CAD: Photorealistic 3D Generation via Adversarial Distillation

Ziyu Wan[1,2] Despoina Paschalidou[2] Ian Huang[2] Hongyu Liu[3] Bokui Shen[2]
Xiaoyu Xiang Jing Liao[1*] Leonidas Guibas[2]
[1]City University of Hong Kong [2]Stanford University [3]HKUST

raywzy.com/CAD

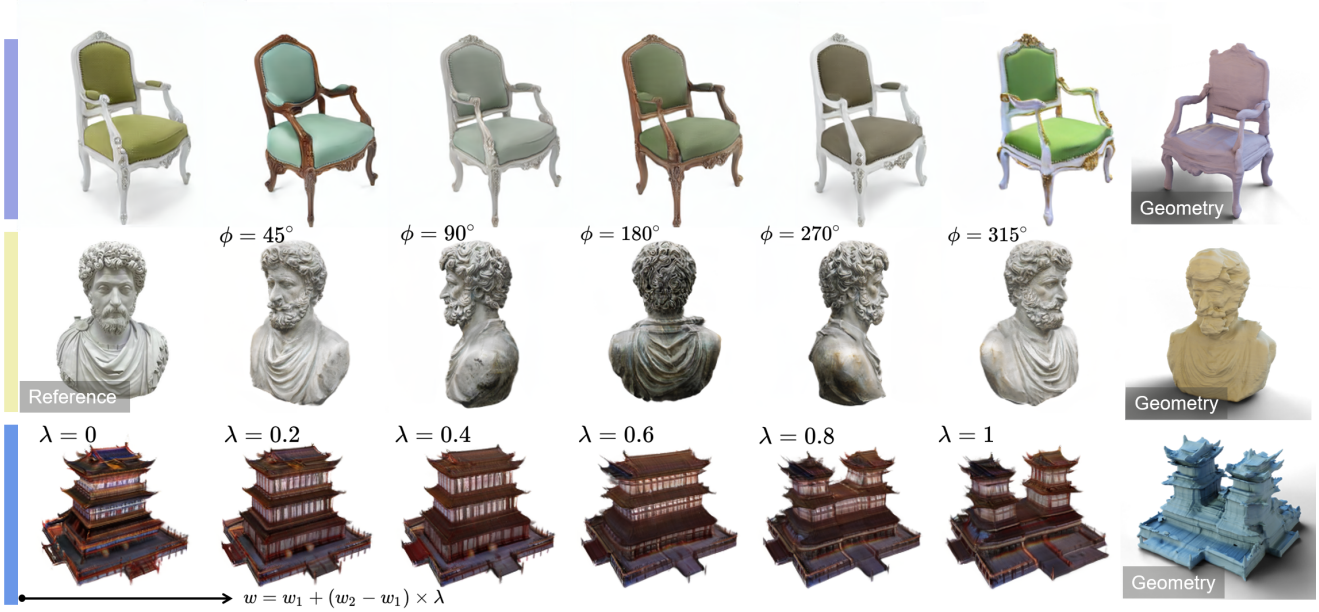


Figure 1. CAD leverages pretrained diffusion models to generate photorealistic 3D contents based on a single input image and the text prompt, enabling different applications including • diversified generation, • single-view reconstruction by inversion and • 3D interpolation.

Abstract

The increased demand for 3D data in AR/VR, robotics and gaming applications, gave rise to powerful generative pipelines capable of synthesizing high-quality 3D objects. Most of these models rely on the Score Distillation Sampling (SDS) algorithm to optimize a 3D representation such that the rendered image maintains a high likelihood as evaluated by a pre-trained diffusion model. However, finding a correct mode in the high-dimensional distribution produced by the diffusion model is challenging and often leads to issues such as over-saturation, over-smoothing, and Janus-like artifacts. In this paper, we propose a novel learning paradigm for 3D synthesis that utilizes pre-trained diffusion models. Instead of focusing on mode-seeking, our method directly models the distribution discrepancy between multi-view renderings and diffusion priors in an adversarial manner, which unlocks the generation of high-fidelity and photorealistic 3D content, conditioned on a single image and prompt. Moreover, by harnessing the latent space of GANs and expressive diffusion model priors, our method facilitates a wide variety of 3D applications including single-view reconstruction, high diversity generation and continuous 3D interpolation in the open domain. The experiments demonstrate the superiority of our pipeline compared to previous works in terms of generation quality and diversity.

1. Introduction

In recent years, we have witnessed an unprecedented explosion in generative models that can synthesize intelligible text [12, 83, 109], photorealistic images [50, 57, 69, 79, 89, 90, 92, 92, 94, 110, 115, 125], video sequences [6, 16, 101, 102, 111], music [3, 9, 24] and 3D data [14, 15, 27, 34, 46, 60, 67, 78, 85, 107, 116, 124, 130–132, 135]. In particular, when dealing with 3D data, manually creating them is a laborious endeavor that necessitates technical skills from highly experienced designers. Therefore, having systems capable of automatically generating realistic and diverse 3D contents could significantly facilitate the workflow of artists and product designers and could

* Corresponding author.

enable new levels of creativity through "generative art" [8].

Recently diffusion models [40, 105, 106] have emerged as a powerful class of tools that can produce photorealistic images conditioned on versatile user inputs [90, 92, 93, 132] such as text, depth maps, semantic masks etc. However, naively adopting them on 3D synthesis tasks is not trivial, due to the lack of large-scale, richly annotated 3D datasets. Despite the recently introduced larger 3D datasets [21, 120], they remain significantly smaller compared to contemporary image-text datasets that typically contain billions of examples. Therefore, a large body of works [15, 60, 85, 117] explored using pre-trained 2D text-to-image diffusion models [92, 93] to generate 3D data. Among the first works was DreamFusion [85] that introduced the Score Distillation Sampling (SDS) algorithm for learning a 3D representation, such that the rendered image from any view looks similar (i.e. has high likelihood) to a sample from the pre-trained 2D diffusion model, given a text description.

Despite their impressive performance [15, 58, 60, 72, 114], SDS-based arts frequently exhibit issues such as oversaturation, over-smoothness, non-photorealism, and limited diversity, primarily attributed to the quality degradation of mode-seeking behaviors in deep generative models [74]. To alleviate these drawbacks, in a concurrent work, Wang et al. introduced Variational Score Distillation (VSD) [117], which is a generalized version of SDS that aims to optimize a 3D distribution to approximate the distribution defined by the diffusion model. While VSD [117] addressed some issues of SDS, its rendering quality still suffers from highly saturated colors hence resulting in less photorealistic generations.

In this paper, we propose Consistent Adversarial Distillation (CAD), a new approach for generating 3D objects conditioned on a text prompt and a single image, to overcome the mentioned issues. Instead of optimizing a single NeRF through score distillation, our key idea is to train a 3D generator that directly models the conditional distribution of a pre-trained diffusion model, through adversarial learning. Although modeling the distribution of generic concepts with Generative Adversarial Networks (GANs) [30] may be challenging, utilizing input conditions such as text or images could effectively constrain the data distribution, since the 3D objects following specific conditions should share similar scale, shape and appearance in the canonical space, which GANs could handle pretty well [14, 17, 39, 80, 95]. Moreover, as the generator learns a mapping from the latent to the continuous 3D distribution, our model becomes applicable to various downstream tasks, including diversified sampling, single-view reconstruction and 3D interpolation (see Fig. 1).

However, distilling prior knowledge from a pre-trained diffusion model into a 3D GAN is not trivial. Existing attempts typically rely on large amount of high-quality and aligned data [13, 80, 95, 100, 103, 119] with evenly-distributed poses [13, 14, 31, 95, 104, 121]. When sampling novel images from a pre-trained 2D diffusion model, there is no guarantee that the data distribution will adequately cover all azimuth angles that fully define the shape's geometry. On the contrary, due to the existing inductive bias of the diffusion model, it is more likely to produce frontal-facing data, even after using prompt engineering or negative prompts. We resolve this by leveraging the view-dependent diffusion model of [64] and further propose several distribution pruning and refinement strategies that ensure stable training as well as diverse and high-quality samples. Finally, we employ a 3D-aware GAN [14] to learn a 3D generator that models the conditional distribution of a pre-trained diffusion model. We evaluate our model on several datasets and showcase that it can generate high-quality, diverse and photorealistic 3D objects conditioned on a single image and a text prompt.

