# Generating Material-Aware 3D Models from Sparse Views

Shi Mao[1,2] Chenming Wu[2] Ran Yi[3] Zhelun Shen[2] Liangjun Zhang[2] Wolfgang Heidrich[1]

[1] KAUST [2] Baidu Research [3] Shanghai Jiao Tong University

wuchenming@baidu.com

## Abstract

*Image-to-3D diffusion models have significantly advanced 3D content generation. However, existing methods often struggle to disentangle material and illumination from coupled appearance, as they primarily focus on modeling geometry and appearance. This paper introduces a novel approach to generate material-aware 3D models from sparse-view images using generative models and efficient pre-integrated rendering. The output of our method is a relightable model that independently models geometry, material, and lighting, enabling downstream tasks to manipulate these components separately. To fully leverage information from limited sparse views, we propose a mixed supervision framework that simultaneously exploits view-consistency via captured views and diffusion prior via generating views. Additionally, a view selection mechanism is proposed to mitigate the degenerated diffusion prior. We adapt an efficient yet powerful pre-integrated rendering pipeline to factorize the scene into a differentiable environment illumination, a spatially varying material field, and an implicit SDF field. Our experiments on both real-world and synthetic datasets demonstrate the effectiveness of our approach in decomposing each component as well as manipulating the illumination. Source codes are available at https://github.com/Sheldonmao/MatSparse3D.*

## 1. Introduction

3D model generation is crucial for various industrial applications such as AR/VR, filming, and gaming. However, the traditional pipeline for 3D model generation is a laborious manual task that relies on artistic modeling skills and technical expertise. Automating the 3D designing process has the potential to significantly reduce production costs and enable faster and more diverse content creation. Inverse rendering, a technique that analyzes captured images to recover the object's geometry, material properties, and external illuminations, holds promise in assisting this process.

The recent advancements in differentiable rendering and neural representation have showcased their capabilities not only in generating novel views [16] but also in tackling inverse rendering challenges [17, 29]. However, these methods requires significant number of input views, preferably encompassing a 360° range, to effectively learn a consistent model that can disentangle the complex rendering process. Recent progress in the generative model, especially the 2D diffusion model, has demonstrated the viability of generating a 3D model by making use of prior knowledge from as few as a single image [12, 19]. However, these methods generate 3D models without disentangling the underlying rendering process and only model the geometry and the outgoing radiance.

In this work, we present a novel approach for generating material-aware 3D models from sparse view images, departing from the conventional practice of generating 3D models with fixed entangled radiance. Our method produces a relightable representation that separately models geometry, material, and lighting, enabling downstream tasks to manipulate these components separately. To achieve this, we propose a differentiable inverse rendering pipeline that leverages RGB and estimated depth information from captured images, along with a 3D prior derived from a 2D diffusion model conditioned on the input images. Given that the 2D diffusion model imposes supervision for the entire image as opposed to individual pixels, an efficient rendering pipeline becomes crucial. We opt for pre-integrated rendering as it effectively renders the outgoing radiance without limiting the illumination frequency. By representing lighting as pre-integrated mipmaps and using a split-sum approximation for rendering, we simultaneously learn all-frequency direct illumination and compact material properties. For optimizing smooth surfaces, we incorporate an implicit surface geometry representation into the rendering pipeline, which demonstrates superior convergence when compared to its discrete counterpart. Additionally, to address degenerated diffusion prior notable for large baseline image generation, we introduce a simple yet effective view selection mechanism that mitigates excessive noise from the diffusion model. In summary, the contributions of our work are as follows:

- A mixed supervision framework that simultaneously exploits RGB/depth information via captured views and diffusion prior via generating views for sparse view inverse rendering.

- An efficient yet powerful pre-integrated rendering pipeline with implicit surface representation for jointly optimizing lighting, material and geometry under the mixed supervision framework.
- A view selection mechanism to mitigate unwanted noise arising from a degenerated diffusion prior.

## 2. Related Work

### 2.1. Neural Inverse Rendering

Inverse rendering has garnered significant attention in computer vision and graphics due to its potential for extracting scene properties from observed images. However, this task is inherently under-constrained and computationally demanding, involving the inference of scene geometry, material properties, and lighting conditions. NeRFactor [28] improves upon existing NeRF models by integrating Multilayer Perceptrons (MLPs) that describe various surface attributes for inverse rendering, but it is constrained to using a low-resolution (16 × 32) environment map for lighting representation. In contrast, PhySG [27] employs up to 128 Spherical Gaussians (SG) to model illumination, extending the model representation ability. NeRD [1] further extends this approach to handle varying illumination conditions by using network to learn 24 Spherical Gaussians representing the illumination. The Neural-PIL approach [2] offers a neural pre-integrated lighting method as an alternative to Spherical Gaussians, enabling accurate estimation of high-frequency lighting details. However, the lighting for different roughness level lacks consistency as they are predicted independently by the neural network. NVDiffrec [17] address this inconsistency by introducing differentiable formulation that build lighting mipmaps of different roughness levels from the same high-resolution (1024 × 512) environment map. Its successor, NVDiffrecmc [8], introduces ray tracing and Monte Carlo integration for more realistic shading. However, both methods adopt discrete deep marching tetrahedra (DMTET) for geometry reconstruction, which leads to unstable training performance especially when input views are limited. Our method adopts efficient pre-integrated rendering techniques similar to [17]. However, we adapt it using neural implicit surface representation to ensure stable training that jointly learn high-quality geometry, material, and illumination under sparse view settings, similar to [15]. NeRO[13] employs a similar geometry representation and learning strategy but primarily focuses on recovering reflective objects. Although NeRO provide valuable insights into the intricate lighting effect including indirect illuminations, the introduction of more complex modeling poses additional challenges when only limited, sparse views are available. The recent progress in 3D Gaussian Splatting (3DGS) [11] has introduced novel representations and factorizations in inverse rendering, as demonstrated in works such as [6, 10, 21]. But obtaining good quality initial point cloud from sparse-view images remains difficult for these methods.

