

# Learning Indoor Inverse Rendering with 3D Spatially-Varying Lighting

Zian Wang<sup>1,2,3</sup>   Jonah Philion<sup>1,2,3</sup>   Sanja Fidler<sup>1,2,3</sup>   Jan Kautz<sup>1</sup>  
NVIDIA<sup>1</sup>   University of Toronto<sup>2</sup>   Vector Institute<sup>3</sup>  
{zianw, jphilion, sfidler, jkautz}@nvidia.com

## Abstract

In this work, we address the problem of jointly estimating albedo, normals, depth and 3D spatially-varying lighting from a single image. Most existing methods formulate the task as image-to-image translation, ignoring the 3D properties of the scene. However, indoor scenes contain complex 3D light transport where a 2D representation is insufficient. In this paper, we propose a unified, learning-based inverse rendering framework that formulates 3D spatially-varying lighting. Inspired by classic volume rendering techniques, we propose a novel Volumetric Spherical Gaussian representation for lighting, which parameterizes the exitant radiance of the 3D scene surfaces on a voxel grid. We design a physics-based differentiable renderer that utilizes our 3D lighting representation, and formulates the energy-conserving image formation process that enables joint training of all intrinsic properties with the re-rendering constraint. Our model ensures physically correct predictions and avoids the need for ground-truth HDR lighting which is not easily accessible. Experiments show that our method outperforms prior works both quantitatively and qualitatively, and is capable of producing photorealistic results for AR applications such as virtual object insertion even for highly specular objects.

## 1. Introduction

The task of inverse rendering, originally proposed by Barrow and Tenenbaum [3] in 1978, aims to reverse the rendering process by estimating reflectance, shape and lighting from a single image. Estimating these intrinsic properties enable downstream applications in augmented and mixed reality, such as realistic insertion of 3D objects into a given 2D image. Inverse rendering also facilitates semantic scene analysis such as object segmentation [5].

Given only observed pixel values, the problem of disambiguating reflectance, geometry and their complex interactions with lighting is challenging and ill-posed. Classic optimization-based methods leverage hand-crafted priors to constrain the ill-posed nature of the problem. However, these priors do not always hold for complex real world images and can lead to artifacts. Indoor scenes commonly encountered in AR applications are considered especially challenging due



Figure 1: From a single image, our model jointly estimates albedo, normals, depth, and the HDR lighting volume. Key to our method is inferring continuous HDR 3D spatially-varying lighting, which is critical in producing high quality virtual object insertion with realistic cast shadows and angular high-frequency details.

to complex 3D light transport that occurs indoors.

In this work, we address the problem of scene-level inverse rendering, focusing specifically on producing high dynamic range (HDR) 3D spatially-varying lighting with high-frequency details, as shown in Fig. 1. Estimating both HDR and 3D spatially-varying lighting is critical for photorealistic virtual object insertion; HDR enables realistic cast shadows and 3D spatially-varying lighting enables high-frequency details. We use the HDR lighting inferred by our model to insert highly specular objects and produce realistic cast shadows and high-frequency details, which were not possible in previous works [11, 22, 33, 38].

Existing learning-based methods usually exploit powerful 2D CNNs and formulate the inverse rendering problem as image-to-image translation. Lighting is usually represented with spherical lobes such as spherical Harmonics and spherical Gaussian [2, 41], and environment maps [10, 33], which ignores spatially-varying effects. Recent works attempt to predict 2D spatially-varying spherical lobes [11, 22], but still lack one degree of freedom (depth) and compromise in terms of angular high-frequency effects. As a consequence, the 2D representation of the scene lighting is not sufficiently performant for many downstream applications.

In this paper, we propose a holistic inverse rendering framework for jointly estimating reflectance, shape and 3D

spatially-varying lighting, by formulating the complete rendering process in an end-to-end trainable way with a 3D lighting representation. We propose a novel Volumetric Spherical Gaussian representation for lighting, which is a voxel representation for the scene surfaces. Spherical Gaussian parameters in each voxel control the emission direction and sharpness of the light source, which captures view-dependent effects and can handle strong directional lighting. Since ground-truth for HDR lighting is not easily available, we design a raytracing based differentiable renderer that leverages our lighting representation and formulates the energy-conserving image formation process. We use the renderer to jointly train all intrinsic properties by enforcing the re-rendering constraint, ensuring that predictions are physically correct. To the best of our knowledge, our approach is the first to estimate a complete continuous light field function from a single image, including both HDR and high-frequency spatial and angular details, despite being trained with only LDR images.

We experimentally show that our approach outperforms existing state-of-the-art inverse rendering and lighting estimation methods. We demonstrate that our method learns to produce complex lighting effects of real-world indoor scenes and better disambiguates intrinsic properties. Our lighting representation enables realistic cast shadows and angular high-frequency details and is therefore capable of producing significantly more realistic object insertion results for AR applications that were not possible previously, most importantly including the insertion of highly specular objects.

## 2. Related Work

**Inverse Rendering.** The task of inverse rendering dates back to Barrow and Tenenbaum [4], with the goal of jointly estimating intrinsic properties of the scene, *i.e.* reflectance, shape and lighting. Classic approaches usually tackle sub-tasks of inverse rendering, such as intrinsic image decomposition [3, 8, 13, 18, 44] and shape from shading [32, 43]. These methods primarily defined hand-crafted priors over the content of a scene and formulate the task as a per-image energy minimization problem. Recently, SIRFS [2] proposed a statistical inference framework that jointly estimates the intrinsic properties. However, these methods rely on assumptions that are not always true for real scenes, leading to artifacts when applied on real-world images. The need to perform test-time optimization also raises the computational burden, limiting these approaches to offline applications.

Recent works utilize 2D CNNs for learning data-driven priors from sparse human reflectance annotations [6], calibrated multi-view and multi-illumination data [24, 41], and most commonly, synthetic data [7, 22, 23, 25, 33] that comes with dense ground-truth labels. Among these works, [7, 25] are limited to single object inputs and do not address complex light transport. NIR [33] and Li *et al.* [22] are most

similar to our work, and address general indoor scenes. Both formulate the task as image-to-image translation and train on synthetic datasets. NIR employs an environment map to represent lighting and introduces a non-interpretable neural renderer to account for spatially-varying lighting effects. Li *et al.* predicts a spherical Gaussian lighting for each pixel location to get 2D spatially-varying lighting, but it still lacks one degree of freedom and sacrifices angular frequency at each pixel location. In our work, we tackle a more challenging task of estimating 3D lighting in a holistic inverse rendering framework, and learn to disentangle complex lighting effects with a physics-based representation.

