Inverse Rendering of Glossy Objects via the Neural Plenoptic Function and Radiance Fields

Haoyuan Wang[1], Wenbo Hu[2†], Lei Zhu[1], Rynson W.H. Lau[1†]
[1]City University of Hong Kong [2]Tencent AI Lab
† Joint corresponding authors

Input image NeRO Ours NeRO Ours

Figure 1. Inverse rendering results of the cutting-edge method, NeRO [10], and ours from calibrated multi-view images of a glossy object. Geometries are shown as a rendered mesh in the second and third images, and materials (metalness & roughness) are shown as a color map in the fourth and fifth images. We can see that our results not only have smoother and more accurate geometry but also present a more reasonable material (since the material of this object should be uniform).

Abstract

Inverse rendering aims at recovering both geometry and materials of objects. It provides a more compatible reconstruction for conventional rendering engines, compared with the neural radiance fields (NeRFs). On the other hand, existing NeRF-based inverse rendering methods cannot handle glossy objects with local light interactions well, as they typically oversimplify the illumination as a 2D environmental map, which assumes infinite lights only. Observing the superiority of NeRFs in recovering radiance fields, we propose a novel 5D Neural Plenoptic Function (NeP) based on NeRFs and ray tracing, such that more accurate lighting-object interactions can be formulated via the rendering equation. We also design a material-aware cone sampling strategy to efficiently integrate lights inside the BRDF lobes with the help of pre-filtered radiance fields. Our method has two stages: the geometry of the target object and the pre-filtered environmental radiance fields are reconstructed in the first stage, and materials of the target object are estimated in the second stage with the proposed NeP and material-aware cone sampling strategy. Extensive experiments on the proposed real-world and synthetic datasets demonstrate that our method can reconstruct high-fidelity geometry/materials of challenging glossy objects with complex lighting interactions from nearby objects. Project webpage: https://whyy.site/paper/nep

1. Introduction

Although Neural Radiance Fields (NeRFs) [1–5, 8, 16–18, 21, 23, 27, 32] have achieved remarkable progress in photo-realistic reconstruction, it is still a challenge to integrate NeRFs into conventional rendering engines since NeRFs represent the object and illumination in an entangled manner. Disentangling the representation into geometry, materials, and environmental lighting, *i.e.* inverse rendering, is crucial for the applicability in game production and extended reality.

Recent works have explored geometry reconstruction [9, 11, 20, 26, 28, 30, 31] and further extended to the materials estimation [7, 10, 19, 23, 33], *e.g.*, albedo, roughness, and metalness. However, they typically represent the illumination as 2D environmental maps [7, 10, 19], which oversimplifies the complicated real-world lighting distribution to *infinite lights* only. In many practical scenarios where the target object is surrounded by other objects, a considerable amount of light actually comes from the radiance of those nearby objects. Neglecting these common scenarios results in inferior reconstruction of both geometry and materials, especially for glossy objects, such as the improper results of NeRO [10] in Fig. 1.

In this paper, we propose a *Neural Plenoptic Function (NeP)* to represent the global illumination as a 5D function, $f_p(\mathbf{x}, \mathbf{d})$, which describes the color of each light ob-

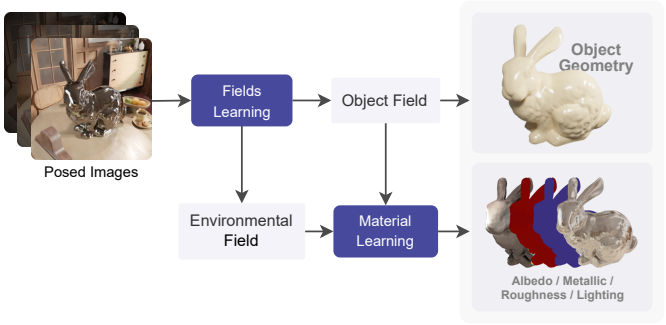

Figure 2. The pipeline of the proposed method. Our method has two stages: the *fields learning stage* for object geometry reconstruction and the neural radiance fields optimization, and the *material learning stage* using ray tracing.

served at position $\mathbf{x}$ with direction $\mathbf{d}$, in line with the definition of traditional plenoptic function [15]. Observing the superiority of NeRFs in recovering radiance fields from multi-view images, we construct the NeP from neural radiance fields based on a ray tracing procedure. However, directly doing so is computationally intensive, because rendering a ray's color from NeRF via ray marching is expensive, and ray tracing requires sampling a large number of rays inside the BRDF lobe to approximate the integration in the rendering equation. Thus, instead of sampling the lobe with rays, we propose an efficient material-aware cone-sampling strategy, where the cone's angle is derived from the predicted roughness. The color of lights inside a cone can be directly rendered from pre-filtered radiance fields, thanks to the anti-aliasing techniques in Mip-NeRF [1].

Overall, our method is divided into two stages: geometry reconstruction and material estimation of the target object. In the first stage, our model consists of an object field with a decoupled color representation, *i.e*. albedo and the color modulated by the lighting, and a pre-filtered environmental field for capturing scene radiance. To promote high-quality geometry reconstruction for glossy objects, we design a dynamic weighting loss mechanism from the decoupled colors to reduce the impact of highly uncertain reflective regions while amplifying the significance of diffuse areas. In the second stage, we adopt the Physically-Based Rendering (PBR) to estimate high-fidelity materials with our proposed NeP and material-aware cone-sampling strategy, based on the extracted triangular mesh of the target object and pre-trained environmental fields in the previous stage. Our two-stage method can faithfully reconstruct both the geometry and material properties of the target object, as shown in Fig. 1, solely from calibrated multi-view images. And importantly, the results can be seamlessly integrated into conventional rendering engines for relighting, as our method can produce compatible triangle meshes with physically-based materials.

