Dream-in-Style: Text-to-3D Generation Using Stylized Score Distillation

Hubert Kompanowski
Trinity College Dublin

Binh-Son Hua
Trinity College Dublin

Abstract

We present a method to generate 3D objects in styles. Our method takes a text prompt and a style reference image as input and reconstructs a neural radiance field to synthesize a 3D model with the content aligning with the text prompt and the style following the reference image. To simultaneously generate the 3D object and perform style transfer in one go, we propose a stylized score distillation loss to guide a text-to-3D optimization process to output visually plausible geometry and appearance. Our stylized score distillation is based on a combination of an original pretrained text-to-image model and its modified sibling with the key and value features of self-attention layers manipulated to inject styles from the reference image. Comparisons with state-of-the-art methods demonstrated the strong visual performance of our method, further supported by the quantitative results from our user study.

1. Introduction

Creating 3D content has been a key but demanding task in computer graphics. Traditional interactive tools such as Maya [1], Blender [3] are among the most popular choices for novices and professionals to perform 3D modeling. In the wave of generative AI development, there have been increased interests in automatic synthesis of 3D content using generative models [45, 59]. This is an open research area with tremendous progress in recent years, with the rise of language models enabling the widespread adoption of natural languages to condition the automatic generation of data in different modalities.

This trend has stimulated the development of text-to-3D generation methods [59, 69], where 3D objects can be generated by simply prompting an input sentence that describes the desired object content. These methods are generic to the appearance of the generated objects, which means that the final look and feel of the 3D content is barely controllable. This is in contrast to a common requirement in traditional 3D modeling, where a visual artist might aim to decorate a 3D object in particular styles. For example, one might be interested in creating a 3D object with low polygon count, making its geometry appear as a collection of flat surfaces, or a 3D object with stylized textures in photorealistic or cartoon styles. Performing such a stylization using traditional tools is a tedious task. Therefore, integrating stylization into generative models is a promising idea to explore.

At its core, text-to-3D generation [59, 71] performs an iterative update on a 3D representation such that its rendering converges to a photorealistic image scored by a pretrained text-to-image model. So far, most development of text-to-3D generation has focused on objects with generic appearance. Creating 3D content with particular styles remains challenging to achieve. A straightforward approach is to incorporate style description into text prompts used to generate 3D content, but this approach is not effective due to the ambiguity in how styles can be described using natural languages.

In this paper, we propose a new stylization method for 3D content creation from text prompts. Our method uses a style reference image to guide text-to-3D generation, transferring the detailed visual elements such as color, tone, or texture in the reference image to the final 3D object. This design choice is made to maximize style guidance so that the desired style can be described by both the reference image and the text prompt. Our method follows an optimization-based text-to-3D framework, performing a gradient descent update to optimize a 3D representation. Our update is regularized by a style-based score distillation that works as a critique to the rendered 3D content using a style-aware text-to-image model modified from the original pretrained model without any finetuning. We formulate this process using a stylized score distillation gradient, which dynamically combines scores from both the original and modified pretrained model. Our experiments demonstrate the effectiveness of the proposed method along with its flexibility and robustness in the styles of the generated 3D content. Some example results are shown in Figure 1.

In summary, our contributions are threefold. First, we propose to adapt a generic pretrained text-to-image model with a reference style image and build a training-free modified pretrained model for stylized text-to-3D generation. Second, we propose a stylized score distillation gradient to steer the

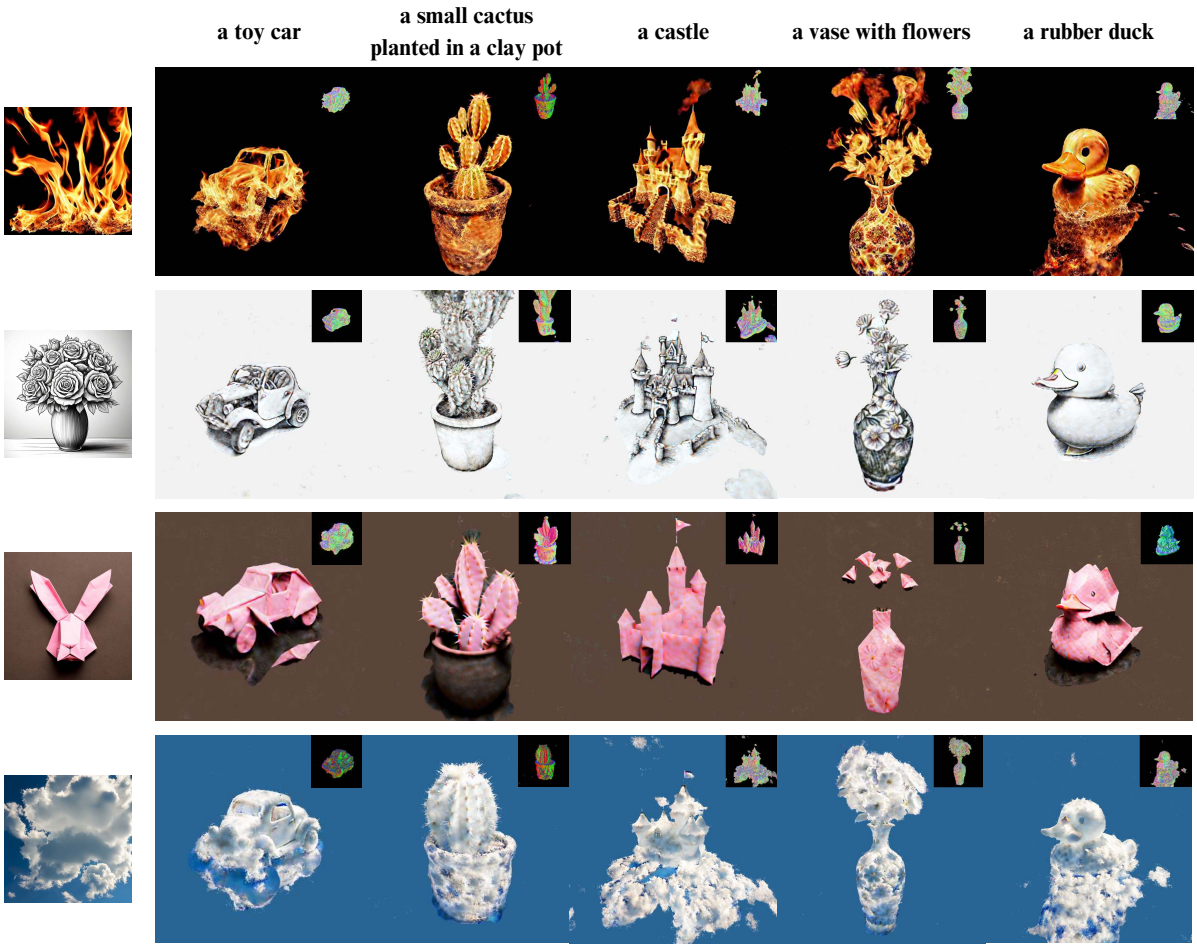

Figure 1. We aim to generate a 3D object jointly from a text prompt and a style reference image so that the object integrates the descriptive elements in the text prompt and the aesthetic style in the reference image. Our method employs a stylized score distillation to steer a text-to-3D optimization using combined scores from a pretrained text-to-image model and its modified variant on attention layer features, generating stylized 3D objects in a single-stage optimization.

