PRM: Photometric Stereo based Large Reconstruction Model

Wenhang Ge[1]* Jiantao Lin[1]* Guibao Shen[1]* Jiawei Feng[1] Tao Hu[3]
Xinli Xu[1] Ying-Cong Chen[1, 2] †
[1] HKUST-GZ [2] HKUST [3] Pico, Bytedance

Abstract

We propose PRM, a novel photometric stereo based large reconstruction model to reconstruct high-quality meshes with fine-grained details. Previous large reconstruction models typically prepare training images under fixed and simple lighting, offering minimal photometric cues for precise reconstruction. Furthermore, images containing specular surfaces are treated as out-of-distribution samples, resulting in degraded reconstruction quality. To handle these challenges, PRM renders images by varying materials and lighting, which not only improves the local details by providing rich photometric cues but also increases the model's robustness to variations in the appearance of input images. To offer enhanced flexibility, we incorporate a real-time physically-based rendering (PBR) method and mesh rasterization for ground-truth rendering. By using an explicit mesh as 3D representation, PRM ensures the application of differentiable PBR for predicted rendering. This approach models specular color more accurately for images with varying materials and illumination than previous neural rendering methods and supports multiple supervisions for geometry optimization. Extensive experiments demonstrate that PRM significantly outperforms other models.

1. Introduction

Recent advancements in generative models [12, 27, 38] have spurred notable progress in 2D content creation, driven by fast growth in data volumes. In contrast, the development in 3D field remains encumbered due to limited 3D assets, which are essential for diverse applications including game modeling [8], computer animation [23, 32], and virtual reality [35]. Traditional approaches to generating 3D assets have utilized optimization-based techniques from multi-view posed images [42, 49, 50] or have harnessed SDS-based distillation methods from 2D diffusion models [25, 26, 34]. Despite their effectiveness, these methods often require increased computational costs, making them less suitable for rapid deployment in real world.

Feed-forward 3D generative models [13, 15, 51] have been developed to address the limitations of per-scene optimization by training a generalizable model on large-scale 3D assets. Notably, the Large Reconstruction Model (LRM) [13] has demonstrated promising results, exhibiting exceptional reconstruction speeds. The subsequent LRM series [13, 39, 43, 46, 48] utilizes a Transformer-based architecture to encode either single or multi-view images, and decoding them into 3D representations, such as triplanes [3], Flexicubes [36] or 3D Gaussians [39]. These 3D representations enable differentiable neural rendering from arbitrary viewpoints, which is crucial for calculating multi-view reconstruction loss for optimization.

While the LRM series effectively reconstructs globally coherent 3D assets, it struggles with capturing fine-grained details. This limitation stems from its reliance on images rendered under fixed, simple lighting, which provide insufficient photometric information for detailed surface reconstruction. Additionally, the LRM series is sensitive to variations in input image appearance, particularly with glossy surfaces, as these are out-of-distribution samples that rarely occur in the training data. Moreover, the use of neural rendering for predicted rendering exacerbates this issue [41], since the view-dependent effect is not considered.

To address the challenges, we introduce PRM, a photometric stereo based large reconstruction model. This model is adept at capturing fine-grained local details and ensures robustness against the complex appearances of input images. We achieve these objectives by leveraging images with varying materials and illumination [10]. Specifically, we render ground-truth images by varying camera pose, materials (i.e., metallic and roughness), and lighting. However, rendering these images is not trivial since there are infinite possible combinations of camera pose, materials and lighting. In the recent LRM series, images are typically rendered offline using Blender's Cycles engine [11]. While this approach produces high-quality, noise-free images, it requires numerous samples of lighting directions, significantly increasing the time cost and making it expensive to maximize the training sample distribution.

*Equal contribution
†Corresponding author

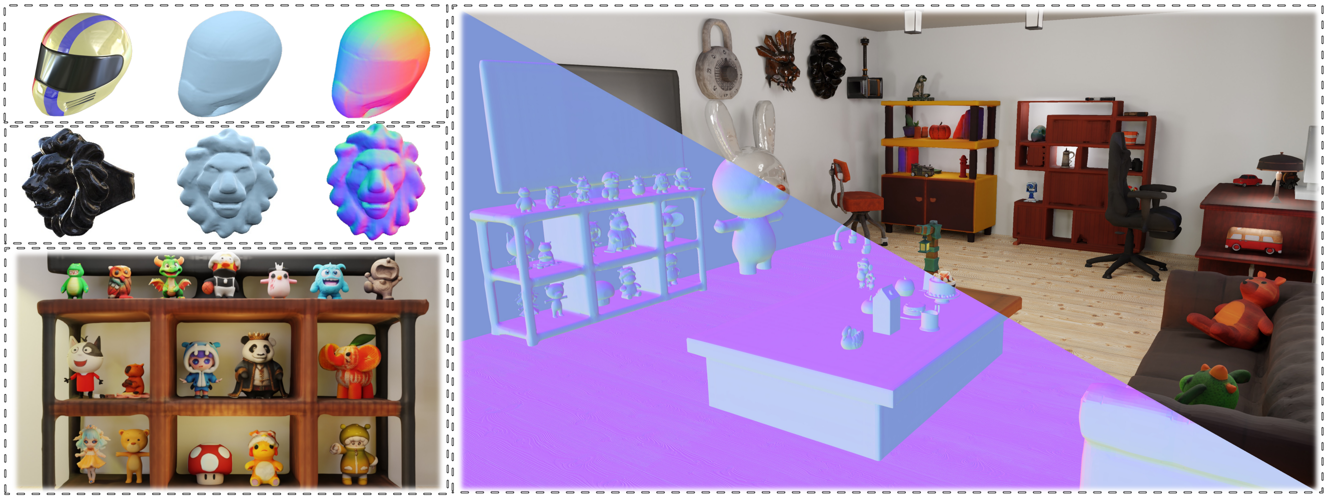

Figure 1. Top left: PRM is capable of reconstructing high-quality meshes with fine-grained local details even under complex image appearances, such as specular highlights and dark appearances. Right: We demonstrate a scene comprising diverse 3D objects generated by our models. Bottom left: A zoomed-in visualization of the scene highlights these details more clearly.

To handle this issue, we incorporate a real-time, physically based rendering technique known as split-sum approximation [19], along with mesh rasterization for online rendering. This approach offers greater flexibility compared to traditional offline methods. We discuss two training strategies in the Supplement. Images with varying materials and illumination offer two distinct advantages. First, they provide additional photometric cues, thereby enhancing the capacity to recover fine-grained local details. Second, the PRM model demonstrates remarkable robustness to variations in the appearances of inputs. For instance, it is capable of accurately reconstructing the geometry of images with glossy surfaces. Moreover, by utilizing mesh as 3D representation, we can employ differentiable PBR for predicted rendering, producing intermediate shading variables such as albedo, specular and diffuse light maps, along with geometric cues like normals and depth. These variables provide multiple supervisions, including photometric and geometric supervision for high-quality reconstruction. Furthermore, differentiable PBR can better disentangle the specular component, making the geometry be correctly recovered when the images are characterized with glossy surfaces.

