A Pre-convolved Representation for Plug-and-Play Neural Illumination Fields

Yiyu Zhuang[1]* Qi Zhang[2]*, Xuan Wang[3], Hao Zhu[1],
Ying Feng[2], Xiaoyu Li[2], Ying Shan[2], Xun Cao[1]
[1]Najing University, Nanjing, China
[2]Tencent AI Lab, Shenzhen, China
[3]Ant Group, Hangzhou, China
yiyu.zhuang@smail.nju.edu.cn, nwpuqzhang@gmail.com, xwang.cv@gmail.com, zh@nju.edu.cn,
yfeng.von@gmail.com, xliea@connect.ust.hk, yingsshan@tencent.com, caoxun@nju.edu.cn

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 Illumination Fields (NeIF) that uses radiance fields 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-convolved background for accurate 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.

Introduction

Modeling and representing the environment illumination from multi-view images is the fundamental issue that has been extensively studied throughout the development of rendering algorithms (Park, Holynski, and Seitz 2020; Yao et al. 2022; Zhang et al. 2022). This task is inherently related to materials decomposition, since the observed scene appearance is affected by the interactions between environment 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 environment illumination is of high-dimensional information and strongly coupled with materials.

Recent methods employ approximated illumination representations (*e.g.*, environment map (Zhang et al. 2021b), and spherical Gaussian (SG) models (Zhang et al. 2021a; Boss et al. 2021a,b; Zhang et al. 2022)) to simplify the interaction between environment illumination and the object and reduce computational expense. Unfortunately, the assumption used in doing so is that the environment illumination of the scene is infinitely far away. Neither the environment map nor SG models take the position of 3D illumination into account so they are unable to handle environment occlusion and directional lighting practically. To address this problem, NeILF (Yao et al. 2022) 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 environment illumination?*

Recent advances in Neural Radiance Fields (NeRF) and its variants (Mildenhall et al. 2020; Barron et al. 2021; Verbin et al. 2022; Zhang et al. 2021b) 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 environmental illumination. To the best of our knowledge, little in the literature has ever tapped into using volumetric radiance fields 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.

This paper presents the *Neural Illumination Fields* (NeIF) 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. We first acquire the object's geometry from the input images using the existing method (Yariv et al. 2020). We focus solely on the decomposition of environment illumination and object material. Specifically, 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

*Both authors contributed equally to this work.

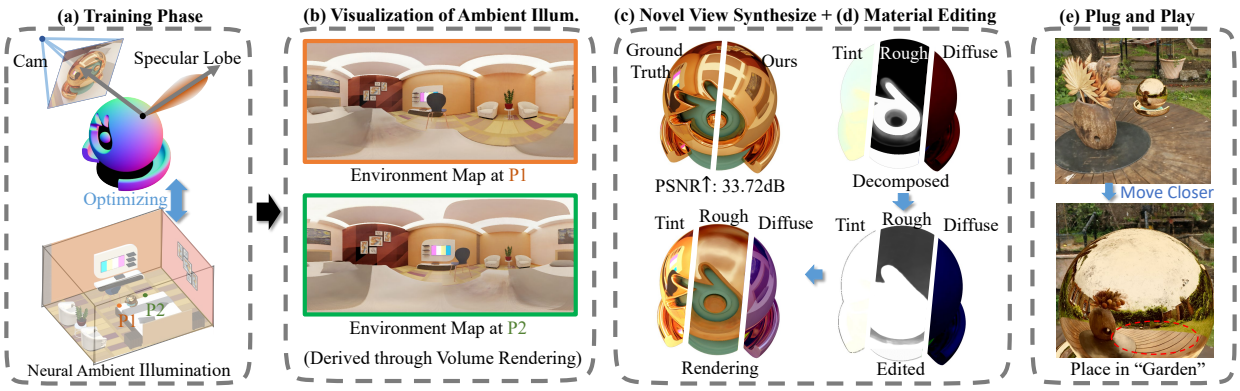

Figure 1: We introduce a 'plug-and-play' Neural Illumination Fields (NeIF) that uses volumetric radiance fields to portray 3D environment samples as light emitters, naturally re-rendering new objects under any NeRF-style environments. (a) Using a set of images and masks, our method first optimizes the geometry and then jointly optimizes the NeIF and an object's materials in two stages. (b) It illustrates the environment map at arbitrary samples in 3D scenes via volume rendering, which handles environment occlusions and directional lighting naturally. The proposed method provides photo-realistic novel views (c) and visually reasonable material editing (d). Furthermore, (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 (red circle) is exactly reproduced, which is almost impossible for the previous illumination representation.