### 2.2. Diffusion Models for 3D Generation

The recent advancements in the 2D diffusion model have demonstrated its effectiveness as a prior for 3D generation. Dreamfusion [18] introduced the Score Distillation Sampling (SDS) loss, enabling text-to-3D generation using the 2D diffusion model as a prior. Zero123 [12] trained a diffusion model using a large-scale 3D model dataset conditioned on posed images, allowing for image-to-3D generation. They trained the model to generate images by considering a reference image and the relative camera view position as conditions, which inherently learned the geometry prior from the large-scale 3D dataset. Subsequent works have further improved upon these results. Prolific Dreamer [25] introduced the Variation Score Distillation (VSD) technique, extending the SDS loss with low-rank adaptation (LoRA) to generate diverse and high-quality objects. Magic123 [19] proposed a coarse-to-fine framework (NeuS to DMTET) that utilized both text conditions (2D prior) and image conditions (3D prior) to generate high-quality objects in a balanced manner. Make-it-3D [22] also presented a two-stage framework (NeRF to textured point cloud) for object generation based on the text-conditioned diffusion model. In recent developments, Zero123++[20] introduced a multiview diffusion model that generates multiple images simultaneously and incorporates attention mechanisms to enforce multi-view consistency. Similarly, Wonder3D[14] employs a similar approach by introducing cross-domain normal attention. However, these methods do not address the entangled environment lights, which limits their usefulness for downstream tasks. Fantasia3D [3] adopts the modeling technique from NVDiffrec [17] and employs a two-stage optimization process to generate geometry and appearance separately. By utilizing DMTET as the geometry representation and employing pre-integrated rendering in the pipeline, they make significant progress. However, they always set the metallic factor as 0 and use a fixed bright environment map, which restricts their ability in inverse rendering scenarios. On the other hand, MATLABER [26] presents a solution for generating material-aware 3D models from text prompts. In our work, we propose a method to generate material-aware 3D models using sparse view images. Comparing with text-hinted generation, our approach enables 3D digitization of real-world objects in an efficient way.

## 3. Method

We present a method to generate material-aware 3D models from sparse-view images, making use of generative mod-

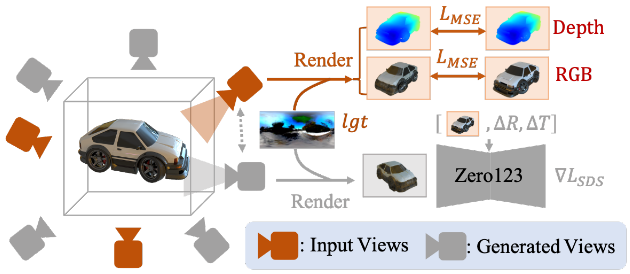

Figure 1. Our mixed supervision framework. Given the input view poses of captured images, RGB and depth images are rendered with trainable illumination, and MSE loss is applied to supervise these rendered images. Additionally, RGB images from generated view poses (with a relative camera rotation $\Delta R$, and translation $\Delta T$ to the referenced input views) are rendered similarly and supervised using SDS loss from the Zero123 diffusion model.

els and pre-integrated rendering. The output of our method is a relightable model that models geometry, material, and lighting separately. We introduce our pipeline in Section 3.1. Then, the modeling of surface geometry is discussed in detail in Section 3.2, followed by the modeling of material and lighting in Section 3.3. The diffusion prior that supervise the generated views is presented in Section 3.4.

### 3.1. Method Pipeline

In our proposed method, depicted in Figure 1, we introduce a framework that jointly optimizes geometry, material, and environment illumination based on limited sparse view inputs, leveraging the prior knowledge provided by the 2D diffusion model. Given the input views of sparsely captured images, we employ trainable illumination to render RGB images. The rendered images are supervised using the mean squared error (MSE) loss with respect to the captured images. To make the most of the limited information available from the captured views, we utilize a monocular depth estimation algorithm to obtain estimated depth values, which are then used to supervise the rendered depth. As the absolute depth values may have arbitrary scaling and bias, we address this issue by fitting a least-square solution, minimizing the mean squared error, before calculating the MSE loss.

In addition to the captured input views, we also leverage geometry priors from the diffusion model. Specifically, we utilize the Score Distillation Sampling (SDS) loss with the Zero123 [12] diffusion model. This loss takes captured images as reference images and incorporates the relative camera pose as conditions to generate novel view images. By imposing this SDS loss, we can effectively supervise the randomly generated views and align them with the geometry priors obtained from the diffusion model.

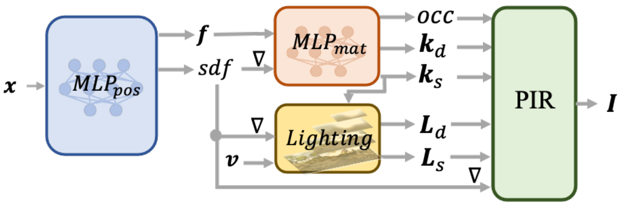

Figure 2. Models for geometry, material and light. Both $MLP_{mat}$ and $MLP_{pos}$ are multi-layer perceptions that models the corresponding mapping for SDF and material properties respectively. The lighting model is represented as differentiable environment map, from which both diffuse and specular irradiance $\boldsymbol{L}_d$ and $\boldsymbol{L}_s$ can be queried effectively using its pre-computed mipmaps. Finally, a pre-integrated rendering (PIR) module render the outgoing radiance $I$ from using the intermediate values.

