Learning Pseudo 3D Guidance for View-consistent Texturing with 2D Diffusion

Kehan Li[1,4⋆], Yanbo Fan[2,6⋆⋆], Yang Wu[2], Zhongqian Sun[2], Wei Yang[2]
Xiangyang Ji[5], Li Yuan[1,3,4], and Jie Chen[1,3,4⋆⋆]

[1] School of Electronic and Computer Engineering, Peking University, Shenzhen, China
[2] Tencent AI Lab, Shenzhen, China
[3] Peng Cheng Laboratory, Shenzhen, China
[4] AI for Science (AI4S)-Preferred Program, Peking University, Shenzhen, China
[5] Department of Automation and BNRist, Tsinghua University, Beijing, China
[6] Ant Group, Hangzhou, China

Abstract. Text-driven 3D texturing requires the generation of high-fidelity texture that conforms to given geometry and description. Recently, the high-quality text-to-image generation ability of 2D diffusion model has significantly promoted this task, by converting it into a texture optimization process guided by multi-view synthesized images, where the generation of high-quality and multi-view consistency images becomes the key issue. State-of-the-art methods achieve the consistency between different views by treating image generation on a novel view as image inpainting conditioned on the texture generated by previously views. However, due to the accumulated semantic divergence of local inpainting and the occlusion between object parts on sparse views, these inpainting-based methods often fail to deal with long-range texture consistency. To address these, we present **P3G**, a texturing approach based on learned **P**seudo **3**D **G**uidance. The key idea of P3G is to first learn a coarse but consistent texture, to serve as a global semantics guidance for encouraging the consistency between images generated on different views. To this end, we incorporate pre-trained text-to-image diffusion models and multi-view optimization to achieve propagating accurate semantics globally for leaning the guidance, and design an efficient framework for high-quality and multi-view consistent image generation that integrates the learned semantic guidance. Quantitative and qualitative evaluation on variant 3D shapes demonstrates the superiority of our P3G on both consistency and overall visual quality.

Keywords: Texture synthesis · 3D synthesis · Diffusion model

1 Introduction

High-quality 3D assets are crucial for applications such as virtual reality, gaming, and the movie industry. As a promising direction to greatly improve efficiency,

⋆ Work done during an internship at Tencent.
⋆⋆ Corresponding author. Project page: https://lkhl.github.io/P3G

Fig. 1: We focus on the text-to-texture generation task. Given a text description and 3D mesh, our method can generate high-quality and view-consistent 3D textures.

automatic 3D content generation has aroused great interest in computer graphics and computer vision. In this work, we focus on text-driven texturing of 3D meshes, which aims to generate high-quality texture of 3D meshes matching the given geometry and text description, as shown in Fig. 1.

Due to the lack of large-scale datasets of high-quality 3D assets and the corresponding text descriptions, most of the existing methods are built on large-scale visual-language models, such as CLIP [25] and text-to-image diffusion models [27], for realizing text-driven texture generation. Among them, some works [18, 20] optimize the texture by maximizing the CLIP matching scores of the rendered 2D images and the input text. While the CLIP scores compare the high-level semantic consistency between text and images, their generated texture lacks visual details [26]. Recently, with the help of photorealistic text-to-image generation ability of diffusion models [9], the visual quality of 3D texturing has been significantly promoted by optimizing texture using multi-view images generated by pre-trained 2D diffusion models [26]. Although the quality of the image generated from a single view is impeccable, the consistency of images between different views is difficult to guarantee due to the natural randomness of the generation process, which significantly hurts the fidelity of the final texture. To tackle this problem, TEXTure [26] and Text2Tex [3] proposed to take previously generated texture into account when generating the image of a novel view through inpainting. In this way, since the semantics are only propagated locally, the semantic shift caused by the randomness of the inpainting operation are accumulates as the number of views grows. At the same time, the inevitable occlusion between object parts on the selected sparse views weakens the effect of propagating semantics. The limitations of these inpainting-based methods finally lead to unsatisfactory long-range consistency of the whole object.

In this work, for the purpose of encouraging global consistency on individual views, we propose to guide the 2D image generation process by globally consistent semantics, which are carried by an initial texture, as shown in Fig. 2. As for guidance acquisition, since the perfect guidance known as ground truth is unavailable, we instead seek a pseudo one through accessible data or models. To precisely align the global semantics with the input geometry and text description, we build the initial texture from its 2D renderings with the help of

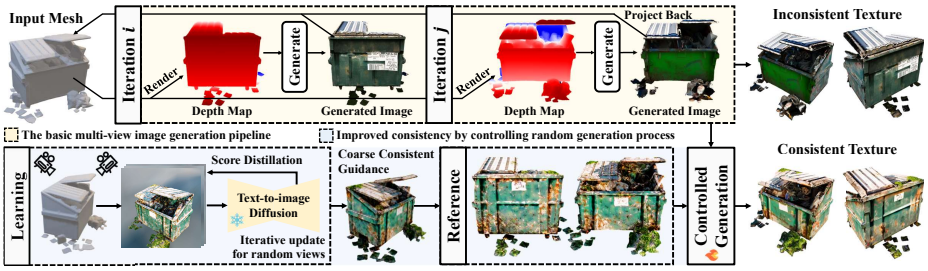

Fig. 2: The mainstream solutions (the first row) are based on generating images from multiple views with text-to-image diffusion models, where the consistentcy between views are the core issue. We propose to first learn a coarse but consistent texture to serve as a global semantics guidance for encouraging the consistency (the second row).

a depth-guided text-to-image diffusion model. Under this premise, we facilitate the global consistency by building a globally related texture model and gradually learning the texture on continuous views from random initialization, which synchronizes the semantic updates bit by bit locally and finally achieve global semantic association. Although this optimization process destroys the generation quality of diffusion model, the resulting coarse texture is sufficient as a guide to represent global semantics that match the input. For integrating the guidance, we control the image generation process by truncating the denoising process at a time step in the late stage and manually set the foreground part of the output image to be the same as the guidance texture, considering that the early and middle stages of the denoising process mainly focus on layout and semantics. Moreover, the guidance of global semantics empowers the deprecation of substantial overlap between views, enabling a more efficient and effective view selection strategy that encourages the generation of large and complete areas.

