

A Pre-convoluted Representation for Plug-and-Play Neural Ambient Illumination

Yiyu Zhuang^{1*}, Qi Zhang^{2*}, Xuan Wang², Hao Zhu¹, Ying Feng², Xiaoyu Li², Ying Shan², Xun Cao¹

¹ Nanjing University, Nanjing, China ² Tencent AI Lab, Shenzhen, China

<https://yiyuzhuang.github.io/NeAI/>

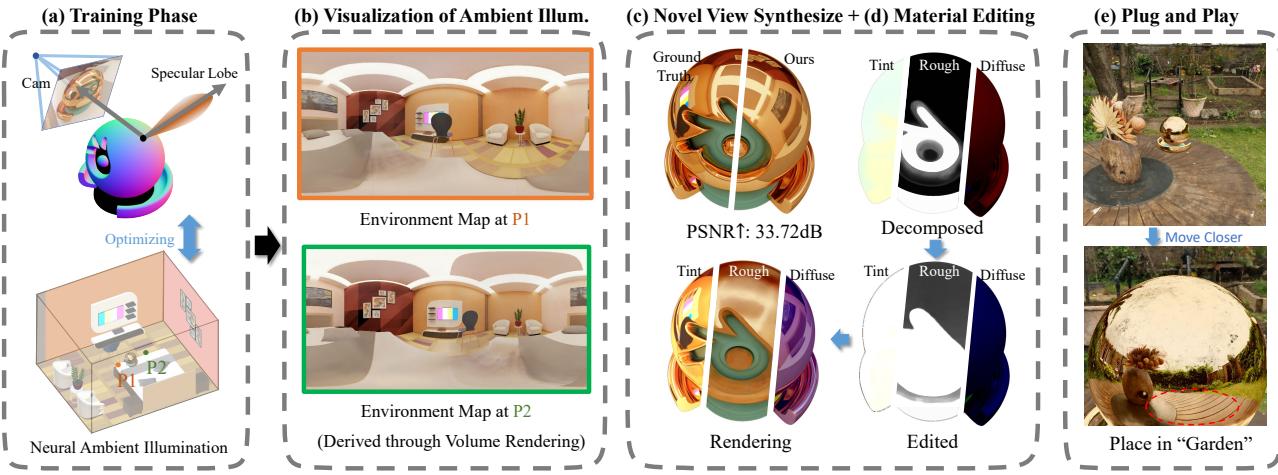


Figure 1. We propose a “plug-and-play” neural ambient illumination (NeAI) that uses the volumetric radiance fields to express the sample of the surrounding 3D environments as a light emitter and re-renders a new object under arbitrary NeRF-style environments naturally. (a) Using a set of images and masks, our method first optimizes the geometry and then jointly optimizes the NeAI and an object’s materials in two stages. (b) Illustrates the environment map at arbitrary samples in 3D scenes via volume rendering, which addresses the environment occlusion and directional lighting naturally. The proposed method provides photo-realistic novel views (c) and visually reasonable material editing (d). More importantly, (e) applies our model to a pre-trained NeRF scene, and produces realistic specular reflections with nice directional illumination. Notably, the Fresnel effect of the table (circle) is exactly reproduced, which is almost impossible for the previous lighting representation.

Abstract

Recent advances in implicit neural representation have demonstrated the ability to recover detailed geometry and material from multi-view images. However, the use of simplified lighting models such as environment maps to represent non-distant illumination, or using a network to fit indirect light modeling without a solid basis, can lead to an undesirable decomposition between lighting and material. To address this, we propose a fully differentiable framework named neural ambient illumination (NeAI) that uses Neural Radiance Fields (NeRF) as a lighting model to handle complex lighting in a physically based way. Together with integral lobe encoding for roughness-adaptive specular lobe and leveraging the pre-convoluted background for accu-

rate decomposition, the proposed method represents a significant step towards integrating physically based rendering into the NeRF representation. The experiments demonstrate the superior performance of novel-view rendering compared to previous works, and the capability to re-render objects under arbitrary NeRF-style environments opens up exciting possibilities for bridging the gap between virtual and real-world scenes. The project and supplementary materials are available at <https://yiyuzhuang.github.io/NeAI/>.

1. Introduction

Modeling and representing the ambient illumination from multi-view images is the fundamental issue that has been extensively studied throughout the development of

*Authors contributed equally to this work.

rendering algorithms in computer vision and graphics[42, 61, 70]. This task is inherently related to materials decomposition, since the observed scene appearance is affected by the interactions between ambient illumination and scene materials. It has drawn significant attention in this era of blowout VR and AR applications, where there is a high demand for photo-realistic rendering of the scene with visually natural illumination in a realistic environment. However, this problem is hard to solve because the ambient illumination is of high-dimensional information and strongly coupled with materials.

Recent methods employ approximated illumination representations (*e.g.*, environment map [69], and spherical Gaussian (SG) models [67, 10, 11, 70]) to simplify the interaction between environment lighting and the object and reduce computational expense. Unfortunately, the assumption used in doing so is that the environmental lighting of the scene is infinitely far away. Neither the environment map nor SG models take the position of 3D environments into account so they are unable to handle environment occlusion and directional lighting practically. To address this problem, NeILF [61] models an incident illumination map for each surface point to handle environment occlusion, indirect light and directional light. However, the constraint of spatial information in lighting is ignored, which is leading to worse material-lighting ambiguity under complex 3D environments. A common follow-up question to ask is: *can we find an elegant representation to express the complex ambient illumination?*

Recent advances in Neural Radiance Fields (NeRF) and its variants [37, 6, 52, 69] have shown great potential to recover underlying scene proprieties (*e.g.* geometry, materials, and lighting) from a set of images. NeRF uses a continuous volumetric function to represent the outgoing ray observed by a viewer. The recent development of NeRF provides new possibilities to model complex ambient illumination. To the best of our knowledge, little in the literature has ever tapped into using NeRF to express lighting. By doing so we could achieve *plug-and-play*: re-render an object with natural illumination of the NeRF-style real-world environment. However, directly gathering thousands of incoming rays through volume rendering to compute the color of each surface point is computationally expensive and may seem impossible.

In this paper, we present the *Neural Ambient Illumination* (NeAI) to express the incoming ray of each surface point as the volumetric radiance fields (density and color) that have the efficient capability of handling environment occlusions and directional lighting naturally, as shown in Fig. 1. The first step of the proposed scheme is to acquire the object’s 3D geometry. Fortunately, state-of-the-art geometry recovery methods[54, 63, 60, 67] can provide almost accurate geometry information for most scenes. So in

this work, we focus solely on optimizing ambient illumination and object material and further application. In particular, the pixel’s specular color is equivalent to the interaction between object materials and the integral of incoming rays within the specular lobe, whose size is related to material roughness. Inspired by environment convolution maps used in traditional image-based lighting, we consider the rough refractive surfaces are inherently related to the convoluted background, although the background is ignored in most methods. Our contributions are summarized as:

- *Neural Ambient Illumination* (NeAI) is proposed to express the radiance field of incoming rays such that treats each sample in the 3D scene as a light emitter.
- A fully *differentiable rendering pipeline* is presented to seamlessly illuminate meshes using a NeRF-style environment.
- *Integrated Lobe Encoding* (ILE) is proposed to feature incoming rays within roughness-adaptive specular lobe to reduce computational cost.
- A *multiscale pre-convoluted representation* for the background is proposed to assist in the decomposition of object materials and ambient illumination.

Extensive experiments and applications demonstrate the superior performance of our method. We strongly recommend browsing our project <https://yiyuzhuang.github.io/NeAI/> for better visualization.

2. Related Work

Ambient Illumination Representation. Ambient illumination representation is an essential component of photo-realistic rendering that aims to provide an immersive experience with visually natural illumination for a variety of view synthesis and relighting applications [20, 58, 18, 8, 30, 10, 49, 70, 61, 21, 11, 69, 67]. Considering that the observed surface appearance in an image is the result of the interactions between ambient illumination and object materials, the ambient illumination is often jointly inferred with object materials from images, also known as inverse rendering [47, 35, 65, 45, 5]. Since inverse rendering is well-known to be severely ill-posed, previous methods usually apply simplified material models [67] and varying lighting conditions [40, 9, 8, 28, 59] to mitigate the problem. In this related work, we will specifically focus on techniques for ambient illumination representations.

The seminal work [15] proposes an omnidirectional radiance map, also known as *environment map*, to represent ambient illumination, which can be applied to render novel objects into the scene realistically. Follow-up methods [56, 20, 5, 51, 48] use the environment map to handle inverse rendering problems naturally. Furthermore, given high-quality geometry, prior works [33, 42] factorize scene

appearance into the diffuse image and the environment map from multi-view images. To be extended beyond a constant term, the environment map is expressed as *spherical Gaussians* (SGs) formulation and integrated its product with surface material BRDF in the same representation to perform illumination calculations [19, 53, 67]. However, it has a significant approximation that light captured by the environment map is emitted infinitely far away.

Considering the illumination from nearly all real-world light sources varies by direction as well as distance, *global illumination representations* (e.g., ray-tracing [38, 49] and path-tracing [4, 66]) are proposed to use ray casting to express the interactions between ambient illumination and surface materials [3]. It is intrinsically described as a ray from a position to determine what objects are in a particular direction. However, global illumination representation is computationally expensive with the pre-computed process, and difficult to reconstruct to 3D real-world illumination for relighting without hand manner. Our work takes inspiration from this line of work in graphics and presents a new ambient illumination representation. We represent the 3D sample of the surrounding environment as a light emitter, such that both position and direction of illumination are taken into account.

Neural Radiance Fields. Neural rendering [37, 63, 31, 62, 54, 41, 10, 13, 57, 14], the task of learning to recover the properties of 3D scenes from observed images, has seen significant success. In particular, Neural Radiance Fields (NeRF) [37] recover the radiance fields (volume density and view-dependent color) of a ray using a continuous volumetric function. Numerous works are extended from NeRF based on its continuous neural volumetric representation for generalizable models [55, 12, 25, 32, 24], dynamic scene [16, 29, 44], non-rigidly deformable objects [50, 43, 71], camera model [23, 34, 22] and acceleration [64, 46, 39, 17]. Recently, Mip-NeRF [6, 7] uses the integral along a cone instead of a ray to recover an anti-aliasing radiance field from a set of multi-scale downsampling images. Besides, Ref-NeRF [52] is proposed for better reflected radiance interpolation. NeRF and its variants have demonstrated remarkable performance in rendering photo-realistic views, but they only model the outgoing radiance of the surface without considering the underlying interaction between ambient illumination and surface material.

