# MAGE 🧙 : Single Image to Material-Aware 3D via the Multi-View G-Buffer Estimation Model

Haoyuan Wang[1*], Zhenwei Wang[1*], Xiaoxiao Long[2], Cheng Lin[3], Gerhard Hancke[1], Rynson W.H. Lau[1†]
[1]City University of Hong Kong, [2]Nanjing University, [3]The University of Hong Kong
* Joint first authors † Corresponding author

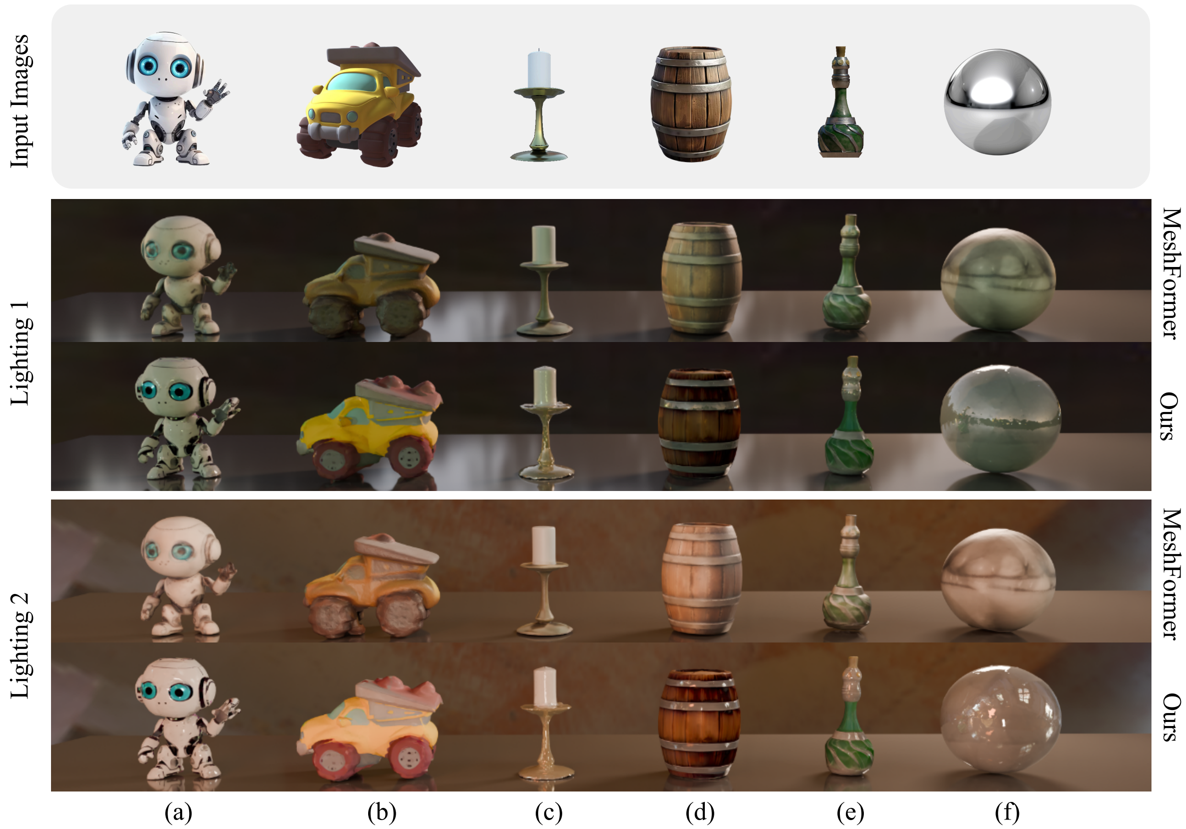

Figure 1. Comparison between the single image to 3D generation results of MeshFormer [27] and our MAGE, under two novel lighting conditions. Our method successfully decomposes material properties into clean albedo and accurate roughness & metallic properties, enabling realistic rendering under different illuminations. In contrast, MeshFormer bakes lighting effects of input images (*e.g.*, shadows in (b) and specular reflections in (f)) directly into the texture, resulting in incorrect appearances when lighting conditions change.

## Abstract

*With advances in deep learning models and the availability of large-scale 3D datasets, we have recently witnessed significant progress in single-view 3D reconstruction. However, existing methods often fail to reconstruct physically based material properties given a single image, limiting their applicability in complicated scenarios. This paper presents a novel approach (named MAGE) for generating 3D geometry with realistic decomposed material properties given a single image as input. Our method leverages inspiration from traditional computer graphics deferred rendering pipelines to introduce a multi-view G-buffer estimation model. The proposed model estimates G-buffers for various views as multi-domain images, including XYZ coordinates, normals, albedo, roughness, and metallic properties from a single-view RGB image. To address the inherent ambiguity and inconsistency in generating G-buffers simultaneously,* *we also formulate a deterministic network from the pretrained diffusion models and propose a lighting response loss that enforces consistency across these domains using PBR principles. Finally, we propose a large-scale synthetic dataset rich in material diversity for our model training. Experimental results demonstrate the effectiveness of our method in producing high-quality 3D meshes with rich material properties. Our code and dataset can be found at* `https://www.whyy.site/paper/mage`.

## 1. Introduction

3D content creation has seen remarkable progress in recent years, driven by advancements in deep learning, large-scale generative models, and 3D datasets. Recent works have explored various approaches for 3D generation from a single image. Some works [28, 31, 42] improve 3D consistency by generating novel views using diffusion models to guide the following 3D reconstruction process. Another line of research [18, 20] focuses on learning direct single image to 3D mapping using transformer-based architectures.

However, existing approaches that estimate 3D models from a single input image often ignore the physically based rendering (PBR) material properties. This limitation in capabilities is particularly significant for 3D content creation demands nowadays. Reconstructing PBR material properties is crucial for creating high-quality, photorealistic 3D content that accurately represents object appearance under various lighting conditions. Predicting 3D models with PBR materials presents the following challenges: **(1) Appearance complexity.** The appearance of an object in an image results from the complex interaction between its shape, material properties, and lighting conditions, which makes directly reasoning the final 3D model with materials from a single image intricate for neural networks to learn. **(2) Decomposition ambiguity.** Material, geometry, and lighting are interdependent. Disentangling and identifying these individual properties from a single image under unknown illumination is inherently ambiguous, making the process ill-defined.

In this paper, we aim to bridge this gap by introducing a novel approach, namely MAGE, to generate 3D models with realistic materials given a single input image. To tackle the challenge of appearance complexity, we observe that the intermediate rasterization images of the deferred rendering pipeline, known as G-buffers [8], provide a structured, view-dependent 3D representation that naturally decouples geometric information (XYZ coordinates and normals) and material properties (albedo, roughness, and metallic) into regular 2D maps, making them more amenable for deep learning prediction. Inspired by this, we propose a multi-view G-buffer estimation network to estimate various G-buffer attributes for different viewpoints, which helps to decompose the complex task of material-aware single image to 3D generation into a more manageable G-buffers estimation task.