With above framework, the visual quality is guaranteed by the powerful text-to-image diffusion model, while the randomness of each generation is controlled through the pseudo 3D guidance and the inpainting operation, thus significantly improving the global consistence. We conduct extensive comparisons on variant 3D shapes with previous methods to demonstrate the effectiveness of our P3G. Quantitative, qualitative evaluation and user study shows that our P3G improves the consistency while maintaining the powerful generation ability of 2D diffusion model. In summary, our contributions include:

- We point out the long-range inconsistency of inpainting-based texturing pipelines with 2D diffusion, and propose to guide the 2D image generation on individual views by globally consistent semantics.
- We implement the generation pipeline with view selection and semantic guidance which hacks the late denoising process with an coarse but consistent texture which is learned using 2D diffusion models.
- We evaluate the generated texture quantitatively and qualitatively in terms of both visual quality and global consistency, demonstrating the effectiveness of the proposed pseudo 3D guidance.

2 Related Work

Text-to-texture Generation. Compared to 2D image generation, text-to-texture generation of 3D shapes is much more complicated, and requires attention to both shapes and text description. While early works adopt probabilistic models or study geometric texture synthesis for some specific categories [1, 7], recent advances explore data-driven approaches for zero-shot text-driven texturing of 3D shapes. Yet, unlike the massive text-to-image datasets, high-fidelity 3D data is relatively scarce. This has inspired several works to explore 3D texture generation with pre-trained 2D text-to-image models. For instance, CLIP-Mesh [20] and Text2Mesh [18] utilize CLIP matching scores as criteria for texture optimization. Yet the CLIP scores compare the high-level semantic consistency between text and images, their generated texture is of low quality and lacks details [26]. For better visual quality, recent works explore pre-trained text-to-image diffusion models for 3D texture generation [3, 17, 26]. Latent-paint [17] proposed to optimize the 3D latent texture with score distillation sampling (SDS) [24] with a pre-trained 2D text-to-image diffusion model, but suffering from low visual quality due to the defects of SDS and latent texture decoding. The pioneering work of TEXTure [26] projects the high-quality 2D images generated on different views by the 2D diffusion model [27] back to the mesh. However, due to the stochastic nature of the generation process and the inevitable occlusions between object parts on sparse views, this iterative inpainting strategy suffers from long-range view inconsistency, which is the main focus of our work.

Another branch of texture generation methods are based on training a generative model using specific 3D datasets [23, 28, 36]. Due to the limited 3D data and the difficulty of 3D texture representation, these methods can only be applied to specific classes, thus lossing the ability to match input text. And the generated texture is relatively simple due to the quality of the dataset.

Text-to-3D Generation. The task of text-to-texture generation is also related to text-to-3D generation, which requires to generate both 3D geometry and texture with the given text. With the development of large-scale vision-language models, the leading text-to-3D methods are based on distilling knowledge from pre-trained vision-language models. For example, Dream Fields [10] proposes to optimize Neural Radiance Fields (NeRF) [19] with CLIP [25] loss. The breakthrough of text-to-image diffusion models [22,27] further compensates the unsatisfactory quality caused by the insufficient generation ability of CLIP. DreamFusion [24] first proposes Score Distillation Sampling (SDS), which optimizes the parameters of a NeRF through probability density distillation on its multi-view renderings with text-to-image diffusion models as prior, and demonstrates the promise of creating high-quality 3D assets with text guidance. Following it, subsequent methods improve the quality by improving the representation and optimization strategy [4, 5, 14, 17, 30] and eliminating the over-saturation effect of SDS [13, 32, 37]. Although these methods these methods can handle the text-to-texture to some extent, the fixed geometry is not exploited and constrained in them, resulting in unsatisfactory quality and geometry matching.

3 Method

3.1 Preliminary

Latent Diffusion Model (LDM). Diffusion model [9] and its variants have achieved state-of-the-art text-guided image generation performance. Let $\boldsymbol{x} \in R^{H\times W\times C}$ be the input image and $\mathcal{P}(\boldsymbol{x})$ be the unknown real data distribution. Diffusion models are latent variable models that learn $\mathcal{P}(\boldsymbol{x})$ by gradually denoising a Gaussian distributed variable. There are *forward process* and *reverse process* in general diffusion models. The *forward process* is defined as a Markov chain that gradually adds random noise to the $\boldsymbol{x}$ towards a Gaussian noise, as follows:

$$\boldsymbol{x}_t = \sqrt{\alpha_t}\boldsymbol{x} + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}, \tag{1}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(0,1)$ and $\alpha_t > 0$ is a pre-defined variance schedule. While the *reverse process* starts from a random sampled Gaussian noise and conducts iterative denoising towards a real data point. The training objective of diffusion models is to learn the denoising function *w.r.t.* the reverse process. LDM [27] proposes to perform diffusion process in the latent space, rather than the high-dimensional pixel space. To this end, it first learns an encoder $\mathcal{E}$ that encodes $\boldsymbol{x}$ into a low-dimensional latent representation $\boldsymbol{x}^{lat} = \mathcal{E}(\boldsymbol{x})$, and an decoder $\mathcal{D}$ that reconstructs $\boldsymbol{x}$ from the latent $\boldsymbol{x}^{lat}$, *i.e.*, $\tilde{\boldsymbol{x}} = \mathcal{D}(\mathcal{E}(\boldsymbol{x}))$. The training objective of LDM is

$$\mathcal{L}_{LDM} = \mathbf{E}_{\mathcal{E}(\boldsymbol{x}),y,\boldsymbol{\epsilon}\sim\mathcal{N}(0,1),t}[||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_{\boldsymbol{\phi}}(\boldsymbol{x}_t^{lat}, t, \boldsymbol{c})||_2^2], \tag{2}$$

where $\boldsymbol{\epsilon}_{\boldsymbol{\phi}}$ represents a neural network parameterized by $\boldsymbol{\phi}$ that predicts the added noise at time t, $\boldsymbol{c}$ is an optional conditional information, *i.e.*, the text prompt in this work. After training, one can generate real images by performing the reverse process with the text guidance.

