

ImageDream: Image-Prompt Multi-view Diffusion for 3D Generation

Peng Wang Yichun Shi

ByteDance, USA

{peng.wang, yichun.shi}@bytedance.com

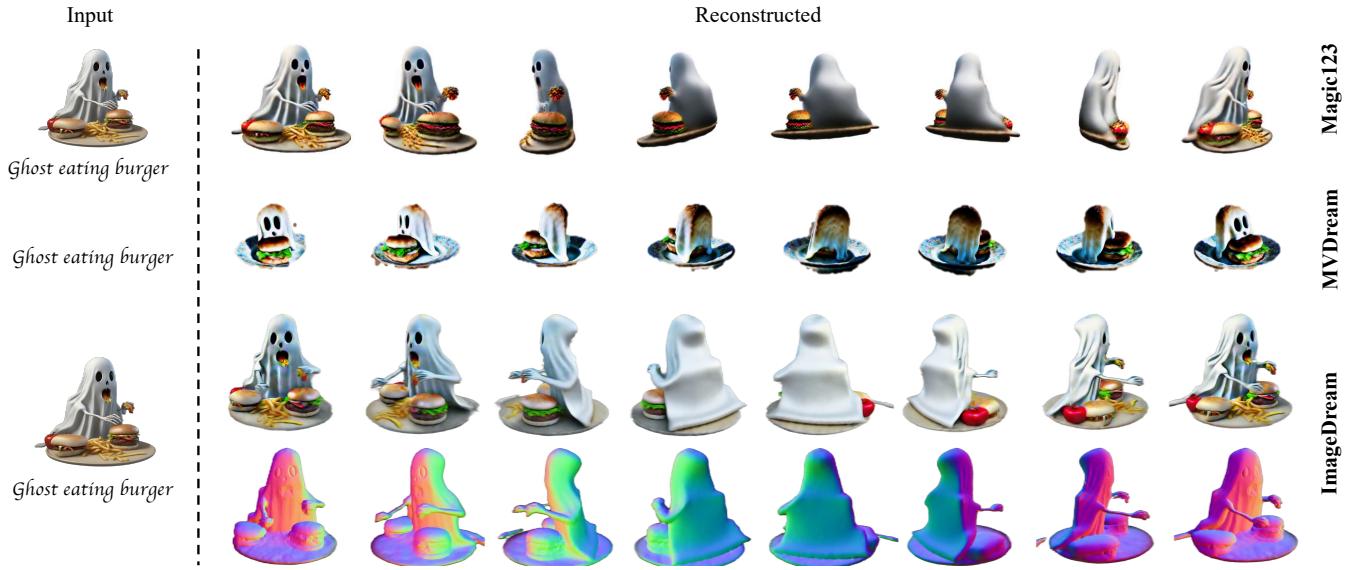


Figure 1. **ImageDream** is a novel framework that generates high quality 3D model from any viewpoint given a single image. It vastly improves the 3D geometry quality comparing to previous SoTA, e.g. Magic123 [31], and more importantly, it inherits the great text image alignment from the generated image-prompt, comparing with MVDream [37]. Here, we provide 8 views of a generated object from different methods, and in the last row, we show the corresponding normal maps rendered with ImageDream generated model.

Abstract

We introduce *ImageDream*, a novel Image-Prompt Multi-view diffusion model devised for 3D object generation. *ImageDream* excels in delivering 3D models of superior quality comparing with other State-of-the-Art (SoTA) image-conditioned endeavors. Specifically, we consider a canonical camera coordination of the object in image rather than relative. This enhancement significantly augments the visual geometry correctness. Our models are meticulously crafted, taking into account varying degrees of control granularity derived from the provided image: wherein, the global control predominantly influences the object layout, whereas the local control adeptly refines the image appearance. The prowess of *ImageDream* is empirically showcased through a comprehensive evaluation predicated on a common prompt list as delineated in MVDream [37]. Our project page is <https://Image-Dream.github.io>.

1. Introduction

In the domain of 3D generation, incorporating images as an additional modality for 3D generation, compared to methods relying solely on text [30], offers significant advantages, as the common saying, *An image is worth a thousand words*. Primarily, images convey rich, precise visual information that text might ambiguously describe or entirely omit. For instance, subtle details like textures, colors, and spatial relationships can be directly and unambiguously captured in an image, whereas a text description might struggle to convey the same level of detail comprehensively or might require excessively lengthy descriptions. This visual specificity aids in generating more accurate and detailed 3D models, as the system can directly reference actual visual cues rather than interpret textual descriptions, which can vary greatly in detail and subjectivity. Moreover, using images allows for a more intuitive and direct way for users to communicate their desired outcomes, particularly for those who may find it challenging to articulate their visions textually. This multimodal approach, combining the richness of vi-

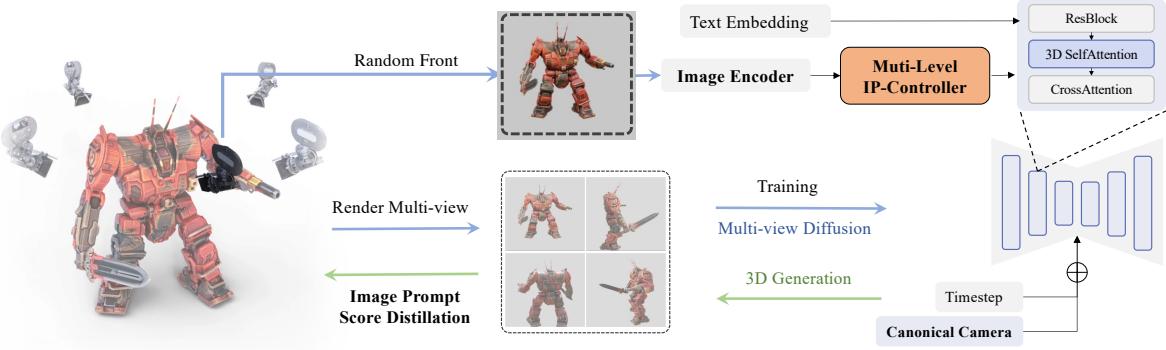


Figure 2. The training pipeline of ImageDream. The blue arrow indicates training of the diffusion network and the green arrow indicates training of NeRF model. In diffusion training, given a 3D object, we first render multiple views based on canonical camera coordination (bottom), and render another image-prompt front-view images with a random setting (top). The multi-view images are fed as training targets for multi-view diffusion networks, and image-prompt is encoded with a multi-level controller as input to the diffusion. In NeRF training, we use the trained diffusion for image-prompt score distillation.

sual data with the contextual depth of text, leads to a more robust, user-friendly, and efficient 3D generation process, catering to a wider range of creative and practical applications.

