MPOD123: One Image to 3D Content Generation Using Mask-enhanced Progressive Outline-to-Detail Optimization

Jimin Xu[1] Tianbao Wang[1] Tao Jin[1†] Shengyu Zhang[1†] Dongjie Fu[1] Zhe Wang[2]
Jiangjing Lyu[2] Chengfei Lv[1,2] Chaoyue Niu[3] Zhou Yu[4] Zhou Zhao[1†] Fei Wu[1†]
[1]Zhejiang University, [2]Alibaba Group, [3]Shanghai Jiaotong University, [4]Hangzhou Dianzi University
{xujimin, jint_zju, sy_zhang, zhaozhou, wufei}@zju.edu.cn

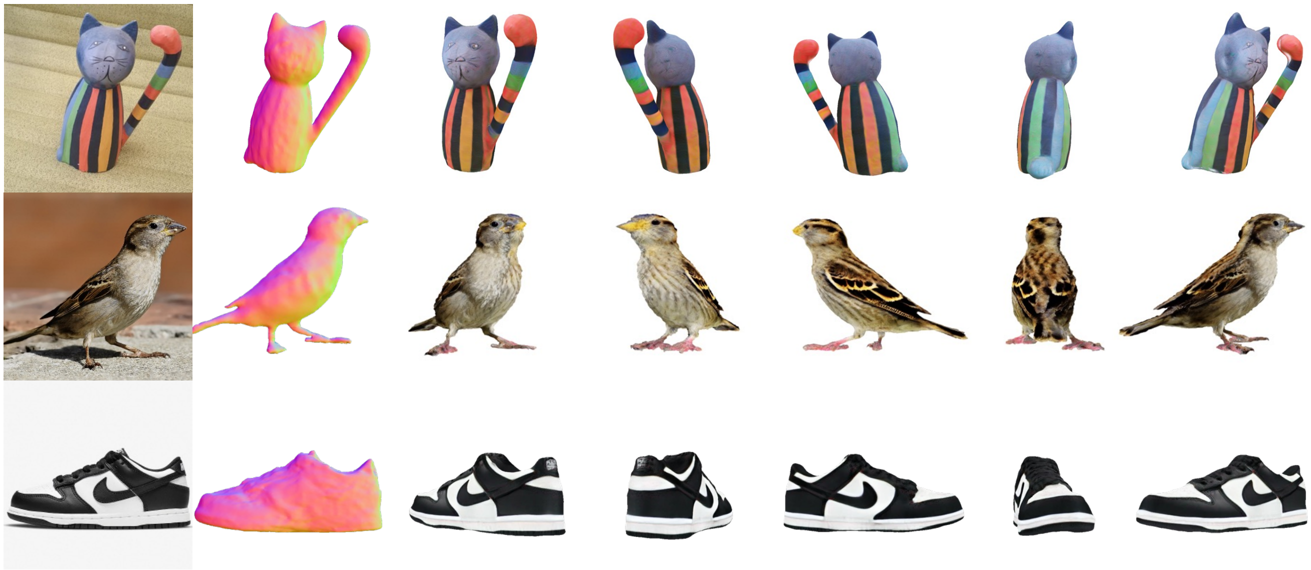

Figure 1. **MPOD123** generates high-quality 3D content from a single image (left column). We show the normal map and renderings from full 360° viewpoint transformations, showcasing plausible geometry and visually pleasing textures at novel views. Visit our project page https://mpod-123.github.io/ for an immersive visualization.

Abstract

Recent advancements in single image driven 3D content generation have been propelled by leveraging prior knowledge from pretrained 2D diffusion models. However, the 3D content generated by existing methods often exhibits distorted outline shapes and inadequate details. To solve this problem, we propose a novel framework called Mask-enhanced Progressive Outline-to-Detail optimization (aka. MPOD123), which consists of two stages. Specifically, in the first stage, MPOD123 utilizes the pretrained view-conditioned diffusion model to guide the outline shape optimization of the 3D content. Given certain viewpoint, we estimate outline shape priors in the form of 2D mask from the 3D content by leveraging opacity calculation. In the second stage, MPOD123 incorporates Detail Appearance Inpainting (DAI) to guide the refinement on local geometry and texture with the shape priors. The essence of DAI lies in the Mask Rectified Cross-Attention (MRCA), which can be conveniently plugged in the stable diffusion model. The MRCA module utilizes the mask to rectify the attention map from each cross-attention layer. Accompanied with this new module, DAI is capable of guiding the detail refinement of the 3D content, while better preserves the outline shape. To assess the applicability in practical scenarios, we contribute a new dataset modeled on real-world e-commerce environments. Extensive quantitative and qualitative experiments on this dataset and open benchmarks demonstrate

[†]Corresponding Authors.

the effectiveness of MPOD123 over the state-of-the-arts.

1. Introduction

3D content has been in high demand for a variety of applications, including architecture, animation, gaming, and the rapidly evolving fields of virtual and augmented reality. In the e-commerce scenery, the application of 3D content has also been growing, enabling richer online interactive experiences between users and businesses, such as 3D displays and personalized customization of products. However, the creation of such 3D content often requires the expertise of professional 3D artists, resulting in significant time and cost investments. As a result, there has been a growing interest in the development of automated 3D content generation techniques [13, 14, 27, 31, 43].

Single image driven 3D content generation is a crucial task for automated 3D content generation in various applications, while inferring complete 3D structures from a single image without any pre-existing prior knowledge is inherently challenging. Therefore, to acquire the necessary prior knowledge, One line of approaches [25, 47] leverage prior knowledge from pretrained 2D diffusion models, such as stable diffusion [38], to guide the generation of 3D content from a single image. The 3D content generated by these approaches often exhibits shape distortion and only demonstrates high-quality textures within a limited range of viewpoint changes, typically attributed to viewpoint biases in stable diffusion [22]. Some works [22, 55] introduce viewpoint control to diffusion models by training on 3D datasets [4, 8], which help the guidance of generating 3D content with faithful shape. However, the texture quality of the generated 3D content declines, particularly for input image that is out-of-distribution of the synthetic 3D datasets.