Texturing as Image Generation. Considering the fixed geometry, state-of-the-art methods [26] treat texture generation as multi-view depth-conditioned image generation, and achieve considerable visual quality with the help of the photorealistic image generation ability of text-to-image diffusion models. Specifically, the texturing process is performed in an iterative manner under this basic framework. In each iteration, a viewpoint that overlaps with previous ones is selected and the corresponding image and depth map of the given object is rendered. Then the part to be generated is determined by some rules (*e.g.* the part that has not been seen yet), which is indicated by a 2D mask. Based on the mask, a new image is generated by inpainting the rendered image for harmony, and the geometry matching is implemented by conditioned on the depth map. Finally, the generated image is projected back to the texture by inverse rendering.

Score Distillation Sampling (SDS). Based on the 2D generation quality of text-to-image diffusion models, SDS is a technique to distill the knowledge of pre-trained text-to-image diffusion models to generate high-quality 3D objects,

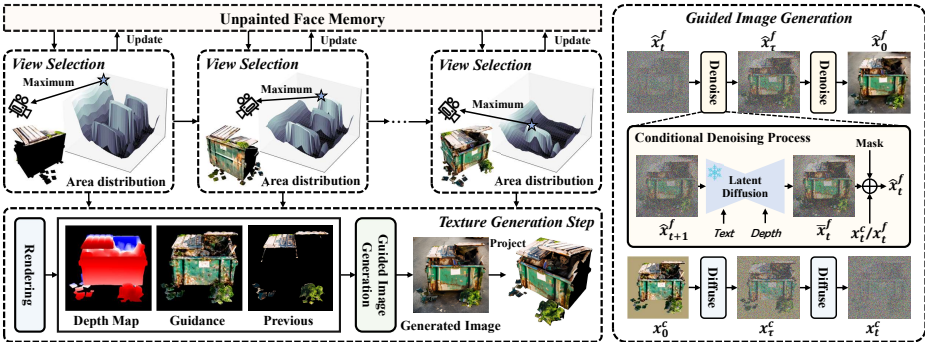

Fig. 3: The overall texture generation pipeline. We build our texturing pipeline upon the multi-view image generation paradigm with adaptively selected viewpoint that encourages the generation of large and complete areas, and improving the consistency between different views by the semantic guidance.

which has gradually become the mainstream in the fields of 3D shape generation, texturing, and editing. Let $g_{\boldsymbol{\theta}}$ be an underlying 3D model with parameters $\boldsymbol{\theta}$ that needs to be generated according to the given text description and $\boldsymbol{x} = g(v)$ be an image rendered by g under a certain view v. For SDS optimization, a random noise is first added to $\boldsymbol{x}$ at time step t by

$$\boldsymbol{x}_t = \sqrt{\alpha_t}\boldsymbol{x} + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}. \tag{3}$$

Then, the per-pixel gradient *w.r.t.* $\boldsymbol{x}$ is calculated as

$$\nabla_{\boldsymbol{x}}\mathcal{L}_{SDS} = w(t)(\boldsymbol{\epsilon}_{\boldsymbol{\phi}}(\boldsymbol{x}_t; t, \boldsymbol{c}) - \boldsymbol{\epsilon}), \tag{4}$$

where $w(t)$ is a constant weighting factor that depends on α_t, $\boldsymbol{c}$ is the text condition, and $\boldsymbol{\epsilon}_{\boldsymbol{\phi}}$ is a pre-trained diffusion model. The gradient *w.r.t.* the parameters of model g can be calculated via the chain rule as

$$\nabla_{\boldsymbol{\theta}}\mathcal{L}_{SDS} = \nabla\boldsymbol{x}\mathcal{L}_{SDS} \cdot \frac{\partial \boldsymbol{x}}{\partial \boldsymbol{\theta}}. \tag{5}$$

3.2 Texturing Pipeline with Semantic Guidance

Overall Pipeline. Due to the high visual quality, we build our texturing pipeline upon the multi-view image generation paradigm as described in Sec. 3.1. For implementation, we adopt a depth-conditioned diffusion model $\mathcal{F}_{depth}$ for image generation. In the image generation process, to avoid the potential artifacts at the seam between the area to be generated and the area that has been generated, we first consider the inpainting operation, where the area to be generated in the i-th view is defined as by pixels that satisfy

$$\cos\langle \boldsymbol{n}, \boldsymbol{s}_i\rangle > \max_p \cos\langle \boldsymbol{n}, \boldsymbol{s}_p\rangle + 0.2,\ p \in \{1, 2, \ldots, i-1\}, \tag{6}$$

where $\boldsymbol{n}$ is the normal of the corresponding face and $\boldsymbol{s}_i$ is the sight direction of this view. This strategy encourages each of the texture to be updated at the best possible view where the corresponding face is parallel to the imaging plane, reducing the effects of distortion when projecting 3D faces to 2D.