**Lighting Estimation** is a sub-task of inverse rendering. Most existing works tackle simplified problem settings and ignore spatially-varying effects, such as outdoor scenes [16, 17, 42] and objects [7, 25, 40]. Prior works on indoor lighting estimation explored lighting representations such as environment maps [10, 20, 33], per-pixel spherical lobes [11, 22, 45] and light source parameters [9]. However, these methods either do not account for spatially-varying effects or do not preserve high-frequency details. Recent works explore 3D spatially-varying lighting [37, 38]. Neural Illumination [37] predicts an environment map with 2D CNNs given each queried 3D location, but it suffers from spatial instability. Lighthouse [38] tackles this problem from the perspective of view synthesis and confirms the necessity for a 3D lighting representation. However, it does not address HDR information, and there is no guarantee that the inpainted lighting is physically correct. In this work, we leverage a holistic inverse rendering framework to enable physically correct HDR lighting prediction, whereby we only train on the LDR ground-truth data as in Lighthouse.

**Neural Scene Representations.** Efficient 3D representation is a rapid area of research, such as voxels [27, 35], point clouds [1, 29] and implicit functions [28, 30, 36, 39]. DeepVoxels [35] and NeuralVolumes [27] use neural networks to predict voxel representations of scenes, where each voxel grid contains neural features or  $RGB\alpha$  values respectively. Zhou *et al.* [46] proposed to use multi-plane images, which is an voxelized representation for the camera frustum. When combined with volume rendering, these works enable predicting 3D properties but require only 2D supervision. Recent methods also show promising results with neural implicit functions [28, 30, 36], which represent the scene as a continuous volumetric function and approximate with a neural network. Our proposed Volumetric Spherical Gaussian draws inspiration from 3D scene representations and augments the  $RGB\alpha$  representation with view-dependent effects to better handle directional lighting.

## 3. Lighting Representation

To invert the rendering process, we require a flexible, structured representation of 3D spatially-varying lighting.

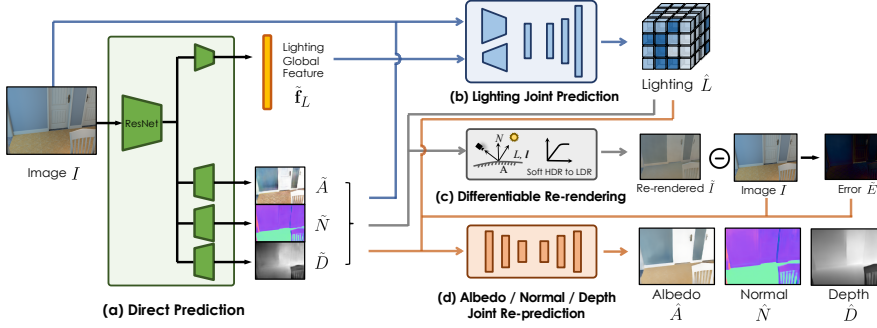


Figure 2: **Model overview.** Our model consists of 4 submodules (a-d). Direct Prediction Module (a) takes a single image as input and jointly predicts initial guess of intrinsic properties. Lighting Joint Prediction Module (b) consumes the initial prediction and predicts a 3D lighting volume. With Differentiable Re-rendering Module (c) re-renders the input image by raytracing, Joint Re-prediction Module (d) finally jointly refines the initial prediction.

Ideally, a model of the light field should capture variation in radiance due to changes in both spatial location and viewing angle. Prior works [11, 22] represent lighting with the radiance *incident* at the visible surface, making the direct extension to 3D intractable. To address this issue, we propose to use Volumetric Spherical Gaussian (VSG) to represent the surface radiance *exitant* from the full scene, including both visible surfaces and surfaces outside the FoV. Illumination at any spatial location and viewing angle can then be rendered using standard volume rendering techniques.

VSG is a voxel-based representation of a scene. We assign an opacity  $\alpha \in [0, 1]$  to each voxel, as well as a set of spherical Gaussian parameters  $\mathbf{c} \in \mathbb{R}^3$ ,  $\boldsymbol{\mu} \in \mathbb{R}^3$ ,  $\sigma \in \mathbb{R}_+$  such that the radiance at a viewing angle  $\mathbf{v} \in \mathbb{R}^3$  is defined:

$$G(\mathbf{v}; \mathbf{c}, \boldsymbol{\mu}, \sigma) = \mathbf{c}e^{-(1-\mathbf{v} \cdot \boldsymbol{\mu})/\sigma^2} \quad (1)$$

Intuitively, every voxel is a light source, where  $\mathbf{c}$  represents HDR RGB intensity,  $\boldsymbol{\mu}$  is the lobe axis and  $1/\sigma^2$  indicates sharpness. For a voxel grid of size  $X \times Y \times Z$ , VSG represents the lighting as an 8-channel tensor  $L \in \mathbb{R}^{8 \times X \times Y \times Z}$ .

To calculate incident radiance for a point  $\mathbf{p} \in \mathbb{R}^3$  with direction  $\mathbf{l} \in \mathbb{R}^3$  in a VSG  $L$ , we select  $N$  equi-spaced points along the ray. We then calculate the radiance  $\mathcal{R}(\mathbf{p}, \mathbf{l}, L) \in \mathbb{R}^3$  using alpha compositing

$$\mathcal{R}(\mathbf{p}, \mathbf{l}, L) = \sum_{k=1}^N \prod_{i=1}^{k-1} (1 - \alpha_i) \alpha_k G(-\mathbf{l}; \mathbf{c}_k, \boldsymbol{\mu}_k, \sigma_k). \quad (2)$$

where  $\mathbf{c}_k$ ,  $\boldsymbol{\mu}_k$ ,  $\sigma_k$  and  $\alpha_k$  are determined by indexing into the lighting volume  $L$ . With  $\alpha$ -channel indicating the opacity of voxels, we can also render “depth” with Eq. 2 by replacing the spherical Gaussian with the voxel depth values.

Compared to recent works that use an  $\text{RGB}\alpha$  volume to render appearance [30, 38], the VSG lighting representation additionally controls the emission direction and sharpness of light sources, and thus can capture view-dependent effects and handle strong directional lighting. Note that for  $\sigma \gg 1$ , our VSG reduces to an  $\text{RGB}\alpha$  representation.

## 4. Method

Our monocular inverse rendering model jointly estimates albedo, normals, depth and a 3D Volumetric Spherical Gaussian lighting representation. To jointly predict intrinsic prop-

erties and 3D VSG lighting from a single image, we split our pipeline into four submodules, with three neural network modules and one differentiable rendering module. The overall architecture is shown in Fig. 2.

First, the Direct Prediction Module makes an initial prediction of the intrinsic properties and extracts a global lighting feature from an input image. The Joint Prediction Module lifts these properties into 3D and jointly predicts a lighting volume. Then, the Differentiable Re-rendering Module re-renders the input image using the current predicted intrinsics and lighting. Finally, the Joint Re-prediction Module conditions on the current prediction and re-rendering error, and jointly refines the initial predictions.

We describe the architecture of each submodule in Sec. 4.1 and present our training scheme in Sec. 4.2.

### 4.1. Model Design

**Direct Prediction Module.** The purpose of the Direct Prediction Module is to make an initial prediction of the albedo, normal, depth and a global feature which encodes lighting information, given a single image as input.