the rough refractive surfaces are inherently related to the convolved background, although the background is ignored in most methods. Our contributions are summarized as:

- Neural Illumination Fields (NeIF) is proposed to express radiance fields of incoming rays such that treats each sample in the 3D scene as a light emitter.
- Fully differentiable rendering pipeline is presented to seamlessly illuminate meshes using a NeRF-style environment.
- Integrated Lobe Encoding (ILE) is proposed to featurize incoming rays within roughness-adaptive specular lobe to reduce computational cost.
- Multiscale pre-convolved representation for the background is proposed to assist in the decomposition of object materials and environmental illumination.

Related Work

Illumination Representation. Illumination representation is essential for photo-realistic rendering in various view synthesis and relighting applications (Haber et al. 2009; Xu et al. 2018; Bi et al. 2020a; Li et al. 2020; Boss et al. 2021a; Srinivasan et al. 2021; Zhang et al. 2022; Yao et al. 2022; Boss et al. 2021b; Zhang et al. 2021b,a). Considering that the observed surface appearance 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 (Sato, Wheeler, and Ikeuchi 1997; Marschner 1998; Yu et al. 1999; Ramamoorthi and Hanrahan 2001). Since it is an ill-posed problem, previous methods mitigate this issue by using simplified material models (Zhang et al. 2021a) and varying lighting conditions (Nam et al. 2018; Bi et al. 2020b,a; Yang et al. 2022). This related work specifically focuses on ambient illumination representation techniques.

The seminal work (Debevec 1998) 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 (Wen, Liu, and Huang 2003; Haber et al. 2009; Barron and Malik 2014; Valgaerts et al. 2012; Song and Funkhouser 2019) use the environment map to handle inverse rendering problems naturally. Furthermore, given high-quality geometry, prior works (Lombardi and Nishino 2016; Park, Holynski, and Seitz 2020) 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 (Green, Kautz, and Durand 2007; Wang et al. 2009; Zhang et al. 2021a). 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 (Miyazaki and Ikeuchi 2007; Srinivasan et al. 2021) and path-tracing (Azinovic et al. 2019; Zhang et al. 2020)) are proposed to use ray casting to express the interactions between ambient illumination and surface materials (Akenine-Moller, Haines, and Hoffman 2019). 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 illumination representation. We represent the 3D sample of the surrounding environment as a light emitter, such that both

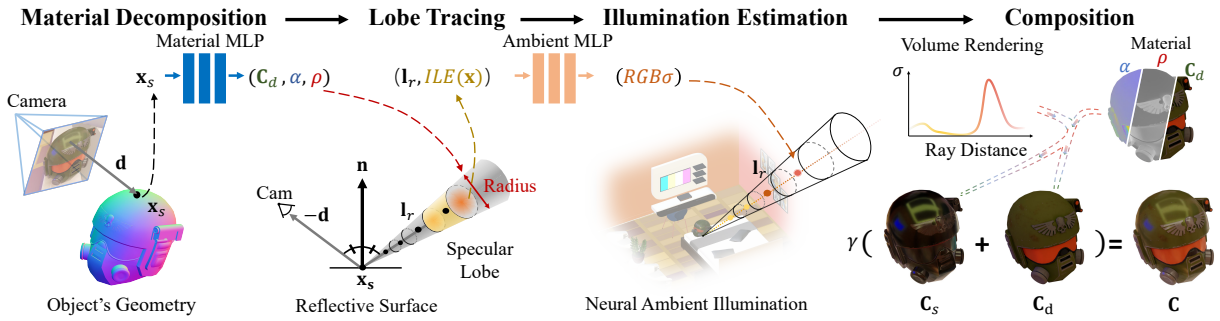

Figure 2: The overview of the proposed method that decomposes materials $(\mathbf{C}_d, \alpha, \rho)$ of an object and NeIF radiance fields from a set of images and a geometry reconstructed by (Yariv et al. 2020). For a specific surface point $\mathbf{x}_s$, NeIF traces its incoming ray within a 3D 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 NeIF representation and object's materials.

position and direction of illumination are taken into account.

