Diffusion Posterior Illumination for Ambiguity-aware Inverse Rendering

LINJIE LYU, Max-Planck-Institut für Informatik, Germany
AYUSH TEWARI, MIT CSAIL, USA
MARC HABERMANN, Max-Planck-Institut für Informatik, Germany
SHUNSUKE SAITO, Reality Labs Research, USA
MICHAEL ZOLLHÖFER, Reality Labs Research, USA
THOMAS LEIMKÜHLER, Max-Planck-Institut für Informatik, Germany
CHRISTIAN THEOBALT, Max-Planck-Institut für Informatik, Germany

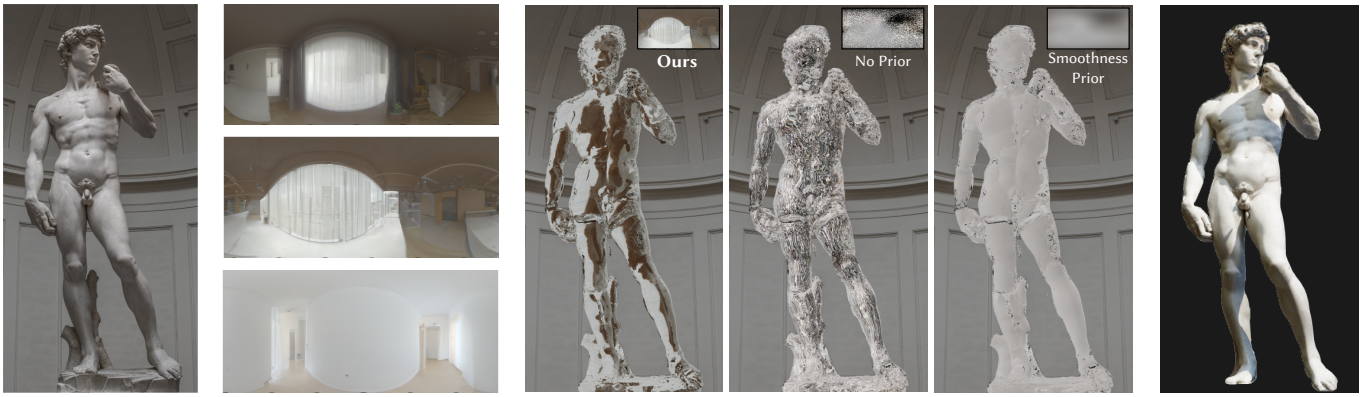

Fig. 1. Given one or multiple images of a scene (*Input*), our method infers surface materials and illumination. We model the inherent ambiguities involved in this process: The statue could have been lit by an infinite number of environment maps, all giving rise to the same image. Our diffusion prior allows to sample from a joint distribution of materials and natural illumination conditions (*Illumination Samples*) that all explain the input. This enables high-quality *Material Editing* (here demonstrating a change to a translucent material; environment map reconstructions are shown as insets) and *Relighting*. Without any prior, the recovered illumination is noisy, the commonly used smoothness prior converges to a single blurry result. In contrast, our samples are diverse and natural.

Inverse rendering, the process of inferring scene properties from images, is a challenging inverse problem. The task is ill-posed, as many different scene configurations can give rise to the same image. Most existing solutions incorporate priors into the inverse-rendering pipeline to encourage plausible solutions, but they do not consider the inherent ambiguities and the multi-modal distribution of possible decompositions. In this work, we propose a novel scheme that integrates a denoising diffusion probabilistic model pre-trained on natural illumination maps into an optimization framework involving a differentiable path tracer. The proposed method allows sampling from combinations of illumination and spatially-varying surface materials that are, both, natural and explain the image observations. We further conduct an extensive comparative study of different priors on illumination used in previous work on inverse rendering. Our method excels in recovering materials and producing highly realistic and diverse environment map samples that faithfully explain the illumination of the input images.

ACM Reference Format:
Linjie Lyu, Ayush Tewari, Marc Habermann, Shunsuke Saito, Michael Zollhöfer, Thomas Leimkühler, and Christian Theobalt. 2023. Diffusion Posterior Illumination for Ambiguity-aware Inverse Rendering. *ACM Trans. Graph.* 42, 6 (December 2023), 14 pages. https://doi.org/10.1145/3618357

Authors' addresses: Linjie Lyu, llyu@mpi-inf.mpg.de, Max-Planck-Institut für Informatik, Germany; Ayush Tewari, ayusht@mit.edu, MIT CSAIL, USA; Marc Habermann, mhaberma@mpi-inf.mpg.de, Max-Planck-Institut für Informatik, Germany; Shunsuke Saito, shunsuke.saito16@gmail.com, Reality Labs Research, USA; Michael Zollhöfer, zollhoefer@meta.com, Reality Labs Research, USA; Thomas Leimkühler, thomas.leimkuehler@mpi-inf.mpg.de, Max-Planck-Institut für Informatik, Germany; Christian Theobalt, theobalt@mpi-inf.mpg.de, Max-Planck-Institut für Informatik, Germany.

0730-0301/2023/12-ART
https://doi.org/10.1145/3618357

1 INTRODUCTION

Inverse rendering is the process of inferring scene properties such as geometry, lighting, and surface materials from images. It is a long-standing challenge with applications in scene understanding, image editing, urban planning, virtual reality, and many more. Typically, inverse-rendering pipelines employ an analysis-by-synthesis approach, in which scene parameters are optimized via a differentiable renderer based on an image reconstruction objective. Unfortunately, determining scene properties from (multi-view) images alone is a severely ill-posed inverse problem since light transport is governed

by the rendering equation [Kajiya 1986], which connects geometry, materials, and illumination via complex integral relationships. This naturally leads to ambiguities: A red pixel might arise from a red material under white illumination, or from a white material under red illumination; a smooth-looking surface might arise from a diffuse material or low-frequency illumination, etc.

