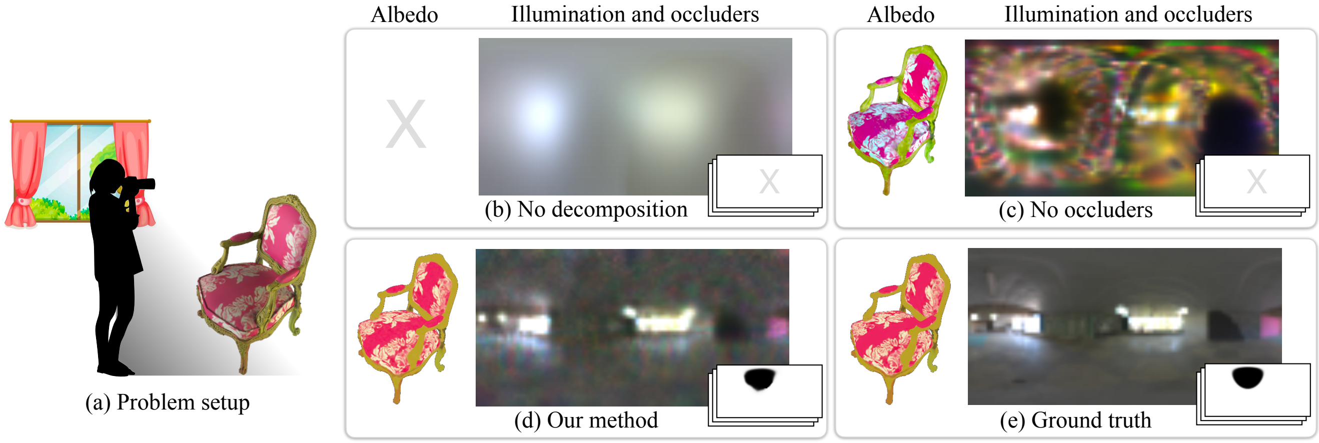
Eclipse: Disambiguating Illumination and Materials using Unintended Shadows

Dor Verbin[1] Ben Mildenhall[1] Peter Hedman[1]
Jonathan T. Barron[1] Todd Zickler[1,2] Pratul P. Srinivasan[1]
[1]Google Research [2]Harvard University



(a) Problem setup

(b) No decomposition

(c) No occluders

(d) Our method

(e) Ground truth

Figure 1. (a) We exploit unintended shadows cast by camera operators (or other unobserved moving occluders) to recover high-fidelity environment lighting and object materials from a set of images. Without modeling such unobserved occluders, prior methods can only: (b) recover convolved lighting without explicit material decomposition [31]; or (c) exploit lighting occlusions that occur internally among the observed objects [37]. (d) We show that additionally modeling and recovering external, unobserved occluders enables lighting and material reconstructions that are closer to ground truth (e).

Abstract

Decomposing an object's appearance into representations of its materials and the surrounding illumination is difficult, even when the object's 3D shape is known beforehand. This problem is especially challenging for diffuse objects: it is ill-conditioned because diffuse materials severely blur incoming light, and it is ill-posed because diffuse materials under high-frequency lighting can be indistinguishable from shiny materials under low-frequency lighting. We show that it is possible to recover precise materials and illumination—even from diffuse objects—by exploiting unintended shadows, like the ones cast onto an object by the photographer who moves around it. These shadows are a nuisance in most previous inverse rendering pipelines, but here we exploit them as signals that improve conditioning and help resolve material-lighting ambiguities. We present a method based on differentiable Monte Carlo ray tracing that uses images of an object to jointly recover its spatially-varying materials, the surrounding illumination environment, and the shapes of the unseen light occluders who inadvertently cast shadows upon it.

1. Introduction

In this work, we show that the long-standing inverse rendering problem of recovering an object's material properties and the surrounding illumination environment from a set of images can greatly benefit from considering the effects of unobserved moving light occluders, such as the camera operator, which partially block incoming light and inadvertently cast shadows onto the object being imaged.

Joint recovery of materials and lighting is a challenging task because the BRDF (bidirectional reflectance distribution function, which represents how a material at any point on an object's surface maps from incoming to outgoing light) acts as a directional filter on incoming light. BRDFs for specular (shiny) materials act as all-pass directional filters, while BRDFs for diffuse materials act as low-pass directional filters. As a result, inverse rendering is fundamentally ambiguous since a shiny object illuminated by blurry lighting can be indistinguishable from a diffuse object illuminated by sharp lighting (Figure 2). Furthermore, since diffuse materials act as low-pass filters and strongly blur incoming light, even the simpler problem of recovering lighting from images of a known diffuse material can be severely ill-conditioned, precluding the recovery of high-frequency illumination [33].

We observe that both of these issues can be ameliorated by exploiting an effect that naturally exists whenever an object is imaged from a sequence of viewpoints under static environment lighting. Namely, between one image and the next, the positions of the camera and its operator(s) must change, and in doing so they occlude different portions of the surrounding environment, casting distinct sets of shadows onto the object being imaged. This effect is very noticeable under strongly-directional lighting or for shiny objects because the cast shadows are sharp or the specular highlights are misshapen and hidden. But even when the effect is barely perceptible, such as for diffuse objects under well-distributed lighting, there exists a subtle signal that one can take advantage of.

In this paper, we make first steps toward a practical algorithm that exploits this cue by using gradient descent with differentiable Monte Carlo ray tracing to jointly recover explicit representations of (i) spatially-varying reflectance, (ii) environment illumination, and (iii) the per-image shapes of any unseen, light-blocking occluders. We evaluate our algorithm using a challenging variety of simulated scenes, including scenes with diffuse-only materials and with object geometry that is not known beforehand. We also apply our method to an existing off-the-shelf dataset designed for inverse rendering, where the (unseen) camera rig moving to capture the scene acts as an accidental occluder. Our results demonstrate that external shadowing effects provide a strong and useful cue, even when they are subtle, and even without relying on strong domain-specific priors for the materials, illumination, or external occluder shapes. This suggests that incorporating unintended shadows into inverse rendering pipelines for real, captured data is valuable for improving the quality of material and lighting assets.

2. Related Work

We build on recent developments in differentiable ray tracing and on a long history of inverse rendering work, including the joint estimation of reflectance and lighting, and the estimation of lighting alone. We are also inspired by a separate line of work on passive non-line-of-sight imaging, which exploits similar light-blocking effects.