To evaluate our method, we compiled two challenging glossy object datasets, from the rendering engine and real-world captures, respectively. Extensive experiments both quantitatively and qualitatively demonstrate that our method is robust, adaptable, and capable of handling diverse challenging illuminations. We also demonstrate the possible applications of our reconstructions, *e.g*., relighting, confirming the compatibility of our results with conventional rendering engines. Our contributions are summarized as follows:

- We design a simple yet effective dynamic weighting loss mechanism from the color decomposition for geometry reconstruction of challenging glossy objects.
- We propose a novel neural plenoptic function (NeP) to represent the global illumination and a material-aware cone-sampling method to effectively integrate NeP over BRDF lobes for high-fidelity material estimation.
- We constructed benchmarks (including both synthetic and real-world data) for the inverse rendering of challenging glossy objects with complex lighting interactions, and conducted extensive experiments to demonstrate the effectiveness of our method.

2. Related Work

Neural Radiance Fields. The introduction of Neural Radiance Fields (NeRF) [16] has marked a significant milestone in the field of 3D scene reconstruction, offering a new perspective on synthesizing novel views of complex scenes with details and realism. NeRF utilizes a fully connected neural network to model the volumetric scene radiance function. Building upon the foundation of NeRF, numerous works have sought to enhance its capabilities, addressing limitations and expanding its application range. Some methods [1, 2, 8] utilize enhanced cone sampling strategies rather than ray sampling, to address the aliasing problem and improve the captured details. Some methods [4, 5, 18] focus on optimizing NeRF for faster training and inference times. Other methods [13, 17, 25] improve NeRF for more robust training under degraded imaging conditions. These advancements collectively contribute to the evolution of NeRF, and solidify its position as a pivotal tool for 3D scene representation.

While NeRF models have demonstrated remarkable capabilities in synthesizing novel views, a key limitation lies in their inability to directly export reconstructed objects to rendering engines. A few NeRF-based inverse rendering approaches have been proposed to address this limitation.

NeRF for Inverse Rendering. Leveraging the potential of NeRF for inverse rendering tasks has garnered substantial attention, aiming to retrieve the intrinsic properties, especially physically-based properties of objects or scenes from 2D images with camera poses. Specifically, inverse rendering of a target object in NeRF aims to reconstruct both the geometry structure and the material information of the object. In this realm, a number of methods have been pro-

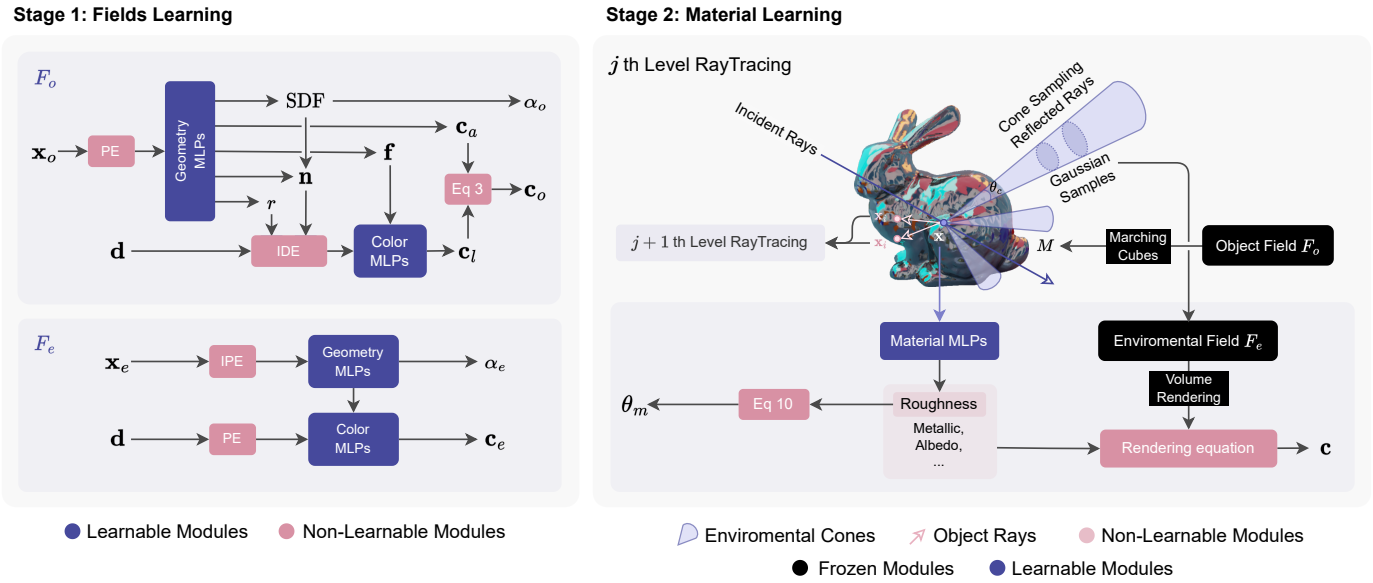

Figure 3. The detailed structure of our proposed method. Fields learning stage consists of an SDF-based object field and a Mip-NeRF as the environmental field. Based on them, we construct our neural plenoptic function via ray tracing and material-aware cone sampling method to represent the global illumination.

posed to unravel the complex interplay among geometry, material properties, and lighting. High-quality inverse rendering results on real-world data are dependent on realistic rendering process simulation, which can be modeled by the framework of the rendering equation, to help decompose the physically based properties and reduce ambiguity.

Previous methods mimic the rendering equation in different manners. [23, 33] represent the early attempts in this direction, utilizing NeRF to infer the surface normal, reflectance, and coarse lighting conditions simultaneously. Subsequently, some methods [7, 19, 29] try to improve geometry representation and propose a more accurate material estimating pipeline using Image-Based Lighting, facilitating a more robust and accurate inverse rendering process. They further improve the reconstruction quality by approximating the rendering equation with Monte Carlo sampling. Recently, some methods [10] took a step further by independently estimating the geometry and material properties of the object, and incorporating occlusion-aware constraints into the NeuS-based inverse rendering framework, ensuring that the reconstructed geometry has fine details and material properties adhere to real-world physics.