2. Related Work

3D Generative Models. Over the years, several works explored combining generative models [30, 40, 53, 91, 105] with different 3D representations such as voxel grids [26, 39, 44, 66, 119], meshes [28, 29, 75, 134], point clouds [2, 88, 122, 127] or neural implicits [19, 54, 128, 129]. Although most of these pipelines can generate plausible 3D geometries, they require explicit 3D supervision. To this end, many works investigated learning the 3D scene geometry and appearance through volumetric [73, 82, 123] and differentiable rendering [59, 112, 134]. The key advantage of these pipelines is that they can recover 3D information only from images. In this work, we introduce a model capable of generating high-quality 3D objects conditioned on a single image and a text prompt, hence effectively alleviating the need for both 3D data and multi-view images during training.

3D-Aware Generative Models. GANs [30] have demonstrated impressive capabilities on several image synthesis [10, 20, 48, 50] and editing [4, 11, 20, 45, 61, 98, 113] tasks. However, adopting them to 3D data is non-trivial as they ignore the physics of the image formation process, hence failing to produce 3D consistent renderings. To address this, several works [23, 31, 35, 37, 38, 70, 76, 77, 81, 136] explored incorporating explicit 3D representations or combining GANs [13, 14, 95, 103] with Neural Radiance Fields (NeRFs) [73]. Due to their compelling results, various follow-up works further improved various aspects of the synthesis process such as the rendering quality [14, 31, 96, 103, 121] , the underlying geometry [84, 99], the editing capabilities [36, 52, 55, 65, 80, 108, 126]. Our generator architecture is similar to EG3D [14], however our training pipeline that leverages 2D diffusion priors is novel and enables the generation of objects from arbitrary categories based on textual descriptions and image guidance.

3D Generation Guided by 2D Prior Models. Our work falls into the category of methods that leverage priors from

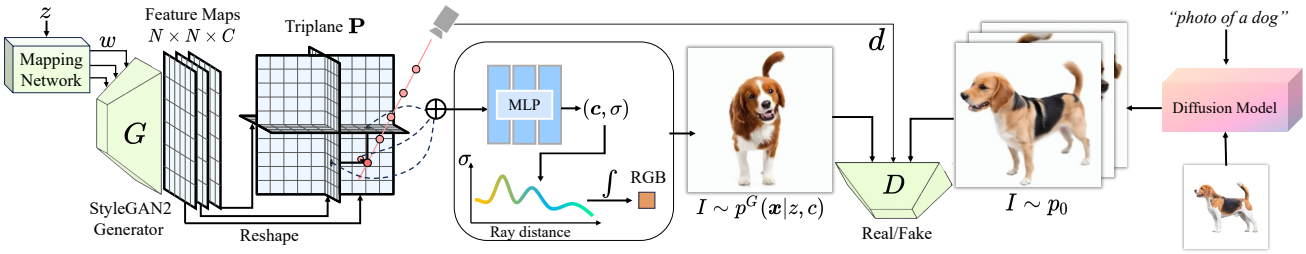


Figure 2. Our adversarial distillation framework mainly comprises three parts: a StyleGAN2-based generator for approximating the target distribution, a pre-trained diffusion model for providing 2D priors based on the given input image and text prompt, and a discriminator for minimizing the distribution gap between $p^G(\boldsymbol{x} \mid \mathbf{z}, \mathbf{c})$ and $p_0(\boldsymbol{x}_0 \mid y)$. For brevity, we omit some details of the triplane generator training. Our method could effectively overcome the issues of score distillation and achieve highly photorealistic and diverse 3D generation.

text-to-image diffusion models [42, 79, 90, 92, 93] for generating 3D data. DreamFusion [85] was among the first that proposed to distill a pre-trained diffusion model [93] into a NeRF [73] for text-guided 3D synthesis. In particular, the objective was to ensure that the rendered images, would match the distribution of the sampled photorealistic images conditioned on a specific text prompt. This process of sampling through optimization is referred to as Score Distillation Sampling (SDS). However, naively applying SDS for 3D synthesis poses several challenges such as over-smoothing, saturated colors as well as Janus-like issues. Concurrently, ProlificDreamer [117] proposed to address these issues with Variational Score Distillation (VSD). The key difference between SDS and VSD is that the latter treats a 3D scene as a random variable, as opposed to a single data point. While [117] addresses some of the issues of SDS, it still suffers from over-saturated colors. In contrast, our work mitigates these challenges by training a 3D-aware generator that directly models the distribution of a diffusion model.

Our work is related to approaches that generate 3D objects conditioned on a single input image. Among the first works in this direction were [5, 32, 47, 118, 137] that proposed training a 3D-aware diffusion model for novel view synthesis. Despite their competitive results, they could only be evaluated on objects from categories seen during training. To mitigate this, an alternative line of research explored using pre-trained 2D diffusion models [22, 68]. To learn control over the camera viewpoint, recently, Zero-1-to-3 [64] fine-tuned a pre-trained image-to-text diffusion model [92] on synthetic data [21]. In a follow-up work, One-2-3-45 [63] employed the view-dependent diffusion priors of [64] to train a multi-view 3D reconstruction pipeline to enable faster inference. Similar to our work Magic123 [86] uses 2D diffusion together with view-dependent diffusion priors [64] for generating 3D textured meshes from a single image.

3. Method

Given a single image and a text prompt, our goal is to generate high-quality, photorealistic and diverse 3D content by distilling pre-trained diffusion models. First, we show how the concept of continuous distribution modeling avoids the quality degradation brought by mode-seeking (Sec. 3.1). Then we introduce CAD, our framework that distills knowledge from pre-trained diffusion models to a 3D GAN with multi-view consistent rendering (Sec. 3.2). However, due to the inherent inductive biases embedded in 2D diffusion models, they tend to only generate images from frontal viewpoints and can not describe the full 3D geometry. We further mitigate this issue by introducing a series of novel strategies to sample multi-view and diversified data from the conditional distribution of diffusion models (Sec. 3.3).

3.1. Representing 3D Distributions

Preliminaries. We first review prior work that leverages diffusion priors for 3D generation. Given a pre-trained text-to-image diffusion model $p_t(\boldsymbol{x}_t \mid y)$ with the noise prediction network $\boldsymbol{\epsilon}_{\text{pretrain}}(\boldsymbol{x}_t, t, y)$, SDS [85] tries to optimize a single NeRF [73] with parameters θ, s.t. its rendered results $\boldsymbol{x}$, given a camera pose $\mathbf{c}$ sampled from a camera distribution $p_{\mathbf{c}}(\cdot)$, could minimize the following objective:

$$L_{\text{SDS}} = \mathbb{E}_{t,c}[D_{\text{KL}}(q_t^{\theta}(\boldsymbol{x}_t \mid \mathbf{c}) \| p_t(\boldsymbol{x}_t \mid y))], \tag{1}$$

where y is the text prompt and $\boldsymbol{x}_t$ is constructed by adding Gaussian noise $\boldsymbol{\epsilon}$ to $\boldsymbol{x}$ according to a specific timestep t and variance schedules. Note that the gradient of Eq. (1) could be approximated by calculating the noise discrepancy as:

$$\nabla_{\theta} L_{\text{SDS}} = \nabla_{\theta} \mathbb{E}_{t,\boldsymbol{\epsilon},c}\left[\omega(t) \| \boldsymbol{\epsilon}_{\text{pretrain}}(\boldsymbol{x}_t, t, y) - \boldsymbol{\epsilon} \|^2\right], \tag{2}$$

where $\omega(t)$ is a weighting function. Notably, Eq. (2) resembles the diffusion training loss, thus intuitively SDS could be explained as optimizing the NeRF renderings to look similar to samples generated from a pre-trained diffusion model.