Physics-based differentiable ray tracing. Modern differentiable ray tracers [22, 28] enable the computation of gradients of rendered images with respect to scene parameters (*i.e*., geometry, materials, and lighting) by differentiating through light transport simulation. Recent works have focused on improving efficiency and performance [14, 29] and on accurately differentiating through visibility discontinuities—such as those caused by shadow and occlusion boundaries—with respect to shape and lighting [1, 42]. Our implementation of differentiable ray tracing leverages insights from these works as well as from Zeltner *et al*. [45], who provide design intuition for Monte Carlo differentiable ray tracers.

Inverse rendering of materials and lighting. Decomposing an object's appearance into representations of material and lighting is a long-standing problem in computer vision and graphics. In their foundational work, Ramamoorthi and Hanrahan [33] developed a signal processing approach by describing the outgoing light at a surface point as the spherical convolution of the BRDF and the incoming lighting. This formulation elucidates why inverse rendering is ill-posed and often ill-conditioned: it is ill-posed because there are multiple illumination-material pairs that convolve into the same image, and it is ill-conditioned because diffuse BRDFs act as a low-pass filters on lighting, which causes the estimates of medium- and high-frequency illumination to be very sensitive to noise.

Because of this, most approaches to inverse rendering rely on strong priors on materials and lighting. Single-view approaches have used hand-designed priors [3] or priors in the form of neural networks trained with supervision from large datasets of material and lighting labels [23, 24]. More relevant to us are multi-view approaches, which typically either assume known lighting [4, 5, 36] or rely on strong priors, such as assuming a single, highly-specular BRDF for the entire object [46] or lighting from a prior distribution that was pre-trained on a dataset of environment maps [6, 47]. The most closely related work is that for which lighting information is not provided as input [17, 27].

For the sake of generality, we do not use such strong priors in our experiments because by avoiding them we can more directly measure the benefits that are gained by modeling unintended shadows as an additional cue. Because our model uses a generic gradient-descent framework, stronger application-specific priors can be added to it to improve performance on a specific domain of interest.

Estimating environment lighting. Another thread of inverse rendering research focuses on the task of recovering high-fidelity lighting environments, typically for augmented reality applications where virtual objects are rendered into photographs with consistent reflections and shadows. Prior work by Debevec [9] used images of a chrome sphere to measure environment lighting directly, and subsequent works have demonstrated that plausible environment lighting can be estimated without chrome spheres, using images of indoor [12, 35] and outdoor scenes [13, 19, 20], or images of a specific class of objects like faces [21, 30]. These subsequent techniques use strong priors in the form of deep neural network weights that are trained with supervision to map from images to lighting, and they do not enforce physical rendering consistency between the recovered lighting and the observations. In contrast, we avoid such strong priors, and we explicitly enforce consistency.

More related to our approach is the work of Park *et al.* [31], which recovers physically-consistent environment maps from RGBD videos of shiny objects. However, their algorithm can only recover the convolution of the environment map with the BRDF of the observed object (*e.g.*, Figure 1(b)), which precludes the recovery of high illumination frequencies unless the object is highly specular.

Also related is the work of Swedish *et al.* [37], which places a known, diffuse object on a ground plane (a scenario first studied by Sato *et al.* [34]) and uses its cast shadows to recover high-quality lighting. Their formulation is linear and it improves lighting estimates by leveraging self-shadowing among diffuse objects that have known geometry and albedo (*e.g.*, Figure 1(c)). We generalize this by replacing the linear formulation with differentiable Monte Carlo ray tracing, which allows exploiting the additional shadowing effects caused by moving external occluders and leads to substantially improved results (*e.g.*, Figure 1(d)). Our formulation also handles objects with more general spatially-varying BRDFs that are not known beforehand.

Passive non-line-of-sight imaging with occlusions. Passive non-line-of-sight techniques can also be seen as recovering environment illumination: They observe a reflective surface (which is typically diffuse and planar) and recover the appearance of a "hidden scene" from these reflections. Similar to us, prior works in this area have observed that the presence of an occluder between the observed surface and the hidden scene/environment aids recovery by introducing sharp angular variations into the rendering integral. This insight was first leveraged in settings with simple occluder-shapes like pinholes and pinspecks [39] or corners joining walls [7]. Subsequent work by Baradad *et al.* [2] generalizes this by considering light-occlusion effects caused by an arbitrary but known 3D shape (a task that is closely related to Swedish *et al.* above). Yedidia *et al.* [44] additionally recover the unobserved occluder's shape assuming it is a planar mask parallel to the observed planar surface.

We generalize these prior works by using differentiable Monte Carlo rendering to replace their deconvolution-based algorithms, which are tailored to the specific case where the observed reflector is planar and diffuse. This allows using reflective objects that have arbitrary shapes, and arbitrary spatially-varying BRDFs which are not known beforehand. It also allows recovering the shapes of unobserved occluders that are arbitrary and time-varying.

3. Motivation and Problem Setup

A well-known ambiguity in computer vision and graphics occurs when decomposing an image into its component lighting and materials. The top of Figure 2 recreates a common depiction of this ambiguity, where an image of a known shape (a sphere) is explained equally well by a diffuse material in a complicated lighting environment or a shiny, mirror-like material under low-frequency lighting. Now, imagine we capture a second image from the same viewpoint, after some external object enters the scene. This second object remains beyond the field of view and so is not directly observed, but it acts as an external occluder that prevents some of the environment's light from reaching the sphere. This second image, shown at the bottom of the figure for each case, clearly reveals which of the two material-lighting explanations is the correct one.

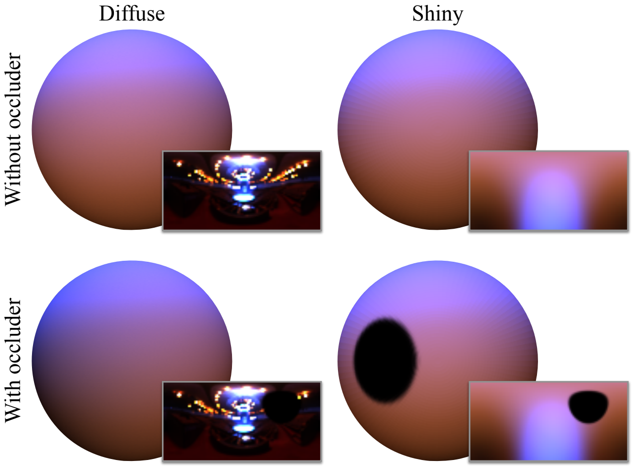


Figure 2. Decomposing appearance into lighting and materials can be inherently ambiguous. Top row: a diffuse sphere lit by high-frequency lighting (inset) is visually indistinguishable from a shiny, mirror-like sphere lit by low-frequency lighting. Bottom row: A second image captured in the presence of a dark external occluder resolves the ambiguity. The occluder's effect is clearly visible on the shiny sphere while on the diffuse sphere it produces shadows that are soft and very subtle.