3D generation toward the desired style specified in the reference style image. Third, we demonstrate the flexibility of our method by applying our stylized distillation to different text-to-3D generation losses, including score distillation sampling (SDS) [59], noise-free score distillation [34], and variational score distillation (VSD) [71]. We also demonstrate the robustness of our 3D stylization via a diverse set of text prompts and styles.

2. Related works

2.1. Style transfer

Style transfer is a traditional computer graphics problem that aims to synthesize images in artistic styles. Example-based methods such as image analogies [23, 41] learn to filter a pair of source images so that when applied to a new target image, the filter can generate analogous filtered results. Image analogies are built upon texture synthesis, which requires source pairs to approximately align to learn effective filters. This restriction is alleviated in modern deep-learning based style transfer methods. Neural style transfer [17] optimizes the target image so that it shares the content of an input image and the style of a reference image with feature correspondences characterized by a pretrained neural network. Several extensions of style transfer follow, notably for efficiency improvement with a feed-forward neural network [32] and image-to-image translation networks [29, 44, 80], style representation with statistical features [27], style representation using text-image features [15, 36, 57], and video style transfer [40]. We refer the reader to the comprehensive survey paper [31] for a broader coverage on visual style transfer methods.

2.2. Generative content creation

Generative models such as generative adversarial networks [18, 33] and diffusion models [25, 58] are notable tools to synthesize realistic data of different modalities, notably text and images when trained on large-scale datasets. Recently, text-to-image models like DALL-E [53] and Stable Diffusion [60] have shown great promise in generating photorealistic images from arbitrary text prompts. Text-to-image diffusion models can be used for image editing [48] including style transfer by learning a copy of a pretrained diffusion model [79], bridging the latent space of two diffusion models for image-to-image translations [37, 64], and instruction-based translations supervised using image-prompt-image datasets [4] and test-time editing directions [56]. Personalized text-to-image methods [61] can generate images in similar subjects defined by a small set of reference images, but these methods requires multiple images to finetune the pretrained diffusion models. Textual inversion methods [16, 20] instead only optimize the text embedding to obtain a text prompt that preserves the subject identity of an input image. Training-free methods [22, 30] inject reference features into the denoising process of a pretrained diffusion model through manipulating features input to attention layers of the denoising U-Net to influence the stylized generation.

In the 3D domain, generative models can be trained to sample 3D data represented by voxels [74] and point clouds [52], but high-quality 3D training datasets are relatively scarce and are in smaller scales compared to language and image datasets [10, 11]. A recent approach to sidestep this issue is to generate 3D data by learning from images. 3D-aware GANs [5] learn to generate 3D-consistent images by incorporating a neural radiance field [50] as the intermediate 3D representation, but their results are limited to a few categories of objects. Instead of focusing on 3D generation, novel view synthesis [6, 66] predicts 3D-consistent views from a sparse set of input views, with support from pretrained image diffusion models [45]. Large reconstruction models [26, 77] demonstrate effective image-to-3D generation by scaling up the training to millions of 3D objects. Text-to-3D generation [59, 69, 71] is a recent advance aiming to generate 3D objects directly from text prompts by using a pretrained text-to-image model to score the rendering of 3D objects at random angles. Text-to-3D methods have witnessed rapid development recently, with significant advances made toward improved distillation [34], shape quality [7, 42, 76, 81] with textures [49], fast rendering [67], amortized sampling [47, 75], 3D editing [21, 35], and animated models [2]. Our method belongs to the text-to-3D family, but focuses on stylized 3D generation.

2.3. 3D stylization

Several approaches for 3D stylization exist. Traditionally, image analogies can be adapted to stylize 3D rendering while preserving physically based illumination effects [14, 65]. Recent notable advances in neural radiance fields (NeRFs) [50] allow us to perform stylization on implicit neural scene representations. NeRF stylization methods [13, 43, 51, 55, 78] often assume a two-stage process in which a radiance field is first reconstructed on photorealistic images and then stylized based on style reference images. This process can be improved by several techniques including optimizing semantic correspondences between the radiance field and the style reference image [55, 78] or generalizing the stylization across scenes and styles with generalizable NeRFs [9, 28]. Two-stage stylization methods tend to change only the appearance of the NeRF while keeping the geometry intact [9, 28] because these methods lack a generative prior and therefore cannot generate style-related geometry. For example, given a car model and a fire style image in Fig. 1, two-stage NeRF stylization methods cannot generate a car made of fire with additional geometry representing fire. We refer the reader to a recent survey [8] for more techniques on 3D stylization.

With the rapid development of image-to-3D based on large reconstruction models [26, 46, 68, 72, 77], one can also consider generating stylized 3D objects by lifting a single stylized image to 3D. However, these pretrained image-to-3D models do not generalize well to stylized images with complex visual effects, e.g., fire around a car, resulting in 3D models with limited geometry quality. By contrast, our method integrates styles into 3D generation by considering style guidance using reference images in a text-to-3D generation framework. Our method follows a single-stage generation principle, simultaneously optimizing 3D geometry, appearance, and styles.

