DreamCraft3D++: Efficient Hierarchical 3D Generation with Multi-Plane Reconstruction Model

Jingxiang Sun, Cheng Peng, Ruizhi Shao, Yuan-Chen Guo, Xiaochen Zhao, Yangguang Li, YanPei Cao, Bo Zhang, Yebin Liu, *Member, IEEE*

Abstract—We introduce DreamCraft3D++, an extension of DreamCraft3D that enables efficient high-quality generation of complex 3D assets. DreamCraft3D++ inherits the multi-stage generation process of DreamCraft3D, but replaces the time-consuming geometry sculpting optimization with a feed-forward multi-plane based reconstruction model, speeding up the process by 1000x. For texture refinement, we propose a training-free IP-Adapter module that is conditioned on the enhanced multi-view images to enhance texture and geometry consistency, providing a 4x faster alternative to DreamCraft3D's DreamBooth fine-tuning. Experiments on diverse datasets demonstrate DreamCraft3D++'s ability to generate creative 3D assets with intricate geometry and realistic 360° textures, outperforming state-of-the-art image-to-3D methods in quality and speed. The full implementation will be open-sourced to enable new possibilities in 3D content creation.

Index Terms—3D generation, diffusion model, score distillation, single-view 3D reconstruction

1 Introduction

The remarkable success of 2D generative modeling [1], [2], [3], [4] has profoundly shaped the way that we create visual content. 3D content creation, which is crucial for applications like games, movies and virtual reality, still presents a significant challenge for deep generative networks. While 3D generative modeling has shown compelling results for certain categories [5], [6], [7], generating general 3D objects remains formidable due to the lack of extensive 3D data.

Recent research in 3D generation has evolved into two main paradigms: 2D lifting [8], [9], [10], [11], [12] and feed-forward 3D generation [13], [14], [15], [16], [17], [18]. 2D lifting methods leverage pretrained 2D vision models to guide 3D optimization, with DreamFusion [8] introducing Score Distillation Sampling (SDS) loss to align 3D renderings with text-conditioned image distributions. Subsequent works have improved photo-realism through stage-wise optimization and enhanced distillation losses. However, these approaches often struggle with complex content synthesis and suffer from the "Janus issue," where individual renderings appear plausible but lack holistic consistency. While recent advancements have significantly improved quality and speed, SDS-based methods remain computationally expensive, typically requiring minutes to hours for single object generation.

Recent advancements in the feed-forward 3D generation [13], [14], [15], [16], [17], [18] have significantly improved efficiency through the use of 3D large reconstruction models(LRM) [15]. These approaches typically employ a two-step process: first, generating multi-view images from a single input image or text prompt, and then directly regressing 3D representations from these generated images using sparse-view reconstructors. However, these methods invariably require 3D shapes or multi-view data for training, which presents challenges when generating in-the-wild 3D assets due to the relative scarcity of diverse 3D data compared to 2D data. Furthermore, the reconstruction results often exhibit limitations, such as a lack of fine geometric structures and detailed texture patterns.

To recap, DreamCraft3D [12] falls within the category of 2D lifting and draw inspiration from the manual artistic process, breaking down the challenging 3D generation into manageable steps. Starting with a high-quality 2D reference image generated from a text prompt, DreamCraft3D lifts it into 3D via stages of geometry sculpting and texture boosting. In the geometry sculpting stage, DreamCraft3D adopts joint 2D-3D SDS loss for novel views and photometric loss at the reference view for producing plausible and consistent 3D geometry. In the texture boosting stage, in order to obtain consistent texture, a pretrained text-to-image diffusion model is finetuned on multi-view renderings of the 3D instance, resulting a personalized 3D-aware generative prior for texture boosting.

While DreamCraft3D significantly improves 3D generation quality, its main drawback is the lengthy processing time, requiring approximately 3 hours per case. In this paper, we introduce DreamCraft3D++, an enhanced extension

- *Jingxiang Sun, Cheng Peng, Ruizhi Shao, Xiaochen Zhao and Yebin Liu are with Department of Automation, Tsinghua University, Beijing, China. E-mail: {starkjxsun, chengpeng002, jia1saurus}@gmail.com; zhaoxc19@mails.tsinghua.edu.cn; liuyebin@mail.tsinghua.edu.cn*
- *Yuan-chen Guo, Yangguang Li and Yanpei Cao are with VAST, Beijing, China. E-mail: {imbennyguo, liyangguang256, caoyanpei}@gmail.com*
- *Bo Zhang is with Zhejiang University, Hangzhou, China. E-mail: bo.zhang@zju.edu.cn*
- *Corresponding author: Yebin Liu.*

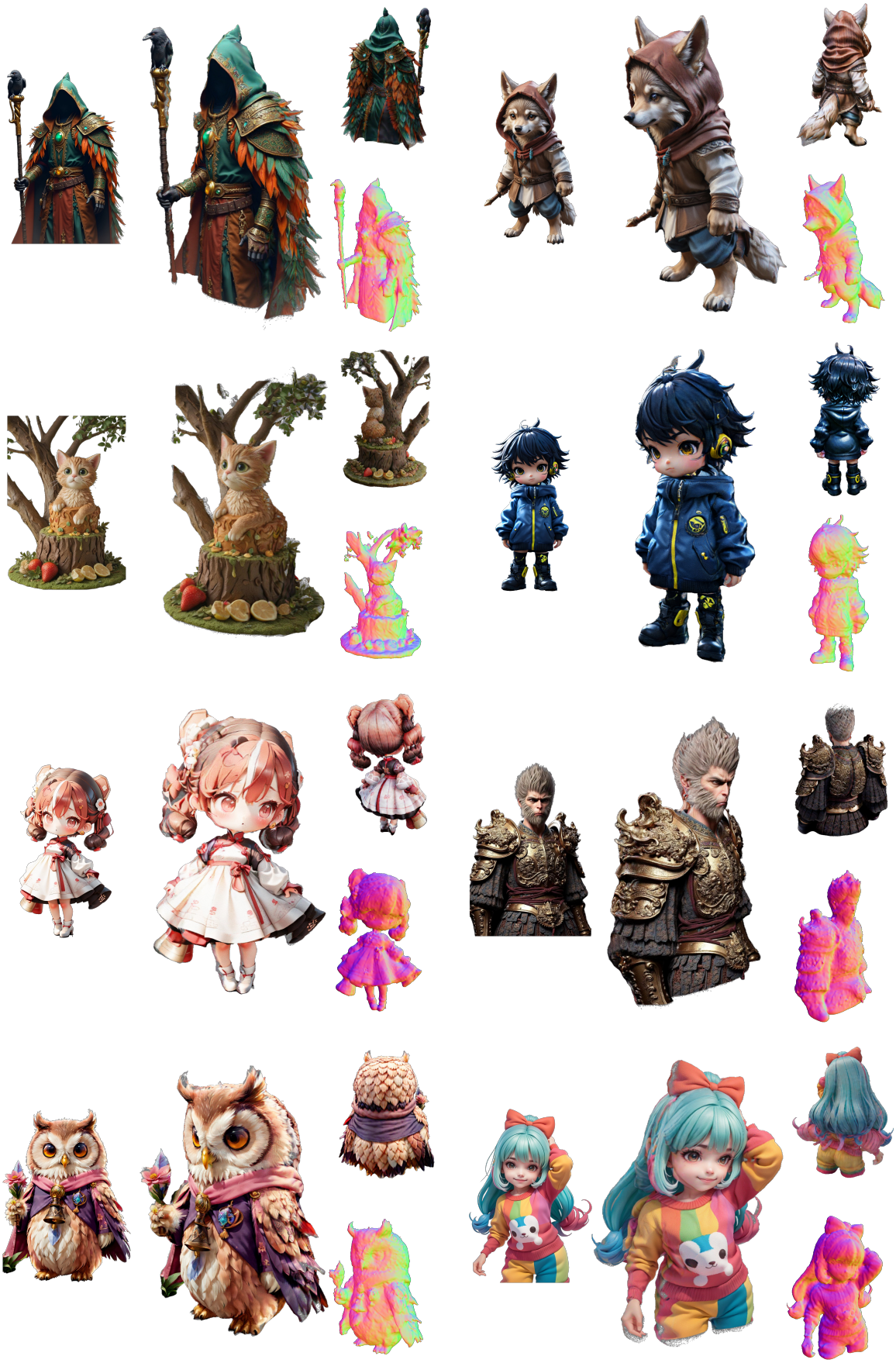


Fig. 1: By lifting 2D images to 3D, DreamCraft3D++ achieves 3D generation with rich details and holistic 3D consistency. Please refer to the demo video for more results.

of DreamCraft3D that enables efficient and high-quality complex 3D asset generation. Our approach maintains the high generation quality of DreamCraft3D while accelerating the process by a factor of 20, reducing the generation time from three hours to just 10 minutes. Through a detailed analysis of the optimization time for individual stages, we identified that the geometry sculpting stage accounts for over 70% of the total time, as optimizing 3D structures from scratch without any shape prior is computationally expensive. Inspired by [15], we propose training a large reconstruction model to provide a coarse textured mesh, thereby replacing the original time-consuming geometry sculpting stage.

To achieve this, our approach utilizes multi-view images as input to a UNet-based architecture for predicting pixel-aligned multi-plane features, which is crucial for bridging 2D and 3D. Our method diverges from current counterparts [15], [16], [19] in two key aspects. Firstly, our approach predicts a nonorthogonal multi-plane representation from fixed-view images, instead of predicting orthogonal axis-aligned planes (e.g. tri-planes) in [15], [16], [20], [21], [22]. We opt for nonorthogonal planes instead of orthogonal ones because directly learning generalized canonical 3D features (e.g., axis-aligned triplanes or grids) from 2D posed images is difficult, particularly given limited training data and network capacity. By predicting nonorthogonal image-aligned plane features, the network can concentrate more effectively on translating 2D high-frequency details into 3D representations. Secondly, we leverage a U-Net convolutional architecture to map input images to multi-planes, exploiting the strong pixel-level alignment between input and output. The substantial bandwidth capacity of our U-Net enables direct transformation of multi-view images into multi-planes, yielding highly detailed results. Additionally, we incorporate normal maps into the reconstruction network, enhancing the model's comprehension of spatial relationships and geometry.