To summarize, our contributions are listed as follows.

- We introduce PRM, a model that is capable of reconstructing geometry with fine-grained local details and robust to variations in the appearance of input image by utilizing images with varying materials and illumination.
- We employ differentiable PBR for predicted rendering by utilizing mesh as the 3D representation. This approach is advantageous for modeling reflective components and enables the incorporation of multiple supervisions for high-quality geometry reconstruction.
- Extensive experiments on multiple datasets demonstrate the effectiveness of the proposed framework, which significantly outperforms previous methods, particularly on glossy objects.

2. Related Work

2.1. Feed-forward 3D Generative Models

Large-scale 3D assets [5] facilitate the training of generalizable reconstruction models. Recent works have focused on generating 3D objects using feed-forward models [13, 14, 24, 39, 43, 47, 48, 51], demonstrating impressive results in terms of speed and quality. Specifically, Clay [51] utilizes occupancy for direct supervision. X-ray [15] explores novel 3D representations by converting a 3D object into a series of surface frames at different layers. The LRM series [13, 24, 43, 46, 48] shows that a transformer backbone can effectively map image tokens to 3D triplanes, benefiting from multi-view supervision. Instant3D [24] employs multi-view images to provide additional 3D information for triplane prediction, yielding promising outcomes. CRM [43] and InstantMesh [46] opt for an explicit mesh representation, supporting mesh rasterization and rendering additional geometric cues for supervision. Despite these achievements, the existing LRM series typically render low-frequency images under fixed and simple lighting, which compromises the model's adaptability to complexity in the appearance of input images and the model's capability to recover local details due to limited photometric cues. In response, we render images with varying materials and illumination that significantly enhance the photometric cues necessary for the recovery of fine-grained local details.

2.2. Photometric Stereo

Photometric stereo (PS) is a technique for recovering surface normals from the appearance of an object under varying lighting conditions. Traditional methods, inspired by the seminal work [44], assume calibrated, directional lighting. Recently, uncalibrated photometric stereo methods have emerged, which assume Lambertian integrable surfaces and aim to resolve the General Bas-Relief ambiguity [9] between light and geometry. However, these methods are still constrained to single directional lighting. More contemporary research [16, 17, 29] has shifted focus towards natural lighting conditions. Despite the significant progress, these approaches generally concentrate on single-view photometric stereo, relying solely on photometric cues and neglecting multi-view information, which is crucial for accurately reasoning geometric features. Some studies [10, 20–22, 33, 53] leverage both photometric and geometric cues for reconstruction. These cues are complementary: photometric stereo provides precise local details, while multi-view information yields accurate global shapes [53]. For example, UA-MVPS [20] utilizes complementary strengths of PS and multi-view stereo for geometry reconstruction. NeRF-MVPS [21] utilizes surface normal estimated from images with varying materials and illumination to enhance the reconstruction performance of NeRF. Our approach integrates the principles of photometric stereo into LRM, aiming to harness the strengths of photometric cues for enhanced reconstruction accuracy.

2.3. Physically-based Rendering

Physically based rendering (PBR) is a computer graphics approach that renders photo-realistic images. PBR offers a physically plausible approach to modeling radiance by simulating the interaction between lighting and materials. PBR has proven to be effective in improving the geometry for multi-view reconstruction task. For example, incorporating the principles of PBR into volume rendering significantly improves the accuracy [7, 28, 41], especially for glossy surfaces. Since geometry and predicted radiance are closely entangled, improving radiance modeling can also enhance geometry reconstruction. Besides, PBR is also widely used in inverse rendering task [1, 4, 31]. The task aims at decomposing image appearance into intrinsic properties. Unlike previous LRM methods that predict radiance without explicitly considering the interactions between materials and lighting, we leverage advancements from the multi-view reconstruction field and employ PBR for improved radiance modeling and geometry reconstruction. To this end, we predict albedo instead of color, which is more reasonable as albedo is view-independent. The final color is derived using the predicted albedo and the ground truth metallic, roughness, and lighting.

3. Method

We begin with a succinct overview of large reconstruction model, physically based rendering and photometric stereo in Section 3.1. Then, we introduce how to prepare images with varying materials and illumination in Section 3.2. Subsequently, we introduce PRM in Section 3.3, with our proposed comprehensive objectives and applications. An overview of our framework is provided in Figure 2.

3.1. Preliminaries

Large Reconstruction Model aims to reconstrcut 3D assets given a single image. LRM first utilizes a pre-trained visual transformer [2], to encode the images into image tokens. Subsequently, it employs an image-to-triplane transformer decoder that projects these 2D image tokens onto a 3D triplane using cross-attention [13]. Following this, images can be differentiable rendered from any viewpoint, supporting photometric supervision and optimization.

Physically-based Rendering aims to produce 2D images using specified geometry, materials, and lighting. Central to this process is the rendering equation [18] formulated by

$$\boldsymbol{C}(\boldsymbol{x}, \boldsymbol{\omega_o}) = \int_{\Omega} f(\boldsymbol{x}, \boldsymbol{\omega_o}, \boldsymbol{\omega_i}) L_i(\boldsymbol{x}, \boldsymbol{\omega_i})(\boldsymbol{\omega_i} \cdot \mathbf{n}) d\boldsymbol{\omega_i}, \quad (1)$$

where $\boldsymbol{\omega_o}$ is the viewing direction of the outgoing light, L_i is the incident light of direction $\boldsymbol{\omega_i}$ sampled from the upper hemisphere Ω of the surface point $\boldsymbol{x}$, and $\mathbf{n}$ is the surface normal. f is the BRDF properties. The function f consists of a diffused and a specular component

$$f(\boldsymbol{x}, \boldsymbol{\omega_o}, \boldsymbol{\omega_i}) = (1-m)\frac{\boldsymbol{a}}{\pi} + \frac{DFG}{4(\boldsymbol{\omega_i} \times \mathbf{n})(\boldsymbol{\omega_o} \times \mathbf{n})}, \quad (2)$$

where $m \in [0,1]$ is the metallic, $\boldsymbol{a} \in [0,1]^3$ is the albedo. We detail the expression of D, F and G in the Supplement. With Eq.(1) and Eq.(2), the outgoing radiance is given by