The backbone of the Direct Prediction Module is a multi-branch ResNet [14]  $h_{\text{DP}}$  that consumes a single linearized RGB image  $I \in \mathbb{R}^{3 \times H \times W}$  and predicts albedo  $\hat{A} \in \mathbb{R}^{3 \times H \times W}$ , surface normals  $\tilde{N} \in \mathbb{R}^{3 \times H \times W}$ , depth  $\tilde{D} \in \mathbb{R}^{H \times W}$  and a global feature vector  $\tilde{\mathbf{f}}_L \in \mathbb{R}^C$  that is used by the downstream Lighting Joint Prediction Module as a “Lighting Global Feature”

$$\tilde{A}, \tilde{N}, \tilde{D}, \tilde{\mathbf{f}}_L = h_{\text{DP}}(I; \Theta_{\text{DP}}). \quad (3)$$

**Lighting Joint Prediction Module.** Unlike albedo, normals, and depth, our lighting representation as defined in Sec. 3 is volumetric:  $\hat{L} \in \mathbb{R}^{8 \times X \times Y \times Z}$ . Our lighting decoding network is shown in Fig. 3 and described below.

We extract features to predict the lighting volume from two different sources. The first source is the global feature vector  $\tilde{\mathbf{f}}_L$ . We use an MLP decoder  $h_{\text{GFD}}$  to map the global feature into a Scene Global Feature Volume. Let  $(x, y, z)$  be the center coordinates of a given voxel. The feature at the corresponding voxel is computed as:

$$\mathbf{z}_g = h_{\text{GFD}}(x, y, z, \tilde{\mathbf{f}}_L).$$

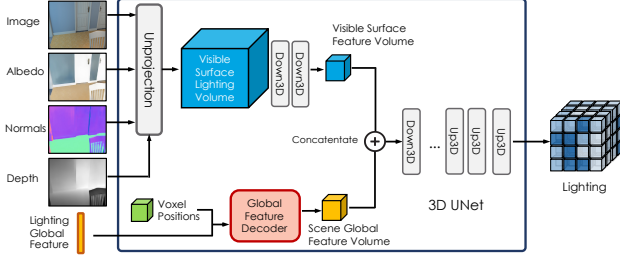


Figure 3: **Architecture of Lighting Joint Prediction Module.** We fuse the unprojected visible FoV information (top) and global scene information (bottom), and process them with a 3D UNet. The output is the Volumetric Spherical Gaussian lighting.

Compared to a sequence of 3D transpose convolution which can achieve similar functionality, this MLP module is more flexible and can naturally extend to multi-view input. We refer to Appendix for more implementation details.

The other sources of features for the lighting volume are the properties within the visible FoV, including input image  $I$ , predicted albedo  $\hat{A}$ , normal  $\hat{N}$ , and depth  $\hat{D}$ . We unproject this visible FoV information into a Visible Surface Lighting Volume and process it with a 3D UNet. Given camera intrinsics, let  $(u_p, v_p)$  be the projection of the center point of the voxel onto the input image with depth  $d_p$ , let  $\tilde{D}_p, \tilde{N}_p, \tilde{A}_p$  be the depth, normal, and albedo predicted by  $h_{DP}$  at the pixel  $(u_p, v_p)$ , and let  $I_p$  be the RGB values of the input image at that pixel. For each voxel, we define its “local” features as:

$$\mathbf{z}_l = kI_p, k\tilde{N}_p, k\tilde{A}_p \quad (4)$$

where  $k = e^{-\frac{(d_p - \tilde{D}_p)^2}{2\sigma_d^2}}$  is the Gaussian distance of depth between the voxel and the corresponding pixel, and  $\sigma_d$  is a hyper-parameter with units of length that we set to 0.15 meters. Intuitively, the factor  $k$  zeros out local features for voxels that are far from 2D surface manifold as determined by the depth output from  $h_{DP}$ .

We fuse the global features  $\mathbf{z}_g \in \mathbb{R}^{C \times X_g \times Y_g \times Z_g}$  and local features of the visible FoV  $\mathbf{z}_l \in \mathbb{R}^{9 \times X \times Y \times Z}$ , and process them through a 3D CNN  $h_{JP}$ :

$$\hat{L} = h_{JP}(\mathbf{z}_g, \mathbf{z}_l; \Theta_{JP}), \quad \hat{L} \in \mathbb{R}^{8 \times X \times Y \times Z} \quad (5)$$

where  $\hat{L}$  is the output VSG containing HDR intensity.

**Differentiable Re-rendering Module.** A valuable source of supervision and inductive bias in inverse rendering comes from the fact that the input image should be retrieved if the predicted geometry and reflectance is re-rendered.

We use the Lambertian reflectance model for re-rendering. For each pixel, let  $\mathbf{p} = (x_p, y_p, \tilde{D}_p)$  be the 3D location of the pixel  $p$  with predicted depth  $\tilde{D}_p$ . To compute lighting at each pixel, we choose  $K$  equi-angular lighting directions  $\{\mathbf{l}\}_K$  on the upper hemisphere with Fibonacci lattice [12]. For each lighting direction  $\mathbf{l}$ , we query the lighting volume and compute the radiance along the ray  $\mathcal{R}(\mathbf{p}, \mathbf{l}, \hat{L})$  with Eq. 2.

Let  $\tilde{A}_p$  and  $\tilde{N}_p$  be albedo and normal predicted in pixel  $p$ . We render LDR RGB values following Lambertian model:

$$\tilde{I}_p = \varphi \left( \sum_{\mathbf{l} \in \{\mathbf{l}\}_K} \frac{\tilde{A}_p}{\pi} \odot \mathcal{R}(\mathbf{p}, \mathbf{l}, \hat{L}) \max(\mathbf{l} \cdot \tilde{N}_p, 0) \Delta\Omega \right) \quad (6)$$

where  $\odot$  is an element-wise product,  $\frac{1}{\pi}$  is the energy conservation ratio and  $\Delta\Omega$  is the differential solid angle. Here,  $\varphi$  is the function to clip the HDR lighting intensity values to LDR values within  $[0, 1]$ . To make this process differentiable, we use a soft clipping with the exponential function:

$$\varphi(x) = \begin{cases} x & \text{if } x \leq \tau \\ 1 - (1 - \tau)e^{-\frac{x - \tau}{1 - \tau}} & \text{if } x > \tau \end{cases} \quad (7)$$

We use  $\tau = 0.9$  in our model.