### 3.2. Geometry Modeling

As shown in Figure 2, we use the implicit neural surface for geometry reconstruction. The implicit neural surface represents surfaces as the zero-level set of implicit SDF representation. Following NeuS[24], we use multi-layer perceptrons (MLPs) to model the mapping of both the SDF and material properties from spatial location $\boldsymbol{x} \in \mathbb{R}^3$. Specifically, the positional mapping $MLP_{pos} : \boldsymbol{x} \mapsto s$ that maps a 3D position $\boldsymbol{x} \in \mathbb{R}^3$ to a feature $\boldsymbol{f}' \in \mathbb{R}^{M+1}$, where the first element of $\boldsymbol{f}'$ have a geometrical meaning of SDF. The gradient of the SDF field is calculated as $\nabla_{sdf}$ and serves as the surface normal $\boldsymbol{n}$. The material mapping $MLP_{mat} : \{\boldsymbol{f}, \nabla_{sdf}\} \mapsto \boldsymbol{m}$ maps the rest channels of feature, i.e. $\boldsymbol{f} = \boldsymbol{f}'[1 : M + 1)$, and the gradient of SDF field $\nabla_{sdf}$ to material properties $\boldsymbol{m} \in \mathbb{R}^6$. Details on the material properties will be elaborated in Section 3.3. NeuS renders an image by accumulating the radiance along the rays cast by pixels, following the standard volume rendering scheme. The occlusion-aware unbiased weight function is calculated from the SDF value referring to Equations. 5 and 10 in NeuS[24].

### 3.3. Material and Lighting Modeling

To decompose the radiance field into geometry, material, and lighting components, we exploit an efficient *pre-integrated rendering* pipeline for differentiable rendering.

**Lighting Modeling.** The pre-integrated illumination, denoted as $\boldsymbol{L}(\boldsymbol{\omega}_r; r)$ is the integration of incident light radiance according to different surface roughness $r$. This lighting function relies on the reflection direction $\boldsymbol{\omega}_r$, which is calculated using the viewing direction $\boldsymbol{\omega}_v$ and the surface normal $\boldsymbol{n}$ as $\boldsymbol{\omega}_r = 2(\boldsymbol{\omega}_v \cdot \boldsymbol{n})\boldsymbol{n} - \boldsymbol{\omega}_v$. A trainable environment cubemap ($6 \times 3 \times H \times W$, with 6 faces and 3 color channel) is used to parameterize the lighting condition. From this cubemap, mipmaps are pre-integrated differentiably for various levels of roughness $r$ to enable fast interpolation when querying the diffuse irradiance $\boldsymbol{L}_d =$

$\boldsymbol{L}(\boldsymbol{n};1)$ and specular irradiance $\boldsymbol{L}_s = \boldsymbol{L}(\boldsymbol{\omega}_r;r)$ for specific surface point [17]. Importantly, the pre-integration and interpolation procedures remain differentiable, allowing for the learning of the trainable environment map.

**Material Modeling.** In our approach, we express the material representation as a combination of different components. This includes the diffuse color $\boldsymbol{k}_d \in \mathbb{R}^3$, the specular term $\boldsymbol{k}_s = r, m \in \mathbb{R}^2$, where $r$ represents roughness and $m$ represents metallic properties, and an additional occlusion term $occ \in \mathbb{R}$. Collectively, these material properties can be represented as $\boldsymbol{m} = \boldsymbol{k}_d, \boldsymbol{k}_s, occ$.

**Pre-integrated Rendering.** Following NVDiffrec [17], the rendering equation is a blend of diffuse and specular terms, and the outgoing radiance $\boldsymbol{I}(\boldsymbol{x}, \boldsymbol{\omega}_v)$ for RGB color from a surface point $\boldsymbol{x}$ with viewing direction $\boldsymbol{\omega}_v$ can be formulated as follows.

$$\boldsymbol{I}(\boldsymbol{x}, \boldsymbol{\omega}_v) = (1 - occ)\left[\boldsymbol{k}_d \boldsymbol{L}_d + \boldsymbol{L}_s\left(\boldsymbol{F}_0 F_s + F_b\right)\right] \qquad (1)$$

The reflectance of specular irradiance is calculated as the specular reflectance at normal incidence $\boldsymbol{F}_0$ modulated by its scale $F_s$ and bias $F_b$. Both the scaling and bias term are a function of the roughness $r$ and the cosine value between viewing direction $\boldsymbol{\omega}_v$ and surface normal $\boldsymbol{n}$, i.e. $F_s = F_s\left(\boldsymbol{\omega}_v \cdot \boldsymbol{n}, r\right)$ and $F_b = F_b\left(\boldsymbol{\omega}_v \cdot \boldsymbol{n}, r\right)$. We adopt the convention from UE4 that sets $\boldsymbol{F}_0$ as an interpolation from $0.04$ (non-metallic material's specular reflectance) to diffuse color $\boldsymbol{k}_d$ (metallic material's specular reflectance) using the material's metallic value $m$.

$$\boldsymbol{F}_0 = 0.04 \times (1 - m) + m\boldsymbol{k}_d. \qquad (2)$$

The final outgoing radiance is mediated by $(1 - occ)$ to account for the shadowing and inter-reflection effect that cannot modeled by the pre-integration rendering.