$$\boldsymbol{C}(\boldsymbol{x}, \boldsymbol{\omega_o}) = \boldsymbol{C}_{\mathrm{d}}(\boldsymbol{x}, \boldsymbol{\omega}_o) + \boldsymbol{C}_{\mathrm{s}}(\boldsymbol{x}, \boldsymbol{\omega}_o), \quad (3)$$

$$\boldsymbol{C}_{\mathrm{d}}(\boldsymbol{x}, \boldsymbol{\omega_o}) = (1-m)\boldsymbol{a} \int_{\Omega} L_i(\boldsymbol{x}, \boldsymbol{\omega_i}) \frac{(\boldsymbol{\omega_i} \cdot \mathbf{n})}{\pi} d\boldsymbol{\omega_i}, \quad (4)$$

$$\boldsymbol{C}_{\mathrm{s}}(\boldsymbol{x}, \boldsymbol{\omega_o}) = \int_{\Omega} \frac{DFG}{4(\boldsymbol{\omega_i} \times \mathbf{n})(\boldsymbol{\omega_o} \times \mathbf{n})} L_i(\boldsymbol{x}, \boldsymbol{\omega_i})(\boldsymbol{\omega_i} \cdot \mathbf{n}) d\boldsymbol{\omega_i}, \quad (5)$$

$\boldsymbol{C}_{\mathrm{s}}$ and $\boldsymbol{C}_{\mathrm{d}}$ are specular and diffuse color, respectively.

Photometric Stereo aims to estimate the surface normals by observing an object under varying lightings [44]. When considering a Lambertian surface illuminated by a single

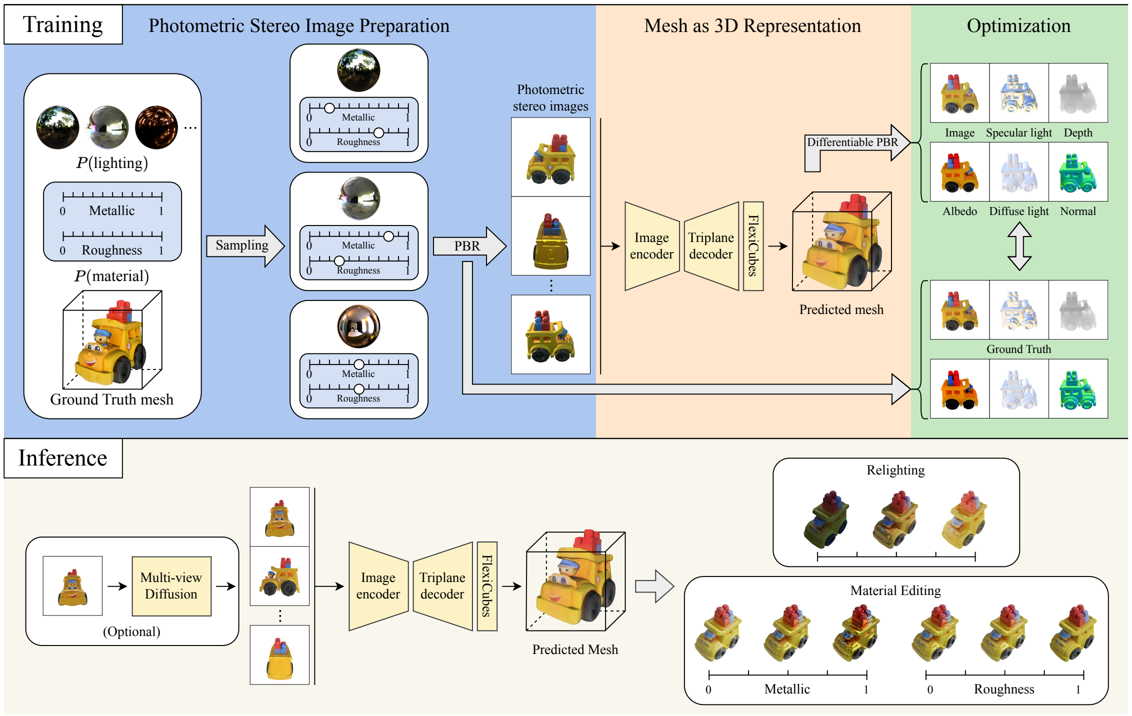

Figure 2. Overview of our framework. During training, images with varying materials and illumination are rendered using PBR with randomly varied materials, lighting, and camera poses, along with depth, normal, albedo, and lighting maps. Images are encoded as a mesh through the network. All associated maps are used for supervision. During inference, an optional multi-view diffusion model outputs multi-view images, which are then fed into the network for mesh prediction. Relighting and material editing functionalities are also supported.

point-like, distant light, the color is determined solely by the diffuse term, which can be formulated by

$$\boldsymbol{C}(\boldsymbol{x}) = \boldsymbol{a}(\boldsymbol{L} \cdot \boldsymbol{n}), \tag{6}$$

where $\boldsymbol{L} = L \cdot \boldsymbol{\omega}$, $\boldsymbol{\omega}$ and L are the direction and the intensity of the lighting, respectively. When we observe the object under different lighting conditions $\boldsymbol{L}_i$, we have multiple such equations. We assume that both the direction and intensity of the lighting are known, and the shading colors $\boldsymbol{C}(\boldsymbol{x})$ is also observed. Under these conditions, determining the surface normals $\boldsymbol{n}$ and the albedo $\boldsymbol{a}$ is effectively equivalent to solving a system of equations derived from these observations. More lighting indicates more equations, which effectively constrain the solution space.

3.2. images with varying materials and illumination

Previous methods typically prepared training data by rendering multi-view images with fixed, simple lighting and materials in Blender [11]. This approach resulted in images with low-frequency appearances, offering limited photometric cues. Consequently, these methods struggled to reconstruct geometry with precise local details. Additionally, they often failed when processing images with glossy surfaces, as these are out-of-distribution samples for the model. We provide some examples in the Supplement.

In contrast, we prepare images with varying materials and illumination by varying materials and lighting. A naive solution is to prepare these images offline, as with previous methods, but this approach poses significant challenges due to the infinite number of potential combinations of materials and lighting. Moreover, rendering high-quality images requires large sample counts, making traditional data preparation methods infeasible. To overcome these issues, we incorporate a real-time rendering method known as split-sum approximation [19] along with mesh rasterization, which facilitates rapid rendering. This method enables online data preparation and significantly enhances flexibility.