Adopting images as an additional modality for 3D object generation, while beneficial, also introduces several challenges. Unlike text, images contain a multitude of features like color, texture, spatial relationships that are more complex to analyze and interpret accurately with a solely encoder like CLIP [32]. In addition, high variant of light, shape or self-occlusion of the object can lead to inaccurate and in-consistent view synthesis, therefore leading blurry or incomplete 3D models.

The complexity of image processing necessitates advanced, computationally intensive algorithms to accurately decode visual information and ensure consistent appearance across multiple views. Researchers have employed various strategies with diffusion models, such as Zero123 [19], and other recent works [20, 31], to elevate a 2D object image to a 3D model. However, a limitation of image-only solutions is that, although the synthesized views are visually impressive, the reconstructed models often lack geometric accuracy and detailed textures, particularly in the object’s rear views. This issue primarily stems from significant geometric inconsistencies across the generated or synthesized views. Consequently, during reconstruction, non-matching pixels are averaged in the final 3D model, leading to indistinct textures and smoothed geometry.

Fundamentally, image-conditioned 3D generation represents an optimization problem with more stringent constraints compared to text-conditioned generation. Hence, achieving optimized 3D models with clear details is more challenging, as the optimization process is prone to deviating from the trained distributions due to the limited amount of 3D data. For example, generating a horse based solely on text descriptions may yield detailed models if the training

dataset includes a variety of horse styles. However, when an image specifies particular textures, shapes, and fur details, the novel-view texture generation may easily deviate from the trained distributions.

In this paper, we introduce ImageDream to address these challenges. Our approach involves considering a canonical camera coordination across different object instances and designing a multi-level image-prompt controller that can be seamlessly integrated into the existing architecture. Specifically, the canonical camera coordination mandates that the rendered image, under default camera settings (i.e., identity rotation and zero translation), represents the object’s centered front-view. This significantly simplifies the task of mapping variations in the input image to 3D. The multi-level controller offers hierarchical control, guiding the diffusion model from the image input to each architectural block, thereby streamlining the path of information transfer.

As illustrated in Fig.1, ImageDream excels in generating objects with correct geometry from a given image, enabling users to leverage well-developed image generation models[29] for better image-text alignment than purely text-conditioned models like MVDream [37]. Furthermore, ImageDream surpasses existing state-of-the-art (SoTA) zero-shot single image 3D model generators, such as Magic123 [31], in terms of geometry and texture quality. Our comprehensive evaluation in the experimental section (Sec. 4), which includes both qualitative comparisons through user studies and quantitative analyses, demonstrates ImageDream’s superiority over other SoTA methods.

2. Related Works

We recognize that 3D generation is a well-established field; this review focuses on significant advancements closely related to our research.

Text-to-3D Generation with Diffusion. The emergence of

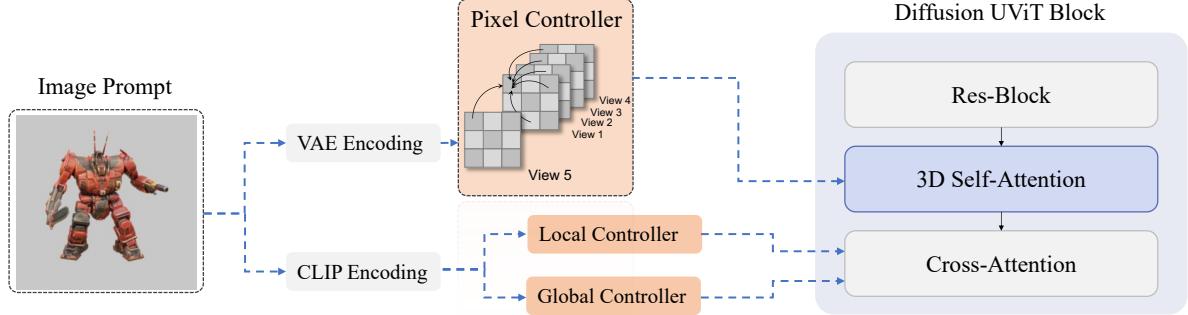


Figure 3. The Multi-level Controller of ImageDream. Given an image prompt, global controller and local controller takes input of image features after CLIP encoding, and then output adapted features to cross-attention layers. It represents image semantic information. The pixel controller send the VAE encoded feature to diffusion, and perform pixel-level dense self-attention with corresponding hidden features at each layer of the four-view MVDiffusion.

deep generative models has significantly impacted 3D generation. Early methods targeted the reconstruction of simple objects using multi-view rendered images [11, 12]. The evolution of these techniques, from Generative Adversarial Networks (GANs)[3, 9, 10, 24, 25, 27] to diffusion-based frameworks[33], marks a notable progression.

Recent 3D diffusion models, specifically for triplanes [38, 45] and feature grids [15], have emerged. However, these models often focus on specific objects like faces and ShapeNet [5] objects. Concurrently, there’s growing interest in reconstructing object shapes from monocular image inputs [14, 26, 48], demonstrating the evolving stability of image generation methodologies. A significant challenge remains in generalizing these models to the extent of their 2D counterparts, likely due to constraints in 3D data size, representation, and architectural design.