The inherent mode-seeking characteristic of SDS [85] often leads to sub-optimal generation quality. To address this, VSD [117] proposed to replace the single NeRF parametrization θ from Eq. (1) with a 3D distribution $\mu(\theta \mid y)$ as follows:

$$L_{\text{VSD}} = \mathbb{E}_{t,\mathbf{c}}[D_{\text{KL}}(q_t^{\mu}(\boldsymbol{x}_t \mid \mathbf{c}) \| p_t(\boldsymbol{x}_t \mid y))], \tag{3}$$

where $\mu(\theta \mid y)$ is parametrized with a set of NeRF instances, referred to as *particles*. Additionally, VSD relies on a dynamically fine-tuned Low-Rank Adaptation (LoRA) [41] to capture the conditional rendering distribution across different camera poses, thus enabling a similar optimization framework under the guidance of denoising score.

3D Distribution Modeling. Turning from single NeRF optimization to distribution modeling is a key for improved 3D synthesis. However, due to the large computation and memory requirements of VSD [117], the number of NeRF instances that VSD could jointly optimize is quite small (< 10), hence the 3D distribution modeled by [117] remains discrete and unexpressive. In practice, we observe VSD still suffers from similar issues as SDS to a certain degree.

Instead, we propose to leverage a generator $G(\boldsymbol{z})$ to capture the continuous 3D distribution. Considering the training efficiency and stability, we follow common practice [7, 13, 27] and implement our generator using a StyleGAN2 [50] backbone paired with triplane features. In particular, starting from a latent code $\mathbf{z} \in \mathbb{R}^{d_z}$ drawn from a unit Gaussian distribution, our generator network generates a triplane feature representation $\mathbf{P}$ as follows:

$$G : \mathbb{R}^{d_z} \to \mathbb{R}^{3\times N\times N\times C}, \quad G(\mathbf{z}) \to \mathbf{P}, \tag{4}$$

where N, C correspond to the spatial resolution and channel size respectively. Note that using a generator to approximate the conditional distribution of the diffusion model fundamentally resolves mode-seeking issues. Additionally, instead of learning the triplane features directly from $\mathbf{z}$, the generator also incorporates a non-linear mapping network, which maps $\mathbf{z}$ to latent vectors $\mathbf{w} \in \mathcal{W}$, controlling the generator through adaptive instance normalization (AdaIN) [43] at each convolution layer. The intermediate latent space $\mathcal{W}$ shares many desirable properties, enabling us to perform continuous 3D interpolation and single-view reconstruction by inversion.

3.2. Consistent Adversarial Distillation

Adversarial Distillation. It should be noted the actual optimization goal of score distillation is

$$\min_{\mu} D_{\mathrm{KL}}\left(q_0^{\mu}\left(\boldsymbol{x}_0 \mid c\right) \| p_0\left(\boldsymbol{x}_0 \mid y\right)\right). \tag{5}$$

SDS tackles Eq. (5) by breaking it down into multiple sub-optimization problems, each associated with a unique diffused distribution indexed by t, as indicated in Eq. (1) and Eq. (3). However, doing optimization using noisy discrepancy can easily lead to noticeable quality degradation in comparison to the direct denoising sampling approach, as reported in [85, 117].

To achieve high-quality and diversified generation, we introduce adversarial distillation shown in Fig. 2, an innovative approach for optimizing the 3D generator $G(\cdot)$ s.t. its sampled data match the samples from a pre-trained diffusion model, $p_0\left(\boldsymbol{x}_0 \mid y\right)$, as follows:

$$\min_{G} D_{\mathrm{KL}}\left(p^{G}(\boldsymbol{x} \mid \mathbf{z}, \mathbf{c}) \| p_0\left(\boldsymbol{x}_0 \mid y\right)\right), \tag{6}$$

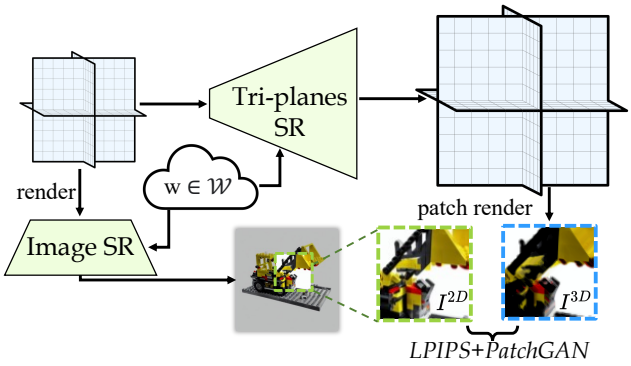


Figure 3. We train a 3D consistent GAN by baking the 2D upsampler to a 3D upsampler through minimizing the patch-level differences between the renderings of the two branches.

where $\boldsymbol{x}$ is the rendered image from triplane $\mathbf{P} = G(\mathbf{z})$ under camera pose $\mathbf{c}$. Instead of directly computing the KL-D, we solve Eq. (6) by learning an implicit likelihood model, namely a GAN [30]. In particular, we train our model using an adversarial objective and adopt the non-saturating GAN loss with R1 regularization [71] from [48] as follows:

$$\begin{aligned} L_{\mathrm{D}} =& \mathbb{E}_{\mathbf{z}\sim p_z, \mathbf{c}\sim p_c}\left[f\left(D(\mathcal{R}(G(\mathbf{z}), \mathbf{c}))\right)\right] + \\ & \mathbb{E}_{I\sim p_0}\left[f\left(-D(I)\right) + \lambda\|\nabla D(I)\|^2\right] \end{aligned} \tag{7}$$

$$L_{\mathrm{G}} = -\mathbb{E}_{\mathbf{z}\sim p_z, \mathbf{c}\sim p_c}\left[f\left(D\left(\mathcal{R}(G(\mathbf{z}), \mathbf{c})\right)\right)\right], \tag{8}$$

where $f(u)$ is defined as $f(u) = -\log(1+\exp(-u))$, I is sampled from the diffusion model, $\mathcal{R}(\cdot)$ denotes the volumetric renderer used to render the generator's output that parametrizes the fake data distribution, $D(\cdot)$ denotes the discriminator and λ is a hyperparameter.

To render an image, for every coordinate along the camera ray, we sample features from the triplanes $\mathbf{P}$ and aggregate them with summation. Next, a light-weight MLP decoder $M(\cdot)$ will explain the aggregated features $\mathbf{f} \in \mathbb{R}^{d_f}$ into radiance $\mathbf{c} \in \mathbb{R}^3$ and volume density $\sigma \in \mathbb{R}^+$. We follow common practice and use volumetric rendering with hierarchical sampling [73] to predict the final RGB colors for every pixel in the image. Unlike score distillation, our distillation scheme is directly based on high-quality and clean samples drawn from the diffusion model, which effectively overcomes oversaturation issues. Furthermore, since adversarial distillation does not minimize the pixel-wise distance to intermediate images or denoising scores that are usually incoherent, our method can reliably generate photorealistic texture details.

Consistency. Directly training a 3D GAN at high resolutions using above-mentioned way is computationally infeasible. Therefore, existing methods [7, 14, 97] rely on image-space convolutions to upscale the resolution of the raw renderings. While 2D upsampling can ensure high-quality and 3D consistent renderings with imperceptible artifacts for forward-facing data [14], we experimentally observe that naively applying this approach to 360° object-level 3D synthesis leads to significant multi-view inconsistencies.

Method	Image Score (↑)		Text Score (↑)	
	ViT-L/14	ViT-B/32	ViT-L/14	ViT-B/32
DreamFusion	65.71	72.61	24.53	28.81
Zero-1-to-3	68.10	77.42	21.35	27.90
ProlificDreamer	72.72	74.46	24.51	29.39
Magic123	75.61	84.02	23.95	**29.89**
Ours	**82.56**	**89.16**	**25.32**	29.81

Table 1. **Quantitative Evaluation.** We measure the CLIP similarity score [87]. For the Image Score we compute the CLIP distance between rendered and reference views, while for the Text Score, we compute the CLIP distance between rendered and text promots. We color each row as best, second best, and third best.