Recent advances in differentiable rendering make it possible to reconstruct ambient illumination under casual lighting conditions. Specifically, PhySG [67] and NeRD [10] use SG representation to decompose the scene under complex and unknown illumination. Zhang *et al.* [70] model the indirect illumination via SGs without considering the environment occlusion. Neither the environment map nor SG models take the position of 3D environments into account so they are unable to handle environment occlusion and in-

direct lighting realistically. NeILF [61] proposes the local environment map of each surface point for environment occlusion but ignores the distance of lighting. Overall, recent methods cannot construct a detailed illumination that takes into account both near-field lighting and environment occlusion. We propose neural ambient illumination representation that uses the volumetric radiance fields to express arbitrary illumination in the environment, such that the environment occlusion and directional lighting could be handled naturally.

3. Method

Given a set of posed images of an object captured under static illumination, we learn to decompose the shape, material, and lighting. As previously mentioned, our main objective is the representation of environment lighting and its interaction with the object surface. While the geometry reconstruction is an important aspect of the pipeline, it is *not* the main focus of our work. Thus, we will introduce the core of our pipeline (see Fig. 2) after the object geometry has been reconstructed using external methods in stage one.

3.1. Preliminaries

Prior knowledge of NeRF and physically based rendering such as BRDF is recommended for readers to fully comprehend the modeling presented in this paper.

NeRF. NeRF [37] represents traditional discrete sampled geometry with a continuous volumetric radiance field (*i.e.* density σ and color \mathbf{c}). Given a sampled point \mathbf{x} along a single ray \mathbf{r} originating at \mathbf{o} with direction \mathbf{d} , a positional MLP predicts its corresponding density $\sigma(\mathbf{x})$, and a direction MLP outputs color $\mathbf{c}(\mathbf{x}, \mathbf{d})$ of that point along the ray direction. To render a pixel’s color, NeRF casts a single ray $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$ through that pixel and out into its volumetric representation, accumulates (σ_i, \mathbf{c}_i) into a single color $\mathbf{C}(\mathbf{r})$ of the pixel via numerical quadrature [36],

$$\mathbf{C}(\mathbf{r}) = \sum_i \exp \left(- \sum_{k=0}^{i-1} \sigma_k \right) (1 - \exp(-\sigma_i)) \mathbf{c}_i. \quad (1)$$

The rendering equation. In contrast to NeRF, we replace the pixel’s color of outgoing radiance from a surface point \mathbf{x}_s along a view direction \mathbf{d} with the diffuse color \mathbf{C}_d of that point and an interaction (specular color \mathbf{C}_s) between the incoming radiance \mathbf{L}_{in} of ambient illumination and scene material based on the rendering equation [26],

$$\mathbf{C}(\mathbf{x}_s, \mathbf{d}) = \mathbf{C}_d(\mathbf{x}_s) + \int_{\Omega} f(\mathbf{l}, -\mathbf{d}, \mathbf{x}_s) \mathbf{L}_{in}(\mathbf{x}_s, \mathbf{l})(\mathbf{l} \cdot \mathbf{n}) d\mathbf{l}, \quad (2)$$

where \mathbf{n} and \mathbf{l} denote the normal vector at \mathbf{x}_s and the direction of \mathbf{L}_{in} respectively. The f is the bidirectional reflectance distribution function (BRDF) that focuses on the local reflectance phenomena. Eq. (2) integrate all incoming direction \mathbf{l} on the hemisphere Ω where $\mathbf{l} \cdot \mathbf{n} > 0$.

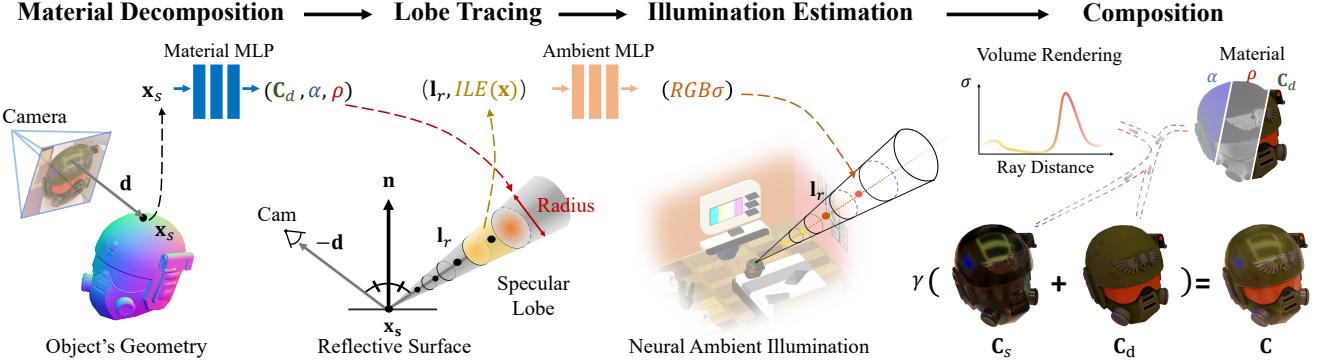


Figure 2. The overview of the proposed method that decomposes materials ($\mathbf{C}_d, \alpha, \rho$) of an object and NeAI radiance fields from a set of images and a reconstructed geometry. For a specific surface point \mathbf{x}_s , NeAI traces its incoming ray within a 3D specular lobe, whose center axis and width are defined by the reflection direction \mathbf{l}_r and roughness ρ . We then feature samples along that lobe with our *integral lobe encoding* (ILE) and feed into the ambient MLP to predict the color and volume density of each sample. Using volume rendering techniques, we integrate these values direction-wise multiplied by the material’s function into the specular color \mathbf{C}_s . We combine this with the diffuse color \mathbf{C}_d provided by the material MLP to render photo-realistic novel views. This rendering procedure is end-to-end differentiable, so we can jointly optimize our NeAI representation and object’s materials.

3.2. Neural Ambient Illumination

Although previous methods attempted to model diverse types of lighting in various ways [67, 69, 10, 11, 70, 61], there still remains a discrepancy between virtual and real-world scenes. However, NeRF has the potential to bridge this gap by enabling the accurate modeling of spatially and directionally varying light.

By expressing the *neural ambient illumination* (NeAI) of an object directly as the continuous volumetric radiance field which includes the volume density and directional emitted radiance at any point in the 3D environment, we can achieve more precise modeling of the light. Given a sample point \mathbf{x} in the environment, we approximate this 5D volumetric radiance field function with the ambient MLP network $\mathcal{R} : (\mathbf{x}, \mathbf{l}) \rightarrow (\mathbf{c}, \sigma)$, where \mathbf{l} is the direction of incoming ray pass through \mathbf{x} . To further integrate the incoming radiance \mathbf{L}_{in} of an incoming ray $\mathbf{r}(t) = \mathbf{x}_s + t\mathbf{l}$ to a surface point \mathbf{x}_s along with the direction \mathbf{l} , we accumulate the corresponding densities and directional emitted colors of $(\mathbf{r}(t), \mathbf{l})$ according to volume rendering,

$$\mathbf{L}_{in}(\mathbf{x}_s, \mathbf{l}) = \int_{t_n}^{t_f} T(t) \sigma(\mathbf{r}(t)) \mathbf{c}(\mathbf{r}(t), \mathbf{l}) dt, \quad (3)$$

where $T(t) = \exp\left(-\int_{t_n}^t \sigma(\mathbf{r}(s)) ds\right)$,

where T indicates the accumulated transmittance. In general, the near and far bounds t_n and t_f are ideally set to infinitely close to zero and infinitely distant, respectively. Eq. (3) indicates that we treat each sample in the 3D environment as a light emitter. This allows the proposed NeAI to model the directional emitted rays and environment occlusions of any static 3D environments.

To define how light derived from a NeRF-style environ-

ment is reflected at an opaque surface, we parameterize the spatially-varying material of surface point \mathbf{x}_s as roughness $\rho \in [0, 1]$, diffuse color $\mathbf{C}_d \in [0, 1]^3$ and specular tint $\alpha \in [0, 1]$, which are output by a material MLP network (using a softplus activation), i.e., $\mathcal{M} : \mathbf{x}_s \rightarrow (\rho, \mathbf{C}_d, \alpha)$. We assume that the BRDF f is rotationally symmetric with respect to the reflection direction \mathbf{l}_r around the specular lobe, such as Phong [3]. \mathbf{l}_r is computed by $2(-\mathbf{d} \cdot \mathbf{n})\mathbf{n} + \mathbf{d}$. Consequently, we approximate the BRDF as von Mises-Fisher (vMF) distribution which is defined on the unit lobe with a normalized spherical Gaussian [3],

$$f(\mathbf{l}, -\mathbf{d}, \mathbf{x}_s) \approx G(\mathbf{l}, \mathbf{l}_r, \mathbf{x}_s) = \alpha \exp(\rho(\mathbf{x}_s)(\mathbf{l} \cdot \mathbf{l}_r - 1)), \quad (4)$$

where \mathbf{l} and \mathbf{l}_r are the unit vector, and the value is positively correlated to $\mathbf{l} \cdot \mathbf{l}_r$. Specifically, \mathbf{l}_r refers to the center axis of the lobe, and spatially-varying roughness $\rho(\mathbf{x}_s)$ controls its angular width (also called the concentration parameter or spread). Noting that, α could be considered as the lobe amplitude, which is learned from the material MLP.

By substituting Eq. (4) and (3) into Eq. (2), we obtain the specular term of Eq. (2) as:

$$\mathbf{C}_s(\mathbf{x}_s, \mathbf{d}) = \iint \alpha e^{\rho(1 \cdot \mathbf{l}_r - 1)} T(\mathbf{x}) \sigma(\mathbf{x}) \mathbf{c}(\mathbf{x}, \mathbf{l})(\mathbf{l} \cdot \mathbf{n}) d\mathbf{x} d\mathbf{l}. \quad (5)$$

According to Eq. (4), a larger ρ value corresponds to a rougher surface with a wider vMF distribution. Therefore, Eq. (5) is equivalent to the integration of the radiance field of each sampled point in the specular lobe defined by the reflection direction. By doing so, we can more effectively and stereographically represent the reflection.