Neural Radiance Fields. Neural rendering (Mildenhall et al. 2020; Yariv et al. 2020; Liu et al. 2020; Yariv et al. 2021; Wang et al. 2021a; Zhuang et al. 2023), the task of learning to recover the properties of 3D scenes from observed images, has seen significant success. In particular, Neural Radiance Fields (NeRF) (Mildenhall et al. 2020) 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 (Wang et al. 2021b; Chen et al. 2021; Johari, Lepoittevin, and Fleuret 2022; Huang et al. 2023), non-rigidly deformable objects (Tretschk et al. 2021; Park et al. 2021; Zhuang et al. 2022; Wu et al. 2023), imaging processing (Huang et al. 2022; Ma et al. 2022; Chen et al. 2022) Recently, Mip-NeRF (Barron et al. 2021, 2022) 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 (Verbin et al. 2022) 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 material.

Recent advances in differentiable rendering make it possible to reconstruct environment illumination under casual lighting conditions. Specifically, PhySG (Zhang et al. 2021a) and NeRD (Boss et al. 2021a) use SG representation to decompose the scene under complex and unknown illumination. Zhang *et al.* (Zhang et al. 2022) 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 indirect lighting realistically. NeILF (Yao et al. 2022) 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 the NeIF that uses the volumetric radiance fields to express arbitrary illumination in the environment, such that environment occlusion and directional lighting could be handled naturally.

Method

Given a set of posed images of an object captured under static illumination, our goal is to decompose the shape, material, and lighting, with a primary focus on representing the environment lighting and its interaction with the object surface. Initially, we represent the shape as a zero-level set (Yariv et al. 2020) by learning a Signal Distance Function (SDF) to reconstruct the geometry. Our main contribution, as shown in Fig. 2, is introduced after the object geometry reconstruction in stage one.

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 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 $(\sigma_i, \mathbf{c}_i)$ into a single color $\mathbf{C}(\mathbf{r})$ of the pixel via numerical quadrature (Max 1995),

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

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 environment illumination and scene material based on the rendering equation (Kajiya 1986),

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

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 $\mathbf{l}\cdot\mathbf{n} > 0$.

Neural Illumination Fields

Although previous methods have attempted to model diverse types of lighting in various ways (Zhang et al. 2021a,b; Boss et al. 2021a; Yao et al. 2022), 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 illumination.

By expressing the *Neural Illumination Fields* (NeIF) 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,

$$\begin{gathered}\mathbf{L}_{in}(\mathbf{x_s}, \mathbf{l}) = \int_{t_n}^{t_f} T(t)\sigma(\mathbf{r}(t))\mathbf{c}(\mathbf{r}(t), \mathbf{l})\mathrm{d}t, \\ \text{where } T(t) = \exp\left(-\int_{t_n}^{t}\sigma(\mathbf{r}(s))\mathrm{d}s\right),\end{gathered} \quad (3)$$

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 NeIF to model the directional emitted rays and environment occlusions of any static 3D environments.

To define how light derived from a NeRF-style environment 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 (Akenine-Moller, Haines, and Hoffman 2019). $\mathbf{l}_r$ 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 lope with a normalized spherical Gaussian (Akenine-Moller, Haines, and Hoffman 2019),

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

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:

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

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.

Integrated Lobe Encoding

To better learn the high-frequency variation of NeIF 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 (Barron et al. 2021), 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 $(\mu, \mathbf{\Sigma})$ of the conical frustum, which is obtained by $\mathbf{x}_s$ and $\mathbf{l}_r$. Note that the radius variance's part in $\mathbf{\Sigma}(\rho)$ is decided by the material roughness of the surface point, which is different with IPE (Barron et al. 2021). 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 $(\mu, \mathbf{\Sigma})$ within the conical frustum,

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

These features encoded by ILE are used as input to the MLP network $\mathcal{R}$ to output the density and color of our NeIF. 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,

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

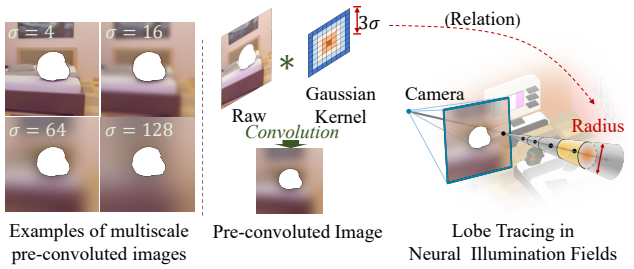

Figure 3: The pipeline demonstrates how pre-convolved images can be used to construct a pre-convolved illumination representation.

