Fantasia3D: Disentangling Geometry and Appearance for High-quality Text-to-3D Content Creation

Rui Chen[*], Yongwei Chen[*], Ningxin Jiao, Kui Jia[†]
South China University of Technology

Abstract

Automatic 3D content creation has achieved rapid progress recently due to the availability of pre-trained, large language models and image diffusion models, forming the emerging topic of text-to-3D content creation. Existing text-to-3D methods commonly use implicit scene representations, which couple the geometry and appearance via volume rendering and are suboptimal in terms of recovering finer geometries and achieving photorealistic rendering; consequently, they are less effective for generating high-quality 3D assets. In this work, we propose a new method of Fantasia3D for high-quality text-to-3D content creation. Key to Fantasia3D is the disentangled modeling and learning of geometry and appearance. For geometry learning, we rely on a hybrid scene representation, and propose to encode surface normal extracted from the representation as the input of the image diffusion model. For appearance modeling, we introduce the spatially varying bidirectional reflectance distribution function (BRDF) into the text-to-3D task, and learn the surface material for photorealistic rendering of the generated surface. Our disentangled framework is more compatible with popular graphics engines, supporting relighting, editing, and physical simulation of the generated 3D assets. We conduct thorough experiments that show the advantages of our method over existing ones under different text-to-3D task settings. Project page and source codes: `https://fantasia3d.github.io/`.

1. Introduction

Automatic 3D content creation [43, 18, 33, 44] powered by large language models has drawn significant attention recently, due to its convenience to entertaining and gaming industries, virtual/augmented reality, and robotic applications. The traditional process of creating 3D assets typically involves multiple, labor-intensive stages, including geometry modeling, shape baking, UV mapping, material creation, and texturing, as described in [15], where different software tools and the expertise of skilled artists are often required. Imperfections would also accumulate across these stages, resulting in low-quality 3D assets. It is thus desirable to automate such a process, and ideally to generate high-quality 3D assets that have geometrically fair surfaces, rich materials and textures, and support photorealistic rendering under arbitrary views.

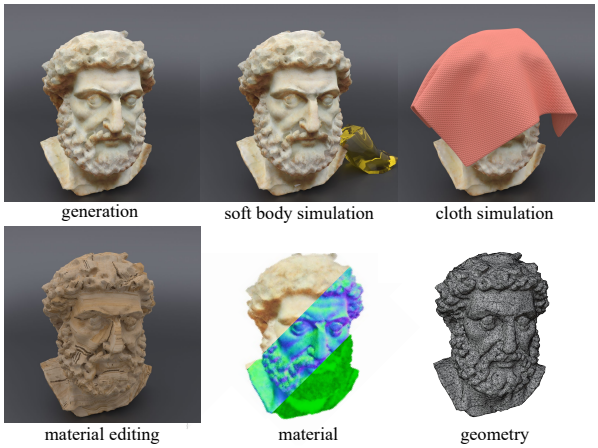

Figure 1. Provided with a textual description of "a highly detailed stone bust of Theodoros Kolokotronis", our method produces high-quality geometry as well as disentangled materials, and enables photorealistic rendering.

In this work, we focus on automatic 3D content creation given text prompts encoded by large language models, i.e., the text-to-3D tasks [18, 33]. Text-to-3D is inspired by the tremendous success of text-to-image research [34, 36, 30, 35]. To enable 3D generation, most existing methods [33, 23] rely on the implicit scene modeling of Neural Radiance Field (NeRF) [25, 4, 27], and learn the NeRFs by back-propagating the supervision signals from image diffusion models. However, NeRF modeling is less effective for surface recovery [43, 46], since it couples the learning of surface geometry with that of pixel colors via volume rendering. Consequently, 3D creation based on

[*]Equal contribution.
[†]Corresponding author.

NeRF modeling is less effective for recovery of both the fine surface and its material and texture. In the meanwhile, explicit and hybrid scene representations [43, 46, 38] are proposed to improve over NeRF by modeling the surface explicitly and performing view synthesis via surface rendering.