As shown in Fig. 3, we resolve this issue by baking the 2D upsampler branch into a 3D triplane upsampler via patch-wise similarity following [18]. The motivation here is two-fold: 1) By sharing the same latent code $\mathbf{w}$, the image rendered through the 3D upsampler has the potential to capture high-frequency details similar to those produced by the 2D branch output. 2) Once trained, renderings from the same triplane trivially preserve 3D consistency. Therefore, given a latent code $\mathbf{z}$ and a camera pose $\mathbf{c}$, we use the LPIPS [133] loss to optimize the 3D upsampler as:

$$L_{\text{consistency}} = \text{LPIPS}\left(I^{3D}, \text{sg}\left(I^{2D}\right)\right), \tag{9}$$

where I^{2D} and I^{3D} are images generated from the 2D and the 3D upsampling branches respectively and sg denotes stopping gradient. For efficiency we only calculate Eq. (9) in 64^2 patch-level. We also train a small patch discriminator to ensure the renderings of upsampled triplane preserving high-frequency details, using a similar objective as in Eq. (7).

3.3. Multi-View Sampling

In this section, we discuss our strategy for sampling and filtering high-quality images from $p_0(\cdot)$ in order to effectively perform adversarial distillation and calculate Eq. (7).

Sampling. Diffusion models [92, 93] can produce high-quality images, but often exhibit a bias towards generating frontal-facing images, thus failing to adequately span the entire azimuth $\phi \in [0, 2\pi]$ and elevation $\theta \in [0, \pi]$ angles. However, this bias can make the discriminator overfit to particular viewpoints, resulting in inaccurate gradients when optimizing the 3D generator. To resolve this we leverage the view-dependent diffusion model of Zero-1-to-3 [64] to produce more diverse views given a single input image.

Pruning. Although the view-conditioned diffusion model could assist us towards obtaining free-view diffusion priors in 360°, we still observe several issues. For instance, the generated viewpoints may not adhere to the target camera pose, or there may be substantial geometric degradation in the generated images, resulting in visible 3D misalignment compared to the reference image as shown in the supplementary. Although the discriminator can tolerate some errors to a certain extent, the frequent occurrence of misalignment will significantly undermine the stability of GAN training.

To this end, we propose a camera pose pruning strategy aiming at filtering out "bad diffusion samples". Specifically, given a reference image and a target camera pose, we synthesize N samples in parallel with reverse denoising. To capture the essential geometric characteristics of the target pose, we compute a shared overlapping mask m according to the generated samples, which is then dilated to allow a certain degree of geometry deformation. We consider the pixels within the mask as "good", and compute the maximum number of "bad pixels" (i.e. pixels outside the mask) among the generated samples. If this maximum count exceeds a predefined threshold, we discard this viewpoint.

In addition to the geometric consistency check, we also assess the semantic and structural similarity between sampled I^i_{syn} and reference views I_r using the pre-trained CLIP image encoder as follows:

$$\min_{i=1:N} \texttt{CLIP}_{\text{image}}(I^i_{\text{syn}}, I_r). \tag{10}$$

If the score of (10), for a specific view, is below a threshold, we reject this viewpoint. Viewpoints that fail to pass any of these two checks are discarded. Among images from the same view, we select the sample with the highest CLIP score.

Distribution Refinement. Although the generated images, produced by the view-dependent Zero-1-to-3 [64] are more diverse in terms of camera poses, they often lack appearance variations and exhibit blurred textures. To mitigate this, we propose to further refine the prior extracted by [64] with a powerful text-guided 2D diffusion model [1, 132]. Specifically, for the sample generated with [64], we add noise over the image with specific strength and then denoise it by [1] to generate high-quality and diversified refinement, which still retains similar pose and semantics with the input. In general, using a higher noise level typically yields more diverse generations, however this also poses a challenge in maintaining the original pose information. Therefore, we also try ControlNet [132] conditioned on depth maps to achieve better 3D diversity while preserving pose information. We show more analysis in the supplementary.

4. Experimental Evaluation

Datasets. In our evaluation, we consider images from three sources: (i) high-quality real-world images from the Internet, (ii) synthetic scenes from Blender including *Synthetic-NeRF* [73] and *Synthetic-NSVF* [62], and (iii) images generated by a text-to-image diffusion model. For the case of objects from [62, 73], we only utilize a single image and ignore its associated pose information. When no text descriptions are provided, we employ an off-the-shelf image caption pipeline [56] to obtain a per-image text description.

Baselines. We assess the performance of our approach against two significant image-to-3D methods including Magic123 [86] and Zero-1-to-3 [64], as well as two notable text-to-3D techniques, DreamFusion [85] and Prolific-

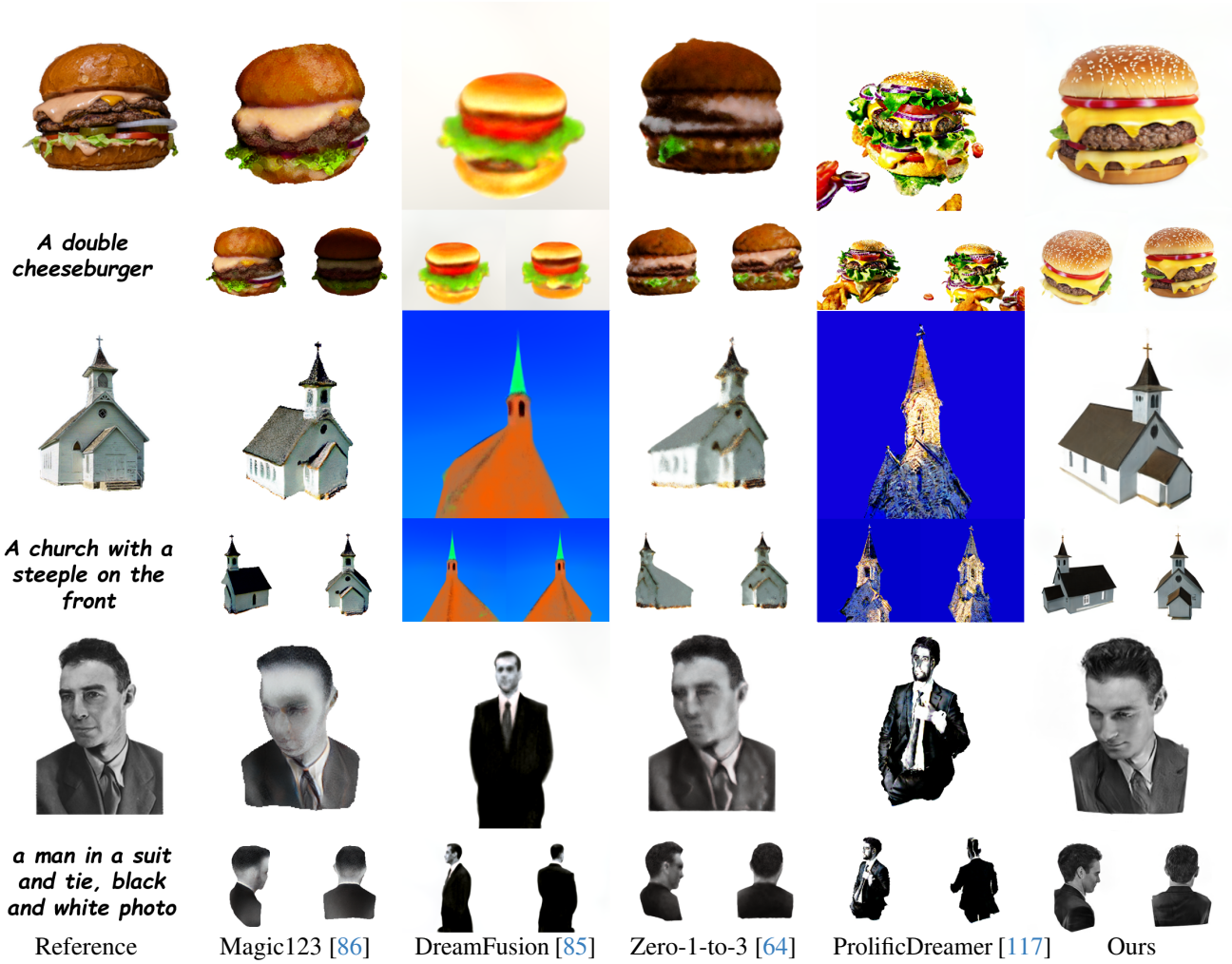


Figure 4. **Qualitative Evaluation.** Our method yields more photo-realistic 3D generations conditioned on a single image (left-most column), with significantly less artifacts compared existing SDS-based pipelines [85, 117]. In comparison to the single-view conditioned 3D generation approaches [64, 86], our generations have significantly fewer artifacts and look more realistic.