From the generated multi-plane features, we further decode Flexicubes [23] parameters and predict the textured mesh. In contrast to NeRF's memory-intensive volume rendering, our approach leverages efficient mesh rasterization, enabling the use of full-resolution images and additional geometric information for supervision. This results in superior reconstruction quality. While alternative 3D representations such as Gaussian Splatting [24] also facilitate efficient high-resolution rendering, they lack explicitly defined surfaces, making them less suitable for geometric modeling.

The output textured meshes of the above geometry sculpting stage are highly consistent to the multi-view input images, though, still suffer from detail distortion because 1) limited network generalbility due to lacking data; 2) inconsistent and blurry generated multi-view renderings from mv diffusion models. Therefore, we introduce a novel refinement algorithm to enhance both texture and geometry concurrently. The core insight lies in that the pretrained 2D diffusion priors lack object awareness and lead to inconsistent optimization, shared insight with DreamCraft3D. But different from it which finetunes a pretrained diffusion model with the multi-view image renderings using DreamBooth, we instead integerate a lightweight image prompt adaptation module, named IP-Adapter [25], enabling the model to form a concept about the scene's surrounding views. Different from DreamBooth, Ip-adapter is training-free so it is much more efficient. Meanwhile, conditioning the IP-adapter solely on the source image can lead to ambiguity and the texture multi-head problem. To address this, we condition the IP-adapter on both the source image and augmented multi-view renderings. During training, the appropriate IP-adapter image embedding is dynamically selected based on the position of the randomly sampled camera.

As shown in Figure 1, our method is capable of producing creative 3D assets with intricate geometric structures and realistic textures rendered coherently in 360°. Compared to optimization-based approaches [8], [9], our method offers substantially improved texture and efficiency. Meanwhile, compared to image-to-3D techniques [11], [26], our work excels at producing unprecedentedly realistic renderings in 360°renderings. These results suggest the strong potential of DreamCraft3D++ in enabling new creative possibilities in 3D content creation. The full implementation will be made publicly available.

While preserving the core multi-stage framework and object-awareness, this paper extends DreamCraft3D, the conference version of this work in the following key aspects:

- For the coarse geometry sculpting stage, we introduce a feed-forward multi-plane based large reconstruction model to replace time-expensive optimization in DreamCraft3D, speeding up 1000 times with comparable results;
- We introduce a training-free IP-Adapter to enhance texture and geometry, achieving comparable results to DreamCraft3D's DreamBooth fine-tuning while being 4 times faster. Our IP-Adapter's dynamic embedding selection based on camera position addresses texture inconsistency and preserves fidelity, offering an efficient alternative to DreamCraft3D's approach.
- Compared to DreamCraft3D, we conduct experiments on a wider range of datasets, demonstrating the robustness and superiority of our model over other image-to-3D methods.

2 Related Works

In this work, we focus on the task of generating high-quality geometric and richly textured 3D models from single images. We categorize the main research efforts in this area into three aspects: Novel-View Synthesis, which involves generating views from different perspectives of an object based on input text or generating new viewpoint views from a single input image; Progressively Optimized Reconstruction, which leverages 2D generative models to progressively optimize implicit 3D neural fields; and Feed-Forward Generation, which involves generating 3D representations in a feed-forward manner based on given text or image instructions.

2.1 Novel-View Synthesis

Recently, direct novel views synthesis(NVS) from single images of a 3D object has been explored, these works [27],

[28], [29] often rely on a pretrained monocular depth prediction model to synthesize view-consistent images. While some models achieve photo-realistic renderings for ImageNet categories, they struggle with large views. Recent attempts [30], [31], [32], [33], [34], [35], [36], [37] training view-dependent diffusion models on 3D data show promising results for open-domain novel view synthesis but, as inherently 2D models, can't ensure perfect view consistency.

Based on this rationale, diffusion models for NVS are utilized as inputs for feed-forward 3D generation models, aiming to leverage the generative capabilities of large models to eliminate inconsistencies present in multi-view images while preserving reasonable texture and geometric information found in those views. Inspired by these works, we adopt the generative outputs of the multi-view generation model Zero123++ [32] as prior inputs for our mp-lrm model.

2.2 Progressively Optimized Reconstruction

Progressively Optimized Reconstruction improve a 3D scene representation by seeking guidance using established 2D text-image foundation models. Early works [38], [39], [40] utilize the pretrained CLIP [41] model to maximize the similarity between rendered images and text prompt. DreamFusion [8] and SJC [42], on the other hand, propose to distill the score of image distribution from a pretrained diffusion model and demonstrate promising results. Recent works have sought to further enhance the texture realism via coarse-to-fine optimization [9], [43], improved distillation loss [10], [44], [45], shape guidance [46] or lifting NVS 2D images to 3D [11], [12], [26], [33], [47], [48], [49], [50], [51]. Recently, [52] proposes to finetune a personalized diffusion model for 3D consistent generation. However, producing globally consistent 3D remains challenging. DreamCraft3D [12] meticulously design 3D priors through the whole hierarchical generation process, achieving unprecedented coherent 3D generation. However, the process necessitates approximately three hours per case, which is highly inefficient.

2.3 Feed-forward Generation

3D generative models have been intensively studied to generate 3D assets without tedious manual creation. Generative adversarial networks (GANs) [6], [28], [53], [53], [54], [54], [55], [56], [57], [58], [59] have long been the prominent techniques in the field. Auto-regressive models have been explored [60], [61], [62], [63], [64], which learn the distribution of these 3D shapes conditioned on images or texts. Diffusion models [5], [65], [66], [67], [68], [69], [70], [71], [72], [73], [74], [75], [76] have also shown significant recent success in learning probabilistic mappings from text or images to 3D shape latent.

Recently, a series of large-scale 3D reconstruction models based on transformers have emerged [15], [16], [18], [21], [22], [77], [78]. Some of these models use single images [15], [16], [18], [21], [77], while others utilize multiple images generated by NVS models [22], [78]. These models employ an end-to-end approach to generate implicit 3D representations using triplane NeRF. In addition, UNet has also been proven effective in generating multiplane and Gaussian representations [17], [79].

However, all these methods require 3D shapes or multi-view data for training, raising challenges when generating in-the-wild 3D assets due to the scarcity of diverse 3D data [80], [81], [82] compared to 2D. At the same time, the reconstruction results of these methods often face issues such as the lack of fine geometric structures and the lack of exquisite texture patterns.

3 Preliminaries

DreamFusion [8] achieves text-to-3D generation by utilizing a pretrained text-to-image diffusion model $\boldsymbol{\epsilon}_\phi$ as an image prior to optimizing the 3D representation parameterized by θ.

The image $\boldsymbol{x} = g(\theta)$, rendered at random viewpoints by a volumetric renderer, is expected to represent a sample drawn from the text-conditioned image distribution $p(\boldsymbol{x}|y)$ modeled by a pretrained diffusion model. The diffusion model ϕ is trained to predict the sampled noise $\boldsymbol{\epsilon}_\phi(\boldsymbol{x}_t; y, t)$ of the noisy image $\boldsymbol{x}_t$ at the noise level t, conditioned on the text prompt y. A *score distillation sampling* (SDS) loss encourages the rendered images to match the distribution modeled by the diffusion model. Specifically, the SDS loss computes the gradient:

$$\nabla_\theta \mathcal{L}_{\text{SDS}}(\phi, g(\theta)) = \mathbb{E}_{t,\boldsymbol{\epsilon}}\Big[\omega(t)(\boldsymbol{\epsilon}_\phi(\boldsymbol{x}_t; y, t) - \boldsymbol{\epsilon})\frac{\partial \boldsymbol{x}}{\partial \theta}\Big], \quad (1)$$

which is the per-pixel difference between the predicted and the added noise upon the rendered image, where $\omega(t)$ is the weighting function.

One way to improve the generation quality of a conditional diffusion model is to use the classifier-free guidance (CFG) technique to steer the sampling slightly away from the unconditional sampling, i.e., $\epsilon_\phi(\boldsymbol{x}_t; y, t) + \omega\epsilon_\phi(\boldsymbol{x}_t; y, t) - \omega\epsilon_\phi(\boldsymbol{x}_t, t, \varnothing)$, where $\varnothing$ represents the "empty" text prompt. Typically, the SDS loss requires a large CFG guidance weight for high-quality text-to-3D generation, yet this will bring side effects like over-saturation and over-smoothing [8].

Recently, Wang et al. [10] proposed a variational score distillation (VSD) loss that is friendly to standard CFG guidance strength and better resolves unnatural textures. Instead of seeking a single data point, this approach regards the solution corresponding to a text prompt as a random variable. Specifically, VSD optimizes a distribution $q^\mu(\boldsymbol{x}_0|y)$ of the possible 3D representations $\mu(\theta|y)$ corresponding to the text y, to be closely aligned with the distribution defined by the diffusion timestep $t = 0$, $p(\boldsymbol{x}_0|y)$, in terms of KL divergence:

$$\mathcal{L}_{\text{VSD}} = D_{\text{KL}}(q^\mu(\boldsymbol{x}_0|y)||p(\boldsymbol{x}_0|y)). \quad (2)$$

[10] further shows that this objective can be optimized by matching the score of noisy real images and that of noisy rendered images at each time t, so the gradient of $\mathcal{L}_{\text{VSD}}$ is

$$\nabla_\theta \mathcal{L}_{\text{VSD}}(\phi, g(\theta)) = \mathbb{E}_{t,\boldsymbol{\epsilon}}\Big[\omega(t)(\boldsymbol{\epsilon}_\phi(\boldsymbol{x}_t; y, t) - \boldsymbol{\epsilon}_{\text{lora}}(\boldsymbol{x}_t; y, t, c))\frac{\partial \boldsymbol{x}}{\partial \theta}\Big]. \quad (3)$$

Here, $\boldsymbol{\epsilon}_{\text{lora}}$ estimates the score of the rendered images using a LoRA (Low-rank adaptation) [83] model. The obtained variational distribution yields samples with high-fidelity textures. However, this loss is applied for texture enhancement and is helpless to the coarse geometry initially learned by SDS. Moreover, both the SDS and VSD attempt to distill

from a fixed target 2D distribution which only assures per-view plausibility rather than a global 3D consistency. Consequently, they suffer from the same appearance and semantic shift issue that hampers the perceived 3D quality.

