-and-shape guided framework for generating high-quality 3D human avatars with controllable poses. While encouraging results have been reported by recent methods on text-guided 3D common object generation, generating high-quality human avatars remains an open challenge due to the complexity of the human body's shape, pose, and appearance. We propose DreamAvatar to tackle this challenge, which utilizes a trainable NeRF for predicting density and color for 3D points and pretrained text-to-image diffusion models for providing 2D self-supervision. Specifically, we leverage the SMPL model to provide shape and pose guidance for the generation. We introduce a dual-observation-space design that involves the joint optimization of a canonical space and a posed space that are related by a learnable deformation field. This facilitates the generation of more complete textures and geometry faithful to the target pose. We also jointly optimize the losses computed from the full body and from the zoomed-in 3D head to alleviate the common multi-face "Janus" problem and improve facial details in the generated avatars. Extensive evaluations demonstrate that DreamAvatar significantly outperforms existing methods, establishing a new state-of-the-art for text-and-shape guided 3D human avatar generation.*

1. Introduction

The creation of 3D graphical human models has received great attention in recent years due to its wide-ranging applications in fields such as film-making, video games, AR/VR, and human-robotic interaction. Traditional methods for building such complex 3D models require thousands of man-hours of trained artists and engineers [10, 12], making the process both time-consuming and highly expert-

dependent. With the development of deep learning methods, we have witnessed the emergence of promising methods [5, 54, 61] which can reconstruct 3D human models from monocular images. These techniques, however, still face challenges in fully recovering details from the input images and rely heavily on the training dataset. To tackle these challenges and simplify the modeling process, adopting generative models for 3D human avatar modeling has recently received increasing attention from the research community. This approach has the potential to alleviate the need for large 3D datasets and facilitate easier and more accessible 3D human avatar modeling.

To leverage the potential of 2D generative image models for 3D content generation, recent methods [8, 29, 31, 35, 46] have utilized pretrained text-guided image diffusion models to optimize 3D implicit representations (e.g., NeRFs [37] and DMTet [40, 56]). DreamFusion [46] introduces a novel Score Distillation Sampling (SDS) strategy to self-supervise the optimization process and achieves promising results. However, human bodies, which are the primary focus of this paper, exhibit a complex articulated structure, with head, arms, hands, trunk, legs, feet, etc., each capable of posing in various ways. As a result, while Dream-Fusion [46] and subsequent methods (e.g., Magic3D [29], ProlificDreamer [58], Fantasia3D [8]) produce impressive results, they lack the proper constraints to enforce consistent 3D human structure and often struggle to generate detailed textures for 3D human avatars. Latent-NeRF [35] introduces a sketch-shape loss based on the 3D shape guidance, but it still faces challenges in generating reasonable results for human bodies.

In this paper, we present **DreamAvatar**, a novel framework for generating high-quality 3D human avatars from text prompts and shape priors. Inspired by previous works [29, 46], DreamAvatar employs a trainable NeRF as the base representation for predicting density and color features for each 3D point. Coupled with pretrained text-to-image diffusion models [50, 68], DreamAvatar can be trained to generate 3D avatars using 2D self-supervision. *The key innovation of DreamAvatar lies in three main aspects. Firstly*, we leverage the SMPL model [32] to provide
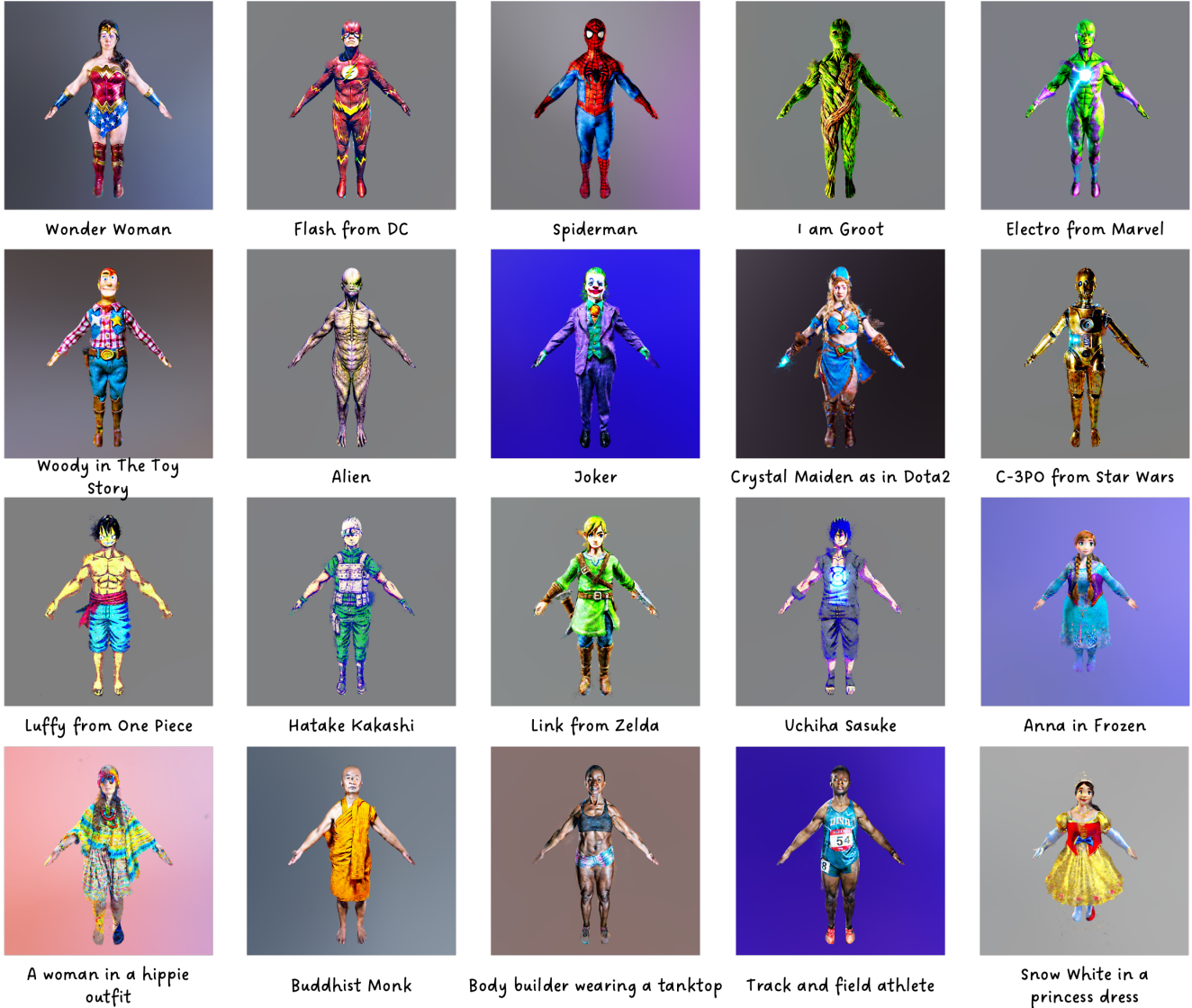