3.3. Integrated Lobe Encoding

To better learn the high-frequency variation of NeAI related to roughness, we introduce a featurized representation,

which we call an *Integrated Lobe Encoding* (ILE), that efficiently and simply constructs positional encoding of all coordinates that lie within specular lobe around \mathbf{l}_r . Our ILE is inspired by IPE used in Mip-NeRF [6], which enables the spatial MLP to represent volume density inside the cone along with view direction. We feature all coordinates inside a roughness-adaptive lobe which considers both the lobe width decided by the roughness and the vMF distribution correlated to the incoming direction.

We could divide the specular lobe of Eq. (5) into a series of conical frustums. Similarly, we approximate this featurized procedure with a set of sinusoids via a multivariate Gaussian. Specifically, we compute the mean and covariance (μ, Σ) of the conical frustum, which is obtained by \mathbf{x}_s and \mathbf{l}_r . Note that the radius variance's part in $\Sigma(\rho)$ is decided by the material roughness of the surface point, which is different with IPE [6]. Given that the integral value of Eq. (5) is under a vMF distribution, it attenuate with the weights of $\mathbf{l} \cdot \mathbf{l}_r$. Our ILE then formulates the encoding of those coordinates (μ, Σ) within the conical frustum,

$$\text{ILE} = \left\{ \begin{bmatrix} \sin(\mu) \exp(-2^{\ell-1}) \text{diag}(\Sigma(\rho)) \\ \cos(\mu) \exp(-2^{\ell-1}) \text{diag}(\Sigma(\rho)) \end{bmatrix} \right\}_{\ell=0}^{L-1}, \quad (6)$$

These features encoded by ILE are used as input to the MLP network \mathcal{R} to output the density and color of our NeAI. This encoding allows the MLP to parameterize the incoming illumination inside the roughness-varying specular lobe, whose strength is variable with the incoming direction under vMF distribution, to behave as an interpolation function. As a result, our ILE efficiently maps continuous input coordinates into a high-frequency space. Please refer to our *supplement* for detail derivation.

According to Eq. (2), the pixel's color \mathbf{C} captured by a camera is equivalent to the diffuse color \mathbf{C}_d of that point and the volumetric integration \mathbf{C}_s of the incoming rays within the specular lobe,

$$\mathbf{C}(\mathbf{x}_s, \mathbf{d}) = \gamma (\mathbf{C}_d(\mathbf{x}_s) + \mathbf{C}_s(\mathbf{x}_s, \mathbf{d})), \quad (7)$$

where γ is a learned HDR-to-LDR mapping. We approximate it as the gamma correction with a learned parameter considering the transformation (*e.g.*, exposure and white balance) learned into the radiance of incoming rays.

3.4. Multiscale Pre-convoluted Representation

Due the complexity of high-dimension representation of NeAI, it's easy to integrate the diffuse color into ambient illumination, causing ambiguous decomposition. Consequently, we propose a multiscale pre-convoluted technique to introduce the background of the object to stabilize the convergence. As shown in Eq. (5), the integral's results of incoming rays within a specular lobe can be considered as convolution, and the width of the lobe is related to the

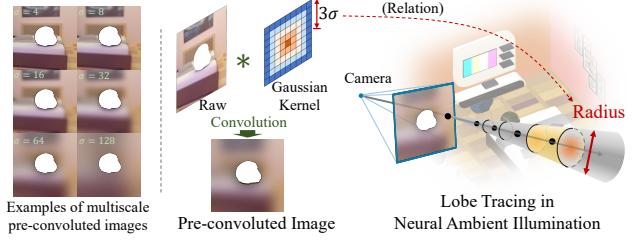


Figure 3. The pipeline demonstrates how pre-convoluted images can be utilized to construct a pre-convoluted representation of illumination.

roughness. As the roughness increases, the ambient illumination is convoluted with more scattered samples within a wider lobe, creating blurrier reflections. By applying a set of different discrete Gaussian blur kernels to the background of the object, we obtain the pre-convoluted results of the incoming rays that correspond to different levels of roughness.

In training phase as shown in Fig. 3, the pixels of the background are evaluated through Eq. (5) and used to supervise the decomposition. We manually set radius $r_p = 3\sigma r_0$, $\alpha = 1$, $\mathbf{C}_d = [0]^3$, where the σ is the variance of the Gaussian kernel and r_0 is the radius of the raw pixel size.

This strategy looks similar to the pre-filtered environment map in CG rendering. However, we originally apply it to the neural rendering framework and take the view direction into account, which allows for more accurate and realistic specular reflections that are consistent with the object's roughness level.

3.5. Loss

To alleviate the ambiguity of the material and illumination, we constrain the roughness $\rho(\mathbf{x}_s)$ and specular tint $\alpha(\mathbf{x}_s)$ of the surface point \mathbf{x}_s to be relatively smooth. With the guidance of the image gradient of pixel \mathbf{p} , we defined the Bilateral Smoothness regularization [61] as:

$$l_s = \frac{1}{|S_I|} \sum_{\mathbf{p} \in S_I} (\|\nabla_{\mathbf{x}_s} \alpha(\mathbf{x}_s)\| + \|\nabla_{\mathbf{x}_s} \rho(\mathbf{x}_s)\|) e^{-\|\nabla_{\mathbf{p}} \mathbf{I}(\mathbf{p})\|}, \quad (8)$$

which forces the material gradient of the surface point \mathbf{x}_s and its projected image pixel \mathbf{p} to be corresponding. The image gradient $\|\nabla_{\mathbf{p}} \mathbf{I}(\mathbf{p})\|$ is pre-computed, and S_I is the set of the pixel on the object.

We compute the L2-norm reconstruction loss between the predicted color $\hat{\mathbf{C}}$ and the ground truth color \mathbf{C} to jointly optimize the ambient illumination and object's materials:

$$l_{rec} = \sum_{\mathbf{p} \in S_I} \|\hat{\mathbf{C}}(\mathbf{x}_s, \mathbf{d}) - \mathbf{C}(\mathbf{p})\|_2^2. \quad (9)$$

The regularization l_{pre} of the pre-convoluted representation is performed in the same manner as the Eq. (9), where the pixel \mathbf{p} is replaced with a pixel from the pre-convoluted

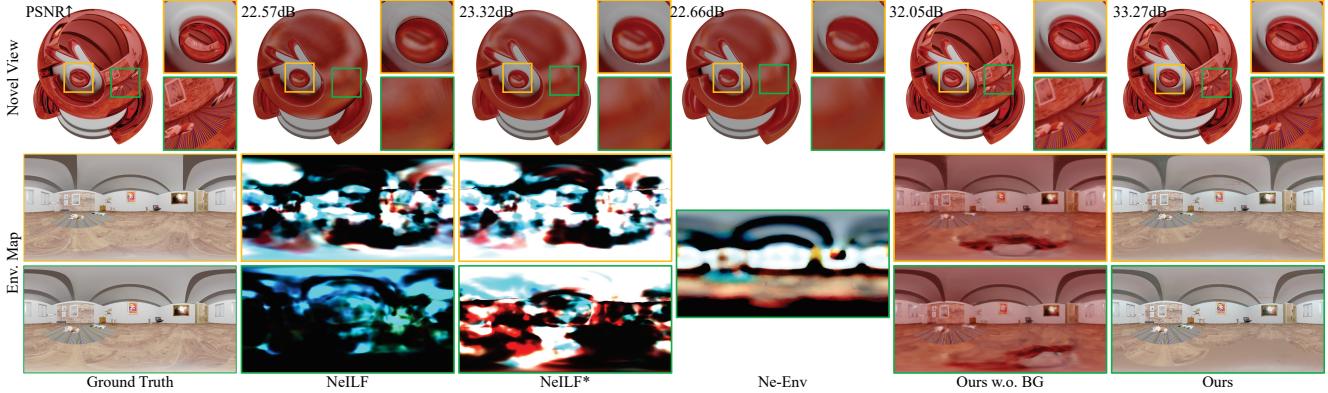


Figure 4. Qualitative comparison of our method against baseline NeILF [61] and its variants on self-rendered Glossy Blender dataset. The zoom-in details labeling with yellow and green boxes and corresponding reconstructed environment map surrounding the surface point of each box center are shown. Our method significantly outperforms other methods, both in the rendering quality of novel views and the reconstruction of ambient illumination. Noted that, although ours w/o background obtains reconstructs geometrically correct results, but fails on the decomposition of diffuse color and ambient illumination due to the incorrect white balance and less background supervision.

background S_B , rather than from the object S_I . Similar to the hierarchical sampling procedure in NeRF, the proposed method also uses “coarse” and “fine” networks for further promising results and sampling efficiency. Overall, the entir loss in our method is $l = l_{rec} + \lambda_s l_s + \lambda_p l_{pre}$, where the weights λ_s and λ_p are empirically set to 10^{-4} and 10^{-1} in all our experiments.

4. Experiment

Implementation Details. Our method is implemented on top of Mip-NeRF [6] with PyTorch, and we discretize the Eq. (5) as Mip-NeRF. The number of samples for both the “coarse” and “fine” phases is 128. We use the same architecture as Mip-NeRF (8 layers, 256 hidden units, ReLU activations) to train our ambient MLP network, but we apply the ILE module to featurize the input coordinates of incoming rays. We also use an 8-layer MLP with a feature size of 512 and a skip connection in the middle to represent the material MLP network. Please refer to our *supplement* for more details about network setting and training schemes.

Through evaluating the results of novel view synthesis, the performance of the decomposition and illumination quality will be measured. We report the image quality with three metrics: PSNR, SSIM and LPIPS [68], on both synthetic and real-world datasets.

Baselines. We compare our method with the following methods: 1) NeILF [61] modeled by the Disney BRDF and incident light field of each surface point; 2) NeILF* extended from NeILF which replaces the Disney BRDF to ours; and 3) Ne-ENV extended from NeILF* that uses a neural environment map instead of the incident light field. These methods using different illumination models could demonstrate the performance of our NeAI. The baseline implementation details could be found in our *supplement*.

Table 1. Quantitative comparison of our method and its ablations against baseline NeILF [61] and its variants on Glossy Blender dataset. Metrics are averaged over all scenes. Our method shows far superior performance in terms of rendering quality.