4 DreamCraft3D++: Overview

We propose a hierarchical pipeline for high-quality and efficient 3D content generation as illustrated in Figure 2. Given a single image, our method first leverages state-of-the-art multi-view diffusion models to produce several orthogonal and consistent multi-view images. Then, we build a feed-forward sparse-view 3D reconstruction model (Sec. 5) to efficiently infer the underlying textured meshes from those input images. For this stage, we prioritize multi-view consistency and global 3D structure, allowing for some compromise on detailed textures and geometry. Finally, we focus on jointly optimizing realistic and coherent texture as well as detailed geometry, with a training-free object-aware diffusion prior (Sec. 6).

5 MP-LRM: Multi-Plane Large Reconstruction Model

As illustrated in Figure 2, the Multi-Plane Large Reconstruction Model (MP-LRM) takes as input multi-view images and their corresponding normal maps with known camera poses. A convolutional U-Net is employed to map the input images and normal maps to a set of non-orthogonal multiple planes (Sec. 5.1 and 5.2). Subsequently, lightweight multi-layer perceptrons (MLPs) are utilized to decode the triplane features into signed distance field (SDF) values, texture colors, and Flexicubes parameters (Sec. 5.3). Finally, these decoded values are used to obtain a textured mesh via the dual marching cubes algorithm. In the following subsections, we elaborate on the key components of MP-LRM (Sec. 5.4).

5.1 Nonorthogonal multi-plane representation with U-Net based backbone

The core of our framework is a U-Net-based backbone $\mathcal{U}$ that predicts a nonorthogonal multi-plane representation from multi-view images. Figure 2 illustrates the network architecture. The input consists of multi-view images N multi-view images $\{\mathbf{I}_i \in \mathbb{R}^{H\times W\times 3} | i = 1, 2, ..., N\}$ and their corresponding normal maps $\{\mathbf{N}_i \in \mathbb{R}^{H\times W\times 3} | i = 1, 2, ..., N\}$ with known poses. Here we adopt Zero123++ [32] to generate 6 fix-viewed images and the number of planes $N = 6$. Following previous works [77], [84], we employ Plücker ray embedding to densely encode the camera poses. The RGB values and ray embeddings are concatenated into 12-channel feature maps, which are then fed through $\mathcal{U}$ to predict a set of nonorthogonal planes $\{\mathbf{\Pi}_i \in \mathbb{R}^{H\times W\times C} | i = 1, 2, ..., N\}$. The term "nonorthogonal" refers to coordinate systems where the axes are not perpendicular to each other, unlike the axis-aligned triplanes used in other methods. We opt for nonorthogonal planes instead of orthogonal axis-aligned planes because we believe that learning the mapping from 2D posed images to orthogonal plane features is inefficient. If the network capacity is insufficient, the mapping result is prone to blurriness. Considering that multi-view diffusion models always generate fixed-view images, we allow the network to focus on learning pixel-aligned features without the need to unpose them in space.

We employ a convolutional U-Net $\mathcal{U}$ to learn the mapping from 2D posed images to non-orthogonal plane features. Compared to transformer-based methods [15], [16], [19], [77], our U-shape design has a larger bandwidth for preserving input information, resulting in highly detailed triplane features and, ultimately, elaborate textured meshes. Moreover, the convolutional network fully utilizes the geometric prior of the spatial correspondence between triplanes and the input six orthographic images, which greatly accelerates convergence and stabilizes training. Our model $\mathcal{M}$ can achieve reasonable reconstruction results at a very early stage of training (around 20 minutes of training from scratch).

5.2 Learnable plane embeddings

To address the lack of information in top and bottom views due to the limited elevation range of generated multi-view images, which often results in meshes with holes or artifacts, we introduce two additional learnable plane embeddings. These embeddings, denoted as $\mathbf{E}_{\text{top}}$ and $\mathbf{E}_{\text{bottom}}$, are strategically placed at the top and bottom, respectively. They are processed through a U-Net-based backbone $\mathcal{U}_E$. Specifically, these learnable plane embeddings and input images are processed separately by the corresponding U-Net model's down-sampling and up-sampling paths, while the intermediate feature maps from both U-Nets are concatenated and passed through the self-attention module. Our experiments demonstrate that this approach effectively mitigates holes, enhancing the quality of the reconstructed meshes.

5.3 Decoding planes to Flexicubes

Previous generic 3D generation methods predominantly employ NeRF [85] or Gaussian splatting [24] as the geometry representation. These methods rely on additional procedures, such as Marching Cubes (MC), to extract the iso-surface, leading to topological ambiguities and challenges in representing high-fidelity geometric details. In this work, we utilize Flexicubes [23] as our geometry representation. Flexicubes allow for mesh extraction from grid features using dual marching cubes during training. The features include signed distance function (SDF) values, deformation, and weights. Texture is obtained by querying the color at the surface. Flexicubes enable the training of our reconstruction model to produce textured meshes as the final output in an end-to-end manner. Given the surface vertices, we project them onto each nonorthogonal plane and query the feature using bilinear sampling. The features from all planes are aggregated through channel-wise concatenation.

5.4 Loss Function

RGB loss. During training, we render the images at the K supervision views, and minimize the image reconstruction loss. Let $\{I_i^{gt} | i = 1, 2, ..., K\}$ be the set of groundtruth views, and $\{\hat{I}_i\}$ be the rendered images, our loss function

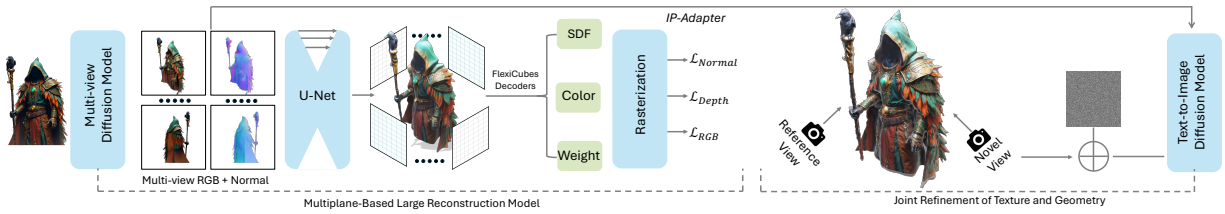


Fig. 2: DreamCraft3D++ pipeline. A single input image is processed by multi-view diffusion models to generate orthogonal, consistent views and normal maps. A feed-forward sparse-view 3D reconstruction model (Sec. 5) infers textured meshes from the multi-view images using a convolutional U-Net to map input to non-orthogonal planes, decoded into Flexicubes. Finally, a training-free object-aware diffusion prior enhances high-frequency geometry and texture details via score distillation (Sec. 6).

is a combination of MSE (Mean Squared Error) loss and Perceptual loss:

$$\begin{aligned}\mathcal{L}_{\text{rgb}} = \lambda_{\text{mse}} \cdot \frac{1}{K}\sum_{i=1}^{K}\left\|\hat{I}_i, I_i^{gt}\right\|_2^2 \\ + \lambda_{\text{lpips}} \cdot \frac{1}{K}\sum_{i=1}^{K}\mathcal{L}_{\text{lpips}}\left(\hat{I}_i, I_i^{gt}\right)\end{aligned} \tag{4}$$

where λ_{lpips} are the weight of the Perceptual loss. During training, we set $\lambda_{\text{lpips}} = 2.0$.

Mask loss. During training, we also render the masks at the K supervision views, and minimize the mask loss. Let $\{M_i^{gt}|i = 1, 2, ..., K\}$ be the set of groundtruth views, and $\{\hat{M}_i\}$ be the rendered masks, and we adopt Binary Cross Entropy (BCE) loss as mask loss:

$$\mathcal{L}_{\text{mask}} = \lambda_{\text{mask}} \cdot \frac{1}{K}\sum_{i=1}^{K}\mathcal{L}_{\text{BCE}}\left(\hat{M}_i, M_i^{gt}\right) \tag{5}$$

where λ_{mask} are the weight of the mask loss. During training, we set $\lambda_{\text{mask}} = 0.1$.

Depth and normal loss. In addition, akin to Nerdi [47], we fully exploit the geometry prior inferred from the reference image, and enforce the consistency with the depth and normal map computed for the reference view. The corresponding depth and normal loss are respectively computed as:

$$\mathcal{L}_{\text{depth}} = \lambda_{\text{depth}} \cdot \frac{1}{K}\sum_{i=1}^{K} M_i^{gt} \otimes \left\|\hat{D}_i, D_i^{gt}\right\|_1 \tag{6}$$

$$\mathcal{L}_{\text{normal}} = \lambda_{\text{normal}} \cdot \frac{1}{K}\sum_{i=1}^{K} M_i^{gt} \otimes \left(1 - \hat{N}_i \cdot N_i^{gt}\right) \tag{7}$$

The overall loss function is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rgb}} + \mathcal{L}_{\text{mask}} + \mathcal{L}_{\text{depth}} + \mathcal{L}_{\text{normal}} \tag{8}$$

6 Accelerated Joint Refinement of Texture and Geometry

6.1 Training-free object-aware diffusion prior

The textured meshes generated by our MP-LRM, while highly consistent with the multi-view input images, still suffer from detail distortion due to two main factors: 1) limited network generalization caused by insufficient training data; and 2) inconsistent and blurry multi-view renderings produced by the multi-view diffusion models. To address these issues, we introduce a refinement algorithm that simultaneously enhances both texture and geometry.