Figure 1. **Results of DreamAvatar.** DreamAvatar can generate high-quality geometry and texture for any type of human avatar.

a *shape prior*, which yields robust shape and pose guidance for the generation process. *Secondly*, we introduce a *dual-observation-space (DOS) design* consisting of a canonical space and a posed space that are related by a learnable deformation field and are jointly optimized. This facilitates the generation of more complete textures and geometry faithful to the target pose. *Thirdly*, we propose to jointly optimize the losses computed from the full body and from the zoomed-in 3D head to alleviate the multi-face "Janus" problem and improve facial details in the generated avatars.

We extensively evaluate DreamAvatar on generating movie/anime/video game characters, as well as general people, and compare it with previous methods. Experimental results show that our method significantly outperforms existing methods and can generate high-quality 3D human avatars with text-consistent geometry and geometry-consistent texture. We will make our code publicly available after publication.

2. Related Work

Text-guided 2D image generation. Recently, the CLIP model [47] (Contrastive Language-Image Pre-training) was proposed with the aim of classifying images and text by mapping them to a shared feature space. However, this model is not consistent with the way human perceives language, and it may not fully capture the intended meanings. With the improvements in neural networks and text-image datasets, the diffusion model has been introduced to handle more complex semantic concepts [2, 35, 48, 53]. Follow-up methods are designed to improve computational efficiency, for instance, through utilizing a cascade of super-resolution models [2, 53] or sampling from a low-resolution latent
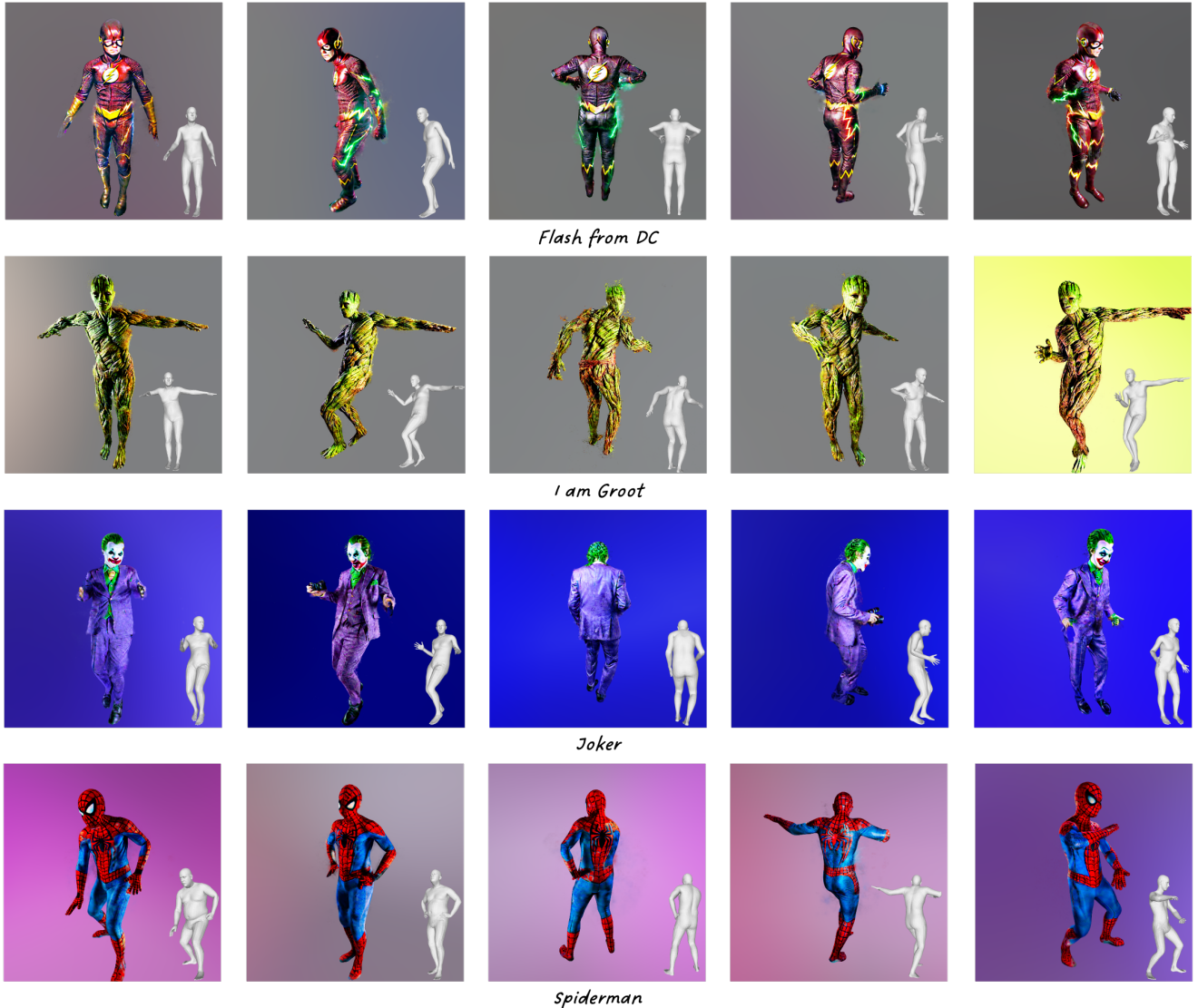
Figure 2. **Avatar generation with different poses.** Our method can handle and control the 3D generation with any pose.

space and decoding the latent features into high-resolution images [35]. DreamBooth [51] fine-tunes the diffusion model for certain subjects, while ControlNet [67] and T2I-Adapter [39] propose controlling the pretrained diffusion models with additional information. However, text-to-3D generation remains a challenge due to the lack of text-3D paired datasets and the associated high training cost .

Text-guided 3D content generation. Text-guided 3D content generation methods have emerged based on the success of text-guided 2D image generation. Earlier works, such as CLIP-forge [55], generate objects by learning a normalizing flow model from textual descriptions, but these methods are computationally expensive. DreamField [24], CLIP-mesh [27], AvatarCLIP [20], Text2Mesh [36], and Dream3D [62] rely on a pretrained image-text model [47] to optimize the underlying 3D representation (e.g., NeRF or mesh).