3. Background

3.1. Text-to-3D generation

The basic concept in text-to-3D generation is to use a pretrained text-to-image diffusion model to score the rendering of a 3D object described by a text prompt. Particularly, given a text prompt y and a pretrained text-to-image diffusion model with the noise prediction network $\epsilon_\phi(\mathbf{z}_t \mid y)$, we aim to generate a 3D object, parameterized by θ, such that its rendering $\mathbf{x} = g(\theta)$ follows the image distribution generated by the pretrained diffusion model. This generation can be formulated as an optimization problem with a score distillation sampling (SDS) gradient [59]:

$$\nabla_\theta \mathcal{L}_{SDS} = \mathbb{E}_{t,\epsilon}\left[\omega(t)\left(\epsilon_\phi(\mathbf{z}_t \mid y) - \epsilon\right)\frac{\partial \mathbf{x}}{\partial \theta}\right], \quad (1)$$

where $\omega(t)$ is a weighting function with $t \sim \mathcal{U}(0.02, 0.098)$, $\epsilon \sim \mathcal{N}(0, I)$, $\mathbf{z}_t = \alpha_t \mathbf{x} + \sigma_t \epsilon$. In practice, the score function $\epsilon_\phi(\mathbf{z}_t \mid y)$ is implemented with classifier-free guidance [24] to steer the denoising process toward conditional generation to align the generated samples with text prompt

y. Several variants of score distillation have been explored to improve the fidelity of generated 3D objects, e.g., variational score distillation (VSD) that expresses the generated objects as probabilistic distributions [71], noise-free score distillation [34] that decomposes distillation scores into interpretable components.

3.2. Baseline methods for 3D stylization

This section discusses three baseline methods for incorporating style into text-to-3D generation. We focus on supporting arbitrary styles in our synthesis, and therefore do not consider techniques that can only support a limited number of styles such as generation guided from style-dependent LoRAs of pretrained diffusion models.

Style-in-prompt. A straightforward baseline method is to use prompt engineering to add a style description directly to the input text prompt. For example, instead of having "ironman" as the original prompt, we can change to "golden ironman" to indicate the desired style of the generation. Although this approach can work for simple styles and objects, style representation using text prompts is generally ambiguous and can only capture high-level styles. It remains challenging to describe detailed visual elements in styles using text prompts, e.g., styles of sketches. Empirically, increasing the complexity of text prompts tends to make text-to-3D optimization more challenging to converge.

Neural style loss. The challenges encountered from the first method motivate us to use a reference image to describe detailed visual elements for style transfer. Our second baseline method involves using a neural style loss [17] to enforce style consistency between the 3D rendering and a style reference image. The style loss is defined by

$$\mathcal{L}_{style}(\theta) = \|f(\mathbf{x}) - f(\mathbf{s})\|_2^2, \tag{2}$$

where f represents the style features extracted by VGG-19 [63]. Style features were extracted from images using the conv1_1, conv2_1, conv3_1, conv4_1, conv4_2, and conv5_1 layers. We apply the style loss as a regularization to an existing score distillation loss.

Textual inversion. The third baseline method is specialized for text-to-3D generation by using textual inversion [16, 20] to map a style reference image to the text embedding of a text-to-image pretrained model, resulting in an augmented text prompt that implicitly encodes the style reference image. Particularly, we follow [16] to optimize a token h to reproduce the style reference image $\mathbf{s}$ so that the augmented prompt can be defined by $y' = [y + \text{"in the style of"} + h]$. We can then use the augmented prompt y' instead of y in a standard text-to-3D optimization. This baseline method depends on the accuracy of textual inversion that might affect the final 3D generation, and also requires additional computation to perform the textual inversion.

Inspired by the challenges of existing baseline methods, let us now describe our method that aims to circumvent these limitations and generate stylized 3D objects in a robust manner.

4. Method

4.1. Overview

Our method seeks a 3D object such that its rendering aligns to an input text prompt and a style reference image. We optimize a 3D neural representation using score distillation, where the rendering of the 3D object is scored by a pretrained text-to-image diffusion model [59]. Compared to the generic text-to-3D generation, one particular challenge here is to integrate the style reference image into the optimization process to generate stylized 3D objects. Our method is designed to be a single-stage optimization, where both stylized geometry and appearance are generated simultaneously. This differs from some existing 3D stylization methods where only geometry or appearance is optimized to stylize a preconstructed neural representation [55, 78].

We propose to consume our style reference image using an attention swapping mechanism on the denoising U-Net of the pretrained diffusion model [22, 30] so that the modified diffusion model can generate images analogously to the style reference image. We show that this modified pretrained model remains suitable for score distillation, which we then leverage to guide the 3D optimization.

4.2. Style-based score distillation

Mathematically, given a text prompt y and a style reference image $\mathbf{s}$, we seek a 3D object parameterized by θ, with $\mathbf{x} = g(\theta)$ being the rendered image from a differentiable rendering function g. We apply diffusion on $\mathbf{x}$, with the forward process q and reverse process p as follows. The forward process $q(\mathbf{z}_t \mid \mathbf{x} = g(\theta)) = \mathcal{N}(\alpha_t \mathbf{x}, \sigma_t^2 \mathbf{I})$ generates a noisy version $\mathbf{z}_t$ of $\mathbf{x}$ at time step t by adding Gaussian noise to $\mathbf{x}$ to remove its structure. The reverse process p predicts the noise from the intermediate state $\mathbf{z}_t$ to reconstruct $\mathbf{x}$.

We aim to synthesize the 3D object via optimizing its parameter θ by minimizing the following KL loss:

$$\mathcal{L}(\theta) = KL(q(\mathbf{z}_t \mid \mathbf{x} = g(\theta)) \parallel p_\phi(\mathbf{z}_t \mid y, \mathbf{s})), \tag{3}$$

where $p_\phi(\mathbf{z}_t \mid y, \mathbf{s})$ is a probability distribution with score function parameterized by ϕ that conditions on both the text prompt y and the style reference $\mathbf{s}$.

To model $p_\phi(\mathbf{z}_t \mid y, \mathbf{s})$, we take inspiration from training-free methods for style transfer using diffusion models [22, 30]. We assume that there are two denoising processes: one process for generating an image using the original text

prompt, and another process for generating a style reference image. Here, the style reference image can be generated by its own text prompt, or from textual inversion of a real style image. Our goal is to influence the former process so that its generated image has the original content but shares the style in the latter process. This can be achieved by sharing features in self-attention blocks [22] or swapping key and value features at self-attention blocks of the latter process with those of the original process [30], allowing features from the style reference images to propagate into image synthesis of the former process. As no finetuning is done on the diffusion model itself, this leaves the parameters of the original diffusion model intact, only the score predictions are updated due to the feature changes in the self-attention blocks. We adopt this concept for text-to-3D generation as it allows us to use the same pretrained model for original and stylized score distillation for text-to-3D generation. In our implementation, we follow the swapped attention in visual style prompting [30] but similar methods such as shared attention [22] should work as well.