**Joint Re-prediction Module.** Reflectance, shape and lighting are inherently correlated through the rendering process. To refine albedo, normals, and depth, we use a fully-convolutional network that takes as input the initial prediction  $\hat{A}, \hat{N}, \hat{D}$ , the re-rendering error  $\tilde{E} = I - \tilde{I}$ , and input image  $I$ . To incorporate the predicted lighting  $\hat{L}$ , we also concatenate shading  $\tilde{S}$ , which indicates how albedo changes affect the re-rendered image, and the Jacobian of shading with respect to the normals  $\frac{\partial \tilde{S}}{\partial \tilde{N}}$ , which indicates how the normals’ change affect the output shading. At each pixel  $p$ ,  $\tilde{S}$  and  $\frac{\partial \tilde{S}}{\partial \tilde{N}}$  have values given by the analytic formulas

$$\begin{aligned} \tilde{S}_p &= \sum_{\mathbf{l} \in \{\mathbf{l}\}_K} \mathcal{R}(\mathbf{p}, \mathbf{l}, \hat{L}) \max(\mathbf{l} \cdot \tilde{N}_p, 0) \Delta\Omega \\ \frac{\partial \tilde{S}}{\partial \tilde{N}_p} &= \sum_{\mathbf{l} \in \{\mathbf{l}\}_K} \mathbb{1}_{\mathbf{l} \cdot \tilde{N}_p > 0} \mathcal{R}(\mathbf{p}, \mathbf{l}, \hat{L}) \otimes \mathbf{l} \Delta\Omega \end{aligned} \quad (8)$$

where  $\otimes$  is the outer product. Note that much of the computation required for shading can be cached during Lambertian rendering given by Eq. 6. The fully-convolutional network predicts an updated albedo  $\hat{A}$ , normals  $\hat{N}$ , and depth  $\hat{D}$

$$\hat{A}, \hat{N}, \hat{D} = h_{JR}(I, \tilde{E}, \tilde{A}, \tilde{N}, \tilde{D}, \tilde{S}, \frac{\partial \tilde{S}}{\partial \tilde{N}}). \quad (9)$$

## 4.2. Training

We train our model on synthetic data with groundtruth  $\{I, A, N, D, \{I_{nv}, P_{nv}\}_N\}$ , where  $A, N, D$  denote albedo, normals, depth, and  $\{I_{nv}, P_{nv}\}_N$  are LDR panoramic images and camera poses from  $N$  novel views. With a volumetric lighting representation, we not only eliminate the need for the densely rendered spherical lobe lighting GT used in past works [22, 26], but also improve angular frequency.

The loss for training comes from two parts: (1) *direct supervision*, which directly enforces consistency with the synthetic groundtruth, and (2) *re-rendering loss* that encourages the re-rendered image to recover the input image.



**Direct Supervision for Reflectance and Shape.** We use L2 loss for albedo. Since albedo is usually piecewise constant, we additionally penalize the gradient of albedo where groundtruth is locally constant. We define

$$\mathcal{L}_{\text{albedo}} = \|A - \hat{A}\|_2^2 + \lambda_{\text{local}} \|\nabla \hat{A} \odot M_{\text{local}}\|_1 \quad (10)$$

where  $M_{\text{local}} \in \mathbb{R}^{H \times W}$  is a mask indicating regions in the ground-truth albedo where the albedo is constant.

For normals, the network output is normalized and we use L1 angular error as supervision:

$$\mathcal{L}_{\text{normal}} = \left\| \cos^{-1} \left( N \cdot \frac{\hat{N}}{|\hat{N}|} \right) \right\|_1. \quad (11)$$

Because depth is high dynamic range, we follow [22] and use log-encoded L2 loss. We also use scale invariant L2 loss to encourage relative consistency due to the inherent scale ambiguity of depth:

$$\mathcal{L}_{\text{depth}} = \left\| \log(D + 1) - \log(\hat{D} + 1) \right\|_2^2 + \lambda_{\text{si}} \|D - c_{\text{si}} \hat{D}\|_2^2 \quad (12)$$

where  $c_{\text{si}} = \arg \min_c \|D - c\hat{D}\|_2^2 = \frac{\sum_p \hat{D}_p \cdot D_p}{\sum_p \hat{D}_p \cdot \hat{D}_p}$  is a scale factor computed on-the-fly for each image.

**Direct Supervision for Lighting.** Recall that the pixel values in LDR images reflect the HDR radiance along the corresponding camera ray, after intensity clipping. Thus, we can use photometric loss of LDR panoramic images  $I_{\text{nv}}$  to supervise the LDR part of the predicted lighting  $\hat{L}$ .

For each pixel  $p$ , we use the camera pose  $P_{\text{nv}}$  and camera intrinsics to compute the corresponding camera ray starting from the camera center  $c$  in the direction  $r$ . To render the novel view  $\hat{I}_{\text{nv},p}$  using the predicted lighting volume  $\hat{L}$ , we compute HDR radiance with Eq. 2, and convert to LDR using a soft clipping function  $\varphi$  defined in Eq. 7:

$$\hat{I}_{\text{nv},p} = \varphi(\mathcal{R}(c, r, \hat{L})) \quad (13)$$

We enforce this rendered novel view  $\hat{I}_{\text{nv}}$  to be consistent with groundtruth  $I_{\text{nv}}$  using L2 loss. To encourage realistic details, we also use the adversarial loss  $\mathcal{L}_{\text{adv}}$  with a discriminator  $\mathcal{D}$ .

$$\begin{aligned} \mathcal{L}_{\text{light}} &= \mathcal{L}_{\text{nv}} + \lambda_{\text{adv}} \mathcal{L}_{\text{adv}} \\ &= \|I_{\text{nv}} - \hat{I}_{\text{nv}}\|_2^2 - \lambda_{\text{adv}} \mathcal{D}(\hat{I}_{\text{nv}}) \end{aligned} \quad (14)$$

The loss for the discriminator  $\mathcal{D}$  is

$$\mathcal{L}_{\mathcal{D}} = \max(0, 1 - \mathcal{D}(I_{\text{nv}})) + \max(0, 1 + \mathcal{D}(\hat{I}_{\text{nv}})). \quad (15)$$

Another source of supervision for lighting is consistency with visible FoV, *i.e.* the image  $I$  and depth  $D$ . We define  $\mathcal{L}_{\text{visible}}$  as the L2 loss between  $I$ ,  $D$  and the rendered perspective RGB image and depth ( $\alpha$ -channel) from the lighting volume using Eq. 2. Since the surface is sparse in the scene, we also encourage the  $\alpha$ -channel of the lighting volume to be either 0 or 1 with a regularization loss  $\mathcal{L}_{\text{reg}} = -\alpha \log(\alpha)$ . The training signal is fully differentiable and backpropagated to supervise the predicted VSG lighting parameters.

**Re-rendering Loss.** We use the predicted albedo, normals and lighting to reconstruct an image  $\hat{I}$  and enforce its consistency with the original input image  $I$ . Specifically, we compute Lambertian re-rendered image in the energy conserved form defined in Eq. 6, and use L2 loss as the re-rendering loss  $\mathcal{L}_{\text{re-render}} = \|I - \hat{I}\|_2^2$ .

The re-rendering loss encourages joint reasoning of albedo, geometry and lighting, and can be used for self-supervised training on real world images as discussed in prior works [33]. In our formulation, the re-rendering loss significantly improves lighting prediction by enforcing the model to learn physically correct lighting and to recover HDR information. This benefit comes from the formulation of the energy-conserving image formation process. Any physically incorrect lighting prediction, such as a LDR prediction or a uniform lighting, will lead to an error in the re-rendered image. In Eq. 14, we only supervise LDR lighting appearance with LDR images. With the complementary re-rendering loss, our model automatically learns to recover HDR lighting even when trained only with LDR images.