Recently, DreamFusion [46] proposes score distillation sampling based on the pretrained diffusion model [53] to enable text-guided 3D generation. Magic3D [29] improves it by introducing a coarse-to-fine pipeline to generate fine-grained 3D textured meshes. Point-E [41] and Shap-E [26] optimize the point cloud based on the diffusion model. Latent-NeRF [35] improves training efficiency by directly optimizing the latent features. TEXTure [49] applies a depth-diffusion model [50] to generate texture maps for a given 3D mesh. Fantasia3D [8] proposes a disentangled training strategy for geometry and texture generation. Guide3D [6] proposes to transfer multi-view generated images to 3D avatars. ProlificDreamer [58] introduces Variational Score Distillation (VSD) for better diversity and quality. Despite their promising performance, these methods still struggle to generate text-guided 3D human avatars due to the inherent challenges of this task.

3D human generative models. 3D generative methods based on 3D voxel grids [14, 19, 33, 60], point clouds [1, 34, 38, 63, 64, 70], and meshes [69] often require expensive and limited 3D datasets. In recent years, various methods [9, 30, 42, 45, 59] have been proposed to utilize NeRF and train on 2D human videos for novel view synthesis. Following these works, EG3D [7] and GNARF [3] propose a tri-plane representation and use GANs for 3D generation from latent codes. ENARF-GAN [43] extends NARF [42] to human representation. Meanwhile, EVA3D [21] and HumanGen [25] propose to generate human radiance fields directly from the 2D StyleGAN-Human [13] dataset. Although these methods have produced convincing results, they do not have the ability to "dream" or generate new subjects that have not been seen during training.

3. Methodology

Here, we introduce our text-and-shape guided generative network, DreamAvatar, which utilizes a trainable NeRF and pretrained diffusion models [50, 68] to generate 3D human avatars under controllable poses. DreamAvatar incorporates two observation spaces, namely a canonical space and a posed space, which are related by a learnable deformation field and are jointly optimized through a shared trainable NeRF (see Fig. 3). We jointly optimize the losses computed from the full-body and from the zoom-in 3D head to alleviate the multi-face "Janus" problem and improve the facial details in the generated avatars. In the following subsections, we first provide the preliminaries that underpin our method in Sec. 3.1. Next, we delve into the details of our method and discuss: (1) the density field derived from the SMPL model for evolving the geometry, (2) the dual observation spaces related by a learnable deformation field, and (3) the joint optimization of the losses computed from the full body and from the zoom-in 3D head in Sec. 3.2.

3.1. Preliminaries

Text-guided 3D generation methods Recent text-guided 3D generation models [46, 58] showcase promising results by incorporating three fundamental components:

(1) *NeRF* that represents a 3D scene via an implicit function, formulated as

$$F_\theta(\gamma(\mathbf{x})) \mapsto (\sigma, \mathbf{c}), \quad (1)$$

where $\mathbf{x}$ is a 3D point that is processed by the grid frequency encoder $\gamma(\cdot)$ [37], and σ and $\mathbf{c}$ denote its density value and color respectively. Typically, the implicit function $F_\theta(\cdot)$ is implemented as an MLP with trainable parameters θ.

(2) *Volume Rendering* technique that effectively renders a 3D scene onto a 2D image. For each image pixel, the rendering is done by casting a ray $\mathbf{r}$ from the pixel location into the 3D scene and sampling 3D points $\boldsymbol{\mu}_i$ along $\mathbf{r}$. The

density and color of the sampled points are predicted by F_θ. The RGB color C of each image pixel is then given by

$$C(\mathbf{r}) = \sum_i W_i \mathbf{c}_i, \quad W_i = \alpha_i \prod_{j<i}(1 - \alpha_j) \quad (2)$$

where $\alpha_i = 1 - e^{(-\sigma_i \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_{i+1}\|)}$.

(3) *Variational Score Distillation (VSD)* derived on text-guided diffusion models ϕ [50, 53]. We employ a pretrained diffusion model [50, 68] with a learned denoising function $\epsilon_{\text{pre}}(\boldsymbol{x_t}; y, t)$. Here $\boldsymbol{x_t}$ denotes the noisy image at timestep t, and y is the text embedding. Given an image $\boldsymbol{g}(\theta, c)$ rendered from the NeRF with camera parameters c, we add random noise ϵ to obtain a noisy image $\boldsymbol{x}_t = \alpha_t \boldsymbol{g}(\theta, c) + \sigma_t \epsilon$ (α_t and σ_t are hyperparameters). We then parameterize ϵ using LoRA (Low-Rank Adaptation [22, 52]) of the pretrained model $\epsilon_{\text{pre}}(x_t; y, t)$ to obtain ϵ_ϕ, and add camera parameters c to the condition embeddings in the network. The gradient for updating the NeRF is then given by

$$\nabla_\theta \mathcal{L}_{\text{VSD}}(\theta) \triangleq \mathbb{E}_{t,\epsilon,c} \left[\omega(t) \left(\epsilon_{\text{pre}}(\boldsymbol{x}_t, t, y) - \epsilon_\phi(\boldsymbol{x}_t, t, c, y) \right) \frac{\partial \boldsymbol{g}(\theta, c)}{\partial \theta} \right], \quad (3)$$

where $w(t)$ is a weighting function that depends on the timestep t.

SMPL [4, 44] 3D parametric human model It builds a 3D human shape using 6,890 body vertices. Formally, by assembling pose parameters ξ and shape parameters β, we can obtain the 3D SMPL human model by:

$$T_P(\beta, \xi) = \bar{T} + B_S(\beta; \mathcal{S}) + B_P(\xi; \mathcal{P}), \quad (4)$$
$$M(\beta, \xi) = \texttt{LBS}(T_P(\beta, \xi), J(\beta), \xi, \mathcal{W}), \quad (5)$$

where $T_P(\cdot)$ represents the non-rigid deformation from the canonical model $\bar{T}$ using the shape blend shape function B_S and pose blend shape function B_P. $\mathcal{S}$ and $\mathcal{P}$ are the principal components of vertex displacements. $\texttt{LBS}(\cdot)$ denotes the linear blend skinning function, corresponding to articulated deformation. It poses $T_P(\cdot)$ based on the pose parameters ξ and joint locations $J(\beta)$, using the blend weights $\mathcal{W}$, individually for each body vertex:

$$\mathbf{v}_p = \mathcal{G} \cdot \mathbf{v}_c, \quad \mathcal{G} = \sum_{k=1}^{K} w_k \mathcal{G}_k(\xi, j_k), \quad (6)$$

where $\mathbf{v}_c$ is an SMPL vertex under the canonical pose, $\mathbf{v}_p$ denotes the corresponding vertex under the given pose, w_k is the skinning weight, $\mathcal{G}_k(\xi, j_k)$ is the affine deformation that transforms the k-th joint j_k from the canonical space to posed space, and K is the number of neighboring joints.