Dreamer [117]. We use the improved implementation from ThreeStudio [33] for the baseline training with shared input view, prompt, pose and resolution across all the methods.

Evaluation Metrics. We follow common practice [63, 85, 86] and report the CLIP similarity score [87] with different image-based ViT [25] and text-based architectures trained by OpenAI. Specifically, we create a 360° camera trajectory orbiting the object with 120 frames that cover different views. For each view, we calculate the image and text similarity with a reference image and a text prompt respectively. In our evaluation, we consider a total of $1,200$ images gathered from our data sources. Additional details as well as a user study that evaluates the subjective synthesis quality and diversity are provided in the supplementary.

Implementation Details. In our framework, we adopt the 3D generator architecture of EG3D [14] with a few key modifications tailored to our specific 3D adversarial distillation formulation. First, we remove the pose conditioning from the generator, as we are interested in modeling generic objects. To guide the generator to learn the correct 3D pose prior, we follow [14] and inject the absolute camera pose into the discriminator. During optimization we generate 10K samples per object and enable the adaptive discriminator augmentation (ADA) [49] to stabilize the adversarial training. During the distribution refinement step, we employ a noise strength of 0.8 for the depth-conditioned ControlNet [132] and noise strengths of 0.3 and 0.7 for the low and high-resolution branches of DeepFloyd [1]. Moreover, we leverage the view-dependent prompting to avoid the multi-face issue while dealing with asymmetric objects. To obtain a 3D distribution based on the image and prompt, the training process takes approximately 3 days for images of 256^2 resolution, where training the 2D upsampler branch requires 1.5 days and fine-tuning together with the 3D upsampler

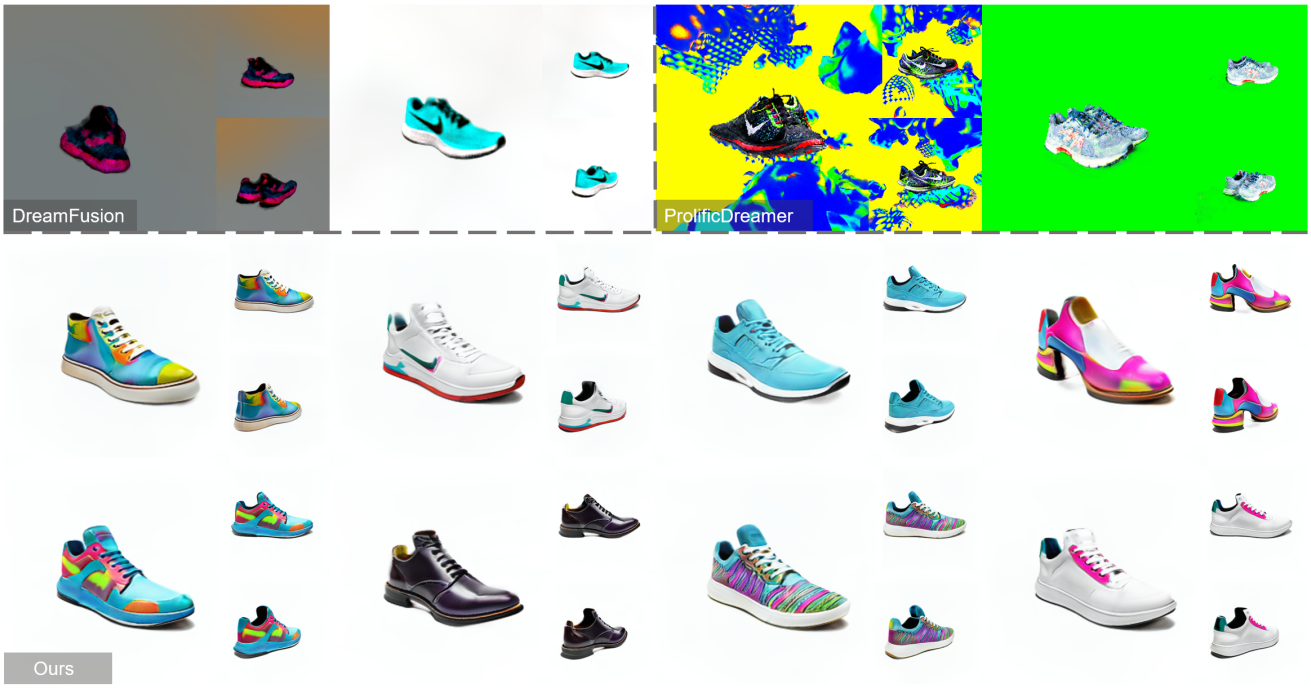


Figure 5. **Diversity and quality comparisons with the baselines.** By distilling the diffusion prior into a 3D generator, our method is better in terms of diversity, quality and photorealism. Note that in contrast to DreamFusion [85] and ProlificDreamer [117] that require re-optimizing for more than 10 hours for each novel sample, our model can generate diverse and novel 3D objects within a second.

requires another 1.5 days on 4×V100 GPUs. Additional implementation details are provided in the supplementary.

4.1. Comparisons

We now compare our adversarial distillation framework with existing score distillation methods and demonstrate that it performs better in terms of generation quality and diversity.

Quantitative Comparison. In Table 1, we compare all methods wrt. their consistency in terms of appearance and semantic similarity. In particular, we note that our method significantly outperforms all baselines in terms of the CLIP image score. This is expected, as our renderings are more photorealistic and do not suffer from over-saturation and over-smoothing issues apparent in existing SDS-based pipelines. In terms of text similarity metrics, our approach again outperforms all baselines when using the ViT-L/14 pre-trained model, while performing on-par with [86] when using the ViT-B/32 pre-trained model. The superiority of our model in both metrics indicates its ability to better match the conditioning signal and generate more high-quality 3D contents.

Qualitative Comparison. In Fig. 4, we qualitatively compare our model with several baselines. Compared to Dreamfusion [85] that generates blurry 3D objects with highly saturated colors, our model produces more photorealistic objects with fine details. ProlificDreamer [117] further proposes VSD, which replaces a single NeRF fitting into distribution matching, demonstrating improved texture details. However, VSD tends to generate multi-face 3D like shown in the second row of Fig. 4. Even for the symmetric objects like burger, its 3D rendering exhibits non realistic colors similar to [85]. In contrast, both Magic123 [86] and Zero-1-to-3 [64] partially mitigate the color shift issues, but still generate objects with blurry appearance. Instead, our adversarial distillation framework attains a higher level of photorealism, while preserving the multi-view consistent fine details.

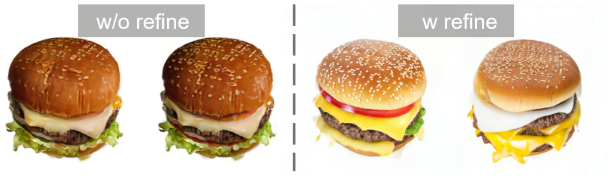
Diversity. A key benefit of our approach is that it can generate diverse 3D objects after distillation. To showcase this, in this section, we compare our model with DreamFusion [85] and ProlificDreamer [117] in terms of their ability to produce diverse renderings conditioned on the text prompt "*running shoes*". For our model, we condition our generation on the same text prompt and an input image showing a white shoe, which we provide in the supplement. Note that currently ProlificDreamer doesn't share the multi-particle implementation thus to obtain the diversity we have to use different seeds for re-optimization. As shown in Fig. 5, even with variations, both DreamFusion [85] and ProlificDreamer [117] produce smooth objects with highly-saturated colors that exhibit Janus-like issues. By contrast, our method obtains significantly better quality and diversity. Moreover, since the the learned 3D generator models the target distribution, our model could generate diverse and novel 3D objects significantly faster than all the baselines without re-optimization.

4.2. Ablation Study

Effectiveness of Pose Pruning. To demonstrate the importance of the camera pose pruning (see Section 3.3), we

Reference – Pruning –Pruning† + Pruning

Figure 6. **Impact of Pose Pruning.** The first column, shows the reference image, the second shows the output when no pose pruning is used, the third shows a variant of our model without pose pruning and ADA [49] but with a larger number of diffusion samples 100K and the last shows the output using the proposed pose pruning.