To enhance the generation quality of both shape and texture, a straightforward approach is to combine the prior knowledge from view-conditioned diffusion model and the stable diffusion model. However, a direct combination strategy often leads to entanglement between the two types of prior knowledge. To address these issues, we propose a novel framework called Mask-enhanced Progressive Outline-to-Detail optimization (aka. MPOD123). The key insight of our framework lies in disentangling the outline shape optimization and detail appearance optimization during the generation process of 3D content. Firstly, MPOD123 manages to derive shape priors from pretrained 2D diffusion models to guide the outline shape optimization of the 3D content. While this stage ensures a faithful representation of the outline shape, it might result in relatively inferior details in terms of local geometry and textures. Subsequently, MPOD123 employs appearance priors derived from 2D diffusion models, to guide the refinement of the detail appearance in the 3D content.

Specifically, in the first stage, we utilize pretrained view-conditioned diffusion model (Zero-1-to-3 [22]) to guide the outline shape optimization of the 3D content represented by Neural Radiance Fields (NeRF) [27]. Given a particular viewpoint, we leverage opacity calculation to construct 2D mask from the optimized 3D NeRF. Outline shape priors can be transferred to the second stage, with these 2D masks. In the second stage, we initialize a textured 3D mesh from the NeRF using DMTet [41], and devise Detail Appearance Inpainting (DAI) approach to guide the optimization on local geometry and texture. The essence of DAI lies in the Mask Rectified Cross-Attention (MRCA) module, which can be conveniently plugged in the stable diffusion model. The MRCA module utilizes the mask to rectify the attention map from each cross-attention layer. It enhances the injection of semantics from diffusion priors into the outline shape mask, ensuring overall semantic consistency between the generated content and the mask. Accompanied with this new module, DAI is capable of guiding the detail refinement of the 3D content, while better preserve the outline shape. We evaluate MPOD123 on existing dataset and our collected e-commerce setting dataset. Qualitative and quantitative experiments show that MPOD123 significantly outperforms state-of-the-art method.

We summarize our contributions as follows:

- To generate 3D content with plausible geometry and visually pleasing textures, we propose a novel framework called MPOD123, which disentangles the outline shape optimization and detail appearance optimization during the generation process.
- By leveraging opacity calculation, we build the 2D masks, transferring outline shape priors between two optimization stages.
- Based on the shape priors, we utilize Detail Appearance Inpainting approach with novel mask rectification mechanism to guide the refinement on local geometry and textures.
- Qualitative and quantitative experiments show that MPOD123 significantly outperforms state-of-the-art method on existing dataset and our collected e-commerce setting dataset.

2. Related work

2.1. Few-view Reconstruction

The early works in 3D reconstruction primarily rely on the principles of multi-view geometry [12, 33, 45]. In recent years, as data-driven paradigms like Neural Radiance Fields (NeRF) [27] gain prominence, there has been a surge in research aimed at unifying the reconstruction of object texture and geometry in scenarios with limited input images. In situations where only a few input images are available, several approaches attempt to enhance the efficiency of NeRF-like models through the incorporation of various priors [14, 31, 48]. Some methods train generalized NeRF models capable of accommodating multiple object categories [18, 37, 56, 58, 59, 62]. However, these methods often suffer from limitations in terms of generalization and precision when it comes to synthesizing novel views.

2.2. End-to-end single-view reconstruction

Generating novel views from a single image is a highly challenging task due to its inherently ill-posed nature, which requires accurate geometric estimation and occlusion handling for both geometry and texture [61]. To recover complete radiance fields from single images, some researchers train category-specific models with multi-view data. For example, [7, 9, 57, 60] learn volumetric representations for reconstruction from synthetic 3D datasets like ShapeNet [4]. Furthermore, from the perspective of incorporating prior knowledge, some works [11, 42, 42, 44, 49, 50] explore the utilization of depth information from images as prior conditions. From the perspective of model architecture design [54, 64], some works introduce methods based on 3D generative adversarial networks [2, 3, 10, 30, 32, 40] and 3D diffusion models [1, 3, 5, 6, 15, 29, 53]. However, the performance of these methods is often constrained by the availability of 3D training data.

2.3. 3D generation guided by 2D diffusion models

To address the challenge of limited 3D data, researchers explore the use of more accessible 2D data to aid in the training of 3D models. [20, 34, 52] utilize Score Distillaion Sampling (SDS) loss to enable the use of a 2D diffusion model [39] as a prior for optimization of a NeRF model. RealFusion [25] is an early attempt to apply diffusion priors to guide the single image to 3D task. Make-it-3D [47] introduces a two-stage optimization pipeline from the perspective of 3D scene representation: from coarse NeRF model to textured point clouds. The 3D content generated by Make-it-3D demonstrates high-quality textures only within a limited range of viewpoint changes, typically attributed to shape distortion. Zero-1-to-3 [22] fine-tunes a pretrained stable diffusion model on the 3D dataset Objaverse [8], resulting in the view-conditioned diffusion model, which helps the guidance of generating 3D content with faithful shape. Magic123 [35] enhances the quality of geometry and texture generation by combining the prior knowledge from Zero-1-to-3 and the stable diffusion model.

Existing approaches give relatively less consideration to effectively leverage prior knowledge from diffusion models to facilitate the generation of high-quality 3D content, encompassing both plausible geometry and visually pleasing textures. Our approach addresses this shortcoming by disentangling the outline shape optimization and detail appearance optimization during the generation process.

3. Method

3.1. Overall

We propose a two-stage progressive optimization framework for generating 3D content from a single reference image as shown in Figure 2. The Score Distillaion Sampling (SDS) loss [34] based on probability density distillation is employed for the optimization process. The key insight of our framework lies in disentangling the outline shape optimization and detail appearance optimization during the generation process of 3D content. Accompanied with our framework, knowledge from pretrained 2D diffusion models can be leveraged progressively, facilitating the high-quality 3D content generation with plausible geometry and visually pleasing textures.

Outline shape optimization. At the first stage, we aim to obtain a coarse 3D content with faithful outline shape. Since generating 3D content from a single reference image is insufficient without any priors, we leverage Zero-1-to-3 [21], a view-conditioned diffusion model, as our first stage priors. Zero-1-to-3 takes input image, relative camera rotation R and translation T for target view as conditional inputs. In contrast to stable diffusion, which exhibits viewpoint bias [21], Zero-1-to-3 enables more precise control over the object pose (identity and orientation) during the generation process. Consequently, Zero-1-to-3 is well-suited to guide the outline shape optimization, aligned with the objective of generating 3D content with faithful outline shape in this stage.