	PSNR↑	SSIM↑	LPIPS↓
NeILF	25.851	0.930	0.088
NeILF*	27.279	0.941	0.084
Ne-Env	25.971	0.935	0.086
w/o background	36.404	0.979	0.033
w/o pre-convolution	37.748	0.981	0.032
w/o regularization	38.442	0.983	0.029
Ours	37.985	0.981	0.031

Glossy Blender Dataset. Although previous NeRF variants have proposed various datasets containing diverse materials, the objects are all synthesized under the environment map, which ideally ignores the ambient distance. It is unrealistic and unnatural that an image renders without background and neglects the 3D environment. Therefore, we propose a new dataset closer to the natural condition. There are 7 synthetic scenes rendered in Blender [1], each containing one glossy object placed in a 3D simulated environment with natural illumination. In our setting, 390 views are rendered around the upper hemisphere of the object, with 180 for training, 10 for validation, and 200 for testing. We adopt the rendered depth and normal maps as the result of stage one to isolate the influence of geometry quality and create an ideal test for evaluating lighting reconstruction and material decomposition.

Tab. 1 and Fig. 4 demonstrate quantitative and qualitative comparisons of our method against baseline methods, respectively. Metrics in Tab. 1 are averaged over all scenes while the full experiments are accessible in *supplement*. Considering the inputs of NeILF [61] is the masked images of the objects without background, we test our method

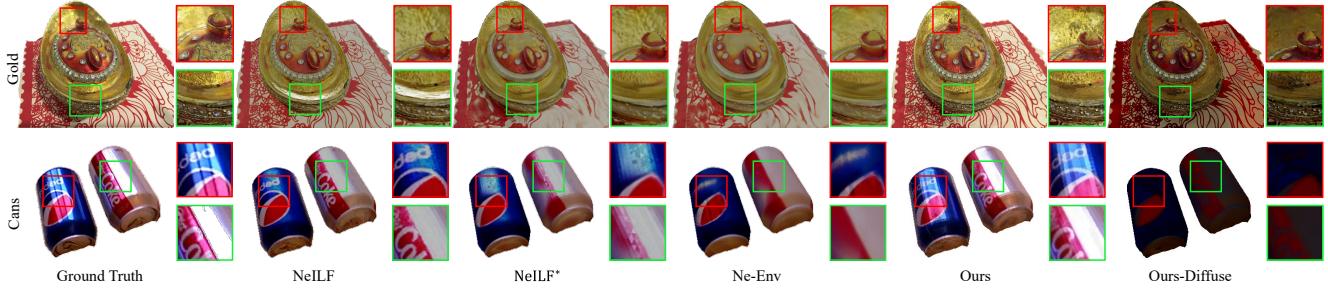


Figure 5. Qualitative comparison with zoom-in details of our method against baselines on real-world [60, 42] dataset. Our method produces the most visually pleasing novel views, especially on reconstructing the highlight regions and detailed textures, as shown in zoom-in details. Note that the decomposed diffuse results of our method are also presented, where the high light is successfully removed.

on the same setting for fairness, referring to our w/o background in the experiments. Besides, another two ablations are contained: w/o pre-convolution, which does not take multiscale pre-convoluted representation into account; w/o regularization, which excludes the Bilateral smoothness regularization (Eq. (8)). Tab. 1 and Fig. 4 show the significantly superior performance of our method compared to baselines in terms of rendering quality on novel views, even without the guidance of background (w/o background).

In particular, as for NeILF and NeILF*, the derived environment maps indicates that they can't handle the nearby illumination, which changes dramatically with the position. Another reason is that they don't gather the information across different views, which is added to the ambiguity. Despite NeILF* simplifies the BRDF function, the same as our method, the performance compared with NeILF is slightly improved, but still has a large gap with our method, as shown in Fig. 4 and Tab. 1. It also verifies the performance of our NeAI on ambient illumination decomposition, especially the nearby lighting and complex 3D environments. As for Ne-Env, we share the same environment map across different positions. That reduces the uncertainty of the solution and therefore constructs a more meaningful environment map. However, the decomposition of materials and ambient illumination is poor. The reason is the simplification of the illumination model, *i.e.*, all radiance incident upon the object being shaded comes from an infinite distance. Compared to NeILF and its variants, our method could render photo-realistic views and recover precise ambient illumination. Even without the background guidance, our method still reconstructs geometrically correct results. However, the white balance is not correct due to the majority color of the object is red, which causes ambiguity about the red object or the red incident light. This side-fact indicates it's necessary to use the background as guidance to disentangle the material and light.

Real-world. We then test our method on 8 real-world scenes from BlendedMVS [60] and Bag of Chips [42], which provide images and depth maps reconstructed by MVS methods or RGBD camera. We selected the last ten

Table 2. Quantitative comparison of our method against baselines on Real-world Datasets. The table shows PSNR metrics of novel view renderings of test images. Our method consistently outperforms the other methods in terms of rendering quality.

	Vase	Cam	Stone	Gold	Conch	Bull	Cans	chips
NeILF	26.270	17.182	23.607	18.458	19.327	19.437	26.408	22.630
NeILF*	25.225	19.205	22.269	17.369	19.272	19.337	25.565	21.571
Ne-Env	24.564	18.668	21.378	17.511	19.347	19.354	23.560	21.906
Ours	28.927	21.864	24.473	20.549	20.557	22.865	31.574	29.366



Figure 6. Our model decomposes the roughness and diffuse color from ambient lighting that enables editing. We can manipulate the roughness of by multiplying a global scale, which demonstrates the specular reflections naturally. We can also edit the diffuse color without affecting its glossy paint's specular reflections.

images as the test set for BlendedMVS, while for Bag Of Chips, we left the last 1/5 as the test set. The qualitative and quantitative comparisons are shown in Tab. 2 and Fig. 5 respectively. All the results validate the performance of our method on illumination reconstruction and material decomposition, representing the robustness of our method in geometric perturbation. Compared to baselines, the performance gap degrades for two reasons: 1) the images with varying exposures and rough geometry in real-world datasets make the decomposition difficult; 2) not enough background in original images could be used to help the ambient illumination reconstruction.

5. Applications

Object's material editing. As our model disentangles the object's material well, our components behave intuitively and enable visually reasonable material editing results. Fig. 6 shows material decomposition (*e.g.* roughness and diffuse color) and natural example edits of the object's ma-

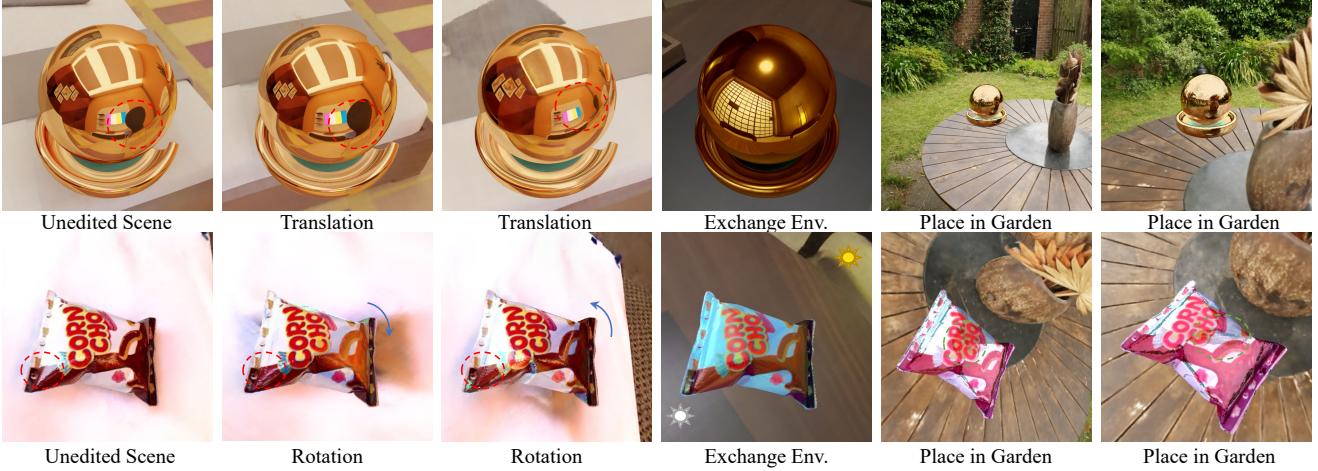


Figure 7. Visualizations of illumination manipulations. Our method recovers environment lighting that enables novel view rendering with high-fidelity specularities and natural illumination under different manipulations. Our method naturally handles the occlusion of a chair in translation cases and the high lighting on the cornchos in rotation cases. We place our decomposed objects in a virtual environment or pre-trained Mip-NeRF environment and produce novel views with more realistic reflections (**Better viewed on screen with zooming in**).



Figure 8. Visualizations of the roughness editing. The first row shows the corresponding environment maps for the first three balls, and the second row shows that the ball gradually becomes blurred with increasing roughness. Noted that, although the roughness is beyond the Gaussian kernel used in the pre-convolved representation, our method could also render visually realistic results.

terials on the ‘Car’ dataset. Furthermore, Fig. 8 shows convincing results of roughness editing (second row) and their corresponding environment maps (first row) on the ‘Metal Ball’ dataset. As the roughness increases, novel views gradually become blurred, even when the roughness becomes extremely large that exceeds the maximum scale of pre-convolved representation.

Illumination manipulation. Fig. 7 illustrates several types (rotation, translation, exchange) of manipulation to the ambient illumination around the object. Our method produces natural photo-realistic views, especially for the environment occlusions (*e.g.* reflected chair) and highlights (*e.g.* chips) as shown in the translation and rotation cases respectively. More importantly, with a slight scale to the decomposed materials, we place our plug-and-play NeAI model in a pre-trained Mip-NeRF environment [7], as shown in the last two columns of Fig. 7. Although there is complex and detailed illumination, visually harmonious novel views are re-rendered with more realistic reflections in the “Garden”

dataset. It verifies that our NeAI is an easy-to-use plugin for NeRF that gives objects a better sense of belonging in the new 3D NeRF-style environments. Please refer to our supplement video for better visualization.

6. Conclusion

This paper presents a novel neural approach to efficiently and stereographically modeling 3D ambient illumination. Previous methods focus on simplified lighting models (*e.g.* environment map and spherical Gaussian) to represent no-distant illumination. Instead, we propose NeAI to model illumination as volumetric radiance fields such that each sample of the surrounding 3D environments is equivalent to a light emitter. We show that, together with our integral lobe encoding and pre-convolved representation, our method can accurately recover ambient illumination and naturally re-render high-quality views for a decomposed object under new NeRF-style environments. We believe that with this high-fidelity and fully differentiable lighting representation, it can be easily extended to downstream tasks and bring us closer to bridging the gap between virtual and real-world scenes.