In this work, we are motivated to use 3D scene representations that are more amenable to the generation of high-quality 3D assets given text prompts. We present an automatic text-to-3D method called Fantasia3D. Key to Fantasia3D is a disentangled learning of geometry and appearance models, such that both a fine surface and a rich material/texture can be generated. To enable such a disentangled learning, we use the hybrid scene representation of DMTET [38], which maintains a deformable tetrahedral grid and a differentiable mesh extraction layer; deformation can thus be learned through the layer to explicitly control the shape generation. For geometry learning, we technically propose to encode a rendered normal map, and use shape encoding of the normal as the input of a pre-trained, image diffusion model; this is in contrast different from existing methods that commonly encode rendered color images. For appearance modeling, we introduce, for the first time, the spatially varying Bidirectional Reflectance Distribution Function (BRDF) into the text-to-3D task, thus enabling material learning that supports photorealistic rendering of the learned surface. We implement the geometry model and the BRDF appearance model as simple MLPs. Both models are learned through the pre-trained image diffusion model, using a loss of Score Distillation Sampling (SDS) [33]. We use the pre-trained stable diffusion [35, 40] as the image generation model in this work.

We note that except for text prompts, our method can also be triggered with additional inputs of users' preferences, such as a customized 3D shape or a generic 3D shape of a certain object category; this is flexible for users to better control what content is to be generated. In addition, given the disentangled generation of geometry and appearance, it is convenient for our method to support relighting, editing, and physical simulation of the generated 3D assets. We conduct thorough experiments to verify the efficacy of our proposed methods. Results show that our proposed Fantasia3D outperforms existing methods for high-quality and diverse 3D content creation. We summarize our technical contributions as follows.

- We propose a novel method, termed Fantasia3D, for high-quality text-to-3D content creation. Our method disentangles the modeling and learning of geometry and appearance, and thus enables both a fine recovery of geometry and photorealistc rendering of per-view images.
- For geometry learning, we use a hybrid representation of DMTET, which supports learning surface deformation via a differentiable mesh extraction; we propose to render and encode the surface normal extracted from DMTET as the input of the pre-trained image diffusion model, which enables more subtle control of shape generation.
- For appearance modeling, to the best of our knowledge, we are the first to introduce the full BRDF learning into text-to-3D content creation, facilitated by our proposed geometry-appearance disentangled framework. BRDF modeling promises high-quality 3D generation via photorealistic rendering.

2. Related work

Text-to-3D content creation. Motivated by the desire to generate high-quality 3D content from simple semantics such as text prompts, text-to-3D has drawn considerable attention in recent years [33, 13, 18]. Existing methods either use pre-trained 2D text-to-image models [35, 3, 36], together with score distillation sampling [33], to generate 3D geometries [18] or synthesize novel views [33], or train a text-conditioned 3D generative model from scratch [39, 49, 17, 29]. These methods generate 3D geometries with little exploration of generating high-quality lighting and surface materials. On the other hand, TANGO [8] is able to generate high-quality surface materials given text prompts; unfortunately, the method requires as input a 3D surface mesh. Our proposed method addresses the shortcomings of the above methods, and is able to generate both high-quality surface geometries and their corresponding materials, both of which are crucial for photorealistic rendering of the generated 3D content. Our method thus, for the first time, closes the loop of object-level text-to-3D content creation.

Surface material estimation. The estimation of surface materials is a long-standing challenge in computer vision and graphics research. Earlier methods [2, 45] focus on recovering physically based materials under known lighting conditions, whose usefulness is, however, limited in real-world scenarios. Subsequent methods [12, 11, 5, 1] try to estimate materials under natural lighting conditions, assuming the availability of complete geometry information. More recently, the joint reconstruction of geometry and materials is proposed given calibrated multi-view images [6, 7, 19, 48]. Alternative to these methods, we explore the novel creation of surface materials and geometries from trained language models.

3. Preliminary

In this section, we present a few preliminaries that are necessary for presenting our proposed method in Section 4.

Figure 2. **Results of our method.** The upper portion of this figure showcases the generation results obtained from solely text prompts. The lower portion showcases user-guided generation results given guiding meshes with the corresponding textual descriptions.