To leverage the priors from pretrained Zero-1-to-3 model, we adopt the score distillation sampling (SDS) loss proposed by DreamFusion [34]. Given an image $\boldsymbol{I}$ rendered from a novel viewpoint, SDS loss is formulated as:

$$\nabla_\theta \mathcal{L}_{\text{SDS}} = \mathbb{E}_{t,\boldsymbol{\epsilon}} \left[w(t)(\boldsymbol{\epsilon}_\phi(\boldsymbol{z}_t; \boldsymbol{y}, t) - \boldsymbol{\epsilon}) \frac{\partial \boldsymbol{z}}{\partial \boldsymbol{I}} \frac{\partial \boldsymbol{I}}{\partial \theta} \right], \quad (1)$$

where $\boldsymbol{y}$ is the conditional inputs, $\boldsymbol{z}_t$ is the noisy latent by adding a random Gaussian noise of a time step t to the latent $\boldsymbol{z}$ of image $\boldsymbol{I}$. $\boldsymbol{\epsilon}$, $\boldsymbol{\epsilon}_\phi$, ϕ, θ are the added noise, predicted noise, parameters of the 2D diffusion model, and the parameters of the 3D model.

Further, to ensure the image rendered from the reference view is fitted to the input image, we minimize the mean squared error (MSE) as follows:

$$\mathcal{L}_{ref} = \|\mathbf{I}^r - \mathcal{G}_\theta(\mathbf{v}^r)\|_2^2, \quad (2)$$

where $\mathbf{I}^r$ is the reference input image, θ is the 3D model parameters to be optimized, $\mathcal{G}_\theta(\mathbf{v}^r)$ is 3D model rendered view from reference viewpoint $\mathbf{v}^r$.

By leveraging opacity calculation, we can build the 2D mask of the generated coarse 3D content, given certain viewpoint. Outline shape priors can be transferred to the

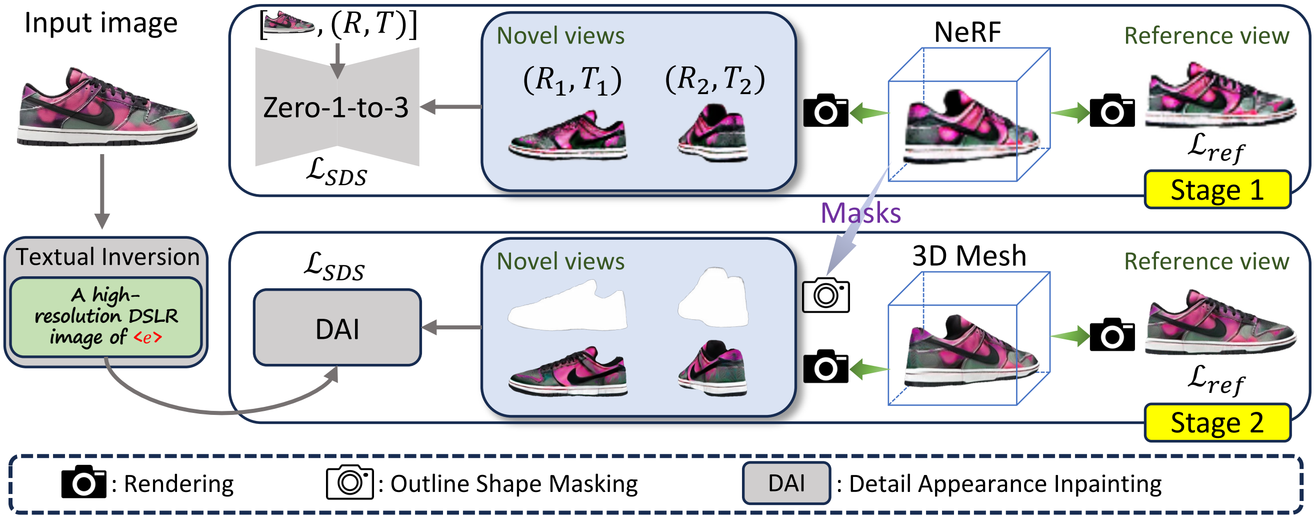

Figure 2. **Overview of MPOD123.** We generate high-quality 3D content from an input image in a progressive optimization manner. At the first stage, we utilize a view-conditioned diffusion model (Zero-1-to-3) to guide the optimization of neural radiance field (NeRF) in novel views. For a certain viewpoint, relative viewpoint transformation (R, T) and input image are used as conditional information of Zero-1-to-3. At the second stage, we initialize a textured 3D mesh from the NeRF. We utilize our Detail Appearance Inpainting approach (DAI) to guide the optimization of the 3D mesh in novel views. For a certain viewpoint, DAI takes two conditional inputs: 2D mask built from the NeRF in the same viewpoint and text prompt derived from the input image. We impose a loss $\mathcal{L}_{ref}$ in both stages to ensure the image rendered from the reference view is fitted to the input image.

second stage, with these 2D masks. More details in Outline Shape Masking can be found in 3.2.

Detail appearance optimization. After the outline shape optimization, we obtain a coarse 3D content with faithful outline shape, but it often displays worse details in local geometry and textures. Further optimization is thus desired for refinement of the coarse 3D content. To rectify detail appearance while preserving the outline shape of the coarse 3D content, we introduce a novel approach, Detail Appearance Inpainting (DAI), to guide the further optimization of coarse 3D content. Our key insight for the DAI approach involves deriving appearance priors from a pretrained 2D diffusion model. By leveraging the priors, DAI can effectively guide the enhancement of the detail appearance in the coarse 3D content. More details in Detail Appearance Inpainting can be found in 3.3.

3.2. Outline Shape Masking

Given a certain viewpoint, Outline Shape Masking aims to build the 2D mask of the generated coarse 3D content. The 2D mask represents the outline shape of the 3D content from that viewpoint.