Limitations. It is difficult to model ambient illumination and decompose the object’s material, our approach has the following limitations. First, our pipeline relies on the geometry reconstructed through stage one, similar to [61, 69]. Second, most illumination is HDR for better harmonization. Although we use a gamma correction to approximate the LDR-HDR mapping, our method could use a similar tone-mapping module in HDR-NeRF [23] to recover HDR-NeAI. We leave those problems in future works.

Supplementary Material

A. Overview

The supplementary material provides the derivation of integral lobe encoding (ILE, Eq. (6) in the submission) (Section B), the detailed architecture of our method, and the implementation details of the baseline methods (Section C). We also introduce more details of our plug-and-play rendering process. The additional experiments demonstrate the superior performance of our method compared to the baseline methods (Section D). We **strongly** encourage the reader to see our website-video supplementary <https://yiyuzhuang.github.io/NeAI/>, in which we present our results on various applications, especially our plug and play: place meshes in the public 360v2 datasets* [7].

B. Derivation of Integral Lobe Encoding

Here, we introduce our ILE featurization procedure, in which we cast a lobe whose size is defined by the spatially-varying roughness. We feature conical frustums along that roughness-varying lobe. Similar to Mip-NeRF [6], images in our method are rendered one pixel at a time so that we can describe our procedure in terms of an individual pixel of interest being rendered. Mip-NeRF casts a constant cone from the camera’s center along the view direction that passes through the pixel’s center, where the cone’s size is spatially-constant and defined by the width of the pixel in world coordinates. As shown in Eq. (2) of the submission, we instead decompose the pixel’s color into diffuse color and specular color of surface point \mathbf{x}_s , where \mathbf{x}_s is the intersection between the object surface and the ray from the camera’s center along the view direction. As shown in Eq. (5) of the submission, the specular color is the integration of the radiance field of each sampled point within the specular lobe combined with surface material. The apex of that specular lobe lies at \mathbf{x}_s , and the radius of the specular lobe is defined by $\hat{\rho}\hat{r}$. According to the pre-filtered environment map used in traditional CG rendering [3], we set $\hat{\rho}$ to the material roughness $[0, 1]$ scaled by the pixel’s size r_0 , and set \hat{r} to the width of the pixel r_0 scaled by mipmap levels (*i.e.* resolution multi-scaling). In the implementation, we discretely divide the specular lobe into several conical frustums according to the discrete distance. The set of positions \mathbf{x} that lie within a conical frustum between two distant value $[t_0, t_1]$ is then defined as,

$$F(\mathbf{x}, \mathbf{x}_s, \mathbf{l}_r, \rho, \hat{r}, t_0, t_1) = \mathbb{I} \left\{ \begin{array}{l} \left(t_0 < \frac{\mathbf{l}_r^\top (\mathbf{x} - \mathbf{x}_s)}{\|\mathbf{l}_r\|_2^2} < t_1 \right) \\ \wedge \left(\frac{\mathbf{l}_r^\top (\mathbf{x} - \mathbf{x}_s)}{\|\mathbf{l}_r\|_2 \|\mathbf{x} - \mathbf{x}_s\|_2} > \frac{1}{\sqrt{1 + (\hat{\rho}\hat{r}/\|\mathbf{l}_r\|_2)^2}} \right) \end{array} \right\}, \quad (10)$$

*<https://jonbarron.info/mipnerf360/>

where $\mathbb{I}\{\cdot\}$ is an indicator function: $F(\mathbf{x}, \cdot) = 1$ iff \mathbf{x} is within the conical frustum defined by $(\mathbf{x}_s, \mathbf{l}_r, \rho, \hat{r}, t_0, t_1)$. $\|\cdot\|_2$ refers to the Euclidean norm of a vector. Note that the conical frustum’s radius is defined not only by the pixel width but also by the roughness, which means the roughness-varying specular lobe instead of constant.

Besides, the success of NeRF in high-fidelity rendering results is attributed to the proposed position encoding for learning high-frequency information. Each sample point \mathbf{x} goes through this featurization procedure before being fed into the MLP, which is formulated as,

$$\text{PE}(\mathbf{x}) = \left[\sin(\mathbf{x}), \cos(\mathbf{x}), \dots, \sin(2^{L-1}\mathbf{x}), \cos(2^{L-1}\mathbf{x}) \right]^T, \quad (11)$$

where L is a hyperparameter and determines the bandwidth of the interpolation kernel. This encoding enables the MLP parameterizing the scene to behave as an interpolation function, and maps the position into a high-frequency space. Consequently, we establish a featurized representation of Eq. (11) inside the conical frustum defined by Eq. (10),

$$\begin{aligned} \text{ILE}(\mathbf{x}_s, \mathbf{l}_r, p_l) &= \frac{\int A(\mathbf{x}, \mathbf{x}_s, \mathbf{l}_r) \text{PE}(\mathbf{x}) F(\mathbf{x}, \mathbf{x}_s, \mathbf{l}_r, p_l) d\mathbf{x}}{\int F(\mathbf{x}, \mathbf{x}_s, \mathbf{l}_r, p_l) d\mathbf{x}}, \\ \text{where } A(\mathbf{x}, \mathbf{x}_s, \mathbf{l}_r) &= \exp \left(\frac{\mathbf{l}_r^\top (\mathbf{x} - \mathbf{x}_s)}{\|\mathbf{l}_r\|_2 \|\mathbf{x} - \mathbf{x}_s\|_2} - 1 \right) \end{aligned} \quad (12)$$

where $p_l = \{\hat{\rho}, \hat{r}, t_0, t_1\}$ indicates the shape parameters about the conical frustum. According to Eq. (5) in the submission, a larger roughness corresponds to a rougher surface under a wider vMF distribution, and the integral volume is attenuated with the angle between incoming direction $\mathbf{l} = \mathbf{x} - \mathbf{x}_s$ and the reflection direction \mathbf{l}_r of the ambient illumination. Consequently, we consider the ILE must be under a vMF distribution, and it attenuates with the weights of $\mathbf{l}_r^\top (\mathbf{x} - \mathbf{x}_s)$. We then set the $A(\mathbf{x}, \mathbf{x}_s, \mathbf{l}_r)$ to the attenuation function, which can be well-approximated using a simple exponential function. Considering attenuation function A is related to the direction of incoming illumination under a vMF distribution, the integral in the numerator can be defined via a multivariate Gaussian to efficiently represent the vMF attenuation and compute that features one time. According to integral position encoding (IPE) in Mip-NeRF [6], the integration within a conical frustum in the numerator has no closed-form solution, only Gaussian approximation. However, this approximation will be inaccurate if there is a significant difference between the center and the boundary of the frustum. Compared to IPE, our ILE is the vMF attenuation, which is more suitable and accurate to describe as a multivariate Gaussian expression.

Eq. (12) can be directly simplified as a multivariate Gaussian. We first represent the shape of the conical frustum by the mean distance along the ray μ_t , the variance along the ray σ_t^2 , and the variance perpendicular to the ray

σ_r^2 . According to a midpoint $t_\mu = (t_0 + t_1)/2$ and a half-width $t_\sigma = (t_1 - t_0)/2$, these variables can be parameterized as,

$$\begin{aligned} \mu_t &= t_\mu + \frac{2t_\mu t_\delta^2}{3t_\mu^2 + t_\delta^2}, \sigma_t^2 = \frac{t_\sigma^2}{3} - \frac{4t_\delta^4(12t_\mu^2 - t_\sigma^2)}{15(3t_\mu^2 + t_\delta^2)^2}, \\ \sigma_r^2(\hat{\rho}) &= \hat{\rho}^2 \hat{r}^2 \left(\frac{t_\mu^2}{4} + \frac{5t_\sigma^2}{12} - \frac{4t_\sigma^4}{15(3t_\mu^2 + t_\delta^2)} \right), \end{aligned} \quad (13)$$

where the width of conical frustum $\sigma_r^2(\rho)$ is spatially varied with the roughness of surface point.

We then compute the shape of a set of conical frustums along the specular lobe. We transform Eq. (13) from the coordinate frame of conical frustums into the world coordinates, of which the mid-point μ and variance Σ are defined as follows,

$$\mu = \mathbf{x}_s + \mu_t \mathbf{l}_r, \Sigma(\hat{\rho}) = \sigma_t^2 \left(\mathbf{l}_r \mathbf{l}_r^\top \right) + \sigma_r^2(\hat{\rho}) \left(\mathbf{I} - \frac{\mathbf{l}_r \mathbf{l}_r^\top}{\|\mathbf{l}_r\|_2^2} \right), \quad (14)$$

where the variance Σ is related to spatially-varying roughness ρ . The next derivation is similar to Mip-NeRF [6], Eq. (11) can be rewritten as matrix form ,

$$\text{PE}(\mathbf{x}) = [\sin(\mathbf{P}\mathbf{x}), \cos(\mathbf{P}\mathbf{x})]^\top \quad (15)$$

$$\mathbf{P} = \begin{bmatrix} 1 & 0 & 0 & 2 & 0 & 0 & \dots & 2^{L-1} & 0 & 0 \\ 0 & 1 & 0 & 0 & 2 & 0 & \dots & 0 & 2^{L-1} & 0 \\ 0 & 0 & 1 & 0 & 0 & 2 & \dots & 0 & 0 & 2^{L-1} \end{bmatrix}^\top \quad (16)$$