We aim to make use of this unintended shadowing that naturally occurs whenever a camera and its operators move around an object while capturing images from different viewpoints. They affect each image by blocking different portions of the surrounding light, and this provides a helpful signal. We will show that this signal is helpful even when the occluder shapes and locations are quite arbitrary and not known beforehand, and when their shadowing effects are very subtle, like between the top and bottom images on the left of Figure 2.

In addition to helping resolve material-lighting ambiguities, shadowing from moving external occluders also improves the conditioning of our inverse rendering problem. This is analyzed in Figure 3, which considers the simplified hypothetical 1D case of recovering an environment illumination from images of a diffuse disk-shaped object, with known occluder positions. Convolution with a dif-

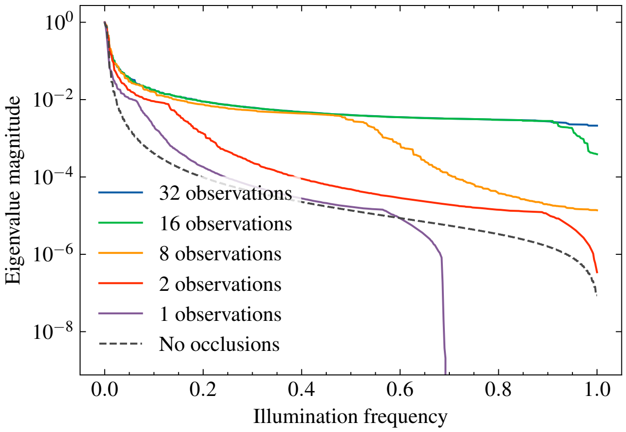


Figure 3. Here we plot eigenvalue magnitudes as a function of illumination angular frequency for a one-dimensional circular Lambertian object with known albedo, under the simplifying assumption of known occluder locations and shapes. For a single input image without occlusion, almost all lighting frequencies are strongly attenuated. In the single-occluder case (similar to the scenario considered by Baradad *et al.* [2]), some low frequencies become recoverable but intermediate and high frequencies do not. Observing additional images enables the recovery of more lighting information, especially at the highest frequencies.

fuse BRDF acts as a low-pass filter of the incoming light, which means that the linear inverse problem's eigenvalues are vanishingly low for higher frequencies. Moving external occluders introduce a sharp angular variation in the per-image illumination, which better-conditions these high frequency components, and allows their recovery. As shown in Figure 3, in this simplified scenario a single image containing an occluder is better than a single image without any occluders, but using observations of more occluders makes the problem even better-conditioned. We demonstrate this effect in practice in the supplement.

3.1. Formulation

Let $\{I_t\}_{t=1}^T$ be a collection of T images of a scene with known camera poses; for every t, I_t is an RGB image with spatial size $H \times W$. We assume the observed scene $\mathcal{S} \subset \mathbb{R}^3$ is illuminated by a far-field environment map $L(\hat{\boldsymbol{\omega}}_i)$ that does not change over time, so the only temporal changes in the incident light field are caused by unobserved light-blockers (called *occluders* hereafter) that move around the scene outside the field of view. We can then write the incident light at any observed position $\mathbf{x} \in \mathcal{S}$ as:

$$L_t(\mathbf{x}, \hat{\boldsymbol{\omega}}_i) = L(\hat{\boldsymbol{\omega}}_i) M_t(\mathbf{x}, \hat{\boldsymbol{\omega}}_i) V(\mathbf{x}, \hat{\boldsymbol{\omega}}_i)\,, \tag{1}$$

where for any t, $M_t(\mathbf{x}, \hat{\boldsymbol{\omega}}_i)$ is a binary signal with value 0 for directions blocked by the occluder and 1 otherwise, and the binary signal $V(\mathbf{x}, \hat{\boldsymbol{\omega}}_i)$ models visibility effects that are *internal* to the scene, with value 0 for directions from $\mathbf{x}$ that are blocked by other scene elements, and value 1 otherwise.

Omitting global illumination effects, a pixel $\mathbf{u}$ corresponding to a surface point $\mathbf{x}$ with BRDF f viewed from direction $\hat{\boldsymbol{\omega}}_o$ at time t has color:

$$I_t(\mathbf{u}) = \int_{\mathbb{S}^2} L_t(\mathbf{x}, \hat{\boldsymbol{\omega}}_i) f(\mathbf{x}, \hat{\boldsymbol{\omega}}_i, \hat{\boldsymbol{\omega}}_o)(\hat{\mathbf{n}}(\mathbf{x}) \cdot \hat{\boldsymbol{\omega}}_i)_+ \, d\hat{\boldsymbol{\omega}}_i\,, \tag{2}$$

where $\hat{\mathbf{n}}(\mathbf{x})$ is the surface normal corresponding to $\mathbf{x}$, and $(\cdot)_+$ clamps negative values to zero.

Given the set of observed images, we would like to recover the environment map L, a spatially-varying BRDF f at every point on the surface of the object $\mathcal{S}$, as well as the set of unobserved occluder masks $\{M_t\}_{t=1}^T$, one for each observation time t. Our goal is therefore to solve the following optimization problem:

$$\underset{\boldsymbol{\phi}^{(o)}, \boldsymbol{\phi}^{(m)}, \boldsymbol{\phi}^{(\ell)}}{\arg\min} \sum_{t,\mathbf{u}} \left\| I_t(\mathbf{u}) - \mathcal{R}_t\Big(\mathbf{u}; \boldsymbol{\phi}^{(o)}, \boldsymbol{\phi}^{(m)}, \boldsymbol{\phi}^{(\ell)}\Big) \right\|^2, \tag{3}$$

where $\mathcal{R}_t(\mathbf{u}; \cdot)$ renders the pixel location $\mathbf{u}$ at time t using the occluder parameters $\boldsymbol{\phi}^{(o)}$, material parameters $\boldsymbol{\phi}^{(m)}$, and illumination parameters $\boldsymbol{\phi}^{(\ell)}$.

At first glance, it may seem that solving for occluders in addition to materials and illumination should make the resulting inverse problem more difficult. However, as illustrated in Figure 3 and demonstrated by our experimental results, modeling and solving for occluders actually makes the full problem *more* well-conditioned and thus improves recovery of material and illumination.