Mask building. Rather than employing volume rendering [26] to determine the color of each pixel from a rendered view, we compute opacity using Eq. (3) to mitigate the influence of color bias on the outline shape.

$$\mathrm{OP}(\mathbf{r}) = \int_{t_n}^{t_f} \exp\left(-\int_{t_n}^{t} \sigma(\mathbf{r}(s))ds\right) \sigma(\mathbf{r}(t))dt, \quad (3)$$

where $\mathbf{r}(t)$ denotes the camera ray from near bounds t_n to far bounds t_f, and $\sigma(\mathbf{x})$ denotes volume density obtained from the NeRF representation of the coarse 3D content.

To numerically estimate this continuous integral, we sample a set of 3D points $\{t_i\}_{i=1}^N$ along camera ray from near bounds t_n to far bounds t_f, using Occupancy Grid approach [28]. We use these samples to estimate $\mathrm{OP}(\mathbf{r})$:

$$\widehat{\mathrm{OP}}(\mathbf{r}) = \sum_{i=1}^{N} \exp\left(-\sum_{j=1}^{i-1} \sigma_j \delta_j\right) (1 - \exp(-\sigma_i \delta_i)), \quad (4)$$

where $\delta_i = t_{i+1} - t_i$ is the distance between adjacent samples.

After obtaining opacity of each pixel, we get an image displaying outline shape from a certain viewpoint. Since opacity value calculated by Eq. (3) are in $[0, 1]$, we can build mask as follow:

$$\mathcal{M}_{ij} = \begin{cases} 0, & \mathcal{OP}_{ij} < 0.5, \\ 1, & \mathcal{OP}_{ij} \geq 0.5, \end{cases} \quad (5)$$

where $\mathcal{OP}_{ij}$ denotes the opacity values at position (i, j) of the image.

3.3. Detail Appearance Inpainting

Given a rendered view with an outline shape mask, we aim to replace the regions specified by the mask with new content. Through this mask-enhanced inpainting, we can

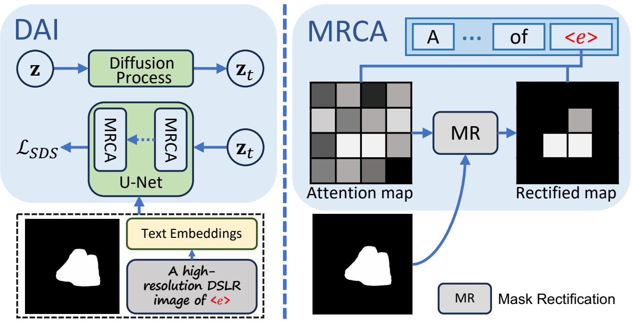

Figure 3. **Illustration of DAI.** To optimize the 3D model, DAI diffuses the rendering and backpropagates $\mathcal{L}_{\text{SDS}}$ calculated by our modified stable diffusion model. By employing Mask Rectified Cross-Attention (MRCA) in stable diffusion, we utilize outline shape mask to rectify the attention maps calculated between the special token $\langle \mathbf{e} \rangle$ and image features.

rectify undesired appearance while preserving the faithful shape. In contrast to most existing inpainting approaches that are trained for specific mask distributions, our approach is well-suited for handling various outline shape masks from different novel viewpoints. Specifically, we modify the pretrained text-to-image diffusion model (Stable Diffusion [38]) with following strategies for Detail Appearance Inpainting.

Textual Inversion. Following RealFusion [24], we use textual inversion to get a text prompt from reference image as the conditional input of our diffusion model. Specifically, we use the templates like "A high-resolution DSLR image of $\langle \mathbf{e} \rangle$" to derive the text prompt. To get the special token $\langle \mathbf{e} \rangle$ introduced to the vocabulary of the text encoder of our diffusion model, we optimize the diffusion loss with respect to the special token $\langle \mathbf{e} \rangle$ on a mini-dataset consisting random augmentations of the reference image, while freezing all other text embeddings and model parameters. As shown in RealFusion [24], text prompt derived from textual inversion helps 3D generation resemble the reference image from all views. Further, special token $\langle \mathbf{e} \rangle$ provides us a more explicit and convenient way to rectify the attention map between object tokens and image features.

Mask Rectified Cross-Attention. Given the text prompt like "A high-resolution DSLR image of $\langle \mathbf{e} \rangle$", the next step is to inject the semantics into the outline shape mask. Before diving into our solution, we revisit the concept of cross-attention [51] within diffusion models. The cross-attention layer takes two inputs: image features $x \in \mathbb{R}^{HW \times d_i}$ and text embeddings $y \in \mathbb{R}^{L \times d_\tau}$. It calculates attention maps between them as follow:

$$\mathcal{AM} = \frac{Q\left(x\right) K(y)^T}{\sqrt{d}} \in \mathbb{R}^{HW \times L}, \tag{6}$$

where Q, K are linear transformations used to obtain image queries $Q\left(x\right) \in \mathbb{R}^{HW \times d}$ and text keys $K(y) \in \mathbb{R}^{L \times d}$.

After getting the attention maps, the cross-attention layer calculates the output image feature $\mathcal{O}$ by:

$$\mathcal{O} = \mathrm{softmax}(\mathcal{AM})V(y), \tag{7}$$

where V are linear transformations used to obtain text values $V(y) \in \mathbb{R}^{L \times d}$.

Each channel of the attention maps is related to a token embedding of the input text prompt. Here, our focus lies on the channel corresponding to the special token $\langle \mathbf{e} \rangle$ derived by textual inversion. This special token $\langle \mathbf{e} \rangle$ contains almost all information about the object in the 3D content. It can be regarded as the object token in the input text prompt, whose semantics we want to inject to the outline shape mask. By modifying this channel, we can change the shape and position of this object in the generated image.

Inspired by this, we employ Mask Rectified Cross-Attention (MRCA) to integrate the outline shape mask into the image generation process. The key insight is to rectify the attention map of the special token $\langle \mathbf{e} \rangle$ using the outline shape mask. In this way, spatial distribution of the attention map is constrained by the outline shape mask, forcing special token $\langle \mathbf{e} \rangle$ to affect only pixels inside the outline shape mask. Specifically, we use the outline shape mask $\mathcal{M}$ as the binary map and resize its spatial size to match the attention maps $\mathcal{AM}^{\langle \mathbf{e} \rangle}$, which is then rectified by:

$$\widehat{\mathcal{AM}}_{ij}^{\langle \mathbf{e} \rangle} = \begin{cases} \mathcal{AM}_{ij}^{\langle \mathbf{e} \rangle}, & \mathcal{M}_{ij} = 1, \\ -inf, & \mathcal{M}_{ij} = 0, \end{cases} \tag{8}$$

where $\mathcal{M}_{ij}$ and $\mathcal{AM}_{ij}^{\langle \mathbf{e} \rangle}$ denote the values at position (i, j) in the outline shape mask $\mathcal{M}$ and attention maps $\mathcal{AM}^{\langle \mathbf{e} \rangle}$, respectively.