Previous methods refine 3D details by optimizing the SDS loss, However, these pretrained 2D diffusion priors lack object awareness, resulting in inconsistent optimization during 3D reconstruction. DreamCraft3D tackles this problem by fine-tuning a pretrained diffusion model using multi-view image renderings through DreamBooth [86]. This approach allows the model to form a concept of the scene's surrounding views and promotes consistent texture generation. Nevertheless, training DreamBooth is computationally inefficient due to its reliance on extensive iterative processes. Its susceptibility to overfitting and the need for careful parameter tuning further increase resource demands and training time.

To overcome this limitation, we are the first to propose the use of IP-Adapter [25], a training-free, lightweight image prompt adaptation method that employs a decoupled cross-attention strategy for existing text-to-image diffusion models. Unlike DreamBooth, Ip-adapter does not require training, making it significantly more efficient and suitable for our refinement algorithm. By leveraging Ip-adapter, our method can efficiently refine the textured meshes, enhancing both texture and geometry while maintaining consistency with the multi-view input images. Notably, utilizing IP-Adapter allows our method to refine textured meshes four times faster than DreamCraft3D using DreamBooth, reducing the optimization time from 40 minutes to just 10 minutes.

To mitigate ambiguity and texture multi-head problems that arise when conditioning IP-Adapter on only the source image, we condition the model on both the source image $\mathbf{I}_{source}$ and the generated multi-view renderings $\mathbf{I}_i, i = 1, 2, ..., N$. Similar to DreamCraft3D, we employ an off-the-shelf upsampler [87] to augment multi-view renderings before conditioning. We obtain image embeddings from the IP-Adapter for all view images. During training, the IP-Adapter image embedding is selected based on the location of the randomly sampled camera, called *view-dependent image prompting*. We use a weighted combination of the image embeddings of different views depending on the value of

the azimuth angle θ_{cam}. This approach ensures that the most relevant image embedding is utilized based on the camera's position, enhancing the consistency and quality of the generated 3D reconstructions.

6.2 Joint texture-geometry refinement

Camera and light augmentations. We follow Magic3D [9] to add random augmentations to the camera and light sampling for rendering the shaded images. Differently, we sample the point light location such that the angular distance from the random camera center location (w.r.t. the origin) is sampled with a random point light distance r_{cam} in [7.5, 10], and we freeze the material augmentation unlike Dreamfusion and Magic3D, as we found it is bad for training convergence. During training, we fix the Field-of-View angle to $40°$, and sample elevation angle ϕ_{cam} from $\mathcal{U}(-10, 45)$ and azimuth angle θ_{cam} from $\mathcal{U}(-180, 180)$, and distance from the origin in [1, 1.2].

Iterative RGB and Normal rendering. DreamCraft3D alternately renders normal maps $\hat{N}$ and RGB images $\hat{I}$ as the input for diffusion guidance $\hat{I}_g$ to enhance texture and geometry disentanglement. For convenience, we ignore the view subscript. While this strategy leads to finer geometry, it can result in geometry-texture misalignment due to the absence of RGB cues in normal maps, causing divergent optimization directions for geometry and texture. To address this issue, we propose a new strategy that blends normal maps and RGB images using a random weight $\alpha \in [0, 1]$, instead of solely relying on normal maps for guidance input:

$$\hat{I}_g = \begin{cases} \hat{I} & \text{if } r < 0.5 \\ \alpha \cdot \hat{I} + (1-\alpha) \cdot \hat{N} & \text{if } r \geq 0.5 \end{cases} \tag{9}$$

where $r \sim \text{Uniform}(0, 1)$ and $\alpha \sim \text{Uniform}(0, 0.5)$ is a random variable used to weight the RGB and normal components of the output when $r \geq 0.5$. This blending approach incorporates both geometric information from the normal maps and color cues from the RGB images, promoting a more coherent optimization of geometry and texture.

6.3 Loss functions

We supervise the refinement by two parts: pixel-level loss under the reference view and the diffusion distillation loss at random views. For the diffusion distillation loss, since we adopt a training-free customized diffusion prior, the standard VSD loss in [10] is used instead of BSD in DreamCraft3D:

$$\nabla_\theta \mathcal{L}_{\text{VSD}}(\theta) = \mathbb{E}_{t,\boldsymbol{\epsilon},c}[\omega(t)(\boldsymbol{\epsilon}_{\text{ipa}}(\hat{I}_t; y^c, t) - \boldsymbol{\epsilon}_{\text{lora}}(\hat{I}_t; t, c, y))\frac{\partial \hat{I}}{\partial \theta}], \tag{10}$$

where $\hat{I}_t = \alpha_t \hat{I} + \sigma_t \boldsymbol{\epsilon}$. $\boldsymbol{\epsilon}_{\text{ipa}}(\hat{I}_t; y^c, t)$ is a pretrained diffusion model adapted to multi-view image prompts using IP-Adapter. $\boldsymbol{\epsilon}_{\text{lora}}$ is parameterized by a LoRA (Low-rank adaptation [83]) of $\boldsymbol{\epsilon}_{\text{ipa}}(\hat{I}_t; y^c, t)$, conditioned on additional camera parameter c.

The pixel-level reconstruction loss at reference view $\mathcal{L}_{recon}$ is the combination of RGB MSE loss and LPIPS loss, as well as consistency loss [88]:

$$\mathcal{L}_{total} = \lambda_{rgb}\mathcal{L}_{mse} + \lambda_{lpips}\mathcal{L}_{lpips} + \lambda_{consistency}\mathcal{L}_{consistency} \tag{11}$$

The total loss is:

$$\mathcal{L}_{total} = \mathcal{L}_{recon} + \lambda_{vsd}\mathcal{L}_{vsd}, \tag{12}$$

where $\lambda_{rgb} = \lambda_{lpips} = 10000$, $\lambda_{consistency} = 100$, $\lambda_{vsd} = 0.1$.

7 Experiments

7.1 Implementation Details

Dataset. We train our model on selected partitation of Objaverse dataset [89], which contains 240k 3D models. We use Blender to render ground-truth 512×512 images, depths and normals for an object. We normalize the shape to the box [-0.5, 0.5] in world space and render 100 random views and 6 fixed views aligned with Zero123++ V2 setting, lit with randomly selected environment maps.

For Evaluation, we adopt Google Scanned Objects (GSO) dataset [90], which includes a wide variety of high-quality scanned household items, to evaluate the performance of our method and other baselines. Following InstantMesh [19], we randomly choose 300 shapes and render 21 images of each object in an orbiting trajectory with uniform azimuths and varying elevations in $30°, 0°, 30°$. Besides, we establish a test benchmark that includes 300 images, which is a mix of real pictures and those produced by Stable Diffusion [91] and Deep Floyd. Each image in this benchmark comes with an alpha mask for the foreground, a predicted depth map, and a text prompt. For real images, the text prompts are sourced from an image caption model. We intend to make this test benchmark accessible to the public.

Architectural details. Our model MP-LRM contains 600M parameters, with two U-Nets for the multi-view images and learnable plane embeddings. The U-Nets have [64, 128, 128, 256, 256, 512, 512] channels and attention blocks at resolutions [64, 32, 16]. We generate 6-view 256×256×36 images and normal maps using Zero123++ [32]. The two learnable 256×256×32 plane embeddings are placed at the bottom and top. The plane decoder is of 5 layers with hidden dimensions 64. The Flexicubes grid size is 96 and the rendering resolution is 512 for both training and inference.

We employ DreamShaper [92] as the pretrained text-to-image diffusion model for guidance. For image prompt adaptation, we initially utilize SUPIR [87] to enhance the resolution of the multi-view images generated by Zero123++ and then apply IP-Adapter [25] to adapt the diffusion model to those upscaled images and the original input image. The IP-Adapter scale is configured to 0.8.

Training details. We train MP-LRM using 64 NVIDIA A100 (80G) GPUs with a batch size of 256 for 100 epochs, which takes approximately 5 days to complete. For each training sample, we use 6 fixed views aligned with Zero123++ as input images. To supervise the shape reconstruction, we utilize a total of 4 views: 3 randomly selected from a set of 100 views and 1 randomly chosen from the input views. During training, we use random background color augmentation. We optimize the model using the AdamW optimizer with a learning rate of 4×10^{-4} and a cosine learning rate schedule.

For the refinement, we build our system based on the foundation of threestudio [93]. We improve the Flexicubes grid size from 96 to 192 for detail sculpting. We optimize each case for 2000 iterations on one NVIDIA A100 (80G)

TABLE 1: Quantitative results on Google Scanned Objects (GSO) orbiting views.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	CD ↓	FS ↑
TripoSR	14.152	0.829	0.213	0.168	0.730
LGM	13.814	0.821	0.218	0.191	0.642
CRM	16.401	0.838	0.201	0.166	0.754
InstantMesh	16.697	0.850	0.157	0.140	0.804
DreamCraft3D	16.302	0.835	0.230	0.177	0.656
Ours (Coarse)	17.729	0.843	0.151	0.131	0.821
Ours	20.251	0.862	0.132	0.136	0.795

TABLE 2: Ablation study on the learnable plane embeddings

Method	PSNR ↑	SSIM ↑	LPIPS ↓	CD ↓	FS ↑
w/o $\mathbf{E}$	17.119	0.838	0.174	0.152	0.754
Ours	17.729	0.843	0.151	0.131	0.821

GPU with batch size of 1. We optimize the model using the AdamW optimizer with separate learning rates for different modules: $\eta_{\epsilon_{\text{lora}}} = 1e^{-4}$, $\eta_{\mathcal{G}_e} = 1e^{-2}$, $\eta_{\mathcal{F}} = 1e^{-3}$. ϵ_{lora}, $\mathcal{G}_e$, $\mathcal{F}$ represent the LoRA layer of the guidance, geometry encoding network, and Flexicubes parameters. During training, the diffusion timestep t is sampled from range [0.02, 0.4].