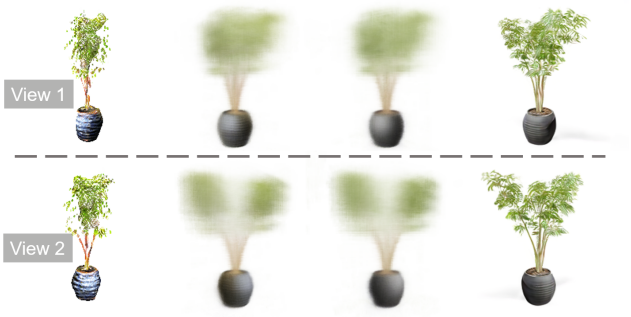
Seed 1 Seed 2 Seed 1 Seed 2

Figure 7. **Impact of Distribution Refinement.** Removing the Distribution Refinement results in inferior diversity and quality.

investigate its impact on the Lego scene from Synthetic-NeRF [73]. We start by removing the camera pose pruning and observe that the the rendering quality deteriorates significantly, as shown in the second column of Fig. 6. To prevent this behavior from being caused by the information leakage of Adaptive Discriminator Augmentation (ADA) [49] , we further increase the diffusion samples from 10K to 100K while also removing ADA [49]. In this scenario, the pose and color seem plausible but the generation quality is still relatively low. In contrast, using camera pose pruning enables the training of a superior 3D GAN, even with limited amount of data, highlighting its efficacy and critical role in enhancing the overall performance of adversarial distillation.

Distribution Refinement. Distribution refinement is a key step to ensure the generation quality and diversity of our distillation pipeline. As shown in Fig. 7, optimizing the 3D generator using the original samples from the view-dependent Zero-1-to-3 [64] without enough diversities could learn to capture a basic shape and appearance but fails to produce high-quality and diverse generations.

Single Mode v.s. Distribution Modeling. We would like to further demonstrate the importance of distribution matching to achieve photorealistic and diversified 3D generation. We now consider the Ficus scene from Synthetic-NeRF [73]. This scene is particularly challenging as it contains thin structures with high-frequency texture details. As shown in Fig. 8, Magic123 [86], which tries to find a single mode under the guidance of score direction, shows poor geometry and rendering quality. Moreover, we also try to fit a single NeRF [73] directly using the raw sampled images from Zero-1-to-3 [64] without any refinement (see second



Magic123 [86] One Mode One Mode† Ours

Figure 8. **Impact of Single Mode Fitting and Distribution Modeling.** We compare our model with Magic123 [86], as well as with two NeRF variants optimized using the raw sampled images [64] (second column) and the refined samples (third column). Our model could yield realistic and 3D consistent renderings.

column) and the improved samples using proposed distribution refinement strategy (see third column). As we could see, both these variants yield low-quality rendering with blurry artifacts. In contrast, our method CAD could render realistic and highly detailed frames with multi-view consistency.

5. Conclusion

In this paper, we proposed CAD, a new approach for generating high-quality, photoreaslisitc and diverse 3D objects conditioned on a single image and a text prompt. Despite the promising results of our model, it still has several limitations. One of the main bottlenecks of our framework is the optimization speed, which is hindered by the inherently slow process of volumetric rendering required to generate high-resolution frames (~ 0.1 FPS). A potential solution to address this limitation is the adoption of more efficient rendering techniques, such as Gaussian Splatting [51]. Moreover, currently we only consider a single conditioning input, joint training with multiple conditions potentially could result in more diverse geometry and appearance variations. Finally, although we mainly focused on the object-level 3D synthesis, extending CAD to scene-level scenarios is also a very promising research direction.

Acknowledgements: We appreciate helpful discussions with Guandao Yang, Boxiao Pan, Zifan Shi, Chao Xu, Xuan Wang, Ivan Skorokhodov, Jingbo Zhang, Xingguang Yan and Connor Zhizhen Lin. We also appreciate the insightful comments and suggestions from the reviewers. The work described in this paper was substantially supported by a GRF grant from the Research Grants Council (RGC) of the Hong Kong Special Administrative Region, China [Project No. CityU11208123]. This research was also supported by an ARL grant W911NF21-2-0104. Despoina Paschalidou is supported by the Swiss National Science Foundation under grant number P500PT 206946. Leonidas Guibas is supported by a Vannevar Bush Faculty Fellowship.

References

[1] Deepfloyd. https://github.com/deep-floyd/IF, 2023. 5, 6

[2] Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas J. Guibas. Learning representations and generative models for 3d point clouds. In *Proc. of the International Conf. on Machine learning (ICML)*, 2018. 2

[3] Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse H. Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matthew Sharifi, Neil Zeghidour, and Christian Havnø Frank. Musiclm: Generating music from text. *arXiv.org*, 2023. 1

[4] Yazeed Alharbi and Peter Wonka. Disentangled image generation through structured noise injection. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[5] Titas Anciukevicius, Zexiang Xu, Matthew Fisher, Paul Henderson, Hakan Bilen, Niloy J. Mitra, and Paul Guerrero. Renderdiffusion: Image diffusion for 3d reconstruction, inpainting and generation. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[6] Sherwin Bahmani, Jeong Joon Park, Despoina Paschalidou, Hao Tang, Gordon Wetzstein, Leonidas J. Guibas, Luc Van Gool, and Radu Timofte. 3d-aware video generation. *arXiv.org*, 2022. 1

[7] Sherwin Bahmani, Jeong Joon Park, Despoina Paschalidou, Xingguang Yan, Gordon Wetzstein, Leonidas J. Guibas, and Andrea Tagliasacchi. CC3D: layout-conditioned generation of compositional 3d scenes. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2023. 4

[8] Jason Bailey. The tools of generative art, from flash to neural networks. *Art in America*, 8, 2020. 2

[9] Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matthew Sharifi, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour. Audiolm: a language modeling approach to audio generation. *arXiv.org*, 2022. 1

[10] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *Proc. of the International Conf. on Learning Representations (ICLR)*, 2019. 2

[11] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[12] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 1

[13] Eric R. Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. Pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 4

[14] 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. 1, 2, 4, 6

[15] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2023. 1, 2

[16] Shu-Yu Chen, Jia-Qi Zhang, You-You Zhao, Paul L Rosin, Yu-Kun Lai, and Lin Gao. A review of image and video colorization: From analogies to deep learning. *Visual Informatics*, 6(3):51–68, 2022. 1

[17] Xi Chen, Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems (NIPS)*, 2016. 2

[18] Xingyu Chen, Yu Deng, and Baoyuan Wang. Mimic3d: Thriving 3d-aware gans via 3d-to-2d imitation. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2023. 5

[19] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[20] Yunjey Choi, Min-Je Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[21] 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. *arXiv.org*, 2023. 2, 3

[22] Congyue Deng, Chiyu Max Jiang, Charles R. Qi, Xinchen Yan, Yin Zhou, Leonidas J. Guibas, and Dragomir Anguelov. Nerdi: Single-view nerf synthesis with language-guided diffusion as general image priors. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[23] Terrance DeVries, Miguel Ángel Bautista, Nitish Srivastava, Graham W. Taylor, and Joshua M. Susskind. Unconstrained scene generation with locally conditioned radiance fields. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2021. 2

[24] Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. Jukebox: A generative model for music. *arXiv.org*, 2020. 1

[25] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proc. of the International Conf. on Learning Representations (ICLR)*, 2021. 6

[26] Matheus Gadelha, Subhransu Maji, and Rui Wang. 3d shape induction from 2d views of multiple objects. In *Proc. of the International Conf. on 3D Vision (3DV)*, 2017. 2

[27] Jun Gao, Tianchang Shen, Zian Wang, Wenzheng Chen, Kangxue Yin, Daiqing Li, Or Litany, Zan Gojcic, and Sanja Fidler. GET3D: A generative model of high quality 3d textured shapes learned from images. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 1, 4