Novel View Guidance. The introduced MRCA module, which enhances the injection of semantics from diffusion priors into the outline mask, ensures overall semantic consistency between the generated content and the mask. Nonetheless, minor discrepancies between the generated content and the outline mask may still persist, particularly at certain edges. During the optimization process of 3D content using the SDS loss, these minor discrepancies tend to magnify. To better preserve the outline shape during the optimization, we impose a loss term $\mathcal{L}_{nvm}$ to ensure the outlines of rendered images from the novel views is as close to the corresponding outline shape masks as possible. Specifically, given a novel viewpoint, $\mathcal{L}_{nvm}$ is imposed between the rendered opacity and outline shape mask from the corresponding novel view as

$$\mathcal{L}_{nvm} = \| \mathcal{M} - \mathcal{M}_\theta(\mathbf{v}) \|_2^2, \tag{9}$$

where θ is the 3D model parameters to be optimized, $\mathcal{M}_\theta(\mathbf{v})$ is rendered opacity from novel viewpoint $\mathbf{v}$. By imposing $\mathcal{L}_{nvm}$, we expect the detail refinement of the 3D

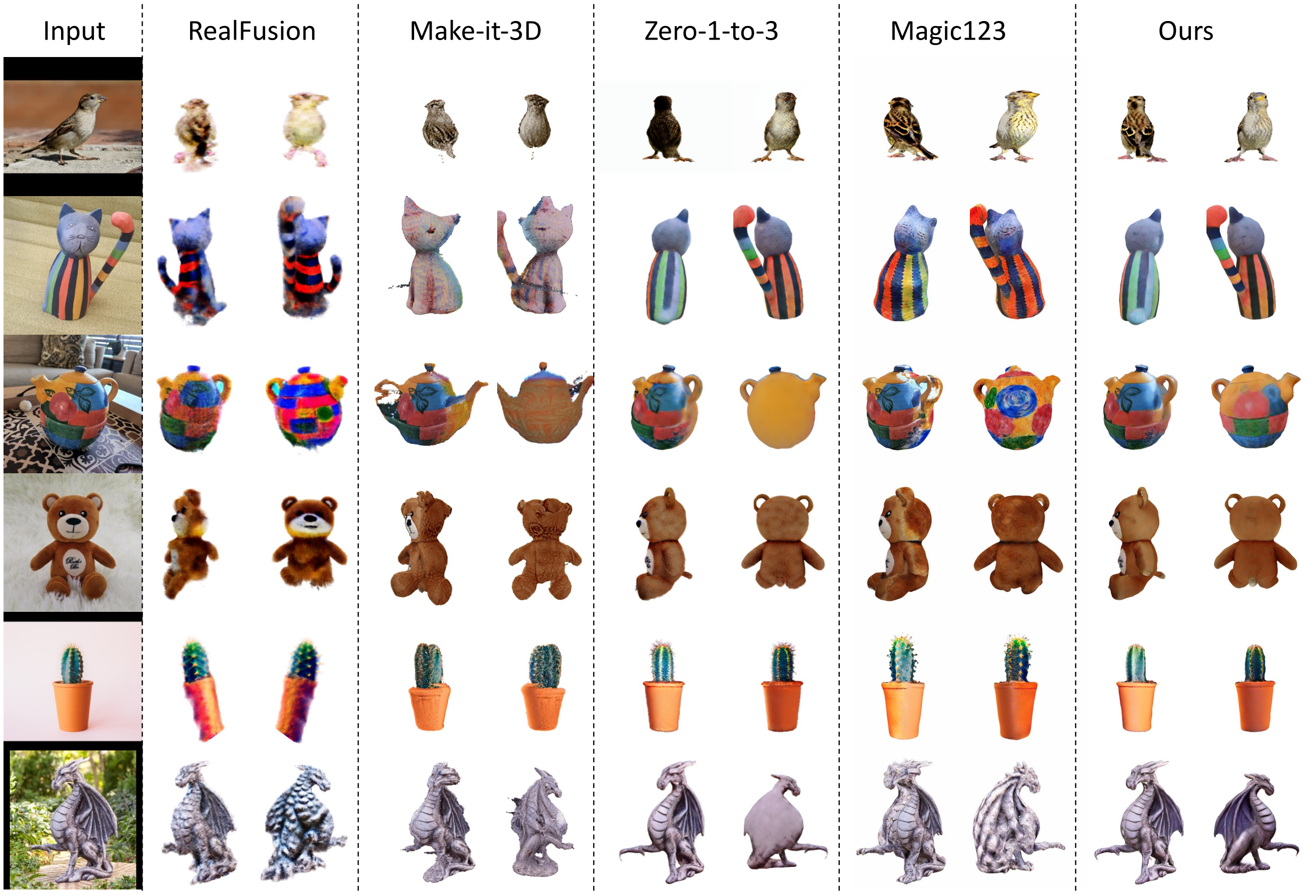

Figure 4. **Qualitative comparison on RealFusion15 [24].** We compare our MPOD123 to diffusion-based approaches for generating 3D contents from a single unposed image (the leftmost column). We show our results with plausible geometry and visually pleasing textures.

content are only inside the outline shape, leaving the area outside the outline shape untouched. The overall loss to guide the novel view optimization can be formulated as a combination of $\mathcal{L}_{nvm}$ and $\mathcal{L}_{SDS}$.

4. Experiments

4.1. Datasets

RealFusion15. We conduct experiments on the dataset released by RealFusion [24]. RealFusion15 consists of 15 testing examples covering a variety of subjects, including birds, cat statues, teapots, and dragon statues etc.