However, the existing methods either mainly utilize a 2D environmental map for lighting representation in the rendering equation, implicitly assuming that alllights over the scene come from infinity, or models direct lighting only [34]. These assumptions often lead to less realistic renderings, especially in scenarios where light sources or other objects are in closer proximity to the target object. In contrast, our approach utilizes neural plenoptic function based on neural radiance fields for more realistic shape and material learning, overcoming this fundamental limitation.

3. Method

Our approach targets the inverse rendering of objects from calibrated multi-view images. As our method is based on Neural Radiance Fields (NeRFs) and physically-based rendering (PBR), we start by briefly revisiting the relevant concepts in Sec. 3.1. We then introduce our two-stage pipeline, which involves a fields learning stage (Sec. 3.2) for geometry reconstruction and environmental lighting learning, and a material learning stage (Sec. 3.3) for material estimation, as illustrated in Fig. 2.

3.1. Preliminaries

Neural Radiance Fields (NeRF) models the scene as a continuous function that maps a 5D vector (spatial location $\mathbf{x}$ and viewing direction $\mathbf{d}$) to a color $\mathbf{c} = f_c(\mathbf{x}, \mathbf{d})$ and a volume density $\sigma = f_d(\mathbf{x})$, where f_c and f_d are MLPs for predicting color and density, respectively. While training, NeRF (which is parameterized as Θ_F) casts camera rays for each pixel, and samples points or Gaussian samples [1] along the ray. The color of the ray is computed as:

$$L_{\text{NeRF}}(\Theta_F, \mathbf{r}) = \int_{t_n}^{t_f} T(t) f_d(\mathbf{r}(t)) f_c(\mathbf{r}(t), \mathbf{d})\, dt \approx \sum_{i=1}^{n} w_i \mathbf{c}_i, \tag{1}$$

where $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$ is the parametric representation of the camera ray, $T(t)$ is the accumulated transmittance along the ray, and w_i is the weights for volume rendering. In practice, $T(t)$ can be defined as $T_i(t) = \prod_{j=1}^{i-1}(1 - \alpha_j)$. Instead

of the density, NeuS [26] predicts Signed Distance Field (SDF) via SDF $= f_{\text{SDF}}(\mathbf{x})$, where the surface of the object is modeled by the zero-level set of SDF. We represent the high-quality target object surfaces and environmental radiance based on both NeuS and NeRF.

Rendering Equation aims to simulate the interaction of light and surfaces in a way that adheres to physical laws. Rendering equation, which is an integral equation describing the equilibrium of light in a scene, is the core of PBR. It is given by:

$$L(\mathbf{x}, \mathbf{d}) = \int_{\Omega} f_r(\mathbf{x}, \mathbf{d}_i, \mathbf{d}) L_i(\mathbf{x}, \mathbf{d}_i)(n \cdot \mathbf{d}_i)\, d\omega, \quad (2)$$

where $L(\mathbf{x}, \mathbf{d})$ is the outgoing radiance from point $\mathbf{x}$ in the view direction $\mathbf{d}$, $L_i(\mathbf{x}, \mathbf{d}_i)$ is the incoming radiance from direction $\mathbf{d}_i$, f_r is the BRDF (Bidirectional Reflectance Distribution Function), n is the surface normal at point $\mathbf{x}$, and $d\omega$ represents an infinitesimal solid angle.

Based on NeRF and PBR techniques, our approach improves the neural inverse rendering of glossy objects by innovatively applying neural radiance fields and neural plenoptic function (NeP).

3.2. Fields Learning

Since simultaneously modeling geometry, lighting, and materials would lead to ambiguity, we construct a two-stage pipeline to optimize the geometry and materials separately. Our primary objective of the first stage is to simultaneously reconstruct the precise geometry of the target object and the environmental radiance field. This is the foundation for the subsequent material learning stage, which requires accurate light-surface intersection and surface normal.

Recent works [7, 10] have explored geometry reconstruction for glossy objects by incorporating physical priors like image-based rendering and split-sum approximations. However, they oversimplify the illumination as a 2D environmental map, which would cause suboptimal geometry, as shown in Sec. 4. Although NeRO [10] extends the environmental map to a directional function defined on the sphere, it may not capture the depth information of the scene, which struggles to reconstruct high-fidelity geometry for high-detailed objects in some cases. To this end, we first propose to decouple the final color $\mathbf{c}_o$ into albedo color $\mathbf{c}_a$ and color modulated by the lighting $\mathbf{c}_l$, as:

$$\mathbf{c}_o = \mathbf{c}_a \circ \mathbf{c}_l, \quad (3)$$

(in line with [23, 25]). Based on the decomposed colors, we employ a dynamic weighting loss mechanism, inspired by [6, 14], to strategically reduce the impact of highly uncertain reflective regions while amplifying the significance of diffuse areas when computing the photometric loss. This mechanism ensures high-quality geometry reconstruction of glossy objects. Although it may distort the learned reflective color, it is inconsequential as we will estimate more accurate decomposed physical materials in the second stage. Unlike Ref-NeuS [6], which determines the loss weights for rays by correlating pixels across views, we propose a simplier yet potent strategy, *i.e.*, leveraging the discrepancy between albedo and final color to derive the weights:

$$w_s = \min\left(\frac{1}{(\mathbf{c}_o - \mathbf{c}_a)^2 + \epsilon}, u\right), \quad (4)$$

where ϵ denotes a small value. u is a hyperparameter (set to 1.5 by default) to cap the upper bound of the weights. The intuition of our strategy lies is that the discrepancy between albedo and final color is higher in reflective regions.