Mathematically, we represent the modified denoising process by a modified score function $\hat{\epsilon}_\phi(\mathbf{z}_t \mid y, \mathbf{s})$ that shares the same network parameters ϕ as the original score function $\epsilon_\phi(\mathbf{z}_t \mid y)$. Note that the modified score function has an additional parameter $\mathbf{s}$ which is the style reference image. Specifically, assume that the style image can be generated by a prompt y_s so that $\mathbf{s} \sim p_0(\mathbf{z}_0 \mid y_s)$. Here we abuse the notation to rewrite the score functions to include self-attention features, namely $\epsilon_\phi(\mathbf{z}_t \mid y; att(y))$ for the original denoising process and $\epsilon_\phi(\mathbf{z}_t \mid y_s; att(y_s))$ for the process generating the style reference image. We define the modified score function as

$$\hat{\epsilon}_\phi(\mathbf{z}_t \mid y, \mathbf{s}) = \epsilon_\phi(\mathbf{z}_t \mid y; att(y_s)), \tag{4}$$

where the condition now includes the original prompt y and style features $att(y_s)$.

It is tempting at first to use the modified score function as a standalone distillation for 3D generation, but we very soon realize that this does not work well because the modified score function steers the generated samples toward stylized rendering. Predicting 3D shapes from stylized images is highly ambiguous, which often results in low-quality 3D geometry. By contrast, the original score function remains useful to steer the denoising process to construct meaningful object shapes. This inspires us to propose a combined score function that balances between the original and modified scores, as follows.

Combined score function. We define $p_\phi(\mathbf{z}_t \mid y, \mathbf{s})$ as a mixture of two distributions in the log space:

$$\log p_\phi(\mathbf{z}_t \mid y, \mathbf{s}) = (1-\lambda) \log p_\phi(\mathbf{z}_t \mid y) + \lambda \log \hat{p}_\phi(\mathbf{z}_t \mid y, \mathbf{s}), \tag{5}$$

where $p_\phi(\mathbf{z}_t \mid y)$ is the conditional probability distribution of the original pretrained model that only conditions on the text prompt y, and $\hat{p}_\phi(\mathbf{z}_t \mid y, \mathbf{s})$ is the conditional probability distribution of the modified pretrained model that conditions on both the text prompt and the style reference. $\lambda \in [0, 1]$ is the style ratio to control the mixture. Using this definition, minimizing $KL(q \parallel p_\phi)$ is equivalent to:

$$\min_\theta \mathbb{E}_\epsilon \left[\log(q(\mathbf{z}_t \mid \mathbf{x} = g(\theta))) \right. \\ \left. -(1-\lambda)\log(p_\phi(\mathbf{z}_t \mid y)) - \lambda \log(\hat{p}_\phi(\mathbf{z}_t \mid y, \mathbf{s}))\right]. \tag{6}$$

Taking the derivative w.r.t. θ results in our stylized score distillation (SSD) gradient:

$$\nabla_\theta \mathcal{L}_{SSD} = \mathbb{E}_{t,\epsilon}\left[\omega(t)\left((1-\lambda)\epsilon_\phi(\mathbf{z}_t \mid y) \right.\right. \\ \left.\left. +\lambda \hat{\epsilon}_\phi(\mathbf{z}_t \mid y, \mathbf{s}) - \epsilon\right) \frac{\partial \mathbf{x}}{\partial \theta}\right]. \tag{7}$$

Notably, our stylized score distillation results in linearly interpolated scores of the original and modified pretrained diffusion model, resembling SDS-family gradients. This makes extensions on SDS become applicable on our method as well, e.g., classifier-free guidance [24], and noise-free score distillation [34], as demonstrated subsequently.

Adaptation to noise-free score distillation. Following the noise-free score distillation loss decomposition and taking into account classifier-free guidance [34], we can represent the score function as a composition of a domain direction δ_D, a noise direction δ_N, and a conditioning direction δ_C. The noise-free version of our stylized score distillation can be written as

$$\nabla_\theta \mathcal{L}_{SNF} = \mathbb{E}_t\left[\omega(t)\left((1-\lambda)\left(\delta_D + \beta\delta_C\right) \right.\right. \\ \left.\left. +\lambda\left(\hat{\delta}_D + \beta\hat{\delta}_C\right)\right) \frac{\partial \mathbf{x}}{\partial \theta}\right], \tag{8}$$

where the domain directions are defined by

$$\delta_D = \begin{cases} \epsilon_\phi(\mathbf{z}_t \mid y = \emptyset), & \text{if } t < 200 \\ \epsilon_\phi(\mathbf{z}_t \mid y = \emptyset) - \epsilon_\phi(\mathbf{z}_t \mid y = p_{neg}), & \text{otherwise,} \end{cases} \tag{9}$$

and

$$\hat{\delta}_D = \begin{cases} \hat{\epsilon}_\phi(\mathbf{z}_t \mid y = \emptyset, \mathbf{s}), & \text{if } t < 200 \\ \hat{\epsilon}_\phi(\mathbf{z}_t \mid y = \emptyset, \mathbf{s}) - \hat{\epsilon}_\phi(\mathbf{z}_t \mid y = p_{neg}, \mathbf{s}), & \text{otherwise,} \end{cases} \tag{10}$$

where p_{neg} is a negative prompt to represent out-of-distribution samples such as “unrealistic, blurry, low quality”.

The conditioning directions are defined by

$$\delta_C = \epsilon_\phi(\mathbf{z}_t \mid y) - \epsilon_\phi(\mathbf{z}_t \mid y = \emptyset), \tag{11}$$

and

$$\hat{\delta}_C = \hat{\epsilon}_\phi(\mathbf{z}_t \mid y, \mathbf{s}) - \hat{\epsilon}_\phi(\mathbf{z}_t \mid y = \emptyset, \mathbf{s}), \tag{12}$$

where β is the classifier-free guidance (CFG) scale.

4.3. Optimization

We found that the style ratio λ greatly affects the convergence of the optimization as it controls the gradients that steer the denoising process toward generating a generic 3D object and its stylized version. We devise a dynamic schedule to adapt the style ratio during optimization as follows. We aim for a small style ratio in early iterations so that basic structures in the 3D object can be generated following the vanilla scores. In subsequent iterations, we increase the style ratio to favor stylized score distillation, emphasizing the importance of generating stylized 3D objects. We explore two dynamic schedules using a square root function:

$$\tau_{\text{sqrt}}(\lambda; \lambda_{\text{max}}, k, K) = \lambda_{\text{max}} \sqrt{\frac{k}{K}}, \tag{13}$$

and a quadratic function:

$$\tau_{\text{quad}}(\lambda; \lambda_{\text{max}}, k, K) = \lambda_{\text{max}} \left(\frac{k}{K}\right)^2, \tag{14}$$

where k and K are the current and total iterations in the optimization, respectively. λ_{max} is the maximum value that the style ratio parameter λ will reach at the end of the scaling process when $k = K$.