Let the region to be generated be indicated by a binary mask $\boldsymbol{m}_{inp}$, inpainting is achieved by setting the unmasked region to the reference image obtained from previously learned texture at the end of each denoising step [15]. Formally, starting from a Gaussian noise $\hat{\boldsymbol{x}}_T^f$, where $T = 1000$ is the maximum time step specified by the pretrained diffusion model, the modified denoising step is

$$\hat{\boldsymbol{x}}_{t-1}^f = \boldsymbol{m}_{inp} \odot \mathcal{F}_{depth}(\hat{\boldsymbol{x}}_t^f; t, \boldsymbol{c}, \boldsymbol{d}) + (1 - \boldsymbol{m}_{inp}) \odot \boldsymbol{x}_{t-1}^f, t > \tau, \tag{7}$$

where $\hat{\boldsymbol{x}}_t^f$ is the predicted image at time step t. Considering that this general pretrained diffusion-based inpainting strategy performs poor for distant areas [26], we further introduce an additional inpainting-specific denoiser $\mathcal{F}_{inp}$ to enhance the inpainting effect. Then the denoising step becomes

$$\hat{\boldsymbol{x}}_{t-1}^f = \boldsymbol{m}_{inp} \odot \tilde{\boldsymbol{x}}_{t-1}^f + (1 - \boldsymbol{m}_{inp}) \odot \boldsymbol{x}_{t-1}^f, t > \tau, \tag{8}$$

where

$$\tilde{\boldsymbol{x}}_{t-1}^f = \begin{cases} \mathcal{F}_{inp}(\hat{\boldsymbol{x}}_t^f; t, \boldsymbol{c}, \boldsymbol{m}_{inp}), & t > \tau \wedge t \bmod 2 = 1, \\ \mathcal{F}_{depth}(\hat{\boldsymbol{x}}_t^f; t, \boldsymbol{c}, \boldsymbol{d}), & otherwise. \end{cases} \tag{9}$$

These two inpainting operators complement each other as the $\mathcal{F}_{inp}$ ignores the depth guidance and may generate geometrically inconsistent textures.

Semantic Guidance. Within the basic pipeline mentioned above, the long-range consistency of the generated texture is still hard to achieve due to the accumulated semantic divergence of local inpainting operation and the occlusion between object parts on sparse views. For encouraging global consistency on individual views, we propose to guide the image generation process by globally consistent semantics, which are carried by a guidance texture $\mathcal{H}$ (Sec. 3.3).

We dive into the intermediate results in the generation process to incorporate the guidance with an off-the-shelf diffusion model. Specifically, based on the phenomena that the early and middle stages of the denoising process focus on layout and semantics [16,35], we hijack the denoising process at time step τ of the late stage by setting the foreground region to $\boldsymbol{x}_\tau^c$, which is obtained by diffusing the guidance image $\boldsymbol{x}^c = \mathcal{R}(\mathcal{H}, \boldsymbol{v}_i)$, where $\mathcal{R}$ is the renderer, $\mathcal{H}$ is the guidance texture with global consistent semantic, and $\boldsymbol{v}_i$ is the view transformation. In this way, Eq. (8) becomes

$$\hat{\boldsymbol{x}}_{t-1}^f = \begin{cases} \boldsymbol{m}_{obj} \odot \boldsymbol{x}_{t-1}^c + (1 - \boldsymbol{m}_{obj}) \odot \tilde{\boldsymbol{x}}_{t-1}^f, & t \leq \tau, \\ \boldsymbol{m}_{inp} \odot \tilde{\boldsymbol{x}}_{t-1}^f + (1 - \boldsymbol{m}_{inp}) \odot \boldsymbol{x}_{t-1}^f, & t > \tau, \end{cases} \tag{10}$$

where $\boldsymbol{m}_{obj}$ is the foreground mask, $\boldsymbol{x}_t^c$ and $\boldsymbol{x}_t^f$ are obtained by adding noise to image $\boldsymbol{x}^c$ rendered from the initial texture and image $\boldsymbol{x}^f$ rendered from already generated texture, respectively. For time step $t < \tau$, we fix the foreground object region to the coarse texture and create a background matching the guidance.

View selection strategy. Due to the presence of our pseudo 3D guidance, it is not necessary to select sequential views like TEXTure [26] in order to take as much the generated textures into account for consistency, which allows for a more flexible and efficient design. Our design principle is that each perspective covers as large an area of ungenerated part as possible to improve efficiency and avoid artifacts caused by multiple inpainting operations (as illustrated in Fig. 4), while maintaining the requirement that every part of the texture is generated at a relatively good view. Therefore, we first define what a relatively good view is, and then dynamically select a view that covers as many textures to be generated as possible by estimating the distribution of faces that can be updated well under different views.

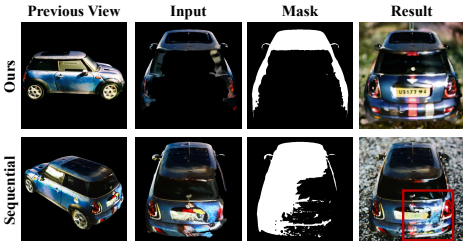

Fig. 4: Our strategy encourages large continuous area, reducing potential artifacts from inpainting.

Specifically, we restrict the camera to be on a sphere with a fixed radius centered on the target, and directed towards the center of the sphere. Then the view is determined by two parameters: the azimuth angle α and the elevation angle β. For each view i, we define its covered texture area as

$$A_i = \sum_{f \in \mathbb{F}_k} a_f, \ \mathbb{F}_k = \{f | f \in \mathbb{U}_j \wedge \cos\langle \boldsymbol{n}_f, \boldsymbol{s}_i \rangle > \delta\}, \tag{11}$$

where a_f is the area of the f-th face in the 3D mesh, $\mathbb{U}_k$ is the unpainted faces at iteration k, and δ is a threshold to limit on faces with smaller distortion. For the k-th texture generation iteration, we select a view by $\arg\max_i A_i$, and update $\mathbb{U}$ by subtracting the updated faces.