The field architectures in the first stage are illustrated in the left part of Fig. 3. We represent the target object in a NeuS-based field F_o and employ a neural radiance field F_e to model the background environment. During the volume rendering step, we divide samples along a ray into two sets with the border of the target object: object samples $\mathbf{x}_o$ and environmental samples $\mathbf{x}_e$. Object samples $\mathbf{x}_o$ are first featurized by the position encoding (PE) [16] and then fed into the Geometry MLPs to predict the SDF value, which is further mapped to opacity α_o as in NeuS [26], albedo $\mathbf{c}_a$, roughness r, and a feature vector $\mathbf{f}$. Next, we encode the view direction $\mathbf{d}$, normal vector $\mathbf{n}$ (derived from the SDF), and the roughness r by Integrated Directional Encoding (IDE) [23]. The IDE features and the feature vector $\mathbf{f}$ are then fed into the Color MLPs to predict the colors modulated by the lighting $\mathbf{c}_l$. The final colors $\mathbf{c}_o$ for samples $\mathbf{x}_o$ are produced by Eq. 3 from the predicted $\mathbf{c}_a$ and $\mathbf{c}_l$. On the other hand, environmental samples $\mathbf{x}_e$ are mapped to opacity α_e and color $\mathbf{c}_e$ by a Mip-NeRF [1] model. Finally, the opacities (α_o, α_e) and colors ($\mathbf{c}_o$, $\mathbf{c}_e$) of object and environmental samples are rendered together into pixel color $\hat{\mathbf{c}_p}$ by volume rendering as Eq. 1. The whole fields are trained jointly by a weighted photometric loss:

$$L_c = w_l \cdot |\mathbf{c}_p - \hat{\mathbf{c}}_p|, \quad (5)$$

where $\hat{\mathbf{c}}_p$ is the GT pixel color, and w_l is the pixel-wise loss weight accumulated from w_s along the rays via volume rendering. Besides the photometric loss, we also utilize a loss function to constrain the curvature of normal vectors. Refer to the Supplemental for more details.

3.3. Material Learning

After obtaining the precise geometry of the target object, our goal for the second stage is to estimate the physically-based materials of the object. To achieve this, we adopt a more comprehensive rendering process, *i.e.*, ray tracing, to evaluate the rendering equation in Eq. 2. We can employ the marching cubes algorithm [12] to extract the triangle mesh

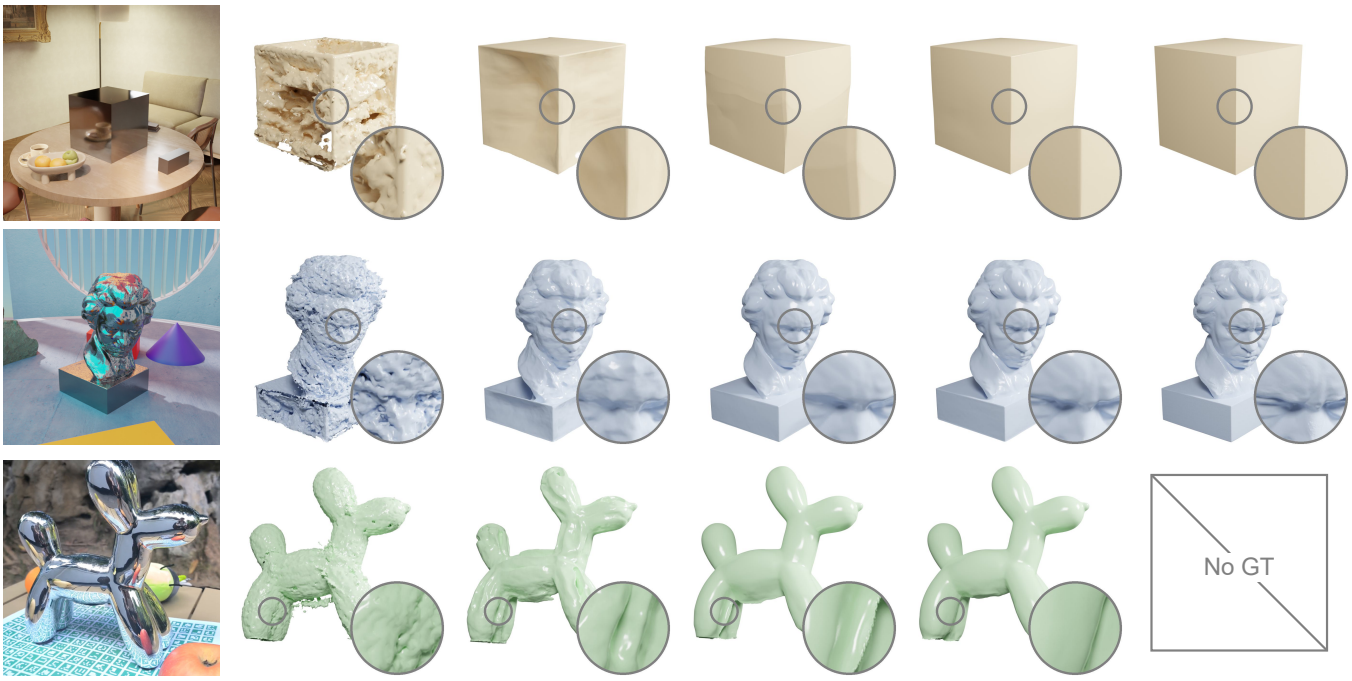

Figure 4. Comparison of the geometry reconstruction among cutting-edge methods and ours. For each method, we utilize marching cubes to extract the triangle meshes for comparison.

Figure 5. Samples from the proposed dataset. The first two examples are synthetic and the last two samples are real-world captured.

of the object by efficiently determining the ray-surface intersection $\mathbf{x}$ and surface normal $\mathbf{n}$. Thus, the key to estimating the materials is to faithfully represent the global illumination L_i in the rendering equation.