### 3.4. Diffusion Prior

To introduce prior knowledge pre-trained in a large-scale 3D model dataset, we embed Zero123[12] into our pipeline using SDS loss to supervise generated novel views. Zero123 is a 2D diffusion model that comes with a learned denoising function $\epsilon_\phi(\boldsymbol{z}_t; \widetilde{I}, \Delta R, \Delta T, t)$ that predicts the sampled noise $\epsilon$ given the noisy image latent $\boldsymbol{z}_t$, noise level $t$, and conditions $[\widetilde{I}, \Delta R, \Delta T]$, where $\widetilde{I}$ is the conditioning image and $[\Delta R, \Delta T]$ is the relative camera pose between a generated novel viewpoint to the reference image. The SDS loss[18] is formulated as:

$$\nabla_\theta L_{sds} = \mathbb{E}_{\epsilon,t}\left[w(t)\left(\epsilon_\phi(\boldsymbol{z}_t; \widetilde{I}, \Delta R, \Delta T, t) - \epsilon\right)\frac{\partial I}{\partial \boldsymbol{\theta}}\right] \qquad (3)$$

where $I = g(\theta)$ is a rendered view with parameter $\theta$ including all the models of geometry, material, and illumination, and $\boldsymbol{z}_t$ is the noisy latent by adding a random Gaussian noise of a time step $t$ to the latent of $I$.

### 3.5. Regularizations and camera selection

**Regularizations** In addition to RGB and depth values, the accumulated opaque density is also supervised by the object mask by MSE and binary cross entropy loss. To regularize the SDF field, we sampled nearby normal in 3D space and enforced the 3D smoothness of SDF field by regularizing the $L1$ distance between nearby normals. To regularize the environment cubemap, we apply a white environment prior and regularize it using its mean absolute error (MAE). Specifically, MAE is applied to the learned the base environment cubemap. For the material, we regularize the occlusion term to be small to limit the outlier of the rendering model. This is enforced by minimizing the mean value of the rendered occlusion term.

**Camera Selection** We observe that the images generated by Zero123 tend to degrade when the relative camera position $\Delta T$ is excessively large. This makes sense as neighboring views share more information and are are easy to generate, while the diffusion model is more likely to go wrong for a drastically different view. Therefore, during training, we select the closest captured view as a reference image for each generated view. We perform an ablation study in Section 4.7.

## 4. Experiments

### 4.1. Dataset

We generated a synthetic dataset using Blender, which consisted of 10 models sampled from NeRFactor [28], RefNeRF [23], and the large open-source 3D dataset Objaverse [5]. To create the training dataset, we randomly sampled a collection of 10 environment maps to render the scenes. For each scene, we randomly selected 5 positions in the upper hemisphere to render the sparse-view images. To evaluate the relighting performance, we relit each scene using 10 different environment maps from 10 evenly distributed circling views, resulting in 100 relighting images. The corresponding albedo images were also rendered as reference. Additionally, we synthesized images using positions from the testing set and illuminations from the training set to evaluate novel view synthesis. The reported results are the average over 10 scenes. Furthermore, we assessed the generalization capabilities of our method by evaluating it on the real-world DTU dataset [9], demonstrating how well our approach can extrapolate to real-world scenarios beyond the synthetic dataset.

### 4.2. Implementation Details

The rendered data in our experiments has a resolution of $800 \times 800$. During training, we downscale the scene to a resolution of $200 \times 200$ for the generated views. For each captured view, we sample 4096 rays. The training process alternates between the captured and generated views, taking

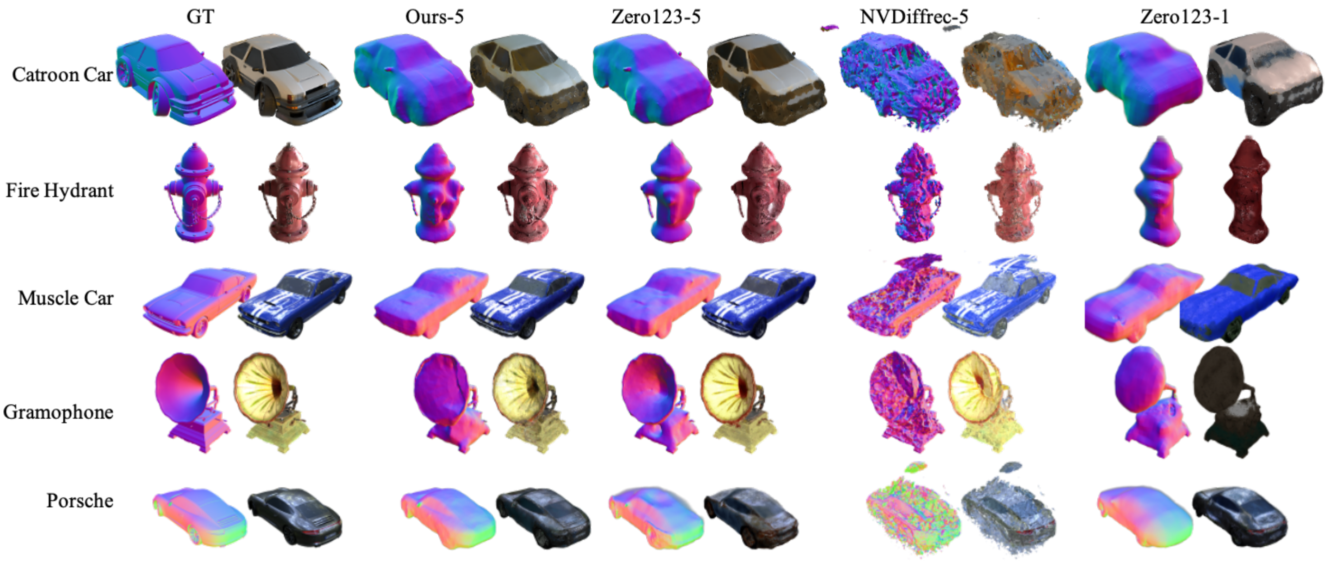

Figure 3. Qualitative comparison on novel view synthesis and geometry reconstruction for different methods. The suffix -n denotes the number of sparsely sampled input views.