Lifting 2D Diffusion for 3D Generation. In light of the limited generalizability of direct 3D generative models, a parallel line of research has explored the elevation of 2D diffusion priors into 3D generation, often integrating with 3D representations like NeRF [23]. A pivotal approach in this area is the score distillation sampling (**SDS**) introduced by Poole et al.[30], using diffusion priors as score functions to guide 3D representation optimization. Alongside Dreamfusion, works like SJC[44], which utilize stable-diffusion models [33], have emerged. Subsequent studies have focused on enhancing 3D representations [6, 17, 42, 43], refining sampling schedules [13], and optimizing loss designs [46]. Despite their ability to generate photorealistic objects of various types without 3D data training, these methods struggle with multi-view consistency. Moreover, each 3D model requires individualized optimization through prompt and hyper-parameter adjustments. Notably, MVDream [37] enhances generation robustness by joint training with 2D and 3D datasets, producing satisfactory results with uniform parameters, drawing on multi-view diffusion via SDS. Our work builds upon these concepts, applying them to image-prompt generation and retaining the robustness characteristic of MVDream.

Image-based Novel View Synthesis. Direct synthesis of novel 3D views from single images has also been explored, bypassing traditional reconstruction processes. Watson et al.[47] pioneered diffusion model applications in view synthesis as the pipeline in Sitzmann et al.[39] using the ShapeNet dataset. Subsequent advancements include Zhou et al.’s [53] extension to latent space with an epipolar feature transformer and Chan et al.’s [4] approach to enhance view consistency. Szymanowicz et al.[41] proposed a multi-view reconstructor using unprojected feature grids. A common limitation across these methods is their dependency on specific training data, with no established adaptability to diverse image inputs. Fine-tuning pre-trained image diffusion models[2] on extensive 3D render datasets for novel view synthesis, as proposed by Zero123 [19], remains constrained by geometric consistency issues. Later works, including SyncDreamer [20], Consistent 1-to-3 [52], and Zero123plus [36], have sought to enhance multi-view consistency through joint diffusion processes, but the reconstruction of geometrically coherent 3D models remains a challenge.

Single Image-conditioned Reconstruction. Recent advances in deriving 3D models from single or few images predominantly leverage NeRF representations. Techniques such as RegNeRF [28], which uses geometry loss from depth patches, and SinNeRF [49], RealFusion [22], and NeuralLift [50], which combine depth maps or Score Distillation Sampling during NeRF training, represent significant steps forward. Despite their effectiveness, the quality of these generated models remains suboptimal for real-world applications. Magic123 [31] combines single-view and novel-view diffusion networks, achieving impressive texture quality in 3D models. However, our tests reveal limitations in understanding correct object geometry.

We also note recent parallel developments, such as Wonder3D [21], which incorporate normal diffused outputs into original diffusion models, and DreamCraft3D [40], which employ a second-stage DreamBooth-like model fine-tuning for enhanced texture modeling. These works, while promis-

ing, remain distinct from our contributions.

3. Methodology

In this section, we first talk about the MVDream [37] pipeline and then describe our method to input the image prompt.

3.1. Preliminary

In MVDream, there are two stage for 3D model production. The first stage is training a multi-view diffusion network that produces four orthogonal and consistent multi-view images from a text-prompt given respective camera embedding. In the second stage, a multi-view score distillation sampling (MV-SDS) is adopted to produce a detailed 3D NeRF model.

In the first stage, each block of the multi-view network contains a densely connected 3D attention on the four view images, which allows a strong interaction in learning the correspondence relationship between different views. To train such a network, it adopts a joint training with the rendered dataset from the Objaverse [8] and a larger scale text-to-image (t2i) dataset, LAION5B [35], to maintain the generalizability of the fine-tuned model. Formally, given text-image dataset $\mathcal{X} = \{\mathbf{x}, y\}$ and a multi-view dataset $\mathcal{X}_{mv} = \{\mathbf{x}_{mv}, y, \mathbf{c}_{mv}\}$, where \mathbf{x} is an latent image embedding from VAE [16], y is a text embedding from CLIP [32], and \mathbf{c} is their self-designed camera embedding, we may formulate the the multi-view (MV) diffusion loss as,

$$\mathcal{L}_{MV}(\theta, \mathcal{X}, \mathcal{X}_{mv}) = \mathbb{E}_{\mathbf{x}, y, \mathbf{c}, t, \epsilon} [\|\epsilon - \epsilon_\theta(\mathbf{x}^p; y, \mathbf{c}^p, t)\|_2^2] \quad (1)$$

$$\text{where, } (\mathbf{x}^p, \mathbf{c}^p) = \begin{cases} (\mathbf{x}, \mathbf{0}) & \text{with probability } p \\ (\mathbf{x}_{mv}, \mathbf{c}_{mv}) & \text{with probability } 1 - p \end{cases}$$

here, \mathbf{x} is the noisy latent image generated from a random noise ϵ and image latent, the ϵ_θ is the multi-view diffusion (MVDiffusion) model parametrized by θ .

After the model is trained, the MVDiffusion model can be inserted to the DreamFusion [30] pipeline, where the authors adopt a score-distillation sampling (SDS) based on the four generated views. Specifically, in each iteration step, a random 4 orthogonal views are rendered from a NeRF $g(\phi)$ with a random 4 view camera extrinsic and intrinsic \mathbf{c} . Then, they are encoded to latents \mathbf{x}^{mv} and inserted to the multi view diffusion network to compute a diffusion loss in the image space which is back propagated to optimize the NeRF parameters. Formally,

$$\mathcal{L}_{MV-SDS}(\phi, \mathbf{x}^{mv}) = \mathbb{E}_{t, \mathbf{c}, \epsilon} [\|\mathbf{x}^{mv} - \hat{\mathbf{x}}_0^{mv}\|_2^2]. \quad (2)$$

Here, $\hat{\mathbf{x}}_0^{mv}$ s the denoised MV image at timestep 0 from MVDiffusion. After fusion, MVDream shows significant improvement of object geometry correctness without the Janus issues.