Neural Plenoptic Function. Instead of simplifying the representation of the lighting as a 2D environment map, we propose a 5D neural plenoptic function (NeP), symbolized by $f_p(\mathbf{x}, \mathbf{d})$, to represent the color of the light observed at the spatial location $\mathbf{x}$ with the direction $\mathbf{d}$, *i.e.* $L_i = f_p(\mathbf{x}, \mathbf{d}_i)$. However, it is challenging to directly learn NeP due to the high dimensionality. Because of the superiority of NeRF in representing the radiance field, we propose to construct NeP from the pre-trained environmental radiance fields F_e in the first stage. Specifically, we start by modeling the intersection between the incoming light $\mathbf{r}_i(t) = \mathbf{x} + t\mathbf{d}_i$ and the object mesh M as a function $I(\mathbf{r}_i, M)$ to determine the point of contact $\mathbf{x}_i$. If there are no intersections, which means that the incoming light is a direct light from the environment, we utilize Eq. 1 with the pretrained F_e to define the lighting color. We also utilize a simple MLP to learn the residual value of L_{NeRF}, permitting the model to learn the distant lighting that is not captured by training images. Conversely, if the intersection point $\mathbf{x}_i$ exists, we employ a ray-tracing approach based on Eq. 2 to obtain the color of the incoming light as it is an indirect light in this situation. The plenoptic function is defined as:

$$f_p(\mathbf{x}, \mathbf{d}_i) = \begin{cases} L(\mathbf{x}_i, \mathbf{d}_i), & \text{if } \exists\, \mathbf{x}_i, \quad \text{(6a)} \\ L_{\text{NeRF}}\left(\Theta_{F_e}, \mathbf{r}_i(t)\right), & \text{otherwise}, \quad \text{(6b)} \end{cases}$$

where L is a discretized rendering equation discussed next.

Ray Tracing. The discretized rendering equation L is derived via the ray tracing algorithm as:

$$L(\mathbf{x}, \mathbf{d}) = \sum_{k=1}^{m} f_r(\mathbf{x}, \mathbf{d}_k, \mathbf{d}) f_p(\mathbf{x}, \mathbf{d}_k)(\mathbf{n} \cdot \mathbf{d}_k), \quad (7)$$

where m is the number of sampled incoming rays per intersection point. Note that the formulation of f_p in Eq. 6a also incorporates the computation of L. Thus, a recursive ray tracing process is constructed. The ray tracing algorithm unfolds in N levels. At each level, it samples a set of incoming lights emitted from the current shading point $\mathbf{x}$, where the color of each light is given using Eq. 6b if no intersections are found with the object, and the tracing process is terminated. Otherwise, the ray tracing delves deeper

into the next level, permitting the light to do an additional bounce. Upon the N-th level tracing, if there still exists an intersection point with the object mesh, we employ an MLP to predict the lighting color: $L^{(N)}(\mathbf{x}, \mathbf{d}) = \text{MLP}(\mathbf{x}, \mathbf{d})$, thus concluding the journey of the light.

Material-Aware Cone Sampling. Representing light via the proposed 5D neural plenoptic function offers a fidelity improvement. However, directly applying it is not practical, as it introduces a significant computational cost in the ray marching of NeRF, which demands sampling along rays to get the color information for each individual light. The complexity is further increased when doing important ray sampling for ray tracing within the BRDF lobe to satisfy the rendering equation.

To address this challenge, we introduce a material-aware cone sampling technique. In the first stage, we adopt a Mip-NeRF as our environmental field, which samples cones instead of rays like a vanilla NeRF. The introduction of Mip-NeRF is informed by the innate congruence of its cone sampling with the BRDF lobe-centric importance ray sampling. During the training of Mip-NeRF in the fields learning stage, the pre-integrated Gaussian samples are employed through Integral Positional Encoding (IPE), yielding better rendering results than a vanilla NeRF without incurring obvious extra costs. During the fields learning stage, the environmental field F_e is trained as a cone pre-filterer supervised by the training images.

In the material learning stage, we fix F_e and sample cones from the BRDF lobe. Because of the correlation between surface roughness and GGX distribution, we derive the cone angle directly from the predicted roughness. Subsequent ray marching of F_e yields the pre-filtered color for the incoming light of each integral component in the rendering equation. Specifically, considering the roughness parameter r at a shading point, we adopt the GGX distribution function $D(\mathbf{m})$ to describe the probability distribution of the microfacet normals $\mathbf{m}$. From this, we have the probability density function (PDF) for the azimuth angle ϕ and elevation angle θ [24]:

$$p_m(\theta, \phi) = \frac{r^4 \cos\theta \sin\theta}{\pi((r^4 - 1)\cos^2\theta + 1)^2}. \tag{8}$$

Our aim is to determine θ_m, which bounds the range of sampling the orientation of the microfacet normal $\mathbf{m}$, allowing the BRDF lobe to capture a predefined portion β of the light energy over the hemisphere. For practical purposes, we select $\beta = 0.9$ to encompass 90% of the radiance energy. This leads us to establish the cumulative distribution function (CDF) P_m and to resolve the following equation:

$$P_m(\theta_m) = \frac{r^4}{\cos^2\theta_m(r^4 - 1)^2 + (r^4 - 1)} - \frac{1}{r^4 - 1}, \tag{9}$$

wherein the solution of $P_m = \beta$ is presented as:

$$\theta_m = \arctan\left(r^2\sqrt{\frac{\beta}{1-\beta}}\right). \tag{10}$$

The angle θ_m, indicative of the spatial extent of the BRDF lobe, is thus directly correlated with the surface roughness. Consequently, we can sample the cone with the apex angle θ_c using θ_m via a simple math transform, a method without the need for learnable parameters. The details of the derivation can be found in the Supplemental.

Leveraging the pre-filtered lighting significantly reduces the number of lighting samples. Typically, employing a GGX distribution-based important sampling requires about 256 diffuse and 128 specular rays to get satisfactory results. In contrast, the proposed method achieves competitive results with merely 8 diffuse and 4 specular cones.