To address the challenge of decomposition ambiguity for consistent G-buffers estimation, we introduce a set of image space loss functions, including a proposed lighting response loss based on the PBR principles. The lighting response loss constrains the similarity between the ground truth colors and the shaded pixel colors using physically based image-based rendering (IBR) under a given lighting condition, which enforces consistency between the various estimated attributes and addresses the ambiguity inherent in generating multiple attributes simultaneously without predicting the illumination explicitly. Using principles from PBR, our generated textures are visually plausible and physically consistent. To effectively use image space loss functions for G-buffer learning, we propose a deterministic single-step architecture for the G-buffer estimation network. This architecture builds upon a pretrained multi-view diffusion U-Net and is carefully designed to leverage the knowledge from the pretrained model for faster convergence while maintaining deterministic behavior without denoising processes.

To facilitate model training, we create a diverse synthetic dataset based on the existing large-scale 3D dataset Objaverse [9]. We present a procedural material generation method that produces a comprehensive dataset of multi-view G-buffers while enriching the material properties of the original meshes. Our dataset provides valuable supervision for the model to learn to disentangle and reconstruct material properties from a single image. It allows our network to learn the complex relationships between geometry, lighting, and material properties. Extensive experiments demonstrate that our model, trained on the proposed dataset, generalizes well to wild input images.

Our contributions can be summarized as follows:

1. Inspired by deferred rendering, we propose a novel framework for a material-aware single image to 3D generation by simultaneously predicting G-buffers with image space loss functions.
2. We propose an efficient neural network to deterministically predict G-buffers with a lighting response loss that constrains the consistency across material properties.
3. We propose a diverse synthetic dataset for training models on PBR textured 3D generation tasks.
4. We conduct extensive experiments to validate the effectiveness of our proposed method and compare it with existing approaches.

## 2. Related Works

**Single Image to 3D Generation.** Early attempts [33, 37, 47, 55] use score distillation sampling (SDS) [36] on pretrained diffusion models for image-conditioned 3D generation. With the advent of large-scale 3D datasets [9, 10], recent studies have proposed feed-forward models for image-to-3D generation based on various representations, such as point clouds [35, 70], neural field [15, 20, 54, 63, 65], SDF [5, 6, 69], and Gaussian splatting [64]. Other works leverage the power of the transformer model [49] to generate 3D models autoregressively [3, 45]. Additionally, some approaches combine multi-view diffusion models [4, 28, 42, 51, 53] with sparse-view 3D reconstruction [27, 48, 56, 58, 59] to enhance generalization capabilities. Despite these promising results, none of these methods can tackle the challenging task of predicting physically accurate multi-view material properties from input images for material-aware 3D generation.

**Material-Aware 3D Reconstruction.** In the realm of material-aware 3D generation, some works focus on generating 3D shapes with materials from a text prompt. Fantasia3D [1] generates geometry and appearance with PBR materials in two separate optimization stages, supervised by the SDS loss. MATLABER [60] leverages both SDS and a novel latent BRDF auto-encoder for text-to-3D generation with the material. However, their results are entangled with lighting, making it difficult to obtain satisfactory relighted objects. Another line of works apply inverse rendering and decomposition learning in 3D reconstruction tasks from multi-views images based on NeRF [34, 50, 52], neural SDF [16, 30], neural implicit surface representation [32] or Gaussian splatting [13, 21, 44]. Other studies focus on generating exquisite textures with materials for existing 3D meshes based on text [2, 11, 38, 65, 68] and large multimodal model [68]. While these approaches focus on the generation of material-aware 3D models from text, sparse real-world views, and existing 3D meshes, we aim for material-aware 3D generation from a single wild image, which is more challenging due to the inherent ambiguities with unknown illuminations.

**Image Intrinsic Decomposition.** Image intrinsic decomposition aims to recover lighting, materials, and geometry from captured images. Early studies [17, 22, 26, 67] rely on time-consuming optimization processes, typically reconstructing the surface geometry first and then recovering materials and environmental lighting next. However, these approaches require dense multi-view inputs and case-specific optimization, making them computationally expensive for real-world applications. With the advent of data-driven generative models, learning-based methods have shown significant improvements in both the quality and efficiency of decomposition. Initial studies [41, 57] utilize physically-based deep networks for jointly learning inverse rendering and relighting, but they suffer from over-smoothed results due to limited data scale and model capacity. Recent works [14, 24] leverage pretrained image diffusion models [19, 40] to enhance the generative quality with high-frequency details. However, these methods face inherent inconsistency across intrinsic components. Our work leverages both diffusion priors and physically-based guidance to generate multi-view intrinsic attributes (G-buffers), enabling material-aware 3D generation from a single image.

## 3. Method

This section presents the methodology of our MAGE framework. We first provide an overview of our method (Sec.3.1), then detail our multi-view G-buffer estimation network (Sec.3.2) and the proposed lighting response loss (Sec. 3.3).

### 3.1. Overview

We aim to reconstruct a 3D model with physically-based material properties from a single view input. Inspired by deferred rendering in computer graphics, we observe that these pipelines' intermediate rasterization results, *i.e.*, G-buffers, offer a structured way to represent 3D information as 2D maps, decoupling geometry and material properties. This allows us to break down the complex 3D reconstruction problem into more manageable sub-tasks. Given the estimated G-buffers, we can unproject them back to 3D space to get the final 3D meshes. Thus, we approach this problem by estimating multi-view G-buffers from the single-view input image, including geometric information (XYZ coordinates $X$, surface normals $N$) and material properties (albedo $A$, roughness $R$, and metallic $M$) for rendering.

While a naive method is to finetune a SOTA multi-view diffusion model on a G-buffers dataset, *i.e.*, the "w/o Single-Step" model in our ablation study (Sec. 5.3), it is challenging to maintain consistency across G-buffers. Our G-buffers are equivalent to 5-domain maps, surpassing the complexity typically addressed in existing methods [31]. Besides, since geometry, materials, and lighting determine the final rendered RGB images jointly, directly predicting G-buffers supervised by ground truth might cause ambiguity and mutually conflicting due to the lack of physical consistency.

To address this problem while still using the generation capability of diffusion models, we focus on two aspects: (1) proposing a deterministic single-step G-buffer estimation network while still initialized from pretrained diffusion model weights, eliminating the denoising process to enhance consistency, and (2) using image space loss functions, including a proposed physically based lighting response loss to constrain physical consistency across estimated G-buffers and resolve ambiguities based on physical guidance and lighting conditions.