3.2. Canonical Camera

In the context of MVDream, a critical observation is the diffusion of multi-view images using a global aligned camera coordination. In other words, the image from a default camera (no azimuth rotation) is always the front view of the object. This is done by asking the CLIP image feature of a view best match the "front view" CLIP text feature embeddings. This alignment facilitates the fusion of diffused images in the fusion step, reducing ambiguity regarding their viewpoints in relation to the provided text prompt.

As emphasized in the introduction, this alignment also reduced the difficulties in learning the accurately reconstructing the geometry of objects. Thereby, in image prompt cases, in contrast to previous image-conditioned approaches like Zero123 [19], which attempt to recover object 3D geometry based on image camera coordination system, ImageDream adopts canonical/world camera coordination as in MVDream. Our diffusion model aims to regress towards the canonical multiple view image of the object as depicted in the image. This approach is expected to yield superior geometric accuracy compared to systems that utilize relative camera coordination.

Formally, for an image of an object rendered from a random viewpoint with a random camera, denoted as \mathbf{x}_r , we create the ImageDream diffusion multi-view (MV) dataset as $\mathcal{X}_{mv} = \{\mathbf{x}_{mv}, y, \mathbf{x}_r, \mathbf{c}_{mv}\}$, where \mathbf{c}_{mv} is the introduced canonical cameras in MVDream. Then, the rest of diffusion loss is the same as Eqn.(1).

3.3. Multi-level Controllers

In order to insert the image prompt to control the output MV images, we consider a multi-level strategy. The overall structure of the multi-level controller from an image prompt can be seen in Fig. 4, and we elaborate the details of each component in the following.

Global Controller. In our initial approach, we integrated global CLIP image features into MVDream, akin to how text features are used, by fine-tuning the model's already well-established training. Recognizing that MVDream is primarily trained on text embeddings, we introduced a multi-layer perceptron (MLP) θ_g , functioning as an adaptor similar to IP-Adaptor [51], following the CLIP image global embedding. This step aims to align image features with text features, ensuring compatibility within the MVDream framework. Specifically, CLIP image encoding encodes image feature to a 1024 vector with a token length of 4, which we named as \mathbf{f}_g . And, θ_g further adapts the image feature to be 1024 as the input to cross-attention.

On the MVDiffusion side, inside of an attention layer l , we add a new set of MLPs, $\theta_{k_g, l}$ and $\theta_{v_g, l}$, that takes the input the adapted features and output its attention key and value matrix, which then aggregated based on the query feature matrix, \mathbf{q}_l , yielding a corresponding image cross-

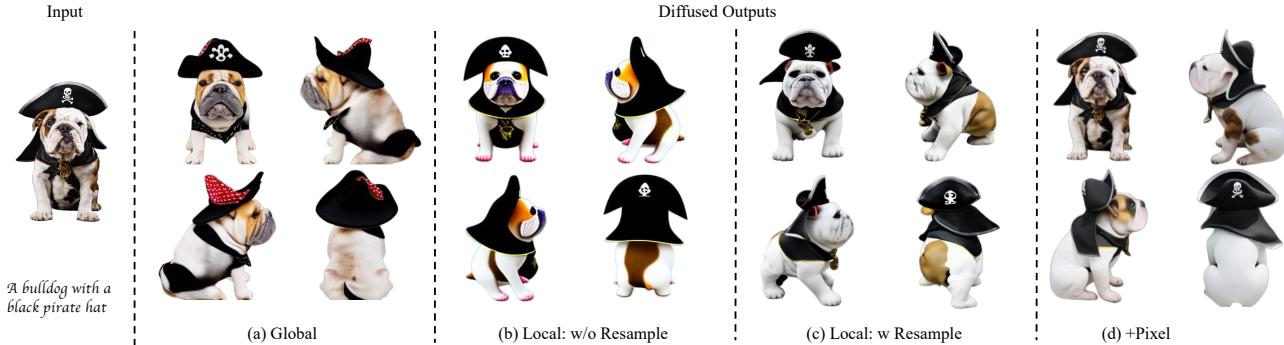


Figure 4. An example of the diffused results from different settings of multi-level controllers in ImageDream (see Sec. 3.3).

attention feature $\mathbf{h}_{g,l}$. Here, a weight $\lambda = 1.0$ is introduced to balance the hidden from text and image, and the final output of layer l is $\mathbf{h}_l = \mathbf{h}_{t,l} + \lambda \mathbf{h}_{g,l}$. We refer to decoupled cross-attention in IP-Adaptor for additional details.

To train such a model, we freeze the diffusion model, and only fine-tune $\{\theta_g, \theta_{k_g,l}, \theta_{v_g,l}\}_l$. We follow the training setting of MVDream by considering both 3D rendered datasets and 2D image datasets together, which will be elaborated in our experimental section.

After the model is tuned, we found the model is able of absorb some informations from the image such as structure of the object etc. As illustrated in Fig. 4 (a), comparing with the input, the diffused output is able to put the pirate hat similar to the image on the bulldog, while some detailed pose and appearance information is lost, which we think is not enough for a good control from the input image.

Local Controller. To enhance control, we try to utilize the hidden feature from the CLIP encoder before its global pooling, which likely contains more detailed structural information. This hidden feature, denoted as \mathbf{f}_h , has a token length of 257 and a feature dimension of 1280. A MLP adaptor θ_h is introduced to feed \mathbf{f}_h into the diffusion network’s cross-attention module, with $\theta_{k_h,l}$ and $\theta_{v_h,l}$ forming the key and values matrix. These parameters, $\{\theta_{k_h,l}, \theta_{v_h,l}, \theta_h\}_l$, are then jointly trained as learnable elements similar to the global controller. Post-training, we observed that the results were overly sensitive to image tokens, leading to overexposed and unrealistic images, especially with higher class free guidance (CFG) settings [33], as shown in Fig. 4(b).

To mitigate this, we implemented a resampling module θ_r , following the approach of IP-Adaptor, reducing the hidden token count from 257 to 16, resulting in a more balanced local image feature \mathbf{f}_r . The corresponding local controller parameters are $\{\theta_r, \theta_{k_r,l}, \theta_{v_r,l}\}_l$. As Fig. 4(c) illustrates, after this resampling, the diffused images more realistic, even at higher CFG levels. From the generated images, it’s evident that the model captures the overall layout and object shape, but also struggles with finer identity details like object skin texture.