4. Experiments and Analysis

In this section, we detail the experiments conducted to validate the efficacy of our proposed method. We assess our method's performance in both geometry and material reconstruction, and demonstrate its practical applications.

Dataset. Our dataset is constructed to benchmark the inverse rendering of challenging glossy objects with diverse lighting interactions with nearby objects. It comprises 20 scenes in total: 10 real-world scenes captured through photographing glossy objects under varying lighting scenarios to produce multi-view images and 10 synthetic scenes. The synthetic scenes are crafted with glossy objects and manually designed background environments, and the multi-view images are rendered using photo-realistic path-tracing rendering engine. A distinctive aspect of our synthetic dataset is that we did not use environmental maps as the background directly, which is commonly used in prior works. Instead, we focus on more realistic settings where the objects are situated within tangible environments consisting of other objects. Samples from the proposed dataset are shown in Fig. 5.

Geometry Evaluations. To evaluate the quality of our method in geometry reconstruction, we compare our mesh outputs against those generated by several cutting-edge approaches, including Nerfacto [22], NeuS [26], and NeRO [10].

The visual comparison results are displayed in Fig. 4, which provides a side-by-side comparison of the reconstructed objects. We also compare the Chamfer Distance (CD) with other methods on the synthetic scenes with GT meshes, as shown in Tab. 1. Our method consistently outperforms the others across all tested scenes, achieving the lowest average CD and better visual quality with better geometric reconstruction fidelity. The comparison of the geometry reconstruction highlights the strengths of our proposed

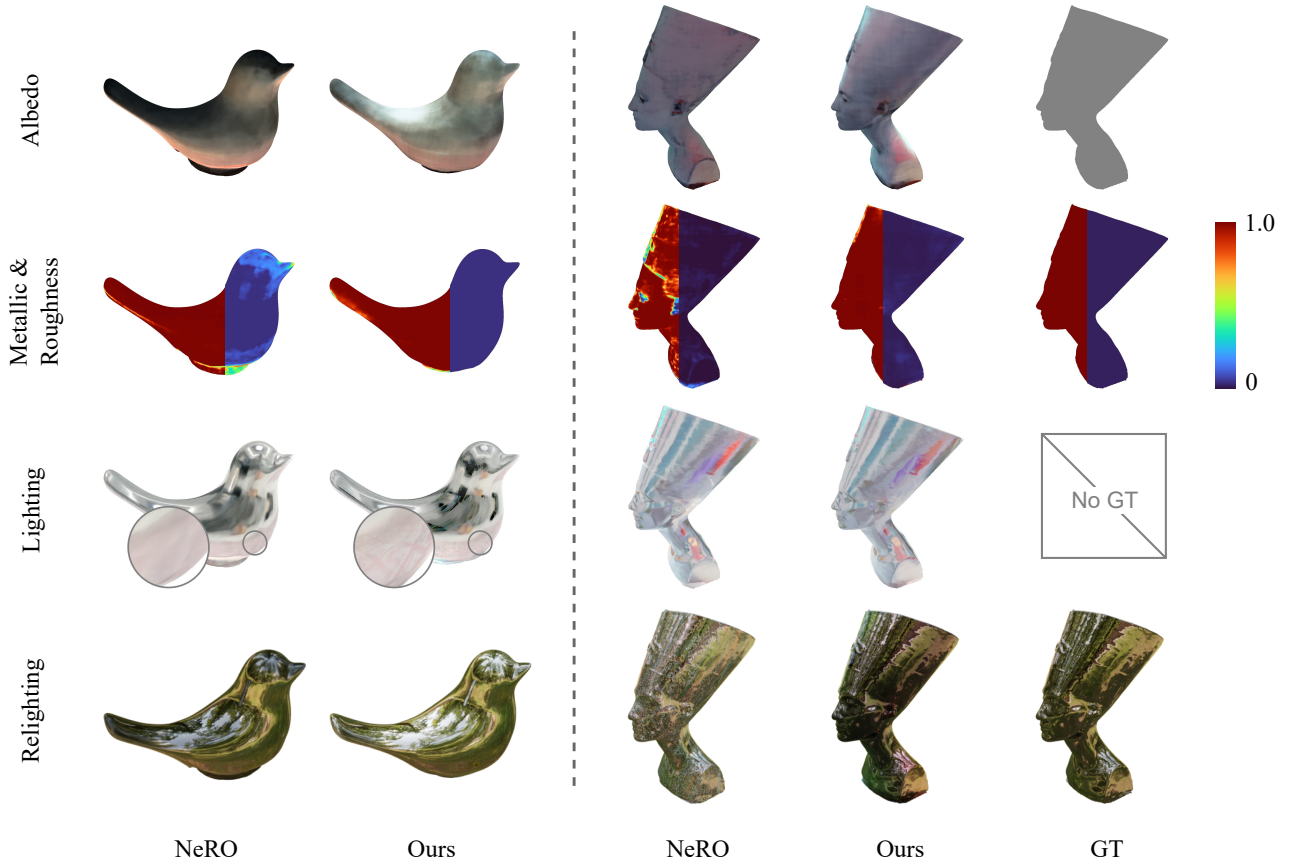

Figure 6. Comparison on the material estimation results. We display one real-world data (bird) and one synthetic data (Nefertiti) with GT for visual comparison. In the images of metallic & roughness row, metallic is displayed on the left half and roughness on the right half.

