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

Yukang Cao1* Yan-Pei Cao2* Kai Han† Ying Shan2 Kwan-Yee K. Wong††
1The University of Hong Kong 2ARC 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 DreamFusion [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...
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.
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).

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

\[ F_\theta(\gamma(x)) \rightarrow (\sigma, c), \]  

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

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 \( r \) from the pixel location into the 3D scene and sampling 3D points \( \mu_i \) along \( r \). The density and color of the sampled points are predicted by \( F_\theta \). The RGB color \( C \) of each image pixel is then given by

\[ C(r) = \sum_i W_i c_i, \quad W_i = \alpha_i \prod_{j<i} (1 - \alpha_j) \]  

where \( \alpha_i = 1 - e^{-(\sigma_i||\mu_i - \mu_{i+1}||)} \).

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

\[ \nabla_\theta L_{VSD}(\theta) \approx E_{t, \epsilon, c} \left[ \omega(t) \left( \epsilon_{pr}(x_t; y, t) - \epsilon_{pr}(x_t; y, t, c) \right) \frac{2m(x, \omega)}{m(x, \omega) + 1} \right], \]  

where \( \omega(t) \) 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 \( \xi \) and shape parameters \( \beta \), we can obtain the 3D SMPL human model by:

\[ T_p(\beta, \xi) = \hat{T} + B_S(\beta; S) + B_P(\xi; P), \]  
\[ M(\beta, \xi) = \text{LBS}(T_p(\beta, \xi), J(\beta), \xi, W), \]  

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

\[ v_p = G \cdot v_c, \quad G = \sum_{k=1}^{K} w_k G_k(\xi, j_k), \]  

where \( v_c \) is an SMPL vertex under the canonical pose, \( v_p \) denotes the corresponding vertex under the given pose, \( w_k \) is the skinning weight, \( G_k(\xi, j_k) \) 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).
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 \( x_c \) 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 \( \sigma_c \) using

\[
\sigma_c = \max(0, \text{softplus}^{-1}(\frac{1}{\alpha} \text{sigmoid}(-d_c/a))),
\]

where \( \text{sigmoid}(x) = 1/(1 + e^{-x}) \), \( \text{softplus}^{-1}(x) = \log(e^x - 1) \), and \( \alpha \) is a predefined hyperparameter [62] which is set to 0.001 in our experiments. Similarly, given a point \( x_p \) in the posed space (the upper branch in Fig. 3), we compute its density value \( \sigma_p \) 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 \( x_p \) from the posed space to their corresponding points \( \hat{x}_c \) 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 LBS(·) (Sec. 3.1), and (2) non-rigid motion modelled by an MLP to learn the corrective offset:

\[
\hat{x}_c = G^{-1} \cdot x_p + \text{MLP}_{\theta_{\text{LBS}}} (\gamma(G^{-1} \cdot x_p)),
\]

where \( G \) is obtained from posed SMPL vertex closest to \( x_p \).

Dual observation spaces (DOS) Given a 3D point \( x_c \) in the canonical space and \( x_p \) in the posed space, we compute their density values \( \sigma_c, \sigma_p \) and latent color features \( c_c, c_p \) by

\[
F(x_c, \sigma_c) = F_\theta(\gamma(x_c)) + (\sigma_c, 0) \mapsto (\sigma_c, c_c),
\]

\[
F(x_p, \sigma_p) = F_\theta(\gamma(x_c)) + (\sigma_p, 0) \mapsto (\sigma_p, c_p),
\]

where \( \hat{x}_c \) denotes the corresponding point of \( x_p \) in the canonical space, and \( \sigma_c \) and \( \sigma_p \) 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
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> token [15] to compute the head VSD loss. The gradient for updating the NeRF now becomes

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

(10)

where $\mathcal{P}_x$ 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.
Figure 6. **Text manipulation over the avatar generations.** † indicates “Joker”

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.
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


[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


[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