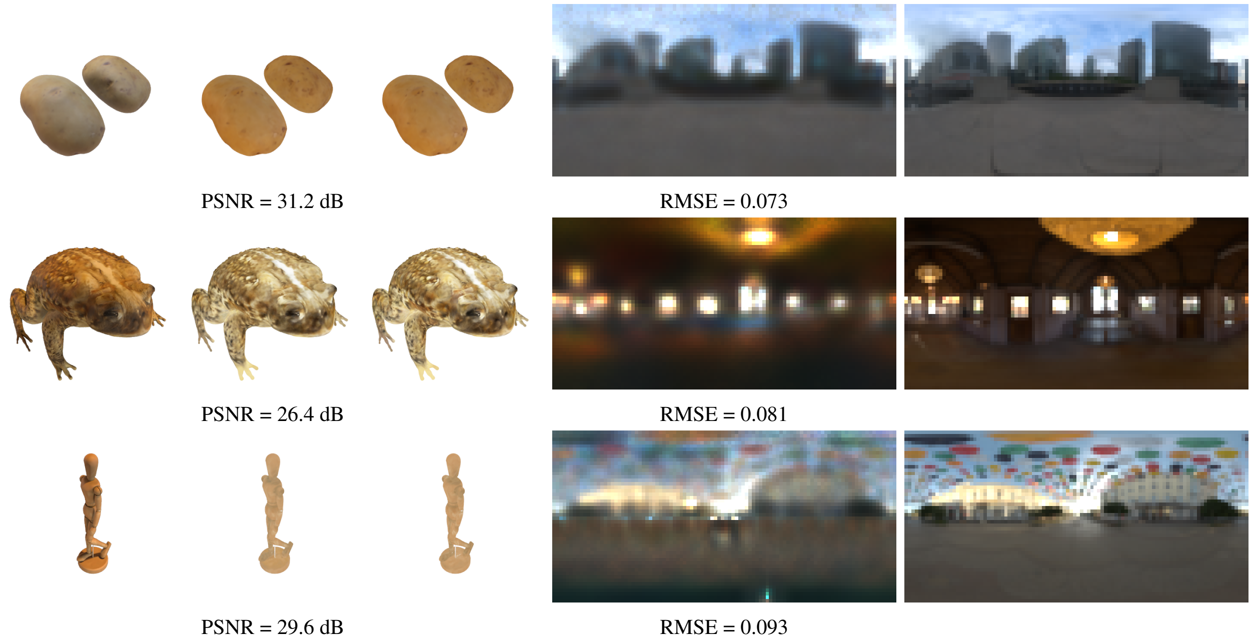
4. Method

We first describe our parameterizations of the unseen occluders, illumination, and materials (Sections 4.1–4.3). Section 4.4 then relates these parameters to rendered pixel colors, and Section 4.5 describes how we optimize to solve the inverse problem in Equation 3.

We emphasize that our method is designed to exploit external shadows across as many scene-types as possible, and so its only priors are those implicit to our parameterizations. In our experiments, we verify that these relatively weak priors are sufficient for recovering high-fidelity occluder shapes, illumination maps, and material maps.

4.1. Occluders

Define a world coordinate system with origin at the scene's center, and let $\{r_t\}_{t=1}^T$ be the radial distances from the origin to each (known) camera's center. Then we model each of the T occluders M_t as an independent binary signal defined on the surface of a finite, radius-r_t sphere that is centered at the origin. The value of the binary occluder signal



(a) Sample input image (b) Recovered albedo (c) True albedo (d) Recovered illumination (e) True illumination

Figure 4. The results of our method on three additional diffuse objects. We report the RMSE of each environment map in linear color space but plot the images after tonemapping for better evaluation of the full dynamic range. The albedo PSNR values are reported on object pixels only.

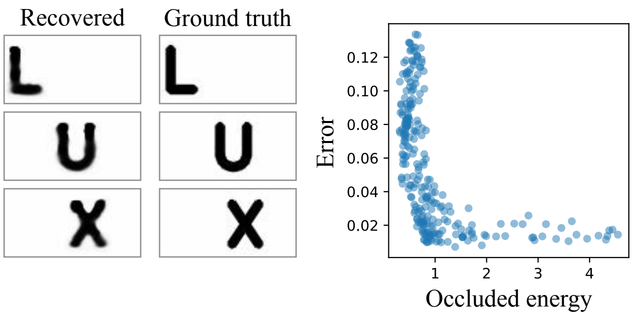


Figure 5. Our occluders extracted from the *potatoes* scene on the top of Figure 4 resemble the true ones (left). However, as expected, inaccuracies in the recovered masks tend to occur at relatively-dark regions of the environment map: there is a very strong inverse correlation (right) between the mean error in the recovered masks and the amount of illumination energy that they block.

for a shadow ray emitted from surface point $\mathbf{x} \in \mathcal{S}$ in direction $\hat{\boldsymbol{\omega}}_i$ at time t can be computed by intersecting the shadow ray with the occluder's spherical shell:

$$M_t(\mathbf{x}, \hat{\boldsymbol{\omega}}_i) = \tilde{M}_t\left(S(\mathbf{x}, \hat{\boldsymbol{\omega}}_i, r_t)\right) \tag{4}$$

where $S(\mathbf{x}, \hat{\boldsymbol{\omega}}_i, r)$ is the intersection of the ray with a sphere of radius r_t, normalized to have unit length:

$$S(\mathbf{x}, \hat{\boldsymbol{\omega}}_i, r) = \frac{\mathbf{x} + \left(\sqrt{(\mathbf{x} \cdot \hat{\boldsymbol{\omega}}_i)^2 + r^2 - \|\mathbf{x}\|^2} - \mathbf{x} \cdot \hat{\boldsymbol{\omega}}_i\right) \hat{\boldsymbol{\omega}}_i}{r}. \tag{5}$$

Note that there is a single intersection, since we assume that the surface points are always inside the occluder spheres.

