

# NeFI: Inverse Rendering for Reflectance Decomposition with Near-Field Indirect Illumination

Haoqian Wu<sup>1</sup>, Zhipeng Hu<sup>1,2</sup>, Lincheng Li<sup>1\*</sup>, Yongqiang Zhang<sup>1</sup>, Changjie Fan<sup>1</sup>, Xin Yu<sup>3</sup>

<sup>1</sup> NetEase Fuxi AI Lab<sup>2</sup> Zhejiang University<sup>3</sup> The University of Queensland

f wuhaoqian, zphu, lilincheng, zhangyongqiang02, fanchangjie

g@corp.netease.com

xin.yu@uq.edu.au

## Abstract

Inverse rendering methods aim to estimate geometry, materials and illumination from multi-view RGB images. In order to achieve better decomposition, recent approaches attempt to model indirect illuminations reflected from different materials via Spherical Gaussians (SG), which, however, tends to blur the high-frequency reflection details. In this paper, we propose an end-to-end inverse rendering pipeline that decomposes materials and illumination from multi-view images, while considering near-field indirect illumination. In a nutshell, we introduce the Monte Carlo sampling based path tracing and cache the indirect illumination as neural radiance, enabling a physics-faithful and easy-to-optimize inverse rendering method. To enhance efficiency and practicality, we leverage SG to represent the smooth environment illuminations and apply importance sampling techniques. To supervise indirect illuminations from unobserved directions, we develop a novel radiance consistency constraint between implicit neural radiance and path tracing results of unobserved rays along with the joint optimization of materials and illuminations, thus significantly improving the decomposition performance. Extensive experiments demonstrate that our method outperforms the state-of-the-art on multiple synthetic and real datasets, especially in terms of inter-reflection decomposition.

## 1. Introduction

Inverse rendering, i.e., recovering geometry, material and lighting from images, is a long-standing problem in computer vision and graphics. It is important for digitizing our real world and acquiring high quality 3D contents in many applications such as VR, AR and computer games.

Recent methods [7, 41, 44, 45] represent geometry and materials as neural implicit fields, and recover them in an

Figure 1. Our method integrates lights through path tracing with Monte Carlo sampling, while Invrender [45] uses Spherical Gaussians to approximate the overall illumination. In this way, our method simultaneously optimizes indirect illuminations and materials, and achieves better decomposition of inter-reflections.

analysis-by-synthesis manner. However, how to decompose the indirect illumination from materials is still challenging. Most methods [6, 7, 25, 41, 44] model the environment illuminations but ignore indirect illuminations. As a result, the inter-reflections and shadows between objects are mistakenly treated as materials. Invrender [45] takes the indirect illumination into consideration, and approximates it with Spherical Gaussian (SG) for computation efficiency. Since SG approximation cannot model the high frequency details, the recovered inter-reflections tend to be blurry and contain artifacts. Besides, indirect illuminations estimated by an SG network cannot be jointly optimized with materials and environment illuminations.

In this paper, we propose an end-to-end inverse render-

\* Corresponding author.

ing pipeline that decomposes materials and illumination, while considering near-eld indirect illumination. In contrast to the method [45], we represent the materials and the indirect illuminations as neural implicit elds, and jointly optimize them with the environment illuminations. Furthermore, we introduce a Monte Carlo sampling based path tracing to model the inter-reflections while leveraging SG to represent the smooth environment illuminations. In the forward rendering, incoming rays are sampled and integrated by Monte Carlo estimator instead of being approximated by a pretrained SG approximator, as shown in Fig. 1. To depict the radiance, the bounced secondary rays are further traced once and computed based on the cached neural indirect illumination. During the joint optimization, the gradients could be directly propagated to revise the indirect illuminations. In this way, high frequency details of the inter-reflection can be preserved.

Specifically, to make our proposed framework work, we need to address two critical techniques:

(i) The Monte Carlo estimator is computationally expensive due to the significant number of rays required for sampling. To overcome this, we use importance sampling to improve integral estimation efficiency. We also find that SG is a better representation of environment illuminations and adapt the corresponding importance sampling techniques to enhance efficiency and practicality.

(ii) Neural implicit elds often suffer generalization problems when the view directions deviate from the training views, which is the common case of indirect illumination. This would lead to erroneous decomposition between materials and illuminations. It is hard to determine whether radiance comes from material albedos or indirect illuminations as the indirect illuminations from unobserved directions are unconstrained or could have any radiance. To learn indirect illuminations from unobserved directions, we introduce a radiance consistency constraint that enforces the implicit neural radiance produced by the neural implicit elds and path tracing results of unobserved directions. In this fashion, the ambiguity between materials and indirect illuminations has been significantly mitigated. Moreover, they can be jointly optimized with environment illuminations, leading to better decomposition performance.