**Training Scheme.** Our model is end-to-end trainable. We adopt a progressive training scheme to ensure the model components act as expected. We first pretrain albedo, normal and depth branches of our Direct Prediction Module. This is because our Lighting Joint Prediction Module (Fig. 3) depends on these properties, and pretraining these branches can make sure they produce reasonable values. Then we jointly train the Direct Prediction Module and Lighting Joint Prediction Module with the multi-task loss

$$\begin{aligned} \mathcal{L} &= \lambda_A \mathcal{L}_{\text{albedo}} + \lambda_N \mathcal{L}_{\text{normal}} + \lambda_D \mathcal{L}_{\text{depth}} + \lambda_L \mathcal{L}_{\text{light}} \\ &\quad + \lambda_{\text{visible}} \mathcal{L}_{\text{visible}} + \lambda_{\text{reg}} \mathcal{L}_{\text{reg}} + \lambda_{\text{re-render}} \mathcal{L}_{\text{re-render}}. \end{aligned} \quad (16)$$

After the first two submodules are trained, we freeze their weights and train the Joint Re-prediction Module for albedo, normals and depth. Finally, we jointly finetune all three modules end-to-end with a multi-task loss on both the the joint prediction  $\hat{A}, \hat{N}, \hat{D}, \hat{L}$  and intermediate output  $\tilde{A}, \tilde{N}, \tilde{D}$ .

## 5. Experiments

We compare our method with prior methods both qualitatively and quantitatively, and validate the effectiveness of our unified inverse rendering framework. We also compare against prior methods on lighting estimation and showcase application on virtual object insertion, demonstrating our method’s ability to create high-quality insertion results.

### 5.1. Experiment Settings

**Implementation Details.** The resolution of initialized visible surface volume and predicted VSG lighting is  $128^3$ . Despite using a  $128^3$  volume, our model is much lighter than Lighthouse [38] which contains six 3D UNet subnetworks. With a batch size of 1, our model consumes 7.5G

Method	Albedo si-MSE	Normal Angular Error	Depth si-MSE
SIRFS [2]	0.0453	56.75°	-
NIR [33]	0.0188	20.35°	-
Ours (w/o JR)	0.0190	19.09°	0.217
Ours (JR w/o lighting)	0.0177	18.63°	0.189
Ours	<b>0.0175</b>	<b>18.40°</b>	<b>0.181</b>

Table 1: Evaluation of albedo, normals and depth on InteriorNet dataset.

Method	PSNR (dB)
NIR [33]	15.39
Lighthouse* [38]	17.29
Ours ( $\mathcal{L}_{\text{albedo}}, \mathcal{L}_{\text{normal}}, \mathcal{L}_{\text{depth}}, \mathcal{L}_{\text{light}}$ only)	16.43
+ $\mathcal{L}_{\text{visible}}$	17.06
+ $\mathcal{L}_{\text{reg}}$	17.33
+ $\mathcal{L}_{\text{re-render}}$	<b>17.37</b>
Ours (w/o SG)	16.94

Table 2: Evaluation of lighting on InteriorNet dataset. \* indicates use of a stereo pair as input.

Method	WHDR
SIRFS [2]	31.4
NIR [33]	18.5
Ours (w/o JR)	18.7
Ours	<b>18.2</b>

Table 3: Evaluation of albedo on IIW dataset.

Method	Normal Angular Error	Depth si-MSE
NIR [33]	23.94°	0.3216
Ours (w/o JR)	23.89°	0.3196
Ours	<b>22.95°</b>	<b>0.2827</b>

Table 4: Evaluation of normals and depth on NYUv2 dataset.

Re-rendering MSE ( $\times 10^{-2}$ )	InteriorNet [21]	NYUv2 [31]
NIR (env. map only) [33]	2.36	4.02
NIR [33]	0.99	2.61
Ours (w/o Re-render Loss)	2.18	5.26
Ours (w/o SG)	1.41	2.72
Ours	<b>0.89</b>	<b>2.33</b>
Ours (w/ real-world tuning)	0.92	<b>1.98</b>

Table 5: Quantitative results of re-rendering error.

GPU memory compared to Lighthouse’s 15G during training. For Differentiable Re-rendering Module, we sample  $K = 50$  rays per pixel and  $N = 128$  points per ray. For each pixel, we share the rays of its 8-neighbors to get  $K' = 450$  rays per pixel. The re-rendering is implemented considering parallelism. Resolution of re-rendering is 60x80 during training. Inference times of Direct Prediction, Lighting Joint Prediction, Differentiable Re-rendering and Joint Re-prediction Modules are 20ms, 130ms, 140ms, 12ms respectively, clocked on a TITAN V GPU.

**Training Data.** We train our model on the InteriorNet dataset [21], which contains realistic renderings of camera sequences in diverse indoor scenes. Each camera sequence contains 1000 rendered LDR perspective images with albedo, normals and depth groundtruth, and 1000 LDR panoramic images rendered in the corresponding viewpoint using path tracing. We use LDR perspective images as input and supervise albedo and geometry with paired GT. To supervise lighting, we sample adjacent panoramic images rendered at locations that are visible in the perspective input image. This makes the environment maps lie in our region of interest, *i.e.* in front of the camera. We follow the data split and preprocessing from [38]. We use 90% (1472) of the scenes to train our model and reserve 10% (162) for evaluation.

When evaluating on real world data, we also finetune our model on IIW dataset [6] for albedo and NYUv2 dataset [31] for depth and normal. We also collected 120 indoor LDR panoramas from the internet and jointly train lighting on these panoramas. More details are included in the Appendix.

**Evaluation.** We evaluate albedo, normals and depth prediction on InteriorNet [21] and real-world datasets IIW [6] and NYUv2 [31]. For quantitative comparison, we use scale-invariant MSE (si-MSE) for albedo and depth due to scale ambiguity. We use mean angular error for normals and

PSNR for lighting. We report the re-rendering MSE between re-rendered image and input image, which indicates whether the prediction is physically correct. The most effective lighting evaluation is through qualitative results. We compare with prior works by visualizing object insertion results and the predicted environment map at a given location.

**Baselines.** We quantitatively compare our methods with the state-of-the-art NIR [33] and the classic optimization-based method SIRFS [2]. In all experiment settings, we re-train NIR on the same data, *i.e.* InteriorNet, to ensure a fair comparison. Li *et al.* [22] requires dense per-pixel lighting supervision and cannot train on the same data sources, and thus we provide qualitative comparison on the lighting prediction. For lighting estimation, we also compare with current state-of-the-art method Lighthouse [38], which uses a stereo image pair as input instead of a monocular image.

## 5.2. Evaluation of Albedo and Shape

**Evaluation on InteriorNet.** We compare with baseline methods and ablate our model choices on InteriorNet. As shown in Table 1, the performance of SIRFS exemplifies the limits of optimization-based methods on images of complex scenes. Our method outperforms NIR, indicating that our method better disambiguates the intrinsic properties. We also ablate our method. Predicted output from Direct Prediction Module is shown as “Ours (w/o JR)”. Results show that the Joint Re-prediction Module helps improve the performance with the benefits of joint reasoning over the initial prediction. By comparing “Ours” and “Ours (JR w/o lighting)”, the quantitative evaluation shows that the properties related to lighting (Eq. 8) helps improve the overall performance. We provide qualitative results of albedo, normals and depth in Fig. 4. Results show that the Joint Re-prediction Module further disambiguates the intrinsic properties based on initial prediction and produces higher quality prediction.