	Glossy Blender Dataset			Real-word		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
NeILF	25.851	0.930	0.088	21.664	0.797	0.203
NeILF*	27.279	0.941	0.084	21.226	0.760	0.242
Ne-Env	25.971	0.935	0.086	20.786	0.761	0.251
w/o ILE	37.362	0.979	0.036	24.472	0.801	0.191
w/o BG	36.404	0.979	0.033	24.248	0.801	0.193
w/o Conv.	37.748	0.981	0.032	24.854	0.808	0.185
w/o Reg.	**38.442**	**0.983**	**0.029**	24.973	**0.816**	**0.176**
Ours	37.985	0.981	0.031	**25.022**	0.814	0.179

Table 1: Our method and its ablations outperform the baseline NeILF and its variants on the Glossy Blender dataset, showing superior rendering quality in averaged metrics.

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.

Multiscale Pre-convolved Representation

Due the complexity of high-dimension representation of NeIF, it's easy to integrate the diffuse color into environment illumination, causing ambiguous decomposition. Consequently, we propose a multiscale pre-convolved 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 roughness. As the roughness increases, the environment illumination is convolved 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-convolved 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 consider the view direction, which allows for more accurate and realistic specular reflections that are consistent with the object's roughness level.

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 (Yao et al. 2022) as:

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

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 $\widehat{\mathbf{C}}$ and the ground truth color $\mathbf{C}$ to jointly optimize the environment illumination and object's materials:

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

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

Experiment

Implementation Details. Our method is implemented on top of Mip-NeRF (Barron et al. 2021) 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 (Zhang et al. 2018), on both synthetic and real-world datasets.

Baselines. We compare our method with the following methods: 1) NeILF (Yao et al. 2022) 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 NeIF. The implementation details could be found in our *supplement*.

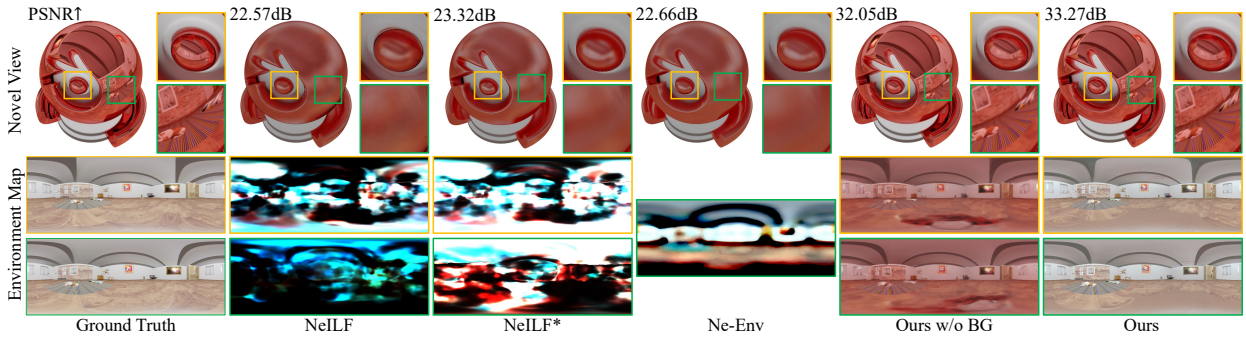

Figure 4: Qualitative comparison of our method against baseline NeILF (Yao et al. 2022) and its variants on self-rendered Glossy Blender dataset. The zoom-in details labeling with yellow and green boxes and corresponding reconstructed environment map surrounding the surface point of each box center are shown. Our method significantly outperforms other methods, both in the rendering quality of novel views and the reconstruction of environment illumination. Noted that, although ours w/o background obtains reconstructs geometrically correct results, but fails on the decomposition of diffuse color and environment illumination due to the incorrect white balance and less background supervision.