Fortunately, real scenes exhibit statistical regularities [Dror et al. 2004]. Exploiting the fact that some scene configurations are more likely than others, most existing solutions incorporate *priors* into the inverse-rendering pipeline. They are meant to encourage plausible solutions. When considering the joint distribution of illumination and materials, applying a prior over illumination conditions naturally narrows down the distribution over materials, and vice versa. Priors vary in their degree of sophistication, ranging from simple heuristics such as smoothness [Zhang et al. 2021b] to the use of data-driven deep generative models [Gardner et al. 2022]. However, the inherent multi-solution ambiguities of the inverse-rendering problem are usually not considered: With a naive prior, the optimization tends to converge to *one* local optimum, which is determined by the weight given to the prior, effectively balancing the trade-off between re-rendering accuracy and the naturalness of the obtained solution [Blau and Michaeli 2018]. Ideally, an inverse-rendering prior should not only *faithfully capture* the distribution of scene properties, such as illumination or material, but also enable the *the generation of diverse solutions from the multi-modal posterior distribution* of scene properties given the input images. Typically, one must choose between achieving high quality or diversity, but a well-designed prior can achieve both simultaneously.

Diffusion models [Sohl-Dickstein et al. 2015] excel at capturing data distributions and have demonstrated the ability to generate realistic and diverse samples. A trained unconditional diffusion model can be utilized to solve non-linear inverse problems [Chung et al. 2022b]. Therefore, in this paper, we investigate the use of diffusion posterior sampling (DPS) as a prior for inverse rendering. To this end, we train a denoising diffusion probabilistic model (DDPM) [Ho et al. 2020] on natural environment maps and then integrate it into an optimization framework involving a differentiable path tracer. Crucially, we extend DPS to incorporate a measurement function with trainable parameters. Our novel optimization scheme allows sampling from combinations of illumination and spatially-varying materials that are *natural*, *diverse*, and *explain the image observations*.

We further conduct a thorough investigation into the effectiveness of various illumination priors for inverse rendering. We demonstrate that our novel formulation not only surpasses the state of the art in its ability to produce highly realistic and diverse environment map samples while fitting the input data in the form of monocular and multi-view images. It also simultaneously facilitates disentanglement and reduces ambiguities in the optimized materials.

In summary, our contributions are:

- A method for recovering the posterior distribution of illumination and materials from image observations under an unknown lighting condition, hence generating samples with both high realism and high diversity.
- A novel strategy combining the denoising process in DDPM and differentiable rendering for the joint optimization of scene illumination and materials.
- A comprehensive study of different priors on illumination used in inverse rendering.

2 RELATED WORK

We review closely related literature on differentiable rendering methods (Sec. 2.1) and popular priors on illumination that are widely used in inverse rendering (Sec. 2.2). Then, we briefly discuss the usage of diffusion generative models for general inverse problems (Sec. 2.3).

2.1 Differentiable and Inverse Rendering

Inverse rendering is commonly performed by means of an analysis-by-synthesis approach: A forward rendering model turns scene parameters such as geometry, illumination, and surface materials into an image, which is compared to existing image observations. For gradient-based optimization of scene parameters to work, the rendering model needs to be (made) differentiable.

Approximate differentiable rendering approaches that rely on mesh rasterization have been proposed for joint optimization of material and illumination. By design, they either ignore global light transport [Kato et al. 2018; Laine et al. 2020; Liu et al. 2019; Loper and Black 2014] or approximate the rendering function of indirect illumination to account for isolated effects like soft shadows [Lyu et al. 2021].

In contrast, more physically-based approaches [Jakob et al. 2022a; Li et al. 2018; Loubet et al. 2019; Nimier-David et al. 2019] account for global illumination by differentiating Monte Carlo-sampled multi-bounce light transport. Special precautions are required to obtain unbiased gradients arising from non-differentiable visibility [Li et al. 2018; Loubet et al. 2019]. A physically-based differentiable path tracer is powerful in inverse rendering with complex global light transport effects [Hasselgren et al. 2022], but can suffer from high-frequency noise in the gradients.

Recent years have witnessed a shift towards neural implicit representations in differentiable and inverse rendering. Continuous volume densities [Mildenhall et al. 2021] and signed distance functions (SDFs) [Park et al. 2019] have many advantages over classical representations, in particular when it comes to differentiability. However, the resulting improved disentanglement of geometry, material, and illumination in the neural rendering process comes at the cost of increased computational complexity. This has spurred research interest in hybrid representations [Munkberg et al. 2022]. In particular, the incorporation of global illumination into the (inverse) rendering process is costly and demands approximations [Srinivasan et al. 2021].

In this work, we build upon the state-of-the-art differentiable path tracer Mitsuba 3 [Jakob et al. 2022b,a], which allows for efficient differentiation of its path-tracing process. Providing a principled and scalable solution for modeling global illumination, we combine it with a strong diffusion-based illumination prior to recover a distribution over spatially-varying materials and illumination.

