Diffusion Reflectance Map:
Single-Image Stochastic Inverse Rendering of Illumination and Reflectance

Yuto Enyo   Ko Nishino

Graduate School of Informatics, Kyoto University
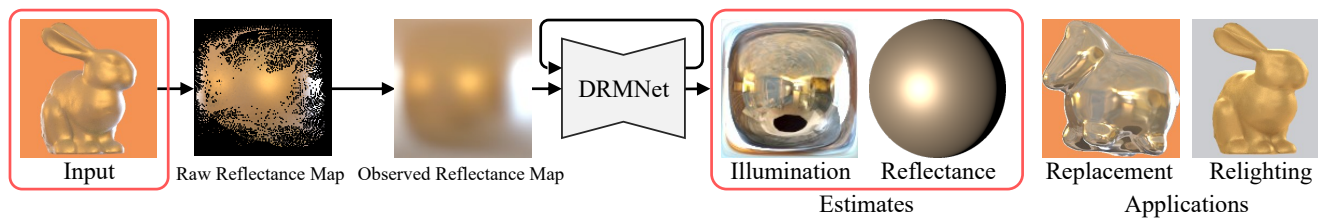
https://vision.ist.i.kyoto-u.ac.jp/

Figure 1. We introduce the first single-image stochastic inverse rendering method, a principled approach for recovering the attenuated frequency spectrum of the illumination and reflectance by seamlessly integrating a neural generative process in inverse rendering. Our key idea is to recover the illumination as a reflectance map of a perfect mirror. The results enable arbitrary object insertion and relighting.

Abstract

Reflectance bounds the frequency spectrum of illumination in the object appearance. In this paper, we introduce the first stochastic inverse rendering method, which recovers the attenuated frequency spectrum of an illumination jointly with the reflectance of an object of known geometry from a single image. Our key idea is to solve this blind inverse problem in the reflectance map, an appearance representation invariant to the underlying geometry, by learning to reverse the image formation with a novel diffusion model which we refer to as the Diffusion Reflectance Map Network (DRMNet). Given an observed reflectance map converted and completed from the single input image, DRMNet generates a reflectance map corresponding to a perfect mirror sphere while jointly estimating the reflectance. The forward process can be understood as gradually filtering a natural illumination with lower and lower frequency reflectance and additive Gaussian noise. DRMNet learns to invert this process with two subnetworks, IllNet and RefNet, which work in concert towards this joint estimation. The network is trained on an extensive synthetic dataset and is demonstrated to generalize to real images, showing state-of-the-art accuracy on established datasets.

1. Introduction

Light interacting with surfaces narrate our rich visual world. The light from illuminants reflect off object surfaces deeply entangling the surface material and geometry in its process. Deciphering the surface geometry, reflectance properties, and the incident illumination from this reflected radiance has been the central focus of physics-based vision research. Our focus is achieving this from a single image.

The ability to explicitly recover rich situational information from a single glance at an object goes to the very heart of embodied perception and enables effective use of visual information both in human perception and robot vision. For instance, decisions on how to act on an object needs to be made as the visual information is captured, and cannot be stalled until a hundred images are captured.

Even when the shape is known, *i.e.*, with an RGB-D smartphone sensor, disentangling the reflectance and illumination from a single image remains ill-posed due to two key problems. The first is color ambiguity—it is unclear whether to associate, for instance, the green object appearance to the illumination or the reflectance which cannot be resolved without assumptions on the color of either. The bigger and fatal obstacle is the loss of information. The reflectance acts as an angular filter on the incident illumination while the geometry determines its orientation [56]. As a result, the frequency spectrum of the object appearance is bounded by the highest frequency of either the reflectance or the illumination, the former of which is usually lower [8].

The frequency bound limits the utility of the estimated reflectance and illumination of past methods which do not pay attention to it. Due to low-pass reflectance, the illumination estimate is blurry and can only be used to insert an object in the scene that has lower frequency characteristics

than the object used to recover the illumination. The same can be said about relighting using the illumination estimate.

How should we recover the true illumination and reflectance from object appearance? The frequency bound of the object appearance suggests that there is no such thing as a true estimate. Given the object appearance, the illumination must be generated together with the reflectance as only its combined, partial information is encoded in it. Inverse rendering, particularly from a single image, as such, is a generative process. We argue for *stochastic inverse rendering* in which image formation is explicitly modeled as a stochastic process and its inversion becomes a stochastic generative reverse process. Stochasticity needs to be seamlessly integrated into the radiometric disentanglement.

In this paper, we derive a principled stochastic inverse rendering method for a single image. We assume homogeneous reflectance, which is the most challenging in terms of lacking visual information and serves as a foundation for multiple materials and objects. Our key idea is to learn to generate the illumination from the object appearance with a diffusion model on the reflectance map. By formulating the task on a geometry-invariant reflectance map, it enables radiometric disentanglement in the same domain eliminating the need for complex differentiable rendering.

Our problem is inherently blind, *i.e.*, the reflectance acting as the forward operator is also unknown. We introduce a novel diffusion model, Diffusion Reflectance Map Network (DRMNet), that generates a reflectance map corresponding to a perfect mirror sphere from the observed reflectance map while jointly estimating the reflectance. DRMNet consists of two subnetworks to learn to invert the forward radiometric image formation process with Gaussian noise.

The first, IllNet, is a U-Net that stochastically converts the current reflectance map into that of a sharper reflectance until it becomes perfectly mirror. The second, RefNet, is a simple network that takes in the observed and current reflectance maps and the current time-step and estimates the object reflectance as a parametric model. RefNet, in concert with IllNet, learns to iteratively refine the reflectance estimate. We introduce another diffusion model to transform the single input image into a reflectance map for DRMNet input by completing missing regions and resolving the many-to-one mapping from pixels to surface normals.

We train these diffusion models with a large synthetic dataset of reflectance maps of known reflectance and complex natural illumination and evaluate the accuracy of our method on a number of real and synthetic data. Through extensive comparisons with relevant works, we show that our method achieves state-of-the-art accuracy on this challenging task. We believe this explicitly generative approach to single-image inverse rendering would benefit a wide range of downstream applications not just for image synthesis but also for situational awareness and embodied interactions

with the objects and surroundings.

2. Related works

Inverse Rendering The disentanglement of object appearance into its radiometric constituents, namely the geometry, reflectance, and illumination, may inform about the object and the surroundings to a wide range of applications in computer vision, graphics, and robotics. This task is often referred to as inverse rendering [50], radiometric decomposition [47], appearance modeling [29], and inverse optics [3].

Many recent methods model this entanglement solely for novel view image synthesis, a representative approach of which is Neural Radiance Fields (NeRF) [54]. The results are impressive, demonstrating high-fidelity novel view synthesis of intricate objects. These methods, however, inherently rely on many multi-view ray samples of the scene [7, 18, 28, 45, 54, 61, 67, 76] or many images under different scene conditions captured from the same viewpoint [40, 41, 44, 73], typically on the order of tens of images. In sharp contrast, our focus is on recovering as much as possible about the object and scene from a single observation by leveraging the key physical characteristics of the interaction of light with the object surface.

Single-image inverse rendering is severely ill-posed even when the object shape is known. Lombardi *et al*. [46] introduce an MAP estimation framework with analytical reflectance and illumination priors. Chen *et al*. [14] model reflectance and illumination using deep neural networks and estimate them through optimization with differentiable rendering. Meka *et al*. [53] estimate the reflectance parameters with an image-space U-Net to convert the object appearance into a mirror and recovers illumination as a linear combination of low-order spherical harmonics estimated from diffuse shading and mirror reflection. This results in patched-up frequency spectrum as the mirror reflection only gives sparse high frequency details and the spherical harmonics only low frequency components. Our method achieves joint estimation of the reflectance and illumination in a space invariant to the surface geometry, *i.e.*, reflectance map. A few past works estimate a reflectance map [57] or an appearance map [52]. These methods, unlike ours, do not disentangle the illumination and reflectance.