We evaluate our method on synthetic and real data. Experiments show that our approach achieves better performance than others. Our method can render sharp inter-reflection and recover accurate roughness as well as diffuse albedo. Our contributions are summarized as follows:

- We propose an end-to-end inverse rendering pipeline that decomposes materials and illumination, while considering near-eld indirect illumination.
- We introduce the Monte Carlo sampling based path tracing and cache the indirect illumination as neural radiance, resulting in a physics-faithful and easy-to-

optimize inverse rendering process.

- We employ SG to parameterize smooth environment illumination and apply importance sampling techniques to enhance efficiency and practicality of the pipeline.
- We introduce a new radiance consistency in learning indirect illuminations, which can significantly alleviate the decomposition ambiguity between materials and indirect illuminations.

## 2. Related Work

### 2.1. Implicit Neural Representation

Implicit neural representations [24, 36, 39] have achieved impressive performance. NeRF [24] represents scenes as radiance elds and volumetric density elds, and achieves photo-realistic novel view synthesis. To better model geometry, some methods, such as IDR [39] and NeuS [36], further represent geometry as Signed Distance Functions (SDFs). However, the object appearance is represented as a radiance eld, which simply outputs outgoing radiance of each 3D point given a view direction. Thus, the surface points can be treated as emissive lighting sources. These methods are not suitable for relighting and material editing.

### 2.2. Material and Illumination Estimation

To estimate object materials, most of previous capture systems rely on constrained settings, such as by light-stages with controlled lights and cameras [10, 20, 43], using moving cameras with co-located flashlight [4, 5], placing objects on a turntable platform, or capturing in special lighting patterns [16]. Apart from those hardware-specific systems, some data-driven methods [3, 21–23, 29, 31, 38, 40] try to directly estimate materials from a single image by neural networks with priors from large-scale datasets. However, they fail to generalize beyond the training datasets and are often restricted to the planar geometry. Differentiable rendering methods [1, 26] aim to make graphic rendering process differentiable, and recover material and illumination by optimization. However, they suffer from demanding computation cost and challenging optimization complexity.

Recent works have been extended to more flexible capture settings by implicitly representing geometry and materials and optimizing them in differentiable pipelines. Most methods adopt differentiable rendering algorithms and only consider direct illuminations, such as Spherical Gaussians (e.g., NeRD [6] and PhysSG [41]), Spherical Harmonics (e.g., NeROIC [17]), point lighting of low resolution environment maps (e.g., NeRFactor [44]), and pre-filtered approximations (e.g., Neural-PIL [7] and NVDiffrec [25]). Some methods [11, 32] integrated with Monte Carlo sampling but still ignore modeling multiple light bouncing, e.g., NVDiffrecmc [11] only considers direct illumination and NeRV [32] considers only one indirect bounce.

Figure 2. Proposed Rendering Pipeline. To render a camera ray intersecting with surface at location  $x$ , we first sample incoming rays and trace them to obtain their second surface intersection and visibility  $V(x; w_i)$  for light source (environment illumination). Then, SVBRDF values at location  $x^0$  and outgoing radiance  $L_o(x^0; w_i)$  of second intersection  $x^0$ , i.e., indirect illumination, are obtained by neural SVBRDF  $M$  and neural radiance  $L$ , respectively. Besides, radiance of incoming rays from light source  $E(w_i)$ , i.e., direct illumination, is obtained by SG environment illumination  $E$ . Finally, a Monte Carlo Estimator is used for rendering the final results as described in Eq. (2). Materials, indirect illumination and environment illumination are jointly optimized by the reconstruction loss.

Invrندر [45], most close to our method, considers a reflected light integration over hemisphere around multi-bounce indirect illuminations. It adopts SG rendering the surface normal. Incoming radiance may come directly from light source, known as direct illumination, or indirectly from other surface after multiple light bouncing, at the first stage. Besides, incoming light of adjacent surface known as indirect illumination. For indirect illumination, face points may vary drastically because incoming light is recursive rendering is often needed.

represented as SGs and modeled by a coordinate-based network trained at the second stage. In contrast, our method considers indirect illuminations and proposes a joint learning approach. Therefore, we can render sharp and complex self-reflection effects and recover material properties with higher quality, as shown in Fig. 1.