Pixel Controller. To optimally integrate object appearance

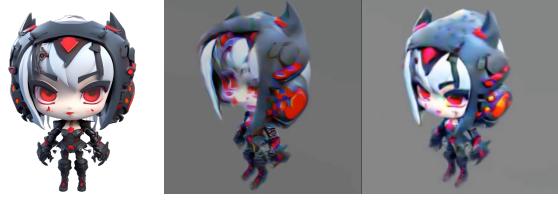
texture, we propose embedding the image prompt pixel latent \mathbf{x} across all attention layers in ImageDream. Specifically, MVDream employs a 3D dense self-attention mechanism with a shape of $(bz, 4, c, h_l, w_l)$ across four views within a transformer layer. In contrast, ImageDream introduces an additional frame by concatenating the input image, resulting in a feature shape of $(bz, 5, c, h_l, w_l)$. This enables similar 3D self-attention processes between the four-view images and the input image.

During the training of our diffusion network, we refrain from adding noise to the latent from the input image prompt, ensuring the network clearly captures the image information. Additionally, to differentiate the input image features and avoid confusion, we assign an all-zero vector to the camera embedding of the input image. Given that the pixel controller is integrated into the multi-view diffusion without extra parameters, we fine-tune all feature parameters in unison, adopting the same training regime as the global/local controllers but with a learning rate reduced by a factor of ten. This approach preserves the original feature representations more effectively. Post-training, as depicted in Fig 4(d), the generated multi-view images not only ethically maintain the appearance from the input image but also uphold the multi-view consistency characteristic of MVDream, resulting in satisfactory 3D model fusion.

Finally, our multi-level controller is a combined one with local and pixel, since we think the global one do not have too much additional information. There might be potential queries regarding the necessity of a pixel controller, given that IP-Adaptor, relying solely on CLIP features, can capture extensive image texture details: we posit that while IP-Adaptor is effective for modifying objects in the same view as the input, decoding the same view is comparatively simpler. Multi-view diffusion, however, presents a more complex challenge. Decoding from a highly compressed CLIP feature could necessitate prolonged training on larger datasets. Therefore, at this stage, we find the pixel controller significantly beneficial for rapidly training a robust multi-view diffusion model.



(a) The floating artifacts w/o background alignment



(b) The geometry collapse w/o and with camera alignment

Figure 5. Example of artifacts we fixed with image-prompt score distillation (see Sec. 3.4).

3.4. Image-Prompt Score Distillation

Implementing the image-prompt multi-view diffusion network in ImageDream follows the multi-view score distillation framework of MVDream (Sec. 3.1), with the addition of an image prompt as an input to the diffusion network. However, we need to consider few key differences in NeRF optimization to achieve accurate results.

Background Alignment. During SDS optimization, the NeRF-rendered image includes a randomly colored background to differentiate the interior and exterior of the 3D object. This random background, when input into the diffusion network alongside the object, can conflict with the background from the image prompt, leading to floating artifacts in the generated NeRF model, as shown in Fig. 5(a). To resolve this, we adjusted the image-prompt background to match the rendered background color from NeRF, successfully eliminating these artifacts.

Camera Alignment. Our diffusion network tends to generate multi-view images mirroring the camera parameters (e.g., elevation, field of view (FoV)) of the input image prompt, parameters which remain unknown during NeRF rendering. Randomly sampling parameters for rendering, as done in MVDream, can result in images incongruent with the image prompt’s rendering settings, affecting the geometry of detailed image structures. To mitigate this, we narrowed the parameter sampling range from MVDream’s [15, 60], [0, 30] for camera FoV and elevation to [45, 50] and [0, 5], respectively, a range more typical for a generated user photos. This adjustment significantly improved the geometric accuracy of the 3D objects, as demonstrated in Fig. 5(b).

We acknowledge this solution’s limitations; when the image prompt’s camera parameters greatly differ from our selected range in canonical camera setting, the resulting 3D object shape may be unpredictable. Future improvements could include a camera parameter estimation module or increased randomness in the image prompt rendering during diffusion training, to better synchronize the settings between NeRF rendering and diffusion.

4. Experiments

In this section, we detail the experimental setup for ImageDream, designed to enable replication of our model. We will release both the model and code following this submission.

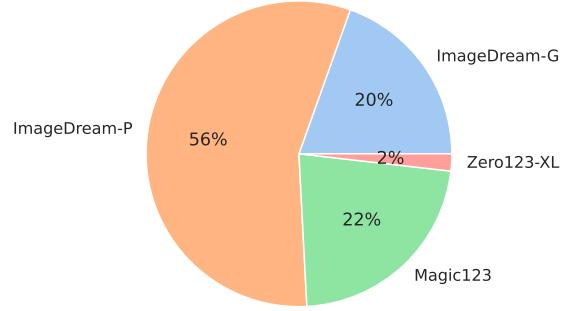


Figure 6. User study of different methods. ImageDream-P: our full model w pixel controller, and ImageDream-G: without pixel controller (see Sec. 4.1).

Implementation Details. Adhering to the dataset configuration of MVDream (Sec. 3.3), we used a combined dataset from Objaverse for 3D multi-view rendering and a 2D image dataset for training controllers (Sec. 3.3). For image prompts in the 3D dataset, we randomly selected one of the 16 front-side views, with azimuth angles ranging from $[-90, 90]$ degrees, out of the total 32 circle views. For the 2D dataset, we used the input image as the image prompt. A random dropout rate of 0.1 was set for the image prompt during training, replacing dropped prompts with a random uni-colored image. For all experiments, i.e. with global controller, local controller and local plus pixel controllers, we trained for 60K steps with a batch size of 256 and a gradient accumulation of 2, using the AdamW optimizer. The model is initialized from stable diffusion (2.1) checkpoint of MVDream. The learning rate was set to 1e-4, except for the model with pixel controller, where it was reduced to 1e-5. Test image prompts were resized to 256×256 , and we set the diffusion CFG to 5.0. The training takes ~ 2 days with 8 A100. For NeRF optimization, we followed MVDream’s configuration but introduced a three-stage optimization at resolutions [64, 192, 256] which switched at [5K, 10K] steps, setting the camera distance between [0.6, 0.85] for better NeRF model coverage. The NeRF training is about 1hr with A100.