Split-sum approximation. High-quality estimation of physically based rendering typically requires Monte Carlo sampling to approximate the integral in Eq.(1). However, this process demands large sample counts, making it time-consuming. Instead, we employ a real-time rendering method known as the split-sum approximation [19]. According to the split-sum approximation, the specular component in Eq.(5) can be rewritten as:

$$\boldsymbol{C}_{\mathrm{s}}(\boldsymbol{x}, \boldsymbol{\omega}_o) \approx \int_\Omega \frac{DFG}{4(\boldsymbol{\omega}_o \cdot \boldsymbol{n})} d\boldsymbol{\omega}_i \int_\Omega L(\boldsymbol{x}, \boldsymbol{\omega}_i) D(\hat{\boldsymbol{d}}, \rho) d\boldsymbol{\omega}_i, \tag{7}$$

The first term is the integral of the BRDF, which is approximated by specular albedo $\boldsymbol{a}_{\text{s}} = ((1-m)*0.04+m*\boldsymbol{a})*F_1+F_2$, where F_1 and F_2 are pre-computed scalars and stored in a 2D lookup texture related to $\rho, \boldsymbol{n}$ and $\boldsymbol{\omega}_o$. The second term is the integral of lights on the normal distribution function $D(\hat{\boldsymbol{d}}, \rho)$, which can also be pre-computed and stored as mipmaps $\boldsymbol{M}_{\text{s}}$. $\hat{\boldsymbol{d}} = 2(-\boldsymbol{\omega}_o \cdot \boldsymbol{n})\boldsymbol{n} + \boldsymbol{\omega}_o$ is the reflective direction. Then the Eq.(7) is modified as

$$\boldsymbol{C}_{\text{s}}(\boldsymbol{x}, \boldsymbol{\omega}_o) = \boldsymbol{a}_{\text{s}}(\boldsymbol{a}, m, \boldsymbol{n}, \boldsymbol{\omega}_o)\boldsymbol{L}_{\text{spec}}(\boldsymbol{x}, \boldsymbol{n}, \boldsymbol{\omega}_o, \rho, \boldsymbol{M}_{\text{s}}), \quad (8)$$

where $\boldsymbol{L}_{\text{spec}} = \text{Tex\_Sample}(\boldsymbol{x}, \hat{\boldsymbol{d}}, \rho, \boldsymbol{M}_{\text{s}})$, Tex_Sample indicates texture sampling based on different levels of roughness ρ in pre-computed lighting map $\boldsymbol{M}_{\text{s}}$. A low-resolution map $\boldsymbol{M}_{\text{d}}$ is also created to represent low-frequency diffuse lighting and the diffuse part in Eq.(4) is simplified as

$$\boldsymbol{C}_{\text{d}}(\boldsymbol{x}, \boldsymbol{\omega}_o) = \boldsymbol{a}_{\text{d}}(\boldsymbol{a}, m)\boldsymbol{L}_{\text{diff}}(\boldsymbol{x}, \boldsymbol{n}, \boldsymbol{M}_{\text{d}}), \quad (9)$$

where $\boldsymbol{a}_{\text{d}} = (1-m)\boldsymbol{a}$ indicates the diffuse albedo and $\boldsymbol{L}_{\text{diff}} = \text{Tex\_Sample}(\boldsymbol{x}, \boldsymbol{n}, \boldsymbol{M}_{\text{d}})$. We show some rendered examples in the Supplement.

Discussion. We render images with varying materials and illumination using varied camera poses, materials, and lighting conditions rather than solely changing the lighting. Please refer to the Supplement for more details. This approach offers two distinct advantages. Firstly, multi-view images provide more geometric cues than single-view images, which are crucial for reconstructing globally reasonable geometry [20–22]. Secondly, by varying materials, we can create images with glossy appearances, particularly when the metallic component is high and roughness is low. These varied images serve as inputs, enhancing the model's robustness to variations in appearance. Furthermore, the core principles of photometric stereo remain applicable. In equations Eq.(8) and Eq.(9), the observation direction $\boldsymbol{\omega}_o$, metallic value m, roughness ρ, and mipmaps $\boldsymbol{M}_{\text{d}}$ and $\boldsymbol{M}_{\text{s}}$ are all known. Predicting the surface normals $\boldsymbol{n}$ and the albedo $\boldsymbol{a}$ still equates to solving these equations. Moreover, compared to merely changing lighting, altering the metallic and roughness values allows for diverse shading color rendering, which produces a richer set of equations.

Mesh Rasterization Rendering. Given an object with explicit mesh $\boldsymbol{O}$, rasterization is utilized to determine surface points $\boldsymbol{x}$, along with corresponding depth $\boldsymbol{d}$, surface normals $\boldsymbol{n}$, and mask $\boldsymbol{m}$. After obtaining the surface points $\boldsymbol{x}$ and their surface normals $\boldsymbol{n}$ along with selected camera pose, materials and lighting, we leverage split-sum approximation to estimate the specular and diffuse color as Eq.(8) and Eq.(9), respectively. During the process, besides the shading color, we can also render albedo, specular light, and diffuse light maps. The entire process can be formulated as

$$\{\boldsymbol{C}, \boldsymbol{n}, \boldsymbol{d}, \boldsymbol{m}, \boldsymbol{a}, \boldsymbol{L}_{\text{spec}}, \boldsymbol{L}_{\text{diff}}\} = \text{PBR}(\text{Rasterization}(\boldsymbol{O})), \quad (10)$$

where $\boldsymbol{C}$, $\boldsymbol{n}$, $\boldsymbol{d}$, $\boldsymbol{m}$, $\boldsymbol{a}$, $\boldsymbol{L}_{\text{spec}}$, and $\boldsymbol{L}_{\text{diff}}$ are the rendered color, normal, depth, mask, albedo, specular light, and diffuse light maps, respectively.

3.3. PRM

Mesh as 3D Representation. The previous LRM series typically integrate triplane as 3D representation. In contrast, we opt for an explicit representation using mesh as our 3D format, which enables the use of differentiable PBR method for better radiance modeling. As a result, specular and diffuse lighting maps are also renderable, providing extra photometric cues that are only related to surface normals. Moreover, PBR can effectively model the specular component, leading to improved geometry reconstruction results [7, 41]. Specifically, we leverage differentiable isosurface extraction module, namely FlexiCubes [36].