| Method-Views | Novel View | | Geometry |
|---|---|---|---|
| | PSNR↑ | SSIM ↑ | Chamfer Dist. ↓ |
| Zero123-1 [12] | 15.14 | 0.817 | 7.31 |
| NVDiffrec-5 [17] | 15.92 | 0.817 | 3.55 |
| Zero123-5 | 20.93 | **0.870** | 2.33 |
| ours-5 | **20.94** | 0.866 | **2.09** |

Table 1. Novel view synthesis and geometry reconstruction. (In bold: best; underline: second best)

them one by one. To generate random views, we sample positions from the upper hemisphere with a random radius ranging from 1 to 1.5. We employ the Adam optimizer with a learning rate of 0.01 to train our model. The training process consists of 10,000 steps, and on a single V100 GPU, it typically takes less than 2 hours to train a model.

### 4.3. Baseline Methods

We compare our method with Zero123[12], a diffusion-based image-to-3D method that can generate 3D shapes from a single image. We use the updated Zero123-XL[4] as the diffusion prior and use the neural surface as the geometry backbone following the implementation of [7]. Using our framework, we extend it to a multi-view setting and term it Zero123-n, where n denotes the number of input views. We additionally evaluate the relighting and material reconstruction results by comparing them with NVDiffrec[17], an inverse rendering method adopting a similar pre-integrate rendering pipeline.

### 4.4. Novel View Synthesis and Geometry Reconstruction

We conducted a comparative analysis of different methods based on their performance in novel view synthesis and geometry reconstruction. To evaluate how accurately the reconstructed geometry aligns with the ground truth model, we first extracted the mesh from the implicit signed distance function (SDF) using the marching cubes algorithm. Subsequently, we sampled 50,000 points uniformly from the mesh and calculated the Chamfer distance between these points and the corresponding points on the ground truth mesh. For simplicity, we multiplied the vertex positions by 10 in the result. For novel view synthesis, PSNR and SSIM are reported. The results, presented in Table 1, indicate that under 5-view setting, our method achieves comparable performance to Zero123-5 on both novel view synthesis and geometry reconstruction, while significantly better than NVDiffrec-5. It is worth noting that we share the same training framework as Zero123-5 but differ in modeling approaches. This suggests that although we impose a more complex material-illumination factorization model, the overall geometry and visual quality are not degraded.

A more detailed qualitative comparison is illustrated in Figure 3. When solely relying on captured images, NVDiffrec-5 tends to reconstruct noisy surfaces with floating artifacts. In contrast, solely rely on diffusion prior without multi-view cross-reference, Zero123-1 generates smoother geometry, but it is often incorrect due to the limited information provided by a single image. This indicates the effectiveness of the mixed supervision framework that simultaneously exploits captured views and diffusion prior

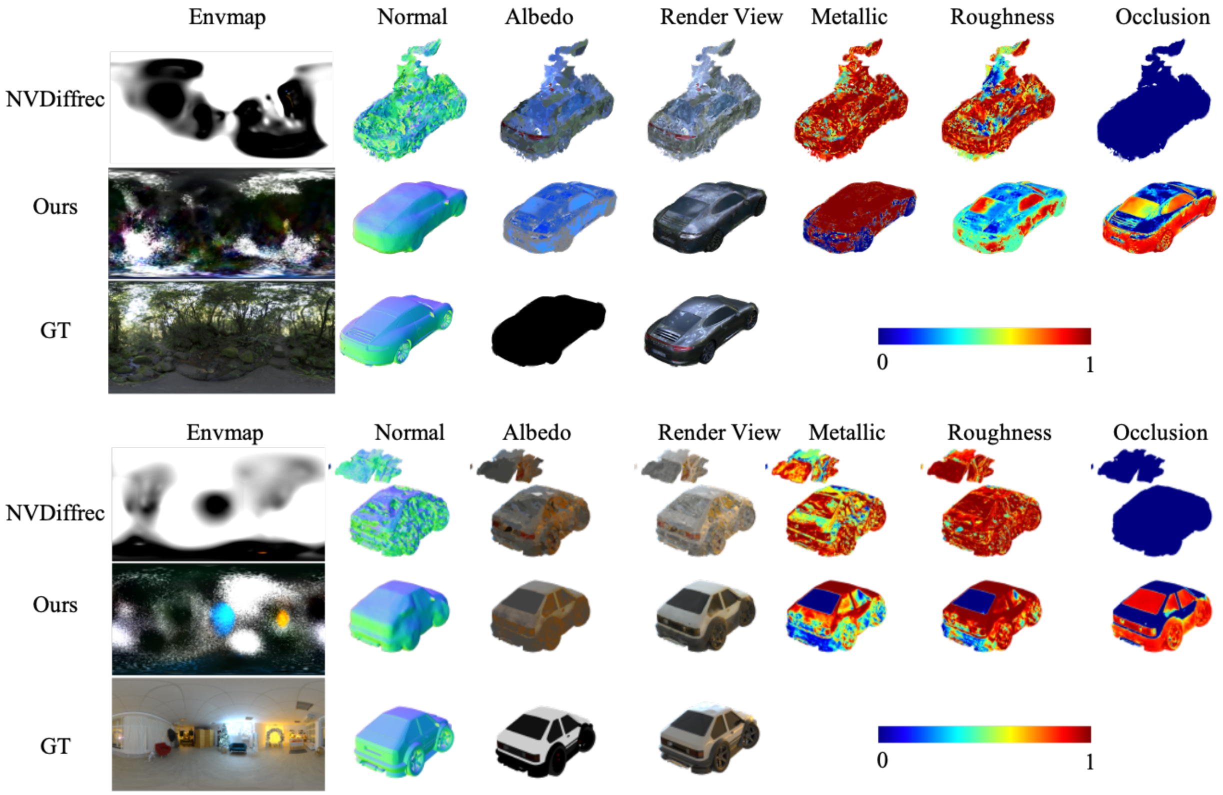

Figure 4. Material and illumination factorization on synthetic Porsche and Cartoon Car scene. Our method correctly learned the blue an yellow highlights in the environment map for cartoon car scene. Material, roughness and occlusion are visualized using jet colormap.