Test Dataset. Our primary focus was evaluating ImageDream outside the Objaverse distribution to ensure its practical applicability. We selected 39 well-curated prompts

Model	Synthesized Image			Re-rendered		
	QIS(256)↑	CLIP(TX)↑	CLIP(IM)↑	QIS(320)↑	CLIP(TX)↑	CLIP(IM)↑
SD-XL [1]	52.0 ± 30.5	34.6 ± 3.09	100	-	-	-
MVDream [37]	23.05 ± 14.4	31.64 ± 2.99	78.41 ± 5.32	29.02 ± 10.24	32.69 ± 3.39	79.63 ± 4.15
Zero123 [19]	22.16 ± 11.16	30.42 ± 3.19	84.88 ± 5.12	-	-	-
Zero123-XL [19]	34.07 ± 11.64	30.80 ± 2.59	84.10 ± 4.76	28.66 ± 5.03	29.19 ± 3.60	79.92 ± 6.59
Magic123 [31]	-	-	-	24.23 ± 7.68	29.56 ± 4.73	82.50 ± 8.78
SyncDreamer [20]	22.04 ± 11.9	27.96 ± 3.01	78.17 ± 6.13	19.84 ± 6.64	25.82 ± 3.39	73.20 ± 6.30
ImageDream						
- global	22.31 ± 7.59	32.01 ± 2.84	84.50 ± 3.96	22.51 ± 5.86	31.48 ± 3.32	82.58 ± 4.35
- local (-G)	22.49 ± 9.57	31.32 ± 2.86	82.99 ± 6.03	22.30 ± 4.47	31.71 ± 2.96	84.34 ± 3.13
- +pixel (-P)	27.10 ± 12.8	32.39 ± 2.78	85.69 ± 3.77	25.16 ± 6.49	31.59 ± 3.23	84.83 ± 4.08

Table 1. Quantitative assessment of image synthesis quality using the test prompt list from MVDream. The DDIM sampler was employed for testing, and Implicit Volume from threestudio was utilized for re-rendering 3D model images. Here Zero123 we took their checkpoint of 165K which including all training instances. 'Zero123-XL' indicates Zero123 trained with the larger Objaverse-xl 10M dataset [7], while others were trained solely on the standard Objaverse dataset [8], which is ∼10 times smaller. (-P) and (-G) are correspondent to ImageDream-P and ImageDream-G in Fig.6.

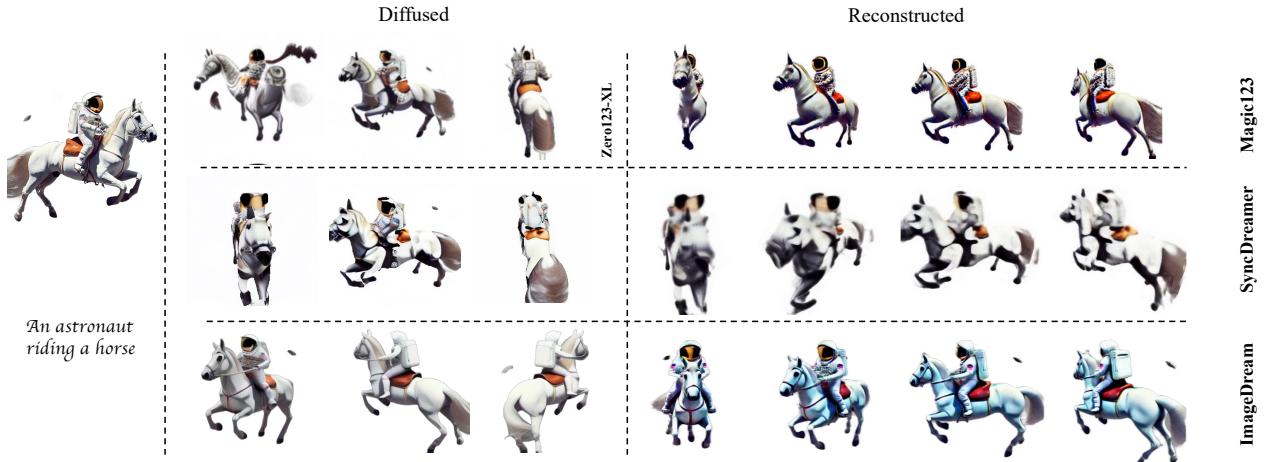


Figure 7. The illustration of synthesized images from different baselines. Diffused: From diffusion models. Reconstructed: Re-rendered images from correspondent fused NeRF model. Please check the webpage for more results.

from MVDream, covering a diverse range of objects with relatively complex geometries and appearances, surpassing datasets like ShapeNet [5] or CO3D [18]. Using SDXL [29], we generated multiple images from each prompt, selecting ones with aesthetically pleasing objects. The backgrounds of these images were then removed, and the objects re-centered, akin to the approach used in Zero123.

4.1. Comparisons

In our evaluation, we compared ImageDream’s performance against several SoTA baselines, including Zero123-XL [19] (trained on 10x larger data than ours), Magic123 [31], and SyncDreamer [20]. The criteria for

comparison were geometry quality and similarity to the image prompt (IP). ‘Geometry quality’ refers to the generated 3D asset’s conformance to common sense in terms of shape and minimal artifacts, while ‘similarity to IP’ assesses the resemblance of the results to the input image. We executed all baseline tests using default configurations as implemented in threestudio¹.