5. Experimental Results

We perform several experiments to demonstrate the effectiveness of our proposed method. First, we compare our method with three baseline methods for stylized text-to-3D generation. Second, we demonstrate that our method can be adapted to other score distillation losses. Finally, we provide ablation studies to validate the importance of our combined score distillation, as well as perform parameter studies to validate our style ratio scheduling. We perform quantitative evaluation of our method through a human-like user study using large language models [73].

5.1. Implementation details

We use the implementation of neural radiance fields from threestudio [19], which is based on NerfAcc [39], as the 3D representation for our optimization. We follow the implementation of DreamFusion [59], noise-free score distillation [34] and ProlificDreamer [71] to implement the score distillations. In our method, we apply an augmentation of the text prompt y by concatenating it with a BLIP2 generated caption [38] of the style reference image, and use this augmented prompt for the modified score $\hat{\epsilon}$. We set the classifier-free guidance (CFG) scale to 100 for score distillation sampling (SDS) [59], and 7.5 for noise-free score distillation (NFSD) [34], and variational score distillation (VSD) [71]. We use NFSD as the default score distillation for our method.

Our experiments are performed on a NVIDIA RTX 4090 GPU with 24 GB of VRAM. Our method optimized a 3D object in approximately 1.5 hours using SDS [59] or NFSD [34] and 2.5 hours using VSD [71], similar to the training time of the vanilla implementations of these methods.

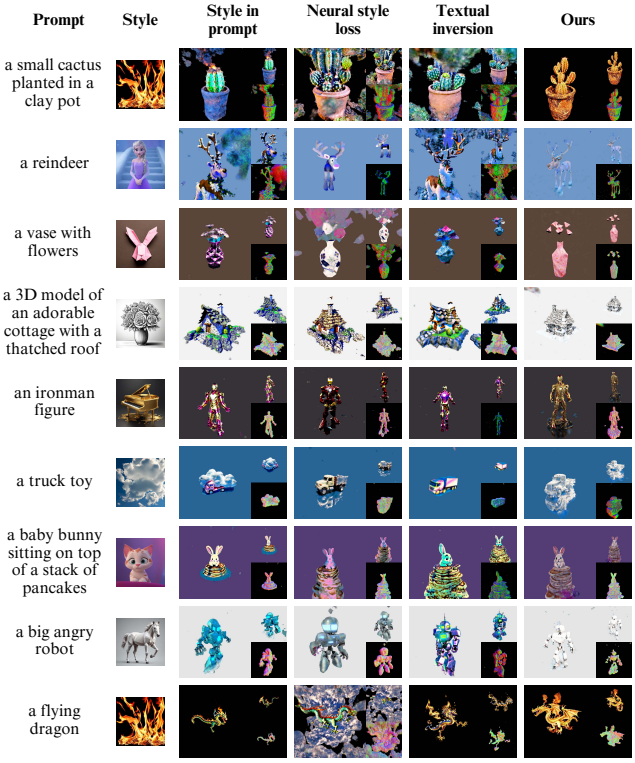

Figure 2. Qualitative comparisons to baseline methods. Our stylized score distillation leads to 3D generation results consistently aligned with the text prompts and the style reference images. Multiple view rendering are shown in the supplementary video.

5.2. Qualitative results

We compare our method with the baseline methods proposed in Section 3.2. Figure 2 presents a list of text prompts and style reference images with the corresponding outputs of all methods. As can be seen, our results have the best visual quality with consistent alignment to the input pairs of text prompts and reference images.

To demonstrate the adaptability of our method to other

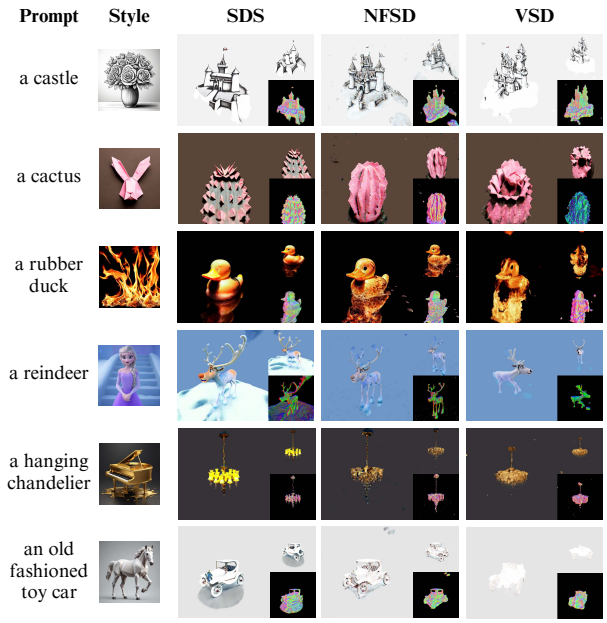

Figure 3. Our method applied to different score distillation losses. Multiple view rendering are shown in the supplementary video.

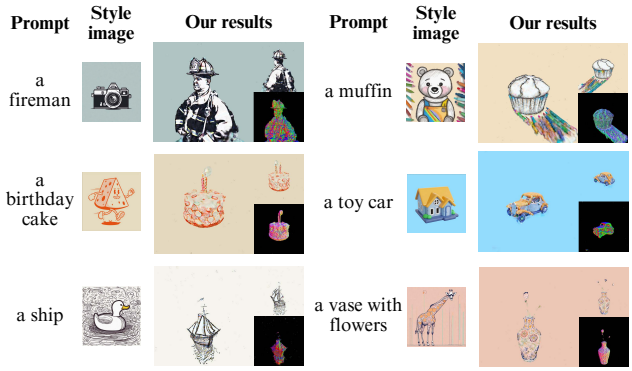

Figure 4. Our results on various style images from [22].