### 2.3. Theoretical Rendering Process

In theory, the rendering process at the intersection location  $x$  of the camera ray with direction  $w_o$  can be expressed by the rendering equation [15]:

$$L_o(x; w_o) = \int_{\Omega} L_i(x; w_i) f_r(x; w_o; w_i) (w_i \cdot n) dw_i; \quad (1)$$

where  $L_i(x; w_i)$  is the incoming radiance at surface point  $x$  along the direction  $w_i$ ,  $f_r$  is the BRDF function and the outgoing radiance  $L_o(x; w_o)$  in observed direction  $w_o$ .

## 3. Proposed Method

### 3.1. Overview

Given a group of multi-view images captured under static illumination, we aim to decompose the geometry and Spatially Varying BRDF (SVBRDF) of the object and the illumination. We take the global illumination effect into consideration, such as shadows and inter-reflections, but consider transparent and translucent objects outside the scope of our work.

The geometry is represented as the zero level set of SDF as in [36, 39, 45], which is modeled by an MLP that maps a 3D location  $x \in \mathbb{R}^3$  to an SDF value and a geometric feature vector  $f \in \mathbb{R}^{512}$ . The material is encoded by another MLP as neural SVBRDF  $M_M(x; f)$ . The environment illumination is parameterized by SG coefficients [41]

$E_E(w_i)$ , where  $w_i \in \mathbb{R}^2$  is the light direction. The radiance is represented as another  $\text{MLP}_L(x; n; w_o; f)$ , which outputs outgoing radiance given location  $x$ , normal  $n \in \mathbb{R}^3$ , viewing direction  $w_o \in \mathbb{R}^3$  and feature  $f$ .

We solve the inverse rendering problem in an analysis-by-synthesis manner by forward rendering with parameterized components. Similar to prior works, we pretrain the geometry SDF by NeuS [36] and freeze the parameters. Given a viewing direction  $w_o$ , we first find the intersection  $x$  on the geometry surface through sphere tracing on the SDF. Then, the path tracing based rendering integrates the outgoing radiance  $L_o(x; w_o)$  in direction  $w_o$ . The rendering results are compared with the input image pixels to optimize  $L$ ,  $M$  and  $E$ .

### 3.2. Cached Path Tracing based Rendering

Theoretical rendering process described in Sec. 2.3 cannot be practically implemented because of its production integration and exponential recursive light bounce. In contrast to simply ignoring light bounce and adopting approximations rendering method such as SG [35], we implement the forward rendering process based on path tracing [18, 19], which is an efficient and differentiable rendering framework that fully incorporates the light bounces. We implement rendering equation in Eq. (1) by Monte Carlo estimator as:

$$L_o(x; w_o) = \frac{1}{N} \sum_{i=1}^N \frac{L_i(x; w_i) f_r(x; w_o; w_i) (w_i \cdot n)}{p(w_i)}. \quad (2)$$

It estimates the production integration by sampling incoming rays with direction  $w_i$  drawn from distribution  $p(w_i)$ .

The incoming radiance includes light rays directly emitted by the light source, i.e., direct illumination, and ones bouncing off of the object surface multiple times, i.e., indirect illumination:

$$L_i(x; w_i) = V(x; w_i) E(w_i) + (1 - V(x; w_i)) L_o(x^0; w_i); \quad (3)$$

where  $E(w_i)$  is the incoming radiance from light source along direction  $w_i$ , and  $L_o(x^0; w_i)$  is the incoming radiance from the second intersection of the ray.  $V(x; w_i)$  is the visibility of location  $x$  for light source and indicates the illumination type, obtained during path tracing.

To obtain indirect illumination, in theory, we should recursively render the outgoing radiance at location along direction  $w_i$  by Eq. (1). This may lead to intractable computation and optimization difficulties for the optimization process. Inspired by [45], we employ the neural radiance  $L_L$  to represent the neural outgoing radiance after multiple light bouncing of the second ray intersection, known as indirect illumination. In such manner, we cache the indirect illumination and avoid the exhaustive ray tracing. The

indirect incoming radiance is calculated as:

$$L_o(x^0; w_i) = L_L(x^0; n^0; w_i; f^0); \quad (4)$$

where  $n^0$  and  $f^0$  are the surface normal and geometric feature vector at  $x^0$  respectively.

The complete pipeline of our path tracing based rendering is shown in Fig. 2. The rendering process is differentiable for optimizing neural radiance  $L_L$ , neural SVBRDF  $M_M$  and SG environment illumination  $E_E$ .