Two Stage Optimization. Inspired by InstantMesh [46], we have similarly designed a two-stage optimization framework. The first stage mirrors Instantmesh, using triplane and volume rendering for optimization with offline rendered data. In the second stage, FlexiCubes is used as the 3D representation. To reuse the knowledge in the first stage, we load the pretrained model as in InstantMesh [46]. The original color MLP is repurposed as an albedo MLP to incorporate color priors. Since our focus is on reconstruction, we directly utilize the ground-truth metallic and roughness during training. Given that an explicit mesh serves as our 3D representation, novel views can be rendered as described in Eq.(10). The difference is that the mesh $\hat{\boldsymbol{O}}$ is extracted using the dual marching cubes algorithm [30], which utilizes predicted SDF values, deformation, and weights derived from the triplane formulated by

$$\{\hat{\boldsymbol{C}}, \hat{\boldsymbol{n}}, \hat{\boldsymbol{d}}, \hat{\boldsymbol{m}}, \hat{\boldsymbol{a}}, \hat{\boldsymbol{L}}_{\text{spec}}, \hat{\boldsymbol{L}}_{\text{diff}}\} = \text{PBR}(\text{Rasterization}(\hat{\boldsymbol{O}})). \quad (11)$$

Optimization. During the training process, our total loss function is

$$\begin{aligned} \mathcal{L} = \mathcal{L}_{\text{MSE}}(\boldsymbol{C}, \hat{\boldsymbol{C}}) + \lambda_{\text{LPIPS}}\mathcal{L}_{\text{LPIPS}}(\boldsymbol{C}, \hat{\boldsymbol{C}}) \\ + \mathcal{L}_{\text{MSE}}(\boldsymbol{a}, \hat{\boldsymbol{a}}) + \lambda_{\text{LPIPS}}\mathcal{L}_{\text{LPIPS}}(\boldsymbol{a}, \hat{\boldsymbol{a}}) \\ + \mathcal{L}_{\text{MSE}}(\boldsymbol{l}_*, \hat{\boldsymbol{l}}_*) + \lambda_{\text{LPIPS}}\mathcal{L}_{\text{LPIPS}}(\boldsymbol{l}_*, \hat{\boldsymbol{l}}_*) \\ + \lambda_{\text{normal}}\hat{\boldsymbol{m}} \otimes (1 - \boldsymbol{n} \cdot \hat{\boldsymbol{n}}) + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}} \\ + \lambda_{\text{depth}}\hat{\boldsymbol{m}} \otimes \|\boldsymbol{d} - \hat{\boldsymbol{d}}\|_1 + \lambda_{\text{mask}}\mathcal{L}_{\text{MSE}}(\boldsymbol{m}, \hat{\boldsymbol{m}}), \end{aligned} \quad (12)$$

where $\mathcal{L}_{\text{MSE}}$ and $\mathcal{L}_{\text{LPIPS}}$ indicates the mean squaree error loss and LPISP loss [52], respectively. $* \in \{\text{spec}, \text{diff}\}$ denotes specular light and diffuse light maps, respectively. $\mathcal{L}_{\text{reg}}$ is the regularization terms used in FlexiCubes [36]. During training, we set $\lambda_{\text{LPIPS}} = 2.0$, $\lambda_{\text{normal}} = 0.2$, $\lambda_{\text{depth}} = 0.5$, $\lambda_{\text{mask}} = 1.0$ and $\lambda_{\text{reg}} = 0.01$.

For both the specular $\boldsymbol{L}_{\text{spec}}$ and the diffuse lighting map $\boldsymbol{L}_{\text{diff}}$, which are directly influenced by surface normals, effectively optimizing these light maps significantly enhances

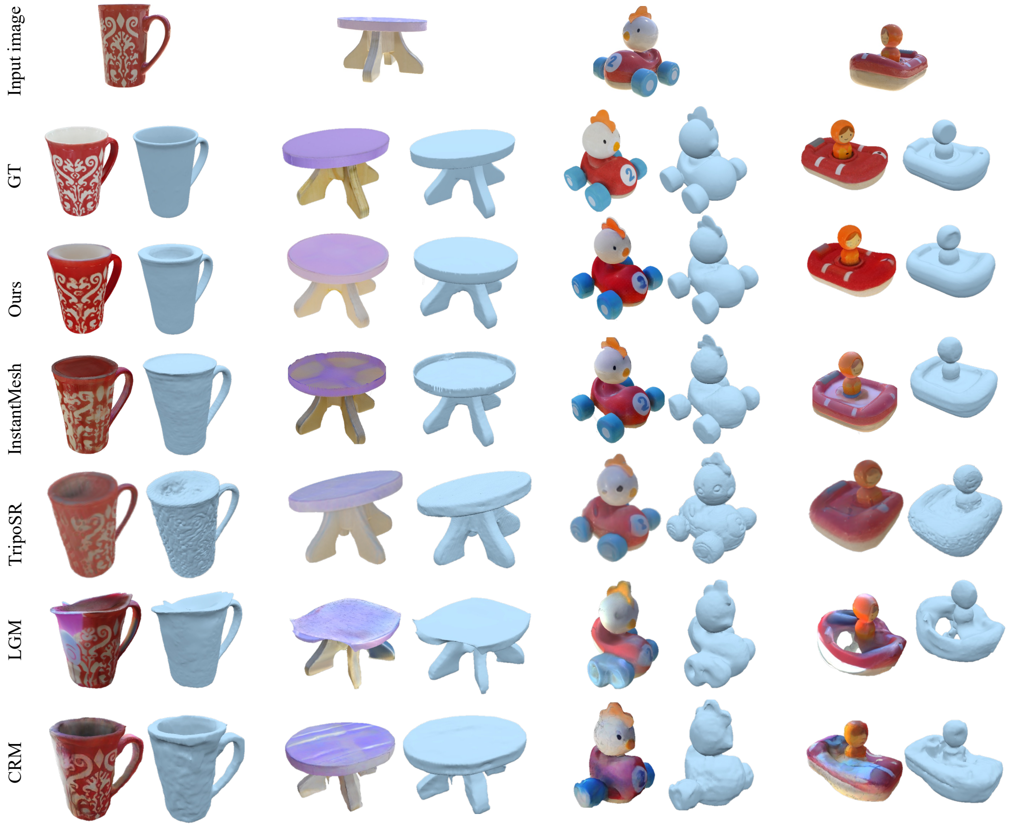

Figure 3. Qualitative comparisons with state-of-the art methods and ground truth for single-view reconstruction task. PRM reconstructs the highest quality 3D mesh and provides a more accurate texture prediction from input photographs compared to the others.

the surface normals, thereby refining the precision of local details. This process is analogous to photometric stereo. The key difference is that these maps are exclusively related to surface normals without considering albedo, as demonstrated in Eq.(8) and Eq.(9), in contrast to shading color.

Applications. PRM achieves high-quality geometry reconstruction with predicted albedo. This capability enables us to render the object under novel lighting conditions and also to modify its material properties. Examples of these applications are provided in the Supplement.