7.2 Comparisons with the State of the Arts

Baselines. We compare our technique against six baselines: four feed-forward methods (TripoSR [94], LGM [17], CRM [79], InstantMesh [19]) and one optimization-based method (DreamCraft3D [12]). TripoSR is the best-performing open-source LRM for single-view reconstruction. LGM and CRM are UNet-based models that reconstruct Gaussians and 3D meshes from multi-view images, respectively. InstantMesh is a Transformer-based LRM using Flexicubes. DreamCraft3D optimizes 3D reconstruction from a single image.

Quantitative comparison. We report the quantitative results on GSO [90] dataset in Table 1. For each metric, we highlight the top three results among all methods, and a deeper color indicates a better result. We use four evaluation metrics: PSNR, SSIM, LPIPS, Chamfer Distance (CD) and F-Score (with a threshold of 0.05). The first four metrics are for novel view synthesis and the last two are for geometry reconstruction quality.

From the 2D novel view synthesis metrics, we can observe that DreamCraft3D++, including our coarse model, outperforms the baselines on PSNR, SSIM and LPIPS significantly, indicating that its generation results have the best perceptually viewing quality.

As for the 3D geometric metrics, DreamCraft3D++ outperforms the baselines on both CD and FS, indicates more reliable generated shapes. Note that the result after refinement has slightly worse geometry metric performance because the objects optimize more surface detail guided by the pretrained diffusion model, which may affect chamfer distance.

Qualitative comparison. To validate the effectiveness of our method, we qualitatively compare our results with the baselines: TripoSR, LGM, CRM, InstantMesh, and DreamCraft3D. For all baselines, we use their official code and checkpoints. We first visualize images from the GSO dataset, as shown in Figure 3. The figure demonstrates that our MP-LRM generates higher quality results compared to all other baselines. This can be attributed to the efficiency of our method, which fully utilizes the spatial alignment of input posed images and outputs a multiplane-based representation. Furthermore, the refinement stage restores texture details, further improving fidelity.

We also visualize the meshes generated from a single input image using these baselines on a set of more complex internet images, as shown in Figure 4. Our improvement in 3D generation quality is even more evident in these complex cases. Compared to the one-stage methods, our MP-LRM produces smoother and more reasonable geometry with better detail. For instance, in the case of the fox and cat (second and eighth rows), LGM, CRM, and InstantMesh generate incomplete meshes, while our coarse results are complete and smooth. More importantly, the significance of our proposed refinement stage is clear. By leveraging a powerful 2D pre-trained diffusion model, it enhances texture and geometric details while maintaining a unified style consistent with the input image through multi-view conditioned object-awareness.

DreamCraft3D generates sharper texture details compared to the one-stage methods; however, it still falls short of our approach. We posit that this is due to the sensitivity of DreamBooth training to hyperparameters and training steps, which can lead to over-saturated textures. Furthermore, the iterative nature of DreamBooth and 3D refinement in DreamCraft3D may exacerbate this issue, causing the accumulation of artifacts over multiple iterations.

7.3 Ablation study

The learnable plane embeddings. In our paper, in order to address the lack of information in the top and bottom views, we introduce two additional learnable plane embeddings placed at the top and bottom of the volume, respectively. Table 2 demonstrates that this design effectively mitigates this issue by incorporating the additional embeddings, thereby reducing holes and enhancing the overall quality of the reconstructed meshes.

The strategies of the refinement. Figure 5 presents an ablation study comparing five texture optimization techniques: (1) Score Distillation Sampling (SDS), (2) Variational Score Distillation (VSD), (3) VSD with a single-view image-conditioned IP-Adapter, (4) VSD with a multi-view image-conditioned IP-Adapter, and (5) VSD with a multi-view image-conditioned IP-Adapter and iterative normal-image rendering. SDS generates novel-view textures that appear overly smooth and saturated. While VSD using standard stable diffusion produces more realistic textures, it suffers from notable inconsistencies, such as the astronaut's face appearing at the back view. Incorporating a single-view image-conditioned IP-Adapter introduces object awareness but still results in multi-face texture issues. Our proposed multi-view image-conditioned approach achieves a balance between realism and consistency, although the resulting geometry appears noisy. To address this, we employ iterative

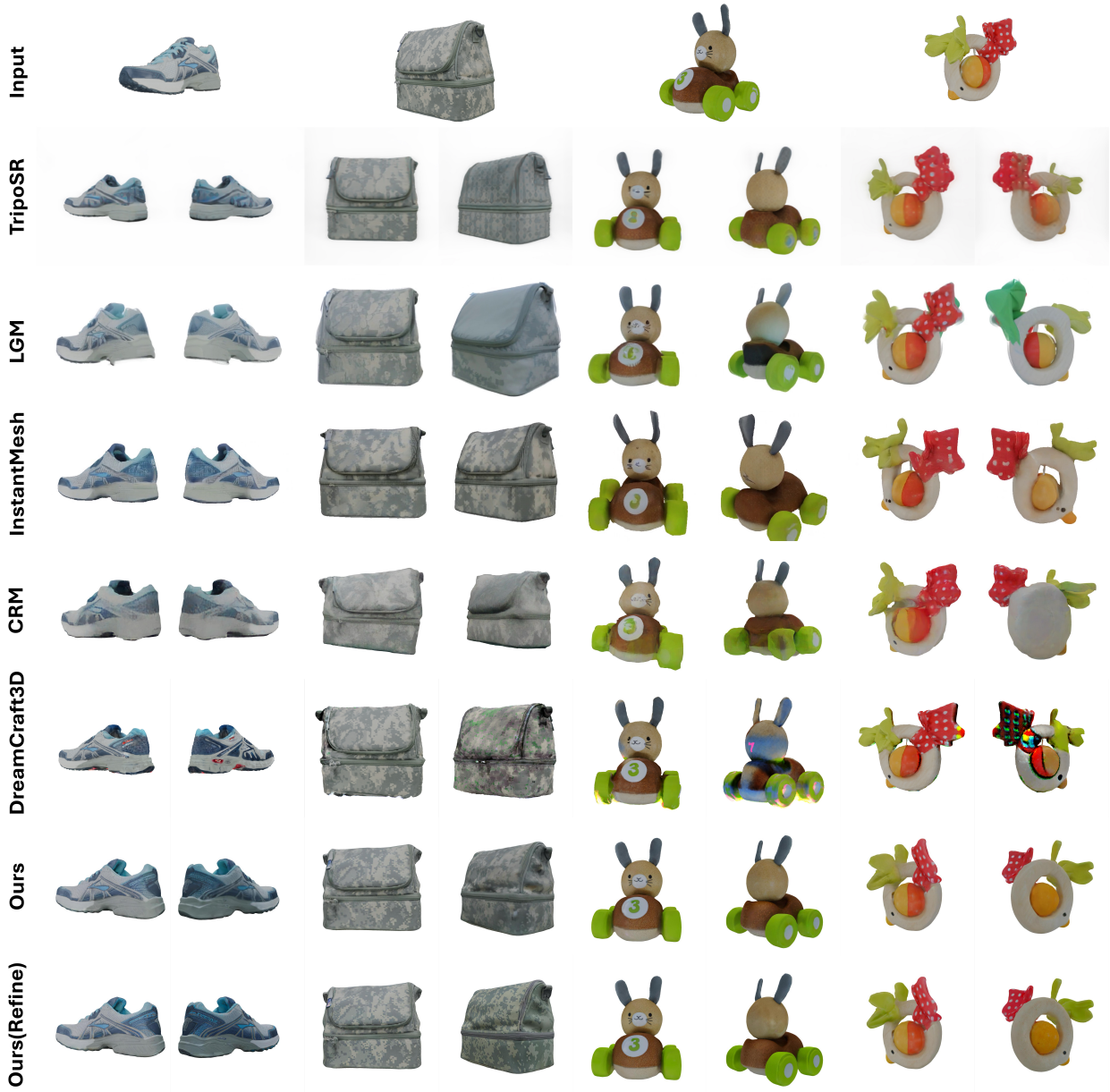


Fig. 3: Qualitative comparison with baselines on the GSO dataset.

RGB and normal rendering, which enhances the disentangled optimization of texture and geometry and refines the surface geometry.

7.4 Applications

3D rigged animation. As shown in Figure 6, we export the high-quality textured meshes for downstream applications, including rigged animation. By utilizing software such as Mixamo [95], we can efficiently employ these meshes for rigging and animation tasks, thereby streamlining the workflow and improving the overall production process.

8 Discussions and Conclusions

Limitations. Though DreamCraft3D++ demonstrates impressive capabilities in high-quality 3D generation, it is not without limitations. A significant drawback lies in the quality of multi-view images produced by Zero123++, which are directly utilized for 3D reconstruction via MP-LRM and IP-Adapter prompt conditioning. Zero123++ struggles to generate satisfactory multi-view images when presented with complex inputs or significant elevation angles. Additionally, DreamCraft3D++ outputs 3D objects with baked illumination, rendering them unsuitable for graphics pipelines that require controlled lighting conditions.