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 (Foundation 1994), 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.

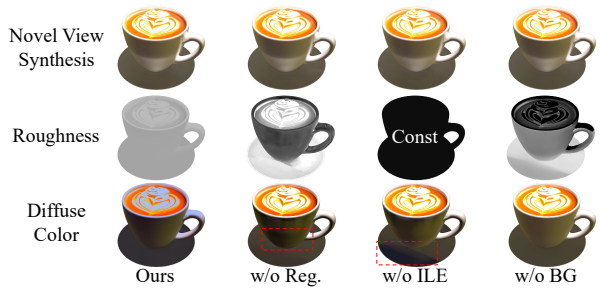

Figure 5: Visualizations of decomposed materials for ablation study. "w/o Reg." generates noisy materials. "w/o ILE" cannot handle the varying in material roughness, resulting in artifacts of diffuse items. "w/o BG" appears incorrect roughness due to lighting.

Tab. 1 and Fig. 4 show 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 (Yao et al. 2022) are the masked images of the objects without background, we test our method on the same setting for fairness, referring to "w/o BG" in the experiments. 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 BG").

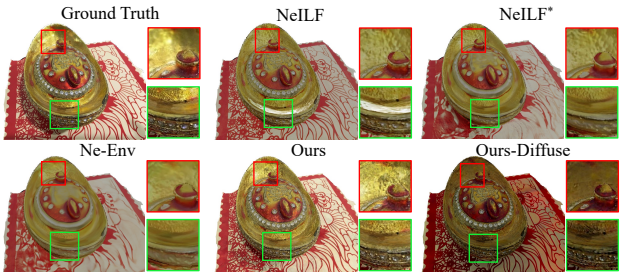

Figure 6: Qualitative comparison with zoom-in details of our method against baselines on real-world dataset. Our method produces the most visually pleasing novel views, especially on reconstructing the highlight regions and detailed textures.

Specifically, NeILF and NeILF* struggle to handle nearby illumination, which varies dramatically with the position, and their inability to gather information across different views exacerbates the ambiguity. Although NeILF* simplifies the BRDF function like ours, its performance only slightly improves compared to NeILF, as shown in Tab. 1. For Ne-Env, sharing the same environment map across different positions reduces uncertainty and constructs a more meaningful map. However, the decomposition of materials and illumination is poor due to the simplification of the illumination model, which assumes all radiance comes from an infinite distance. Compared to NeILF and its variants, our method renders photo-realistic views and recovers precise ambient illumination. Even without background guidance, our method reconstructs geometrically correct results. However, the white balance is incorrect, as the majority color of the object is red, causing ambiguity about the red object or red incident light. This issue highlights the importance of using background guidance to disentangle material and light.

Ablation Studies. Except "w/o BG", three additional ablations are contained: "w/o ILE" refers to ignoring ILE, "w/o Conv." omits multiscale pre-convolved representation, and "w/o Reg." excludes the Bilateral smoothness regulariza-

Figure 7: Illumination manipulation visualizations show our method recovering environment lighting for high-fidelity specularities and natural illumination in various manipulations. It handles occlusion and high lighting effectively, and when placed in virtual or pre-trained Mip-NeRF environments, it produces novel views with realistic reflections.

tion (Eq. (8)). The results are reported in Tab. 1 and Fig. 5. Specifically, while "w/o Reg." achieves the best metric performance, it generates noise and becomes inconsistent across the same material, as shown in Fig. 5. This makes it unsuitable for material editing applications. Besides, in "w/o ILE, the model is trained without the roughness-related encoding, leading to difficulties in handling spatially-varying roughness and resulting in artifacts for diffuse items. The "w/o BG" results show that without background assistance, it becomes challenging to distinguish between light and material, causing ambiguity.

Real-world. We then test our method on 8 real-world scenes from BlendedMVS (Yao et al. 2020) and Bag of Chips (Park, Holynski, and Seitz 2020), which provide images and depth maps reconstructed by MVS methods or RGBD camera. We selected the last ten 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. 1 and Fig. 6 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.