Georgoulis *et al*. [25] regress the reflectance parameters and illumination as an environment map from a reflectance map. This method, however, does not account for the frequency attenuation in the observation and only recovers an arbitrarily bounded illumination. A few recent methods learn to estimate shape, reflectance, and illumination from a single image just for one class of objects (*e.g*., car) [27, 74] or by using supervision of synthetic data [9, 72]. Self-supervised methods use differentiable rendering to decompose single images into shape, reflectance, and illumi-

nation [12, 13, 34, 77, 79, 80]. Most of these methods are limited to simple or low-frequency illumination. Even with a fully-supervised model, directly estimating a single high-frequency solution from observations that only contain low-frequency components will lead to an averaged low-frequency estimate, *i.e.*, it cannot return a full spectrum estimate. Our focus is to break free of this fundamental limitation of inverse rendering by formulating it as a stochastic generative process.

Illumination Estimation Past methods have also tackled estimation of illumination alone [26, 55, 70, 71]. Yu *et al.* [78] estimated the lighting from shiny objects whose shape, reflectance, and texture are known. Our method does not make assumptions on the object reflectance and jointly estimates it. Swedish *et al.* [65] achieves high-frequency illumination estimation by leveraging cast shadows. Our method does not rely on high-frequency illumination that casts sharp shadows, *i.e.*, the illumination can have arbitrary frequency characteristics.

Recent illumination estimation methods extrapolate the full surrounding from a partial scene captured in the limited single view frustum [20, 22, 23, 39, 81–83]. This has also been demonstrated for spatially varying illumination in 2D [5, 24, 63, 66] or 3D [4, 64]. Several methods estimate the geometry as depth or normals, together with the reflectance and illumination from a single scene image [42, 43, 69, 85, 86]. These methods, however, fundamentally assume rich scene information to be present in the image, not just a single object, and high-frequency lighting estimation is conducted mainly through spatial interpolation.

In a concurrent work, Lyu *et al.* [49] propose a generative model of environment maps by applying probabilistic diffusion as a prior on the illumination in a conventional Bayesian framework [14, 46]. The method samples multiple illuminations using this diffusion model. In contrast, our approach reformulates inverse rendering from the ground up as a stochastic inversion process by seamlessly integrating probabilistic diffusion as a forward radiometric formation model on the reflectance map. This eliminates the need of differentiable rendering to bridge the complex domain gap between the image and 3D object surface and enables stochastic generation of the illumination without separate explicit sampling. We show that our method achieves significantly higher accuracy.

Diffusion for Inverse Problems Denoising Diffusion Probabilistic Models [30, 62] have been applied to inverse problems in computer vision including image deblurring, inpainting, and super-resolution [15, 17, 36, 37]. These methods assume known, deterministic forward operators and modify the reverse process based on score-matching. Other methods have used diffusion models by conditioning on the given image [48, 58, 60]. Chung *et al.* [16] extend

these to noisy non-linear inverse problems.

The forward diffusion process can be extended to general signal degradation [6, 19, 31, 38]. Rissanen *et al.* introduced "heat dissipation" as a forward process, and gradually generated low- to high-frequencies in its reverse process [59]. The forward operators in these methods are known. Our problem is similar but significantly more challenging as the operator, *i.e.*, the object reflectance, is unknown—we have a blind inverse problem.

3. Diffusion Reflectance Map

We start by reformulating radiometric image formulation as a forward stochastic process on the reflectance map.

3.1. Forward Reflectance Maps

The reflected radiance L_r at a surface point $\mathbf{x}$ due to incident directional illumination L_i can be described with the rendering equation [35]

$$L_r(\mathbf{x}, \omega_o') = \int_{\Omega'} f_r(\omega_i', \omega_o') L_i(\omega_i')(\omega_i' \cdot \mathbf{n}(\mathbf{x})) d\omega_i', \quad (1)$$

where ω_i' and ω_o' are the incident and outgoing directions of light in the local coordinate frame (*i.e.*, with $\mathbf{n}(\mathbf{x})$ as the north pole) of the surface point, respectively, Ω' is the upper local hemisphere of the surface point, and f_r is the bidirectional reflectance distribution function (BRDF). We assume a surface with homogeneous material properties $f_r(\mathbf{x}, \omega_i', \omega_o') = f_r(\omega_i', \omega_o')$. Spatially varying BRDFs are often handled by pre-segmenting the object surface into regions of different BRDF, or by considering soft assignment indicator values, which we will consider in future work.

Estimation of the reflectance (BRDF f_r) and the illumination (L_i) from a single image (L_r at each point in the viewing direction) is challenging even when the surface geometry (n for every $\mathbf{x}$) is known and the reflectance is homogeneous. The key difficulty stems from the multiplication of the two radiometric ingredients and the integration over the upper hemisphere. These cause frequency attenuation and implicit angular coordinate transform that hamper reflectance and illumination recovery from the object appearance, which past methods tackled with complex non-linear optimization or approximate differentiable rendering.

We can explicitly rotate each local direction into the global coordinate frame and approximate each local integration domain with the hemisphere centered around the viewing direction. This means that, if the surface point $\mathbf{x}$ can be uniquely determined by its surface normal, the surface radiance (*i.e.*, object appearance) can be uniquely defined by $L_r(\mathbf{n})$, the reflectance map [32]

$$L_r(\mathbf{n}) = \int_{\Omega} f_r(\omega_i(\mathbf{n}), \omega_o) L_i(\omega_i(\mathbf{n}))(\omega_i(\mathbf{n}) \cdot \mathbf{n}) d\omega_i. \quad (2)$$
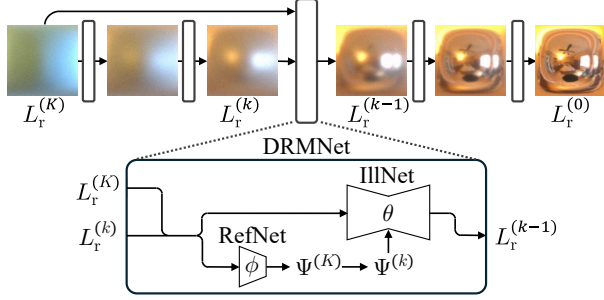
Figure 2. Overall architecture of Diffusion Reflectance Map Network (DRMNet). DRMNet consists of two subnetworks, IllNet for stochastic reverse diffusion to recursively transform the observed reflectance map into a reflectance map of a perfect mirror, and RefNet for jointly and iteratively estimating the reflectance.

Different surface points with the same surface normal can have different radiance due to global light transport, such as cast shadows and interreflection. We assume that these variations can be resolved when we map the object appearance (image) to the reflectance map ($L_r(\mathbf{n})$) (Sec. 4).

3.2. Stochastic Inverse Rendering

Optical image formation adds measurement noise which can be approximated with a zero-mean Gaussian. As such, image formation can be viewed as a Gaussian stochastic forward process on the reflectance map Eq. (2). Now our goal is to recover the reflectance f_r and the illumination L_i in the global coordinate frame, *i.e.*, as an environment map. This clearly necessitates a stochastic generative approach that reverts Gaussian added reflectance map to recover the attenuated frequency components of L_i by f_r.