Compact low-frequency illumination models based on spherical harmonics (SH) [Basri and Jacobs 2003; Ramamoorthi and Hanrahan 2001] are popular in traditional real-time rasterization methods. Spherical Gaussians [Wang et al. 2009] excel in representing a sparse set of high-frequency features and have recently been used for neural inverse rendering [Jin et al. 2023; Wu et al. 2023; Zhang et al. 2021a, 2022]. However, this representation does not scale favorably to complex natural illumination conditions. In this work, we focus on a regularly sampled 2D environment map under equirectangular projection, which matches well with current CNN-based diffusion models.

2.2 Priors for Inverse Rendering

Inverse-rendering problems are notoriously ill-posed, and various priors have been proposed to address this issue. Data-*agnostic* priors include smoothness regularization, as utilized by NeRFactor [Zhang et al. 2021b] and NRTF [Lyu et al. 2022], which mitigates high-frequency noise in estimated environment maps. NvDiffrecMC [Hasselgren et al. 2022] proposes a prior that assumes illumination is mostly monochrome, promoting high-frequency lighting and sharp shadows represented by lighting instead of material textures. The Deep Image Prior (DIP) [Ulyanov et al. 2018] demonstrates the effectiveness of randomly initialized neural networks as handcrafted priors, achieving excellent results in standard inverse problems such as denoising, super-resolution, and inpainting. Data-agnostic priors are effective heuristics, but they may not accurately reflect the true underlying distribution of the data. The performance of these priors is highly dependent on the choice of hyper-parameters, which can be difficult to tune. Additionally, using these priors often leads to optimization algorithms converging to a single local minimum, resulting in a loss of diversity in the final solutions.

On the other hand, data-*driven* priors have the advantage of more accurately capturing the true distribution of real illumination environments. They have the potential to enable sampling from the multi-modal posterior distribution, providing a more diverse set of solutions. Egger et al. [2018] as well as Yu and Smith [2021] utilize a linear statistical model with a Gaussian prior in the space of spherical harmonics (SH) coefficients to avoid unrealistic illumination environments. However, this approach can only reproduce low-frequency lighting effects. Emlight [Zhan et al. 2021] decomposes the illumination map into spherical light distribution, light intensity, and the ambient term for natural illumination regression. Stylelight [Wang et al. 2022] and ImmerseGAN [Dastjerdi et al. 2022] propose GAN panorama generation networks for realistic lighting estimation and editing. Gardner et al. [2022] develop a rotation-equivariant, high-dynamic-range (HDR) neural illumination model based on a variational auto-decoder (VAE) [Rezende and Mohamed 2015] that can express complex features of the natural environment distribution and use it for inverse rendering. VAEs and GANs are powerful techniques for generating diverse samples of illumination that can be used to solve inverse problems, given input images. However, while these methods can provide a range of solutions, the quality of the rendered images and the naturalness of the sampled environment maps are inferior to what our approach delivers.

2.3 Generative Diffusion Models for Inverse Problems

Denoising diffusion probabilistic models (DDPMs) [Ho et al. 2020; Sohl-Dickstein et al. 2015; Song et al. 2020] have recently demonstrated remarkable performance on both faithfulness and diversity of samples in tasks like image synthesis [Dhariwal and Nichol 2021]. Also, a pre-trained, task-agnostic DDPM has been demonstrated to be an excellent prior for imaging inverse problems: Typically, one can resort to iterative projections to the measurement space [Choi et al. 2021; Chung et al. 2022c; Song et al. 2020] or estimate the posterior score function [Chung et al. 2022b] to reach feasible solutions from the implicit prior data distribution. This has been used both in linear [Chung et al. 2022c; Kawar et al. 2022; Song et al. 2020] and non-linear [Chung et al. 2022b; Song et al. 2023] imaging inverse problems. While most of the inverse problem solvers with DDPMs are limited to a known and fixed forward measurement operator, BlindDPS [Chung et al. 2022a] proposes to sample the image and the operator parameters from posterior score functions in parallel, to solve blind inverse problems with unknown forward operators. In contrast to our solution, they require an additional DDPM for the operator.

This work sets out to utilize DDPMs as a prior for natural illumination distribution to address complex non-linear inverse rendering problems. Specifically, we aim to sample diverse but realistic combinations of illumination and materials that can accurately explain the given observation, which poses significant challenges.

3 BACKGROUND

In this section, we recap Denoising Diffusion Probabilistic Models (DDPMs) (Sec. 3.1), followed by an introduction of the diffusion posterior sampling method (Sec. 3.2) for general inverse problems.

3.1 Denoising Diffusion Probabilistic Models

Denoising diffusion probabilistic models (DDPMs) [Ho et al. 2020; Sohl-Dickstein et al. 2015] are a type of generative models that define the generation process as a Markov chain of denoising steps. Formally, given the clean data $\mathbf{x}_0 \sim q(\mathbf{x}_0)$, in the forward steps, DDPMs gradually add noise to the data, which is described as a Gaussian transition:

$$q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) := \mathcal{N}\left(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t \mathbf{I}\right), \tag{1}$$

where $0 < \beta_1 < \beta_2 < .... < \beta_T = 1$ are the fixed variance schedule. The distribution of sampling x_t at the time step t has a closed form

$$q(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}\left(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I}\right), \tag{2}$$

with $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$. This enables efficient training since we can randomly sample a single timestep during training and generate the input $\mathbf{x}_t$. At the largest time step T, the noised data is transformed to a standard Gaussian distribution.