### 3.2. Multi-View G-Buffer Estimation Network

Given a single input image $I$, we first generate a set of multi-view RGB images $\{I_{i=1}^{N}\}$ using an off-the-shelf multi-view diffusion model [42] $F_m$, which is fine-tuned based on the commonly used stable diffusion model [39].

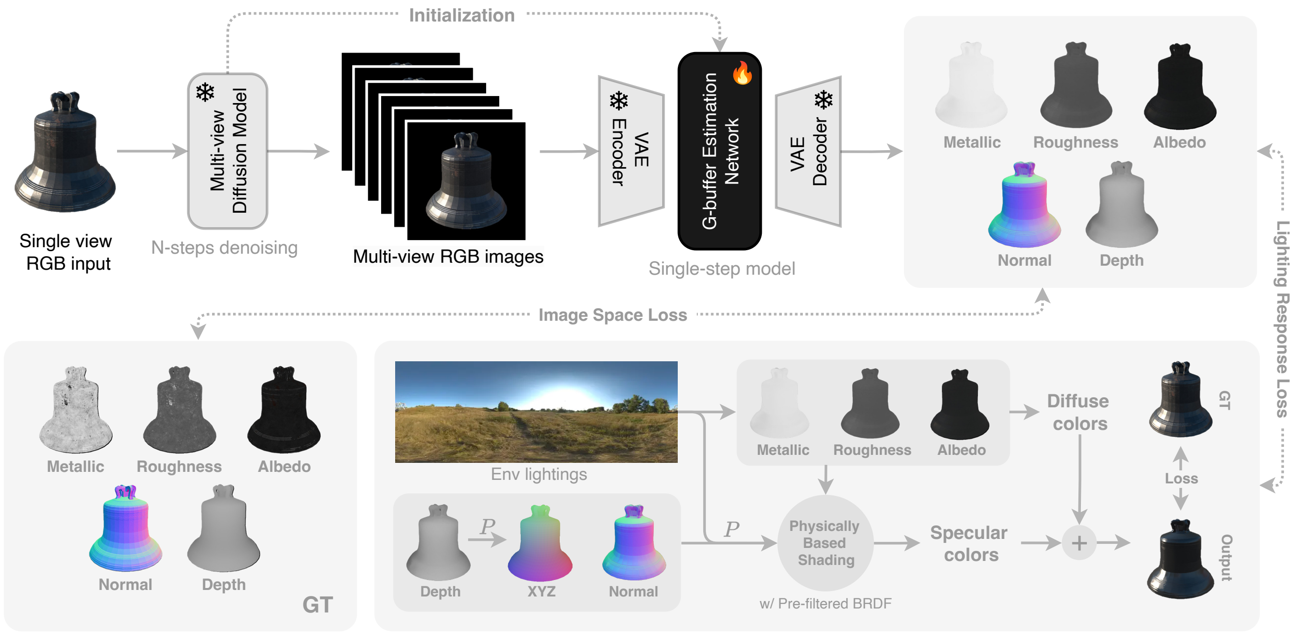

Figure 2. Overview of our framework. We use a pretrained diffusion model to generate multi-view images from a single-view input. Our deterministic single-step G-buffer estimation network will process them to predict G-buffers (metallic, roughness, albedo, normal, and XYZ coordinates, parameterized as a depth ma‘p). The predicted G-buffers are supervised by the image space loss with ground-truth G-buffers and a physically-based lighting response loss to ensure physical consistency across all G-buffer components.

This step provides consistent multi-view images of the object, which is essential for accurate 3D reconstruction. Then, we propose our G-buffer estimation network $F_g$ to predict consistent, high-quality, and physically accurate G-buffers for view $i$, denoted as $G_i = \{X_i, N_i, A_i, R_i, M_i\}$, where $X_i, N_i, A_i, R_i, M_i$ are XYZ coordinates map under the world coordinate system, camera space normal map, and the rasterization results of material properties: albedo, roughness and metallic, respectively. Overall, $F_g$ is based on the pretrained U-Net of $F_m$, and we introduce two key network designs for $F_g$.

**Network Initialization.** First, we design the output image layout of $F_g$ to align with the output layout of $F_m$. We recast our output G-buffers into a single tiled image in a $3 \times 2$ grid containing a recovered RGB image and five components of $G_i$. This form of tiling is widely used in multi-view image generation [42, 43, 53] for its compact representation and stable performance. By aligning our predicted G-buffers with $F_m$, we can fully leverage the powerful priors from the pretrained multi-view diffusion models for faster convergence and more stable training.

**Deterministic Single-Step Inference.** Second, we transform the multi-step denoising U-Net into a deterministic single-step U-Net, inspired by recent single-step diffusion models [14, 46, 61], for three reasons: (1) the multi-step inference process relies on high computational resources with low inference performance; (2) it is infeasible to add task-specific image space loss during training, such as the lighting response loss we will introduce in Sec. 3.3; and (3) the denoising-based inference process has stochastic factors introduced by the noise, which suits creative generation tasks, but is unstable for our task since the estimation of G-buffers should be deterministic from a given image.

To essentially transform the stochastic multi-step diffusion model into a deterministic single-step model, we replace the latent Gaussian noise with the latent representation of the tiled conditioned RGB image, which is obtained by $\mathbf{z}_{I'_i} = \mathcal{E}(I'_i)$, where $\mathcal{E}$ is the pretrained VAE encoder and $I'_i$ is the tiled conditioned RGB image, *i.e.*, copying $I_i$ into a $3 \times 2$ grid to align with the output. Moreover, instead of predicting a less noisy latent progressively, we directly predict the latent of the target G-buffers in a single pass. The deterministic single-step generation process is defined as:

$$G_i = \mathcal{D}(\mathbf{z}_{G_i}), \quad \mathbf{z}_{G_i} = F_g(\mathbf{z}_{I'_i}) \tag{1}$$

where $\mathbf{z}_{G_i}$ is the predicted latent of $G_i$ and $\mathcal{D}$ is the pretrained VAE decoder. In training, the predicted G-buffer $G_i$ can be used to compute task-specific image space loss.

### 3.3. Lighting Response Loss Function

With the deterministic single-step model $F_g$, we can apply image space loss functions instead of denoising training objectives to address the problem of the inherent conflict and ambiguity of the predicted G-buffers. The ambiguity arises from the inconsistency across the predicted G-buffers and the uncertainty of the lighting conditions. Thus, we introduce a physically based lighting response loss, which adopts a differentiable image-based renderer $F_r$ to render a lighting response image $\hat{I}_i$ from the predicted G-buffers $G_i$, using existing environmental lightings in our dataset. Formally, given the environmental lighting $L$ and the predicted G-buffers $G_i$, we obtain the re-rendered image $\hat{I}_i$ by:

$$\hat{I}_i = F_r(G_i, L), \tag{2}$$

where the XYZ coordinates of $G_i$ are transformed from the predicted depth map using the training camera pose $P$, as shown in Fig. 2. The lighting response loss is then computed between the lighting response image $\hat{I}_i$ and the ground truth rendered image $I_i^* = F_r(G_i^*, L)$. This loss ensures that our predicted G-buffers are physically consistent when combined through the rendering process.