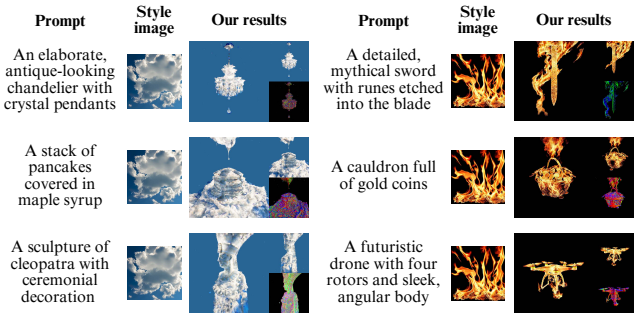

Figure 5. Our results on complex and detailed text prompts.

score distillation losses, we apply our method to the vanilla score distillation sampling (SDS), noise-free score distillation (NFSD), and variational score distillation (VSD). Figure 3 provides the results of our method adapted to these score distillation losses. It can be seen that the NFSD variant works best, outperforming SDS and VSD. Our method, however, does not yet adapt the LoRA model for scoring noisy rendering in VSD for stylization which is the subject of future work.

To demonstrate the robustness of our method across different styles and complex prompts, we tested it on various styles as in [22], as shown in Figure 4. Additionally, we applied our approach to more complex textual prompts, which are illustrated in Figure 5.

To verify the robustness of our stylized score distillation across training-free methods for style transfer, we attempted our method on StyleAligned [22]. Compared to visual style prompting (VSP) [30] that only attends to the style image, StyleAligned attends to both the style image and the resulting image. We observed that our SSD with VSP has more natural results. For example, given a fire style image and prompt 'a toy car', SSD with StyleAligned generates fire around a car, while VSP produces a car made of fire (similar trend with their 2D results). A more detailed discussion is provided in the supplementary material.

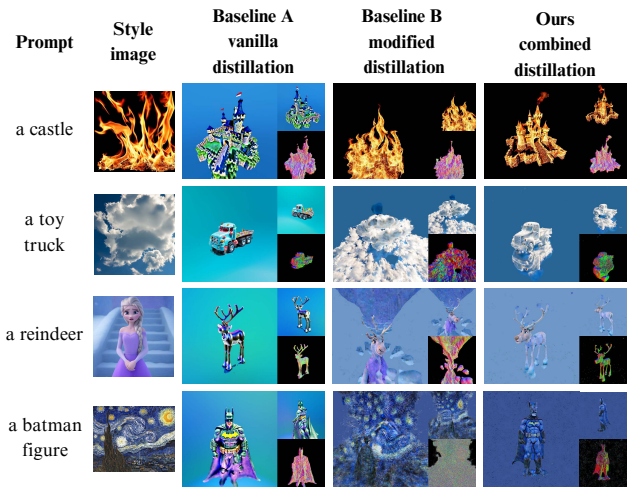

Figure 6. Ablation studies. We confirm the effectiveness of our combined distillation by comparing with two baselines: A) the vanilla text-to-3D generation without styles, B) text-to-3D generation guided by only the modified pretrained model.

5.3. Ablation studies

Figure 6 provides an ablation study on the effectiveness of our method by comparing with two baselines: A) no style reference image (i.e., the vanilla text-to-3D generation), B) text-to-3D generation guided by only the modified pretrained model. It is shown that baseline A has high-quality object geometry and appearance, while baseline B has consistent styles with the reference images but causes corrupted ge-

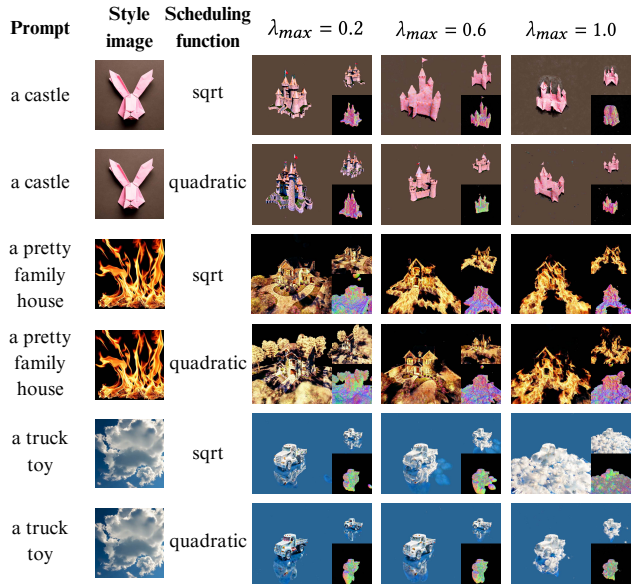

Figure 7. Effects of our schedule function with different style ratios on the stylization results.

ometry. This confirms the need for using our stylized score distillation to generate the desired objects in styles.

Figure 7 provides a study on the choice of the style ratio λ. We found that starting the generation process with an unstyled object and then gradually adding the style works best. We evaluated two scheduling functions, quadratic and square root, which sets λ between 0 and λ_{max}. The results are shown in Figure 7. We found that the $sqrt$ schedule works best with $\lambda_{max} = 0.6$ in general. When the style reference image represents abstract concepts without a specific foreground object, it is preferred to use the $quad$ schedule with $\lambda_{max} = 1.0$.

5.4. Quantitative results

The existence of large language models with vision capability (e.g., GPT-4v model [54]) allows us to prompt a language model for 3D asset evaluation, which had been demonstrated to work well for the text-to-3D generation task with human-like performance [73]. We follow [73] and extend their GPTEval3D tool to incorporate style evaluation, resulting to six evaluation criteria including text-geometry alignment, text-asset alignment, style alignment, geometry details, texture details, and 3D plausibility. We use this tool to compare the results generated by four methods, including style-in-prompt, neural style loss, textual inversion, and our method. We set up the style-in-prompt method as the base/anchor for the evaluation process. We asked GPTEval3D to perform 120 pairwise comparisons, and then calculated the Elo score [12] for each method. The results are presented in Figure 8. It can be seen that our method outperforms all baseline methods in this evaluation.

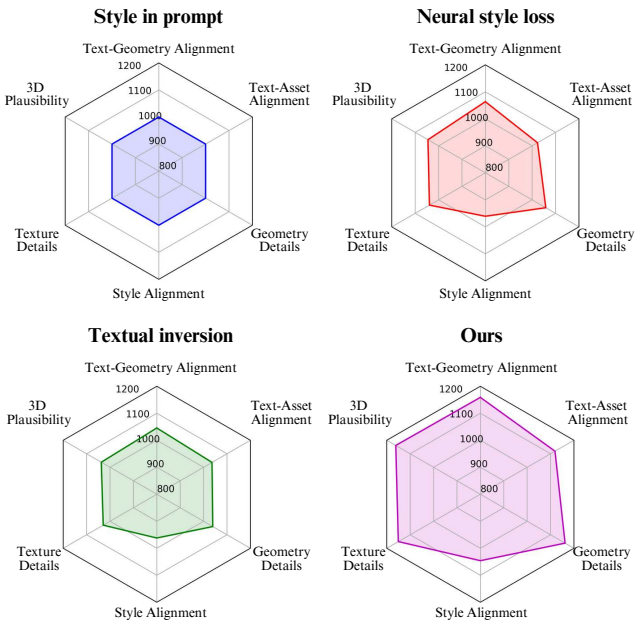

Figure 8. Results of GPTEval3D on text-geometry alignment, text-asset alignment, style alignment, geometry details, texture details, and 3D plausibility, which confirm the effectiveness of our method.