Although our formulation models the occluders as binary signals, a binary representation is not well-suited for gradient-based optimization. Instead, we represent the occluders as a continuously-valued function on the sphere in a spherical harmonic basis, mapped using a sigmoid function σ to lie in $[0, 1]$:

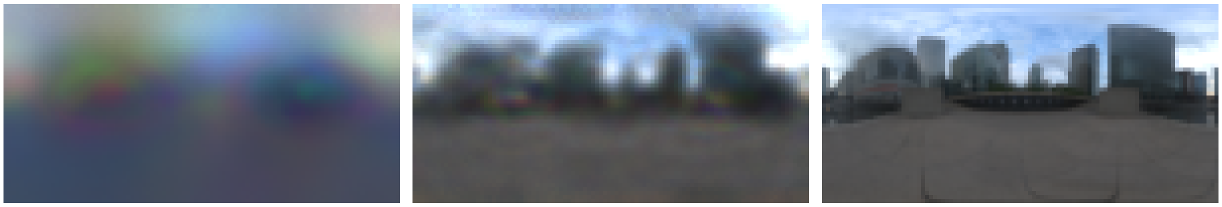
$$\tilde{M}_t(\hat{\boldsymbol{\omega}}) = \sigma\left(\sum_{\ell=0}^{P} \sum_{m=-\ell}^{\ell} a_{t\ell m} Y_\ell^m(\hat{\boldsymbol{\omega}})\right), \tag{6}$$

where $\boldsymbol{\phi}^{(o)} \triangleq \{a_{t\ell m}\}$ are optimizable coefficients that parameterize the occluder at time t, and P is the degree required to span the space of spherical images with the same resolution as our environment illumination (see below) [11].

Note that for radially symmetric BRDFs, the rendering integral in Equation 2 can be written as a spherical convolution [11, 33]. This makes spherical harmonics a natural choice for representing the occluder signals, since spherical convolutions are diagonal in the basis of spherical harmonics. See our supplement for more details.

4.2. Environment Illumination

We assume that the scene's illumination is distant and can be represented as an environment map. The environment map is parameterized as an image pyramid of size $H' \times W'$



(a) Recovered illumination (no occluder) (b) Recovered illumination (c) True illumination

Figure 6. In the case of purely Lambertian material (with the *potatoes* geometry), (a) only using self occlusions is insufficient for the recovery of high-frequency content in the illumination. (b) Modeling and estimating occluders recovers significantly higher-quality illumination that closely resembles the ground truth.

in equirectangular coordinates (50 × 100 in our experiments), with an exponential nonlinearity:

$$L(\hat{\boldsymbol{\omega}}_i) = \exp\left(\sum_{k=0}^{K-1} a^k L_k(\hat{\boldsymbol{\omega}}_i)\right) , \quad (7)$$

where $\boldsymbol{\phi}^{(\ell)} \triangleq \{L_k\}_{k=0}^{K-1}$ comprise a coarse $(k = 0)$ to fine $(k = K - 1)$ representation of the illumination, and we set $a = 2$. For each k, $L_k(\hat{\boldsymbol{\omega}}_i)$ is computed by bilinearly interpolating into a grid whose size exponentially increases with k. The exponential nonlinearity is helpful for obtaining high dynamic range values in the environment map, as noted, *e.g*. in [3].

Unlike the occluder masks, we find that representing the environment map as a pyramid results in better performance than what is achieved with spherical harmonics. See appendix for additional details.

4.3. Materials

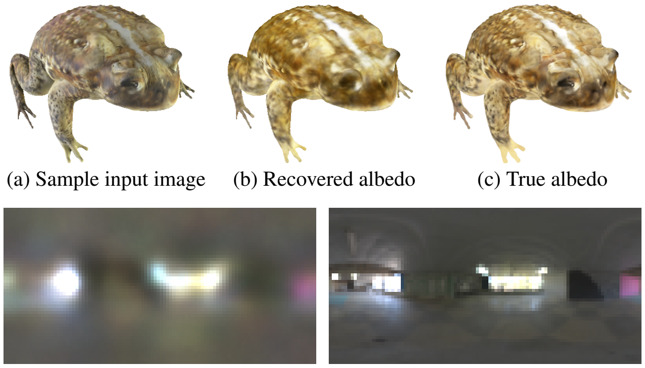
We represent the scene's spatially-varying BRDF as the sum of a diffuse and a specular component. We use a Lambertian model for the diffuse term and a GGX microfacet model [43] for the specular term. The spatially-varying BRDF parameters are represented by a coordinate-based multi-layer perceptron (MLP) with a positional encoding [38] function γ and optimizable parameters $\boldsymbol{\phi}^{(m)}$:

$$\boldsymbol{\alpha}(\mathbf{x}) = \text{MLP}\Big(\gamma(\mathbf{x}); \boldsymbol{\phi}^{(m)}\Big) . \quad (8)$$

Here, $\gamma(\mathbf{x})$ is the positionally-encoded position on the object's surface, and $\boldsymbol{\alpha}(\mathbf{x}) \in \mathbb{R}^5$ are BRDF coefficients, consisting of an RGB diffuse albedo, a spatially-varying scalar microfacet roughness, and a spatially-varying scalar specular reflectance at normal incidence (equivalent to a reparameterization of the material's index of refraction). See the supplement for an exact specification of our BRDF model.

4.4. Rendering

We render the occluders, materials, and illumination, *i.e*. the rendering operator $\mathcal{R}_t$ from Problem 3, by approximating the integral in Equation 2 using Monte Carlo techniques. We use a standard multiple importance sampler [40] based on the illumination and material model. We model shadows cast by the occluders and self-occlusions by the object itself, but for efficiency and due to our focus on diffuse objects, we neglect lower-order effects and global illumination. See the supplemental material for a full description of our renderer.



(a) Sample input image (b) Recovered albedo (c) True albedo

(d) Recovered illumination (e) True illumination

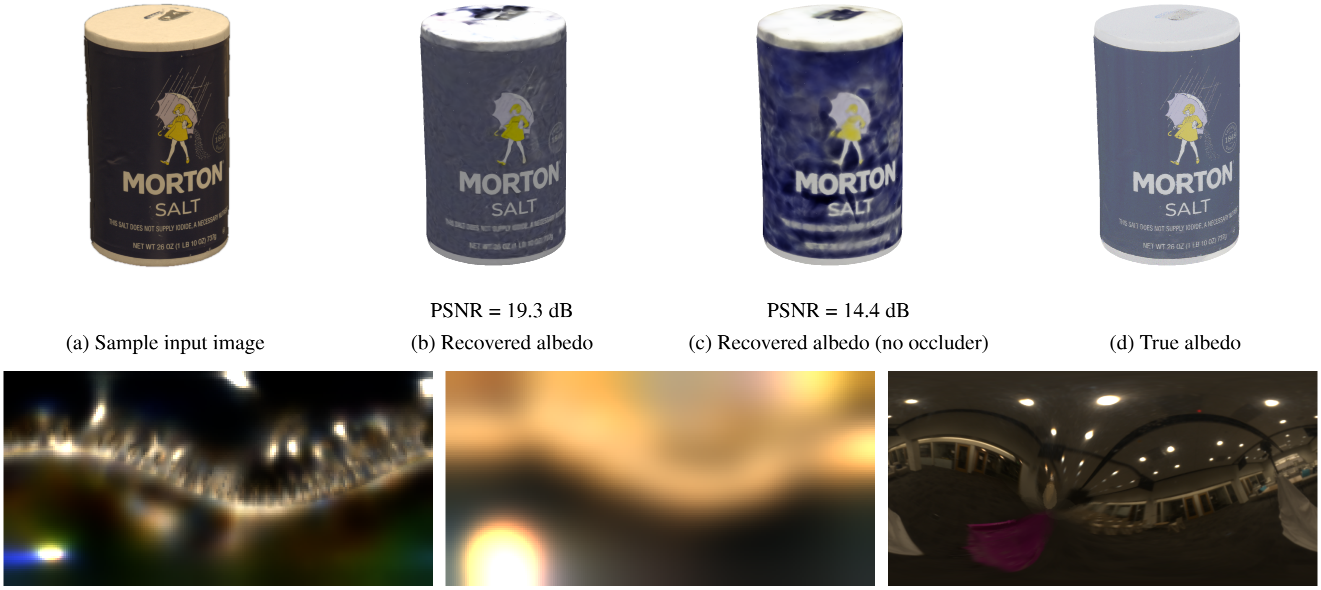
Figure 7. We are able to recover illumination even when geometry is unknown by first optimizing a volumetric representation of geometry using a NeRF-based method. Despite this inaccurate proxy geometry, our method still recovers plausible (albeit blurry) illumination (d) and albedo (b).