To summarize, the texture generation is performed in an iterative manner as shown in Fig. 3. Initializing $\mathbb{U}$ to all faces of the 3D mesh, we iteratively select view (α, β), update the unpainted face set $\mathbb{U}$, generate a high-quality image from this view conditioned on the geometry, the texture from the coarse stage, and the previously generated texture, then project the image to the texture. The iteration terminates when a preset number of times is reached, or when there is very little content to be generated on a view. In this process, the view selection strategy reduces the number of views required and encourages contiguous areas to be updated. Meanwhile, the view-consistent image generation module helps to ensure global consistency via the conditional generation strategy.

3.3 Pseudo 3D Guidance Learning

In this section, we introduce how to obtain the guidance texture $\mathcal{H}$ in detail. An ideal guidance should be a realistic texture for a given geometry, which is unavailable in the generation task. We instead seek a pseudo one through accessible data or models. Since the role of guidance is mainly to carry globally

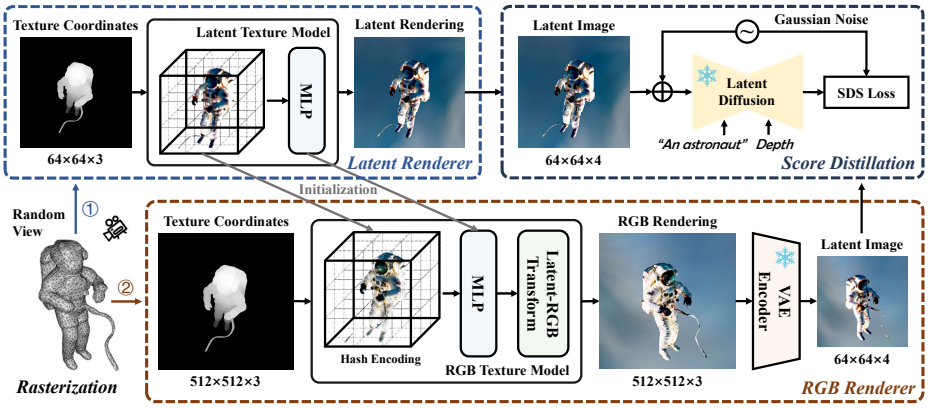

Fig. 5: The detailed pipeline of the pseudo 3D guidance learning. For encouraging consistency, the texture model is updated bit by bit from continuous views, with the depth-guided and latent-to-RGB strategy to improve the quality.

consistent semantics, the first thing is to clarify what the semantics are. On the text-to-texture generation task, the semantics is defined by the input geometry and text. Therefore, to precisely align the global semantics with the input geometry and text description, we build the initial texture from its 2D renderings with the help of a depth-guided text-to-image diffusion model. Under this premise, we facilitate the global consistency by building a globally related texture model, and gradually learning the texture on continuous views from random initialization, which is implemented with score distillation [24] guided by the depth map. In the optimization process, the semantic updates are synchronized between continuous views bit by bit locally, finally achieving global semantic association. Although relatively low visual quality with SDS, the resulting coarse texture is sufficient as a guide to represent global semantics that match the input.

Detailed Pipeline. The overall pipeline for producing pseudo 3D guidance is given in Fig. 5. We first model the texture $\mathcal{H}_{\boldsymbol{\theta}}$ as coordinate-to-color mapping in the 3D space, which is implemented by Instant-NGP [21] with $\boldsymbol{\theta}$ as the parameters. For rendering, we first obtain the depth and 3D coordinate of each pixel in the 2D image using differentiable rasterization [12], and use the texture model to transform the coordinates to colored image.

Starting from random features, the texture is updated iteratively from random viewpoints by score distillation sampling (SDS) [24], based on the ability of the diffusion model to update random samples toward high probability areas. Considering two key issues of texture generation, namely the geometric matching and the sharpness, we employ a depth-conditioned diffusion model [29] and adopt latent-to-RGB optimization. Specifically, in each iteration, we first render an image $\boldsymbol{x}$ and the corresponding depth map from a randomly sampled viewpoint $\boldsymbol{v}$ through a differentiable renderer $\mathcal{R}$. The texture model is updated by

Fig. 6: Multi-style, text-guided, and high-quality texture generated by our method.

propagating the SDS gradient *w.r.t.* the image to $\boldsymbol{\theta}$:

$$\nabla_{\boldsymbol{\theta}}\mathcal{L}_{SDS} = \nabla_{\boldsymbol{x}}\mathcal{L}_{SDS} \cdot \frac{\partial \boldsymbol{x}}{\partial \boldsymbol{\theta}}, \quad \boldsymbol{\theta} = \boldsymbol{\theta} - \gamma \cdot \nabla_{\boldsymbol{\theta}}\mathcal{L}_{SDS}, \tag{12}$$

where $\partial \boldsymbol{x}/\partial \boldsymbol{\theta}$ is calculated through the differentiable renderer and γ is the learning rate. As we use LDM as the denoiser which incorporates a variational autoencoder (VAE) to project an RGB image $\boldsymbol{x}^{rgb} \in \mathbb{R}^{h\times w\times 3}$ to a latent image $\boldsymbol{x}^{lat} \in \mathbb{R}^{\frac{h}{8}\times\frac{w}{8}\times 4}$ and processes in the latent space. Considering this property, we conduct SDS in the latent space in the early stage of optimization for fast converging, and improve the sharpness of the texture as much as possible to provide a good guidance by bringing the SDS to the high-resolution RGB space. To this end, we convert the latent texture model to an RGB one by applying a point-wise color projection [17] which transforms the four-dimensional latent color to three-dimensional RGB color, and render high-resolution RGB images from the converted texture model.

4 Experiment

4.1 Implementation Details

Texture Pipeline. We use the same depth-conditioned diffusion model as in the coarse stage, and an extra inpainting model SD-v2-inpainting. The threshold δ, the initial time step τ, and the maximum number of views are set to $\cos 30°$, 500, and 10, respectively. For rendering, the texture is represented as an atlas through UV mapping calculated by Xatlas [34], and kaolin [8] is used as the renderer for its adaptability to texture atlas. To do that, we first convert the coarse texture model to a texture atlas by rasterization and render x^c from it.