via generating views adopted by our method and zero123-5. Our model performs similarly to the NeuS-based Zero123-5 approach, but shows better performance in reconstructing the shiny convex-shape Porsche car, while it slightly underperforms in the concave-shape Gramophone. We attribute these variations to our illumination modeling, which assumes occlusion-free visibility to the environment map from every point on the surface. Although occlusion terms can partially compensate for this, the inherent modeling encourages a convex shape.

### 4.5. Material Reconstruction and Relighting

We conducted further evaluation to assess the reconstruction quality of material and illumination through relighting. The model was relighted using 10 testing environment maps for each testing camera pose and compared against ground truth renderings. To evaluate the material reconstruction, we rendered the albedo image by ray marching the $k_d$ values and compared it with the diffuse color rendered by Blender. Both comparisons were quantitatively measured using PSNR and SSIM metrics. It is important to note that the material and illumination can only be resolved up to a relative scale. To account for this, we adopted a convention that scales the predicted albedo image by a factor that matches the average ground-truth albedo. The results, presented in Table 2, demonstrate that our method outperforms NVDiffrec, which struggles to generate reasonable geometry under limited views. The material and illumination factorization is visualized in Figure 4. Our method reconstruct the environment maps better, especially correctly learning the blue an yellow highlights for Cartoon Car scene. Comparing with NVDiffrec, more consist material properties are reconstructed for different parts of the objects. Additionally, a qualitative assessment of the relighting results, depicted in Figure 5, reveals that our model aligns better with the ground truth rendering.

| Method-Views | Relight | | Albedo | |
|---|---|---|---|---|
| | PSNR↑ | SSIM ↑ | PSNR↑ | SSIM ↑ |
| NVDiffrec-5 [17] | 16.21 | 0.754 | 15.22 | 0.806 |
| ours-5 | **19.06** | **0.878** | **17.81** | **0.857** |
| ours-wo-depth-5 | 18.79 | 0.874 | 17.17 | 0.853 |
| ours-random-5 | 18.38 | 0.840 | 16.28 | 0.808 |

Table 2. Relighting and albedo reconstruction

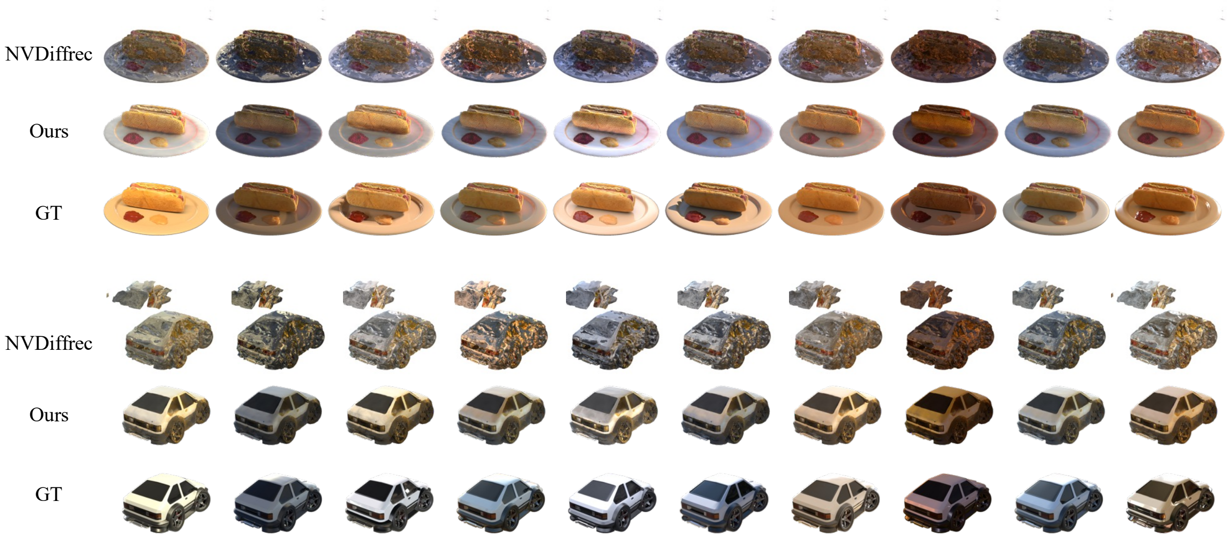

Figure 5. Relight Hot Dog and Cartoon Car model using 10 different environment maps of the test set. Both ground truth rendering and the results of our models and NVDiffrec are shown.