The renderer function $F_r$ should be fully differentiable and physically accurate and produce rendering results close to GT. Some commonly used PBR methods, like differentiable global-illumination algorithms, are typically too slow for deep learning training, while some differentiable rendering methods, like SoftRas [29], lack physical accuracy. As a trade-off between the rendering quality, differentiability, and speed, we choose the split-sum image-based rendering method [23] as $F_r$. As illustrated in Fig. 2, the rendering result is composed of diffuse colors and specular colors, where specular colors can be approximated as the following equation according to [23] using the microfacet BRDF [7]:

$$\hat{I}_p \approx \int_{\Omega} f(\omega_r, \omega_p)(\omega_r \cdot \mathbf{n}_p) d\omega_r \int_{\Omega} L_i(\omega_r) D(\omega_r, \omega_p)(\omega_r \cdot \mathbf{n}_p) d\omega_r, \tag{3}$$

where $f$ is the BRDF function, $\omega_r$ is the reflected ray direction sampled in hemisphere $\Omega$ with the environmental lighting color $L_r(\omega_r)$. $\omega_p$ and $\mathbf{n}_p$ are outgoing ray direction and the normal vector for pixel $p$. $D$ is the geometry distribution function used in microfacet BRDF. Both integrals can be pre-computed given the environmental map $L$. We use a differentiable version of the renderer from [34].

When computing lighting response loss, we isolate the gradients for each G-buffer component by using ground truth values for all other components. Specifically, for each component $c \in \{X, N, A, R, M\}$, we compute:

$$\mathcal{L}_L = \sum_c \frac{1}{HW} \sum_p ||F_r(G^c_{i,p}, L) - I^*_{i,p}||_2^2, \tag{4}$$

where $G_i^c$ represents a hybrid set of G-buffers where only the component $c$ is predicted while all other components use GT values. This approach ensures that each component receives clear supervision signals without ambiguity from potentially incorrect predictions of other components.

**Overall Training Objectives.** We have also adopted a series of other image space losses for each G-buffer attribute. Formally, given ground-truth G-buffers $G_i^* = \{X_i^*, N_i^*, A_i^*, R_i^*, M_i^*\}$ as supervision, we have $L_2$ loss and LPIPS loss for albedo $A_i$ as:

$$\mathcal{L}_A = \frac{1}{HW} \sum_p ||A_{i,p} - A^*_{i,p}||_2^2 + \mathcal{L}_{lpips}(A_{i,p}, A^*_{i,p}), \tag{5}$$

L1 loss for monocular depth (derived from $X_i$), roughness $R_i$, and metallic $M_i$ as:

$$\mathcal{L}_{Q \in \{X,R,M\}} = \frac{1}{HW} \sum_p ||Q_{i,p} - Q^*_{i,p}||_1 \tag{6}$$

and Angle-based loss and LPIPS loss [66] for normal $N_i$ as:

$$\mathcal{L}_N = \frac{1}{HW} \sum_p \arccos\left(\frac{N_{i,p} \cdot N^*_{i,p}}{\|N_{i,p}\| \|N^*_{i,p}\|}\right) + \mathcal{L}_{lpips}(N_{i,p}, N^*_{i,p}). \tag{7}$$

Finally, the overall loss is weighted using $\lambda_W$:

$$\mathcal{L}_G = \sum_W \lambda_W \mathcal{L}_W, \; W \in \{L, X, N, A, R, M\} \tag{8}$$

## 4. Dataset

To facilitate the training of our network, we present a comprehensive data generation pipeline that creates paired training samples with explicit G-buffer supervision for material properties. Our pipeline leverages Blender to process 3D models from Objaverse [9], generating a large-scale dataset as illustrated in Fig. 3.

**Material Enhancement.** While Objaverse provides a large collection of 3D models, most of them lack physically based rendering (PBR) materials. To address this limitation, we implement a filtering and enhancement process. We first filter the dataset by removing models that are untextured, lack PBR material support (*e.g.*, with non-PBR materials), or contain only uniform colors. We then propose a randomized material assignment process to create diverse material appearances. For each 3D model, we randomly assign material properties to each sub-mesh from a predefined range of roughness and metallic values (ranging from 0 to 1), drop the original materials based on the predefined rules and combine the assigned properties into unified outputs while preserving the original texture mappings. While this random assignment might occasionally result in “unreasonable” combinations (*e.g.*, a highly metallic wooden surface), this does not affect our training objectives as the primary goal of the dataset is to provide supervision for 3D reconstruction with physically consistent materials, rather than maintaining semantic accuracy.

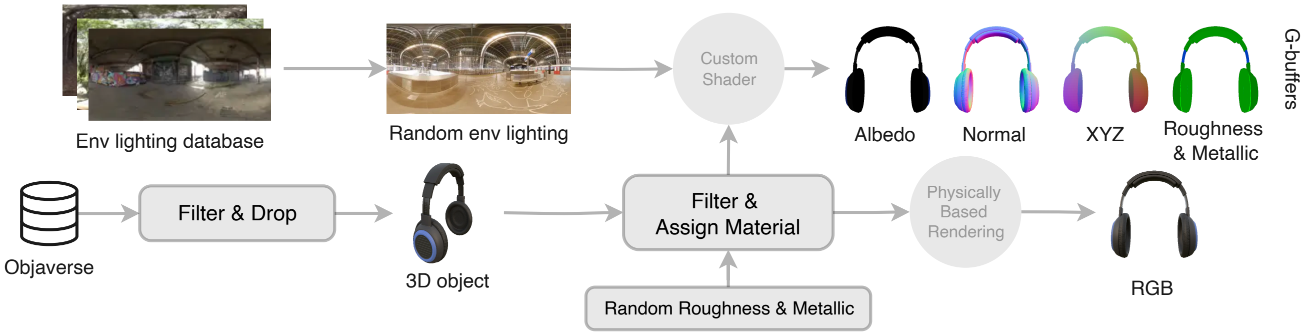

Figure 3. Overview of our dataset generation pipeline. We process 3D models from Objaverse through filtering and material enhancement. Each model is rendered using random environment lighting from the lighting database. Custom shaders extract G-buffer components (albedo, normal, XYZ, roughness, and metallic), while physically-based rendering generates the corresponding RGB images.