The reflectance map enables us to formulate single-image reflectance and illumination disentanglement as a geometry-independent stochastic inversion process. This allows us to derive a stochastic inverse rendering approach, a principaled inverse rendering approach with stochasticity seamlessly integrated in the estimation process. In particular, we formulate it as generating a reflectance map of a perfect mirror surface from the observed image. This inverse diffusion process is not that of a regular noise to signal, but of a deterministic forward process with stochastic observation noise, similar to heat dissipation [59].

The inverse generative process can be considered as steps taken along the reflectance space towards a perfect mirror. If we employ an analytical reflectance model, this can be considered as a trajectory in its parameter space. We employ the Disney principled BSDF model [11]

$$
\begin{aligned}
f_r(\boldsymbol{\omega}'_i, \boldsymbol{\omega}'_o; \Psi &:= \{\boldsymbol{\rho}_d, \rho_s, \alpha, \gamma\}) \\
&= (1-\gamma)\frac{\boldsymbol{\rho}_d}{\pi}(f_{\text{diff}}(\boldsymbol{\omega}'_i, \boldsymbol{\omega}'_o) + f_{\text{retro}}(\omega'_i, \omega'_o; \alpha)) \\
&+ f_{\text{spec}}(\boldsymbol{\omega}'_i, \boldsymbol{\omega}'_o; \boldsymbol{\rho}_d, \rho_s, \alpha, \gamma).
\end{aligned} \tag{3}
$$

where we use a partial set of its BSDF model parameters for simplicity: $\boldsymbol{\rho}_d, \rho_s, \alpha$ and γ, the diffuse color, specular strength, roughness and metallicness, respectively. All of the parameters are normalized to $[0, 1]$. Please see the supplementary material for the full model.

3.3. Diffusion Reflectance Map Network

Figure 2 depicts the architecture of Diffusion Reflectance Map Network (DRMNet) which consists of two subnetworks. The first is a diffusion model, IllNet, that learns to invert the forward reflectance map generation process conditioned on the reflectance. The second is a reflectance estimator, RefNet. Input to DRMNet are the observed and current reflectance maps, $L_r^{(K)}$ and $L_r^{(k)}$, respectively.

IllNet estimates the illumination from the observed reflectance map by iteratively generating a reflectance map $L_r^{(k-1)}$ and reflectance parameters one step closer to perfect mirror $f_r^{(k-1)}(\omega'_i, \omega'_o; \Psi^{(k-1)})$ from the current reflectance map $L_r^{(k)}$ and reflectance parameter estimate $\Psi^{(k)}$. The reflectance estimate of the object is $\Psi^{(K)}$, where K is the maximum time step, *i.e.*, the observation. This recursion traces a trajectory in the reflectance parameter space

$$
\Psi^{(k-1)} - \Psi_0 = \eta(\Psi^{(k)} - \Psi_0), \tag{4}
$$

where η is a constant in $(0, 1)$ that controls the speed towards Ψ_0, *i.e.*, a perfect mirror reflectance $[1, 1, 1, 0, 1]$.

The main parameter that determines the highest frequency of the reflectance is the surface roughness α. When $\alpha = 0$, the surface is perfect mirror reflectance, which we would like to achieve as Ψ_0. If we focus on the surface roughness α, the iterative reverse steps defined in Eq. (4) gradually moves it towards 0 and its change $\left|\alpha^{(k-1)} - \alpha^{(k)}\right| = (1-\eta)\alpha^{(k)}$, asymptotically dampens in the process. As these smaller steps in the surface roughness correspond to higher frequencies in the reflectance map, this means, with the uniform steps in Eq. (4), the higher the frequencies, the denser the sampled steps. This contributes to robust recovery of the illumination.

This iterative generation, learned as a reverse diffusion process, starts with the observed reflectance map $L_r^{(K)}$ and reflectance parameters $\Psi^{(K)}$ and IllNet is applied recursively until it reaches a reflectance map $L_r^{(0)}$ corresponding to a perfect mirror Ψ_0. In practice, by definition (Eq. (4)), $\Psi^{(0)}$ only asymptotically reaches Ψ_0. For this, we set K to satisfy $||\Psi^{(0)} - \Psi_0||_2 < \epsilon$ for a small ϵ. At inference time, we do not need to explicitly define K and instead iterate till $\Psi^{(k)}$ reaches $||\Psi^{(k)} - \Psi_0||_2 < \epsilon$.

RefNet learns to estimate the reflectance parameters of the object in the observed single image: $\Psi^{(K)}$. The reflectance parameter $\Psi^{(k)}$ corresponding to the current reflectance map can be computed from this observed re-
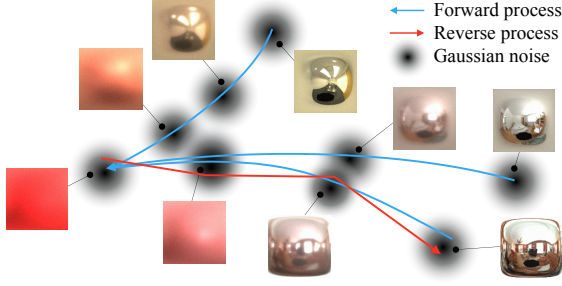
Figure 3. Additive Gaussian observation noise makes each image formation process stochastic, and reflectance attenuates high-frequencies of the illumination whose residues are gradually washed out by the noise. The reverse process is thus necessarily stochastic, effectively generating the lost higher frequency components of the illumination by sampling along learned scores.

flectance estimate by

$$\Psi^{(k)} = \eta^{(K-k)}(\Psi^{(K)} - \Psi_0) + \Psi_0\,, \qquad (5)$$

using the steps taken so far $(K - k)$.

The illumination generation process by IllNet is inherently noisy due to its stochasticity and a direct estimate of the current reflectance parameters $\Psi^{(k)}$ from $L_r^{(k)}$ would be noisy especially as the step size gradually becomes smaller. Instead, by always estimating the observed reflectance and analytically computing the current reflectance, we obtain robust estimates of the current reflectance that are guaranteed to converge. At the same time, this lets RefNet iteratively refine its estimate of the observed object reflectance, *i.e.*, RefNet learns to optimize the object reflectance through its iteration along with IllNet.

We can describe the radiometric formulation of an observed reflectance map $L_r^{(K)}$ from the surrounding illumination $L_r^{(0)}$ and the object reflectance $f_r^{(K)}$ as a forward iterative process

$$q(L_r^{(1:K)}|L_r^{(0)}, f_r^{(K)}) = \prod_{k=1}^{K} q(L_r^{(k)}|L_r^{(0)}, f_r^{(K)})\,. \qquad (6)$$

Let us express the reverse process (DRMNet) using IllNet and RefNet parameters, θ and ϕ, respectively,

$$p_\theta(L_r^{(0:K-1)}|L_r^{(K)}) = \prod_{k=1}^{K} p_{\theta,\phi}(L_r^{(k-1)}|L_r^{(k)})\,. \qquad (7)$$

The conditional $p_{\theta,\phi}$ must invert q in Eq. (6). Although the forward transition q is deterministic, it cannot be inverted analytically, not just because it is highly non-linear, but also because of the information loss. For this, $p_{\theta,\phi}$ must be a stochastic generative model. In order to make this reverse process stochastic, we must in turn make the forward process q also stochastic while at the same time obeying the deterministic radiometric image formation process.
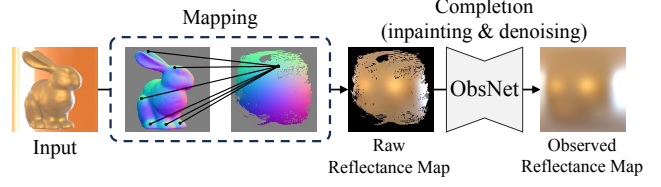

Figure 4. We convert the single input image into an observation reflectance map by "completing" the sparsely mapped reflectance map with another diffusion model, ObsNet.