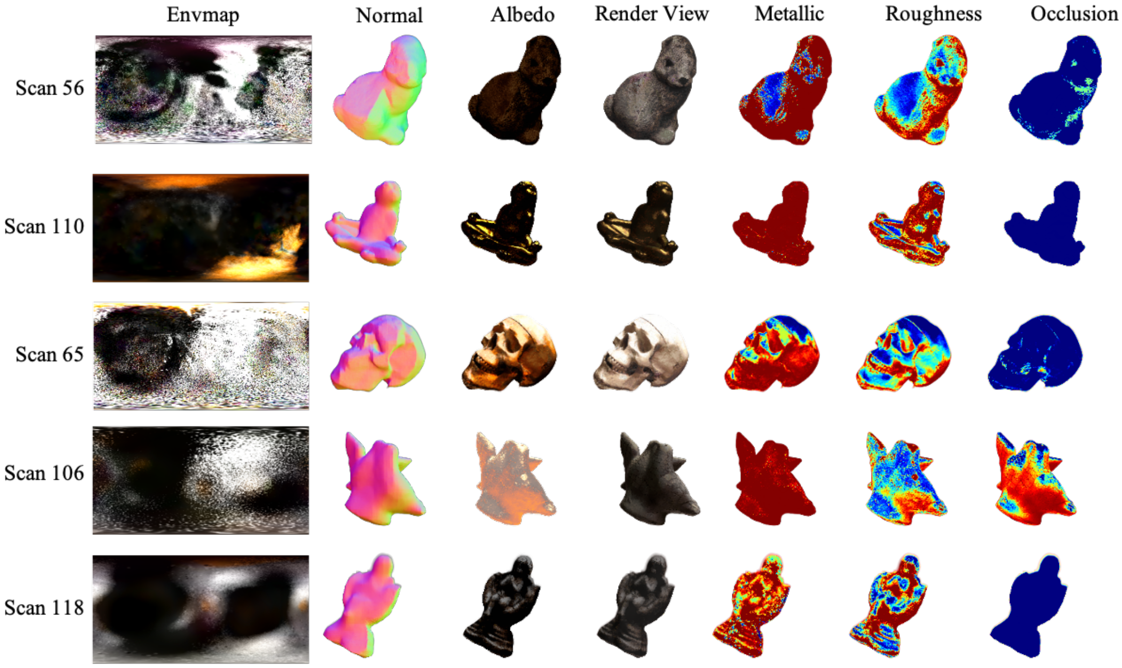

Figure 6. Material and illumination factorization on real-world DTU [9] scans using our method. For each real-world scan, sparsely sampled 5 views are taken as input. Material, roughness and occlusion are visualized using jet colormap.

### 4.6. Real-world Evaluation

To assess the generalization capability of our proposed method, we conducted a comprehensive evaluation using 5 scans from the real-world DTU dataset. For each scan, we sparsely sampled 5 views as input and tested the model on the remaining views. The evaluation results, shown in Figure 6, demonstrate that our method exhibits strong generalization capabilities to real-world scenes. It successfully reconstructs geometry and accurately captures material properties when provided with only 5 input views. In contrast, as depicted in Figure 7, NVDiffrec encounters difficulties in

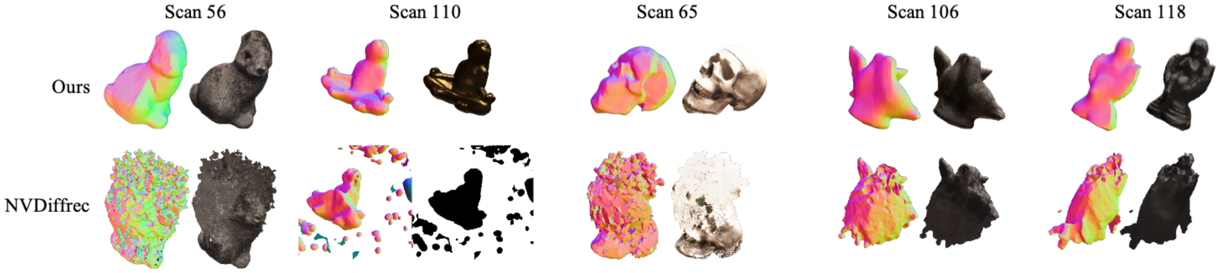

Figure 7. Novel views synthesis and geometry reconstruction comparison on real-world DTU [9] scans, For each real-world scan, sparsely sampled 5 views are taken as input.

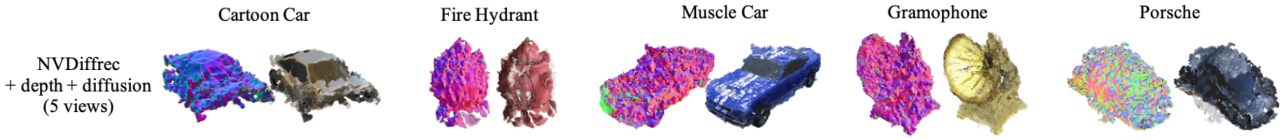

Figure 8. Novel view synthesis and geometry reconstruction achieved using NVDiffrec in combination with depth information and diffusion prior, with 5 views provided as input.

converging towards a satisfactory level of geometry reconstruction. Furthermore, our method achieves notable performance in novel view rendering, with an average PSNR of 21.78 and SSIM of 0.87. These results surpass the performance of NVDiffrec, which achieves a comparatively lower PSNR of 9.31 and SSIM of 0.67. These findings affirm the effectiveness and superior performance of our method in generalizing to real-world scenes.

## 4.7. Ablation Study

**Baseline Comparison.** To assess the effectiveness of utilizing the neural surface as a geometry representation, we conducted additional comparisons using DMTET as a baseline for geometry. In the case of NVDiffrec, we incorporated depth cues and generated views during training. To address the non-differentiability issue associated with rasterized $z/w$ values during rendering, we manually converted the vertex positions to camera space depth and rasterized it as a feature to ensure differentiability. The qualitative assessment of the novel view synthesis and geometry reconstructions, following the same protocols, are presented in Figure 8. These results highlight the challenges associated with optimizing the discrete DMTET representation and validates our decision to employ the neural surface as the geometry representation, as it offers superior optimization capabilities and avoids the degradation caused by noise introduced by alternative representations such as DMTET.