This matrix form could help us to derive the final ILE feature. According to Eq. (16), we could identify the mean and covariance of the conical frustum Gaussian along the specular lobe. Our ILE feature can be finally formulated as the expected sines and cosines of the mean and the diagonal of the covariance matrix:

$$\begin{aligned} \text{ILE}(\mathbf{x}_s, \mathbf{l}_r, p_l) &= E_{x \sim N(\mu, \Sigma(\hat{\rho}))}[\text{PE}(x)] \\ &= \begin{bmatrix} \sin(\mathbf{P}\mu) \odot \exp(-\frac{1}{2} \text{diag}(\mathbf{P}\Sigma(\hat{\rho})\mathbf{P}^T)) \\ \cos(\mathbf{P}\mu) \odot \exp(-\frac{1}{2} \text{diag}(\mathbf{P}\Sigma(\hat{\rho})\mathbf{P}^T)) \end{bmatrix}, \end{aligned} \quad (17)$$

where \odot refers to element-wise multiplication. For better understanding, we formulate Eq. (17) as

$$\text{ILE}(\mathbf{x}_s, \mathbf{l}_r, p_l) = \left\{ \begin{bmatrix} \sin(2^\ell \mu) \exp(-2^{\ell-1} \text{diag}(\Sigma(\hat{\rho}))) \\ \cos(2^\ell \mu) \exp(-2^{\ell-1} \text{diag}(\Sigma(\hat{\rho}))) \end{bmatrix} \right\}_{\ell=0}^{L-1}, \quad (18)$$

$$\text{diag}(\Sigma(\hat{\rho})) = \sigma_t^2 (\mathbf{l}_r \odot \mathbf{l}_r) + \sigma_r^2(\hat{\rho}) \left(\mathbf{1} - \frac{\mathbf{l}_r \odot \mathbf{l}_r}{\|\mathbf{l}_r\|_2^2} \right) \quad (19)$$

where L is a hyperparameter and determines the bandwidth of the interpolation kernel. $p_l = \{\hat{\rho}, \hat{r}, t_0, t_1\}$ refers to the shape parameters about the conical frustum. Our ILE allows the MLP to parameterize the incoming illumination

inside the roughness-varying specular lobe, whose strength is variable with the incoming direction under vMF distribution, to behave as an interpolation function. As a result, our ILE efficiently maps continuous input coordinates into a high-frequency space.

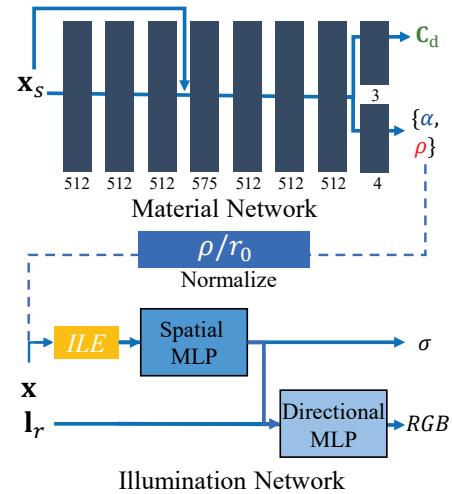


Figure 9. Details of the network architecture. As for the material network, a ReLU is applied for each MLP except the last layer. We choose tanh and sigmoid to the output of \mathbf{C}_d and α, ρ , separately. The parameters of illumination network is the same as [6].

C. Implement Details



Figure 10. Visualisations of the results on our "Glossy Blender" dataset. It's obvious that the reflection and the highlights are correctly reconstructed with our method. Compared to NeILF or other baselines, it shows great robustness in representing complex lighting.

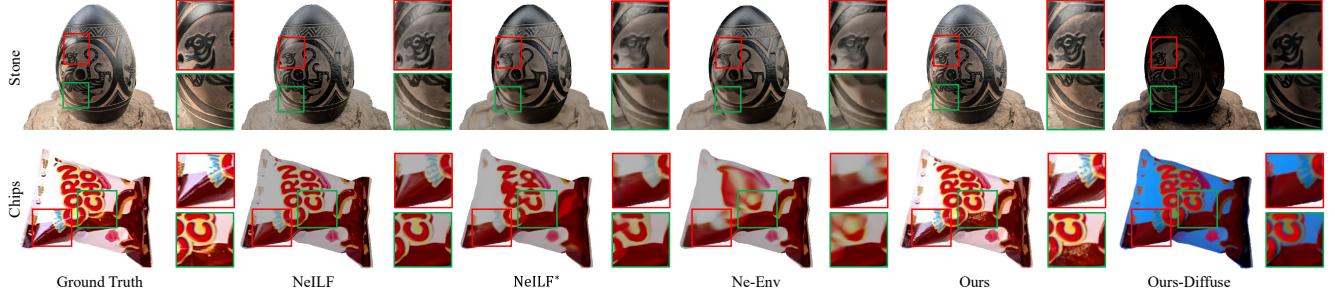


Figure 11. More qualitative comparison with zoom-in details of our method against baselines on real-world [60, 42] dataset. Our method produces the most visually pleasing novel views, especially on reconstructing the highlight regions and detailed textures, as shown in zoom-in details. Note that the decomposed diffuse results of our method are also presented, where the highlights are successfully removed. The diffuse color of our method looks wrong on white balance, probably because lacking of enough background information.

C.1. Architecture

In the paper, we introduce our novel representation of illumination, and the rendering equation for the pixel of the object is given as Eq. (7). The whole pipeline is differentiable, and the architecture of the network is shown in Fig. 9. We believe the architecture can be easily re-implemented or extended by other researchers.

For the pixel of the background, we remove the diffuse term \mathbf{C}_d and the tint α , and directly give the value of radius $r_p = 3\sigma r_0$. σ is set according to the pre-convoluted image. It is worth noting that, different with traditional pre-filtered technique [3], our method does not handle the images with different mipmap levels (multiscale resolution) for simplification, as shown in Fig. 3, so \hat{r} is set to the raw pixel's size r_0 . According to Eq. (3) and ILE Eq. (18), the captured value of along the ray $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$ from t_n to t_f is:

$$\mathbf{C}(\mathbf{o}, \mathbf{d}) = \int_{t_n}^{t_f} T(t)\sigma(\mathcal{L})\mathbf{C}(\mathcal{L}, \mathbf{d})dt, \quad (20)$$

where $\mathcal{L} = \text{ILE}(\mathbf{r}(t), \mathbf{d}, p_l)$,

and $p_l = \{3\sigma, r_0, t_n, t_f\}$ refers to the shape parameters about the conical frustum along that ray.

C.2. More about Plug and Play

With our novel representation of environment lighting and its interaction with the object surface, we lay the foundation for a fancy application: seamless illumination of a decomposed mesh with a NeRF-style environment. For a decomposed object from stage two, we need to manually set the transformation to put it into the environment. Sequentially, the realistic results can be rendered through the rendering equations shown in Eq. (7) of the submission and Eq. (20). Considering that the object is sometimes occluded by the landscape, we modify the rendering Eq.(7) slightly to account for occlusions during the rendering of new views,

$$\mathbf{C}(\mathbf{o}, \mathbf{d}) = \mathbf{C}^*(\mathbf{o}, \mathbf{d})w^*(\mathbf{o}, \mathbf{d}) + \mathbf{C}(\mathbf{x}_s, \mathbf{d})(1 - w^*(\mathbf{o}, \mathbf{d})), \quad (21)$$

where $\mathbf{C}^*(\cdot)$ and $w^*(\cdot)$ with superscript * mean that the t_f is set to 0.99 times the depth of the object's surface. $w^*(\cdot)$ is the cumulative distribution function of density from t_n to t_f along the ray. Additionally, it is recommended to scale the decomposed material $\alpha, \rho, \mathbf{C}_d$ for better visual effect. The rendering videos could be found in *supplement website*.

It is worth noting that, with our rendering pipeline, our method takes about 50 seconds to render an 800×800 image on a single Tesla A100, while the Cycles engine on Blender [1] takes about several minutes to render the ground truth.

C.3. Training Details

Our implementation is based on the re-implementation pytorch [2] of Mip-NeRF [6]. We train our model with Adam [27] for 1 million iterations. For each iteration, 4096 rays are randomly selected from all images including the pre-convoluted images. The training process takes about two days to train on a single Tesla A100 GPU or 16 hours for 600,000 iterations on 4 Tesla A100 GPUs.

C.4. Baseline Implementations

NeILF. [61] We follow the same setup in the author's code[†]. The architecture can be divided into two parts, one for material estimation, called BRDF MLP, and the other, NEILF MLP. We follow their procedure for training on a single scene.

NeILF*. We replace the BRDF MLP with our material network and modify the rendering equation to directly predict the diffuse term as \mathbf{C}_d . Another change is to modify the Fresnel term F to be a position-dependent magnitude α , so the specular term is only related to roughness ρ , magnitude α . The predicted \mathbf{C}_s is finally obtained by integrating with the product of all incident light and the specular term on the hemisphere Ω .

Ne-Env. A variant of NeILF*, named neural environment map. The position point input of NEILF MLP is removed,

[†]<https://github.com/apple/ml-neilf>

so the spatially-varying environment map is modified to the non-spatially varying environment map.

w/o ILE. The roughness ρ is set to a small constant as r_0 , which means the adaptive lobe encoding is abolished, and the specular lobe degenerates into a thin ray. It causes artifacts to the objects with rough material.

D. Additional Results

D.1. Glossy Blender dataset

The error metrics for each individual scene are provided in Tabs. 5, 6 and 7, and the more visualized results are shown in Fig. 10. It demonstrates the significantly superior performance of our lighting representation compared to the baselines through rendering quality on novel views.

We also show visual comparison of the ablation in Fig. 12. We can see that, without the Bilateral Smoothness regularization (Eq. (8) in our submission), although obtains the best metric performance, it generates noise and becomes inconsistent across the same material. It is unsuitable for material editing applications. On the other hand, without the ILE, spatially-varying roughness can not be handled very well, resulting in artifacts of diffuse items, as shown in Fig. 12. The result of ‘w/o BG’ shows that without the assistance of the background, it’s easy to creates ambiguity between light and material.



Figure 12. Visualizations of the decomposed material for ablation study. Without the Bilateral Smoothness regularization, the material becomes noisy. Without the ILE, changes in material roughness cannot be handled well, resulting in artifacts of diffuse items. Without background, the roughness is not correct due to the effects of lighting.

D.2. Real-world dataset

Tabs. 3 and 4 list the average SSIM and LPIPS for each individual scene in real-world dataset [60, 42], and more visualization is shown in Fig. 11. It demonstrates the performance of our method in real scenes compared to NeILF and its variants in terms of rendering quality on novel views.