**Evaluation on real-world data.** We evaluate albedo prediction on the IIW dataset [6], which provides sparse pairwise human annotations. We use the official Weighted Human Disagreement Rate (WHDR) as the metric, which measures the error when albedo prediction disagrees with human perception, and the results are shown in Table 3. We also evaluate normal and depth prediction on the NYUv2 dataset in Table 4. Our method outperforms baselines and validates the effectiveness of our Joint Re-prediction Module. Similar to prior works [22, 33], we focus on the holistic inverse rendering framework and do not compete with state-of-

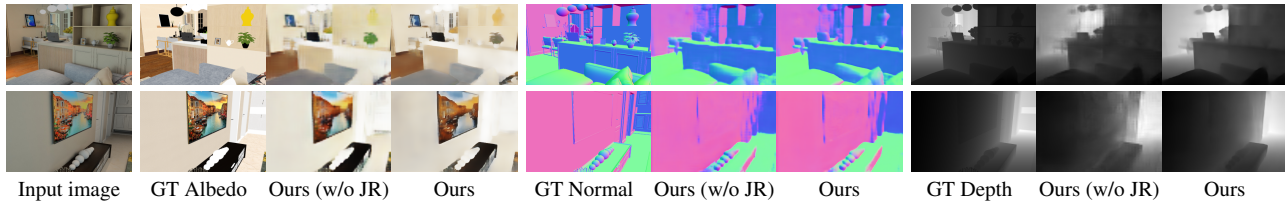


Figure 4: **Qualitative results of predicted albedo, normals and depth.** The results are GT, our model without Joint Re-prediction (JR) Module and our full model. Joint Re-prediction enables joint reasoning and obtains crisper and more accurate results.

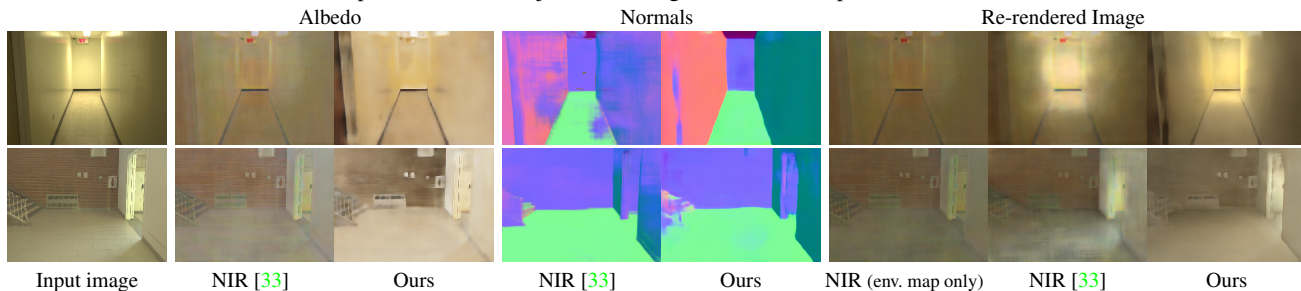


Figure 5: **Qualitative comparison on predicted albedo, normals and re-rendered image.** Our fully physics-based lighting representation and differentiable renderer can better disambiguate and reproduce complex lighting effects with less artifacts.

the-art depth and normal estimation methods with specially designed architecture and other data sources [15, 19].

**Evaluation of re-rendered images.** We compare re-rendering error with baselines on InteriorNet and NYUv2 in Table 5. NIR uses environment map as lighting representation, and employs a non-interpretable neural renderer, named residual appearance renderer (RAR), to account for all other lighting effects. Our method outperforms NIR both with and without the neural rendering module. We also show qualitative results of re-rendered images in Fig. 5. Compared to the environment map re-rendering of NIR, we can handle 3D spatially-varying lighting and can re-render complex lighting effects, while NIR leaves these effects to the RAR module. Though addressing a more challenging task, our fully physics-based rendering process outperforms the RAR module which may easily produce artifacts and hurt the performance of other properties, such as artifacts in normals.

We ablate our design choices in Table 5. The re-rendering loss, which enforces joint reasoning of different properties, is critical for physically correct predictions. By comparing our Volumetric Spherical Gaussian with  $RGB\alpha$  volume (Ours w/o SG), the results show that the spherical Gaussian volume increases the model capacity and leads to better re-rendering. We also tried jointly training on InteriorNet and 120 real-world LDR panoramas. We show that when evaluating on the real-world NYUv2 dataset, training on real-world LDR panoramas further improves performance.

### 5.3. Evaluation of Lighting

**Quantitative Evaluation.** We evaluate our lighting prediction on InteriorNet in Table 2. Our method significantly outperforms NIR [33] due to our 3D lighting representation that can handle spatially-varying lighting. Our method also outperforms Lighthouse [38], the current state-of-the-art

method for lighting estimation. Note that Lighthouse uses stereo pairs as input, which provides more information about depth and visible surface than monocular image. For the ablation study, empirical results indicate that our loss design is key in achieving the best performance. Compared to  $RGB\alpha$  volume, our Volumetric Spherical Gaussian is able to capture view-dependent effects and brings better performance.

**Qualitative Evaluation.** We qualitatively compare the lighting estimation and virtual object insertion results with baselines in Fig. 6 on InteriorNet. Note that the inserted light probe is highly specular. NIR uses a single low-resolution environment map, which cannot handle spatially-varying effects and only recovers low frequency lighting, thus causing severe artifacts. Li *et al.* [22] employs 2D spatially-varying spherical Gaussian which can produce spatially-varying lighting, but the local lighting is still low frequency spherical lobes and cannot account for angular high-frequency details. These methods fail for inserting highly specular objects. Lighthouse [38] is using a volumetric  $RGB\alpha$  lighting representation, which allows for 3D spatially-varying lighting. But the lighting volume in Lighthouse is learned with voxel inpainting from LDR panoramas, without supervision signal to make it *physically correct*, and cannot predict HDR lighting. We can observe Lighthouse’s lighting prediction has significantly less intensity variation, and the predicted LDR lighting cannot produce realistic cast shadows.