Future work. For future work, one promising direction involves the exploration of enhanced multi-view diffusion models that can deliver higher quality outputs and accommodate a broader range of elevation angles in input images. Furthermore, integrating physically-based rendering (PBR) materials into the 3D generation process could yield significant improvements. Finally, expanding the scope of

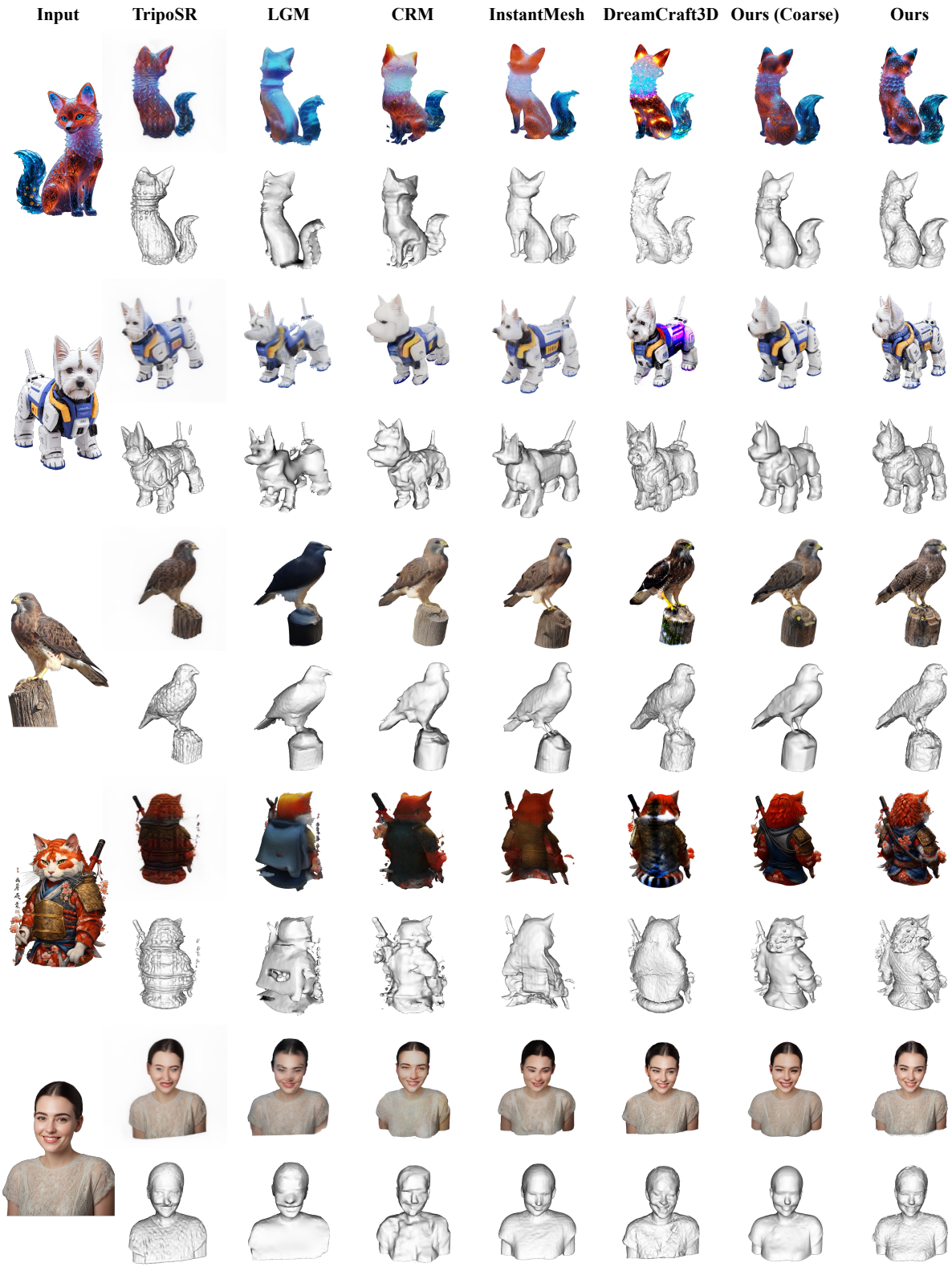


Fig. 4: Qualitative comparison with baselines on the Internet images.

3D generation from individual objects to entire scenes is essential, particularly by supporting flexible input formats such as captured video sequences or multiple unposed images.

Conclusions. In this work, we present DreamCraft3D++, a framework for efficient high-quality generation of complex 3D assets. Building on the multi-stage process of DreamCraft3D, we replace the time-consuming geometry sculpting optimization with a feed-forward, multi-plane reconstruction model, achieving a 1000x speedup. For texture

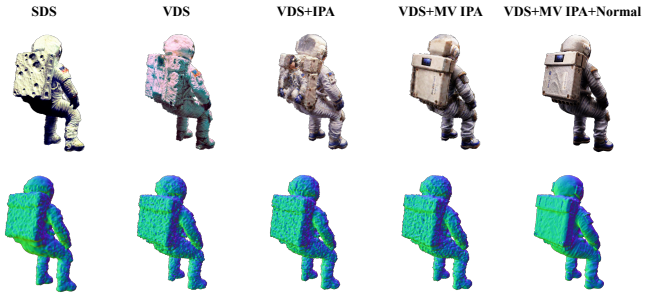


Fig. 5: Ablation study on the strategies of the refinement.

Fig. 6: We export the high-quality textured meshes, enabling seamless integration into downstream applications such as 3D rigged animation.

refinement, our training-free IP-Adapter module utilizes enhanced multi-view images to improve texture and geometry consistency, providing a solution that is four times faster than DreamCraft3D's DreamBooth fine-tuning. Our approach generates intricate 3D assets with realistic 360° textures, significantly outperforming current state-of-the-art image-to-3D methods in quality and speed.

REFERENCES

[1] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *Advances in Neural Information Processing Systems*, vol. 35, pp. 36 479–36 494, 2022. 1

[2] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, vol. 1, no. 2, p. 3, 2022. 1

[3] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695. 1

[4] S. Gu, D. Chen, J. Bao, F. Wen, B. Zhang, D. Chen, L. Yuan, and B. Guo, "Vector quantized diffusion model for text-to-image synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 696–10 706. 1

[5] T. Wang, B. Zhang, T. Zhang, S. Gu, J. Bao, T. Baltrusaitis, J. Shen, D. Chen, F. Wen, Q. Chen *et al.*, "Rodin: A generative model for sculpting 3d digital avatars using diffusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 4563–4573. 1, 4

[6] E. R. Chan, C. Z. Lin, M. A. Chan, K. Nagano, B. Pan, S. De Mello, O. Gallo, L. J. Guibas, J. Tremblay, S. Khamis *et al.*, "Efficient geometry-aware 3d generative adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 123–16 133. 1, 4

[7] H. Zhang, B. Chen, H. Yang, L. Qu, X. Wang, L. Chen, C. Long, F. Zhu, K. Du, and M. Zheng, "Avatarverse: High-quality & stable 3d avatar creation from text and pose," *arXiv preprint arXiv:2308.03610*, 2023. 1

[8] B. Poole, A. Jain, J. T. Barron, and B. Mildenhall, "Dreamfusion: Text-to-3d using 2d diffusion," *arXiv preprint arXiv:2209.14988*, 2022. 1, 3, 4

[9] C.-H. Lin, J. Gao, L. Tang, T. Takikawa, X. Zeng, X. Huang, K. Kreis, S. Fidler, M.-Y. Liu, and T.-Y. Lin, "Magic3d: High-resolution text-to-3d content creation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 3, 4, 7

[10] Z. Wang, C. Lu, Y. Wang, F. Bao, C. Li, H. Su, and J. Zhu, "Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation," *arXiv preprint arXiv:2305.16213*, 2023. 1, 4, 7

[11] G. Qian, J. Mai, A. Hamdi, J. Ren, A. Siarohin, B. Li, H.-Y. Lee, I. Skorokhodov, P. Wonka, S. 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. 1, 3, 4

[12] J. Sun, B. Zhang, R. Shao, L. Wang, W. Liu, Z. Xie, and Y. Liu, "Dreamcraft3d: Hierarchical 3d generation with bootstrapped diffusion prior," *arXiv preprint arXiv:2310.16818*, 2023. 1, 4, 8

[13] L. Zhang, Z. Wang, Q. Zhang, Q. Qiu, A. Pang, H. Jiang, W. Yang, L. Xu, and J. Yu, "Clay: A controllable large-scale generative model for creating high-quality 3d assets," *ACM Transactions on Graphics (TOG)*, vol. 43, no. 4, pp. 1–20, 2024. 1

[14] S. Wu, Y. Lin, F. Zhang, Y. Zeng, J. Xu, P. Torr, X. Cao, and Y. Yao, "Direct3d: Scalable image-to-3d generation via 3d latent diffusion transformer," *arXiv preprint arXiv:2405.14832*, 2024. 1

[15] Y. Hong, K. Zhang, J. Gu, S. Bi, Y. Zhou, D. Liu, F. Liu, K. Sunkavalli, T. Bui, and H. Tan, "Lrm: Large reconstruction model for single image to 3d," *arXiv preprint arXiv:2311.04400*, 2023. 1, 3, 4, 5

[16] P. Wang, H. Tan, S. Bi, Y. Xu, F. Luan, K. Sunkavalli, W. Wang, Z. Xu, and K. Zhang, "Pf-lrm: Pose-free large reconstruction model for joint pose and shape prediction," Nov 2023. 1, 3, 4, 5

[17] J. Tang, Z. Chen, X. Chen, T. Wang, G. Zeng, and Z. Liu, "Lgm: Large multi-view gaussian model for high-resolution 3d content creation," *ArXiv*, vol. abs/2402.05054, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:267523413 1, 4, 8

[18] Y. Xu, H. Tan, F. Luan, S. Bi, P. Wang, J. Li, Z. Shi, K. Sunkavalli, G. Wetzstein, Z. Xu, and K. Zhang, "Dmv3d: Denoising multi-view diffusion using 3d large reconstruction model," *ArXiv*, vol. abs/2311.09217, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:265213192 1, 4

[19] J. Xu, W. Cheng, Y. Gao, X. Wang, S. Gao, and Y. Shan, "Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models," *arXiv preprint arXiv:2404.07191*, 2024. 3, 5, 7, 8

[20] T. Müller, A. Evans, C. Schied, and A. Keller, "Instant neural graphics primitives with a multiresolution hash encoding," *ACM Trans. Graph.*, vol. 41, no. 4, pp. 102:1–102:15, Jul. 2022. [Online]. Available: https://doi.org/10.1145/3528223.3530127 3