[28] Lin Gao, Jie Yang, Tong Wu, Yu-Jie Yuan, Hongbo Fu, Yu-Kun Lai, and Hao Zhang. SDM-NET: deep generative network for structured deformable mesh. In *ACM SIGGRAPH Conference and Exhibition on Computer Graphics and Interactive Techniques in Asia (SIGGRAPH Asia)*, 2019. 2

[29] Lin Gao, Tong Wu, Yu-Jie Yuan, Ming-Xian Lin, Yu-Kun Lai, and Hao (Richard) Zhang. TM-NET: deep generative networks for textured meshes. *ACM Trans. on Graphics*, 2021. 2

[30] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NIPS)*, 2014. 2, 4

[31] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. Stylenerf: A style-based 3d aware generator for high-resolution image synthesis. In *Proc. of the International Conf. on Learning Representations (ICLR)*, 2022. 2

[32] Jiatao Gu, Alex Trevithick, Kai-En Lin, Joshua M. Susskind, Christian Theobalt, Lingjie Liu, and Ravi Ramamoorthi. Nerfdiff: Single-image view synthesis with nerf-guided distillation from 3d-aware diffusion. In *Proc. of the International Conf. on Machine learning (ICML)*, 2023. 3

[33] 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

[34] Jun Han and Chaoli Wang. Vcnet: A generative model for volume completion. *Visual Informatics*, 6(2):62–73, 2022. 1

[35] Zekun Hao, Arun Mallya, Serge J. Belongie, and Ming-Yu Liu. Gancraft: Unsupervised 3d neural rendering of minecraft worlds. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2021. 2

[36] Ayaan Haque, Matthew Tancik, Alexei A. Efros, Aleksander Holynski, and Angjoo Kanazawa. Instruct-nerf2nerf: Editing 3d scenes with instructions. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[37] Paul Henderson and Vittorio Ferrari. Learning single-image 3d reconstruction by generative modelling of shape, pose and shading. *International Journal of Computer Vision (IJCV)*, 2019. 2

[38] Paul Henderson and Christoph H. Lampert. Unsupervised object-centric video generation and decomposition in 3d. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2

[39] Philipp Henzler, Niloy J Mitra, , and Tobias Ritschel. Escaping plato's cave: 3d shape from adversarial rendering. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2019. 2

[40] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[41] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *Proc. of the International Conf. on Learning Representations (ICLR)*, 2022. 4

[42] Ian Huang, Vrishab Krishna, Omoruyi Atekha, and Leonidas Guibas. Aladdin: Zero-shot hallucination of stylized 3d assets from abstract scene descriptions. *arXiv preprint arXiv:2306.06212*, 2023. 3

[43] Xun Huang and Serge J. Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2017. 4

[44] Moritz Ibing, Gregor Kobsik, and Leif Kobbelt. Octree transformer: Autoregressive 3d shape generation on hierarchically structured sequences. *arXiv.org*, 2021. 2

[45] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[46] Heewoo Jun and Alex Nichol. Shap-e: Generating conditional 3d implicit functions. *arXiv.org*, 2023. 1

[47] Animesh Karnewar, Niloy J. Mitra, Andrea Vedaldi, and David Novotný. Holofusion: Towards photo-realistic 3d generative modeling. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2023. 3

[48] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 4

[49] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 6, 8

[50] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. 2020. 1, 2, 4

[51] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. on Graphics*, 2023. 8

[52] Hyunsu Kim, Gayoung Lee, Yunjey Choi, Jin-Hwa Kim, and Jun-Yan Zhu. 3d-aware blending with generative nerfs. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2023. 2

[53] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *Proc. of the International Conf. on Learning Representations (ICLR)*, 2014. 2

[54] Juil Koo, Seungwoo Yoo, Minh Hieu Nguyen, and Minhyuk Sung. SALAD: part-level latent diffusion for 3d shape generation and manipulation. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2023. 2

[55] Jeong-gi Kwak, Yuanming Li, Dongsik Yoon, Donghyeon Kim, David K. Han, and Hanseok Ko. Injecting 3d perception of controllable nerf-gan into stylegan for editable portrait image synthesis. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2022. 2

[56] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proc. of the International Conf. on Machine learning (ICML)*, 2022. 5

[57] Tiemeng Li, Songqian Wu, Yanning Jin, Haopai Shi, and Shiran Liu. X-space: Interaction design of extending mixed reality space from web2d visualization. *Visual Informatics*, 7(4):73–83, 2023. 1

[58] Tingting Liao, Hongwei Yi, Yuliang Xiu, Jiaxiang Tang, Yangyi Huang, Justus Thies, and Michael J. Black. Tada! text to animatable digital avatars. In *Proc. of the International Conf. on 3D Vision (3DV)*, 2023. 2

[59] Yiyi Liao, Katja Schwarz, Lars M. Mescheder, and Andreas Geiger. Towards unsupervised learning of generative models for 3d controllable image synthesis. *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[60] 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 *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 2

[61] Huan Ling, Karsten Kreis, Daiqing Li, Seung Wook Kim, Antonio Torralba, and Sanja Fidler. Editgan: High-precision semantic image editing. In *Advances in Neural Information Processing Systems (NIPS)*, 2021. 2

[62] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 5

[63] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Zexiang Xu, Hao Su, et al. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. *arXiv preprint arXiv:2306.16928*, 2023. 3, 6

[64] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9298–9309, 2023. 2, 3, 5, 6, 7, 8

[65] Steven Liu, Xiuming Zhang, Zhoutong Zhang, Richard Zhang, Jun-Yan Zhu, and Bryan Russell. Editing conditional radiance fields. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2021. 2

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

[67] Fei Luo, Yongqiong Zhu, Yanping Fu, Huajian Zhou, Zezheng Chen, and Chunxia Xiao. Sparse rgb-d images create a real thing: a flexible voxel based 3d reconstruction pipeline for single object. *Visual Informatics*, 7(1):66–76, 2023. 1

[68] Luke Melas-Kyriazi, Iro Laina, Christian Rupprecht, and Andrea Vedaldi. Realfusion 360° reconstruction of any object from a single image. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[69] Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik P. Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1

[70] Quan Meng, Anpei Chen, Haimin Luo, Minye Wu, Hao Su, Lan Xu, Xuming He, and Jingyi Yu. Gnerf: Gan-based neural radiance field without posed camera. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2021. 2

[71] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? In *Proc. of the International Conf. on Machine learning (ICML)*, 2018. 4

[72] Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. Latent-nerf for shape-guided generation of 3d shapes and textures. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[73] 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 *Proc. of the European Conf. on Computer Vision (ECCV)*, 2020. 2, 3, 4, 5, 8

[74] Eric T. Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Görür, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? In *Proc. of the International Conf. on Learning Representations (ICLR)*, 2019. 2

[75] Charlie Nash, Yaroslav Ganin, S. M. Ali Eslami, and Peter W. Battaglia. Polygen: An autoregressive generative model of 3d meshes. In *Proc. of the International Conf. on Machine learning (ICML)*, 2020. 2

[76] Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. Hologan: Unsupervised learning of 3d representations from natural images. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2019. 2

[77] Thu Nguyen-Phuoc, Christian Richardt, Long Mai, Yong-Liang Yang, and Niloy Mitra. Blockgan: Learning 3d object-aware scene representations from unlabelled images. *arXiv.org*, 2020. 2

[78] Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. Point-e: A system for generating 3d point clouds from complex prompts. *arXiv.org*, 2022. 1

[79] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: towards photorealistic image generation and editing with text-guided diffusion models. In *Proc. of the International Conf. on Machine learning (ICML)*, 2022. 1, 3

[80] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[81] Michael Niemeyer and Andreas Geiger. CAMPARI: camera-aware decomposed generative neural radiance fields. In *Proc. of the International Conf. on 3D Vision (3DV)*, 2021. 2

[82] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[83] OpenAI. GPT-4 technical report. *arXiv.org*, 2023. 1

[84] Xingang Pan, Xudong Xu, Chen Change Loy, Christian Theobalt, and Bo Dai. A shading-guided generative implicit model for shape-accurate 3d-aware image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2

[85] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 1, 2, 3, 4, 5, 6, 7

[86] Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov, et al. Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. *arXiv preprint arXiv:2306.17843*, 2023. 3, 5, 6, 7, 8