3.1. Score distillation sampling

DreamFusion [33] presents a method that optimizes 3D scene parameters and synthesizes novel views from textual descriptions, by employing a pre-trained 2D diffusion model. The scene is represented as a differentiable image parameterization [26], where a differentiable generator g

Figure 3. Overview of our method. Our method can generate disentangled geometry and appearance given a text prompt (cf. figure (a)), which are produced by (b) geometry modeling and (c) appearance modeling, respectively. (b) We employ a hybrid 3D representation, DMTet, as our 3D geometry representation, which is initialized as a 3D ellipsoid here. To optimize the parameters of DMTet, we render the normal map (and the object mask in the early training phase) of the extracted mesh from DMTet as the shape encoding of stable diffusion [35, 40]. (c) For appearance modeling, we introduce the spatially-varying Bidirectional Reflectance Distribution Function (BRDF) modeling into text-to-3D generation, and learn to predict three components (namely, k_d, k_{orm}, and k_n) of the appearance. Both geometry and appearance modeling are supervised by Score Distillation Sampling (SDS) loss [33].

renders 2D images $x = g(\theta)$ from a modified Mip-NeRF [4] parameterized as θ. DreamFusion leverages a diffusion model (Imagen [36] in this instance) to provide a score function $\hat{\epsilon}(x_t; y; t)$, which predicts the sampled noise given the noisy image x_t, text-embedding y, and noise level t. This score function guides the direction of the gradient for updating the scene parameters θ, and the gradient is calculated by Score Distillation Sampling (SDS):

$$\nabla_\theta L_{SDS}(\theta; x = g(\theta)) = \mathbb{E}_{t,\epsilon}\left[w(t)(\hat{\epsilon}(x_t; y; t) - \epsilon)\frac{\partial x}{\partial \theta} \right], \quad (1)$$

while $w(t)$ is a weighting function. Since Imagen is not publicly accessible, in this work, we use the released latent space diffusion model of Stable Diffusion [35, 40] as our guidance model, and revise the SDS loss (1) accordingly. Details are given in Section 4.2.

3.2. DMTet

Implicit surface representations [25, 31, 22] are popularly used in novel view synthesis and 3D reconstruction, due to their capabilities to represent complex scenes. However, surfaces of lower quality may be obtained [28] by extracting explicit meshes from these implicit representations using marching cubes [20]. Instead, Shen et al. [38] propose a hybrid representation, termed DMTet, that has two key features, i.e., a deformable tetrahedral grid and a differentiable Marching Tetrahedral (MT) layer. The deformable tetrahedral grid (V_T, T) has vertices V_T in the tetrahedral grid T. For each vertex $v_i \in V_T$, the proposed method predicts the Signed Distance Function (SDF) value $s(v_i)$ and a position offset Δv_i by:

$$(s(v_i), \Delta v_i) = \Phi(v_i; \Phi), \quad (2)$$

where Φ is the parameters of a network, enabling the extraction of explicit meshes through MT layer during each iteration of training. In this work, We use DMTet as our geometry representation and render the mesh extracted from MT layer iteratively by a differentiable renderer [16, 28].

4. The Proposed Method

In this section, we present our proposed method of Fantasia3D for high-quality text-to-3D object generation, by disentangling the modeling and learning of geometry and appearance. For geometry modeling, we rely on the hybrid surface representation of DMTet, and parameterize the 3D geometry as an MLP Φ that learns to predict the SDF value and position offset for each vertex in the deformable tetrahedral grid of DMTet; in contrast to previous methods, we propose to use the rendered normal map (and the object mask in the early training phase) of the extracted mesh from DMTet as the input of shape encoding. For appearance modeling, we introduce, for the first time, the full Bidirectional Reflectance Distribution Function (BRDF) modeling into text-to-3D generation, and learn an MLP Ψ that predicts parameters of surface material and supports high-quality 3D generation via photorealistic rendering. Given the disentangled modeling of Φ and Ψ, the whole pipeline is learned with SDS supervision, and the gradients are back-propagated through the pre-trained stable diffusion model. Our pipeline is initialized either as a 3D ellipsoid or as a customized 3D model provided by users. Fig. 3 gives an illustration of our proposed method. In contrast, previous methods couple the geometry and appearance learning, and are suboptimal in terms of leveraging the powerful pre-trained 2D image diffusion models via SDS loss. Details of the proposed Fantasia3D are presented as follows.