**Generated View Conditioning.** We performed the ablation study that randomly selected a reference frame instead of selecting the nearest captured images as a reference frame. This method is termed *ours-random-5* in Table 2. The results demonstrate a significant decrease in both relighting and albedo reconstruction performance. This indicates that the diffusion prior plays a crucial role in providing higher-quality information, particularly when the relative view difference is small.

**Depth Supervision.** To assess the necessity of supervising the scene using estimated depth, we conducted an ablation study where we removed the depth supervision. This variant is referred to as *ours-wo-depth-5* in Table 2. The results indicate a decrease in both relighting and albedo reconstruction performance, although the drop is relatively small. This suggests that the diffusion model already incorporates geometry priors that mitigate the requirement for explicit depth supervision.

## 5. Conclusion

This paper presents a novel approach to generate material-aware 3D models using sparse view images. Our proposed mixed supervision framework combines RGB and depth information from captured views and leverages the diffusion prior obtained through generating views for sparse view inverse rendering. To effectively represent each component, we utilize adapted pre-integrated rendering and implicit surface techniques within the framework. These techniques enable efficient supervision and reconstruction of geometry, material, and illumination. The experimental on both synthetic and real-world demonstrate the model's ability to reconstruct each component and highlight the effectiveness of the camera selection mechanism. While the current results are promising, there is a need for more detailed geometry reconstruction, as well as the accounting for more complex indirect illumination.

## References

[1] Mark Boss, Raphael Braun, Varun Jampani, Jonathan T. Barron, Ce Liu, and Hendrik P.A. Lensch. Nerd: Neural reflectance decomposition from image collections. In *ICCV*, 2021. 2

[2] Mark Boss, Varun Jampani, Raphael Braun, Ce Liu, Jonathan Barron, and Hendrik Lensch. Neural-pil: Neural pre-integrated lighting for reflectance decomposition. *Advances in Neural Information Processing Systems*, 34: 10691–10704, 2021. 2

[3] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. *arXiv preprint arXiv:2303.13873*, 2023. 2

[4] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, Eli VanderBilt, Aniruddha Kembhavi, Carl Vondrick, Georgia Gkioxari, Kiana Ehsani, Ludwig Schmidt, and Ali Farhadi. Objaverse-xl: A universe of 10m+ 3d objects. *arXiv preprint arXiv:2307.05663*, 2023. 5

[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*, pages 13142–13153, 2023. 4

[6] 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 preprint arXiv:2311.16043*, 2023. 2

[7] Yuan-Chen Guo, Ying-Tian Liu, Ruizhi Shao, Christian Laforte, Vikram Voleti, Guan Luo, Chia-Hao Chen, Zi-Xin Zou, Chen Wang, Yan-Pei Cao, and Song-Hai Zhang. threestudio: A unified framework for 3d content generation. https://github.com/threestudio-project/threestudio, 2023. 5

[8] Jon Hasselgren, Nikolai Hofmann, and Jacob Munkberg. Shape, Light, and Material Decomposition from Images using Monte Carlo Rendering and Denoising. *arXiv:2206.03380*, 2022. 2

[9] Rasmus Jensen, Anders Dahl, George Vogiatzis, Engil Tola, and Henrik Aanæs. Large scale multi-view stereopsis evaluation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 406–413. IEEE, 2014. 4, 7, 8

[10] 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. *arXiv preprint arXiv:2311.17977*, 2023. 2

[11] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM TOG*, 42(4), 2023. 2

[12] 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. 1, 2, 3, 4, 5

[13] 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. 2

[14] 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 preprint arXiv:2310.15008*, 2023. 2

[15] Shi Mao, Chenming Wu, Zhelun Shen, Yifan Wang, Dayan Wu, and Liangjun Zhang. Neus-pir: Learning relightable neural surface using pre-integrated rendering. *arXiv preprint arXiv:2306.07632*, 2023. 2

[16] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 1

[17] 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 *CVPR*, pages 8280–8290, 2022. 1, 2, 4, 5, 6

[18] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 2, 4

[19] 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 preprint arXiv:2306.17843*, 2023. 1, 2

[20] 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. 2

[21] 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 preprint arXiv:2312.05133*, 2023. 2

[22] Junshu Tang, Tengfei Wang, Bo Zhang, Ting Zhang, Ran Yi, Lizhuang Ma, and Dong Chen. Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior. *arXiv preprint arXiv:2303.14184*, 2023. 2

[23] 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 *CVPR*, pages 5481–5490. IEEE, 2022. 4

[24] 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. *Advances in Neural Information Processing Systems*, 34:27171–27183, 2021. 3

[25] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *arXiv preprint arXiv:2305.16213*, 2023. 2

[26] Xudong Xu, Zhaoyang Lyu, Xingang Pan, and Bo Dai. Matlaber: Material-aware text-to-3d via latent brdf auto-encoder. *arXiv preprint arXiv:2308.09278*, 2023. 2

[27] Kai Zhang, Fujun Luan, Qianqian Wang, Kavita Bala, and Noah Snavely. Physg: Inverse rendering with spherical gaussians for physics-based material editing and relighting. In *CVPR*, pages 5453–5462, 2021. 2

[28] 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. 2, 4

[29] Yuanqing Zhang, Jiaming Sun, Xingyi He, Huan Fu, Rongfei Jia, and Xiaowei Zhou. Modeling indirect illumination for inverse rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18643–18652, 2022. 1