DDPMs learn the reverse Gaussian transition that gradually denoises the noised data with the objective:

$$\mathcal{L}(\theta) := \mathbb{E}_{t,\mathbf{x}_0,\boldsymbol{\epsilon}}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\|^2\right], \tag{3}$$

corresponding to the "simple" loss formulation in Ho et al. [2020], where ϵ_θ denotes the residual noise predicted by a network and θ

are the free variables of the network. The data at the next timestep can now be computed as:

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta\left(\mathbf{x}_t, t\right)\right) + \sigma_t \mathbf{z}, \tag{4}$$

where $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and the covariance σ_t is fixed or learned. At test time, a randomly sampled distribution can be gradually denoised into the learned data manifold.

Song et al. [2020] show that there is a continuous stochastic differential equation (SDE) [Anderson 1982] formulation of the generation process, which is equivalent to Eq. 4:

$$d\boldsymbol{x} = \left[-\frac{\beta(t)}{2}\boldsymbol{x} - \beta(t)\nabla_{\boldsymbol{x}_t}\log p_t\left(\boldsymbol{x}_t\right)\right]dt + \sqrt{\beta(t)}d\overline{\boldsymbol{w}}\,. \tag{5}$$

Thus, sampling from a DDPM can also be seen as numerically solving the SDE. Here, $d\overline{\boldsymbol{w}}$ is the standard Wiener process running backward. The score function $\nabla_{\mathbf{x}_t}\log p_t\left(\mathbf{x}_t\right)$ can be approximated by a network $\boldsymbol{s}_\theta\left(\mathbf{x}_t, t\right)$ using score matching [Hyvärinen and Dayan 2005; Vincent 2011], which is connected with the noise predictor in DDPMs via

$$\boldsymbol{s}_\theta\left(\mathbf{x}_t, t\right) = \frac{\boldsymbol{\epsilon}_\theta\left(\mathbf{x}_t, t\right)}{\sqrt{1-\bar{\alpha}_t}}\,. \tag{6}$$

One can also find the posterior estimation of the clean image $\hat{\mathbf{x}}_t$ at the t step by

$$\hat{\mathbf{x}}_t = \frac{1}{\sqrt{\bar{\alpha}_t}}(\mathbf{x}_t + (1-\bar{\alpha}_t)\boldsymbol{s}_\theta(\mathbf{x}_t, t))\,. \tag{7}$$

For a comprehensive overview, we refer to McAllester [2023].

3.2 Diffusion Posterior Sampling

Chung et al. [2022b] proposed Diffusion Posterior Sampling (DPS) to solve general inverse problems with a pre-trained unconditional DDPM. Consider the forward process

$$\mathbf{y} = \mathcal{M}(\mathbf{x}_0) + \mathbf{n}, \quad \mathbf{y}, \mathbf{n} \in \mathbb{R}^n,\ \mathbf{x} \in \mathbb{R}^d\,, \tag{8}$$

where $\mathcal{M}(\cdot) : \mathbb{R}^d \mapsto \mathbb{R}^n$ is the measurement operator, $\mathbf{x}_0$ is the input data, $\mathbf{y}$ is the measurement observation and $\mathbf{n}$ is measurement noise. In an inverse problem, we attempt to estimate the input image $\mathbf{x}_0$ given the observation $\mathbf{y}$ from the posterior distribution $p(\mathbf{x}_0 \mid \mathbf{y})$.

Chung et al. [2022b] approximate the posterior score function given observation $\mathbf{y}$ as:

$$\nabla_{\mathbf{x}_t}\log p_t(\mathbf{x}_t \mid \mathbf{y}) \simeq \boldsymbol{s}_\theta\left(\mathbf{x}_t, t\right) - \frac{1}{\sigma^2}\nabla_{\mathbf{x}_t}\|\mathbf{y} - \mathcal{M}(\hat{\mathbf{x}}_t)\|_2^2\,. \tag{9}$$

They assume a known and fixed forward operator $\mathcal{M}(\cdot)$ and Gaussian measurement noise $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$. This posterior score function is then used to denoise the image in the reverse generation process:

$$d\boldsymbol{x} = \left[-\frac{\beta(t)}{2}\boldsymbol{x} - \beta(t)\nabla_{\boldsymbol{x}_t}\log p_t\left(\boldsymbol{x}_t \mid \mathbf{y}\right)\right]dt + \sqrt{\beta(t)}d\overline{\boldsymbol{w}}\,. \tag{10}$$

The posterior score function in Eq. 9 aims to generate natural outputs while guiding them to fit the observation at the same time. However, DPS requires the operator $\mathcal{M}(\cdot)$ to be known and fixed, which is not the case for our method, where we seek to jointly optimize illumination and material parameters from images. Here, the operator is the rendering function, which naturally encompasses both illumination and materials. We will present a method that enables jointly optimizing for the output of the diffusion model i.e., scene illumination, as well as the parameters of the measurement operator i.e., scene materials.

4 METHOD

Our method takes monocular or multi-view images of an object with a well-initialized geometry under one unknown illumination as input and allows realistic environment map sampling and joint optimization of material that fits the observation. To this end, our method not only considers the faithfulness of the reconstructed environment maps, but also explores the ambiguity of the material–light decomposition and the diversity of the solutions.

We use a differentiable renderer to guide a pre-trained unconditional diffusion model to sample multiple plausible natural environment maps that can explain the ground-truth observations. At the same time, we optimize the spatially-varying BRDFs of the scene for each illumination sample. In the following sections, we first describe the rendering equation, our material model, and our illumination representation (Sec. 4.1). Next, we discuss the challenges to extend diffusion posterior sampling (DPS) with a differentiable renderer for inverse rendering problems (Sec. 4.2), followed by an explanation of the process used to jointly optimize the spatially-varying BRDF and sample plausible environments (Sec. 4.3). We give an overview of our pipeline in Fig. 2.