6. Conclusions and limitations

We present a method for text-to-3D generation in styles. Our method is based on a combined score distillation to balance the influence of the original and the modified pre-trained diffusion model in generating stylized 3D objects. We demonstrated the performance and robustness of our method in various comparisons.

Our method is not without limitations. A particular problem is that our results are prone to the Janus problem [59], which could be mitigated by using pretrained diffusion models for multi-view generation [62, 70]. Additionally, our method is best compatible with SDS-family losses. Adapting our method to non-SDS losses is future work. It is of great interest to extend the investigation of stylized text-to-3D generation to videos and 4D data for style references and outputs, respectively.

7. Acknowledgments

This work was conducted with the financial support of the Research Ireland Centre for Research Training in Digitally-Enhanced Reality (d-real) under Grant No. 18/CRT/6224. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

This project is supported by Research Ireland under the Research Ireland Frontiers for the Future Programme, award number 22/FFP-P/11522.

References

[1] Autodesk. *Maya*, 2019. 1

[2] Sherwin Bahmani, Ivan Skorokhodov, Victor Rong, Gordon Wetzstein, Leonidas Guibas, Peter Wonka, Sergey Tulyakov, Jeong Joon Park, Andrea Tagliasacchi, and David B. Lindell. 4d-fy: Text-to-4d generation using hybrid score distillation sampling. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[3] Blender Foundation. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. 1

[4] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 3

[5] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16123–16133, 2022. 3

[6] Eric R. Chan, Koki Nagano, Matthew A. Chan, Alexander W. Bergman, Jeong Joon Park, Axel Levy, Miika Aittala, Shalini De Mello, Tero Karras, and Gordon Wetzstein. GeNVS: Generative novel view synthesis with 3D-aware diffusion models. In *ICCV*, 2023. 3

[7] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 3

[8] Yingshu Chen, Guocheng Shao, Ka Chun Shum, Binh-Son Hua, and Sai-Kit Yeung. Advances in 3d neural stylization: A survey. *arXiv preprint arXiv:2311.18328*, 2023. 3

[9] Pei-Ze Chiang, Meng-Shiun Tsai, Hung-Yu Tseng, Wei sheng Lai, and Wei-Chen Chiu. Stylizing 3d scene via implicit representation and hypernetwork. In *WACV*, 2022. 3

[10] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, Eli VanderBilt, Aniruddha Kembhavi, Carl Vondrick, Georgia Gkioxari, Kiana Ehsani, Ludwig Schmidt, and Ali Farhadi. Objaverse-XL: A universe of 10m+ 3d objects. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. 3

[11] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 3

[12] A.E. Elo. *The USCF Rating System: Its Development, Theory, and Applications*. United States Chess Federation, 1966. 8

[13] Zhiwen Fan, Yifan Jiang, Peihao Wang, Xinyu Gong, Dejia Xu, and Zhangyang Wang. Unified implicit neural stylization. In *ECCV*, 2022. 3

[14] Jakub Fišer, Ondřej Jamriška, Michal Lukáč, Eli Shechtman, Paul Asente, Jingwan Lu, and Daniel Sýkora. Stylit: Illumination-guided example-based stylization of 3d renderings. *ACM Trans. Graph.*, 2016. 3

[15] Rinon Gal, Or Patashnik, Haggai Maron, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)*, 2022. 2

[16] 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. In *ICLR*, 2023. 3, 4

[17] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016. 2, 4

[18] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, 2014. 3

[19] 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. 6

[20] Inhwa Han, Serin Yang, Taesung Kwon, and Jong Chul Ye. Highly personalized text embedding for image manipulation by stable diffusion. *arXiv preprint arXiv:2303.08767*, 2023. 3, 4

[21] Amir Hertz, Kfir Aberman, and Daniel Cohen-Or. Delta denoising score. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2328–2337, 2023. 3

[22] Amir Hertz, Andrey Voynov, Shlomi Fruchter, and Daniel Cohen-Or. Style aligned image generation via shared attention. In *CVPR*, 2024. 3, 4, 5, 7

[23] Aaron Hertzmann, Charles E Jacobs, Nuria Oliver, Brian Curless, and David H Salesin. Image analogies. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 327–340, 2001. 2

[24] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 3, 5

[25] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in neural information processing systems*, pages 6840–6851, 2020. 3

[26] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. In *ICLR*, 2024. 3

[27] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017. 2

[28] Yi-Hua Huang, Yue He, Yu-Jie Yuan, Yu-Kun Lai, and Lin Gao. Stylizednerf: Consistent 3d scene stylization as stylized nerf via 2d-3d mutual learning. In *CVPR*, 2022. 3

[29] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 2

[30] Jaeseok Jeong, Junho Kim, Yunjey Choi, Gayoung Lee, and Youngjung Uh. Visual style prompting with swapping self-attention. *arXiv preprint arXiv:2402.12974*, 2024. 3, 4, 5, 7

[31] Yongcheng Jing, Yezhou Yang, Zunlei Feng, Jingwen Ye, Yizhou Yu, and Mingli Song. Neural style transfer: A review. *IEEE transactions on visualization and computer graphics*, 2019. 2

[32] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 2

[33] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 3

[34] Oren Katzir, Or Patashnik, Daniel Cohen-Or, and Dani Lischinski. Noise-free score distillation. *ICLR*, 2024. 2, 3, 4, 5, 6

[35] Juil Koo, Chanho Park, and Minhyuk Sung. Posterior distillation sampling. In *CVPR*, 2024. 3

[36] Gihyun Kwon and Jong Chul Ye. Clipstyler: Image style transfer with a single text condition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18062–18071, 2022. 2

[37] Bo Li, Kaitao Xue, Bin Liu, and Yu-Kun Lai. Bbdm: Image-to-image translation with brownian bridge diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1952–1961, 2023. 3

[38] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, pages 19730–19742, 2023. 6

[39] Ruilong Li, Hang Gao, Matthew Tancik, and Angjoo Kanazawa. Nerfacc: Efficient sampling accelerates nerfs, 2023. 6