TBPR-Shoes. We further collect a real e-commerce setting Product Reconstruction (Shoes) dataset from Taobao, named TBPR-Shoes. TBPR-Shoes consists of 17 testing examples from 3 categories: men's shoes, women's shoes, and children's shoes. We collect the dataset by filtering out the images with a front view object from Taobao product pages, and obtaining the foreground object mask for each image using Segment Anything Model (SAM) [17]. Besides, each image is provided with a text prompt obtained from textual inversion.

4.2. Implementation Details

Camera setting. Following the camera sampling method used in [34], we randomly sample novel views with a 75% probability and sample the pre-defined reference view with a 25% probability. The reference view remains fixed at the center of the camera's range, with a camera distance of 2.5 and a field-of-view (FOV) of 40 degrees. And in our experiments, we always assume the reference image is shot from the front view, with polar angle 90 degrees and azimuth angle 0 degrees.

Rendering. We initialize the spatial density following [19], which benefits to stable training. In the first stage, we adopt the multi-scale hash encoding from Instant-NGP [28] to implement the NeRF representation, which enables neural rendering at a computational cost. Similar to Instant-NGP, we maintain an occupancy grid to enable efficient ray sampling by skipping empty space. In the second stage, we convert the coarse NeRF representation to an SDF representation and adopt DMTet [41], which is a hybrid SDF-Mesh rep-

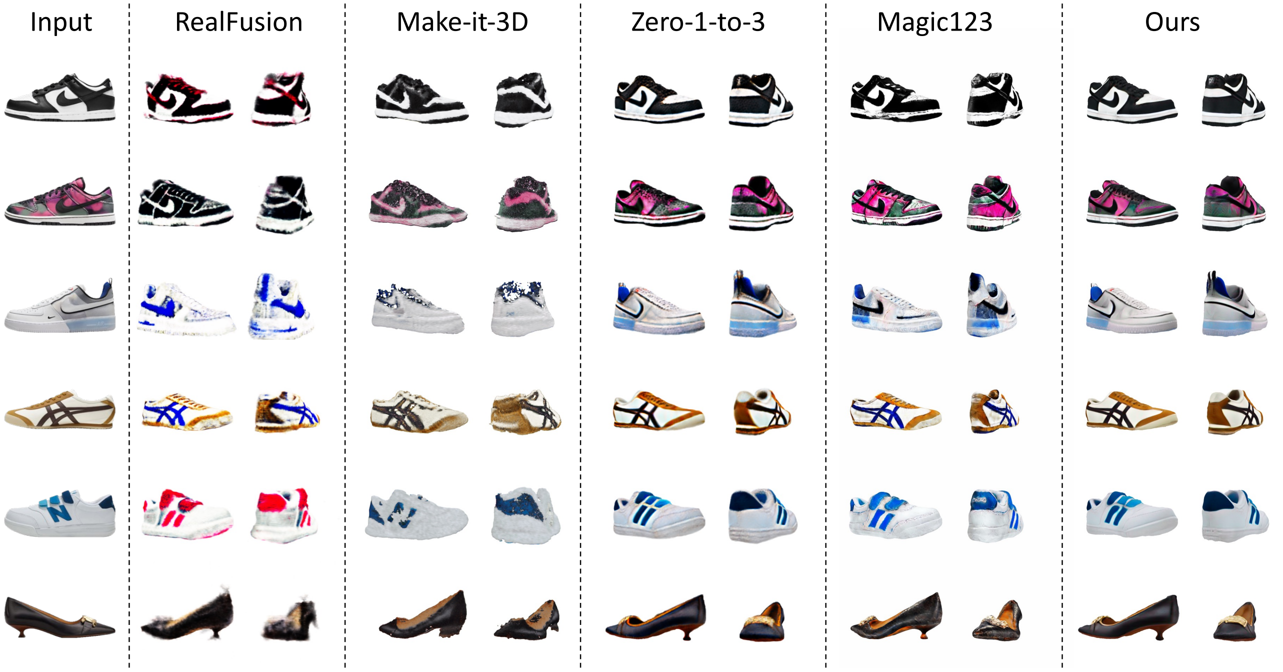

Figure 5. **Qualitative comparison on TBPR-Shoes.** We shows results based on an image collected from Taobao (the leftmost column). Our MPOD123 outperforms the state of the arts for generating high-fidelity 3D contents on real e-commerce setting.

resentation, to implement the textured meshes representation. With high memory efficiency of DMTet, it is capable of generating high-resolution 3D contents.

Two-stage training We use Adam [16] with a learning rate of 0.001 for both stages. The first stage is trained for 5,000 iterations at a rendering resolution of 128×128. The second stage then takes another 5,000 iterations at a rendering resolution of 1024×1024.

4.3. Comparisons with the State of the Arts

Baselines. We compare our method with four representative baselines on single image to 3D generation. 1) RealFusion [24], an early attempt to incorporate pretrained text-to-image diffusion model prior. 2) Make-it-3D [46], a two-stage scheme with NeRF and Textured Point Clouds. 3) Zero-1-to-3 [21], a viewpoint-conditioned diffusion model. For a fair comparison, we modify it using SDS loss optimization with the same experimental settings as ours. 4) Magic123 [35], a method with combination of 2D and 3D diffusion priors.

Qualitative comparison. We compare our method against four representative baselines in both RealFusion15 and TBPR-Shoes datasets. As shown in Figure 4 and Figure 5, the results generated by RealFusion exhibit poor similarity in both shape and appearance compared to the input image. Make-it-3D displays competitive quality in terms of texture, but it falls short in preserving shape consistency under large viewpoint changes. Zero-1-to-3 manages to maintain consistent shape even under large viewpoint changes, but it suffers from a decline in appearance quality. For instance, the teapot and the dragon statue in the Figure 4, exhibit poor textures in the backside viewpoint. Besides, the sparrow in the first row of Figure 4, exhibit blurry textures, making it impossible to clearly discern the sparrow's eyes and patterns. The results generated by Magic123 demonstrate instability, with shape distortions and poor texture consistency among novel views. For instance, the sparrow in the first row of Figure 4 exhibits multi-face Janus problem, and the shoe in the third row of Figure 5 exhibits shape distortions. In contrast, our approach excels in generating high-fidelity results, complying to input reference images. Further, these results exhibit remarkably faithful shape and visually appealing appearance, while ensuring 3D consistency across multiple viewpoints.