Unlike the original SMPL which defines "T-pose" as the canonical model, here, we adopt "A-pose" as the canonical model which is a more natural human rest pose for the diffusion models to understand (see Fig. 3).
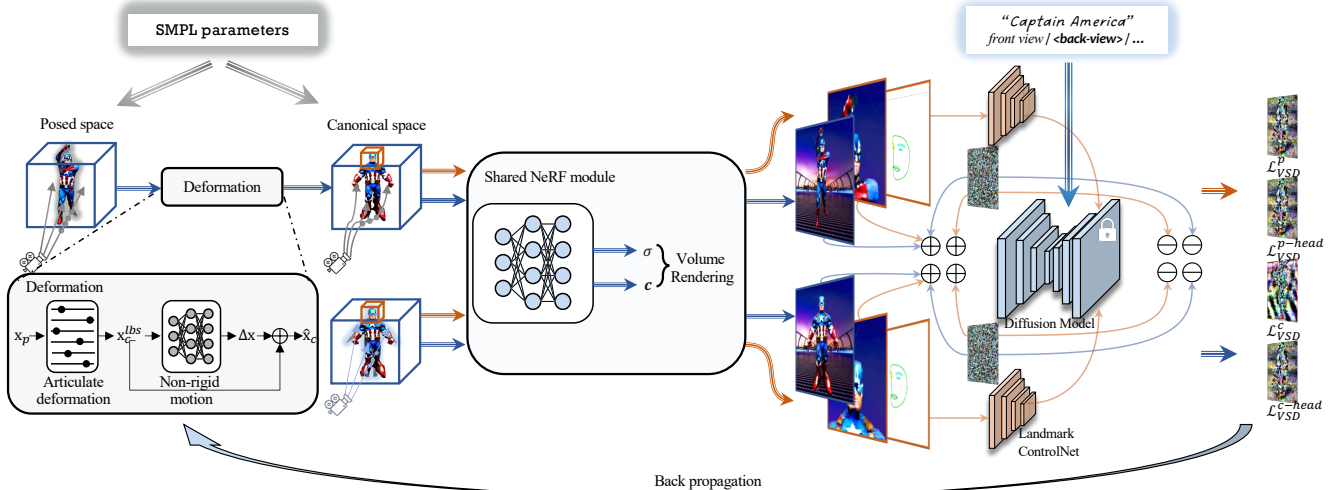
Figure 3. **Overview of DreamAvatar.** Our network takes as input a text prompt and SMPL parameters to optimize a trainable NeRF via a pretrained denoising stable diffusion model. At the core of our network are two observation spaces, namely the canonical space and the posed space, that are related by an SMPL-based learnable deformation field. This dual-observation-space design facilities the generation of more complete textures and geometry faithful to the target pose.

3.2. DreamAvatar

As illustrated in Fig. 3, our proposed framework takes as input a text prompt and SMPL parameters, defining the target shape and pose in the posed space. DreamAvatar conducts Variational Score Distillation (VSD)-based optimization [58] in both the canonical space and posed space simultaneously and learns a deformation field relating the two spaces. To represent the canonical space and posed space, we utilize an extended neural radiance field where the density, RGB color value, and normal direction of each sample point can be queried and optimized. We utilize the input SMPL parameters to handle different body parts separately and derive good initial density values in each space. To alleviate the multi-face "Janus" problem and improve facial details in the generated avatars, we optimize, in addition to the loss computed from the full body, the loss computed from the zoomed-in 3D head using a landmark-based ControlNet [67] and a learned special <back-view> token [15].

SMPL-derived density fields We propose to make our NeRF evolve from the density field derived from the input SMPL model. Specifically, given a 3D point $\mathbf{x}_c$ in the canonical space, we first calculate its signed distance d_c to the SMPL surface in the canonical pose and convert it to a density value $\bar{\sigma}_c$ using

$$\bar{\sigma}_c = \texttt{max}(0, \texttt{softplus}^{-1}(\frac{1}{a}\texttt{sigmoid}(-d_c/a))),$$

where $\texttt{sigmoid}(x) = 1/(1 + e^{-x})$, $\texttt{softplus}^{-1}(x) = \texttt{log}(e^x - 1)$, and a is a predefined hyperparameter [62] which is set to 0.001 in our experiments. Similarly, given a point $\mathbf{x}_p$ in the posed space (the upper branch in Fig. 3), we

compute its density value $\bar{\sigma}_p$ from its signed distance d_p to the SMPL surface in the target pose.

Deformation field Inspired by HumanNeRF [59], we employ a deformation field to map points $\mathbf{x}_p$ from the posed space to their corresponding points $\hat{\mathbf{x}}_c$ in the canonical space. The deformation field is composed of two parts, namely (1) articulated deformation that applies the inverse transformation of SMPL linear blend skinning $\texttt{LBS}(\cdot)$ (Sec. 3.1), and (2) non-rigid motion modelled by an MLP to learn the corrective offset:

$$\hat{\mathbf{x}}_c = \mathcal{G}^{-1} \cdot \mathbf{x}_p + \texttt{MLP}_{\theta_{\text{NR}}}(\gamma(\mathcal{G}^{-1} \cdot \mathbf{x}_p)), \tag{7}$$

where $\mathcal{G}$ is obtained from posed SMPL vertex closest to $\mathbf{x}_p$.