Applications

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 NeIF model in a pre-trained Mip-NeRF environment (Barron et al. 2022), 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

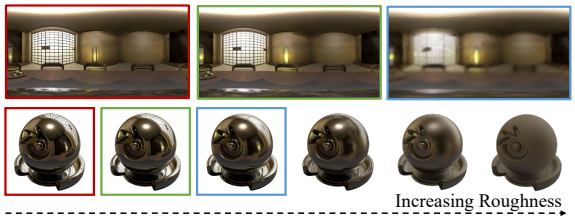

Figure 8: Visualizations of roughness editing are presented. The first row displays environment maps for the initial three balls, while the second row demonstrates increasing blurriness with higher roughness. Despite roughness exceeding the Gaussian kernel used in pre-convolved representation, our method still produces visually realistic outcomes.

in the "Garden" dataset. It verifies that our NeIF 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.

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. 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.

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 NeIF 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 scenes.

Limitations. It is difficult to model ambient illumination and decompose the object's material, our pipeline relies on the geometry reconstructed through stage one.

Acknowledgments

Supported by China's National Key Research and Development Program (Grant 2022YFF0902201), National Natural Science Foundation (Grants 62001213, 62025108), and Tencent Rhino-Bird Research Program. Thanks to anonymous reviewers for valuable feedback.

References

Akenine-Moller, T.; Haines, E.; and Hoffman, N. 2019. *Real-time rendering*. AK Peters/crc Press.

Azinovic, D.; Li, T.-M.; Kaplanyan, A.; and Nießner, M. 2019. Inverse path tracing for joint material and lighting estimation. In *CVPR*, 2447–2456.

Barron, J. T.; and Malik, J. 2014. Shape, illumination, and reflectance from shading. *IEEE transactions on pattern analysis and machine intelligence*, 37(8): 1670–1687.

Barron, J. T.; Mildenhall, B.; Tancik, M.; Hedman, P.; Martin-Brualla, R.; and Srinivasan, P. P. 2021. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *ICCV*, 5855–5864.

Barron, J. T.; Mildenhall, B.; Verbin, D.; Srinivasan, P. P.; and Hedman, P. 2022. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *CVPR*, 5470–5479.

Bi, S.; Xu, Z.; Sunkavalli, K.; Hašan, M.; Hold-Geoffroy, Y.; Kriegman, D.; and Ramamoorthi, R. 2020a. Deep reflectance volumes: Relightable reconstructions from multi-view photometric images. In *ECCV*, 294–311. Springer.

Bi, S.; Xu, Z.; Sunkavalli, K.; Kriegman, D.; and Ramamoorthi, R. 2020b. Deep 3d capture: Geometry and reflectance from sparse multi-view images. In *CVPR*, 5960–5969.

Boss, M.; Braun, R.; Jampani, V.; Barron, J. T.; Liu, C.; and Lensch, H. 2021a. NeRD: Neural reflectance decomposition from image collections. In *ICCV*, 12684–12694.

Boss, M.; Jampani, V.; Braun, R.; Liu, C.; Barron, J.; and Lensch, H. 2021b. Neural-pil: Neural pre-integrated lighting for reflectance decomposition. *NeuIPS*, 34: 10691–10704.

Chen, A.; Xu, Z.; Zhao, F.; Zhang, X.; Xiang, F.; Yu, J.; and Su, H. 2021. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *ICCV*, 14124–14133.

Chen, X.; Zhang, Q.; Li, X.; Chen, Y.; Feng, Y.; Wang, X.; and Wang, J. 2022. Hallucinated neural radiance fields in the wild. In *CVPR*, 12943–12952.

Debevec, P. 1998. Rendering synthetic objects into real scenes: Bridging traditional and image-based graphics with global illumination and high dynamic range photography. In *SIGGRAPH*, 189–198.

Foundation, B. 1994. Blender. https://www.blender.org/. Accessed: 2022-07-15.

Green, P.; Kautz, J.; and Durand, F. 2007. Efficient reflectance and visibility approximations for environment map rendering. In *Computer Graphics Forum*, volume 26, 495–502. Wiley Online Library.

Haber, T.; Fuchs, C.; Bekaer, P.; Seidel, H.-P.; Goesele, M.; and Lensch, H. P. 2009. Relighting objects from image collections. In *CVPR*, 627–634. IEEE.