Pseudo 3D Guidance Learning. We use the depth-conditioned Stable Diffusion [27] (SD-v2-depth). For both latent space and RGB space, the optimization is performed by mini-batch with a batch size of 4 for 500 iterations using the

Table 1: Quantitative evaluation results. We evaluate state-of-the-art methods from the text matching (CLIP Score), the overall quality (FID, KID, and CLIP-IQA), and consistency (CLIP Variance). The results demonstrate the leading quality and consistency of our P3G with comparable text matching.

Method	FID ↓	KID ↓	CLIP Score ↑	CLIP IQA ↑	CLIP Variance ↓
TEXTure [26] SIGGRAPH'23	63.7	23.7	**25.66**	40.19	9.28
Text2Tex [3] ICCV'23	59.1	21.2	24.07	35.48	8.82
Baseline (*w/o* Guidance)	60.1	22.6	25.07	38.14	10.20
P3G (*Ours*)	**58.0**	**20.3**	25.53	**46.17**	**8.13**

Table 2: User study results show the superiority in both consistency and quality.

Method	Inconsistency ↓	Overall Quality ↑
TEXTure [26] SIGGRAPH'23	23.63	36.23
P3G (*Ours*)	**15.55**	**42.45**

Adam [11] optimizer with a learning rate of 0.01. We adopt nvdiffrast [12] for rendering due to its efficiency. For the latent space optimization, we set the rendering resolution to 64×64 and the time step t is sampled from $[20, 980]$. For the RGB space optimization, the rendering resolution is set to 512×512 and the time step t is sampled from a smaller range $[20, 500]$. We add a linear transformation as [17] on the hash encoding and MLP optimized in latent space and use sigmoid activation to limit the RGB value range.

4.2 Main Results

Quantitative Evaluation. Following Text2Tex [3], we select 410 meshes from Objaverse [6] dataset with ground-truth texture for evaluation. The metrics are calculated by the multi-view renderings of the generated texture. Specifically, we set the elevation angel, the distance, and the FOV of the perspective camera to 60°, 2.5 and 60°, respectively, and evenly change the azimuth angle from 0° to 360° to get 10 different views for rendering the texture.

For the metrics, considering the requirements of the text-guided texture synthesis task, we compare our method with previous methods in three dimensions: ❶ *Text matching*. We first evaluate how well the generated texture matches the text input, using the average CLIP score [25] across multiple views. ❷ *Quality*. We evaluate our method on a subset of meshes on the Objaverse [6] dataset with ground-truth texture by FID and KID score following [3], to show the distribution similarity with the ground truth. In addition, we exploit an image-only perception metric CLIP-IQA [31] for assessing the quality, and calculate it from multiple views. ❸ *Consistency*. Since there is no previous work evaluating the multi-view consistency of 3D objects, we develop a metric called CLIP variance

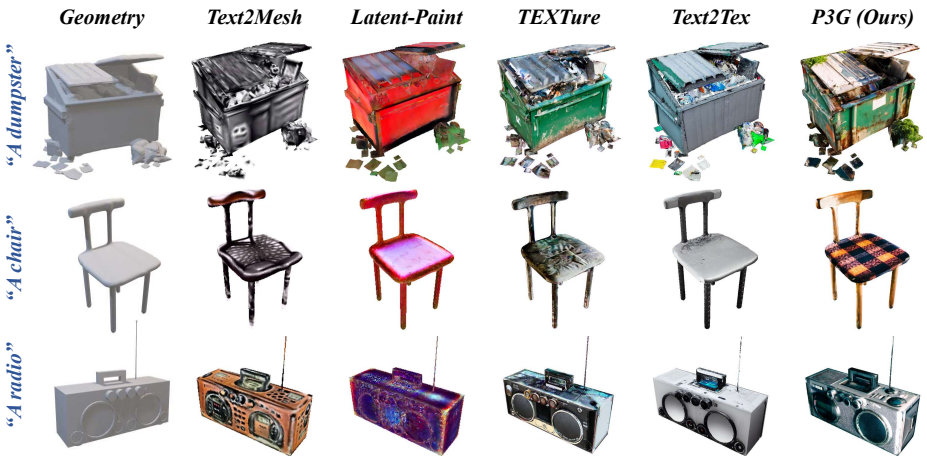

Fig. 7: Visual quality comparisons of text-driven 3D texturing. For each row, the leftmost plot shows the input text and 3D mesh. We present rendering images under the same view for different methods.

based on the idea that images of the same object from multiple views have the same semantics. Specifically, we use the CLIP ViT to extract features for multi-view rendering due to its ability to represent various semantics, and take $(1 - cos\langle \boldsymbol{f}_i, \boldsymbol{f}_j \rangle) * 100$ between these features as the metric.

Tab. 1 shows the comparison with previious state-of-the-art methods including TEXTure [26] and Text2Tex [3]. For text matching, our method get a comparable CLIP score with TEXTure, which demonstrates that the consistency guidance introduced in the multi-view image generation process does not hurt the text-to-image generation ability of 2D diffusion model. Moreover, our P3G significantly outperforms previous methods on CLIP-IQA. This is attributed to the consistency between different view that avoids obvious artifacts on the generated 3D texture. As for the consistency, the texture generated by Text2Tex is relatively simple, resulting in fair consistency score but in the cost of low overall quality, which is reflected in significantly lower CLIP-IQA. Overall, our P3G surpasses the counterpart TEXTure both in overall quality and consistency. It is also worth mentioning that Text2Tex achieves considerable consistency with P3G, because the textures it generates are often simple in color and detail, as shown by the CLIP-IQA in Tab. 1.