### 3.3. Efficient Monte Carlo Estimator

Monte Carlo estimator needs to sample a large number of rays to produce high-quality results without noise, which is not affordable for practical optimization. Although some techniques can tackle the issue, most of them are inappropriate in inverse rendering scenario. For example, denoising techniques [2, 9, 12, 46] require spatial information of the whole rendered image and temporal information from previous frames. These information is not available in inverse rendering, where we randomly pick posed images and sample some pixels for training. Hence, we apply importance sampling techniques, including cosine sampling as well as GGX importance sampling [13], to improve Monte Carlo estimator efficiency and use multiple importance sampling method [28, 34] to fuse all of them.

For light importance sampling, the piecewise-constant 2D distribution sampling [28] is not applicable, since it is designed for known environment illumination represented as the 2D array. As mentioned in [6, 45], parameterizing environment illumination in such a way could make each pixel of environment maps vary independently, lead diffuse albedo baked in illumination and cause illumination inefficient for optimization. In contrast, we parameterize environment illumination as SG coefficients, and introduce and adapt Spherical Gaussian (SG) distribution sampling [14] as the corresponding light importance sampling technique:

$$p_{SG}(w_i) = \sum_{k=1}^M a_k \frac{k}{2(1 - e^{-2k})} e^{-k(w_i \cdot k - 1)}; \quad (5)$$

$$a_k = \frac{k \max(n_k; \epsilon)}{\sum_{j=1}^M j \max(n_j; \epsilon)}; \quad (6)$$

where  $k; \epsilon \in \mathbb{R}^3$  are SG parameters of environment illumination, i.e., lobe axis, lobe sharpness and lobe amplitude of SG respectively, and  $\epsilon$  is the energy of lobe amplitude. Since we only need to sample rays over the hemisphere around  $n$ , we assign a tiny weight to SG components whose lobe axis is beyond the hemisphere. According to the SG distribution, it has a higher probability to sample light rays that belong to brighter SG lobes and are closer to SG lobe centers. The detailed process is described in our supplemental material.

Figure 3. Training with traced rays. We alternatively train with observed rays and unobserved rays.

### 3.4. Training with Traced Rays

We alternatively train our framework with observed rays and unobserved rays. Training with observed rays alone is challenging because some locations or view directions are not observed due to occlusion. Besides, there exists ambiguity between indirect illumination and material properties, since indirect incoming rays with many directions cannot be directly observed by the camera. Hence, neural radiance  $L_L$  is indeterminate with re-render loss alone. We propose to utilize the unobserved rays to provide more information and constraints.

Train with observed rays. As shown in the top of Fig. 3, we optimize  $L$ ,  $M$  and  $E$  with observed rays using the following loss:

$$\begin{aligned} \mathcal{L}_o = & \frac{1}{N_{obj}} \sum_{i=1}^{N_{obj}} \|kc_i^{ob} - c_i^{obgt}\|_2^2 k_1 \\ & + \frac{1}{N_{obj}} \sum_{i=1}^{N_{obj}} \|ke_i^{ob} - c_i^{obgt}\|_2^2 k_1 \\ & + \frac{1}{N_{nobj}} \sum_{i=1}^{N_{nobj}} \|kc_i^{nob} - c_i^{nobgt}\|_2^2 k_2 \end{aligned} \quad (7)$$

The first term is the reconstruction loss of path-tracing-based rendering results of object pixels  $c_i^{ob}$  with

ground truth  $c_i^{obgt}$ . The second term is the reconstruction loss of neural rendering results of object pixels  $ke_i^{ob}$ . The third term is the environment reconstruction loss, which renders non-object pixels  $c_i^{nob}$  to compare with the ground truth  $c_i^{nobgt}$ . Train with unobserved rays. As shown in the bottom of Fig. 3, we additionally optimize components with unobserved rays. Although there is no ground truth of unobserved rays, the consistency of neural rendering and path tracing based rendering of the rays can be used for training:

$$\mathcal{L}_u = \frac{1}{N_{sec}} \sum_{j=1}^{N_{sec}} \|kc_j^0 - e_j^0\|_2^2 k_1; \quad (8)$$

where  $c_j^0 = L_o(x^0, w_i)$  is the path tracing rendering result at the unobserved ray origin  $x^0$  for outgoing direction  $w_i$ , and  $e_j^0 = L_L(x^0, n^0, w_i; f^0)$  is the neural rendering result.  $N_{sec}$  is the amount of unobserved rays.

Unobserved rays are uniformly sampled from the secondary rays, which are generated in path tracing of observed rays, instead of being generated by virtual cameras. We alternatively train the networks with observed rays and unobserved rays rather than aggregate the two losses. The unobserved rays is optimized every steps.