Table 3. Per-Scene quantitative comparison of SSIMs on real-world dataset.

	bull	cam	stone	gold	conch	vase	cans	chips
NeILF	0.779	0.460	0.859	0.771	0.787	0.827	0.957	0.939
NeILF*	0.725	0.573	0.801	0.625	0.744	0.745	0.949	0.917
Ne-Env	0.745	0.567	0.801	0.633	0.750	0.738	0.944	0.915
Ours	0.760	0.705	0.862	0.769	0.655	0.851	0.960	0.953

Table 4. Per-Scene quantitative comparison of LPIPS on real-world dataset.

	bull	cam	stone	gold	conch	vase	cans	chips
NeILF	0.272	0.437	0.131	0.173	0.313	0.213	0.034	0.051
NeILF*	0.318	0.345	0.172	0.297	0.382	0.301	0.044	0.076
Ne-Env	0.324	0.355	0.187	0.311	0.391	0.311	0.050	0.082
Ours	0.257	0.256	0.130	0.186	0.366	0.172	0.028	0.041

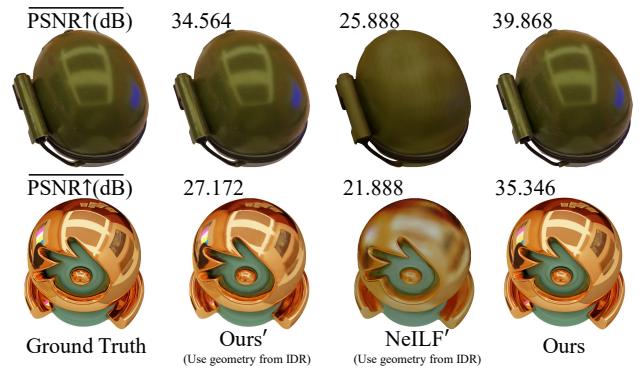


Figure 13. Qualitative and quantitative results based on a geometry reconstructed by IDR[63] (in column 2,3). Compared to the result of the ideal test (in column 4), it is worth noting that the geometric deviation from the stage one would affects the quality of novel-view synthesis. This is because the tracing direction and origin of the specular lobe depends on the geometry. For a rougher surface, the lobe is wider, so is has more tolerance to this perturbation. The opposite is true for a smoother surface. We demonstrate this effect with two examples of different roughness.

D.3. Discussion about the geometric perturbation.

Previous studies have shown that the decomposition between light and material relies on the geometric quality [42, 61, 70]. In Section 4 of the main text, we adopt the rendered depth and normal maps as the result of stage one to isolate the influence of geometry quality. This means that the quality of novel-view synthesis will only reflect the lighting reconstruction and material decomposition. However, in practice, there is always some deviation in geometry. Few studies have discussed the impact and underlying reason for this dependency. Through the experiment, we find that the decomposition of the mirror-like surfaces is more sensitive to the geometric perturbation, as shown in Fig. 13. This because our physically based lobe tracing relies on the reconstructed geometry to calculate the incoming direction. If the estimated geometry deviates from

Table 5. Per-Scene quantitative comparison of PSNRs on Glossy Blender dataset.

	Car Playroom	Helmet Bedroom	Coffee Tearoom	RedBall Playroom	GoldBall Bedroom	GoldBall Bedroom2	MetalBall Tearoom
NeILF	20.863	28.282	26.785	26.630	24.889	24.620	28.886
NeILF*	21.332	29.419	27.036	27.444	25.383	25.175	29.216
Ne-Env	21.326	29.228	26.953	26.777	24.682	24.542	28.287
w/o ILE	35.090	38.843	41.407	38.046	35.489	34.544	38.116
w/o background	35.131	39.835	40.496	37.649	35.562	34.893	31.263
w/o pre-convolution	35.072	39.608	41.559	38.567	35.987	35.110	38.336
w/o regularization	35.983	41.439	42.167	39.004	36.429	35.535	38.540
Ours	35.296	39.868	41.735	38.840	36.301	35.346	38.506

Table 6. Per-Scene quantitative comparison of SSIMs on Glossy Blender dataset.

	Car Playroom	Helmet Bedroom	Coffee Tearoom	RedBall Playroom	GoldBall Bedroom	GoldBall Bedroom2	MetalBall Tearoom
NeILF	0.908	0.942	0.966	0.917	0.917	0.913	0.947
NeILF*	0.926	0.955	0.969	0.924	0.924	0.922	0.950
Ne-Env	0.926	0.954	0.969	0.918	0.917	0.915	0.946
w/o ILE	0.970	0.981	0.982	0.982	0.981	0.978	0.981
w/o background	0.970	0.988	0.981	0.980	0.982	0.981	0.971
w/o pre-convolution	0.969	0.985	0.982	0.985	0.982	0.980	0.982
w/o regularization	0.973	0.990	0.983	0.986	0.984	0.982	0.982
Ours	0.968	0.985	0.981	0.986	0.983	0.981	0.982

Table 7. Per-Scene quantitative comparison of LPIPS on Glossy Blender dataset.

	Car Playroom	Helmet Bedroom	Coffee Tearoom	RedBall Playroom	GoldBall Bedroom	GoldBall Bedroom2	MetalBall Tearoom
NeILF	0.115	0.081	0.056	0.096	0.099	0.108	0.057
NeILF*	0.109	0.073	0.062	0.094	0.095	0.098	0.060
Ne-Env	0.110	0.074	0.063	0.095	0.098	0.102	0.062
w/o ILE	0.049	0.047	0.045	0.025	0.025	0.027	0.030
w/o background	0.048	0.027	0.048	0.026	0.019	0.021	0.038
w/o pre-convolution	0.054	0.034	0.046	0.019	0.021	0.023	0.030
w/o regularization	0.047	0.025	0.045	0.017	0.019	0.020	0.030
Ours	0.057	0.036	0.046	0.018	0.021	0.021	0.030

the true geometry, different tracing lobes may conflict with each other, which becomes more serious when the lobe is thinner. Despite this challenge, our method still produces better results of novel-view synthesis than existing methods, which demonstrates its robustness to geometric perturbation. Moreover, we suggest that this error could be potentially exploited to reconstruct a more accurate geometry in the future work.

D.4. Dataset License

We overview each scene in Fig. 14. Our new “Glossy Blender” scenes were adapted from the following Sketchfab <https://sketchfab.com/feed> models, including object models and ambient illumination models.

The object models are:

1. Car: created by DAVID.3D.ART, CC Attribution-

NonCommercial license (model *AC-Bmw 1m*);

2. Helmet: released by public NeILF dataset [61];
3. Coffee: created by Elen, CC Attribution-ShareAlike license (model *Coffee cup with plate*);
4. GoldBall & RedBall & MetalBall: released by public NeRF blender dataset [37] (model *materials dataset*).

The ambient illumination models are:

1. Tearoom: created by Anex, CC Attribution license (model *Room*);
2. Playroom: created by enteRenders, CC Attribution license (model *Interior Living Room Corner*);
3. Bedroom & Bedroom2: created by ankitk2618, CC Attribution license (model *Bedroom Interior*).

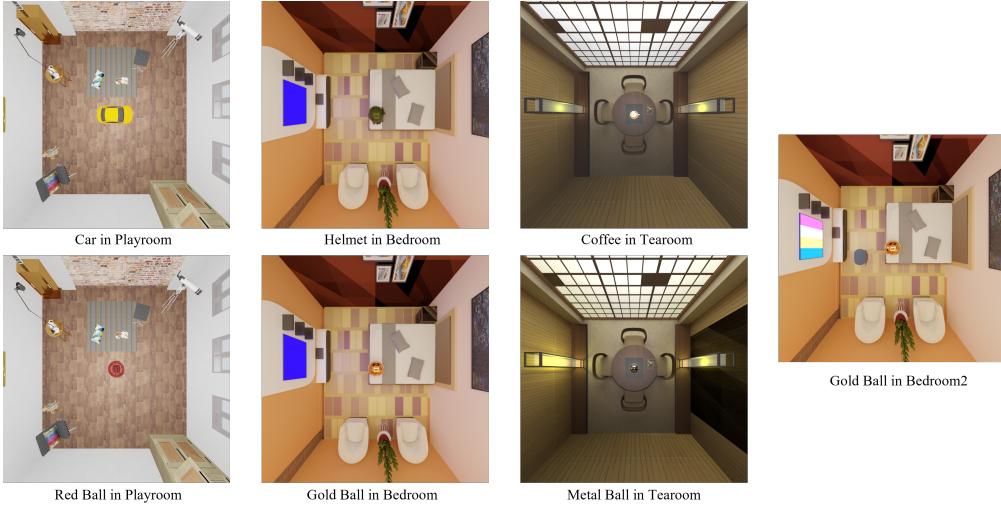


Figure 14. Visualizations of our new “Glossy Blender” scenes. Each one is rendered by a camera under the ceiling with a focal length of around 12mm.

We have slightly remixed, transformed, and built upon the material of models for better visual harmonization between objects and ambient illuminations.

Acknowledgements. This work was supported by the NSFC grant 62001213, 62025108, and a gift funding from Tencent Rhino-Bird Research Program.

References

- [1] Blender. <https://www.blender.org/>. 6, 11
- [2] Pytorch. <https://pytorch.org/>. 11
- [3] Tomas Akenine-Moller, Eric Haines, and Naty Hoffman. *Real-time rendering*. AK Peters/crc Press, 2019. 3, 4, 9, 11
- [4] Dejan Azinovic, Tzu-Mao Li, Anton Kaplanyan, and Matthias Nießner. Inverse path tracing for joint material and lighting estimation. In *CVPR*, pages 2447–2456, 2019. 3
- [5] 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
- [6] 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*, pages 5855–5864, 2021. 2, 3, 5, 6, 9, 10, 11
- [7] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *CVPR*, pages 5470–5479, 2022. 3, 8, 9
- [8] Sai Bi, Zexiang Xu, Kalyan Sunkavalli, Miloš Hašan, Yannick Hold-Geoffroy, David Kriegman, and Ravi Ramamoorthi. Deep reflectance volumes: Relightable reconstructions from multi-view photometric images. In *ECCV*, pages 294–311. Springer, 2020. 2
- [9] Sai Bi, Zexiang Xu, Kalyan Sunkavalli, David Kriegman, and Ravi Ramamoorthi. Deep 3d capture: Geometry and reflectance from sparse multi-view images. In *CVPR*, pages 5960–5969, 2020. 2
- [10] Mark Boss, Raphael Braun, Varun Jampani, Jonathan T Barron, Ce Liu, and Hendrik Lensch. NeRD: Neural reflectance decomposition from image collections. In *ICCV*, pages 12684–12694, 2021. 2, 3, 4
- [11] Mark Boss, Varun Jampani, Raphael Braun, Ce Liu, Jonathan Barron, and Hendrik Lensch. Neural-pil: Neural pre-integrated lighting for reflectance decomposition. *NeurIPS*, 34:10691–10704, 2021. 2, 4
- [12] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *ICCV*, pages 14124–14133, 2021. 3
- [13] Xingyu Chen, Qi Zhang, Xiaoyu Li, Yue Chen, Ying Feng, Xuan Wang, and Jue Wang. Hallucinated neural radiance fields in the wild. In *CVPR*, pages 12943–12952, 2022. 3
- [14] Yue Chen, Xingyu Chen, Xuan Wang, Qi Zhang, Yu Guo, Ying Shan, and Fei Wang. Local-to-global registration for bundle-adjusting neural radiance fields. In *CVPR*, 2023. 3
- [15] Paul Debevec. Rendering synthetic objects into real scenes: Bridging traditional and image-based graphics with global illumination and high dynamic range photography. In *SIGGRAPH*, pages 189–198, 1998. 2
- [16] Yilun Du, Yinan Zhang, Hong-Xing Yu, Joshua B Tenenbaum, and Jiajun Wu. Neural radiance flow for 4d view synthesis and video processing. In *ICCV*, pages 14304–14314. IEEE Computer Society, 2021. 3
- [17] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinrong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *CVPR*, 2022. 3
- [18] Marc-André Gardner, Kalyan Sunkavalli, Ersin Yumer, Xiaohui Shen, Emiliano Gambaretto, Christian Gagné, and