4. Experiments

4.1. Datasets and Evaluation Protocol

Datasets. For training, we used Objaverse [5], a dataset comprised of synthetic 3D assets, that allows us to control the lighting, geometry, and material properties. We first filter Objaverse to get a high-quality subset for training. The filtering process aims to exclude objects lacking texture maps or those of inferior quality, such as those with low-poly properties. Texture maps are essential as they provide detailed albedo maps; in their absence, vertex colors are used as a substitute for albedo. Moreover, low-poly meshes result in uneven surface lighting. During rendering, we maintained the original albedo unchanged and randomly selected a material combination from a total of 121 possibilities, which were derived by varying the metallic and roughness properties from 0 to 1 in increments of 0.1. For lighting, we utilized environment maps sampled from a collection of 679 maps available on Polyhaven.com, thereby ensuring a diverse range of lighting conditions.

For evaluation, We performed quantitative comparisons using two public datasets, including Google Scanned Objects (GSO) [6] and OmniObject3D (Omni3D) [45]. We randomly picked out 300 objects as the evaluation set both datasets, respectively. To show the robust capabilities of our model on appearance variations, we rendered the input view of each object with a randomly sampled combination of materials and lighting. We also report extra comparison results, following previous methods that utilized fixed lighting and did not change materials.

Evaluation Protocol. We evaluated both the 2D visual quality and the 3D geometric quality. For the 2D visual evaluation, we rendered novel views from the reconstructed 3D mesh and compared them with the ground truth views, using PSNR, SSIM, and LPIPS. Since other methods do not

Table 1. Quantitative comparison with methods on the GSO and Omni3D datasets, showcasing 3D reconstruction and 2D rendering metrics.

Dataset	GSO					OmniObject3D				
Metric	CD↓	FS@0.1↑	PSNR↑	SSIM↑	LPIPS↓	CD↓	FS@0.1↑	PSNR↑	SSIM↑	LPIPS↓
TripoSR	0.109	0.872	15.247	0.859	0.197	0.083	0.940	15.237	0.865	0.176
CRM	0.202	0.707	17.293	0.854	0.142	0.088	0.907	18.293	0.894	0.112
LGM	0.144	0.800	17.643	0.869	0.158	0.138	0.823	17.893	0.884	0.139
InstantMesh	0.076	0.931	19.988	0.901	0.096	0.096	0.892	18.608	0.903	0.096
PRM	0.050	0.981	25.125	0.929	0.061	0.053	0.979	25.063	0.932	0.063

predict albedo, we compared their shading color in novel views. For our method, we rendered the shading color based on the predicted mesh and albedo, then performed the metric calculations. For the 3D geometric evaluation, we first aligned the coordinate systems of the reconstructed meshes with those of the ground truth meshes. Subsequently, we repositioned and rescaled all meshes into a cube of size $[-1, 1]^3$. We reported the Chamfer Distance (CD) and F-Score (FS) at a threshold of 0.1, which were computed by uniformly sampling 16K points from the mesh surfaces.

4.2. Implementation Details

Our model was developed based on InstantMesh [46]. The architecture of the Transformer encoder, triplane transformer, and the FlexiCubes decoder mirrors that of InstantMesh. Our model underwent training for 7 days and 3 days on 32 NVIDIA RTX A800 GPUs for the first stage and second stage, respectively. During inference, we used a single RGB image to generate six fixed views by Zero123++ [37] as the input of PRM. For more details, please see our Supplement.

4.3. Comparison with State-of-the-Art Methods

We compared the proposed PRM with four baselines. These include TripoSR [40], an open-source LRM implementation renowned for its superior single-view reconstruction performance; CRM [43], a UNet-based Convolutional Reconstruction Model that reconstructs 3D meshes from generated multi-view images and canonical coordinate maps (CCMs); LGM [39], a unet-based Large Gaussian Model that reconstructs Gaussians from generated multi-view images; and InstantMesh [46], a transformer-based LRM that employs a two-stage training strategy for direct 3D mesh reconstruction. We reported both quantitative and qualitative comparative results for a complete comparison analysis.

Quantitative Results. We reported quantitative results with randomly selected lighting and materials in two different datasets in Table 1. We also report quantitative results without changing materials and using fixed lighting as previous methods compared with cutting-edge method Instantmesh on GSO in Table 2, where ground-truth rendered multi-view images were used as input.

For 3D reconstruction metrics, PRM achieves significant improvements over all state-of-the-art methods on both datasets, as shown in Table 1. The qualitative comparison with other methods is presented in Figure 3. We attribute these improvements to our use of images with varying materials and illumination for both input and supervision. This approach not only enables the model to learn fine-grained geometric details by providing photometric cues but also enhances the model's robustness to variations in image appearance. Further validation of these results in the test setting without material changes and using fixed lighting is shown in Table 2.

For 2D visual metrics, our approach effectively mitigates the impact of lighting variations to accurately restore the original colors of objects, as shown in Table 1. We surpass all current methods across all metrics.

Table 2. Comparison on GSO dataset. "Random m&r" indicates whether materials and lighting were randomly changed.

Method	Random m&r	CD↓	FS@0.1↑	PSNR↑
InstantMesh	✓	0.061	0.934	21.115
InstantMesh	×	0.048	0.972	23.644
PRM	✓	0.053	0.982	24.602
PRM	×	0.043	0.991	26.377

Table 3. Quantitative results of the ablation study.

Metric	CD↓	FS@0.1↑	PSNR↑
Only albedo supervision	0.099	0.887	17.532
w/o PBR	0.089	0.909	19.143
w/o lighting supervision	0.073	0.919	20.114
w/o change materials	0.089	0.894	19.662
Full model	0.066	0.948	20.992

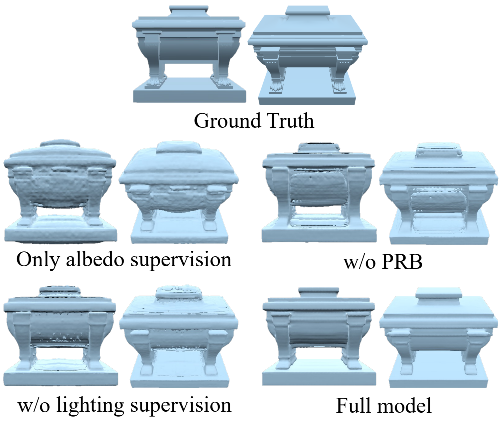

Figure 4. Visualization of our ablation study to validate the effectiveness of each component in our framework.

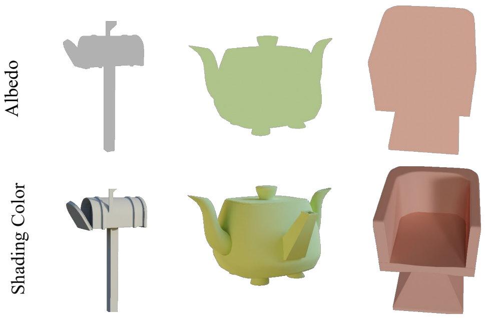

Figure 5. Shading color offers significant photometric cues for perceiving geometry, whereas albedo lacks this information.