### 3.5. Implementation Details

We set the sampled number of rays = 64 and the SG components number  $M = 128$ . We set the loss weights  $w_1 = 1:0$ ,  $w_2 = 1:0$  and unobserved rays training interval  $K = 10$ . SDF and neural SVBRDF contains 8 layers with 512 hidden units and positional encoding [24, 33] is applied to the input 3D locations with 6 and 10 frequency components respectively. Neural radiance contains 4 layers with 512 hidden units and positional encoding of the 3D location and directions with 10 and 4 frequency components respectively. Our approach is implemented in Pytorch [27] and optimized with Adam with learning rate  $10^{-4}$ . We train about 120 epochs on 4 RTX 3090 GPUs and it takes about 5 hours. We use the simplified Disney BRDF model [8] with parameters including roughness, diffuse albedo and specular albedo. The specular albedo is assumed as 0.5, the value of common dielectric surfaces. For stable optimization, we fix roughness at the first 50 epochs to warm up.

## 4. Experiment

### 4.1. Synthetic Data

We collect four synthetic scenes with obvious self-reactions to showcase the quality of the estimated BRDF parameters and illumination. We render 200 images and their masks under a natural HDRI environment map via Blender Cycles and uniformly sample 100 for training and leave the rest for testing. We also render diffuse albedo



Figure 4. Qualitative comparisons with the state-of-the-art. We present synthetic rendering result and specular reflection component as well as estimated aligned diffuse albedo [41, 44] and roughness of each method on two scenes. The roughness of NerFactor [44] is visualized with the BRDF identity latent code. Compared with previous works, our method better simulates sharp self-reflection and separates shadows and indirect illumination from diffuse albedo. Besides, roughness maps recovered by our method are more accurate.

maps, roughness maps, and specular reflection components. Fig. 4 shows that our method could render sharp reflection for test images to evaluate the inverse rendering ability. The reflection effect due to our joint-learning path-tracing-based image resolution is set to 512 512.

#### 4.2. Comparison with the State-of-the-Art

The closest work to ours is Invrender [45] which forms our primary comparisons. We also compare with other methods tackling on the similar inverse rendering settings as this paper for thorough comparisons, including NerFactor [44] and PhySG [41]. We mainly focus the evaluation on material properties and illumination estimation instead with SG and is trained in three stages. They represent visibility of shape reconstruction. We make quantitative comparisons on the synthetic data and directly learn geometry from mesh for every approach to better evaluate material estimation ability without interference of geometry reconstruction quality. Following previous works [41, 45], we adopt Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM) [37], and Learned Perceptual Image Patch Similarity (LPIPS) [42] to evaluate image quality metrics and evaluate the diffuse albedo after aligning.

Invrender [45] approximates the indirect illumination and visibility are modeled more accurately, so less indirect illumination and shadow are baked into diffuse albedo. Tab. 1 shows quantity improvements, especially in roughness estimation and specular reflection synthesis, of our methods. Invrender [45] approximates the indirect illumination and incoming indirect light of each point as SG parameters by neural networks and train in the second stage, then optimize materials with SG rendering at the third stage. SG does not work well for high-frequency lighting, and the visibility and indirect illumination of adjacent surface points may vary drastically, hence, reflection tend to be noisy and rough, as shown in rendering RGB and specular RGB results in Fig. 4. Besides, the radiance field, trained with limited observed rays of multi-view images, could not

Method	Roughness	Aligned Diffuse Albedo			View Synthesis Specular RGB			View Synthesis RGB		
	MSE #	PSNR"	SSIM"	LPIPS#	PSNR"	SSIM"	LPIPS#	PSNR"	SSIM"	LPIPS#
NeRFactor [44]	-	21.8857	0.9159	0.0953	19.2751	0.8695	0.1147	29.9826	0.9597	0.0475
PhySG [41]	0.0481	19.7933	0.8988	0.1109	26.7784	0.9025	0.0693	31.0425	0.9642	0.0436
Invrender [45]	0.0464	27.4026	0.9426	0.0914	26.1370	0.9035	0.0831	30.8743	0.9616	0.0490
Ours	0.0065	28.1094	0.9516	0.0845	34.2930	0.9608	0.0416	31.0909	0.9586	0.0528

Table 1. Quantitative evaluations. We present quantitative comparison with the state-of-the-art. Results show that our method achieves impressive improvements, especially in roughness estimation and specular reflection synthesis. Due to the rendering noise of our path tracing based rendering, some metrics of the view synthesis RGB are slightly worse than other methods.