4.5. Optimization

We optimize the objective in Equation 3 using the L_2 error between rendered values and ground truth ones. We use Adam [16] to optimize all three components of our model, with a learning rate of $3 \cdot 10^{-3}$ for the environment map and materials, and a learning rate of 1 for the occluders.

In each iteration, we compute the L_2 loss using a batch size of 2^{16} pixels. In order to avoid bias in the gradient updates to our model's parameters, we use the same approach as [10]: we render every pixel value twice using independent samples and multiply the per-channel deviations of both from the ground truth to get the expected loss value:

$$\mathcal{L} = \sum_{\mathbf{u}} \left(\tilde{I}_t^{(1)}(\mathbf{u}) - I_t(\mathbf{u})\right) \cdot \left(\tilde{I}_t^{(2)}(\mathbf{u}) - I_t(\mathbf{u})\right) , \quad (9)$$



PSNR = 19.3 dB

PSNR = 14.4 dB

(a) Sample input image (b) Recovered albedo (c) Recovered albedo (no occluder) (d) True albedo

RMSE = 0.035

RMSE = 0.087

(e) Recovered illumination (f) Recovered illumination (no occluder) (g) **Warped** true illumination

Figure 8. Results on the captured `salt004` scene from Stanford-ORB [18]. Note that the "warped true illumination" environment map (g) was captured by a light probe that was not co-located with the object, any may therefore be significantly warped. The recovered RMSE reported is computed with respect to the average of all environment maps provided with the dataset rotated the same coordinate frame.

where $\tilde{I}_t^{(1)}(\mathbf{u})$ and $\tilde{I}_t^{(2)}(\mathbf{u})$ are the two independently-rendered pixel values.

5. Experiments

We implement our entire rendering and optimization pipeline in JAX [8] and run each of our experiments on 8 NVIDIA V100 GPUs. We evaluate our method on two sources of data: synthetic objects, and captured data from the Stanford-ORB dataset [18]. Our synthetic results were generated using 256 input images of size 400×400, although our method works similarly well with as few as 32 captures (see supplement for the effect of the number of images). Unless noted otherwise, the occluders for all experiments are spherical caps with a size of $0.1 \cdot 4\pi$ steradians, centered at the camera.

Estimating illumination, materials, and occluders. Our method can effectively recover environment illumination, spatially-varying material parameters, and the shapes of unseen lighting occluders. We first investigate our algorithm's performance using rendered images of a variety of diffuse objects (roughness 0.6, see BRDF model in appendix) with known geometry. The illumination and albedos recovered by our model are visualized with their true values in Figures 1 and 4. Figure 6 shows our reconstruction for purely Lambertian materials, compared with the reconstruction from the same number of images *without* unobserved occluders (but with self-occlusions), similar to the scenario investigated by Swedish *et al*. [37]. For Lambertian objects, reconstruction is severely ill-conditioned even when self-occlusions are present, which results in significantly blurrier reconstructions. Figure 5 validates that our recovered occluder shapes are accurate, especially for occluders that block more of the illumination energy.

Captured data. We apply our method to scenes from the Stanford-ORB dataset [18]. Each object in the dataset is placed on a platform and captured from multiple directions. As the camera rig and photographer move around the scene, the photographer is hidden under white cloth below the camera. The images are provided along with a scanned mesh which we use for geometry.

Despite not being designed for our task, Figure 8 shows our recovered environment map obtained using the 66 training views in the `salt004` scene, which features a rough cylindrical object. Note that the dataset only provides an estimated illumination obtained by a light probe placed at a large distance from the object, meaning there is a significant unknown non-rigid warp between the true incident light and the image in Figure 8g labeled "warped true illumination". See supplement for additional results.

Without known geometry. Figure 7 presents preliminary evidence that our method can be used to recover lighting and materials even when the object geometry is not known. In this experiment, we first recover an estimate of object geometry using a Neural Radiance Field (NeRF) [25]-based

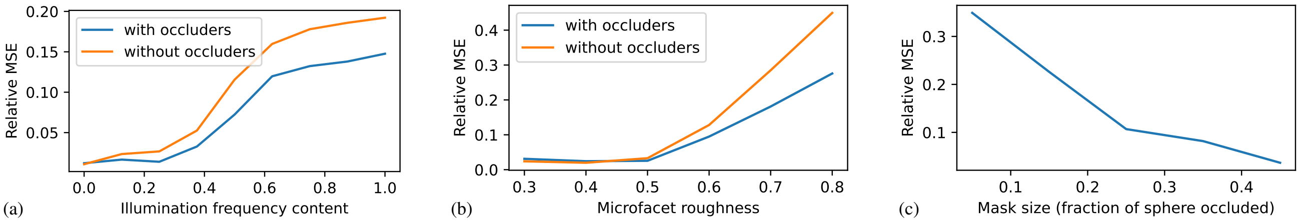
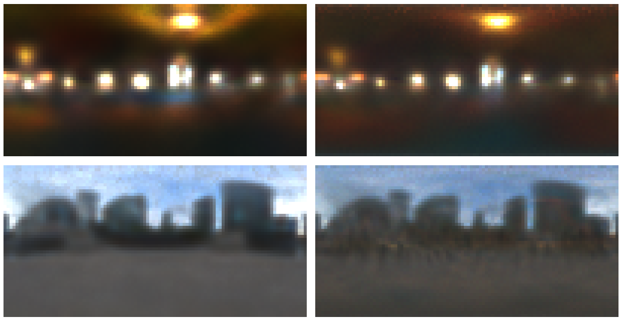