Dual observation spaces (DOS) Given a 3D point $\mathbf{x}_c$ in the canonical space and $\mathbf{x}_p$ in the posed space, we compute their density values σ_c, σ_p and latent color features $\mathbf{c}_c, \mathbf{c}_p$ by

$$F(\mathbf{x}_c, \bar{\sigma}_c) = F_\theta(\gamma(\mathbf{x}_c)) + (\bar{\sigma}_c, \mathbf{0}) \mapsto (\sigma_c, \mathbf{c}_c), \tag{8}$$

$$F(\mathbf{x}_p, \bar{\sigma}_p) = F_\theta(\gamma(\hat{\mathbf{x}}_c)) + (\bar{\sigma}_p, \mathbf{0}) \mapsto (\sigma_p, \mathbf{c}_p), \tag{9}$$

where $\hat{\mathbf{x}}_c$ denotes the corresponding point of $\mathbf{x}_p$ in the canonical space, and $\bar{\sigma}_c$ and $\bar{\sigma}_p$ are the SMPL-derived density values in the canonical space and posed space respectively. Our DOS design serves two main purposes. Firstly, the avatar in rest pose in the canonical space exhibits minimum self-occlusion. Observations in the canonical space therefore facilitate the generation of more complete textures. Observations in the posed space, on the other hand, facilitate the generation of geometry faithful to the target pose. They also provide extra supervision in optimizing the NeRF. Our DOS design also differentiates itself from MPS-NeRF [16] and TADA! [28] by prioritizing joint optimization and mutual distillation between the canonical space and
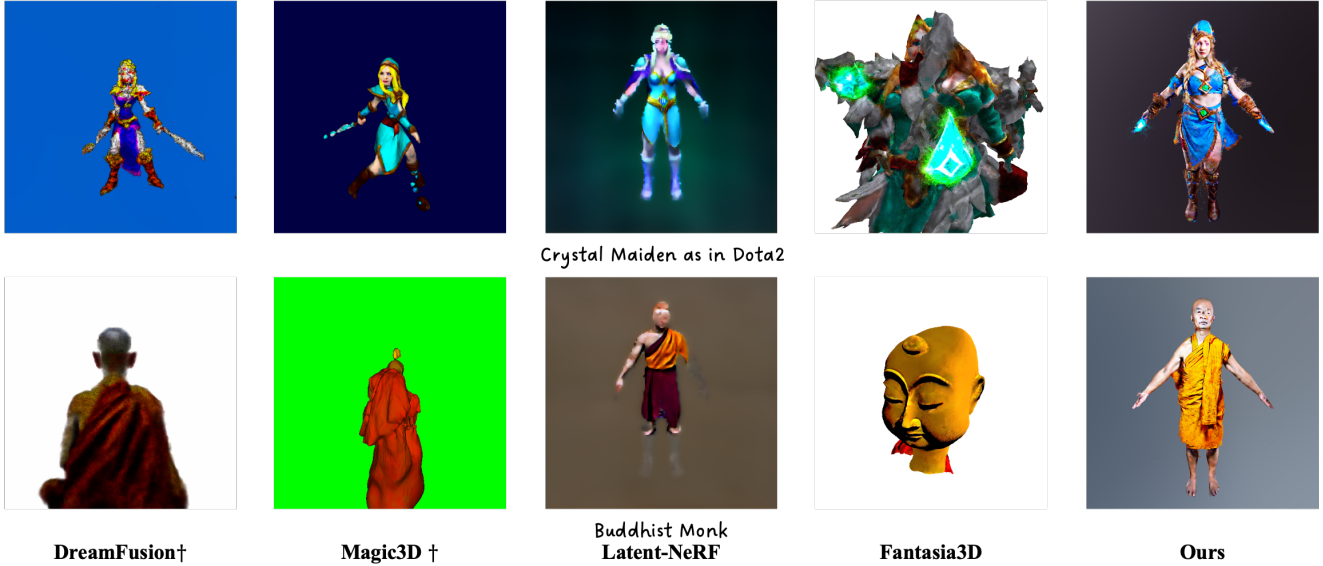
Figure 4. **Comparison with existing text-to-3D baselines.** While baselines struggle to ensure structurally and topologically correct geometry, our method can produce high-quality geometry and texture.
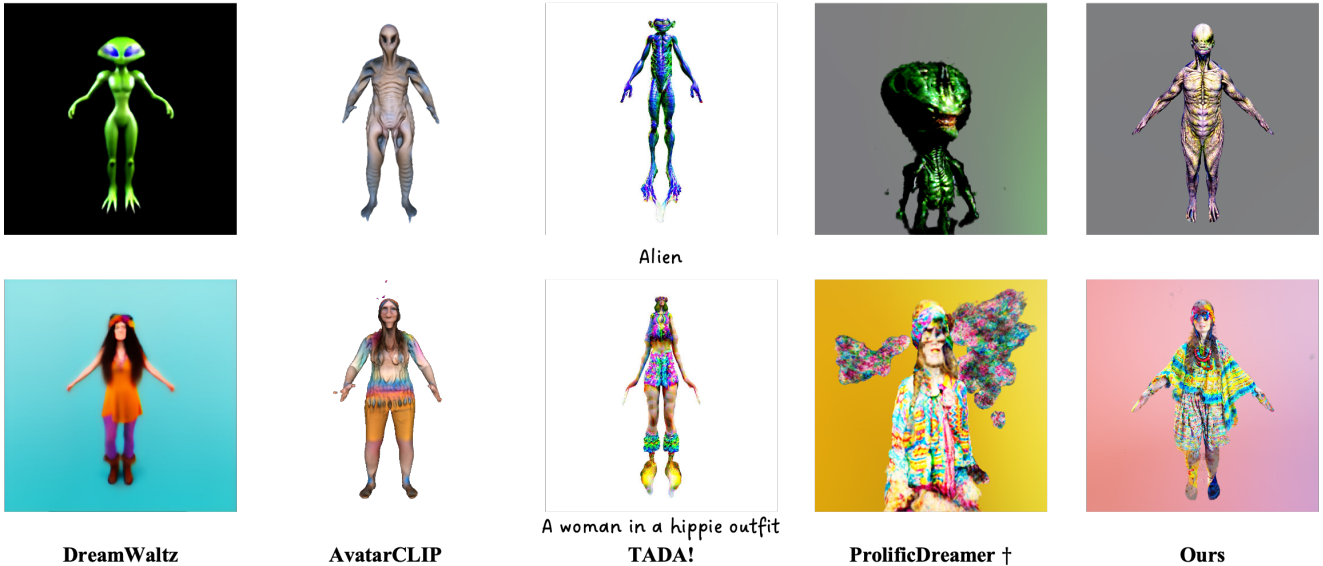

Figure 5. **Comparison with existing avatar-specified baselines.** Our method can generate 3D human avatars with geometry and texture with much higher resolution.

posed space. In contrast, existing methods only utilize observations in the posed space without fully exploiting information in the canonical space.