- Jean-François Lalonde. Learning to predict indoor illumination from a single image. *ACM TOG*, 36(6):1–14, 2017. 2
- [19] Paul Green, Jan Kautz, and Frédo Durand. Efficient reflectance and visibility approximations for environment map rendering. In *Computer Graphics Forum*, volume 26, pages 495–502. Wiley Online Library, 2007. 3
- [20] Tom Haber, Christian Fuchs, Philippe Bekaer, Hans-Peter Seidel, Michael Goesele, and Hendrik PA Lensch. Relighting objects from image collections. In *CVPR*, pages 627–634. IEEE, 2009. 2
- [21] Jon Hasselgren, Nikolai Hofmann, and Jacob Munkberg. Shape, light & material decomposition from images using monte carlo rendering and denoising. In *NeurIPS*, 2022. 2
- [22] Xin Huang, Qi Zhang, Ying Feng, Hongdong Li, and Qing Wang. Inverting the imaging process by learning an implicit camera model. In *CVPR*, 2023. 3
- [23] Xin Huang, Qi Zhang, Ying Feng, Hongdong Li, Xuan Wang, and Qing Wang. HDR-NeRF: High dynamic range neural radiance fields. In *CVPR*, pages 18398–18408, 2022. 3, 8
- [24] Xin Huang, Qi Zhang, Ying Feng, Xiaoyu Li, Xuan Wang, and Qing Wang. Local implicit ray function for generalizable radiance field representation. In *CVPR*, 2023. 3
- [25] Mohammad Mahdi Johari, Yann Lepoittevin, and François Fleuret. Geonerf: Generalizing nerf with geometry priors. In *CVPR*, pages 18365–18375, 2022. 3
- [26] James T Kajiya. The rendering equation. In *SIGGRAPH*, pages 143–150, 1986. 3
- [27] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 11
- [28] Junxuan Li and Hongdong Li. Neural reflectance for shape recovery with shadow handling. In *CVPR*, pages 16221–16230, 2022. 2
- [29] Tianye Li, Mira Slavcheva, Michael Zollhoefer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, Richard Newcombe, et al. Neural 3d video synthesis from multi-view video. In *CVPR*, pages 5521–5531, 2022. 3
- [30] Zhengqin Li, Mohammad Shafiei, Ravi Ramamoorthi, Kalyan Sunkavalli, and Manmohan Chandraker. Inverse rendering for complex indoor scenes: Shape, spatially-varying lighting and svbrdf from a single image. In *CVPR*, pages 2475–2484, 2020. 2
- [31] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. *NeurIPS*, 33:15651–15663, 2020. 3
- [32] Yuan Liu, Sida Peng, Lingjie Liu, Qianqian Wang, Peng Wang, Christian Theobalt, Xiaowei Zhou, and Wenping Wang. Neural rays for occlusion-aware image-based rendering. In *CVPR*, pages 7824–7833, 2022. 3
- [33] Stephen Lombardi and Ko Nishino. Radiometric scene decomposition: Scene reflectance, illumination, and geometry from rgb-d images. In *2016 fourth international conference on 3d vision (3dv)*, pages 305–313. IEEE, 2016. 2
- [34] 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*, pages 12861–12870, 2022. 3
- [35] Stephen Robert Marschner. *Inverse rendering for computer graphics*. Cornell University, 1998. 2
- [36] Nelson Max. Optical models for direct volume rendering. *IEEE TVCG*, 1(2):99–108, 1995. 3
- [37] 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*, pages 405–421. Springer, 2020. 2, 3, 13
- [38] Daisuke Miyazaki and Katsushi Ikeuchi. Shape estimation of transparent objects by using inverse polarization ray tracing. *IEEE TPAMI*, 29(11):2018–2030, 2007. 3
- [39] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *arXiv preprint arXiv:2201.05989*, 2022. 3
- [40] Giljoo Nam, Joo Ho Lee, Diego Gutierrez, and Min H Kim. Practical svbrdf acquisition of 3d objects with unstructured flash photography. *ACM TOG*, 37(6):1–12, 2018. 2
- [41] Michael Oechsle, Songyou Peng, and Andreas Geiger. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *ICCV*, pages 5589–5599, 2021. 3
- [42] Jeong Joon Park, Aleksander Holynski, and Steven M Seitz. Seeing the world in a bag of chips. In *CVPR*, pages 1417–1427, 2020. 2, 7, 11, 12
- [43] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *ICCV*, pages 5865–5874, 2021. 3
- [44] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *CVPR*, pages 10318–10327, 2021. 3
- [45] Ravi Ramamoorthi and Pat Hanrahan. A signal-processing framework for inverse rendering. In *SIGGRAPH*, pages 117–128, 2001. 2
- [46] Christian Reiser, Songyou Peng, Yiyi Liao, and Andreas Geiger. Kilonerf: Speeding up neural radiance fields with thousands of tiny mlps. In *ICCV*, pages 14335–14345, 2021. 3
- [47] Yoichi Sato, Mark D Wheeler, and Katsushi Ikeuchi. Object shape and reflectance modeling from observation. In *SIGGRAPH*, pages 379–387, 1997. 2
- [48] Shuran Song and Thomas Funkhouser. Neural illumination: Lighting prediction for indoor environments. In *CVPR*, pages 6918–6926, 2019. 2
- [49] Pratul P Srinivasan, Boyang Deng, Xiuming Zhang, Matthew Tancik, Ben Mildenhall, and Jonathan T Barron. Nerv: Neural reflectance and visibility fields for relighting and view synthesis. In *CVPR*, pages 7495–7504, 2021. 2, 3
- [50] Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Christoph Lassner, and Christian Theobalt. Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In *ICCV*, pages 12959–12970, 2021. 3

- [51] Levi Valgaerts, Chenglei Wu, Andrés Bruhn, Hans-Peter Seidel, and Christian Theobalt. Lightweight binocular facial performance capture under uncontrolled lighting. *ACM TOG*, 31(6):187–1, 2012. 2
- [52] 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. 2, 3
- [53] Jiaping Wang, Peiran Ren, Minmin Gong, John Snyder, and Baining Guo. All-frequency rendering of dynamic, spatially-varying reflectance. In *SIGGRAPH Asia*, pages 1–10. 2009. 3
- [54] 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*, 34:27171–27183, 2021. 2, 3
- [55] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P Srinivasan, Howard Zhou, Jonathan T Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *CVPR*, pages 4690–4699, 2021. 3
- [56] Zhen Wen, Zicheng Liu, and Thomas S Huang. Face relighting with radiance environment maps. In *CVPR*, volume 2, pages II–158. IEEE, 2003. 2
- [57] Shaowen Xie, Hao Zhu, Zhen Liu, Qi Zhang, You Zhou, Xun Cao, and Zhan Ma. Diner: Disorder-invariant implicit neural representation. In *CVPR*, 2023. 3
- [58] Zexiang Xu, Kalyan Sunkavalli, Sunil Hadap, and Ravi Ramamoorthi. Deep image-based relighting from optimal sparse samples. *ACM TOG*, 37(4):1–13, 2018. 2
- [59] Wenqi Yang, Guanying Chen, Chaofeng Chen, Zhenfang Chen, and Kwan-Yee K. Wong. S^3 -nerf: Neural reflectance field from shading and shadow under a single viewpoint. In *NeurIPS*, 2022. 2
- [60] Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. In *CVPR*, pages 1790–1799, 2020. 2, 7, 11, 12
- [61] 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*, pages 700–716. Springer, 2022. 2, 3, 4, 5, 6, 8, 11, 12, 13
- [62] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. *NeurIPS*, 34:4805–4815, 2021. 3
- [63] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. *NeurIPS*, 33:2492–2502, 2020. 2, 3, 12
- [64] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. Plenoctrees for real-time rendering of neural radiance fields. In *ICCV*, pages 5752–5761, 2021. 3
- [65] Yizhou Yu, Paul Debevec, Jitendra Malik, and Tim Hawkins. Inverse global illumination: Recovering reflectance models of real scenes from photographs. In *SIGGRAPH*, pages 215–224, 1999. 2
- [66] Cheng Zhang, Bailey Miller, Kan Yan, Ioannis Gkioulekas, and Shuang Zhao. Path-space differentiable rendering. *ACM TOG*, 39(4), 2020. 3
- [67] 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, 3, 4
- [68] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018. 6
- [69] 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 TOG*, 40(6):1–18, 2021. 2, 4, 8
- [70] Yuanqing Zhang, Jiaming Sun, Xingyi He, Huan Fu, Rongfei Jia, and Xiaowei Zhou. Modeling indirect illumination for inverse rendering. In *CVPR*, pages 18643–18652, 2022. 2, 3, 4, 12
- [71] Yiyu Zhuang, Hao Zhu, Xusen Sun, and Xun Cao. Mofanerf: Morphable facial neural radiance field. In *ECCV*, 2022. 3