Similar to [59], we resolve this by adding stochastic perturbations to the deterministic forward process (*i.e.*, the Gaussian observation noise)

$$q(L_r^{(k)}|L_r^{(0)}, f_r^{(K)}) = \mathcal{N}(L_r^{(k)}|L_r(\mathbf{n}; L_i, f_r^{(k)}), \sigma^2\mathbf{I})\,, \quad (8)$$

where $f_r(k)$ is the reflectance whose parameter $\Psi^{(k)}$ is computed from the observed reflectance estimate $\Psi^{(K)}$ by Eq. (5), and σ is the standard deviation of the added noise.

The stochastic additive observation noise washes out the high-frequency components of the illumination, the bulk of which is already attenuated by the reflectance. Figure 3 depicts this forward radiometric object appearance process q for different combinations of illumination L_i and reflectance $f_r^{(K)}$. As the reflectance deviates from perfect mirror reflection, the reflected illumination, *i.e.*, the reflectance map, looses high-frequency details and color and all combinations become indistinguishable.

The conditional $p_{\theta,\phi}$ models the reverse process of this stochastic forward radiometric process. Starting from the observed reflectance map, by iteratively reverse transitioning with the conditional $p_{\theta,\phi}$, we can generate one illumination that gives rise to the observed reflectance map for that object reflectance among the infinite possibilities. We model this reverse transition $p_{\theta,\phi}$ with a Gaussian and explicitly condition it on the observed reflectance map (*i.e.*, object appearance) $L_r^{(K)}$

$$p_{\theta,\phi}(L_r^{(k-1)}|L_r^{(k)}, L_r^{(K)}, f_r^{(k)}) = \mathcal{N}(L_r^{(k-1)}|\mu_{\theta,\phi}(L_r^{(k)}, L_r^{(K)}), \delta^2\mathbf{I})\,.$$
$$(9)$$

We learn the expectation $\mu_{\theta,\phi}$ of this Gaussian

$$\mu_\theta(L_r^{(k)}, L_r^{(K)}, \Psi_\phi^{(k)}(L_r^{(k)}, L_r^{(K)}, K - k))\,,$$

and set the variance to a constant δ. We further reparameterize $\mu_\theta(L_r^{(k)}, L_r^{(K)}, \Psi_\phi^{(k)}) = L_r^{(k)} + \varepsilon_\theta(L_r^{(k)}, L_r^{(K)}, \Psi_\phi^{(k)})$ to train IllNet from the residuals.

We maximize the evidence lower bound (ELBO) of the marginalized reverse transition $p_{\theta,\phi}(L_r^{(0)}|L_r^{(K)})$ by minimizing the upper bound on the negative log likelihood. In the supplementary material, we derive this step by step, which leads to a simplified objective

$$\mathcal{L}_i = \mathbb{E}_{L_i, f_r, k}|\mu_{\theta,\phi}(\mathbf{n}; L_r^{(k)}, L_r^{(K)}) - L_r(\mathbf{n}; L_i, f_r^{(k)})|_2^2\,.$$
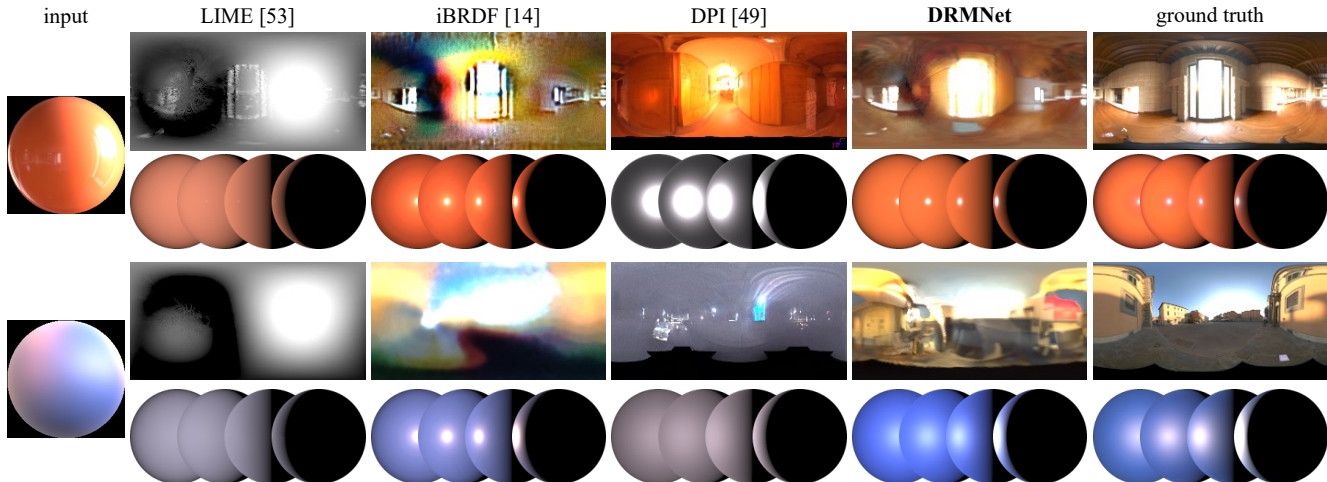$$(10)$$

Figure 5. Qualitative results on iBRDF synthetic dataset. For each input, the top row is the illumination estimate shown as a spherical panorama and the bottom row is the reflectance estimate rendered as a sphere under a point source.

When training DRMNet, the illumination L_i, reflectance $f_r^{(K)}$, and step k are uniformly sampled from a set of environment maps, reflectance parameter values, and $[1, K]$, respectively. As the reflectance map corresponding to perfect mirror reflection $L_r^{(0)}$ is equivalent to the incident illumination L_i, we can use L_i as the expectation instead of $L_r^{(0)}$.

In addition to the loss on the reflectance map Eq. (10), we impose a reflectance loss on the reflectance parameters estimated by RefNet $\Psi_\phi^{(K)}$ using the reflectance parameters corresponding to the forward step and process, Eq. (8) $f_r^{(k)}$ and Eq. (6) $f_r^{(K)}$, respectively,

$$\mathcal{L}_r = \mathbb{E}_{L_i, f_r, k} \left[|\Psi_\phi^{(k)} - \Psi^{(k)}|_2^2 + |\Psi_\phi^{(K)} - \Psi^{(K)}|_2^2 \right] . \tag{11}$$

The final loss is $\mathcal{L} = \lambda_i \mathcal{L}_i + \lambda_r \mathcal{L}_r$, where λ_i and λ_r are hyperparameters weighting the ELBO and reflectance losses. IllNet is implemented with a U-Net and RefNet with a simple MLP. Please see the supplementary material for detailed architectures of IllNet and RefNet.

4. Observed Reflectance Map

How do we get the first input to DRMNet from the single input image? Given the single image of an object whose geometry as surface normals are known, we can compute its reflectance map by mapping each pixel onto a Gaussian hemisphere, *i.e.* a hemisphere parameterized by the surface normals with its north pole pointing towards the viewpoint. Each pixel p gives a pair of surface radiance L_p and surface normal $\mathbf{n}_p$ related by Eq. (2)

$$L_p = L_r(\mathbf{x}_p, \mathbf{R}_{\mathbf{n_p}} w_{o,p}) , \tag{12}$$

	illumination	reflectance
	logRMSE↓/PSNR↑/SSIM↑/LPIPS↓	logRMSE↓
LIME [53]	5.74 / 8.7 / 0.24 / 0.70	1.88
iBRDF [14]	3.20 / 11.5 / 0.34 / 0.67	1.01
DPI [49]	4.22 / 10.8 / 0.25 / 0.68	2.03
DRMNet (Ours)	**2.62 / 14.4 / 0.59 / 0.59**	**0.66**

Table 1. Quantitative results on the iBRDF synthetic dataset [14]. Our DRMNet acheives state-of-the-art accuracy.

where $\mathbf{x}_p$ is the surface point of pixel p. We assume orthographic projection and use the camera coordinate frame so that $w_{o,p} = [0, 0, -1]$ and a homogeneous surface $L_p = L_r(\mathbf{n_p})$. The reflectance map $L_r(\mathbf{n_p})$ can thus be computed by extracting the surface normal of each image pixel.