Qualitative Evaluation. Lacking ground truth for the test image prompts, we conducted a real user study to evaluate the quality of the generated 3D models. Participants were briefed on our evaluation standards and asked to choose their preferred model based on these criteria. The exper-

¹<https://github.com/threestudio-project/threestudio>

iment was double-blind, with participants shown 3D assets generated by different methods without identifying labels. The comparison results, depicted in Fig. 6, show that ImageDream, both with (ImageDream-P) and without (ImageDream-G) the pixel controller, significantly outperformed other baselines. ImageDream-P was particularly favored, while ImageDream-G also received a positive preference rate. SyncDreamer was omitted from the figure due to its NeuS results having a 0% preference rate.

Fig. 7 presents a representative case comparing results from the diffusion models and the final NeRF model. Systems like Magic123 and Zero123, which rely on single-view diffusion with relative camera embedding, often produce incorrect geometry, as illustrated by their inability to accurately represent the span of horse body. In contrast, ImageDream, through its unique design, effectively resolves this issue, resulting in more satisfactory models (more results are list in webpage).

Numerical Evaluation. To thoroughly assess image quality at various stages of our generation pipeline, we employed the Inception Score (IS)[34] and CLIP scores[32] using text-prompt and image-prompt, respectively. The IS evaluates image quality, while CLIP scores assess text-image and image-image alignment. However, since IS traditionally evaluates both image quality and diversity within a set, and our prompt quantity is limited, the diversity aspect makes the score less reliable. Therefore, we modified the IS by omitting its diversity evaluation, replacing the mean distribution with a uniform distribution. Specifically, we set q_i in IS to be $1/N$, making the IS of an image $\sum_i p_i \log(Np_i)$, where N is the inception class count and p_i is the predicted probability for the i^{th} class. We denote this modified metric as Quality-only IS (QIS). For the CLIP score, we calculated the mean score between each generated view and the provided text-prompt or image-prompt.

In Tab.1, we present comparative results. SD-XL, reflecting the score of test images, achieved the highest QIS and CLIP scores. MVDream, listed as a benchmark for final 3D model quality, shows improved synthesized image quality after 3D fusion due to multi-view consistency. In contrast, Zero123 and Zero123-XL experienced a drop in image quality post-3D fusion due to diffusion inconsistency. Magic123 enhanced the CLIP score over Zero123 by integrating a joint diffusion model. SyncDreamer’s quality declined as it diffuses only 16 fixed views, complicating reconstruction. In ImageDream, we evaluated three models for ablation: one with a global controller, another with a local controller, and the last incorporating both local and pixel controllers (Sec.3.3). ImageDream maintained high image quality in both diffusion and post-3D fusion stages. The local controller, in particular, provided better image CLIP scores post-fusion, thanks to richer image feature representations. The pixel controller model excelled in image



Figure 8. Illustration of a failure case. Where Trumps’ face turns blurry after 3D fusion reconstruction (first row). However, the face textures can be generated with model using global-only controller thanks to its semantic global representation.

CLIP scores during both stages. Notably, ImageDream-pixel ranked second in other scores, with Zero123-XL using a significantly larger dataset (Objaverse-XL [7]).

However, these scores don’t fully encapsulate important aspects like multi-view consistency and geometric correctness. For instance, as Fig. 7 demonstrates, zero123-XL, despite having high IS due to easy image classification, showed poorer consistency. Thus, while these scores offer some reliability when consistency is high, future research should aim to develop more comprehensive metrics that accurately capture geometric correctness to better compare different generation algorithms.

4.2. Limitations

While the model incorporating the pixel controller achieves the best scores, we observed certain trade-offs, particularly when the image constraints are overly stringent. For instance, in cases like the small facial details of a full-body avatar (shown in Fig.8), the model struggles to capture these nuances, whereas global control might recover the face based on the text prompt. To address this, as outlined in Sec.3.4, the pixel controller model needs to better estimate image intrinsic and extrinsic properties, or a better balance tuning inside of multi-level controllers. This may be solved by exploring the use of larger models, such as SDXL [29], which be our future work.

5. Conclusion

We introduce ImageDream, an advanced image-prompt 3D generation model utilizing multi-view diffusion. This model innovatively applies canonical camera coordination and multi-level image-prompt controllers, enhancing control and addressing geometric inaccuracies seen in prior methods. Future improvements could focus on increasing randomness in image-prompts during training to further reduce texture blurriness in the generated models. These steps are expected to further advance the capabilities and applications of ImageDream in 3D model generation.

6. Acknowledgements and Ethics Statement

We thank our 3D group members of **Kejie Li**, and our intern **Zeyuan Chen** in joint meeting discussion and setup baselines of SyncDreamer, which help complete this paper. In addition, note that the models proposed in this paper aims to facilitate the 3D generation task that is widely demanded in industry for ethical purpose. It could be potentially applied to unwanted scenarios such as generating violent and sexual content by third-party fine-tuning. Built upon the Stable Diffusion model [33], it might also inherit the biases and limitations to generate unwanted results. Therefore, we believe that the images or models synthesized using our approach should be carefully examined and be presented as synthetic. Such generative models may also have the potential to displace creative workers via automation. That being said, these tools may also enable growth and improve accessibility for the creative industry.