Figure 9. The signal provided by unintended shadows cast by unseen occluders improves the quality of recovered environment illumination. Here, we plot the relative MSE for the recovered environment maps under two scenarios: (blue) using images rendered with unobserved occluders and jointly estimating materials, illumination and occluder shape and (orange) using images rendered without occluders and only estimating materials and illumination (this is the problem setting considered by Swedish *et al*. [37]). The cue of unintended shadows consistently improves the quality of estimated illumination across varying (a) illumination frequency content, (b) object material roughness, and (c) mask sizes. Additionally, we show that larger occluders that block more light improve the reconstructed illumination quality.



(a) Illumination (pyramid) (b) Illumination (direct)

Figure 10. Results from the experiment in Figure 4, where the pyramid environment map has been replaced with a direct optimization of its (pre-exponentiated) values. Directly optimizing the mask values instead of using spherical harmonics also results in worse reconstructions (see Table 1).

model augmented using the orientation loss from Verbin *et al*. [41], and then use this (potentially-imprecise) proxy geometry in our inverse rendering optimization pipeline. Please refer to the supplemental materials for a detailed description of our NeRF-based model.

Object self-occlusion does not contain enough signal. Figure 9 quantitatively demonstrates that only using the signal provided by object self-occlusions ("without occluders" plot in orange) cannot recover illumination as accurately as our method, which leverages the cue of unintended shadows ("with occluders" plot in blue). In particular, exploiting unintended shadows is increasingly important as the illumination contains higher frequencies (Figure 9a) and as the object material becomes increasingly diffuse (Figure 9b).

Larger occluders improve illumination recovery. Figure 9c shows that increasing the size of the unseen occluders improves the accuracy of the recovered environment maps. Increasing the occluder size effectively improves the problem's conditioning by creating a larger variation in the incident lighting across points on the object.

	chair	*mannequin*	*potatoes*	*toad*
Direct spherical harmonics	0.831	0.654	0.997	0.943
Direct environment map	0.042	0.071	0.015	0.073
Ours	**0.038**	**0.035**	**0.011**	**0.046**

Table 1. Errors (RMSE) on all four of our scenes for our full model compared with our model without the spherical harmonic occluder parameterization, and without the pyramid environment parameterization. While the pyramid representation of the environment map improves smoothness (see Figure 10), we find that representing the masks using spherical harmonics is critical for our method.

Parameterization ablation study. Table 1 and Figure 10 show that representing the environment map as an image pyramid and the image masks with spherical harmonics, both perform better than simply optimizing their respective element values directly. See supplement for more details.

6. Discussion

Even though unintended shadows are common in most real-world capture scenarios, they are often treated as outlier data. In this paper, we showed how to explicitly leverage these shadows as a signal to improve the quality of recovered lighting and materials in particularly challenging scenarios, such as objects made of diffuse materials. Furthermore, we have shown that our algorithm can be used with approximate geometry reconstructed by view synthesis techniques, and that it can produce imperfect, yet promising results on real captures, despite not being designed for this type of data which does not satisfy many of our model's assumptions. We believe that this work makes important first steps towards a general inverse rendering algorithm that can recover geometry, materials, and illuminations from images captured under realistic conditions.

Acknowledgements. Our sincere thanks to Delio Vicini, Rick Szeliski, Aleksander Hołyński, and Guy Satat for helpful discussions and suggestions, and to David Salesin, Janne Kontkanen, and Lior Yariv for their help in improving the manuscript before submission.

References

[1] Sai Bangaru, Michael Gharbi, Tzu-Mao Li, Fujun Luan, Kalyan Sunkavalli, Milos Hasan, Sai Bi, Zexiang Xu, Gilbert Bernstein, and Fredo Durand. Differentiable rendering of neural SDFs through reparameterization. *SIGGRAPH Asia*, 2022. 2

[2] Manel Baradad, Vickie Ye, Adam B. Yedidia, Frédo Durand, William T. Freeman, Gregory W. Wornell, and Antonio Torralba. Inferring light fields from shadows. *CVPR*, 2018. 3, 4

[3] Jonathan T. Barron and Jitendra Malik. Shape, illumination, and reflectance from shading. *TPAMI*, 2015. 2, 6

[4] Sai Bi, Zexiang Xu, Pratul P. Srinivasan, Ben Mildenhall, Kalyan Sunkavalli, Milos Hasan, Yannick Hold-Geoffroy, David Kriegman, and Ravi Ramamoorthi. Neural reflectance fields for appearance acquisition. *arXiv*, 2020. 2

[5] Sai Bi, Zexiang Xu, Kalyan Sunkavalli, Milovs Havsan, Yannick Hold-Geoffroy, David J. Kriegman, and Ravi Ramamoorthi. Deep Reflectance Volumes: Relightable Reconstructions from Multi-View Photometric Images. *ECCV*, 2020. 2

[6] Mark Boss, Raphael Braun, Varun Jampani, Jonathan T. Barron, Ce Liu, and Hendrik P. A. Lensch. NeRD: Neural reflectance decomposition from image collections. *ICCV*, 2021. 2

[7] Katherine L. Bouman, Vickie Ye, Adam B. Yedidia, Frédo Durand, Gregory W. Wornell, Antonio Torralba, and William T. Freeman. Turning corners into cameras: Principles and methods. *ICCV*, 2017. 3

[8] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. `http://github.com/google/jax`. 7

[9] Paul Debevec. Rendering synthetic objects into real scenes: Bridging traditional and image-based graphics with global illumination and high dynamic range photography. *SIGGRAPH*, 1998. 2

[10] Xi Deng, Fujun Luan, Bruce Walter, Kavita Bala, and Steve Marschner. Reconstructing Translucent Objects using Differentiable Rendering. *SIGGRAPH*, 2022. 6

[11] James R Driscoll and Dennis M Healy. Computing Fourier Transforms and Convolutions on the 2-Sphere. *Advances in applied mathematics*, 1994. 5

[12] Marc-André Gardner, Kalyan Sunkavalli, Ersin Yumer, Xiaohui Shen, Emiliano Gambaretto, Christian Gagné, and Jean-François Lalonde. Learning to predict indoor illumination from a single image. *ACM Trans. Graph.*, 2017. 2