Our method, with the Volumetric Spherical Gaussian lighting, produces more visually pleasing results with more realistic details. The re-rendering loss in our method enables joint reasoning and ensures that the model predicts physically correct HDR outputs due to the energy constraints. Our method is the only one that both preserves angular details and predicts HDR output for realistic cast shadows. Note the predicted cast shadows by our method are consistent with



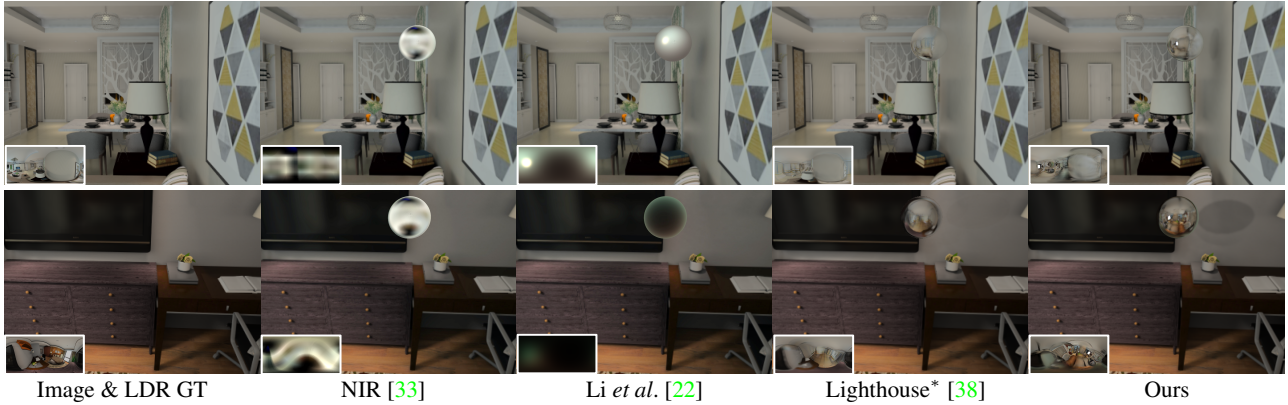


Figure 6: **Qualitative comparison of lighting estimation.** We compare insertion of a *purely specular* sphere, and on the bottom-left of each example displays the estimated environment map at the inserted location. Our method produces both angular details (env. map) and realistic cast shadows with HDR, outperforming all competing methods. (\* indicates use of a stereo pair as input. Best viewed zooming in. )

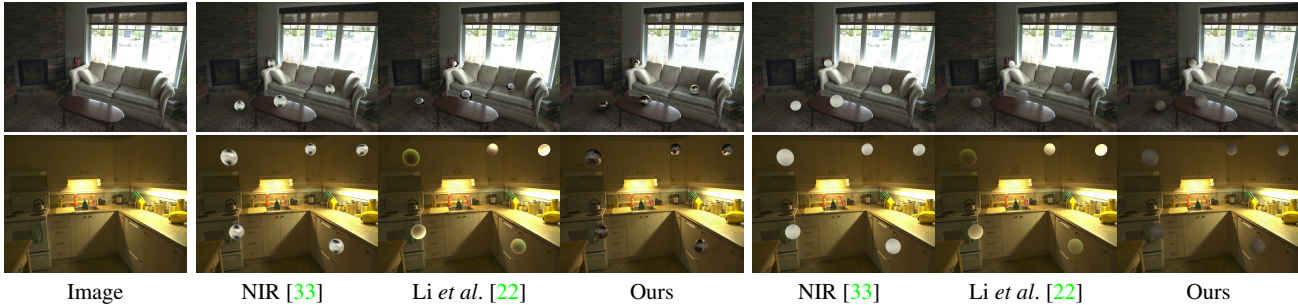


Figure 7: **Qualitative comparison of lighting estimation on real-world images.** We compare *purely specular* object insertion on the left, and on the right is mostly diffuse object. The top row shows insertion on a solid surface while bottom row shows freely inserted objects in 3D. Our method produces more realistic results in both specular and diffuse settings and is spatially consistent. (Best viewed zooming in. )



Figure 8: **Qualitative results of object insertion on real-world images.** From left to right, we insert a bunny, kettle, cart and armchair.

other visual cues in the scene (Top: lamp; Bottom: desk).

We compare with prior works on real world data in Fig. 7. Here, we do not compare with Lighthouse as it requires stereo images. Comparing purely specular spheres on the left, our method preserves angular high-frequency details and is significantly more realistic. These benefits also apply to the diffuse spheres on the right. In the bottom row, spheres are from randomly sampled 3D locations. NIR [34] uses a single environment map and can hardly capture spatial variation. Li *et al.* [22] uses 2D lighting representation, which cannot produce spatially consistent lighting and also cannot handle locations far from the 2D surfaces, leading to severe artifacts. Our method produces spatially coherent lighting with correct HDR intensity. We also show real-world object insertion results in Fig. 8. Our method generalizes well to real-world images and consistently produces realistic shading and shadows. More results are included in the Appendix.

## 6. Conclusion

In this paper, we proposed a holistic monocular inverse rendering framework that jointly estimates albedo, normals, depth, and HDR light field. Our proposed Volumetric Spherical Gaussian representation nicely handles spatial and angular high-frequency details. With the physically based differentiable renderer, our method is capable of learning and reproducing the complex indoor lighting effects, and better disambiguate image intrinsics. Benefiting from joint training with the re-rendering constraint, our model can predict physically correct HDR lighting, despite being trained with only LDR images. We experimentally demonstrate that our model outperforms prior work on standard benchmarks, and is able to realistically render virtual objects in images with realistic shadows and high-frequency details, even when the objects are highly specular. We believe that these results demonstrate great promise of our model for AR applications.