User Study. We collect about 50 multi-category meshes from ModelNet [33], ShapeNet [2], and open source projects [18, 26] for user study. We compare our method with the inpainting-based approach TEXTure here and other methods are omitted because the texture they generate are relatively simple. First, the participants are asked to answer whether there is obvious inconsistency in the generated texture. If the answer is yes, then the inconsistency index is increased by one. Then the user need to score the generated texture 0 or 1 based on their preference, and the scores are added up as the quality indicator. Finally

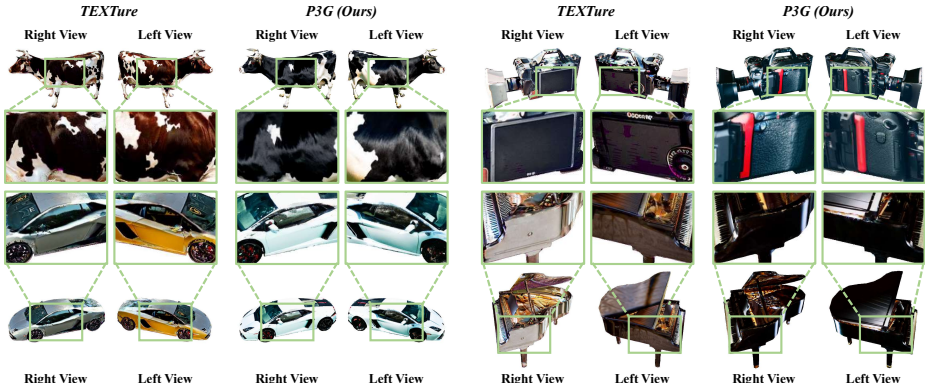

Fig. 8: Consistency comparison. Previous inpainting-based methods fail to ensure long-range consistency. Our P3G resolves it by introducing the guidance.

we calculate the average of these two indicators, and the results are shown in Tab. 2. Based on user feedback, our P3G is significantly better than TEXTure in both consistency and overall quality.

Qualitative Evaluation. Fig. 6 and Fig. 7 show some examples of the generated texture. Thanks to the generation ability of 2D diffusion model and the proposed pseudo 3D guidance for encouraging the consistency, our P3G can generate view-consistent and highly-detailed texture which also matches well with the input geometry and text. Fig. 8 shows some cases that previous inpainting-based methods fail to resolve. With our P3G, the generation process is controlled thus producing consistent results in different part of the object, which is especially helpful for cylindrical objects that require extremely high consistency. Please refer to the appendix for more qualitative results.

4.3 Ablation Study

Effectiveness of Pseudo 3D Guidance. For verifying the role of the pseudo 3D guidance on consistency, we introduce a baseline which dose not incorporate the guidance when generating image and retain the inpainting operation to maintain local consistency. The results in Tab. 1 demonstrate that our guidance successfully controls the randomness of generation thus producing more consistent texture. Please refer to the appendix for more qualitative results.

View Selection Strategy. In Fig. 9 demonstrates the efficiency advantages of our view selection strategy compared to sequential view selection, where the abscissa is the number of views passed, and the ordinate is the average cosine of the angle between the corresponding face and the sight when it is updated, for each pixel on the texture map. A larger value on the vertical axis indicates that a larger area of texture is updated at a better view. The results show that compared to fixed sequential perspectives, our method quickly finds ungenerated areas, and

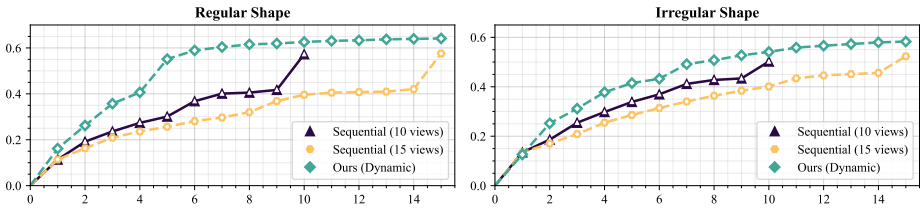

Fig. 9: Effective of the view selection strategy. The change in generated area over the number of iterations demonstrates the efficiency advantages and flexibility of our view selection strategy compared to sequential view selection.

finds the most appropriate perspective for each area as much as possible, for both simple shapes and complex shapes. Moreover, our strategy does not require a prior estimate of the number of views required, and the generation process can be stopped when a pre-defined threshold is reached.

5 Conclusion

We present a novel method for high-quality text-driven 3D texturing. The automatic generation of 3D assets is an intriguing research topic and is of great value in many applications. While the pioneering works utilize the photorealistic 2D image generation ability of large-scale pre-trained generative models, the view inconsistency induced by the natural randomness of the generation process and the fact that 2D generative models are unaware of 3D consistency largely limits its performance. To move step further, we opt to learn a pseudo 3D guidance first and then use it to guide high-quality and view-consistent multi-view image generation. Specifically, we first learn the pseudo 3D guidance based on multi-view optimization using text-to-image diffusion models to align the semantics with the input and simultaneously encourage the synchronization of global semantics through continuous views. Later, we design an efficient conditional generation pipeline that enables high-quality and view-consistent multi-view image generation according to the depth map, the learned pseudo 3D guidance, and the previously generated textures. We conduct both quantitative and qualitative evaluations for text matching, visual quality, and consistency on various 3D shapes and text descriptions which demonstrate the superiority of P3G.

Acknowledgements

This work was supported in part by the National Key R&D Program of China (No. 2022ZD0118201), Natural Science Foundation of China (No. 61972217, 32071459, 62176249, 62006133, 62271465, 62332002, 62202014), the Shenzhen Medical Research Funds in China (No. B2302037), and AI for Science (AI4S)-Preferred Program, Peking University Shenzhen Graduate School, China.