As each surface normal of the reflectance map can be observed at multiple image pixels, the mapping from the input image to the reflectance map is not injective. Global light transport manifests in self-shadows and interreflection causing darker and brighter radiance, respectively, for the same surface normal at other surface points that are directly lit. We minimize global light transport effects in our reflectance map mapping by taking the median pixel intensities of the multiple surface points with the same surface normal

$$r(\boldsymbol{n}) = \text{median}\{L_p | \theta_p := \cos^{-1}(\boldsymbol{n}_p \cdot \boldsymbol{n}) < \epsilon\}, \tag{13}$$

where ϵ is the angular threshold to determine the set of pixels corresponding to a given surface normal.

This reflectance map is inevitably sparse unless the surface normals of the object cover all possible directions uniformly, *i.e.*, it is a hemisphere facing the camera. We "complete" this reflectance map with another diffusion model, which we refer to as ObsNet. Unlike regular diffusion for image inpainting [6], we train the diffusion model to learn to generate the observed regions as they can be inaccurate
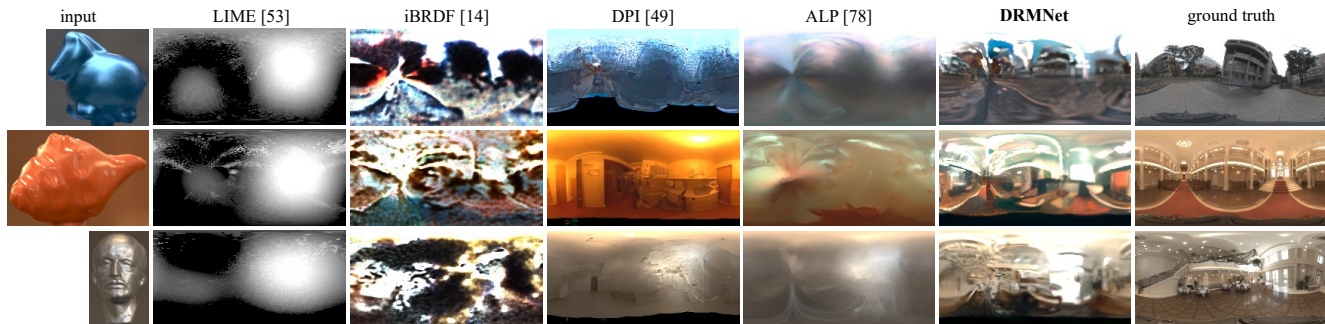
Figure 6. Illumination estimates of the nLMVS-Real dataset for different objects taken in complex environments. DRMNet successfully recovers accurate and plausible detailed illumination from the frequency-attenuated object appearance. Note that ALP knows the reflectance.

	illumination	est. time
	logRMSE↓/PSNR↑/SSIM↑/LPIPS↓	secs↓
LIME [53]	7.05 / 8.11 / 0.14 / 0.71	**1.1**
iBRDF [14]	2.13 / 10.6 / 0.17 / 0.70	5048
DPI [49]	3.57 / 12.6 / 0.27 / <u>0.65</u>	1143
ALP [78]	**0.95** / **15.3** / **0.39** / 0.66	267
DRMNet (Ours)	<u>1.18</u> / <u>14.6</u> / <u>0.33</u> / **0.60**	<u>14.3</u>

Table 2. Quantitative evaluation of the illumination estimates on the nLMVS-Real dataset [75]. Note that ALP requires reflectance which we provide by estimation from multiview images a priori. DRMNet achieves comparable accuracy 20 times faster while jointly estimating the reflectance from a single image.

	illumination	relighting
	logRMSE↓/DSSIM↓	logRMSE↓/DSSIM↓
DelightNet	0.933 / 0.365	1.110 / 0.186
iBRDF	0.864 / 0.329	1.027 / **0.077**
Ours	**0.694** / **0.308**	**0.414** / 0.107

Table 3. Quantitative results on the DeLight-Net dataset. DRM-Net consistently achieves high accuracy in both illumination and reflectance estimation. See text for details.

due to global light transport effects that slipped through the median mapping. ObsNet can be trained on a large number of synthetic reflectance maps rendered for various combinations of reflectance and illumination masked by surface normal distributions of randomly selected objects.

5. Experiments

We evaluate the effectiveness of DRMNet quantitatively and qualitatively and compare with related existing inverse-rendering methods [14, 49, 53, 78] on synthetic and real images. In the supplementary material, we also justify our formulation of stochastic inverse rendering with probabilistic diffusion as well as the specific architecture of DRMNet through ablation studies, and analyze the stochastic behavior of DRMNet over multiple estimations on the same input. Please see supplementary material for more results.
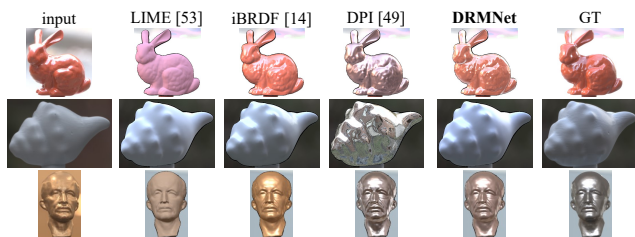

Figure 7. Relighting results using reflectance estimates on the nLMVS-Real dataset. DRMNet results match ground truth well, demonstrating the accuracy of its recovered illumination. ALP cannot be applied as it requires reflectance pre-acquisition.

Datasets We train DRMNet on a large dataset of synthetic reflectance maps rendered with Mitsuba3 [33]. We use environment maps from the Laval Indoor Dataset [10, 22] and Poly Haven HDRIs [2] for illumination. Please see the supplementary material for details about how we randomly sample these illumination and reflectance to create a large training and test dataset. We use the same dataset to train ObsNet. All data and code can be found in the project page.

For thorough comparative studies, we use the iBRDF synthetic dataset [14], the nLVMS real dataset [75], and the DeLight-Net real dataset [27]. The iBRDF dataset consists of images of spheres rendered with HDR environment maps [21] and measured BRDFs [51]. This dataset allows us to quantitatively evaluate the accuracy of reflectance estimates, which is not possible with real images. We use a large enough subset of this dataset to achieve thorough quantitative comparison in realistic time. The nLMVS dataset consists of images captured in six different illumination environments for 20 different objects of 5 different shapes and 4 different reflectances, all with ground truth geometry and HDR environment maps. We test on one view for each environment. The DeLight-Net dataset consists of images capturing real spheres and is used to compare with the method by Georgoulis *et al.* [27] as their code is not available (link broken).
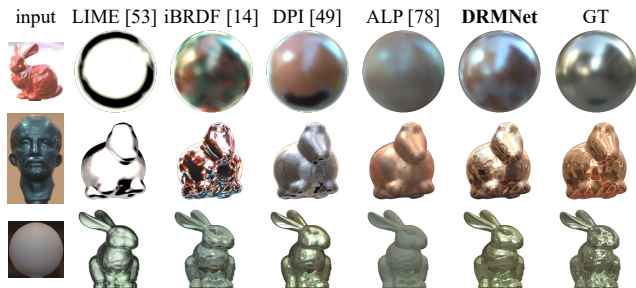
Figure 8. Object replacement results on the nLMVS-Real dataset. Accurate illumination estimates of DRMNet enable rendering of detailed object appearance for objects with arbitrary reflectance.