[87] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proc. of the International Conf. on Machine learning (ICML)*, 2021. 5, 6

[88] Sameera Ramasinghe, Salman H. Khan, Nick Barnes, and Stephen Gould. Spectral-gans for high-resolution 3d point-cloud generation. In *Proc. IEEE International Conf. on Intelligent Robots and Systems (IROS)*, 2020. 2

[89] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *Proc. of the International Conf. on Machine learning (ICML)*, 2021. 1

[90] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *arXiv.org*, 2022. 1, 2, 3

[91] Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *Proc. of the International Conf. on Machine learning (ICML)*, 2015. 2

[92] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2, 3, 5

[93] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2, 3, 5

[94] Axel Sauer, Tero Karras, Samuli Laine, Andreas Geiger, and Timo Aila. Stylegan-t: Unlocking the power of gans for fast large-scale text-to-image synthesis. 2023. 1

[95] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2

[96] Katja Schwarz, Axel Sauer, Michael Niemeyer, Yiyi Liao, and Andreas Geiger. Voxgraf: Fast 3d-aware image synthesis with sparse voxel grids. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2

[97] Bokui Shen, Xinchen Yan, Charles R. Qi, Mahyar Najibi, Boyang Deng, Leonidas J. Guibas, Yin Zhou, and Dragomir Anguelov. GINA-3D: learning to generate implicit neural assets in the wild. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023. 4

[98] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[99] Zifan Shi, Yinghao Xu, Yujun Shen, Deli Zhao, Qifeng Chen, and Dit-Yan Yeung. Improving 3d-aware image synthesis with A geometry-aware discriminator. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2

[100] Zifan Shi, Yujun Shen, Yinghao Xu, Sida Peng, Yiyi Liao, Sheng Guo, Qifeng Chen, and Dit-Yan Yeung. Learning 3d-aware image synthesis with unknown pose distribution. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[101] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data. *arXiv.org*, 2022. 1

[102] Ivan Skorokhodov, Sergey Tulyakov, and Mohamed Elhoseiny. Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1

[103] Ivan Skorokhodov, Sergey Tulyakov, Yiqun Wang, and Peter Wonka. Epigraf: Rethinking training of 3d gans. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2

[104] Ivan Skorokhodov, Aliaksandr Siarohin, Yinghao Xu, Jian Ren, Hsin-Ying Lee, Peter Wonka, and Sergey Tulyakov. 3d generation on imagenet. In *Proc. of the International Conf. on Learning Representations (ICLR)*, 2023. 2

[105] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proc. of the International Conf. on Machine learning (ICML)*, 2015. 2

[106] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *Proc. of the International Conf. on Learning Representations (ICLR)*, 2021. 2

[107] Kunhua Su, Jun Zhang, Deyue Xie, and Jun Tao. Importance guided stream surface generation and feature exploration. *Visual Informatics*, 7(2):54–63, 2023. 1

[108] Konstantinos Tertikas, Despoina Paschalidou, Boxiao Pan, Jeong Joon Park, Mikaela Angelina Uy, Ioannis Z. Emiris, Yannis Avrithis, and Leonidas J. Guibas. Generating part-aware editable 3d shapes without 3d supervision. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[109] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *arXiv.org*, 2023. 1

[110] Ziyu Wan, Bo Zhang, Dongdong Chen, Pan Zhang, Dong Chen, Jing Liao, and Fang Wen. Bringing old photos back to life. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2747–2757, 2020. 1

[111] Ziyu Wan, Bo Zhang, Dongdong Chen, and Jing Liao. Bringing old films back to life. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17694–17703, 2022. 1

[112] Ziyu Wan, Christian Richardt, Aljaž Božič, Chao Li, Vijay Rengarajan, Seonghyeon Nam, Xiaoyu Xiang, Tuotuo Li, Bo Zhu, Rakesh Ranjan, et al. Learning neural duplex radiance fields for real-time view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8307–8316, 2023. 2

[113] Binxu Wang and Carlos R. Ponce. A geometric analysis of deep generative image models and its applications. In *Proc. of the International Conf. on Learning Representations (ICLR)*, 2021. 2

[114] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A. Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[115] Jun Wang, Bohan Lei, Liya Ding, Xiaoyin Xu, Xianfeng Gu, and Min Zhang. Autoencoder-based conditional optimal transport generative adversarial network for medical image generation. *Visual Informatics*, 2023. 1

[116] Yujie Wang, Yixin Zhuang, Yunzhe Liu, and Baoquan Chen. Mdisn: Learning multiscale deformed implicit fields from single images. *Visual Informatics*, 6(2):41–49, 2022. 1

[117] 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. 2023. 2, 3, 4, 6, 7

[118] Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models. In *Proc. of the International Conf. on Learning Representations (ICLR)*, 2023. 3

[119] Jiajun Wu, Chengkai Zhang, Tianfan Xue, Bill Freeman, and Josh Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In *Advances in Neural Information Processing Systems (NIPS)*, 2016. 2

[120] Shangzhe Wu, Ruining Li, Tomas Jakab, Christian Rupprecht, and Andrea Vedaldi. Magicpony: Learning articulated 3d animals in the wild. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[121] Yinghao Xu, Sida Peng, Ceyuan Yang, Yujun Shen, and Bolei Zhou. 3d-aware image synthesis via learning structural and textural representations. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[122] Guandao Yang, Xun Huang, Zekun Hao, Ming-Yu Liu, Serge J. Belongie, and Bharath Hariharan. Pointflow: 3d point cloud generation with continuous normalizing flows. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2019. 2

[123] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Ronen Basri, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2

[124] Liang Yuan and Issei Fujishiro. Multiview svbrdf capture from unified shape and illumination. *Visual Informatics*, 7 (3):11–21, 2023. 1

[125] Liang Yuan, Dingkun Yan, Suguru Saito, and Issei Fujishiro. Diffmat: Latent diffusion models for image-guided material generation. *Visual Informatics*, 2024. 1

[126] Yu-Jie Yuan, Yang-Tian Sun, Yu-Kun Lai, Yuewen Ma, Rongfei Jia, and Lin Gao. Nerf-editing: Geometry editing of neural radiance fields. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[127] Xiaohui Zeng, Arash Vahdat, Francis Williams, Zan Gojcic, Or Litany, Sanja Fidler, and Karsten Kreis. LION: latent point diffusion models for 3d shape generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2

[128] Biao Zhang, Matthias Nießner, and Peter Wonka. 3dilg: Irregular latent grids for 3d generative modeling. In *NeurIPS*, 2022. 2

[129] Biao Zhang, Jiapeng Tang, Matthias Nießner, and Peter Wonka. 3dshape2vecset: A 3d shape representation for neural fields and generative diffusion models. 2023. 2

[130] Jingbo Zhang, Xiaoyu Li, Ziyu Wan, Can Wang, and Jing Liao. Fdnerf: Few-shot dynamic neural radiance fields for face reconstruction and expression editing. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–9, 2022. 1

[131] Jingbo Zhang, Xiaoyu Li, Ziyu Wan, Can Wang, and Jing Liao. Text2nerf: Text-driven 3d scene generation with neural radiance fields. *arXiv preprint arXiv:2305.11588*, 2023.

[132] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2023. 1, 2, 5, 6

[133] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 5

[134] Yuxuan Zhang, Wenzheng Chen, Huan Ling, Jun Gao, Yinan Zhang, Antonio Torralba, and Sanja Fidler. Image gans meet differentiable rendering for inverse graphics and interpretable 3d neural rendering. In *Proc. of the International Conf. on Learning Representations (ICLR)*, 2021. 2

[135] Yue Zhang, Zhenyuan Wang, Jinhui Zhang, Guihua Shan, and Dong Tian. A survey of immersive visualization: Focus on perception and interaction. *Visual Informatics*, 7(4):22–35, 2023. 1

[136] Peng Zhou, Lingxi Xie, Bingbing Ni, and Qi Tian. Cips-3d: A 3d-aware generator of gans based on conditionally-independent pixel synthesis. *arXiv.org*, abs/2110.09788, 2021. 2

[137] Zhizhuo Zhou and Shubham Tulsiani. Sparsefusion: Distilling view-conditioned diffusion for 3d reconstruction. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3