References

1. Aneja, S., Thies, J., Dai, A., Nießner, M.: Clipface: Text-guided editing of textured 3d morphable models. In: ACM SIGGRAPH. pp. 1–11 (2023)
2. Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., et al.: Shapenet: An information-rich 3d model repository. arXiv preprint arXiv:1512.03012 (2015)
3. Chen, D.Z., Siddiqui, Y., Lee, H.Y., Tulyakov, S., Nießner, M.: Text2tex: Text-driven texture synthesis via diffusion models. Proceedings of the IEEE/CVF International Conference on Computer Vision (2023)
4. Chen, R., Chen, Y., Jiao, N., Jia, K.: Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22246–22256 (2023)
5. Chen, Z., Wang, F., Liu, H.: Text-to-3d using gaussian splatting. arXiv preprint arXiv:2309.16585 (2023)
6. Deitke, M., Schwenk, D., Salvador, J., Weihs, L., Michel, O., VanderBilt, E., Schmidt, L., Ehsani, K., Kembhavi, A., Farhadi, A.: Objaverse: A universe of annotated 3d objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13142–13153 (2023)
7. Efros, A.A., Leung, T.K.: Texture synthesis by non-parametric sampling. In: Proceedings of the IEEE International Conference on Computer Vision. vol. 2, pp. 1033–1038 (1999)
8. Fuji Tsang, C., Shugrina, M., Lafleche, J.F., Takikawa, T., Wang, J., Loop, C., Chen, W., Jatavallabhula, K.M., Smith, E., Rozantsev, A., Perel, O., Shen, T., Gao, J., Fidler, S., State, G., Gorski, J., Xiang, T., Li, J., Li, M., Lebaredian, R.: Kaolin: A pytorch library for accelerating 3d deep learning research. `https://github.com/NVIDIAGameWorks/kaolin` (2022)
9. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Advances in Neural Information Processing Systems. vol. 33, pp. 6840–6851 (2020)
10. Jain, A., Mildenhall, B., Barron, J.T., Abbeel, P., Poole, B.: Zero-shot text-guided object generation with dream fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 867–876 (2022)
11. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
12. Laine, S., Hellsten, J., Karras, T., Seol, Y., Lehtinen, J., Aila, T.: Modular primitives for high-performance differentiable rendering. ACM Transactions on Graphics **39**(6) (2020)
13. Liang, Y., Yang, X., Lin, J., Li, H., Xu, X., Chen, Y.: Luciddreamer: Towards high-fidelity text-to-3d generation via interval score matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)
14. Lin, C.H., Gao, J., Tang, L., Takikawa, T., Zeng, X., Huang, X., Kreis, K., Fidler, S., Liu, M.Y., Lin, T.Y.: Magic3d: High-resolution text-to-3d content creation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 300–309 (2023)
15. Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Van Gool, L.: Repaint: Inpainting using denoising diffusion probabilistic models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11461–11471 (2022)
16. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: Sdedit: Guided image synthesis and editing with stochastic differential equations. In: International Conference on Learning Representations (2021)

17. Metzer, G., Richardson, E., Patashnik, O., Giryes, R., Cohen-Or, D.: Latent-nerf for shape-guided generation of 3d shapes and textures. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)
18. Michel, O., Bar-On, R., Liu, R., Benaim, S., Hanocka, R.: Text2mesh: Text-driven neural stylization for meshes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13492–13502 (2022)
19. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: European Conference on Computer Vision. pp. 405–421 (2020)
20. Mohammad Khalid, N., Xie, T., Belilovsky, E., Popa, T.: Clip-mesh: Generating textured meshes from text using pretrained image-text models. In: ACM SIGGRAPH Asia (2022)
21. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. ACM Transactions on Graphics (ToG) **41**(4), 1–15 (2022)
22. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741 (2021)
23. Oechsle, M., Mescheder, L., Niemeyer, M., Strauss, T., Geiger, A.: Texture fields: Learning texture representations in function space. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4531–4540 (2019)
24. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. In: International Conference on Learning Representations (2023)
25. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763 (2021)
26. Richardson, E., Metzer, G., Alaluf, Y., Giryes, R., Cohen-Or, D.: Texture: Text-guided texturing of 3d shapes. arXiv preprint arXiv:2302.01721 (2023)
27. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10684–10695 (2022)
28. Siddiqui, Y., Thies, J., Ma, F., Shan, Q., Nießner, M., Dai, A.: Texturify: Generating textures on 3d shape surfaces. In: European Conference on Computer Vision. pp. 72–88. Springer (2022)
29. StabilityAI: Stable diffusion v2 depth. https://huggingface.co/stabilityai/stable-diffusion-2-depth
30. Sun, J., Zhang, B., Shao, R., Wang, L., Liu, W., Xie, Z., Liu, Y.: Dreamcraft3d: Hierarchical 3d generation with bootstrapped diffusion prior. In: International Conference on Learning Representations (2024)
31. Wang, J., Chan, K.C., Loy, C.C.: Exploring clip for assessing the look and feel of images. In: AAAI. vol. 37, pp. 2555–2563 (2023)
32. Wang, Z., Lu, C., Wang, Y., Bao, F., Li, C., Su, H., Zhu, J.: Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. In: Advances in Neural Information Processing Systems (2023)
33. Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., Xiao, J.: 3d shapenets: A deep representation for volumetric shapes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1912–1920 (2015)
34. Young, J.: Xatlas: Mesh parameterization / uv unwrapping library. `https://github.com/jpcy/xatlas` (2020)

35. Yu, J., Wang, Y., Zhao, C., Ghanem, B., Zhang, J.: Freedom: Training-free energy-guided conditional diffusion model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 23174–23184 (2023)
36. Yu, X., Dai, P., Li, W., Ma, L., Liu, Z., Qi, X.: Texture generation on 3d meshes with point-uv diffusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4206–4216 (2023)
37. Yu, X., Guo, Y.C., Li, Y., Liang, D., Zhang, S.H., Qi, X.: Text-to-3d with classifier score distillation. In: International Conference on Learning Representations (2024)