## References

- [1] Kara-Ali Aliev, Dmitry Ulyanov, and Victor Lempitsky. Neural point-based graphics. *arXiv preprint arXiv:1906.08240*, 2(3):4, 2019. [2](#)
- [2] 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. [1](#), [2](#), [6](#)
- [3] Harry Barrow, J Tenenbaum, A Hanson, and E Riseman. Recovering intrinsic scene characteristics. *Comput. Vis. Syst.*, 2:3–26, 1978. [1](#), [2](#)
- [4] H. Barrow and J. M. Tenenbaum. Recovering intrinsic scene characteristics from images. 1978. [2](#)
- [5] Anil S Baslamisli, Thomas T Groenestege, Partha Das, Hoang-An Le, Sezer Karaoglu, and Theo Gevers. Joint learning of intrinsic images and semantic segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 286–302, 2018. [1](#)
- [6] Sean Bell, Kavita Bala, and Noah Snavely. Intrinsic images in the wild. *ACM Transactions on Graphics (TOG)*, 33(4):159, 2014. [2](#), [6](#)
- [7] Mark Boss, Varun Jampani, Kihwan Kim, Hendrik Lensch, and Jan Kautz. Two-shot spatially-varying brdf and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3982–3991, 2020. [2](#)
- [8] Adrien Bousseau, Sylvain Paris, and Frédo Durand. User-assisted intrinsic images. In *ACM Transactions on Graphics (TOG)*, volume 28, page 130. ACM, 2009. [2](#)
- [9] Marc-André Gardner, Yannick Hold-Geoffroy, Kalyan Sunkavalli, Christian Gagné, and Jean-François Lalonde. Deep parametric indoor lighting estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7175–7183, 2019. [2](#)
- [10] 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. *arXiv preprint arXiv:1704.00090*, 2017. [1](#), [2](#)
- [11] Mathieu Garon, Kalyan Sunkavalli, Sunil Hadap, Nathan Carr, and Jean-François Lalonde. Fast spatially-varying indoor lighting estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6908–6917, 2019. [1](#), [2](#), [3](#)
- [12] Álvaro González. Measurement of areas on a sphere using fibonacci and latitude–longitude lattices. *Mathematical Geosciences*, 42(1):49–64, 2010. [4](#)
- [13] R. Grosse, M. K. Johnson, E. H. Adelson, and W. T. Freeman. Ground truth dataset and baseline evaluations for intrinsic image algorithms. In *2009 IEEE 12th International Conference on Computer Vision*, pages 2335–2342, 2009. [2](#)
- [14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. [3](#)
- [15] Steven Hickson, Karthik Raveendran, Alireza Fathi, Kevin Murphy, and Irfan Essa. Floors are flat: Leveraging semantics for real-time surface normal prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. [7](#)
- [16] Yannick Hold-Geoffroy, Akshaya Athawale, and Jean-François Lalonde. Deep sky modeling for single image outdoor lighting estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6927–6935, 2019. [2](#)
- [17] Yannick Hold-Geoffroy, Kalyan Sunkavalli, Sunil Hadap, Emiliano Gambaretto, and Jean-François Lalonde. Deep outdoor illumination estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7312–7321, 2017. [2](#)
- [18] Edwin H Land and John J McCann. Lightness and retinex theory. *Josa*, 61(1):1–11, 1971. [2](#)
- [19] Jin Han Lee, Myung-Kyu Han, Dong Wook Ko, and Il Hong Suh. From big to small: Multi-scale local planar guidance for monocular depth estimation. *arXiv preprint arXiv:1907.10326*, 2019. [7](#)
- [20] Chloe LeGendre, Wan-Chun Ma, Graham Fyffe, John Flynn, Laurent Charbonnel, Jay Busch, and Paul Debevec. Deeplight: Learning illumination for unconstrained mobile mixed reality. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5918–5928, 2019. [2](#)
- [21] Wenbin Li, Sajad Saeedi, John McCormac, Ronald Clark, Dimos Tzoumanikas, Qing Ye, Yuzhong Huang, Rui Tang, and Stefan Leutenegger. Interiornet: Mega-scale multi-sensor photo-realistic indoor scenes dataset. In *British Machine Vision Conference (BMVC)*, 2018. [6](#)
- [22] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2475–2484, 2020. [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#)
- [23] Zhengqi Li and Noah Snavely. Cgintrinsics: Better intrinsic image decomposition through physically-based rendering. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 371–387, 2018. [2](#)
- [24] Zhengqi Li and Noah Snavely. Learning intrinsic image decomposition from watching the world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9039–9048, 2018. [2](#)
- [25] Zhengqin Li, Zexiang Xu, Ravi Ramamoorthi, Kalyan Sunkavalli, and Manmohan Chandraker. Learning to reconstruct shape and spatially-varying reflectance from a single image. *ACM Transactions on Graphics (TOG)*, 37(6):1–11, 2018. [2](#)
- [26] Zhengqin Li, Ting-Wei Yu, Shen Sang, Sarah Wang, Sai Bi, Zexiang Xu, Hong-Xing Yu, Kalyan Sunkavalli, Miloš Hašan, Ravi Ramamoorthi, et al. Openrooms: An end-to-end open framework for photorealistic indoor scene datasets. *arXiv preprint arXiv:2007.12868*, 2020. [4](#)
- [27] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. Neural volumes: Learning dynamic renderable volumes from images. *arXiv preprint arXiv:1906.07751*, 2019. [2](#)

- [28] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4460–4470, 2019. 2
- [29] Moustafa Meshry, Dan B Goldman, Sameh Khamis, Hugues Hoppe, Rohit Pandey, Noah Snavely, and Ricardo Martin-Brualla. Neural rerendering in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6878–6887, 2019. 2
- [30] 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. *arXiv preprint arXiv:2003.08934*, 2020. 2, 3
- [31] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgb-d images. In *ECCV*, 2012. 6
- [32] Geoffrey Oxholm and Ko Nishino. Shape and reflectance from natural illumination. In *European Conference on Computer Vision*, pages 528–541. Springer, 2012. 2
- [33] Soumyadip Sengupta, Jinwei Gu, Kihwan Kim, Guilin Liu, David W. Jacobs, and Jan Kautz. Neural inverse rendering of an indoor scene from a single image. In *International Conference on Computer Vision (ICCV)*, 2019. 1, 2, 5, 6, 7, 8
- [34] Soumyadip Sengupta, Jinwei Gu, Kihwan Kim, Guilin Liu, David W. Jacobs, and Jan Kautz. Neural inverse rendering of an indoor scene from a single image. In *ICCV*, 2019. 8
- [35] Vincent Sitzmann, Justus Thies, Felix Heide, Matthias Nießner, Gordon Wetzstein, and Michael Zollhofer. Deepvoxels: Learning persistent 3d feature embeddings. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2437–2446, 2019. 2
- [36] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. In *Advances in Neural Information Processing Systems*, pages 1121–1132, 2019. 2
- [37] Shuran Song and Thomas Funkhouser. Neural illumination: Lighting prediction for indoor environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6918–6926, 2019. 2
- [38] Pratul P Srinivasan, Ben Mildenhall, Matthew Tancik, Jonathan T Barron, Richard Tucker, and Noah Snavely. Light-house: Predicting lighting volumes for spatially-coherent illumination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8080–8089, 2020. 1, 2, 3, 5, 6, 7, 8
- [39] Towaki Takikawa, Joey Litalien, Kangxue Yin, Karsten Kreis, Charles Loop, Derek Nowrouzezahrai, Alec Jacobson, Morgan McGuire, and Sanja Fidler. Neural geometric level of detail: Real-time rendering with implicit 3D shapes. *arXiv preprint arXiv:2101.10994*, 2021. 2
- [40] Xin Wei, Guojun Chen, Yue Dong, Stephen Lin, and Xin Tong. Object-based illumination estimation with rendering-aware neural networks. *arXiv preprint arXiv:2008.02514*, 2020. 2
- [41] Ye Yu and William AP Smith. Inverserendernet: Learning single image inverse rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2
- [42] Jinsong Zhang, Kalyan Sunkavalli, Yannick Hold-Geoffroy, Sunil Hadap, Jonathan Eisenman, and Jean-François Lalonde. All-weather deep outdoor lighting estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10158–10166, 2019. 2
- [43] Ruo Zhang, Ping-Sing Tsai, James Edwin Cryer, and Mubarak Shah. Shape-from-shading: a survey. *IEEE transactions on pattern analysis and machine intelligence*, 21(8):690–706, 1999. 2
- [44] Qi Zhao, Ping Tan, Qiang Dai, Li Shen, Enhua Wu, and Stephen Lin. A closed-form solution to retinex with nonlocal texture constraints. *IEEE transactions on pattern analysis and machine intelligence*, 34(7):1437–1444, 2012. 2
- [45] Yiqin Zhao and Tian Guo. Pointar: Efficient lighting estimation for mobile augmented reality. *arXiv preprint arXiv:2004.00006*, 2020. 2
- [46] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018. 2