Figure 5. Ablation on environment illumination representation. Representing environment illumination in environment maps and using 2D piece-constant sampling causes neighbor pixels in the environment map vary independently, and part of the illumination is baked into diffuse albedo. Representing environment illumination in SG coefficients and using SG importance sampling better decomposes the illumination and diffuse albedo.

predict radiance of indirect rays with unobserved directions correctly. Hence, more indirect illumination and shadow are baked in diffuse albedo as shown in Fig. 4.

Other methods [41, 44] ignore indirect illumination and achieve worse results of material recovering. Indirect illumination is baked in diffuse albedo and roughness maps are recovered inaccurately.

### 4.3. Ablation Studies

Ablations on environment illumination representation.

As shown in Sec. 4.2, representing environment illumination in SG coefficients and using SG importance sampling are better for optimization. Representing environment illumination in 2D environment maps and using piece-constant sampling causes neighbor pixels in the environment map vary independently, and it is easier for illumination to be baked into diffuse albedo.

Ablations on training with unobserved rays. We ablate the unobserved training, and compare the results in Fig. 6. We visualize the mean of indirect illumination from all directions at each point and show the recovered diffuse albedo as well as roughness. Without unobserved training, the network predicts wrong indirect illumination at some lo-

Figure 6. Ablations on training with unobserved rays. We visualize the incoming indirect light for each point and present recovered diffuse albedo and roughness under both settings. Training with unobserved rays helps the decomposition of indirect light and diffuse albedo. Besides, roughness at the interstice of objects is recovered more accurately.

cations, especially at the interstice of objects. Due to the incorrect indirect illumination, the recovered diffuse albedo contains some artifacts, e.g., the bread on the side of the sausage of the hotdog. Besides, interstices, areas between the hotdog and the plane, are not visible by cameras from many directions. Hence, the roughness at these areas cannot be estimated correctly and confidently when only trained with observed rays.

Ablations on indirect lighting. We show the influence of modeling indirect illumination and visibility in Fig. 7. Without modeling indirect illumination, indirect illumination would be baked into diffuse albedo, resulting in wrong brightness. Further without modeling visibility, shadows would be baked into diffuse albedo and the roughness is not correctly recovered. These results show the necessity of indirect illumination modeling in inverse rendering.

Figure 7. Ablation on indirect lighting. Without modeling indirect illumination and visibility, indirect illumination and shadows would be baked into diffuse albedo and roughness.

Figure 9. Results on real captures. Our method estimates reasonable materials for real-world objects.

Figure 8. Relighting Results. Our method supports further relighting with the recovered materials.

Figure 10. Failure cases. Shadow might pose challenges for reflectance decomposition in some extreme cases.

#### 4.4. Relighting

We relight the objects with recovered material properties under two environment illuminations and show results in Fig. 8. Our method could recover accurate material properties and support further relighting.

#### 4.5. Results on Real Captures

We test our method on real captured images of 3 objects with different materials. Each scene has about 40 to 60 valid images for training and we use COLOMAP [30] to estimate the camera poses. We train our method without masks. Note that reflection properties of real materials are more complex in contrast to BRDF models and there are more interference in real capturing, e.g., video motion blur caused by moving cameras and illumination changing during capturing. As shown in Fig. 9, our method could estimate reasonable material properties.

#### 4.6. Failure Cases

As shown in Fig. 10, our method has difficulty in estimating roughness in large shadow areas due to the low visibility of scenes. In some extreme cases, the shadow may leak into the albedo because of illumination ambiguity.

### 5. Conclusion

To summarize, our paper presents an end-to-end inverse rendering pipeline that is capable of decomposing materials and illumination from multi-view images, while considering near-eld indirect illumination. Our method utilizes the Monte Carlo sampling based path tracing and cache the indirect illumination as neural radiance, enabling a physics-faithful and easy-to-optimize inverse rendering method. We implement an efficient Monte Carlo estimator and propose a novel radiance consistency constraint of unobserved rays to decrease the ambiguity. Extensive experiments demonstrate that our method models the sharp inter-reflections better and recovers material properties more accurately.

Our method still has some limitations. First, the shape is not joint optimized because visibility gradients are not handled well by current ray tracing technique. Second, to decrease the ambiguity of the inverse problem, the specular albedo is assumed as 0.5, the value of common dielectric surfaces. They will be the subject of our future works.

**Acknowledgements.** This research is funded in part by ARC-Discovery grant (DP220100800 to XY) and ARC-DECRA grant (DE230100477 to XY). We thank Yuanqing Zhang and Lumin Yang for generously sharing their knowledge. We also thank the anonymous reviewers for their constructive suggestions on this manuscript.