| | logRMSE↓/PSNR↑/SSIM↑/LPIPS↓ | |
	relighting	object replacement
LIME [53]	1.73 / 18.6 / <u>0.81</u> / 0.40	1.67 / 14.0 / 0.73 / 0.25
iBRDF [14]	<u>1.69</u> / <u>19.9</u> / **0.83** / <u>0.35</u>	0.58 / 19.2 / 0.80 / 0.22
DPI [49]	1.75 / 17.7 / 0.73 / 0.44	0.49 / 21.0 / <u>0.83</u> / 0.21
ALP [78];	(pre-acquired reflectance)	<u>0.39</u> / <u>22.2</u> / **0.86** / <u>0.19</u>
DRMNet	**1.68** / **20.6** / 0.83 / 0.34	**0.35** / **23.3** / **0.86** / **0.17**

Table 4. Quantitative evaluation of relighting with rotated geometry and object replacement on the nLMVS-Real dataset [75]. Our method achieves highest accuracy in both.

Metrics We use several metrics to quantitatively evaluate the accuracy of the illumination and reflectance estimates. For the reflectance, we use the log-scale RMSE in the non-parametric MERL BRDF representation as in iBRDF [14]. The non-parametric representation enables accuracy comparison regardless of the BRDF model. Scale-invariance absorbs the ambiguity in absolute radiance due to exposure and overall illumination brightness. We also evaluate reflectance estimates through relighting results with log-scale RMSE and SSIM [68]. For quantitative evaluation of the illumination estimates, we compute scale-invariant log-scale RMSE, PSNR, SSIM, and LPIPS [84] at 128×256 resolution. For the PSNR, SSIM, and LPIPS, we compute them on LDR images after tone-mapping HDR illumination.

iBRDF synthetic dataset Table 1 and Fig. 5 show quantitative and qualitative evaluation of the accuracy of illumination and reflectance estimates using the iBRDF synthetic dataset. DRMNet achieves higher accuracy and recovers more natural illumination. As ALP [78] requires estimation of the shape and reflectance parameters at many viewpoints, we do not compare with it here. Figure 5 shows that the concurrent work by Lyu *et al*. [49] results in illumination estimates that significantly deviate from the ground truth, as it is a naive noise seeded diffusion model used as an external prior in a classic Bayesian inverse-rendering formulation.

nLMVS real dataset We evaluate the estimation accuracy on real images using the nLMVS-Real dataset [75]. Table 2 shows quantitative results. Note that we have also included ALP [78] for comparison, but the method requires known reflectance, pre-acquired from multiple images under known lighting. Our method achieves accuracy comparable to this known-reflectance method, but in an order faster computation as DRMNet does not require complex non-linear optimization.

As ground truth reflectance is unknown, we evaluate the accuracy of the reflectance estimate through relighting. Figure 7 shows relighting results under a different illumination using the estimated reflectance and rotated geometry to align with that in the ground truth image. The results show strong consistency in object appearance, suggesting high accuracy of reflectance estimates. Figure 8 shows renderings of different objects under the estimated illumination. Since our method explicitly recovers the high frequency spectrum of the illumination, even objects with higher frequency than the one used to recover the illumination can be relit with natural appearance. Please see supplementary material for more results in higher resolution. Table 4 shows quantitative results corresponding to Figs. 7 and 8. DRMNet achieves highest accuracy in both.

DeLight-Net real dataset Table 3 shows quantitative results on the DeLightNet dataset. "Illumination" is the estimated illumination reflected by a mirror sphere and Georgoulis *et al*. [27] directly regress this from the reflectance map. Note that DeLightNet does not estimate the reflectance. To directly compare with their method, we evaluate using log-scale RMSE and DSSIM [1]. Our DRMNet achieves higher accuracy over the two existing methods in both illumination and reflectance estimates except for one metric. We plan to incorporate a learned BRDF model, *e.g.*, iBRDF [14], in future extensions.

6. Conclusion

We introduced DRMNet, a principled stochastic inverse rendering method that estimates the illumination and reflectance from a single image of an object of known geometry taken under complex natural illumination. By formulating the radiometric disentanglement on the reflectance map and as a recursive inversion of stochastic diffusion but with an underlying deterministic rendering equation, we showed that DRMNet achieves state-of-the-art accuracy and full detailed recovery of surrounding illumination, enabling relighting and replacement of objects with arbitrary reflectance properties. By breaking the fundamental limitations of inverse rendering with a seamlessly integrated generative model, DRMNet opens new possibilities of single-image radiometric understanding.

References

[1] DSSIM - image comparison tool. https://pngquant.org/dssim.html. 8

[2] HDRIs: poly haven. https://polyhaven.com/hdris. 7

[3] Perception viewed as an inverse problem. *Vision Research*, 41(24):3145–3161, 2001. 2

[4] Jiayang Bai, Jie Guo, Chenchen Wang, Zhenyu Chen, Zhen He, Shan Yang, Piaopiao Yu, Yan Zhang, and Yanwen Guo. Deep Graph Learning for Spatially-Varying Indoor Lighting Prediction. *Science China Information Sciences*, 66(3): 132106, 2023. 3

[5] Jiayang Bai, Zhen He, Shan Yang, Jie Guo, Zhenyu Chen, Yan Zhang, and Yanwen Guo. Local-to-Global Panorama Inpainting for Locale-Aware Indoor Lighting Prediction, 2023. 3

[6] Arpit Bansal, Eitan Borgnia, Hong-Min Chu, Jie S. Li, Hamid Kazemi, Furong Huang, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Cold Diffusion: Inverting Arbitrary Image Transforms Without Noise, 2022. 3, 6

[7] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-NeRF: A Multiscale Representation for Anti-Aliasing Neural Radiance Fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5855–5864, 2021. 2

[8] Ronen Basri and David W. Jacobs. Lambertian Reflectance and Linear Subspaces. 25(2), 2003. 1

[9] Tianteng Bi, Junjie Ma, Yue Liu, Dongdong Weng, and Yongtian Wang. SIR-Net: Self-Supervised Transfer for Inverse Rendering via Deep Feature Fusion and Transformation From a Single Image. *IEEE Access*, 8:201861–201873, 2020. 2

[10] Christophe Bolduc, Justine Giroux, Marc Hébert, Claude Demers, and Jean-François Lalonde. Beyond the Pixel: a Photometrically Calibrated HDR Dataset for Luminance and Color Prediction. In *ICCV*, pages 8071–8081, 2023. 7

[11] Brent Burley. Physically-based Shading at Disney. In *siggraph*, 2012. 4

[12] Wenzheng Chen, Jun Gao, Huan Ling, Edward J. Smith, Jaakko Lehtinen, Alec Jacobson, and Sanja Fidler. Learning to Predict 3D Objects with an Interpolation-Based Differentiable Renderer. In *NeurIPS*, Red Hook, NY, USA, 2019. Curran Associates Inc. 3

[13] Wenzheng Chen, Joey Litalien, Jun Gao, Zian Wang, Clement Fuji Tsang, Sameh Khamis, Or Litany, and Sanja Fidler. DIB-R++: learning to predict lighting and material with a hybrid differentiable renderer. In *NeurIPS*, pages 22834–22848, 2021. 3

[14] Zhe Chen, Shohei Nobuhara, and Ko Nishino. Invertible Neural BRDF for Object Inverse Rendering. *IEEE TPAMI*, 44(12):9380–9395, 2022. 2, 3, 6, 7, 8