Huang, X.; Zhang, Q.; Feng, Y.; Li, H.; Wang, X.; and Wang, Q. 2022. HDR-NeRF: High Dynamic Range Neural Radiance Fields. In *CVPR*, 18398–18408.

Huang, X.; Zhang, Q.; Feng, Y.; Li, X.; Wang, X.; and Wang, Q. 2023. Local Implicit Ray Function for Generalizable Radiance Field Representation. In *CVPR*.

Johari, M. M.; Lepoittevin, Y.; and Fleuret, F. 2022. Geonerf: Generalizing nerf with geometry priors. In *CVPR*, 18365–18375.

Kajiya, J. T. 1986. The rendering equation. In *SIGGRAPH*, 143–150.

Li, Z.; Shafiei, M.; Ramamoorthi, R.; Sunkavalli, K.; and Chandraker, M. 2020. Inverse rendering for complex indoor scenes: Shape, spatially-varying lighting and svbrdf from a single image. In *CVPR*, 2475–2484.

Liu, L.; Gu, J.; Zaw Lin, K.; Chua, T.-S.; and Theobalt, C. 2020. Neural sparse voxel fields. *NeuIPS*, 33: 15651–15663.

Lombardi, S.; and Nishino, K. 2016. Radiometric scene decomposition: Scene reflectance, illumination, and geometry from rgb-d images. In *2016 fourth international conference on 3d vision (3dv)*, 305–313. IEEE.

Ma, L.; Li, X.; Liao, J.; Zhang, Q.; Wang, X.; Wang, J.; and Sander, P. V. 2022. Deblur-NeRF: Neural Radiance Fields from Blurry Images. In *CVPR*, 12861–12870.

Marschner, S. R. 1998. *Inverse rendering for computer graphics*. Cornell University.

Max, N. 1995. Optical models for direct volume rendering. *TVCG*, 1(2): 99–108.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2020. NeRF: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 405–421. Springer.

Miyazaki, D.; and Ikeuchi, K. 2007. Shape estimation of transparent objects by using inverse polarization ray tracing. *PAMI*, 29(11): 2018–2030.

Nam, G.; Lee, J. H.; Gutierrez, D.; and Kim, M. H. 2018. Practical svbrdf acquisition of 3d objects with unstructured flash photography. *TOG*, 37(6): 1–12.

Park, J. J.; Holynski, A.; and Seitz, S. M. 2020. Seeing the world in a bag of chips. In *CVPR*, 1417–1427.

Park, K.; Sinha, U.; Barron, J. T.; Bouaziz, S.; Goldman, D. B.; Seitz, S. M.; and Martin-Brualla, R. 2021. Nerfies: Deformable neural radiance fields. In *ICCV*, 5865–5874.

Ramamoorthi, R.; and Hanrahan, P. 2001. A signal-processing framework for inverse rendering. In *SIGGRAPH*, 117–128.

Sato, Y.; Wheeler, M. D.; and Ikeuchi, K. 1997. Object shape and reflectance modeling from observation. In *SIGGRAPH*, 379–387.

Song, S.; and Funkhouser, T. 2019. Neural illumination: Lighting prediction for indoor environments. In *CVPR*, 6918–6926.

Srinivasan, P. P.; Deng, B.; Zhang, X.; Tancik, M.; Mildenhall, B.; and Barron, J. T. 2021. Nerv: Neural reflectance and visibility fields for relighting and view synthesis. In *CVPR*, 7495–7504.

Tretschk, E.; Tewari, A.; Golyanik, V.; Zollhöfer, M.; Lassner, C.; and Theobalt, C. 2021. Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In *ICCV*, 12959–12970.

Valgaerts, L.; Wu, C.; Bruhn, A.; Seidel, H.-P.; and Theobalt, C. 2012. Lightweight binocular facial performance capture under uncontrolled lighting. *TOG*, 31(6): 187–1.

Verbin, D.; Hedman, P.; Mildenhall, B.; Zickler, T.; Barron, J. T.; and Srinivasan, P. P. 2022. Ref-nerf: Structured view-dependent appearance for neural radiance fields. In *CVPR*, 5481–5490. IEEE.