Quantitative comparison. A compelling generated 3D content should exhibit high similarity with the reference image under novel views, both at the pixel-level and semantic-level. Additionally, the generated 3D content should ensure consistency among different novel views, such as maintaining consistent texture features across all viewpoints. We evaluate these two aspects using the following metrics: 1) contextual (CX) distance [23], which measures pixel-level similarity between novel-view rendering and the reference, and 2) CLIP score [36], which evaluates the semantic similarity between the novel view and the reference. 3) multi views consistency (MVC) score, we utilize CLIP score to

Table 1. **Quantitative comparison.** We show quantitative results in terms of CX↓ / CLIP↑ / MVC↑. The results are shown on the Realfusion15 and TBPR-Shoes datasets, while **bold** reflects the best.

Dataset	**Metrics\Methods**	RealFusion [24]	Make-it-3D [46]	Zero-1-to-3 [21]	Magic123 [35]	**MPOD123 (Ours)**
RealFusion15	CX↓	2.04	1.79	1.74	1.65	**1.54**
	CLIP↑	0.72	0.77	0.74	0.83	**0.86**
	MVC↑	0.87	0.84	0.89	0.87	**0.90**
TBPR-Shoes	CX↓	1.63	1.48	1.49	1.33	**1.29**
	CLIP↑	0.51	0.58	0.56	0.71	**0.78**
	MVC↑	0.79	0.68	0.84	0.74	**0.86**

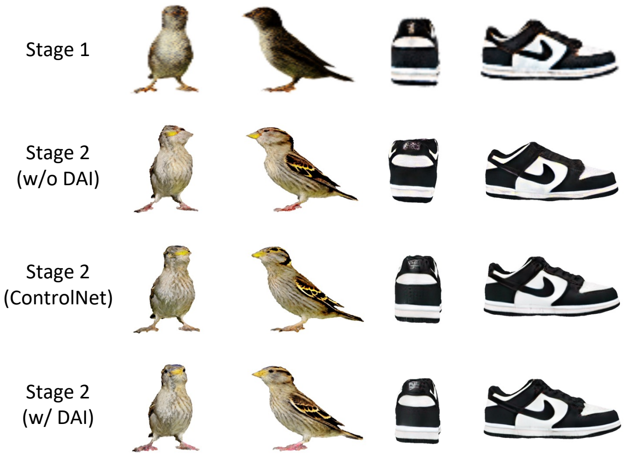

Figure 6. **Ablation Studies.** We compare the results of each stage in our progressive optimization framework. Stage 1 demonstrates faithful shapes, whereas Stage 2 exhibits visual pleasing appearances. We ablate the effect of DAI in the second stage optimization. DAI effectively addresses shape distortions and mitigates the multi-face Janus problem.

evaluates the semantic consistency among different novel views. As shown in Table 1, our approach substantially outperforms baselines in terms of both metrics.

4.4. Ablation Studies

Ablation on Two Stages. We compare the results of each stage in our two-stage progressive optimization framework as shown in Figure 6. It can be observed that the generated novel views in the first stage exhibit faithful outline shapes but suffer from poor local geometry and texture details. After further optimization in the second stage, the generated novel views maintain the faithful outline shapes from the first stage and demonstrate improved quality in terms of local geometry and texture. For example, the eyes and feathers of the bird exhibit high-quality geometry and texture. And the sole of the shoe exhibit bump details, resembling the reference input image.

Ablation on Detail Appearance Inpainting. We ablate the effect of Detail Appearance Inpainting in the second stage optimization, as shown in Figure 6. We compare the results of our Detail Appearance Inpainting approach with the original stable diffusion model (w/o DAI) in the second stage. It can be observed that, after further optimization in the second stage, the generated novel views exhibit improved texture quality in both approaches. However, the results generated by w/o DAI suffer from shape distortions and multi-face Janus problem. In contrast, our DAI approach generating high-quality local geometry and textures, while maintains the faithful outline shape from the first stage. In other words, our DAI approach can mitigate shape distortions and multi-face Janus problem. For example, the head of the bird, generated by w/o DAI, suffers from multi-face Janus problem. And the back view of the shoe exhibits shape distortions. Whereas our DAI approach avoids these issues.

Mask Injection Techniques. Apart from our MRCA module, alternative mask injection techniques through the modulation of attention layers can be seamlessly incorporated into our two-stage progressive optimization framework. Within this framework, we have investigated the effectiveness of ControlNet [63], utilizing the opacity masks as a conditional mechanism. As show in Figure 6, our MRCA module is distinguished by its lightweight design and better performance in refining local texture details.

5. Conclusions

We introduce MPOD123, a novel two-stage progressive optimization framework for generating high-quality 3D content from one image. Accompanied with our framework, knowledge from pretrained 2D diffusion models can be leveraged progressively, facilitating the high-quality 3D content generation with plausible geometry and visually pleasing textures. Future works include specializing the diffusion model for providing better outline shape priors and expanding to the multi-object generation task, where scenes contain multiply 3D objects.

Acknowledgments. This work was supported by National Key R&D Program of China (2022ZD0162000), National Natural Science Foundation of China (62222211, U20A20387) and National Science and Technology Major Project (2022ZD0119100). In addition, this work was supported by Alibaba Group through Alibaba Research Intern Program and Alibaba Innovation Research (AIR) Program.

References

[1] Titas Anciukevičius, Zexiang Xu, Matthew Fisher, Paul Henderson, Hakan Bilen, Niloy J Mitra, and Paul Guerrero. Renderdiffusion: Image diffusion for 3d reconstruction, inpainting and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12608–12618, 2023. 3

[2] Eric R Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5799–5809, 2021. 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16123–16133, 2022. 3

[4] 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. 2, 3

[5] Hansheng Chen, Jiatao Gu, Anpei Chen, Wei Tian, Zhuowen Tu, Lingjie Liu, and Hao Su. Single-stage diffusion nerf: A unified approach to 3d generation and reconstruction. *arXiv preprint arXiv:2304.06714*, 2023. 3

[6] Yen-Chi Cheng, Hsin-Ying Lee, Sergey Tulyakov, Alexander G Schwing, and Liang-Yan Gui. Sdfusion: Multimodal 3d shape completion, reconstruction, and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4456–4465, 2023. 3

[7] Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, pages 628–644. Springer, 2016. 3