Qualitative Results. For the qualitative comparison, we randomly selected four images from the GSO dataset to serve as inputs for 3D model reconstruction. For each reconstructed mesh, we visualized both the albedo (ours) and the shading color (others), as well as the pure geometry.

As shown in Figure 3, the results reconstructed by PRM exhibit more accurate geometry and appearance. Our model can reconstruct precise geometry and accurately predict albedo from images with specular highlights, whereas other methods fail. For instance, InstantMesh often predicts uneven geometric surfaces and tends to reconstruct incorrect geometry. Similarly, both CRM and LGM struggle to produce satisfactory results, falling short in both geometry accuracy and texture prediction. This underscores the robustness and superior performance of our PRM method in handling complex lighting conditions and intricate surface details, making it a more reliable choice for high-quality 3D model reconstruction.

4.4. Ablation Study

We conducted the ablation study to validate the effectiveness of each component in our framework. We reported quantitative results of our ablation study in Table 3. For ablations, we reduce the training set from 120K to 30K for faster experiments due to the heavy cost.

The effectiveness of PBR. PBR plays a crucial role in improving geometry, especially for glossy surfaces. To validate the effectiveness of PBR, we conducted experiments by directly predicting shading colors using an MLP, rather than deriving results with PBR. However, direct prediction of shading color, without accounting for view-dependent appearances, struggles to model such effects and thus fails to accurately reconstruct geometry. We presented these results as “w/o PBR” in Table 3 and Figure 4.

Albedo supervision. To avoid the interference caused by specular color on the surface, an intuitive approach is to directly use albedo instead of shading color for supervision. However, this method proves ineffective in practice, as albedo contains few photometric cues, thereby hindering geometry reconstruction. For example, a concave surface with uniform albedo may appear planar without the presence of cast shadows, while shading color provides significant photometric cues critical for accurately perceiving geometry, as illustrated in Figure 5. We conducted an ablation study within our framework by excluding shading color and light maps from supervision. The qualitative comparison is shown in and Table 3 and Figure 4 denoted as “Only albedo supervision”. The results demonstrate that photometric cues are crucial for accurate geometry reconstruction.

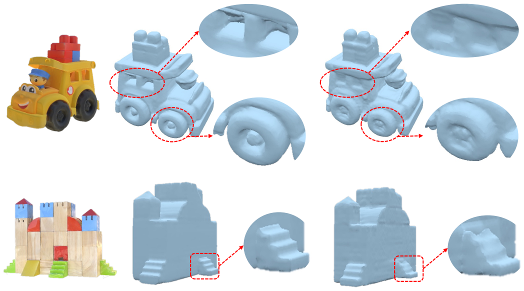

Figure 6. Ablation study of the effect of changing materials.

Lighting maps supervision. We conducted the ablation study within our framework by excluding the lighting maps loss. The results, denoted as “w/o lighting supervision”, are displayed in Table 3 and Figure 4 . Since lighting maps are exclusively related to surface normals and do not include albedo, optimizing these maps proves particularly beneficial for enhancing the optimization of fine-grained local details.

Variations in materials. We conducted an ablation study to illustrate the effectiveness of varying materials during training. The results are shown in Table 3 and Figure 6. Without changing the materials, some details are lost. Moreover, the model has lost the capability to reconstruct glossy surfaces.

5. Conclusion and Limitation

Limitation. Despite the high-quality results achieved in this work, there remain several limitations for future research to explore: 1) Firstly, the reconstructed 3D model is sensitive to the quality of multi-view images. 2) Secondly, the accuracy of the estimated albedo appears to be somewhat entangled with the lighting conditions.

Conclusion. In this work, we introduce PRM, a novel feed-forward framework designed to reconstructed high-quality 3D assets with fine-grained local details, even amidst complex image appearances. To achieve this goal, we utilize images with varying materials and illumination as both input and supervision, providing sufficient photometric cues for fine-grained geometry recovery and enhancing the model’s robustness to variations in image appearance. Using a mesh as our 3D representation, we employ differentiable PBR for predicted rendering, underpinning the utilization of multiple supervisions for optimization. Experiments on public datasets validate that PRM surpasses other methods.

6. Acknowledgement

This research is supported by Guangdong Provincial Key Lab of Integrated Communication, Sensing and Computation for Ubiquitous Internet of Things (No.2023B1212010007).

References

[1] Jonathan T Barron and Jitendra Malik. Shape, illumination, and reflectance from shading. *IEEE transactions on pattern analysis and machine intelligence (TPAMI)*, 2014. 3

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, 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 (CVPR)*, 2022. 1

[4] Zhifei Chen, Tianshuo Xu, Wenhang Ge, Leyi Wu, Dongyu Yan, Jing He, Luozhou Wang, Lu Zeng, Shunsi Zhang, and Yingcong Chen. Uni-renderer: Unifying rendering and inverse rendering via dual stream diffusion. *arXiv preprint arXiv:2412.15050*, 2024. 3

[5] 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)*, 2023. 2, 6

[6] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. In *2022 International Conference on Robotics and Automation (ICRA)*, 2022. 6

[7] Wenhang Ge, Tao Hu, Haoyu Zhao, Shu Liu, and Ying-Cong Chen. Ref-neus: Ambiguity-reduced neural implicit surface learning for multi-view reconstruction with reflection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 3, 5

[8] Jason Gregory. *Game engine architecture*. AK Peters/CRC Press, 2018. 1

[9] Hideki Hayakawa. Photometric stereo under a light source with arbitrary motion. *JOSA A*, 1994. 3

[10] Carlos Hernandez, George Vogiatzis, and Roberto Cipolla. Multiview photometric stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2008. 1, 3

[11] Roland Hess. *Blender foundations: The essential guide to learning blender 2.5*. Routledge, 2013. 1, 4

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 1

[13] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*, 2023. 1, 2, 3

[14] Tao Hu, Wenhang Ge, Yuyang Zhao, and Gim Hee Lee. X-ray: A sequential 3d representation for generation, 2024. 2

[15] Tao Hu, Wenhang Ge, Yuyang Zhao, and Gim Hee Lee. X-ray: A sequential 3d representation for generation. *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. 1, 2

[16] Satoshi Ikehata. Universal photometric stereo network using global lighting contexts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[17] Satoshi Ikehata. Scalable, detailed and mask-free universal photometric stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[18] James T Kajiya. The rendering equation. In *Proceedings of the 13th annual conference on Computer graphics and interactive techniques*, 1986. 3