## References

- [1] Dejan Azinovic, Tzu-Mao Li, Anton Kaplanyan, and Matthias Nießner. Inverse path tracing for joint material and lighting estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2447–2456, 2019. 2
- [2] Steve Bako, Thijs Vogels, Brian McWilliams, Mark Meyer, Jan Novák, Alex Harvill, Pradeep Sen, Tony Derose, and Fabrice Rousselle. Kernel-predicting convolutional networks for denoising monte carlo rendering. *ACM Trans. Graph.* 36(4):97–1, 2017. 4
- [3] Jonathan T Barron and Jitendra Malik. Shape, illumination, and reflectance from shading. *IEEE transactions on pattern analysis and machine intelligence* 37(8):1670–1687, 2014. 2
- [4] Sai Bi, Zexiang Xu, Pratul Srinivasan, Ben Mildenhall, Kalyan Sunkavalli, Mila Hasan, Yannick Hold-Geoffroy, David Kriegman, and Ravi Ramamoorthi. Neural reflectance fields for appearance acquisition. *arXiv preprint arXiv:2008.03824* 2020. 2
- [5] Sai Bi, Zexiang Xu, Kalyan Sunkavalli, Mila Hasan, Yannick Hold-Geoffroy, David Kriegman, and Ravi Ramamoorthi. Deep reflectance volumes: Relightable reconstructions from multi-view photometric images. In *European Conference on Computer Vision*, pages 294–311. Springer, 2020. 2
- [6] Mark Boss, Raphael Braun, Varun Jampani, Jonathan T Barron, Ce Liu, and Hendrik Lensch. Nerd: Neural reflectance decomposition from image collections. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12684–12694, 2021. 1, 2, 4
- [7] 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. 1, 2
- [8] Brent Burley and Walt Disney Animation Studios. Physically-based shading at disney. *ACM SIGGRAPH*, volume 2012, pages 1–7. vol. 2012, 2012. 5
- [9] Chakravarthy R Alla Chaitanya, Anton S Kaplanyan, Christoph Schied, Marco Salvi, Aaron Lefohn, Derek Nowrouzezahrai, and Timo Aila. Interactive reconstruction of monte carlo image sequences using a recurrent denoising autoencoder. *ACM Transactions on Graphics (TOG)* 36(4):1–12, 2017. 4
- [10] Kaiwen Guo, Peter Lincoln, Philip Davidson, Jay Busch, Xueming Yu, Matt Whalen, Geoff Harvey, Sergio Orts-Escolano, Rohit Pandey, Jason Dourgarian, et al. The relightables: Volumetric performance capture of humans with realistic relighting. *ACM Transactions on Graphics (TOG)* 38(6):1–19, 2019. 2
- [11] 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
- [12] Jon Hasselgren, Jacob Munkberg, Marco Salvi, Anjul Patney, and Aaron Lefohn. Neural temporal adaptive sampling and denoising. In *Computer Graphics Forum*, volume 39, pages 147–155. Wiley Online Library, 2020. 4
- [13] Eric Heitz. Sampling the ggx distribution of visible normals. *Journal of Computer Graphics Techniques (JCGT)* 37(4):1–13, 2018. 4
- [14] Wenzel Jakob. Numerically stable sampling of the von mises-scher distribution on  $S^2$  (and other tricks). *Interactive Geometry Lab, ETH Zürich, Tech. Reppage* 6, 2012. 4
- [15] James T Kajiya. The rendering equation. *Proceedings of the 13th annual conference on Computer graphics and interactive techniques*, pages 143–150, 1986. 3
- [16] Kaizhang Kang, Cihui Xie, Chengan He, Mingqi Yi, Minyi Gu, Zimin Chen, Kun Zhou, and Hongzhi Wu. Learning efficient illumination multiplexing for joint capture of reflectance and shape. *ACM Trans. Graph.* 38(6):165–1, 2019. 2
- [17] Zhengfei Kuang, Kyle Olszewski, Menglei Chai, Zeng Huang, Panos Achlioptas, and Sergey Tulyakov. Neroc: Neural rendering of objects from online image collections. *ACM Trans. Graph.* 41(4), jul 2022. 2
- [18] Eric Lafortune. Mathematical models and monte carlo algorithms for physically based rendering. *Department of Computer Science, Faculty of Engineering, Katholieke Universiteit Leuven* 20:74–79, 1996. 4
- [19] Eric P Lafortune and Yves D Willems. Bi-directional path tracing. 1993. 4
- [20] Hendrik PA Lensch, Jochen Lang, Asla M, and Hans-Peter Seidel. Planned sampling of spatially varying brdfs. In *Computer graphics forum*, volume 22, pages 473–482. Wiley Online Library, 2003. 2
- [21] Zhengqin Li, Mohammad Shaei, Ravi Ramamoorthi, Kalyan Sunkavalli, and Manmohan Chandraker. Inverse rendering for complex indoor scenes: Shape, spatially-varying lighting and svbrdf from a single image. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2475–2484, 2020. 2
- [22] Zhengqin Li, Zexiang Xu, Ravi Ramamoorthi, Kalyan Sunkavalli, and Manmohan Chandraker. Learning to reconstruct shape and spatially-varying reflectance from a single image. *ACM Transactions on Graphics (TOG)* 37(6):1–11, 2018. 2
- [23] Daniel Lichy, Jiaye Wu, Soumyadip Sengupta, and David W Jacobs. Shape and material capture at home. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6123–6133, 2021. 2
- [24] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM* 65(1):99–106, 2021. 2, 5
- [25] 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*, pages 8280–8290, 2022. 1, 2