[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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13142–13153, 2023. 2, 3

[9] Shivam Duggal and Deepak Pathak. Topologically-aware deformation fields for single-view 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1536–1546, 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. *Advances In Neural Information Processing Systems*, 35:31841–31854, 2022. 3

[11] Yuxuan Han, Ruicheng Wang, and Jiaolong Yang. Single-view view synthesis in the wild with learned adaptive multiplane images. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–8, 2022. 3

[12] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 2

[13] Philipp Henzler, Niloy J Mitra, and Tobias Ritschel. Learning a neural 3d texture space from 2d exemplars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8356–8364, 2020. 2

[14] Ajay Jain, Matthew Tancik, and Pieter Abbeel. Putting nerf on a diet: Semantically consistent few-shot view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5885–5894, 2021. 2, 3

[15] Heewoo Jun and Alex Nichol. Shap-e: Generating conditional 3d implicit functions. *arXiv preprint arXiv:2305.02463*, 2023. 3

[16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 7

[17] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4015–4026, 2023. 6

[18] Jonáš Kulhánek, Erik Derner, Torsten Sattler, and Robert Babuška. Viewformer: Nerf-free neural rendering from few images using transformers. In *European Conference on Computer Vision*, pages 198–216. Springer, 2022. 3

[19] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 300–309, 2023. 6

[20] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 300–309, 2023. 3

[21] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9298–9309, 2023. 3, 7, 8

[22] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9298–9309, 2023. 2, 3

[23] Roey Mechrez, Itamar Talmi, and Lihi Zelnik-Manor. The contextual loss for image transformation with non-aligned data. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 7

[24] Luke Melas-Kyriazi, Iro Laina, Christian Rupprecht, and Andrea Vedaldi. Realfusion: 360deg reconstruction of any object from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8446–8455, 2023. 5, 6, 7, 8

[25] Luke Melas-Kyriazi, Iro Laina, Christian Rupprecht, and Andrea Vedaldi. Realfusion: 360deg reconstruction of any object from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8446–8455, 2023. 2, 3

[26] 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*, 2020. 4

[27] 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. *Communications of the ACM*, 65(1):99–106, 2021. 2

[28] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4):102:1–102:15, 2022. 4, 6

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

[30] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11453–11464, 2021. 3

[31] Michael Niemeyer, Jonathan T Barron, Ben Mildenhall, Mehdi SM Sajjadi, Andreas Geiger, and Noha Radwan. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5480–5490, 2022. 2, 3

[32] Roy Or-El, Xuan Luo, Mengyi Shan, Eli Shechtman, Jeong Joon Park, and Ira Kemelmacher-Shlizerman. Stylesdf: High-resolution 3d-consistent image and geometry generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13503–13513, 2022. 3

[33] Eunbyung Park, Jimei Yang, Ersin Yumer, Duygu Ceylan, and Alexander C Berg. Transformation-grounded image generation network for novel 3d view synthesis. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 3500–3509, 2017. 2

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

[35] 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. 3, 7, 8

[36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 7

[37] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10901–10911, 2021. 3

[38] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 2, 5

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

[40] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. *Advances in Neural Information Processing Systems*, 33:20154–20166, 2020. 3

[41] Tianchang Shen, Jun Gao, Kangxue Yin, Ming-Yu Liu, and Sanja Fidler. Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. In *Advances in Neural Information Processing Systems*, 2021. 2, 6

[42] Meng-Li Shih, Shih-Yang Su, Johannes Kopf, and Jia-Bin Huang. 3d photography using context-aware layered depth inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8028–8038, 2020. 3

[43] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[44] Pratul P Srinivasan, Richard Tucker, Jonathan T Barron, Ravi Ramamoorthi, Ren Ng, and Noah Snavely. Pushing the boundaries of view extrapolation with multiplane images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 175–184, 2019. 3

[45] Shao-Hua Sun, Minyoung Huh, Yuan-Hong Liao, Ning Zhang, and Joseph J Lim. Multi-view to novel view: Synthesizing novel views with self-learned confidence. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 155–171, 2018. 2

[46] 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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22819–22829, 2023. 7, 8

[47] 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 preprint arXiv:2303.14184*, 2023. 2, 3

[48] Alex Trevithick and Bo Yang. Grf: Learning a general radiance field for 3d representation and rendering. In *Proceed-*

ings of the IEEE/CVF International Conference on Computer Vision, pages 15182–15192, 2021. 3

[49] Richard Tucker and Noah Snavely. Single-view view synthesis with multiplane images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 551–560, 2020. 3

[50] Shubham Tulsiani, Richard Tucker, and Noah Snavely. Layer-structured 3d scene inference via view synthesis. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 302–317, 2018. 3

[51] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 5

[52] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12619–12629, 2023. 3

[53] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4563–4573, 2023. 3

[54] Xiao Wang, Peng Cui, Jing Wang, Jian Pei, Wenwu Zhu, and Shiqiang Yang. Community preserving network embedding. In *Proceedings of the AAAI conference on artificial intelligence*, 2017. 3

[55] Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models. *arXiv preprint arXiv:2210.04628*, 2022. 2

[56] Chao-Yuan Wu, Justin Johnson, Jitendra Malik, Christoph Feichtenhofer, and Georgia Gkioxari. Multiview compressive coding for 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9065–9075, 2023. 3

[57] Haozhe Xie, Hongxun Yao, Xiaoshuai Sun, Shangchen Zhou, and Shengping Zhang. Pix2vox: Context-aware 3d reconstruction from single and multi-view images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2690–2698, 2019. 3

[58] 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. In *European Conference on Computer Vision*, pages 736–753. Springer, 2022. 3

[59] 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 360deg views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4479–4489, 2023. 3

[60] Farid Yagubbayli, Yida Wang, Alessio Tonioni, and Federico Tombari. Legoformer: Transformers for block-by-block multi-view 3d reconstruction. *arXiv preprint arXiv:2106.12102*, 2021. 3

[61] Yufei Ye, Shubham Tulsiani, and Abhinav Gupta. Shelf-supervised mesh prediction in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8843–8852, 2021. 3

[62] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4578–4587, 2021. 3

[63] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 8

[64] Ziwei Zhang, Peng Cui, and Wenwu Zhu. Deep learning on graphs: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):249–270, 2022. 3