[19] Brian Karis and Epic Games. Real shading in unreal engine 4. *Proc. Physically Based Shading Theory Practice*, 2013. 2, 4

[20] Berk Kaya, Suryansh Kumar, Carlos Oliveira, Vittorio Ferrari, and Luc Van Gool. Uncertainty-aware deep multi-view photometric stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3, 5

[21] Berk Kaya, Suryansh Kumar, Francesco Sarno, Vittorio Ferrari, and Luc Van Gool. Neural radiance fields approach to deep multi-view photometric stereo. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision (WACV)*, 2022. 3

[22] Berk Kaya, Suryansh Kumar, Carlos Oliveira, Vittorio Ferrari, and Luc Van Gool. Multi-view photometric stereo revisited. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2023. 3, 5

[23] John Lasseter. Principles of traditional animation applied to 3d computer animation. In *Proceedings of the 14th annual conference on Computer graphics and interactive techniques*, 1987. 1

[24] Jiahao Li, Hao Tan, Kai Zhang, Zexiang Xu, Fujun Luan, Yinghao Xu, Yicong Hong, Kalyan Sunkavalli, Greg Shakhnarovich, and Sai Bi. Instant3d: Fast text-to-3d with sparse-view generation and large reconstruction model. *arXiv preprint arXiv:2311.06214*, 2023. 2

[25] Yixun Liang, Xin Yang, Jiantao Lin, Haodong Li, Xiaogang Xu, and Yingcong Chen. Luciddreamer: Towards high-fidelity text-to-3d generation via interval score matching. *arXiv preprint arXiv:2311.11284*, 2023. 1

[26] 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)*, 2023. 1

[27] Jianman Lin, Haojie Li, Chunmei Qing, Zhijing Yang, Liang Lin, and Tianshui Chen. Geometry-editable and appearance-preserving object compositon. *arXiv preprint arXiv:2505.20914*, 2025. 1

[28] Yuan Liu, Peng Wang, Cheng Lin, Xiaoxiao Long, Jiepeng Wang, Lingjie Liu, Taku Komura, and Wenping Wang. Nero: Neural geometry and brdf reconstruction of reflective objects from multiview images. In *SIGGRAPH*, 2023. 3

[29] Zhipeng Mo, Boxin Shi, Feng Lu, Sai-Kit Yeung, and Yasuyuki Matsushita. Uncalibrated photometric stereo under natural illumination. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[30] Gregory M Nielson. Dual marching cubes. In *IEEE visualization 2004*, 2004. 5

[31] Merlin Nimier-David, Delio Vicini, Tizian Zeltner, and Wenzel Jakob. Mitsuba 2: A retargetable forward and inverse renderer. *ACM Transactions on Graphics (TOG)*, 2019. 3

[32] Rick Parent. *Computer animation: algorithms and techniques*. Newnes, 2012. 1

[33] Jaesik Park, Sudipta N Sinha, Yasuyuki Matsushita, Yu-Wing Tai, and In So Kweon. Multiview photometric stereo using planar mesh parameterization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2013. 3

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

[35] Martijn J Schuemie, Peter Van Der Straaten, Merel Krijn, and Charles APG Van Der Mast. Research on presence in virtual reality: A survey. *Cyberpsychology & behavior*, 2001. 1

[36] Tianchang Shen, Jacob Munkberg, Jon Hasselgren, Kangxue Yin, Zian Wang, Wenzheng Chen, Zan Gojcic, Sanja Fidler, Nicholas Sharp, and Jun Gao. Flexible isosurface extraction for gradient-based mesh optimization. *ACM Trans. Graph.*, 2023. 1, 5

[37] Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*, 2023. 7

[38] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 1

[39] Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. *arXiv preprint arXiv:2402.05054*, 2024. 1, 2, 7

[40] Dmitry Tochilkin, David Pankratz, Zexiang Liu, Zixuan Huang, Adam Letts, Yangguang Li, Ding Liang, Christian Laforte, Varun Jampani, and Yan-Pei Cao. Triposr: Fast 3d object reconstruction from a single image. *arXiv preprint arXiv:2403.02151*, 2024. 7

[41] Dor Verbin, Peter Hedman, Ben Mildenhall, Todd Zickler, Jonathan T Barron, and Pratul P Srinivasan. Ref-nerf: Structured view-dependent appearance for neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 3, 5

[42] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021. 1

[43] Zhengyi Wang, Yikai Wang, Yifei Chen, Chendong Xiang, Shuo Chen, Dajiang Yu, Chongxuan Li, Hang Su, and Jun Zhu. Crm: Single image to 3d textured mesh with convolutional reconstruction model. *arXiv preprint arXiv:2403.05034*, 2024. 1, 2, 7

[44] Robert J Woodham. Photometric method for determining surface orientation from multiple images. *Optical engineering*, 1980. 3

[45] Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Jiawei Ren, Liang Pan, Wayne Wu, Lei Yang, Jiaqi Wang, Chen 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 (CVPR)*, 2023. 6

[46] Jiale Xu, Weihao Cheng, Yiming Gao, Xintao Wang, Shenghua Gao, and Ying Shan. Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. *arXiv preprint arXiv:2404.07191*, 2024. 1, 2, 5, 7

[47] Xinli Xu, Wenhang Ge, Jiantao Lin, Jiawei Feng, Lie Xu, HanFeng Zhao, Shunsi Zhang, and Ying-Cong Chen. Flexgen: Flexible multi-view generation from text and image inputs. *arXiv preprint arXiv:2410.10745*, 2024. 2

[48] Yinghao Xu, Hao Tan, Fujun Luan, Sai Bi, Peng Wang, Jiahao Li, Zifan Shi, Kalyan Sunkavalli, Gordon Wetzstein, Zexiang Xu, et al. Dmv3d: Denoising multi-view diffusion using 3d large reconstruction model. *arXiv preprint arXiv:2311.09217*, 2023. 1, 2

[49] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 1

[50] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 1

[51] Longwen Zhang, Ziyu Wang, Qixuan Zhang, Qiwei Qiu, Anqi Pang, Haoran Jiang, Wei Yang, Lan Xu, and Jingyi Yu. Clay: A controllable large-scale generative model for creating high-quality 3d assets. *arXiv preprint arXiv:2406.13897*, 2024. 1, 2

[52] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2018. 5

[53] Dongxu Zhao, Daniel Lichy, Pierre-Nicolas Perrin, Jan-Michael Frahm, and Soumyadip Sengupta. Mvpsnet: Fast generalizable multi-view photometric stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)*, 2023. 3