4.1 Rendering Model

We are interested in estimating a natural environment map $\mathbf{x}_0$ and spatially-varying surface materials $\mathbf{k}_{\boldsymbol{arm}}$ explaining input image observations using an analysis-by-synthesis approach. Global light transport [Kajiya 1986] under distant illumination can be formulated as

$$L\left(\mathbf{p}, \boldsymbol{\omega}_o\right) = \int_{\Omega_+} \mathbf{x}_0(\boldsymbol{\omega}_i)\mathbf{T}(\mathbf{p}, \boldsymbol{\omega}_o, \boldsymbol{\omega}_i, \mathbf{k}_{\boldsymbol{arm}})\left(\boldsymbol{\omega}_i \cdot \mathbf{n}\right)d\boldsymbol{\omega}_i, \tag{11}$$

where $L\left(\mathbf{p}, \boldsymbol{\omega}_o\right)$ is radiance leaving surface point $\mathbf{p}$ in direction $\boldsymbol{\omega}_o$, Ω_+ is the hemisphere centered at the surface normal $\mathbf{n}$, and $\mathbf{T}(\mathbf{p}, \boldsymbol{\omega}_o, \boldsymbol{\omega}_i, \mathbf{k}_{\boldsymbol{arm}})$ is the radiance transfer function, describing how distant illumination from direction $\boldsymbol{\omega}_i$ is scattered through the scene (via potentially many bounces) to finally leave at $\mathbf{p}$ in direction $\boldsymbol{\omega}_o$. We implement T using path tracing [Kajiya 1986], which stochastically samples light paths to obtain a Monte Carlo estimate of the global illumination, including potentially complex inter-reflections.

We utilize the state-of-the-art differentiable path tracer Mitsuba3 [Jakob et al. 2022b] with path replay backpropagation [Vicini et al. 2021]. As our research focuses on investigating the ambiguity involved in disentangling material and illumination, we assume that the scene geometry is available. To represent the material properties, we adopt the principled BSDF in Mitsuba3 [Jakob et al. 2022b] based on the Disney BSDF [Burley 2015; Burley and Studios 2012]. Our approach utilizes a 256x256 texture $\mathbf{k}_{\boldsymbol{arm}}$ to model RGB albedo $\boldsymbol{a}$, roughness $\boldsymbol{r}$, and metallic properties $\boldsymbol{m}$. We use four light bounces in all our experiments, which we found enough to capture all relevant light transport in the scenes we consider, faithfully modeling global-illumination effects.

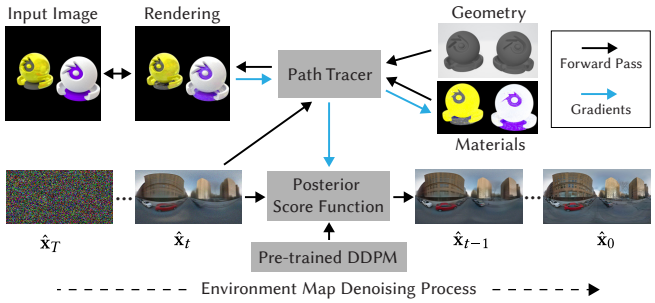

Fig. 2. Overview of our approach. We first pre-train a DDPM that generates realistic environment maps unconditionally. Then, given input images and geometry, we set up a series of denoising processes. In every time step t, our differentiable path tracer takes materials and the posterior estimation of the clean environment map $\hat{x}_t$ as input and produces a rendered image. The gradient from the rendering loss is used to optimize materials and gets incorporated into a posterior score function that enforces the DDPM to generate a natural environment map that faithfully explains the input images.

4.2 DPS for Inverse Rendering

For our application, as the measurement operator $\mathcal{R}(\cdot)$, we use a differentiable path tracer with additional material and camera pose input. Namely, the forward measurement process is described as

$$\mathbf{y} = \mathcal{R}(\mathbf{x}_0, \mathbf{k}_{\boldsymbol{arm}}, \boldsymbol{c}) + \mathbf{n}_r, \quad \mathbf{y}, \mathbf{n}_r \in \mathbb{R}^n, \ \mathbf{x} \in \mathbb{R}^d. \tag{12}$$

This corresponds to the rendering equation (Eq. 11), where $\mathbf{y}$ is the rendered image, $\mathcal{R}(\cdot) : \mathbb{R}^d \mapsto \mathbb{R}^n$ is the rendering function, $\boldsymbol{c}$ are camera parameters, and $\mathbf{n}_r$ is noise caused by Monte Carlo sampling.