[15] Hyungjin Chung, Byeongsu Sim, Dohoon Ryu, and Jong Chul Ye. Improving Diffusion Models for Inverse Problems using Manifold Constraints. pages 25683–25696, 2022. 3

[16] Hyungjin Chung, Jeongsol Kim, Sehui Kim, and Jong Chul Ye. Parallel Diffusion Models of Operator and Image for Blind Inverse Problems. In *CVPR*, pages 6059–6069, 2023. 3

[17] Hyungjin Chung, Jeongsol Kim, Michael Thompson Mccann, Marc Louis Klasky, and Jong Chul Ye. Diffusion Posterior Sampling for General Noisy Inverse Problems. In *ICLR*, 2023. 3

[18] Peng Dai, Yinda Zhang, Xin Yu, Xiaoyang Lyu, and Xiaojuan Qi. Hybrid Neural Rendering for Large-Scale Scenes with Motion Blur. In *CVPR*, pages 154–164, 2023. 2

[19] Giannis Daras, Mauricio Delbracio, Hossein Talebi, Alexandros Dimakis, and Peyman Milanfar. Soft Diffusion: Score Matching with General Corruptions. *TMLR*, 2023. 3

[20] Mohammad Reza Karimi Dastjerdi, Jonathan Eisenmann, Yannick Hold-Geoffroy, and Jean-François Lalonde. EverLight: Indoor-Outdoor Editable HDR Lighting Estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7420–7429, 2023. 3

[21] Paul Debevec. Rendering Synthetic Objects into Real Scenes: Bridging Traditional and Image-Based Graphics with Global Illumination and High Dynamic Range Photography. In *SIGGRAPH*, pages 1–10, New York, NY, USA, 2008. Association for Computing Machinery. 7

[22] Marc-André Gardner, Kalyan Sunkavalli, Ersin Yumer, Xiaohui Shen, Emiliano Gambaretto, Christian Gagné, and Jean-François Lalonde. Learning to Predict Indoor Illumination from a Single Image. *ACM TOG*, 36(6), 2017. 3, 7

[23] Marc-Andre Gardner, Yannick Hold-Geoffroy, Kalyan Sunkavalli, Christian Gagné, and Jean-François Lalonde. Deep Parametric Indoor Lighting Estimation. In *ICCV*, pages 7174–7182, 2019. 3

[24] Mathieu Garon, Kalyan Sunkavalli, Sunil Hadap, Nathan Carr, and Jean-François Lalonde. Fast Spatially-Varying Indoor Lighting Estimation. In *CVPR*, pages 6901–6910, 2019. 3

[25] Stamatios Georgoulis, Konstantinos Rematas, Tobias Ritschel, Mario Fritz, Luc Van Gool, and Tinne Tuytelaars. DeLight-Net: Decomposing Reflectance Maps into Specular Materials and Natural Illumination. *arXiv*, abs/1603.08240, 2016. 2

[26] Stamatios Georgoulis, Konstantinos Rematas, Tobias Ritschel, Mario Fritz, Tinne Tuytelaars, and Luc Van Gool. What Is Around the Camera? In *ICCV*, pages 5170–5178, 2017. 3

[27] Stamatios Georgoulis, Konstantinos Rematas, Tobias Ritschel, Efstratios Gavves, Mario Fritz, Luc Van Gool, and Tinne Tuytelaars. Reflectance and Natural Illumination from Single-Material Specular Objects Using Deep Learning. *IEEE TPAMI*, 40(8):1932–1947, 2018. 2, 7, 8

[28] Michael Goesele, Brian Curless, and Steven M Seitz. Multi-View Stereo Revisited. In *CVPR*, pages 2402–2409, 2006. 2

[29] Pat Hanrahan, Jitendra Malik, Henrik Wann Jensen, and Steve Marschner. Standford CS448c: Appearance Models for Computer Graphics and Vision.

https://graphics.stanford.edu/courses/cs448c-00-fall/, 2000. 2

[30] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. In *NeurIPS*, pages 6840–6851. Curran Associates, Inc., 2020. 3

[31] Emiel Hoogeboom and Tim Salimans. Blurring Diffusion Models. In *ICLR*, 2023. 3

[32] Berthold K.P. Horn. *Robot Vision*. The MIT Press, 1986. 3

[33] Wenzel Jakob, Sébastien Speierer, Nicolas Roussel, Merlin Nimier-David, Delio Vicini, Tizian Zeltner, Baptiste Nicolet, Miguel Crespo, Vincent Leroy, and Ziyi Zhang. Mitsuba 3 renderer, 2022. https://mitsuba-renderer.org. 7

[34] Michael Janner, Jiajun Wu, Tejas D Kulkarni, Ilker Yildirim, and Josh Tenenbaum. Self-Supervised Intrinsic Image Decomposition. page 5938–5948, 2017. 3

[35] James T. Kajiya. The Rendering Equation. 4:143–150, 1986. 3

[36] Bahjat Kawar, Gregory Vaksman, and Michael Elad. SNIPS: Solving noisy inverse problems stochastically. pages 21757–21769, 2021. 3

[37] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising Diffusion Restoration Models. In *NeurIPS*, pages 23593–23606, 2022. 3

[38] Sangyun Lee, Hyungjin Chung, Jaehyeon Kim, and Jong Chul Ye. Progressive Deblurring of Diffusion Models for Coarse-to-Fine Image Synthesis, 2022. 3

[39] 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 *CVPR*, pages 5918–5928, 2019. 3

[40] Chenhao Li, Trung Thanh Ngo, and Hajime Nagahara. Inverse Rendering of Translucent Objects using Physical and Neural Renderers. In *CVPR*, pages 12510–12520, 2023. 2

[41] Zhengqi Li and Noah Snavely. Learning Intrinsic Image Decomposition from Watching the World. In *CVPR*, pages 9039–9048, 2018. 2

[42] 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*, pages 2472–2481, 2019. 3

[43] Zhengqin Li, Jia Shi, Sai Bi, Rui Zhu, Kalyan Sunkavalli, Miloš Hašan, Zexiang Xu, Ravi Ramamoorthi, and Manmohan Chandraker. Physically-Based Editing of Indoor Scene Lighting from a Single Image. In *ECCV*, pages 555–572. Springer, 2022. 3

[44] Zongrui Li, Qian Zheng, Boxin Shi, Gang Pan, and Xudong Jiang. DANI-Net: Uncalibrated Photometric Stereo by Differentiable Shadow Handling, Anisotropic Reflectance Modeling, and Neural Inverse Rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8381–8391, 2023. 2

[45] Yuan Liu, Peng Wang, Cheng Lin, Xiaoxiao Long, Jiepeng Wang, Lingjie Liu, Taku Komura, and Wenping Wang. NeRO: Neural Geometry and BRDF Reconstruction of Reflective Objects from Multiview Images. *ACM Trans. Graph.*, 42(4), 2023. 2

[46] Stephen Lombardi and Ko Nishino. Reflectance and Natural Illumination from a Single Image. In *ECCV*, pages 582–595. Springer, 2012. 2, 3

[47] Stephen Lombardi and Ko Nishino. Radiometric Scence Decomposition: Scene Reflectance, Illumination, and Geometry from RGB-D Images. In *3DV*, 2016. 2

[48] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. RePaint: Inpainting using Denoising Diffusion Probabilistic Models. In *CVPR*, pages 11461–11471, 2022. 3

[49] Linjie Lyu, Ayush Tewari, Marc Habermann, Shunsuke Saito, Michael Zollhöfer, Thomas Leimküehler, and Christian Theobalt. Diffusion Posterior Illumination for Ambiguity-aware Inverse Rendering. *ACM TOG*, 42(6), 2023. 3, 6, 7, 8