Table 1. Quantitative ablation study results.

| | PSNR ↑ | | | | SSIM ↑ | | | | LPIPS ↓ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Albedo | Normal | Roughness | Metallic | Albedo | Normal | Roughness | Metallic | Albedo | Normal | Roughness | Metallic |
| ***Full Model*** | 22.243 | 28.399 | 21.819 | 20.278 | 0.8946 | 0.9387 | 0.9110 | 0.9029 | 0.0780 | 0.0626 | 0.0974 | 0.1014 |
| *- w/o Single-Step* | 18.620 | 24.668 | 18.552 | 16.423 | 0.8644 | 0.9092 | 0.8851 | 0.8364 | 0.1170 | 0.1027 | 0.1181 | 0.1406 |
| *- w/o Image Space Loss* | 17.226 | 23.458 | 18.319 | 15.258 | 0.8071 | 0.8841 | 0.8310 | 0.7829 | 0.1494 | 0.1509 | 0.1584 | 0.1714 |
| *- w/o Lighting Response Loss* | 19.353 | 24.969 | 19.019 | 18.028 | 0.8541 | 0.9066 | 0.8713 | 0.8685 | 0.1271 | 0.1093 | 0.1260 | 0.1409 |

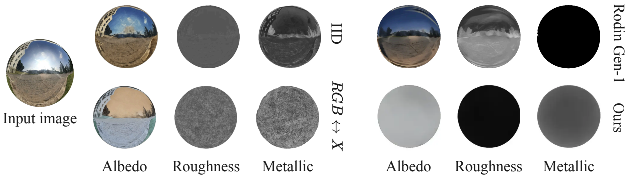

Figure 4. Comparison of the material estimation between our method and Rodin Gen-1. Because Rodin Gen-1 generates materials from text prompts, reference images, and 3D meshes, its decomposition results might be ambiguous.

**G-buffer Generation.** Since Blender does not natively support G-buffer rendering, We developed custom shaders to decompose various components (XYZ coordinates, normal, albedo, roughness, and metallic) as separate channels to get multi-view G-buffers. Each 3D model is rendered from six fixed views on a sphere under different environmental lighting conditions, and the final RGB images are rendered using Blender's cycles engine for physical accuracy.

Our pipeline produces a dataset of 512×512 resolution images for over 17k objects, each with corresponding G-buffers and RGB renders from multiple views. All objects are normalized to a unit box and randomly rotated during input view rendering to improve the training robustness. Note that the dataset can be further expanded to a larger scale by applying our method to more 3D models.

## 5. Experiments

In this section, we conduct extensive experiments to evaluate the performance of our method. We begin by detailing the experimental settings (Sec. 5.1), followed by the qualitative and quantitative results (Sec. 5.2). Finally, we conduct ablation studies (Sec. 5.3).

### 5.1. Experimental Settings

**Implementation details.** We initialize the G-buffer estimation network from the weights of Zero123++ [42] and train the network on our synthetic dataset for 10K steps with a batch size of 16 and learning rate of $1 \times 10^{-5}$. The training process can be conducted on 4×H100 GPUs for only 20 hours. For sparse-view 3D reconstruction, we use Nvdiffrast [25] for mesh optimization from G-buffers.

**Baselines.** To our knowledge, there are no existing open-source methods for material-aware single image to 3D generation tasks. Although some existing methods [65] can generate PBR textured meshes from images, they typically separate the generating of geometry and texture. These approaches are essentially texturing a 3D mesh with additional text input, which can lead to misaligned material properties with the input image, as shown in Fig. 4. Since no existing methods perform the same task as ours, we evaluate our methods from two aspects. **First**, we compare our reconstructed 3D models with *MeshFormer* [27], a state-of-the-art singe image to 3D generation method. **Second**, we compare our generated G-buffers with existing single image intrinsic decomposition methods, including *RGB↔X* [62], which generates albedo, normal, roughness, and metallic; and *IID* [24], generating albedo, roughness, and metallic. Since IID cannot estimate geometry properties, we use the normal maps generated by GeoWizard [12], a SOTA method on single-image normal estimation.

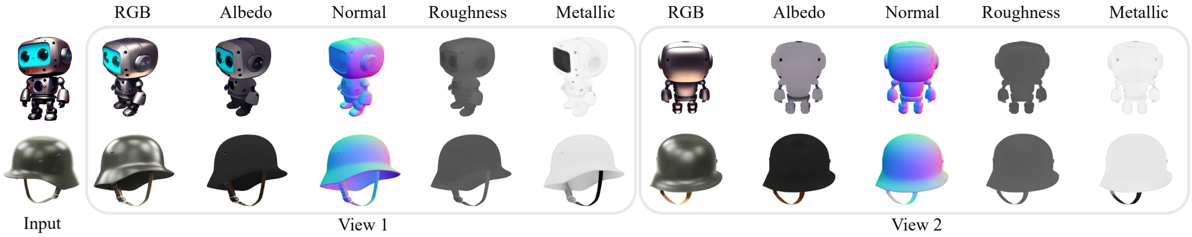

Figure 5. Multi-view RGB images and G-buffers generated by MAGE. The first line is an AI-created image, and the second line is a real captured image. We only show two novel views here, while MAGE generates six novel views at inference.

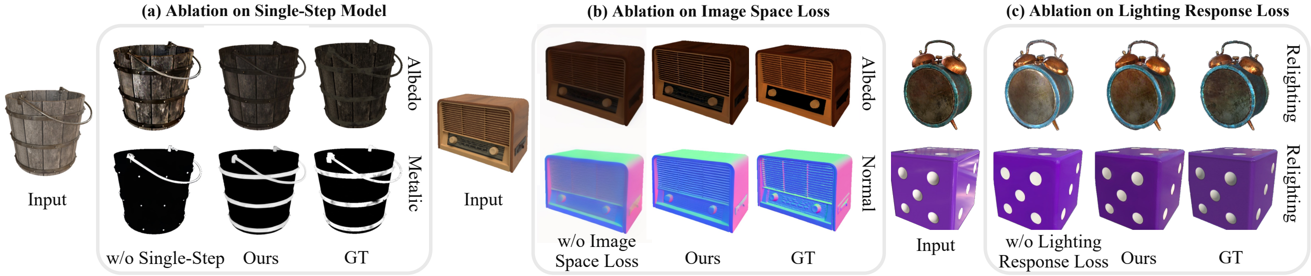
(a) Ablation on Single-Step Model
(b) Ablation on Image Space Loss
(c) Ablation on Lighting Response Loss
Albedo
Metalic
Input
w/o Single-Step
Ours
GT
Albedo
Normal
Input
w/o Image Space Loss
Ours
GT
Relighting
Relighting
Input
w/o Lighting Response Loss
Ours
GT

Figure 6. Ablation study results. We evaluate our proposed design choices by comparing our full model with three variant models: (a) a multi-step diffusion model for G-buffers generation, (b) a single-step model trained with latent space loss, and (c) a single-step model trained with image space loss but without the lighting response loss.