Zoomed-in head Inspired by JIFF [5] which enhances human reconstruction by learning to recover details in the zoomed-in face, we propose to optimize, in addition to the loss computed from the full body, the loss computed from the zoomed-in 3D head to improve the facial details in the generated 3D avatars. Specifically, we render a zoomed-in head image at each iteration for computing the VSD loss on head. To alleviate the common multi-face "Janus" problem, we follow HeadSculpt [18] to employ a landmark-based ControlNet $\mathcal{C}$ [67] and a learned special <back-view> to-

ken [15] to compute the head VSD loss. The gradient for updating the NeRF now becomes

$$\nabla_\theta \mathcal{L}_{\text{VSD}}^{\text{head}}(\theta) \triangleq \mathbb{E}_{t,\epsilon,c} \left[\omega(t) \left(\epsilon_{\text{pre}}(\cdot) - \epsilon_\phi(\cdot) \right) \frac{\partial \boldsymbol{g}(\theta, c)}{\partial \theta} \right], \quad (10)$$

$$\epsilon_{\text{pre}}(\cdot) = \epsilon_{\text{pre}}(\boldsymbol{x}_t, t, y, \mathcal{C}(\mathcal{P}_\pi)), \epsilon_\phi(\cdot) = \epsilon_\phi(\boldsymbol{x}_t, t, c, y, \mathcal{C}(\mathcal{P}_\pi)),$$

where $\mathcal{P}_\pi$ denotes the facial landmark map obtained from the projection of the SMPL head model.

4. Experiments

We now validate the effectiveness and capability of our proposed framework on a variety of text prompts and provide comparisons with existing text-guided 3D generation methods using the same text prompts.

†*wearing green suit*   †*wearing texudo*   †*wearing pink suit*   †*wearing black suit*   †*wearing navy suit*

Figure 6. **Text manipulation over the avatar generations.** † indicates "Joker"

I am Groot

Figure 7. **Shape modification via SMPL shape parameters.** Our method can generate 3D avatars with different shapes based on the input SMPL shape parameters, e.g., "short", "fat", "thin", "tall" as above from left to right.

Implementation details We follow threestudio [17] to implement the NeRF [37] and Variational Score Distillation in our method. We utilize Ver-2.1 of Stable Diffusion [57] and Ver-1.1 of ControlNet [11, 68] in our implementation. Typically, for each text prompt, we train our network for 10, 000 iterations on one single NVIDIA A40 GPU.

Baseline methods We compare our method with DreamFusion [46], Latent-NeRF [35], Fantasia3D [8], Magic3D [29], ProlificDreamer [58], DreamWaltz [23], AvatarCLIP [20], and TADA! [28]. We are not able to compare with AvatarBooth [65] and AvatarVerse [66] as their codes are not publicly available.

4.1. Qualitative Evaluations

Avatar generation with different styles In Fig. 1, we provide a diverse set of 3D human avatars, e.g., real-world human beings, movie, anime, and video game characters, generated by our DreamAvatar. We can consistently observe high-quality geometry and texture from all these examples.

Avatar generation under different poses In Fig. 2, we validate the effectiveness of our method for generating 3D human avatars in various poses, which is not achievable by other existing methods due to the absence of the shape prior. Thanks to our DOS design, DreamAvatar can maintain high-quality texture and geometry for extreme poses, e.g., complex poses with severe self-occlusion.

Comparison with SOTA methods We provide qualitative comparisons with existing SOTA methods in Fig. 4 and Fig. 5. We can observe that our method consistently achieves topologically and structurally correct geometry and texture compared to baseline methods, and outperforms

the avatar-specified generative methods with much better and higher-resolution texture and geometry. See supplementary for more comparisons.

Text manipulation on avatar generation We explore the capabilities of DreamAvatar by editing the text prompt for controlled generation (see Fig. 6). Our method can generate faithful avatars that accurately embody the provided text, incorporating additional descriptive information and capturing the unique characteristics of the main subject.

Shape modification via SMPL shape parameters We further demonstrate the possibility of our method to generate different sizes of 3D human avatars, e.g., thin, short, tall, fat, by editing the SMPL shape parameters in Fig. 7.

Attributes integration: Zoomed-in head and full body Benefiting from the joint modeling of the zoomed-in 3D head and full body, our method seamlessly integrates the unique attributes derived from both head and body characters. See visualizations in Fig. 8. In contrast to the conventional approach of separately modeling head and body parts, DreamAvatar harmoniously combines these elements, resulting in a cohesive model that faithfully captures the essence of both subjects.

4.2. User Studies

We conduct user studies to compare with SOTA methods [8, 20, 23, 29, 35, 46, 58]. 25 volunteers are presented with 20 examples to rate these methods in terms of (1) geometry quality, (2) texture quality, and (3) consistency with the text from 1 (worst) to 8 (best). The final rates in Fig. 9 clearly show that our method achieves the best rankings in all three aspects.
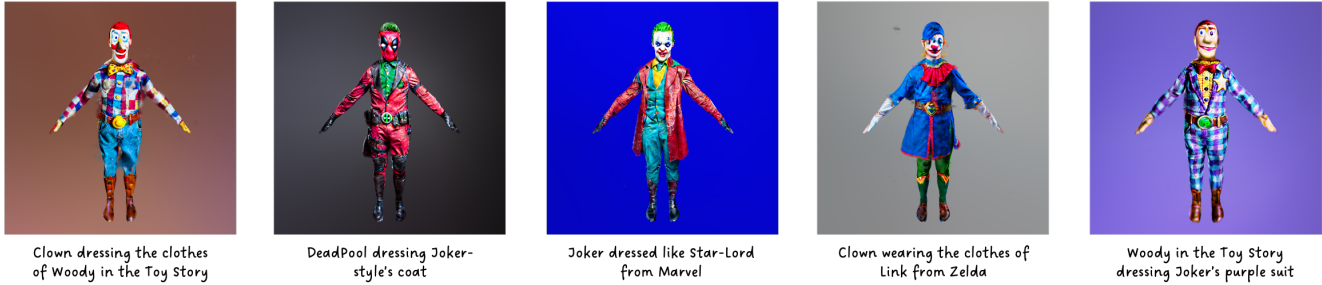
| Clown dressing the clothes of Woody in the Toy Story | DeadPool dressing Joker-style's coat | Joker dressed like Star-Lord from Marvel | Clown wearing the clothes of Link from Zelda | Woody in the Toy Story dressing Joker's purple suit |

Figure 8. **Attributes integration between head and body parts.**
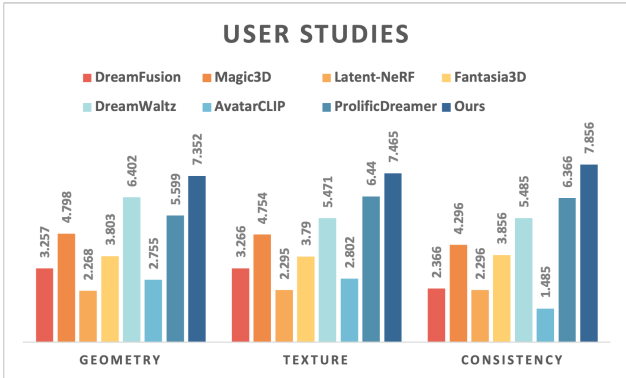
Figure 9. **User study on rotating 3D human avatars.**

Figure 10. **Analysis of the setup for SMPL-derived density** $\bar{\bar{\sigma}}_c, \bar{\bar{\sigma}}_p$**, and dual-space design.**

Figure 11. **Analysis of joint modeling of head region.**

4.3. Further Analysis

Effectiveness of SMPL-derived density We optimize the NeRF without using the SMPL-derived densities $\bar{\sigma}_c, \bar{\sigma}_p$ as the basis for density prediction. We find that without using $\bar{\sigma}_c$ and $\bar{\sigma}_p$, (1) the generated avatars exhibit low-quality geometry and texture with strong outliers, and (2) the generated shapes are not constrained to reasonable human bodies and are not view-consistent (see "w/o SMPL-density field" in Fig. 10).