References

- [1] stable-diffusion-xl-base-1.0. <https://huggingface.co/stabilityai/stable-diffusion-xl-base-1.0>. Accessed: 2023-08-29. 7
- [2] Stable diffusion image variation. <https://huggingface.co/spaces/lambdalabs/stable-diffusion-image-variations>. 3
- [3] 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 *CVPR*, 2022. 3
- [4] 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 *arXiv*, 2023. 3
- [5] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 3, 7
- [6] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. *arXiv:2303.13873*, 2023. 3
- [7] 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. 2023. 7, 8
- [8] 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 *CVPR*, pages 13142–13153, 2023. 4, 7
- [9] Yu Deng, Jiaolong Yang, Jianfeng Xiang, and Xin Tong. Gram: Generative radiance manifolds for 3d-aware image generation. In *CVPR*, pages 10673–10683, 2022. 3
- [10] 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. *NeurIPS*, 2022. 3
- [11] Paul Henderson and Vittorio Ferrari. Learning single-image 3d reconstruction by generative modelling of shape, pose and shading. *International Journal of Computer Vision*, 2020. 3
- [12] Paul Henderson, Vagia Tsimimaki, and Christoph H Lampert. Leveraging 2d data to learn textured 3d mesh generation. In *CVPR*, 2020. 3
- [13] Yukun Huang, Jianan Wang, Yukai Shi, Xianbiao Qi, Zheng-Jun Zha, and Lei Zhang. Dreamtime: An improved optimization strategy for text-to-3d content creation. *arXiv:2306.12422*, 2023. 3
- [14] Heewoo Jun and Alex Nichol. Shap-e: Generating conditional 3d implicit functions. *arXiv:2305.02463*, 2023. 3
- [15] Animesh Karnewar, Andrea Vedaldi, David Novotny, and Niloy J Mitra. Holodiffusion: Training a 3d diffusion model using 2d images. In *CVPR*, 2023. 3
- [16] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014. 4
- [17] 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 *CVPR*, 2023. 3
- [18] Shanchuan Lin, Bingchen Liu, Jiashi Li, and Xiao Yang. Common diffusion noise schedules and sample steps are flawed. *arXiv:2305.08891*, 2023. 7
- [19] 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. *arXiv:2303.11328*, 2023. 2, 3, 4, 7
- [20] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Learning to generate multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023. 2, 3, 7
- [21] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, and Wenping Wang. Wonder3d: Single image to 3d using cross-domain diffusion, 2023. 3
- [22] Luke Melas-Kyriazi, Iro Laina, Christian Rupprecht, and Andrea Vedaldi. Realfusion: 360deg reconstruction of any object from a single image. In *CVPR*, 2023. 3
- [23] 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 *ECCV*, 2021. 3
- [24] Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. Hologan: Unsupervised learning of 3d representations from natural images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019. 3

- [25] Thu H Nguyen-Phuoc, Christian Richardt, Long Mai, Yongliang Yang, and Niloy Mitra. Blockgan: Learning 3d object-aware scene representations from unlabelled images. *NeurIPS*, 2020. 3
- [26] 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
- [27] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *CVPR*, 2021. 3
- [28] Michael Niemeyer, Jonathan T. Barron, Ben Mildenhall, Mehdi S. M. Sajjadi, Andreas Geiger, and Noha Radwan. Regnرف: Regularizing neural radiance fields for view synthesis from sparse inputs. In *CVPR*, 2022. 3
- [29] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv:2307.01952*, 2023. 2, 7, 8
- [30] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *ICLR*, 2023. 1, 3, 4
- [31] Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov, and Bernard Ghanem. Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. *arXiv preprint arXiv:2306.17843*, 2023. 1, 2, 3, 7
- [32] 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 *ICML*, 2021. 2, 4, 8
- [33] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 3, 5, 9
- [34] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *NeurIPS*, 2016. 8
- [35] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *NeurIPS*, 2022. 4
- [36] Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*, 2023. 3
- [37] Yichun Shi, Peng Wang, Jianglong Ye, Long Mai, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv:2308.16512*, 2023. 1, 2, 3, 4, 7
- [38] J Ryan Shue, Eric Ryan Chan, Ryan Po, Zachary Ankner, Jiajun Wu, and Gordon Wetzstein. 3d neural field generation using triplane diffusion. In *CVPR*, 2023. 3
- [39] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. *NeurIPS*, 32, 2019. 3
- [40] Jingxiang Sun, Bo Zhang, Ruizhi Shao, Lizhen Wang, Wen Liu, Zhenda Xie, and Yebin Liu. Dreamcraft3d: Hierarchical 3d generation with bootstrapped diffusion prior, 2023. 3
- [41] Stanislaw Szymanowicz, Christian Rupprecht, and Andrea Vedaldi. Viewset diffusion:(0-) image-conditioned 3d generative models from 2d data, 2023. 3
- [42] Junshu Tang, Tengfei Wang, Bo Zhang, Ting Zhang, Ran Yi, Lizhuang Ma, and Dong Chen. Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior. *arXiv:2303.14184*, 2023. 3
- [43] Christina Tsalicoglou, Fabian Manhardt, Alessio Tonioni, Michael Niemeyer, and Federico Tombari. Textmesh: Generation of realistic 3d meshes from text prompts. *arXiv:2304.12439*, 2023. 3
- [44] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *CVPR*, 2023. 3
- [45] Tengfei Wang, Bo Zhang, Ting Zhang, Shuyang Gu, Jianmin Bao, Tadas Baltrusaitis, Jingjing Shen, Dong Chen, Fang Wen, Qifeng Chen, et al. Rodin: A generative model for sculpting 3d digital avatars using diffusion. In *CVPR*, 2023. 3
- [46] 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. *arXiv:2305.16213*, 2023. 3
- [47] Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models. In *ICLR*, 2023. 3
- [48] Chao-Yuan Wu, Justin Johnson, Jitendra Malik, Christoph Feichtenhofer, and Georgia Gkioxari. Multiview compressive coding for 3d reconstruction. In *CVPR*, 2023. 3
- [49] Dejia Xu, Yifan Jiang, Peihao Wang, Zhiwen Fan, Humphrey Shi, and Zhangyang Wang. Sinnerf: Training neural radiance fields on complex scenes from a single image. 2022. 3
- [50] Dejia Xu, Yifan Jiang, Peihao Wang, Zhiwen Fan, Yi Wang, and Zhangyang Wang. Neurallift-360: Lifting an in-the-wild 2d photo to a 3d object with 360° views. 2022. 3
- [51] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 4
- [52] Jianglong Ye, Peng Wang, Kejie Li, Yichun Shi, and Heng Wang. Consistent-1-to-3: Consistent image to 3d view synthesis via geometry-aware diffusion models. *arXiv preprint arXiv:2310.03020*, 2023. 3
- [53] Zhizhuo Zhou and Shubham Tulsiani. Sparsefusion: Distilling view-conditioned diffusion for 3d reconstruction. In *CVPR*, 2023. 3