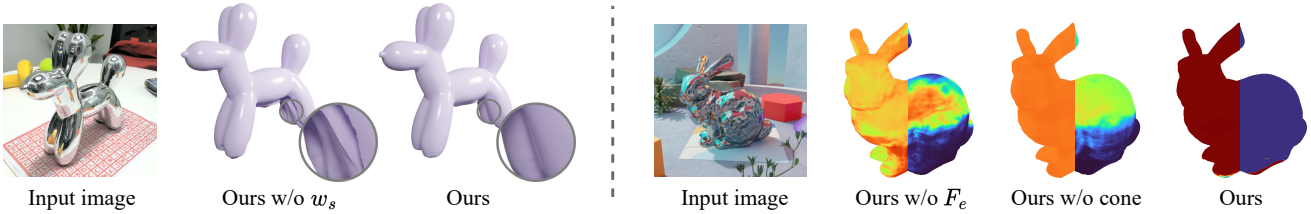

Figure 7. Ablation study results showcasing the impact of key components on inverse rendering quality for both stages. The absence of dynamic weighting (Ours w/o w_s) leads to less smooth surface reconstruction, while the lack of environmental field (Ours w/o F_e) or material-aware cone sampling method (Ours w/o cone) diminishes material fidelity. In each image of Figure (b), metallic is displayed on the left half and roughness on the right half.

dynamic weighting loss mechanism from the decoupled colors in the first stage. The performance of our method in reconstructing accurate and detailed geometries lays the groundwork for the subsequent material property estimation and their applications of seamlessly integrated into the rendering engine.

Material Evaluations. Moving beyond geometry, we compare the material estimation results produced by our method with NeRO [10]. Given our synthetic data subset, we possess the ground truth values for albedo, roughness, and metallic maps, allowing for a direct quantitative assessment via Mean Squared Error (MSE) as illustrated in Tab. 1. For visual comparison, we present two samples to validate compared methods and GT in Fig. 6. Both the quantitative and qualitative results reveal the competitive performances of our method in reconstructing materials with high fidelity, where our proposed neural plenoptic function is able to predict a more precise and smooth roughness, metallic, and

	Geometry Comparison (CD↓)				Roughness / Metallic / Albedo Comparison (MSE↓)	
	Nerfacto [22]	NeuS [26]	NeRO [10]	Ours	NeRO [10]	Ours
Bunny	0.05498	0.00852	0.00153	**0.00147**	**0.002** / 0.022 / 0.044	**0.002 / 0.016 / 0.022**
Box	0.07028	0.03339	0.00412	**0.00135**	0.003 / 0.068 / **0.056**	**0.001 / 0.019** / 0.067
Beethoven	0.04038	0.01706	0.00197	**0.00146**	0.007 / 0.030 / 0.031	**0.004 / 0.026 / 0.027**
Suzanne	0.05753	0.00754	0.00264	**0.00227**	0.004 / 0.030 / **0.024**	**0.001 / 0.022** / 0.029
Nefertiti	0.05299	0.01053	0.00587	**0.00167**	0.020 / 0.103 / 0.040	**0.008 / 0.021 / 0.035**
Avg.	0.05523	0.01541	0.00322	**0.00164**	0.007 / 0.051 / 0.039	**0.003 / 0.020 / 0.036**

Table 1. Quantitative comparison results. We compare the Chamfer Distance (CD) between the meshes from our method and those from other methods for the synthetic objects with GT on the left, and compare the MSE between the materials from of our methods and those from other methods on the right.

slightly better albedo compared with the existing methods.

Ablation Study. To assess the contribution of different components in our model, we perform an ablation study on the two stages of our method. In the first stage, we study the effect of the proposed dynamic weighting method. We compare the result of our full model with that of the variant where the proposed dynamic weighting scheme is excluded (*i.e.*, Ours w/o w_s). In the second stage, we study the results of our model in material estimation (1) by replacing the environmental field with an MLP to predict environmental lighting (*i.e.*, Ours w/o F_e), and (2) by replacing the material-aware cone sampling mechanism by a fixed number of rays (*i.e.*, Ours w/o cone). The ablation results on two samples are shown in Fig. 7.

We can see that by omitting the dynamic weighting strategy, which is crucial in balancing the influence of uncertain regions, we observe a decrease in the model's ability to reconstruct a smooth surface for the target object. Without the environmental field or material-aware cone sampling method for lighting representation, it leads to a degradation in material fidelity and introduces more ambiguity. Overall, we have demonstrated through the ablation study the importance of different parts of our method on the quality of the final inverse rendering results.

Limitation. While our method has been shown to make advancements in physically-based inverse object rendering, it is important to acknowledge certain limitations inherent in our current method.

One of the primary constraints of our approach is that the material learning stage is heavily reliant on the quality of the reconstructed geometry in the field learning stage. This dependency implies that the inaccuracies in the recovered geometry may adversely affect the material estimation process. If the geometry reconstructed in the first stage is not precise, it could lead to suboptimal material estimations in the subsequent stage.

Besides, although our method reduces the ambiguities typically present in inverse rendering tasks. we observe that the decomposition between lighting, geometry, and material properties is not entirely unambiguous. For example, there are scenarios where the color affected by the geometric structure is inaccurately incorporated into the albedo. This suggests an inherent complexity in perfectly disentangling the contributing factors to the final rendered image even with an explicit representation of lighting.

In future works, addressing these limitations will be crucial to further enhance the robustness and accuracy of our neural inverse rendering method. Potential improvements may include introducing more sophisticated constraints for geometry-material interdependence and refining the decomposition process to minimize ambiguities.

5. Conclusion

In this paper, we have presented a novel physically-based inverse rendering method for glossy objects, with the introduction of Neural Plenoptic Function (NeP) based on NeRFs. Our method addresses the limitations of the dependency on the simplified lighting representation in previous NeRF-based inverse rendering approaches. It is formulated as a two-stage model. The fields learning stage enhances the accuracy of 3D geometry reconstruction, especially for glossy objects under complex lighting. In the material learning stage, NeP employs a 5D neural plenoptic function for lighting representation based on the object field and environmental field, leading to higher-fidelity material estimation and inverse rendering. Our proposed material-aware cone sampling strategy further improves the efficiency of material learning. Experiments on real-world and synthetic datasets demonstrate the superior performance of our method.

Acknowledgements

This work is partly supported by a GRF grant from the Research Grants Council of Hong Kong (Ref.: 11205620).