Note that the original DPS formulation requires a known and fixed forward operator $\mathcal{M}(\cdot)$ (Eq. 8) to approximate the posterior score function $\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t|\mathbf{y})$, which is not applicable in an inverse rendering scenario where the material properties are generally unknown. In our case, the posterior score function is rewritten as

$$\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t|\mathbf{y}) \simeq \boldsymbol{s}_\theta(\mathbf{x}_t, t) - \rho \nabla_{\mathbf{x}_t} \|\hat{\mathbf{y}}_i - \mathcal{R}(\hat{\mathbf{x}}_t, \mathbf{k}_{\boldsymbol{arm}}, \boldsymbol{c}_i)\|_2^2. \tag{13}$$

Here, $\hat{\mathbf{x}}_t$ is the posterior estimation of the environment map image from Eq. 7, and ρ is a weight hyper-parameter. Ideally, the sampled environment map $\hat{\mathbf{x}}_t$ should minimize the re-render loss

$$\|\hat{\mathbf{y}}_i - \mathcal{R}(\hat{\mathbf{x}}_t, \mathbf{k}_{\boldsymbol{arm}}, \boldsymbol{c}_i)\|_2^2, \tag{14}$$

and we opt to concurrently optimize the material $\mathbf{k}_{\boldsymbol{arm}}$ with the denoising process for the environment map. To achieve this objective, we propose a novel approach that enables simultaneous sampling of the environment map image and optimization of the material properties in Sec. 4.3.

DPS provides measurements (Eq. 8) that exhibit various types of noise, which correspond to different closed forms of the posterior score function (Eq. 9). As reported by Lehtinen et al. [2018], the use of Monte Carlo sampling in the ray tracer can result in random noise with no specific distribution characteristics. Despite this, we found that applying the posterior score function for Gaussian noise produces satisfactory results for the inverse rendering problem.

4.3 Joint Optimization

Given the posed input image(s) and pre-computed geometry, we first jointly optimize the material and the environment map using Mitsuba3 [Jakob et al. 2022b] with the objective

$$\mathcal{L}_{\text{PT}}(\mathbf{x}, \mathbf{k}_{\boldsymbol{arm}}) = \sum_{i=1}^{m} ||\hat{\mathbf{y}}_i - \mathcal{R}(\mathbf{x}, \mathbf{k}_{\boldsymbol{arm}}, \boldsymbol{c}_i)||^2, \tag{15}$$

as the initialization, where $\mathbf{x}$ is the reconstructed environment map by Mitsuba3 [Jakob et al. 2022b], $\hat{\mathbf{y}}_i$ is an input image and $\boldsymbol{c}_i$ is the corresponding camera pose. During this initialization stage, we use the Adam [Kingma and Ba 2015] optimizer with a learning rate of $1e^{-2}$ for metallic and roughness parameters, the environment map, and the base color.

Then we keep the optimized roughness and metallic texture map, but reset the base color to zero as our initialization condition for the joint optimization. We find that the initialization of roughness and metallic helps the material optimization converge faster. Next, we gradually denoise the environment map images that fit the observation from our DDPM prior and optimize our estimated material at the same time. We use a DDPM posterior sampling process (Eq. 10) for 1000 time steps, with the posterior score function in Eq. 13.

The hyper-parameter ρ in Eq. 13 balances the faithfulness of the generation and the quality of reconstruction. A too small ρ encourages the posterior sampling to be as realistic as possible, disregarding the input observations, while a too large ρ tends to produce results faithful to the observations, yet with inferior realism. In our setting, for the first 500 denoising steps ($t > 500$), we set $\rho = 0.1$. For time steps $t < 500$, $\rho = 0.1 \sim 1$, and in every denoising step, we optimize the base color, roughness, and metallic jointly with the loss function

$$\mathcal{L}_{\text{denoise}}(\mathbf{k}_{\boldsymbol{arm}}, t) = \sum_{i=1}^{m} ||\hat{\mathbf{y}}_i - \mathcal{R}(\hat{\mathbf{x}}_t, \mathbf{k}_{\boldsymbol{arm}}, \boldsymbol{c}_i)||^2. \tag{16}$$

The intuition behind the adjustment of the step size ρ in Eq. 13 is that at the early stage of the generation, we encourage the posterior score function to push the sampling distribution towards a more realistic manifold. And when the DDPM adds more details to the images, we raise the step size ρ to force the sampling to fit the observation. At the late stage of the generation where the noise in the sampling is gradually reduced, the estimated environment map from a realistic prior will prevent the optimization for the material from getting stuck in a local minimum.

Finally, after generating a clean and realistic environment map $\mathbf{x}_0$, the material parameters will be refined by

$$\mathcal{L}_{\text{refine}}(\mathbf{k}_{\boldsymbol{arm}}) = \sum_{i=1}^{m} ||\hat{\mathbf{y}}_i - \mathcal{R}(\mathbf{x}_0, \mathbf{k}_{\boldsymbol{arm}}, \boldsymbol{c}_i)||^2. \tag{17}$$

for 200 additional iterations.

4.4 Seamless Environment Map Generation

We aim to develop a generative model capable of sampling seamlessness spherical environment maps $\mathbf{x}_0$. This entails constructing a model that effectively captures the distribution of the input data while maintaining seamless panoramas under continuous rotations.