Effectiveness of incorporating head VSD loss In order to assess its impact, we conduct ablation studies by disabling the head VSD loss. The results are presented in Fig. 11. Notably, we observe a significant drop in the quality of the generated head region. Moreover, the multi-face "Janus" problem becomes more pronounced in the generated content.

Effectiveness of DOS design To validate our design, we experiment with two degenerated versions of our framework: (1) only the canonical space, and (2) only the posed space without deformation field. Results in Fig. 10 clearly demonstrate that neither of these degenerated designs can match the performance achieved by our DOS design.

5. Conclusions

In this paper, we have introduced DreamAvatar, an effective framework for text-and-shape guided 3D human avatar generation. In DreamAvatar, we propose to leverage the parametric SMPL model to provide shape prior, guiding the generation with a rough 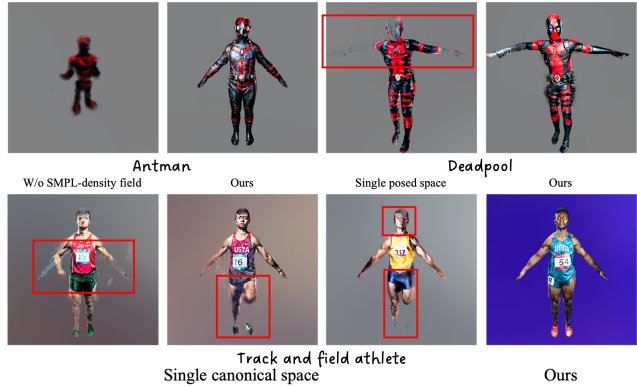shape and pose. We also propose a dual-observation-space design, facilitating the generation of more complete textures and geometry faithful to the target pose. Additionally, we propose to jointly optimize the loss computed from the full body and from the zoomed-in 3D head, effectively alleviating the multi-face "Janus" problem and improving facial details in the generated avatars. Extensive experiments show that our method has achieved state-of-the-art 3D human avatar generation.

Limitations Despite establishing a new state-of-the-art, DreamAvatar encounters limitations: **(1)** Animation was not considered in our current implementation of DreamAvatar; **(2)** The model inherits biases from the pretrained diffusion model due to the text-image data distribution, such that performance on more frequently appeared subjects in the pretraining data may be better than the others.

Societal impact Advancements in 3D avatar generation can streamline metaverse development. However, dangers exist regarding the nefarious use of this technology to generate plausible renderings of individuals. Open and transparent research and usage are encouraged.

References

[1] Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas Guibas. Learning representations and generative models for 3d point clouds. In *International conference on machine learning*, 2018. 4

[2] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022. 2

[3] Alexander W. Bergman, Petr Kellnhofer, Wang Yifan, Eric R. Chan, David B. Lindell, and Gordon Wetzstein. Generative neural articulated radiance fields. In *Advances in Neural Information Processing Systems*, 2022. 4

[4] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *European Conference on Computer Vision*, 2016. 4

[5] Yukang Cao, Guanying Chen, Kai Han, Wenqi Yang, and Kwan-Yee K Wong. Jiff: Jointly-aligned implicit face function for high quality single view clothed human reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 1, 6

[6] Yukang Cao, Yan-Pei Cao, Kai Han, Ying Shan, and Kwan-Yee K Wong. Guide3d: Create 3d avatars from text and image guidance. *arXiv preprint arXiv:2308.09705*, 2023. 3

[7] Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3D generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 4

[8] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. 2023. 1, 3, 7

[9] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 4

[10] Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese, Hugues Hoppe, Adam G. Kirk, and Steve Sullivan. High-quality streamable free-viewpoint video. *ACM Transactions on Graphics*, 2015. 1

[11] CrucibleAI. ControlNetMediaPipeFace. https://huggingface.co/CrucibleAI/ControlNetMediaPipeFace, 2023. 7

[12] Massimiliano Favalli, Alessandro Fornaciai, Ilaria Isola, Simone Tarquini, and Luca Nannipieri. Multiview 3d reconstruction in geosciences. *Computers Geosciences*, 2012. 1

[13] Jianglin Fu, Shikai Li, Yuming Jiang, Kwan-Yee Lin, Chen Qian, Chen Change Loy, Wayne Wu, and Ziwei Liu. Stylegan-human: A data-centric odyssey of human generation. In *European Conference on Computer Vision*, 2022. 4

[14] Matheus Gadelha, Subhransu Maji, and Rui Wang. 3d shape induction from 2d views of multiple objects. In *International Conference on 3D Vision*, 2017. 4

[15] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. 2023. 5, 6

[16] Xiangjun Gao, Jiaolong Yang, Jongyoo Kim, Sida Peng, Liu Zicheng, and Xin Tong. Mps-nerf: Generalizable 3d human rendering from multiview images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 5

[17] Yuan-Chen Guo, Ying-Tian Liu, Ruizhi Shao, Christian Laforte, Vikram Voleti, Guan Luo, Chia-Hao Chen, Zi-Xin Zou, Chen Wang, Yan-Pei Cao, and Song-Hai Zhang. threestudio: A unified framework for 3d content generation. https://github.com/threestudio-project/threestudio, 2023. 7

[18] 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. 6

[19] Philipp Henzler, Niloy J Mitra, and Tobias Ritschel. Escaping plato's cave: 3d shape from adversarial rendering. In *International Conference on Computer Vision*, 2019. 4

[20] Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, and Ziwei Liu. Avatarclip: Zero-shot text-driven generation and animation of 3d avatars. *ACM Transactions on Graphics*, 2022. 3, 7