References

[1] Jonathan T. Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P. Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *ICCV*, 2021. 1, 2, 3, 4

[2] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *CVPR*, 2022. 2

[3] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Zip-nerf: Anti-aliased grid-based neural radiance fields. In *ICCV*, 2023.

[4] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *ECCV*, 2022. 2

[5] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *CVPR*, 2022. 1, 2

[6] Wenhang Ge, T. Hu, Haoyu Zhao, Shu Liu, and Yingke Chen. Ref-neus: Ambiguity-reduced neural implicit surface learning for multi-view reconstruction with reflection. In *ICCV*, 2023. 4

[7] Jon Hasselgren, Nikolai Hofmann, and Jacob Munkberg. Shape, Light, and Material Decomposition from Images using Monte Carlo Rendering and Denoising. In *NeurIPS*, 2022. 1, 3, 4

[8] Wenbo Hu, Yuling Wang, Lin Ma, Bangbang Yang, Lin Gao, Xiao Liu, and Yuewen Ma. Tri-miprf: Tri-mip representation for efficient anti-aliasing neural radiance fields. In *ICCV*, 2023. 1, 2

[9] Zhaoshuo Li, Thomas Müller, Alex Evans, Russell H Taylor, Mathias Unberath, Ming-Yu Liu, and Chen-Hsuan Lin. Neuralangelo: High-fidelity neural surface reconstruction. In *CVPR*, 2023. 1

[10] 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. 1, 3, 4, 5, 6, 7, 8

[11] Xiaoxiao Long, Cheng Lin, Lingjie Liu, Yuan Liu, Peng Wang, Christian Theobalt, Taku Komura, and Wenping Wang. Neuraludf: Learning unsigned distance fields for multi-view reconstruction of surfaces with arbitrary topologies. In *CVPR*, 2023. 1

[12] William E. Lorensen and Harvey E. Cline. Marching cubes: A high resolution 3d surface construction algorithm. *Proceedings of the 14th annual conference on Computer graphics and interactive techniques*, 1987. 4

[13] Li Ma, Xiaoyu Li, Jing Liao, Qi Zhang, Xuan Wang, Jue Wang, and Pedro V. Sander. Deblur-nerf: Neural radiance fields from blurry images. In *CVPR*, 2022. 2

[14] Ricardo Martin-Brualla, Noha Radwan, Mehdi S. M. Sajjadi, Jonathan T. Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *CVPR*, 2021. 4

[15] Leonard McMillan and Gary Bishop. Plenoptic modeling: an image-based rendering system. *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, 1995. 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. In *ECCV*, 2020. 1, 2, 4

[17] Ben Mildenhall, Peter Hedman, Ricardo Martin-Brualla, Pratul P. Srinivasan, and Jonathan T. Barron. Nerf in the dark: High dynamic range view synthesis from noisy raw images. In *CVPR*, 2022. 2

[18] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4):1–15, 2022. 1, 2

[19] 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*, 2022. 1, 3

[20] Michael Oechsle, Songyou Peng, and Andreas Geiger. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *ICCV*, 2021. 1

[21] Christian Reiser, Richard Szeliski, Dor Verbin, Pratul P. Srinivasan, Ben Mildenhall, Andreas Geiger, Jonathan T. Barron, and Peter Hedman. MERF: memory-efficient radiance fields for real-time view synthesis in unbounded scenes. *ACM Trans. Graph.*, 42(4):89:1–89:12, 2023. 1

[22] Matthew Tancik, Ethan Weber, Evonne Ng, Ruilong Li, Brent Yi, Terrance Wang, Alexander Kristoffersen, Jake Austin, Kamyar Salahi, Abhik Ahuja, David McAllister, Justin Kerr, and Angjoo Kanazawa. Nerfstudio: A modular framework for neural radiance field development. In *SIGGRAPH*, 2023. 5, 6, 8

[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*, 2022. 1, 3, 4

[24] Bruce Walter, Stephen R. Marschner, Hongsong Li, and Kenneth E. Torrance. Microfacet models for refraction through rough surfaces. In *Proceedings of the 18th Eurographics Conference on Rendering Techniques*, 2007. 6

[25] Haoyuan Wang, Xiaogang Xu, Ke Xu, and Rynson W.H. Lau. Lighting up nerf via unsupervised decomposition and enhancement. In *ICCV*, 2023. 2, 4

[26] 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. In *NeurIPS*, 2021. 1, 4, 5, 6, 8

[27] Peng Wang, Yuan Liu, Zhaoxi Chen, Lingjie Liu, Ziwei Liu, Taku Komura, Christian Theobalt, and Wenping Wang. F2-nerf: Fast neural radiance field training with free camera trajectories. In *CVPR*, 2023. 1

[28] Yiming Wang, Qin Han, Marc Habermann, Kostas Daniilidis, Christian Theobalt, and Lingjie Liu. Neus2: Fast learning of neural implicit surfaces for multi-view reconstruction. In *ICCV*, 2023. 1

[29] Yao Yao, Jingyang Zhang, Jingbo Liu, Yihang Qu, Tian Fang, David McKinnon, Yanghai Tsin, and Long Quan.

Neilf: Neural incident light field for physically-based material estimation. In *ECCV*, 2022. 3

[30] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. In *NeurIPS*, 2021. 1

[31] Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Sattler, and Andreas Geiger. Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. In *NeurIPS*, 2022. 1

[32] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields. *ArXiv*, abs/2010.07492, 2020. 1

[33] Xiuming Zhang, Pratul P. Srinivasan, Boyang Deng, Paul E. Debevec, William T. Freeman, and Jonathan T. Barron. Nerfactor. *TOG*, 40:1 – 18, 2021. 1, 3

[34] Yiyu Zhuang, Qi Zhang, Xuan Wang, Hao Zhu, Ying Feng, Xiaoyu Li, Ying Shan, and Xun Cao. Neai: A pre-convoluted representation for plug-and-play neural ambient illumination. 2024. 3