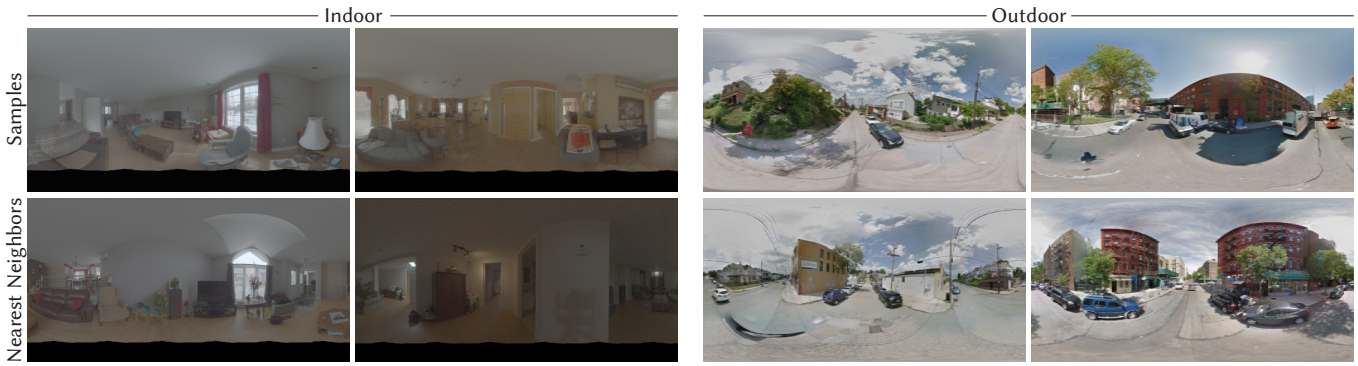

Fig. 3. Unconditional samples from our generative models (top row) alongside their nearest neighbors in the training dataset (bottom row).

To achieve this, we employ two techniques. Firstly, we apply data augmentation during training. Specifically, when training our DDPM using real-world environment map datasets, we introduce horizontal rotations to each sample, randomizing the rotation angle. By incorporating this data augmentation strategy, our training objective becomes

$$\mathcal{L}(\theta) := \mathbb{E}_{t,\mathbf{x}_0,\boldsymbol{\epsilon},\phi}\left[\left\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta\left(\sqrt{\bar{\alpha}_t}\phi(\mathbf{x}_0) + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t\right)\right\|^2\right], \quad (18)$$

where ϕ is the random horizontal rotation operator. This augmentation strategy introduces variations in the rotation angle, effectively reducing potential bias in the input data distribution where the environment maps are predominantly positioned at specific rotation angles. Next, during the denoising generation steps, we employ a similar rotation scheme that replaces the pre-trained unconditional score function (Eq. 6) as:

$$\boldsymbol{s}_\theta^{'}(\mathbf{x}_t, t) = \phi_t^{-1}\left(\boldsymbol{s}_\theta(\phi_t(\mathbf{x}_t), t)\right), \quad (19)$$

where ϕ_t is the random horizontal rotation operator at time step t and ϕ_t^{-1} is its reverse. During the denoising process, the score function network $\boldsymbol{s}_\theta$ takes the rotated noised images as input and preserves continuity and meaningful information around the stitched seam, resulting in panoramas without noticeable disruptions or artifacts at the boundaries.

5 EXPERIMENTS

Here, we report results of the experiments we conducted to evaluate our method. We first provide details on our implementation and training procedure (Sec. 5.1). Then we evaluate the quality of our generative model (Sec. 5.2), before providing an in-depth evaluation of the inverse-rendering capabilities of our method (Sec. 5.3 and Sec. 5.4). We provide further analyses (Sec. 5.5) and conclude with a short section on additional applications (Sec. 5.6).

5.1 Implementation and Training Details

Since indoor and outdoor illumination conditions exhibit markedly different characteristics, we employ two separate DDPM models. For the indoor model, we train on the Laval indoor dataset [Gardner et al. 2017], which includes 2.2k high dynamic range (HDR) environment maps, while for the outdoor model, we use the Streetlearn dataset [Mirowski et al. 2019], containing 143k low dynamic range (LDR) environment maps. We apply an HDR reconstruction pre-process on the outdoor dataset using the method of Santos et al. [2020]. We perform data augmentation by applying random horizontal shifts (corresponding to horizontal rotations of the scene-enclosing sphere) and horizontal flips. For stable training, we compress the HDR content using the global invertible mapping $f(\mathbf{x}) = 0.5\mathbf{x}^{\frac{1}{2.4}}$ for outdoor data and $0.9\mathbf{x}^{\frac{1}{6}}$ for indoor. Before using our samples for path tracing, we apply f^{-1} to recover HDR illumination. We use the network architecture proposed by Nichol and Dhariwal [2021] without modifications. Training one model takes 1 week using four A100 GPUs. Obtaining a single sample from our joint optimization procedure takes 20 minutes, including 1000 iterations of interleaved environment map denoising (290 ms) and material optimization (630 ms). We provide all source code and pre-trained models on https://vcai.mpi-inf.mpg.de/projects/2023-DPE/.

5.2 Quality of the Generative Models

Table 1. Quality (FID↓) of generative models for unconditional sampling.

Model	GAN	RENI	**Ours**
Indoor	45.5	288.2	**12.3**
Outdoor	15.8	323.7	**7.6**

Before we investigate the inverse-rendering capabilities of our method, we seek to gain insights into the quality of our trained generative models. To this end, we consider unconditional sampling. Tab. 1 provides a numerical evaluation using FID [Heusel et al. 2017] scores. In Fig. 3 we show samples of our models. To make sure that our networks do not simply memorize the training data, we also display the respective nearest neighbors in the training dataset. Distances between samples are computed using the LPIPS [Zhang et al. 2018] metric, while also taking into account the augmentations (rotations, flips) performed during training. We see that our samples are markedly different from the closest examples in the training data corpus. We conclude that our models produce high-quality samples, which look realistic and capture the data distribution well.