**Evaluation Dataset.** We evaluate our method using a subset of randomly selected Objaverse objects with built-in material maps. We render each object RGB image as input and corresponding G-buffers as ground truth from 20 randomly sampled viewpoints. To assess our performance on wild images, we also compile a test set of real captured images and AI-generated images from Freepik[1].

## 5.2. Experimental Results

**Visual Comparisons.** We show multi-view G-buffers generated by our method given a single wild input image in Fig. 5. We can see that our method effectively predicts consistent multi-view G-buffers for both AI-created images and real captured images. As shown in Fig. 1, these estimated multi-view G-buffers can be further used to reconstruct material-aware 3D models. Our method successfully disentangles lighting effects and object color to generate 3D models with accurate materials, enabling realistic rendering under different lighting conditions. In contrast, MeshFormer conflates lighting effects with the color of objects, resulting in unrealistic textured meshes.

For G-buffers estimation, we decompose the input image into G-buffers using different methods and present the results in Fig. 7. We also show the relighting results and the predicted G-buffers. As can be seen, RGB↔X fails to capture the correct geometry of the input image, resulting in inaccurate normal maps. In contrast, our method generates accurate normal maps comparable to GeoWizard's, which specifically focuses on geometry estimation. For material properties, RGB↔X fails to predict accurate albedo and distinguish materials of different object parts, while IID tends to bake lighting effects into roughness and metallic maps, *e.g.*, the shadow of the kettle. RGB↔X and IID ignore the interplay between G-buffer attributes and produce unsatisfactory results when relighted under new illuminations. In contrast, our method successfully estimates clean and accurate albedo and part-aware roughness and metallic properties, enabling realistic relighting.

**Quantitative Comparisons.** We use image reconstruction metrics (PSNR, SSIM, and LPIPS) to measure each predicted G-buffer attribute against the ground truth for quantitative comparisons. Results in Tab. 2 demonstrate that our method consistently surpasses all baseline methods across all domains and metrics. For normal estimation, our results are comparable to those of GeoWizard[12], a state-of-the-art method focusing solely on geometry prediction.

## 5.3. Ablation Study

**Ablation on Single-Step Model.** We also train a multi-step diffusion model on our proposed dataset to assess the effect of our deterministic single-step design. This model is trained with the v-prediction as in [42]. As shown in Fig. 6(a), our single-step model significantly outperforms the multi-step model in generating part-ware metallic and light-free albedo with the help of image space losses.

**Ablation on Image Space Loss.** To further evaluate the effectiveness of the image space loss for training our single-

[1] https://www.freepik.com/

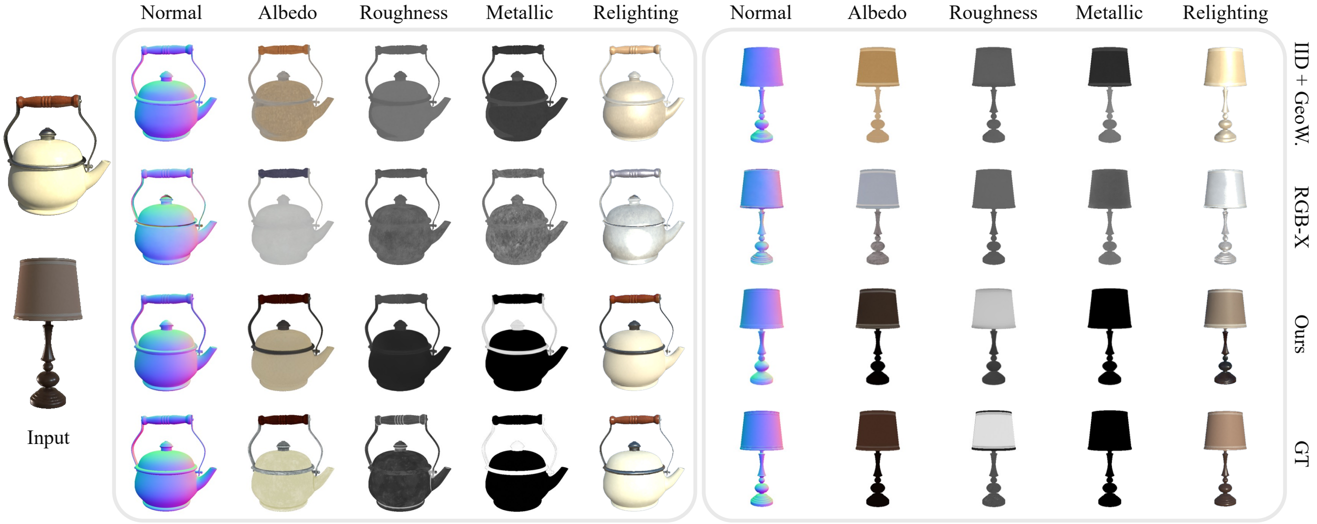

Figure 7. Visual comparison on single-view G-buffers estimation between our method, IID [24], and RGB↔X [62]. Since IID solely focuses on material estimation, we include the normal maps generated by GeoWizard [12] for a comprehensive comparison.

Table 2. Quantitative comparison with baselines on G-buffers estimation.

| | **IID [24] + GeoW. [12]** | | | | **RGB↔X [62]** | | | | **Ours** | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Albedo | Normal | Roughness | Metallic | Albedo | Normal | Roughness | Metallic | Albedo | Normal | Roughness | Metallic |
| **PSNR** ↑ | 18.193 | 26.917 | 19.335 | 16.264 | 15.487 | 20.489 | 19.848 | 15.749 | 22.243 | 28.399 | 21.819 | 20.278 |
| **SSIM** ↑ | 0.8520 | 0.9334 | 0.8828 | 0.8145 | 0.8349 | 0.8841 | 0.8831 | 0.8348 | 0.8946 | 0.9387 | 0.9110 | 0.9029 |
| **LPIPS** ↓ | 0.1145 | 0.0685 | 0.1123 | 0.1381 | 0.1339 | 0.1331 | 0.1129 | 0.1455 | 0.0780 | 0.0626 | 0.0974 | 0.1014 |

step model, we compare our results with those of another single-step model trained with latent space loss in Fig. 6(b). We observe that our model trained with image space loss produces results with more high-frequency details compared to the model without image space loss, demonstrating the effectiveness of image space loss in improving our single-step inference performance.