[50] Stephen Robert Marschner. *Inverse Rendering for Computer Graphics*. Cornell University, 1998. 2

[51] Wojciech Matusik, Hanspeter Pfister, Matt Brand, and Leonard McMillan. A Data-Driven Reflectance Model. *ACM TOG*, 22(3):759–769, 2003. 7

[52] Maxim Maximov, Laura Leal-Taixé, Mario Fritz, and Tobias Ritschel. Deep Appearance Maps. In *ICCV*, pages 8729–8738, 2019. 2

[53] Abhimitra Meka, Maxim Maximov, Michael Zollhoefer, Avishek Chatterjee, Hans-Peter Seidel, Christian Richardt, and Christian Theobalt. LIME: Live Intrinsic Material Estimation. In *CVPR*, pages 6315–6324, 2018. 2, 6, 7, 8

[54] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 2

[55] Jinwoo Park, Hunmin Park, Sung-Eui Yoon, and Woontack Woo. Physically-inspired Deep Light Estimation from a Homogeneous-Material Object for Mixed Reality Lighting. *IEEE TVCG*, 26(5):2002–2011, 2020. 3

[56] Ravi Ramamoorthi and Pat Hanrahan. A Signal-Processing Framework for Inverse Rendering. In *SIGGRAPH*, pages 117–128, 2001. 1

[57] Konstantinos Rematas, Tobias Ritschel, Mario Fritz, Efstratios Gavves, and Tinne Tuytelaars. Deep Reflectance Maps. In *CVPR*, pages 4508–4516, 2016. 2

[58] Mengwei Ren, Mauricio Delbracio, Hossein Talebi, Guido Gerig, and Peyman Milanfar. Multiscale Structure Guided Diffusion for Image Deblurring. In *ICCV*, pages 10721–10733, 2023. 3

[59] Severi Rissanen, Markus Heinonen, and Arno Solin. Generative Modelling With Inverse Heat Dissipation. In *ICLR*, 2023. 3, 4, 5

[60] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *SIGGRAPH*, pages 1–10, 2022. 3

[61] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-Motion Revisited. In *CVPR*, pages 4104–4113, 2016. 2

[62] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. In *ICML*, pages 2256–2265. PMLR, 2015. 3

[63] Shuran Song and Thomas A. Funkhouser. Neural Illumination: Lighting Prediction for Indoor Environments. In *CVPR*, pages 6911–6919, 2019. 3

[64] Pratul Srinivasan, Ben Mildenhall, Matthew Tancik, Jonathan Barron, Richard Tucker, and Noah Snavely. Lighthouse: Predicting Lighting Volumes for Spatially-Coherent Illumination. pages 8077–8086, 2020. 3

[65] Tristan Swedish, Connor Henley, and Ramesh Raskar. Objects as Cameras: Estimating High-Frequency Illumination from Shadows. In *ICCV*, pages 2593–2602, 2021. 3

[66] Jiajun Tang, Yongjie Zhu, Haoyu Wang, Jun Hoong Chan, Si Li, and Boxin Shi. Estimating Spatially-Varying Lighting In Urban Scenes With Disentangled Representation. In *ECCV*, page 454–469, Berlin, Heidelberg, 2022. Springer-Verlag. 3

[67] Dor Verbin, Peter Hedman, Ben Mildenhall, Todd E. Zickler, Jonathan T. Barron, and Pratul P. Srinivasan. Ref-NeRF: Structured View-Dependent Appearance for Neural Radiance Fields. In *CVPR*, pages 5481–5490, 2021. 2

[68] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 8

[69] Zian Wang, Jonah Philion, Sanja Fidler, and Jan Kautz. Learning Indoor Inverse Rendering with 3D Spatially-Varying Lighting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12538–12547, 2021. 3

[70] Henrique Weber, Donald Prévost, and Jean-Francois Lalonde. Learning to Estimate Indoor Lighting from 3D Objects. In *3DV*, pages 199–207, 2018. 3

[71] Xin Wei, Guojun Chen, Yue Dong, Stephen Lin, and Xin Tong. Object-Based Illumination Estimation with Rendering-Aware Neural Networks. In *ECCV*, page 380–396, Berlin, Heidelberg, 2020. Springer-Verlag. 3

[72] Felix Wimbauer, Shangzhe Wu, and C. Rupprecht. De-rendering 3D Objects in the Wild. *CVPR*, pages 18469–18478, 2022. 2

[73] Robert J Woodham. Photometric Method for Determining Surface Orientation from Multiple Images. *Optical engineering*, 19(1):139–144, 1980. 2

[74] Shangzhe Wu, Ameesh Makadia, Jiajun Wu, Noah Snavely, Richard Tucker, and Angjoo Kanazawa. De-rendering the World's Revolutionary Artefacts. In *CVPR*, pages 6338–6347, 2021. 2

[75] Kohei Yamashita, Yuto Enyo, Shohei Nobuhara, and Ko Nishino. nLMVS-Net: Deep Non-Lambertian Multi-View Stereo. In *WACV*, pages 3037–3046, 2023. 7, 8

[76] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview Neural Surface Reconstruction by Disentangling Geometry and Appearance. In *NeurIPS*, pages 2492–2502, 2020. 2

[77] Renjiao Yi, Chenyang Zhu, and Kai Xu. Weakly-supervised Single-view Image Relighting. In *CVPR*, pages 8402–8411, 2023. 3

[78] Hong-Xing Yu, Samir Agarwala, Charles Herrmann, Richard Szeliski, Noah Snavely, Jiajun Wu, and Deqing Sun. Accidental Light Probes. In *CVPR*, pages 12521–12530. IEEE Computer Society, 2023. 3, 7, 8

[79] Ye Yu and W. Smith. InverseRenderNet: Learning Single Image Inverse Rendering. pages 3150–3159, 2018. 3

[80] Ye Yu and William AP Smith. Outdoor Inverse Rendering From a Single Image Using Multiview Self-Supervision. *IEEE TPAMI*, 44(7):3659–3675, 2021. 3

[81] Fangneng Zhan, Changgong Zhang, Wenbo Hu, Shijian Lu, Feiying Ma, Xuansong Xie, and Ling Shao. Sparse Needlets for Lighting Estimation With Spherical Transport Loss. In *ICCV*, pages 12830–12839, 2021. 3

[82] Fangneng Zhan, Changgong Zhang, Yingchen Yu, Yuan Chang, Shijian Lu, Feiying Ma, and Xuansong Xie. EM-Light: Lighting Estimation via Spherical Distribution Approximation. In *AAAI*, pages 3287–3295, 2021.

[83] Fangneng Zhan, Yingchen Yu, Changgong Zhang, Rongliang Wu, Wenbo Hu, Shijian Lu, Feiying Ma, Xuansong Xie, and Ling Shao. GMLight: Lighting Estimation via Geometric Distribution Approximation. *IEEE TIP*, 31:2268–2278, 2022. 3

[84] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 8

[85] Jingsen Zhu, Fujun Luan, Yuchi Huo, Zihao Lin, Zhihua Zhong, Dianbing Xi, Rui Wang, Hujun Bao, Jiaxiang Zheng, and Rui Tang. Learning-Based Inverse Rendering of Complex Indoor Scenes with Differentiable Monte Carlo Raytracing. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–8. ACM, 2022. 3

[86] Rui Zhu, Zhengqin Li, Janarbek Matai, Fatih Porikli, and Manmohan Chandraker. IRISformer: Dense Vision Transformers for Single-Image Inverse Rendering in Indoor Scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2822–2831, 2022. 3