- [26] Merlin Nimier-David, Zhao Dong, Wenzel Jakob, and Anton Kaplanyan. Material and lighting reconstruction for complex indoor scenes with texture-space differentiable rendering. 2021. **2**
- [27] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* 32, 2019. **5**
- [28] Matt Pharr, Wenzel Jakob, and Greg Humphreys. *Physically based rendering: From theory to implementation*. Morgan Kaufmann, 2016. **4**
- [29] Shen Sang and Manmohan Chandraker. Single-shot neural relighting and svbrdf estimation. *European Conference on Computer Vision* pages 85–101. Springer, 2020. **2**
- [30] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition* pages 4104–4113, 2016. **8**
- [31] Soumyadip Sengupta, Jinwei Gu, Kihwan Kim, Guilin Liu, David W Jacobs, and Jan Kautz. Neural inverse rendering of an indoor scene from a single image. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8598–8607, 2019. **2**
- [32] Pratul P Srinivasan, Boyang Deng, Xiuming Zhang, Matthew Tancik, Ben Mildenhall, and Jonathan T Barron. Nerv: Neural radiance fields for relighting and view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pages 7495–7504, 2021. **2**
- [33] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in Neural Information Processing Systems* 33:7537–7547, 2020. **5**
- [34] Eric Veach and Leonidas J Guibas. Optimally combining sampling techniques for monte carlo rendering. *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques* pages 419–428, 1995. **4**
- [35] Jiaping Wang, Peiran Ren, Minmin Gong, John Snyder, and Baining Guo. All-frequency rendering of dynamic, spatially-varying radiance. In *ACM SIGGRAPH Asia 2009 papers* pages 1–10. 2009. **4**
- [36] 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. *NeurIPS* 2021. **2, 3, 4**
- [37] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 13(4):600–612, 2004. **6**
- [38] Xin Wei, Guojun Chen, Yue Dong, Stephen Lin, and Xin Tong. Object-based illumination estimation with rendering-aware neural networks. In *European Conference on Computer Vision* pages 380–396. Springer, 2020. **2**
- [39] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. *Advances in Neural Information Processing Systems* 33:2492–2502, 2020. **2, 3**
- [40] Ye Yu and William AP Smith. Inverserendernet: Learning single image inverse rendering. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pages 3155–3164, 2019. **2**
- [41] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pages 5453–5462, 2021. **1, 2, 3, 6, 7**
- [42] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. **6**
- [43] Xiuming Zhang, Sean Fanello, Yun-Ta Tsai, Tiancheng Sun, Tianfan Xue, Rohit Pandey, Sergio Orts-Escolano, Philip Davidson, Christoph Rhemann, Paul Debevec, et al. Neural light transport for relighting and view synthesis. *ACM Transactions on Graphics (TOG)* 40(1):1–17, 2021. **2**
- [44] Xiuming Zhang, Pratul P Srinivasan, Boyang Deng, Paul Debevec, William T Freeman, and Jonathan T Barron. Nerfactor: Neural factorization of shape and radiance under an unknown illumination. *ACM Transactions on Graphics (TOG)* 40(6):1–18, 2021. **1, 2, 6, 7**
- [45] 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, 2, 3, 4, 6, 7**
- [46] M. Zwicker, W. Jarosz, J. Lehtinen, B. Moon, R. Ramamoorthi, F. Rousselle, P. Sen, C. Soler, and S.-E. Yoon. Recent advances in adaptive sampling and reconstruction for monte carlo rendering. *Comput. Graph. Forum* 34(2):667–681, may 2015. **4**