[40] Xueting Li, Sifei Liu, Jan Kautz, and Ming-Hsuan Yang. Learning linear transformations for fast image and video style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3809–3817, 2019. 2

[41] Jing Liao, Yuan Yao, Lu Yuan, Gang Hua, and Sing Bing Kang. Visual attribute transfer through deep image analogy. *ACM Trans. Graph.*, 2017. 2

[42] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 3

[43] Kunhao Liu, Fangneng Zhan, Yiwen Chen, Jiahui Zhang, Yingchen Yu, Abdulmotaleb El Saddik, Shijian Lu, and Eric Xing. Stylerf: Zero-shot 3d style transfer of neural radiance fields. In *CVPR*, 2023. 3

[44] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Advances in neural information processing systems*, pages 700–708, 2017. 2

[45] 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. In *ICCV*, 2023. 1, 3

[46] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. In *CVPR*, 2024. 3

[47] Jonathan Lorraine, Kevin Xie, Xiaohui Zeng, Chen-Hsuan Lin, Towaki Takikawa, Nicholas Sharp, Tsung-Yi Lin, Ming-Yu Liu, Sanja Fidler, and James Lucas. Att3d: Amortized text-to-3d object synthesis. *ICCV*, 2023. 3

[48] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022. 3

[49] Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. Latent-nerf for shape-guided generation of 3d shapes and textures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12663–12673, 2023. 3

[50] 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*, pages 405–421. Springer, 2020. 3

[51] Thu Nguyen-Phuoc, Feng Liu, and Lei Xiao. Snerf: stylized neural implicit representations for 3d scenes. *ACM Transactions on Graphics*, 2022. 3

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

[53] OpenAI. Dall·e 3, 2023. 3

[54] OpenAI. Gpt-4v, 2023. 8

[55] Hong-Wing Pang, Binh-Son Hua, and Sai-Kit Yeung. Locally stylized neural radiance fields. In *ICCV*, 2023. 3, 4

[56] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023. 3

[57] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2085–2094, 2021. 2

[58] Ryan Po, Wang Yifan, Vladislav Golyanik, Kfir Aberman, Jonathan T. Barron, Amit H. Bermano, Eric Ryan Chan, Tali Dekel, Aleksander Holynski, Angjoo Kanazawa, C. Karen Liu, Lingjie Liu, Ben Mildenhall, Matthias Nießner, Björn Ommer, Christian Theobalt, Peter Wonka, and Gordon Wetzstein. State of the art on diffusion models for visual computing. *Computer Graphics Forum*, 2024. 3

[59] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *The Eleventh International Conference on Learning Representations*, 2023. 1, 2, 3, 4, 6, 8

[60] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 3

[61] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, 2023. 3

[62] Yichun Shi, Peng Wang, Jianglong Ye, Long Mai, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. In *ICLR*, 2024. 8

[63] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015. 4

[64] Xuan Su, Jiaming Song, Chenlin Meng, and Stefano Ermon. Dual diffusion implicit bridges for image-to-image translation. In *The Eleventh International Conference on Learning Representations*, 2023. 3

[65] Daniel Sýkora, Ondřej Jamriška, Ondřej Texler, Jakub Fišer, Michal Lukáč, Jingwan Lu, and Eli Shechtman. StyleBlit: Fast example-based stylization with local guidance. *Computer Graphics Forum*, 38(2):83–91, 2019. 3

[66] Stanislaw Szymanowicz, Christian Rupprecht, and Andrea Vedaldi. Viewset diffusion: (0-)image-conditioned 3D generative models from 2D data. In *ICCV*, 2023. 3

[67] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *ICLR*, 2024. 3

[68] Dmitry Tochilkin, David Pankratz, Zexiang Liu, Zixuan Huang, , Adam Letts, Yangguang Li, Ding Liang, Christian Laforte, Varun Jampani, and Yan-Pei Cao. Triposr: Fast 3d object reconstruction from a single image. *arXiv preprint arXiv:2403.02151*, 2024. 3

[69] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1, 3

[70] Peng Wang and Yichun Shi. Imagedream: Image-prompt multi-view diffusion for 3d generation. *arXiv preprint arXiv:2312.02201*, 2023. 8

[71] 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. In *Advances in Neural Information Processing Systems*, 2023. 1, 2, 3, 4, 6

[72] Zhengyi Wang, Yikai Wang, Yifei Chen, Chendong Xiang, Shuo Chen, Dajiang Yu, Chongxuan Li, Hang Su, and Jun Zhu. Crm: Single image to 3d textured mesh with convolutional reconstruction model. In *ECCV*, 2024. 3

[73] Tong Wu, Guandao Yang, Zhibing Li, Kai Zhang, Ziwei Liu, Leonidas Guibas, Dahua Lin, and Gordon Wetzstein. Gpt-4v(ision) is a human-aligned evaluator for text-to-3d generation. In *CVPR*, 2024. 6, 8

[74] Zhirong Wu, Shuran Song, Aditya Khosla, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shape modeling. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, USA, 2015. 3

[75] Kevin Xie, Jonathan Lorraine, Tianshi Cao, Jun Gao, James Lucas, Antonio Torralba, Sanja Fidler, and Xiaohui Zeng. Latte3d: Large-scale amortized text-to-enhanced3d synthesis. *arXiv preprint arXiv:2403.15385*, 2024. 3

[76] Jiale Xu, Xintao Wang, Weihao Cheng, Yan-Pei Cao, Ying Shan, Xiaohu Qie, and Shenghua Gao. Dream3d: Zero-shot text-to-3d synthesis using 3d shape prior and text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20908–20918, 2023. 3

[77] Yinghao Xu, Hao Tan, Fujun Luan, Sai Bi, Peng Wang, Jiahao Li, Zifan Shi, Kalyan Sunkavalli, Gordon Wetzstein, Zexiang Xu, and Kai Zhang. Dmv3d: Denoising multi-view diffusion using 3d large reconstruction model. In *ICLR*, 2024. 3

[78] Kai Zhang, Nick Kolkin, Sai Bi, Fujun Luan, Zexiang Xu, Eli Shechtman, and Noah Snavely. Arf: Artistic radiance fields. In *ECCV*, 2022. 3, 4

[79] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 3

[80] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 2

[81] Junzhe Zhu, Peiye Zhuang, and Sanmi Koyejo. HIFA: High-fidelity text-to-3d generation with advanced diffusion guidance. In *The Twelfth International Conference on Learning Representations*, 2024. 3