[13] Yannick Hold-Geoffroy, Kalyan Sunkavalli, Sunil Hadap, Emiliano Gambaretto, and Jean-François Lalonde. Deep outdoor illumination estimation. *CVPR*, 2017. 2

[14] Wenzel Jakob, Sébastien Speierer, Nicolas Roussel, and Delio Vicini. Dr.Jit: A Just-In-Time Compiler for Differentiable Rendering. *SIGGRAPH*, 2022. 2

[15] Brian Karis and Epic Games. Real shading in unreal engine 4. *Proc. Physically Based Shading Theory Practice*, 2013. 1

[16] P. Diederik Kingma and Lei Jimmy Ba. Adam: A Method for Stochastic Optimization. *ICLR*, 2015. 6, 7

[17] Zhengfei Kuang, Kyle Olszewski, Menglei Chai, Zeng Huang, Panos Achlioptas, and Sergey Tulyakov. Neroic: Neural rendering of objects from online image collections. *ACM Trans. Graph.*, 2022. 2

[18] Zhengfei Kuang, Yunzhi Zhang, Hong-Xing Yu, Samir Agarwala, Shangzhe Wu, and Jiajun Wu. Stanford-ORB: A real-world 3d object inverse rendering benchmark. *NeurIPS*, 2023. 7, 1, 2, 6

[19] Jean-François Lalonde, Alexei A Efros, and Srinivasa G Narasimhan. Estimating the natural illumination conditions from a single outdoor image. *IJCV*, 2012. 2

[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. *CVPR*, 2019. 2

[21] Chloe LeGendre, Wan-Chun Ma, Rohit Pandey, Sean Fanello, Christoph Rhemann, Jason Dourgarian, Jay Busch, and Paul Debevec. Learning illumination from diverse portraits. *SIGGRAPH Asia 2020 Technical Communications*, 2020. 2

[22] Tzu-Mao Li, Miika Aittala, Frédo Durand, and Jaakko Lehtinen. Differentiable monte carlo ray tracing through edge sampling. *SIGGRAPH Asia*, 2018. 2

[23] Zhengqin Li, Zexiang Xu, Ravi Ramamoorthi, Kalyan Sunkavalli, and Manmohan Chandraker. Learning to reconstruct shape and spatially-varying reflectance from a single image. *SIGGRAPH Asia*, 2018. 2

[24] 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. *CVPR*, 2020. 2

[25] 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. *ECCV*, 2020. 7, 8

[26] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *SIGGRAPH*, 2022. 8

[27] Jacob Munkberg, Jon Hasselgren, Tianchang Shen, Jun Gao, Wenzheng Chen, Alex Evans, Thomas Mueller, and Sanja Fidler. Extracting Triangular 3D Models, Materials, and Lighting From Images. *CVPR*, 2022. 2

[28] Merlin Nimier-David, Delio Vicini, Tizian Zeltner, and Wenzel Jakob. Mitsuba 2: A retargetable forward and inverse renderer. *SIGGRAPH Asia*, 2019. 2

[29] Merlin Nimier-David, Sébastien Speierer, Benoît Ruiz, and Wenzel Jakob. Radiative Backpropagation: An Adjoint Method for Lightning-Fast Differentiable Rendering. *SIGGRAPH*, 2020. 2

[30] Ko Nishino and Shree K Nayar. Eyes for relighting. *ACM Trans. Graph.*, 2004. 2

[31] Jeong Joon Park, Aleksander Holynski, and Steven M Seitz. Seeing the world in a bag of chips. *CVPR*, 2020. 1, 3

[32] Matt Pharr, Wenzel Jakob, and Greg Humphreys. *Physically Based Rendering: From Theory to Implementation (3rd*

ed.). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edition, 2016. 1

[33] Ravi Ramamoorthi and Pat Hanrahan. A signal-processing framework for inverse rendering. *SIGGRAPH*, 2001. 2, 5

[34] Imari Sato, Yoichi Sato, and Katsushi Ikeuchi. Illumination from shadows. *TPAMI*, 2003. 3

[35] Shuran Song and Thomas Funkhouser. Neural illumination: Lighting prediction for indoor environments. *CVPR*, 2019. 2

[36] Pratul P. Srinivasan, Boyang Deng, Xiuming Zhang, Matthew Tancik, Ben Mildenhall, and Jonathan T. Barron. NeRV: Neural reflectance and visibility fields for relighting and view synthesis. *CVPR*, 2021. 2

[37] Tristan Swedish, Connor Henley, and Ramesh Raskar. Objects as cameras: Estimating high-frequency illumination from shadows. *ICCV*, 2021. 1, 3, 7, 8

[38] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *NeurIPS*, 2020. 6

[39] Antonio Torralba and William T. Freeman. Accidental pinhole and pinspeck cameras: Revealing the scene outside the picture. *CVPR*, 2012. 3

[40] Eric Veach and Leonidas J Guibas. Optimally Combining Sampling Techniques for Monte Carlo Rendering. *SIGGRAPH*, 1995. 6, 4

[41] Dor Verbin, Peter Hedman, Ben Mildenhall, Todd Zickler, Jonathan T. Barron, and Pratul P. Srinivasan. Ref-NeRF: Structured View-Dependent Appearance for Neural Radiance Fields. *CVPR*, 2022. 8

[42] Delio Vicini, Sébastien Speierer, and Wenzel Jakob. Differentiable Signed Distance Function Rendering. *SIGGRAPH*, 2022. 2

[43] Bruce Walter, Stephen R Marschner, Hongsong Li, and Kenneth E Torrance. Microfacet models for refraction through rough surfaces. *Eurographics*, 2007. 6, 1

[44] Adam B Yedidia, Manel Baradad, Christos Thrampoulidis, William T Freeman, and Gregory W Wornell. Using unknown occluders to recover hidden scenes. *CVPR*, 2019. 3

[45] Tizian Zeltner, Sébastien Speierer, Iliyan Georgiev, and Wenzel Jakob. Monte Carlo Estimators for Differential Light Transport. *SIGGRAPH*, 2021. 2, 4

[46] Kai Zhang, Fujun Luan, Qianqian Wang, Kavita Bala, and Noah Snavely. PhySG: Inverse rendering with spherical gaussians for physics-based material editing and relighting. *CVPR*, 2021. 2

[47] Xiuming Zhang, Pratul P. Srinivasan, Boyang Deng, Paul Debevec, William T. Freeman, and Jonathan T. Barron. NeRFactor: Neural factorization of shape and reflectance under an unknown illumination. *SIGGRAPH Asia*, 2021. 2