[21] Z.-X. Zou, Z. Yu, Y.-C. Guo, Y. Li, D. Liang, Y.-P. Cao, and S.-H. Zhang, "Triplane meets gaussian splatting: Fast and generalizable single-view 3d reconstruction with transformers," *arXiv preprint arXiv:2312.09147*, 2023. 3, 4

[22] J. Li, H. Tan, K. Zhang, Z. Xu, F. Luan, Y. Xu, Y. Hong, K. Sunkavalli, G. Shakhnarovich, and S. Bi, "Instant3d: Fast text-to-3d with sparse-view generation and large reconstruction model," *ArXiv*, vol. abs/2311.06214, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:265128529 3, 4

[23] T. Shen, J. Munkberg, J. Hasselgren, K. Yin, Z. Wang, W. Chen, Z. Gojcic, S. Fidler, N. Sharp, and J. Gao, "Flexible isosurface extraction for gradient-based mesh optimization." *ACM Trans. Graph.*, vol. 42, no. 4, pp. 37–1, 2023. 3, 5

[24] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering," *TOG*, vol. 42, no. 4, pp. 1–14, 2023. 3, 5

[25] H. Ye, J. Zhang, S. Liu, X. Han, and W. Yang, "Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models," *arXiv preprint arXiv:2308.06721*, 2023. 3, 6, 7

[26] J. Tang, T. Wang, B. Zhang, T. Zhang, R. Yi, L. Ma, and D. Chen, "Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior," *arXiv preprint arXiv:2303.14184*, 2023. 3, 4

[27] K. Sargent, J. Y. Koh, H. Zhang, H. Chang, C. Herrmann, P. Srinivasan, J. Wu, and D. Sun, "Vq3d: Learning a 3d-aware generative model on imagenet," *arXiv preprint arXiv:2302.06833*, 2023. 3

[28] I. Skorokhodov, A. Siarohin, Y. Xu, J. Ren, H.-Y. Lee, P. Wonka, and S. Tulyakov, "3d generation on imagenet," *arXiv preprint arXiv:2303.01416*, 2023. 3, 4

[29] J. Xiang, J. Yang, B. Huang, and X. Tong, "3d-aware image generation using 2d diffusion models," *arXiv preprint arXiv:2303.17905*, 2023. 3

[30] D. Watson, W. Chan, R. Martin-Brualla, J. Ho, A. Tagliasacchi, and M. Norouzi, "Novel view synthesis with diffusion models," *arXiv preprint arXiv:2210.04628*, 2022. 4

[31] R. Liu, R. Wu, B. Van Hoorick, P. Tokmakov, S. Zakharov, and C. Vondrick, "Zero-1-to-3: Zero-shot one image to 3d object," *arXiv preprint arXiv:2303.11328*, 2023. 4

[32] R. Shi, H. Chen, Z. Zhang, M. Liu, C. Xu, X. Wei, L. Chen, C. Zeng, and H. Su, "Zero123++: a single image to consistent multi-view diffusion base model," *arXiv preprint arXiv:2310.15110*, 2023. 4, 5, 7

[33] V. S. Voleti, C.-H. Yao, M. Boss, A. Letts, D. Pankratz, D. Tochilkin, C. Laforte, R. Rombach, and V. Jampani, "Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion," *ArXiv*, vol. abs/2403.12008, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:268531375 4

[34] Y. Shi, P. Wang, J. Ye, M. Long, K. Li, and X. Yang, "Mvdream: Multi-view diffusion for 3d generation," *ArXiv*, vol. abs/2308.16512, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:261395233 4

[35] P. Wang and Y. Shi, "Imagedream: Image-prompt multi-view diffusion for 3d generation," *ArXiv*, vol. abs/2312.02201, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:265659122 4

[36] M. Liu, C. Xu, H. Jin, L. Chen, M. Varma T, Z. Xu, and H. Su, "One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization," in *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36. Curran Associates, Inc., 2023, pp. 22 226–22 246. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/file/4683beb6bab325650db13afd05d1a14a-Paper-Conference.pdf 4

[37] L. Qiu, G. Chen, X. Gu, Q. Zuo, M. Xu, Y. Wu, W. Yuan, Z. Dong, L. Bo, and X. Han, "Richdreamer: A generalizable normal-depth diffusion model for detail richness in text-to-3d," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 9914–9925. 4

[38] A. Jain, B. Mildenhall, J. T. Barron, P. Abbeel, and B. Poole, "Zero-shot text-guided object generation with dream fields," 2022. 4

[39] H.-H. Lee and A. X. Chang, "Understanding pure clip guidance for voxel grid nerf models," *arXiv preprint arXiv:2209.15172*, 2022. 4

[40] F. Hong, M. Zhang, L. Pan, Z. Cai, L. Yang, and Z. Liu, "Avatarclip: Zero-shot text-driven generation and animation of 3d avatars," *ACM Transactions on Graphics (TOG)*, vol. 41, no. 4, pp. 1–19, 2022. 4

[41] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763. 4

[42] H. Wang, X. Du, J. Li, R. A. Yeh, and G. Shakhnarovich, "Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation," *arXiv preprint arXiv:2212.00774*, 2022. 4

[43] R. Chen, Y. Chen, N. Jiao, and K. Jia, "Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation," *arXiv preprint arXiv:2303.13873*, 2023. 4

[44] Z. Liu, Y. Li, Y. Lin, X. Yu, S. Peng, Y.-P. Cao, X. Qi, X. Huang, D. Liang, and W. Ouyang, "Unidream: Unifying diffusion priors for relightable text-to-3d generation," *arXiv preprint arXiv:2312.08754*, 2023. 4

[45] Z. Huang, H. Wen, J. Dong, Y. Wang, Y. Li, X. Chen, Y.-P. Cao, D. Liang, Y. Qiao, B. Dai *et al.*, "Epidiff: Enhancing multi-view synthesis via localized epipolar-constrained diffusion," *arXiv preprint arXiv:2312.06725*, 2023. 4

[46] G. Metzer, E. Richardson, O. Patashnik, R. Giryes, and D. Cohen-Or, "Latent-nerf for shape-guided generation of 3d shapes and textures," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 12 663–12 673. 4

[47] C. Deng, C. Jiang, C. R. Qi, X. Yan, Y. Zhou, L. Guibas, D. Anguelov *et al.*, "Nerdi: Single-view nerf synthesis with language-guided diffusion as general image priors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 20 637–20 647. 4, 6

[48] M. Liu, C. Xu, H. Jin, L. Chen, Z. Xu, H. 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. 4

[49] N. Huang, T. Zhang, Y. Yuan, D. Chen, and S. Zhang, "Customize-it-3d: High-quality 3d creation from a single image using subject-specific knowledge prior," *arXiv preprint arXiv:2312.11535*, 2023. 4

[50] Y. Liu, C.-H. Lin, Z. Zeng, X. Long, L. Liu, T. Komura, and W. Wang, "Syncdreamer: Generating multiview-consistent images from a single-view image," *ArXiv*, vol. abs/2309.03453, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:261582503 4

[51] J. Tang, J. Ren, H. Zhou, Z. Liu, and G. Zeng, "Dreamgaussian: Generative gaussian splatting for efficient 3d content creation," *ArXiv*, vol. abs/2309.16653, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:263131552 4

[52] A. Raj, S. Kaza, B. Poole, M. Niemeyer, N. Ruiz, B. Mildenhall, S. Zada, K. Aberman, M. Rubinstein, J. Barron *et al.*, "Dreambooth3d: Subject-driven text-to-3d generation," *arXiv preprint arXiv:2303.13508*, 2023. 4

[53] E. R. Chan, M. Monteiro, P. Kellnhofer, J. Wu, and G. Wetzstein, "pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 5799–5809. 4

[54] C. Xie, C. Wang, B. Zhang, H. Yang, D. Chen, and F. Wen, "Style-based point generator with adversarial rendering for point cloud completion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4619–4628. 4

[55] X. Zeng, A. Vahdat, F. Williams, Z. Gojcic, O. Litany, S. Fidler, and K. Kreis, "Lion: Latent point diffusion models for 3d shape generation," *arXiv preprint arXiv:2210.06978*, 2022. 4

[56] J. Gao, T. Shen, Z. Wang, W. Chen, K. Yin, D. Li, O. Litany, Z. Gojcic, and S. Fidler, "Get3d: A generative model of high quality 3d textured shapes learned from images," *Advances In Neural Information Processing Systems*, vol. 35, pp. 31 841–31 854, 2022. 4

[57] J. Tang, B. Zhang, B. Yang, T. Zhang, D. Chen, L. Ma, and F. Wen, "Explicitly controllable 3d-aware portrait generation," *arXiv preprint arXiv:2209.05434*, 2022. 4

[58] J. Sun, X. Wang, L. Wang, X. Li, Y. Zhang, H. Zhang, and Y. Liu, "Next3d: Generative neural texture rasterization for 3d-aware head avatars," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 20 991–21 002. 4

[59] J. Sun, X. Wang, Y. Shi, L. Wang, J. Wang, and Y. Liu, "Ide-3d: Interactive disentangled editing for high-resolution 3d-aware portrait synthesis," *ACM Trans. Graph.*, vol. 41, no. 6, nov 2022. [Online]. Available: https://doi.org/10.1145/3550454.3555506 4

[60] A. Sanghi, H. Chu, J. G. Lambourne, Y. Wang, C.-Y. Cheng, M. Fumero, and K. R. Malekshan, "Clip-forge: Towards zero-shot text-to-shape generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 603–18 613. 4

[61] P. Mittal, Y.-C. Cheng, M. Singh, and S. Tulsiani, "Autosdf: Shape priors for 3d completion, reconstruction and generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 306–315. 4

[62] X. Yan, L. Lin, N. J. Mitra, D. Lischinski, D. Cohen-Or, and H. Huang, "Shapeformer: Transformer-based shape completion

via sparse representation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 6239–6249. 4