4.1. DMTET initialization

We adopt DMTET as our 3D scene representation, which is parameterized as the MLP network ϕ. For each vertex $v_i \in V_T$ of the tetrahedral grid (V_T, T), ϕ is trained to predict the SDF value $s(v_i)$ and the deformation offset Δv_i. A triangular mesh can be extracted from (V_T, T) using the MT layer, whose procedure is also differentiable w.r.t. the parameters of ϕ. We initialize DMTET either as a 3D ellipsoid or as a customized 3D model provided by users; the latter choice is useful when the text-to-3D task is to be conditioned on users' preferences. In either case, we initialize ϕ by the following fitting procedure. We sample a point set $\{p_i \in \mathbb{R}^3\}$ whose points are in close proximity to the initialized 3D ellipsoid or customized model, and compute their SDF values, resulting in $\{\mathrm{SDF}(p_i)\}$; we use the following loss to optimize the parameters of ϕ:

$$L_{\mathrm{SDF}} = \sum_{p_i \in P} \| s(p_i; \phi) - \mathrm{SDF}(p_i) \|_2^2. \quad (3)$$

4.2. Geometry modeling

Previous text-to-3D methods [33, 13] commonly use NeRF [25, 4] as the implicit scene representation, which couples the geometry with color/appearance and use volume rendering for view synthesis. Since NeRF modeling is less effective for surface reconstruction [43, 46], these methods are consequently less effective to generate high-quality 3D surfaces by back-propagating the supervision of SDS loss through pre-trained text-to-image models. As a remedy, the method [18] uses a second stage of the refined generation that is based on scene modeling of explicit surfaces.

In this work, we propose to decouple the generation of geometry from that of appearance, based on the hybrid scene representation of DMTET, which enables photorealistic surface rendering to make better use of the powerful pre-trained text-to-image models. More specifically, given the current DMTET with MLP parameters ϕ, we generate a normal map n, together with an object mask o, as:

$$(n, o) = g_n(\phi, c); \quad (4)$$

where g_n is a differentiable renderer (e.g., nvdiffrast [16]), and c is a sampled camera pose. We randomly sample the camera poses in the spherical coordinate system to ensure that the camera poses are distributed uniformly on the sphere. We propose to use the generated (n, o) as the input of shape encoding to connect with stable diffusion. To update ϕ, we again employ SDS loss that computes the gradient w.r.t. ϕ as:

$$\nabla_\phi L_{\mathrm{SDS}}(\phi, n) = \mathbb{E}_{t,\epsilon}\left[w(t)(\hat{\epsilon}_\phi(z_t^n; y; t) - \epsilon) \frac{\partial n}{\partial \phi} \frac{\partial z_t^n}{\partial n} \right], \quad (5)$$

Figure 4. Three components of the material model, namely the diffuse term k_d, the roughness and metallic term k_m, and the normal variation term k_n.

Figure 5. A comparison of UV edge padding (cf. the left column) and original texturing (cf. the right column). UV edge padding removes the white seams that appear in the right rendering.

where ϕ parameterizes the pre-trained stable diffusion model, n^- denotes concatenation of the normal n with the mask o, $z_t^{n^-}$ is the latent code of n^- via shape encoding, $\hat{\epsilon}_\phi(z_t^{n^-}; y; t)$ is the predicted noise given text embedding y and noise level t, and ϵ is the noise added in $z_t^{n^-}$. In practice, we utilize a coarse-to-fine strategy to model the geometry. During the early phase of training, we use the downsampled n^- as the latent code, which is inspired by [23], to rapidly update ϕ and attain a coarse shape. However, a domain gap exists between n^- and the latent space data distribution learned by the VAE encoder in the stable diffusion, which may lead to a mismatch of the generated geometry from the textual description. To mitigate this discrepancy, we implement a data augmentation technique by introducing random rotations to n^-. Our experimental observations reveal that this technique enhances the alignment between the generated geometry and the provided textual description. In the later phase of training, aiming to capture finer geometric details with greater precision, we encode the high-resolution normal n (without the mask o) to derive z^n, using the pre-trained image encoder in stable diffusion.