**Ablation on Lighting Response Loss.** In Fig. 6(c), we show the relighted input image using the estimated G-buffers of our whole model and our variant model trained without the proposed lighting response loss. Compared to the model without lighting response loss, our full model produces more realistic and consistent relighting results, indicating the effectiveness of the lighting response loss in enhancing physical consistency across G-buffer attributes.

## 6. Limitation and Future Work

While our method shows promising results, there are still limitations and opportunities for future work. The current approach relies on synthetic training data, where the distribution of the dataset may be limited. Our method can not handle highly complex materials, like anisotropic surfaces or subsurface scattering effects. The current framework primarily focuses on standard PBR materials defined by albedo, roughness, and metallic properties. Besides, our approach does not handle refractive materials such as glass, liquids, or transparent plastics. The current G-buffer representation and rendering model do not account for light refraction, which is crucial for accurately representing these materials. In addition, our current pipeline estimates G-buffers in 2D space, which is limited to a few sparse viewpoints. In the future, extending our framework to handle more diverse material types or dynamic scenes and jointly learning geometry and materials directly in native 3D space could be valuable directions for future research.

## 7. Conclusion

This paper presented MAGE, a novel approach for generating material-aware 3D models from a single input image. Our method leverages the concept of G-buffers from deferred rendering pipelines. It builds a deterministic single-step network architecture that efficiently predicts multi-view G-buffers while maintaining physical consistency through our proposed lighting response loss. We also developed a comprehensive synthetic dataset with diverse material properties to facilitate training. Experimental results demonstrate that our method successfully generates high-quality 3D models with physically accurate material properties, outperforming existing approaches and baselines in both qualitative and quantitative evaluations.

# References

[1] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22246–22256, 2023. 3

[2] Yongwei Chen, Rui Chen, Jiabao Lei, Yabin Zhang, and Kui Jia. Tango: Text-driven photorealistic and robust 3d stylization via lighting decomposition. *Advances in Neural Information Processing Systems*, 35:30923–30936, 2022. 3

[3] Yiwen Chen, Tong He, Di Huang, Weicai Ye, Sijin Chen, Jiaxiang Tang, Xin Chen, Zhongang Cai, Lei Yang, Gang Yu, Guosheng Lin, and Chi Zhang. Meshanything: Artist-created mesh generation with autoregressive transformers, 2024. 3

[4] Zilong Chen, Yikai Wang, Feng Wang, Zhengyi Wang, and Huaping Liu. V3d: Video diffusion models are effective 3d generators. *arXiv:2403.06738*, 2024. 3

[5] 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

[6] Gene Chou, Yuval Bahat, and Felix Heide. Diffusion-sdf: Conditional generative modeling of signed distance functions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2262–2272, 2023. 3

[7] R. L. Cook and K. E. Torrance. A reflectance model for computer graphics. *ACM Trans. Graph.*, 1(1):7–24, 1982. 5

[8] Michael Deering, Stephanie Winner, Bic Schediwy, Chris Duffy, and Neil Hunt. The triangle processor and normal vector shader: a vlsi system for high performance graphics. *SIGGRAPH Comput. Graph.*, 22(4):21–30, 1988. 2

[9] 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*, pages 13142–13153, 2023. 2, 5

[10] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, et al. Objaverse-xl: A universe of 10m+ 3d objects. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[11] Kangle Deng, Timothy Omernick, Alexander Weiss, Deva Ramanan, Jun-Yan Zhu, Tinghui Zhou, and Maneesh Agrawala. Flashtex: Fast relightable mesh texturing with lightcontrolnet. *arXiv:2402.13251*, 2024. 3

[12] Xiao Fu, Wei Yin, Mu Hu, Kaixuan Wang, Yuexin Ma, Ping Tan, Shaojie Shen, Dahua Lin, and Xiaoxiao Long. Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. In *European Conference on Computer Vision*, pages 241–258. Springer, 2025. 6, 7, 8

[13] Jian Gao, Chun Gu, Youtian Lin, Hao Zhu, Xun Cao, Li Zhang, and Yao Yao. Relightable 3d gaussian: Real-time point cloud relighting with brdf decomposition and ray tracing. *arXiv:2311.16043*, 2023. 3

[14] Gonzalo Martin Garcia, Karim Abou Zeid, Christian Schmidt, Daan de Geus, Alexander Hermans, and Bastian Leibe. Fine-tuning image-conditional diffusion models is easier than you think. *arXiv:2409.11355*, 2024. 3, 4

[15] Anchit Gupta, Wenhan Xiong, Yixin Nie, Ian Jones, and Barlas Oğuz. 3dgen: Triplane latent diffusion for textured mesh generation. *arXiv:2303.05371*, 2023. 3

[16] Lei Zhu Haoyuan Wang, Wenbo Hu and Rynson W.H. Lau. Inverse rendering of glossy objects via the neural plenoptic function and radiance fields. In *CVPR*, 2024. 3

[17] Jon Hasselgren, Nikolai Hofmann, and Jacob Munkberg. Shape, light, and material decomposition from images using monte carlo rendering and denoising. *Advances in Neural Information Processing Systems*, 35:22856–22869, 2022. 3

[18] Zexin He and Tengfei Wang. Openlrm: Open-source large reconstruction models. `https://github.com/3DTopia/OpenLRM`, 2023. 2

[19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3

[20] 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:2311.04400*, 2023. 2, 3

[21] Yingwenqi Jiang, Jiadong Tu, Yuan Liu, Xifeng Gao, Xiaoxiao Long, Wenping Wang, and Yuexin Ma. Gaussianshader: 3d gaussian splatting with shading functions for reflective surfaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5322–5332, 2024. 3

[22] Haian Jin, Isabella Liu, Peijia Xu, Xiaoshuai Zhang, Songfang Han, Sai Bi, Xiaowei Zhou, Zexiang Xu, and Hao Su. Tensoir: Tensorial inverse rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 165–174, 2023. 3

[23] Brian Karis and Epic Games. Real shading in unreal engine 4. *Proc. Physically Based Shading Theory Practice*, 4(3):1, 2013. 5

[24] Peter Kocsis, Vincent Sitzmann, and Matthias Nießner. Intrinsic image diffusion for indoor single-view material estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5198–5208, 2024. 3, 6, 8

[25] Samuli Laine, Janne Hellsten, Tero Karras, Yeongho Seol, Jaakko Lehtinen, and Timo Aila. Modular primitives for high-performance differentiable rendering. *ACM Transactions on Graphics*, 39(6), 2020. 6

[26] Zhihao Liang, Qi Zhang, Ying Feng, Ying Shan, and Kui Jia. Gs-ir: 3d gaussian splatting for inverse rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21644–21653, 2024. 3