[63] B. Zhang, M. Nießner, and P. Wonka, "3dilg: Irregular latent grids for 3d generative modeling," in *NeurIPS*, 2022. 4

[64] W. Yu, X. Qian, J. Huo, T. Huang, B. Zhao, and Y. Fu, "Pushing the limits of 3d shape generation at scale," *arXiv preprint arXiv:2306.11510*, 2023. 4

[65] Y.-C. Cheng, H.-Y. Lee, S. Tulyakov, A. G. Schwing, and L.-Y. Gui, "Sdfusion: Multimodal 3d shape completion, reconstruction, and generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 4456–4465. 4

[66] M. Li, Y. Duan, J. Zhou, and J. Lu, "Diffusion-sdf: Text-to-shape via voxelized diffusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 12 642–12 651. 4

[67] G. Nam, M. Khlifi, A. Rodriguez, A. Tono, L. Zhou, and P. Guerrero, "3d-ldm: Neural implicit 3d shape generation with latent diffusion models," *arXiv preprint arXiv:2212.00842*, 2022. 4

[68] B. Zhang, J. Tang, M. Niessner, and P. Wonka, "3dshape2vecset: A 3d shape representation for neural fields and generative diffusion models," *arXiv preprint arXiv:2301.11445*, 2023. 4

[69] A. Nichol, H. Jun, P. Dhariwal, P. Mishkin, and M. Chen, "Point-e: A system for generating 3d point clouds from complex prompts," *arXiv preprint arXiv:2212.08751*, 2022. 4

[70] H. Jun and A. Nichol, "Shap-e: Generating conditional 3d implicit functions," *arXiv preprint arXiv:2305.02463*, 2023. 4

[71] M. A. Bautista, P. Guo, S. Abnar, W. Talbott, A. Toshev, Z. Chen, L. Dinh, S. Zhai, H. Goh, D. Ulbricht *et al.*, "Gaudi: A neural architect for immersive 3d scene generation," *Advances in Neural Information Processing Systems*, vol. 35, pp. 25 102–25 116, 2022. 4

[72] A. Gupta, W. Xiong, Y. Nie, I. Jones, and B. Oğuz, "3dgen: Triplane latent diffusion for textured mesh generation," *arXiv preprint arXiv:2303.05371*, 2023. 4

[73] X. Long, Y.-C. Guo, C. Lin, Y. Liu, Z. Dou, L. Liu, Y. Ma, S.-H. Zhang, M. Habermann, C. Theobalt *et al.*, "Wonder3d: Single image to 3d using cross-domain diffusion," *arXiv preprint arXiv:2310.15008*, 2023. 4

[74] Y.-T. Liu, G. Luo, H. Sun, W. Yin, Y.-C. Guo, and S.-H. Zhang, "Pi3d: Efficient text-to-3d generation with pseudo-image diffusion," *arXiv preprint arXiv:2312.09069*, 2023. 4

[75] L. Zhang, Z. Wang, Q. Zhang, Q. Qiu, A. Pang, H. Jiang, W. Yang, L. Xu, and J. Yu, "Clay: A controllable large-scale generative model for creating high-quality 3d assets," *ArXiv*, vol. abs/2406.13897, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:270619933 4

[76] S. Wu, Y. Lin, F. Zhang, Y. Zeng, J. Xu, P. Torr, X. Cao, and Y. Yao, "Direct3d: Scalable image-to-3d generation via 3d latent diffusion transformer," *ArXiv*, vol. abs/2405.14832, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:269982452 4

[77] K. Zhang, S. Bi, H. Tan, Y. Xiangli, N. Zhao, K. Sunkavalli, and Z. Xu, "Gs-lrm: Large reconstruction model for 3d gaussian splatting," *arXiv*, 2024. 4, 5

[78] Y. Siddiqui, T. Monnier, F. Kokkinos, M. Kariya, Y. Kleiman, E. Garreau, O. Gafni, N. V. Neverova, A. Vedaldi, R. Shapovalov, and D. Novotny, "Meta 3d assetgen: Text-to-mesh generation with high-quality geometry, texture, and pbr materials," 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:270878069 4

[79] Z. Wang, Y. Wang, Y. Chen, C. Xiang, S. Chen, D. Yu, C. Li, H. Su, and J. Zhu, "Crm: Single image to 3d textured mesh with convolutional reconstruction model," *ArXiv*, vol. abs/2403.05034, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:268297409 4, 8

[80] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su *et al.*, "Shapenet: An information-rich 3d model repository," *arXiv preprint arXiv:1512.03012*, 2015. 4

[81] M. Deitke, D. Schwenk, J. Salvador, L. Weihs, O. Michel, E. VanderBilt, L. Schmidt, K. Ehsani, A. Kembhavi, and A. Farhadi, "Objaverse: A universe of annotated 3d objects," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 13 142–13 153. 4

[82] T. Wu, J. Zhang, X. Fu, Y. Wang, J. Ren, L. Pan, W. Wu, L. Yang, J. Wang, C. Qian *et al.*, "Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 803–814. 4

[83] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021. 4, 7

[84] Y. Xu, Z. Shi, W. Yifan, H. Chen, C. Yang, S. Peng, Y. Shen, and G. Wetzstein, "Grm: Large gaussian reconstruction model for efficient 3d reconstruction and generation," *arXiv preprint arXiv:2403.14621*, 2024. 5

[85] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021. 5

[86] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22 500–22 510. 6

[87] F. Yu, J. Gu, Z. Li, J. Hu, X. Kong, X. Wang, J. He, Y. Qiao, and C. Dong, "Scaling up to excellence: Practicing model scaling for photo-realistic image restoration in the wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 25 669–25 680. 6, 7

[88] J. Munkberg, J. Hasselgren, T. Shen, J. Gao, W. Chen, A. Evans, T. Müller, and S. Fidler, "Extracting triangular 3d models, materials, and lighting from images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8280–8290. 7

[89] M. Deitke, R. Liu, M. Wallingford, H. Ngo, O. Michel, A. Kusupati, A. Fan, C. Laforte, V. Voleti, S. Y. Gadre *et al.*, "Objaverse-xl: A universe of 10m+ 3d objects," *Advances in Neural Information Processing Systems*, vol. 36, 2024. 7

[90] L. Downs, A. Francis, N. Koenig, B. Kinman, R. Hickman, K. Reymann, T. B. McHugh, and V. Vanhoucke, "Google scanned objects: A high-quality dataset of 3d scanned household items," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 2553–2560. 7, 8

[91] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," 2021. 7

[92] -, "Dreamshaper (revision 8c1bfc6)," 2023. [Online]. Available: https://huggingface.co/Lykon/DreamShaper 7

[93] Y.-C. Guo, Y.-T. Liu, R. Shao, C. Laforte, V. Voleti, G. Luo, C.-H. Chen, Z.-X. Zou, C. Wang, Y.-P. Cao, and S.-H. Zhang, "threestudio: A unified framework for 3d content generation," https://github.com/threestudio-project/threestudio, 2023. 7

[94] D. Tochilkin, D. Pankratz, Z. Liu, Z. Huang, A. Letts, Y. Li, D. Liang, C. Laforte, V. Jampani, and Y.-P. Cao, "Triposr: Fast 3d object reconstruction from a single image," *ArXiv*, vol. abs/2403.02151, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:268248244 8

[95] Adobe. (2024) Mixamo. [Online]. Available: https://www.mixamo.com/#/ 9

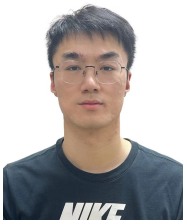
Jingxiang Sun received the MS degree from Department of ECE, University of Illinois at Urbana-Champaign in Jan 2022. He is currently working toward the PhD degree with the Department of Automation, Tsinghua University, advised by Prof. Yebin Liu. His current research interests include computer vision and computer graphics.

Cheng Peng received the BS degree from the Department of Automation, Tsinghua University in July 2024. He is currently working toward the PhD degree with the Department of Automation, Tsinghua University, advised by Prof. Yebin Liu. His current research interests include computer vision and computer graphics.

Ruizhi Shao received the BS degree from Nankai University in July 2020. He is currently working toward the PhD degree with the Department of Automation, Tsinghua University, advised by Prof. Yebin Liu. His current research interests include computer vision and computer graphics.

Yuan-chen Guo is a research scientist at VAST. He received his PhD degree from the Department of Computer Science and Technology of Tsinghua University in 2024. His research interests include computer graphics and computer vision.

Xiaochen Zhao received the B.E. degree from the Automation Department, Tsinghua University, Beijing, China, in 2019. He is currently a Ph.D. student from the Automation Department, Tsinghua University, Beijing, China. His research areas include computer vision, computer graphics, and computational photography.



Yangguang Li is the R&D Director at VAST, with a primary research focus on 3D generation. Prior to this, he conducted research in 3D detection and multi-modal direction at Shanghai Artificial Intelligence Laboratory and SenseTime.

Yanpei Cao (Member, IEEE) received his bachelor's and Ph.D. degrees in computer science from Tsinghua University in 2013 and 2018, respectively. He is currently the Head of Research and co-founder at VAST. His research interests include computer graphics and 3D computer vision.

Bo Zhang is a ZJU 100 Young Professor (Ph.D. supervisor) of Zhejiang University. He received B.E. degree from Zhejiang University in 2013 and Ph.D. degree with the Department of Electronic and Computer Engineering at Hong Kong University of Science and Technology (HKUST) in 2019. His research interest involves 2D/3D content creation, virtual human modeling, multimodal models, and embodied intelligence.

Yebin Liu received the B.E. degree from the Beijing University of Posts and Telecommunications, China, in 2002 and the Ph.D. degree from the Automation Department, Tsinghua University, Beijing, China, in 2009. He is currently a full professor with Tsinghua University. He was a research fellow in the Computer Graphics Group of the Max Planck Institute for Informatik, Germany, in 2010. His research areas include computer vision, computer graphics, and computational photography.