Wang, J.; Ren, P.; Gong, M.; Snyder, J.; and Guo, B. 2009. All-frequency rendering of dynamic, spatially-varying reflectance. In *SIGGRAPH Asia*, 1–10.

Wang, P.; Liu, L.; Liu, Y.; Theobalt, C.; Komura, T.; and Wang, W. 2021a. NeuS: Learning Neural Implicit Surfaces by Volume Rendering for Multi-view Reconstruction. *NeuIPS*, 34: 27171–27183.

Wang, Q.; Wang, Z.; Genova, K.; Srinivasan, P. P.; Zhou, H.; Barron, J. T.; Martin-Brualla, R.; Snavely, N.; and Funkhouser, T. 2021b. Ibrnet: Learning multi-view image-based rendering. In *CVPR*, 4690–4699.

Wen, Z.; Liu, Z.; and Huang, T. S. 2003. Face relighting with radiance environment maps. In *CVPR*, volume 2, II–158. IEEE.

Wu, M.; Zhu, H.; Huang, L.; Zhuang, Y.; Lu, Y.; and Cao, X. 2023. High-Fidelity 3D Face Generation From Natural Language Descriptions. In *CVPR*, 4521–4530.

Xu, Z.; Sunkavalli, K.; Hadap, S.; and Ramamoorthi, R. 2018. Deep image-based relighting from optimal sparse samples. *TOG*, 37(4): 1–13.

Yang, W.; Chen, G.; Chen, C.; Chen, Z.; and Wong, K.-Y. K. 2022. S^3-NeRF: Neural Reflectance Field from Shading and Shadow under a Single Viewpoint. In *NeuIPS*.

Yao, Y.; Luo, Z.; Li, S.; Zhang, J.; Ren, Y.; Zhou, L.; Fang, T.; and Quan, L. 2020. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. In *CVPR*, 1790–1799.

Yao, Y.; Zhang, J.; Liu, J.; Qu, Y.; Fang, T.; McKinnon, D.; Tsin, Y.; and Quan, L. 2022. NeILF: Neural Incident Light Field for Physically-based Material Estimation. In *ECCV*, 700–716. Springer.

Yariv, L.; Gu, J.; Kasten, Y.; and Lipman, Y. 2021. Volume rendering of neural implicit surfaces. *NeuIPS*, 34: 4805–4815.

Yariv, L.; Kasten, Y.; Moran, D.; Galun, M.; Atzmon, M.; Ronen, B.; and Lipman, Y. 2020. Multiview neural surface reconstruction by disentangling geometry and appearance. *NeuIPS*, 33: 2492–2502.

Yu, Y.; Debevec, P.; Malik, J.; and Hawkins, T. 1999. Inverse global illumination: Recovering reflectance models of real scenes from photographs. In *SIGGRAPH*, 215–224.

Zhang, C.; Miller, B.; Yan, K.; Gkioulekas, I.; and Zhao, S. 2020. Path-space differentiable rendering. *TOG*, 39(4).

Zhang, K.; Luan, F.; Wang, Q.; Bala, K.; and Snavely, N. 2021a. Physg: Inverse rendering with spherical gaussians for physics-based material editing and relighting. In *CVPR*, 5453–5462.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 586–595.

Zhang, X.; Srinivasan, P. P.; Deng, B.; Debevec, P.; Freeman, W. T.; and Barron, J. T. 2021b. Nerfactor: Neural factorization of shape and reflectance under an unknown illumination. *TOG*, 40(6): 1–18.

Zhang, Y.; Sun, J.; He, X.; Fu, H.; Jia, R.; and Zhou, X. 2022. Modeling Indirect Illumination for Inverse Rendering. In *CVPR*, 18643–18652.

Zhuang, Y.; Zhang, Q.; Feng, Y.; Zhu, H.; Yao, Y.; Li, X.; Cao, Y.-P.; Shan, Y.; and Cao, X. 2023. Anti-Aliased Neural Implicit Surfaces with Encoding Level of Detail. In *SIGGRAPH Asia*, 1–10.

Zhuang, Y.; Zhu, H.; Sun, X.; and Cao, X. 2022. Mofanerf: Morphable facial neural radiance field. In *ECCV*.