[27] Minghua Liu, Chong Zeng, Xinyue Wei, Ruoxi Shi, Linghao Chen, Chao Xu, Mengqi Zhang, Zhaoning Wang, Xiaoshuai Zhang, Isabella Liu, Hongzhi Wu, and Hao Su. Meshformer: High-quality mesh generation with 3d-guided reconstruction model. *arXiv:2408.10198*, 2024. 1, 3, 6

[28] 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

[29] Shichen Liu, Tianye Li, Weikai Chen, and Hao Li. Soft rasterizer: A differentiable renderer for image-based 3d reasoning. *The IEEE International Conference on Computer Vision (ICCV)*, 2019. 5

[30] 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

[31] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. *arXiv:2310.15008*, 2023. 2, 3

[32] Shi Mao, Chenming Wu, Ran Yi, Zhelun Shen, Liangjun Zhang, and Wolfgang Heidrich. Generating material-aware 3d models from sparse views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1400–1409, 2024. 3

[33] Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. RealFusion: 360 reconstruction of any object from a single image. In *CVPR*, 2023. 2

[34] Jacob Munkberg, Jon Hasselgren, Tianchang Shen, Jun Gao, Wenzheng Chen, Alex Evans, Thomas Müller, and Sanja Fidler. Extracting Triangular 3D Models, Materials, and Lighting From Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8280–8290, 2022. 3, 5

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

[36] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. DreamFusion: Text-to-3D using 2D diffusion. In *International Conference on Learning Representations (ICLR)*, 2023. 2

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

[38] Elad Richardson, Gal Metzer, Yuval Alaluf, Raja Giryes, and Daniel Cohen-Or. Texture: Text-guided texturing of 3d shapes. In *ACM SIGGRAPH 2023 conference proceedings*, pages 1–11, 2023. 3

[39] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 3

[40] 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

[41] Shen Sang and Manmohan Chandraker. Single-shot neural relighting and svbrdf estimation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIX 16*, pages 85–101. Springer, 2020. 3

[42] 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:2310.15110*, 2023. 2, 3, 4, 6, 7

[43] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv:2308.16512*, 2023. 4

[44] Yahao Shi, Yanmin Wu, Chenming Wu, Xing Liu, Chen Zhao, Haocheng Feng, Jingtuo Liu, Liangjun Zhang, Jian Zhang, Bin Zhou, et al. Gir: 3d gaussian inverse rendering for relightable scene factorization. *arXiv:2312.05133*, 2023. 3

[45] Yawar Siddiqui, Antonio Alliegro, Alexey Artemov, Tatiana Tommasi, Daniele Sirigatti, Vladislav Rosov, Angela Dai, and Matthias Nießner. Meshgpt: Generating triangle meshes with decoder-only transformers. *arXiv:2311.15475*, 2023. 3

[46] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. *arXiv:2303.01469*, 2023. 4

[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. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22819–22829, 2023. 2

[48] 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:2402.05054*, 2024. 3

[49] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 3

[50] 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. *CVPR*, 2022. 3

[51] Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitry Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. *arXiv:2403.12008*, 2024. 3

[52] Haoyuan Wang, Xiaogang Xu, Ke Xu, and Rynson W.H. Lau. Lighting up nerf via unsupervised decomposition and enhancement. In *ICCV*, 2023. 3

[53] Peng Wang and Yichun Shi. Imagedream: Image-prompt multi-view diffusion for 3d generation. *arXiv:2312.02201*, 2023. 3, 4

[54] 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

[55] Zhenwei Wang, Tengfei Wang, Gerhard Hancke, Ziwei Liu, and Rynson WH Lau. Themestation: Generating theme-aware 3d assets from few exemplars. *SIGGRAPH*, 2024. 2

[56] Zhenwei Wang, Tengfei Wang, Zexin He, Gerhard Hancke, Ziwei Liu, and Rynson WH Lau. Phidias: A generative model for creating 3d content from text, image, and 3d conditions with reference-augmented diffusion. *arXiv:2409.11406*, 2024. 3

[57] Felix Wimbauer, Shangzhe Wu, and Christian Rupprecht. De-rendering 3d objects in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18490–18499, 2022. 3

[58] Kailu Wu, Fangfu Liu, Zhihan Cai, Runjie Yan, Hanyang Wang, Yating Hu, Yueqi Duan, and Kaisheng Ma. Unique3d: High-quality and efficient 3d mesh generation from a single image, 2024. 3

[59] 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:2404.07191*, 2024. 3

[60] Xudong Xu, Zhaoyang Lyu, Xingang Pan, and Bo Dai. Matlaber: Material-aware text-to-3d via latent brdf auto-encoder. *arXiv:2308.09278*, 2023. 3

[61] Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6613–6623, 2024. 4

[62] Zheng Zeng, Valentin Deschaintre, Iliyan Georgiev, Yannick Hold-Geoffroy, Yiwei Hu, Fujun Luan, Ling-Qi Yan, and Miloš Hašan. Rgb x: Image decomposition and synthesis using material-and lighting-aware diffusion models. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 6, 8

[63] Biao Zhang, Jiapeng Tang, Matthias Niessner, and Peter Wonka. 3dshape2vecset: A 3d shape representation for neural fields and generative diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4):1–16, 2023. 3

[64] Bowen Zhang, Yiji Cheng, Jiaolong Yang, Chunyu Wang, Feng Zhao, Yansong Tang, Dong Chen, and Baining Guo. Gaussiancube: Structuring gaussian splatting using optimal transport for 3d generative modeling. *arXiv:2403.19655*, 2024. 3

[65] 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, 2024. 3, 6

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

[67] Xiuming Zhang, Pratul P Srinivasan, Boyang Deng, Paul Debevec, William T Freeman, and Jonathan T Barron. Nerfactor: Neural factorization of shape and reflectance under an unknown illumination. *ACM Transactions on Graphics (ToG)*, 40(6):1–18, 2021. 3

[68] Yuqing Zhang, Yuan Liu, Zhiyu Xie, Lei Yang, Zhongyuan Liu, Mengzhou Yang, Runze Zhang, Qilong Kou, Cheng Lin, Wenping Wang, et al. Dreammat: High-quality pbr material generation with geometry-and light-aware diffusion models. *ACM Transactions on Graphics (TOG)*, 43(4):1–18, 2024. 3

[69] Xin-Yang Zheng, Hao Pan, Peng-Shuai Wang, Xin Tong, Yang Liu, and Heung-Yeung Shum. Locally attentional sdf diffusion for controllable 3d shape generation. *ACM Transactions on Graphics (ToG)*, 42(4):1–13, 2023. 3

[70] Linqi Zhou, Yilun Du, and Jiajun Wu. 3d shape generation and completion through point-voxel diffusion. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5826–5835, 2021. 3