[21] Fangzhou Hong, Zhaoxi Chen, Yushi LAN, Liang Pan, and Ziwei Liu. EVA3d: Compositional 3d human generation from 2d image collections. In *International Conference on Learning Representations*, 2023. 4

[22] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 4

[23] 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. 2023. 7

[24] Ajay Jain, Ben Mildenhall, Jonathan T. Barron, Pieter Abbeel, and Ben Poole. Zero-shot text-guided object generation with dream fields. 2022. 3

[25] Suyi Jiang, Haoran Jiang, Ziyu Wang, Haimin Luo, Wenzheng Chen, and Lan Xu. Humangen: Generating human radiance fields with explicit priors. *arXiv preprint arXiv:2212.05321*, 2022. 4

[26] Heewoo Jun and Alex Nichol. Shap-e: Generating conditional 3d implicit functions. *arXiv preprint arXiv:2305.02463*, 2023. 3

[27] Nasir Mohammad Khalid, Tianhao Xie, Eugene Belilovsky, and Popa Tiberiu. Clip-mesh: Generating textured meshes from text using pretrained image-text models. 2022. 3

[28] Tingting Liao, Hongwei Yi, Yuliang Xiu, Jiaxiang Tang, Yangyi Huang, Justus Thies, and Michael J Black. Tada! text to animatable digital avatars. *arXiv preprint arXiv:2308.10899*, 2023. 5, 7

[29] 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. *arXiv preprint arXiv:2211.10440*, 2022. 1, 3, 7

[30] Lingjie Liu, Marc Habermann, Viktor Rudnev, Kripasindhu Sarkar, Jiatao Gu, and Christian Theobalt. Neural actor: Neural free-view synthesis of human actors with pose control. *ACM SIGGRAPH Asia*, 2021. 4

[31] 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. 2023. 1

[32] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. SMPL: A skinned multi-person linear model. *TOG*, 2015. 1

[33] Sebastian Lunz, Yingzhen Li, Andrew Fitzgibbon, and Nate Kushman. Inverse graphics gan: Learning to generate 3d shapes from unstructured 2d data. *arXiv preprint arXiv:2002.12674*, 2020. 4

[34] Shitong Luo and Wei Hu. Diffusion probabilistic models for 3d point cloud generation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 4

[35] Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. Latent-nerf for shape-guided generation of 3d shapes and textures. *arXiv preprint arXiv:2211.07600*, 2022. 1, 2, 3, 7

[36] Oscar Michel, Roi Bar-On, Richard Liu, Sagie Benaim, and Rana Hanocka. Text2mesh: Text-driven neural stylization for meshes. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 3

[37] 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*, 2020. 1, 4, 7

[38] Kaichun Mo, Paul Guerrero, Li Yi, Hao Su, Peter Wonka, Niloy Mitra, and Leonidas Guibas. Structurenet: Hierarchical graph networks for 3d shape generation. *ACM Transactions on Graphics*, 2019. 4

[39] 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 arXiv:2302.08453*, 2023. 3

[40] Jacob Munkberg, Jon Hasselgren, Tianchang Shen, Jun Gao, Wenzheng Chen, Alex Evans, Thomas Müller, and Sanja Fidler. Extracting triangular 3d models, materials, and lighting from images. 2022. 1

[41] Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. Point-e: A system for generating 3d point clouds from complex prompts. *arXiv preprint arXiv:2212.08751*, 2022. 3

[42] Atsuhiro Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. Neural articulated radiance field. In *International Conference on Computer Vision*, 2021. 4

[43] Atsuhiro Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. Unsupervised learning of efficient geometry-aware neural articulated representations. In *European Conference on Computer Vision*, 2022. 4

[44] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 4

[45] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 4

[46] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 1, 3, 4, 7

[47] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 2021. 2, 3

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

[49] Elad Richardson, Gal Metzer, Yuval Alaluf, Raja Giryes, and Daniel Cohen-Or. Texture: Text-guided texturing of 3d shapes. *arXiv preprint arXiv:2302.01721*, 2023. 3

[50] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 1, 3, 4

[51] 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. 2022. 3

[52] Simo Ryu. Low-rank adaptation for fast text-to-image diffusion fine-tuning. https://github.com/cloneofsimo/lora. 2023. 4

[53] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 2, 3, 4

[54] 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*, 2019. 1

[55] 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 *IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 3

[56] Tianchang Shen, Jun Gao, Kangxue Yin, Ming-Yu Liu, and Sanja Fidler. Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. 2021. 1

[57] Stability.AI. Stable diffusion. `https://stability.ai/blog/stable-diffusion-public-release`, 2022. 7

[58] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *arXiv preprint arXiv:2305.16213*, 2023. 1, 3, 4, 5, 7

[59] Chung-Yi Weng, Brian Curless, Pratul P Srinivasan, Jonathan T Barron, and Ira Kemelmacher-Shlizerman. Humannerf: Free-viewpoint rendering of moving people from monocular video. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 4, 5

[60] Jiajun Wu, Chengkai Zhang, Tianfan Xue, Bill Freeman, and Josh Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. *Advances in Neural Information Processing Systems*, 29, 2016. 4

[61] Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J. Black. ECON: Explicit Clothed humans Obtained from Normals. In *IEEELearning Transferable Visual Models From Natural Language Supervision Conference on Computer Vision and Pattern Recognition*, 2023. 1

[62] 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. *arXiv preprint arXiv:2212.14704*, 2022. 3, 5

[63] Guandao Yang, Xun Huang, Zekun Hao, Ming-Yu Liu, Serge Belongie, and Bharath Hariharan. Pointflow: 3d point cloud generation with continuous normalizing flows. In *International Conference on Computer Vision*, 2019. 4

[64] Xiaohui Zeng, Arash Vahdat, Francis Williams, Zan Gojcic, Or Litany, Sanja Fidler, and Karsten Kreis. Lion: Latent point diffusion models for 3d shape generation. In *Advances in Neural Information Processing Systems*, 2022. 4

[65] Yifei Zeng, Yuanxun Lu, Xinya Ji, Yao Yao, Hao Zhu, and Xun Cao. Avatarbooth: High-quality and customizable 3d human avatar generation. 2023. 7

[66] 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, 2023. 7

[67] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 3, 5, 6

[68] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023. 1, 4, 7

[69] Yuxuan Zhang, Wenzheng Chen, Huan Ling, Jun Gao, Yinan Zhang, Antonio Torralba, and Sanja Fidler. Image gans meet differentiable rendering for inverse graphics and interpretable 3d neural rendering. 2021. 4

[70] Linqi Zhou, Yilun Du, and Jiajun Wu. 3d shape generation and completion through point-voxel diffusion. In *International Conference on Computer Vision*, 2021. 4