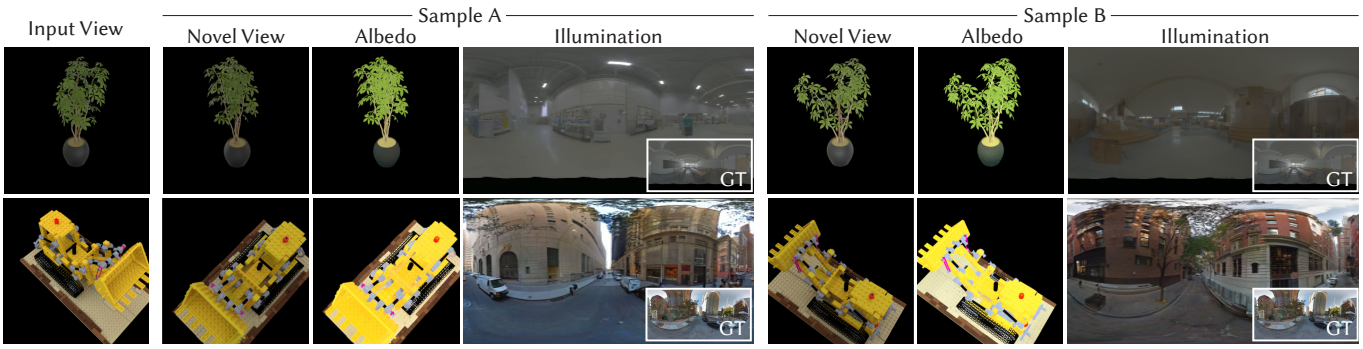

Fig. 4. Qualitative results on synthetic datasets. Here we show two samples per scene, demonstrating indoor (top) and outdoor (bottom) illumination. Notice that the recovered illumination samples are markedly dissimilar in the details, but share the overall illumination structure, faithfully capturing the ambiguity of the inverse problem. Insets show the ground-truth environment maps.

5.3 Inverse Rendering

We now turn to our core task: The ambiguity-aware decomposition of a scene into illumination and material properties, based on one or multiple images. We assume we have access to or can reconstruct the scene geometry before running our method. This step is orthogonal to our approach, and we give details on how geometry is obtained for each experimental setup.

5.3.1 Quantitative Evaluation. We start by considering a set of synthetic scenes, where ground-truth geometry, illumination, and materials are available for evaluation. Specifically, we base our analysis on nine scenes, including six scenes from the NeRF [Mildenhall et al. 2021] dataset with ten illumination conditions per scene (five indoor and five outdoor), all distinct from the training data corpus. The scenes exhibit a broad spectrum of material types. Inputs to our method are 16 $\sim$ 50 multi-view images of resolution 800x600 pixels.

Metrics. To quantitatively evaluate the obtained solutions, we are interested in several properties, listed as columns in Tab. 2 and Tab. 3: First, we consider three different aspects concerning the reconstructed *materials*. We measure albedo reconstruction accuracy using the mean squared error (MSE). We also show how much albedo reconstruction varies when running the method multiple times with exactly the same input by computing diversity in the form of variance σ^2_{sample} across five runs. Further, we evaluate how far the reconstruction of albedo is invariant under different illumination conditions, computed as the variance σ^2_{invar}.

Second, we analyze the reconstructed *environment maps*. We start with the common full-reference quality metrics PSNR, SSIM [Wang et al. 2004], and LPIPS [Zhang et al. 2018] to gain an understanding of how close the reconstructions are to the ground-truth illumination. We are further interested in measuring the naturalness of the obtained illumination conditions. To this end, we employ the FID [Heusel et al. 2017] score, comparing the conditional distribution of reconstructed environment maps given the input images to the unconditional distribution of environment maps in the training data corpus. The distributions are clearly dissimilar by construction, and the number of samples is low for our reconstructions, naturally resulting in relatively high scores, but we found the FID to nevertheless correlate well with the perceived naturalness of the solutions. As an additional indicator of naturalness we compute the non-reference HDR metric PU21-NIQE [Hanji et al. 2022; Mittal et al. 2013]. Analogous to the albedo analysis, we consider the diversity of the illumination estimates by computing variance σ^2_{sample} across different samples.

Finally, we show full-reference quality metrics for the *novel-view synthesis* task on held-out input views. This helps us understand how accurately our decompositions explain the scene's appearance.

Notice that the true values of σ^2_{sample} (both for material and illumination) and σ^2_{invar} should reflect the variance of the true posterior distribution, which is unknown. A reasonably low value for σ^2_{invar} indicates stability, a reasonably high value for σ^2_{sample} indicates diversity. However, blindly aiming for a minimal σ^2_{invar} or maximal σ^2_{sample} is not desirable. As a counter-example, consider a deterministic method that always converges to the same albedo estimation and therefore has $\sigma^2_{\text{invar}} = 0$, which is clearly a sub-optimal solution.

Baselines. We compare our diffusion-based prior on illumination to a broad variety of alternatives, including data-agnostic and data-driven priors. Whenever necessary, we set weighting factors balancing reconstruction and regularization such that both losses have the same magnitude. We consider the following priors:

- `No Prior`: As a baseline, we run path tracing-based inverse rendering without any prior at all.
- `Global Norm.`: To address the inherent *global* ambiguities of illumination–reflectance decomposition, we proceed as in the previous case, but additionally scale the estimated environment maps to enforce their median to match the median of our dataset.
- `Smoothness`: A very common prior used in many inverse problems is the assumption that obtained solutions should be smooth [Lyu et al. 2022; Zhang et al. 2021b]. We adopt the 2D total variation penalty

$$\mathcal{L}_{\text{smooth}} = \sum_{\mathbf{u}} \|\nabla_{\mathbf{u}} \mathbf{x}_0(\mathbf{u})\|_1 ,$$

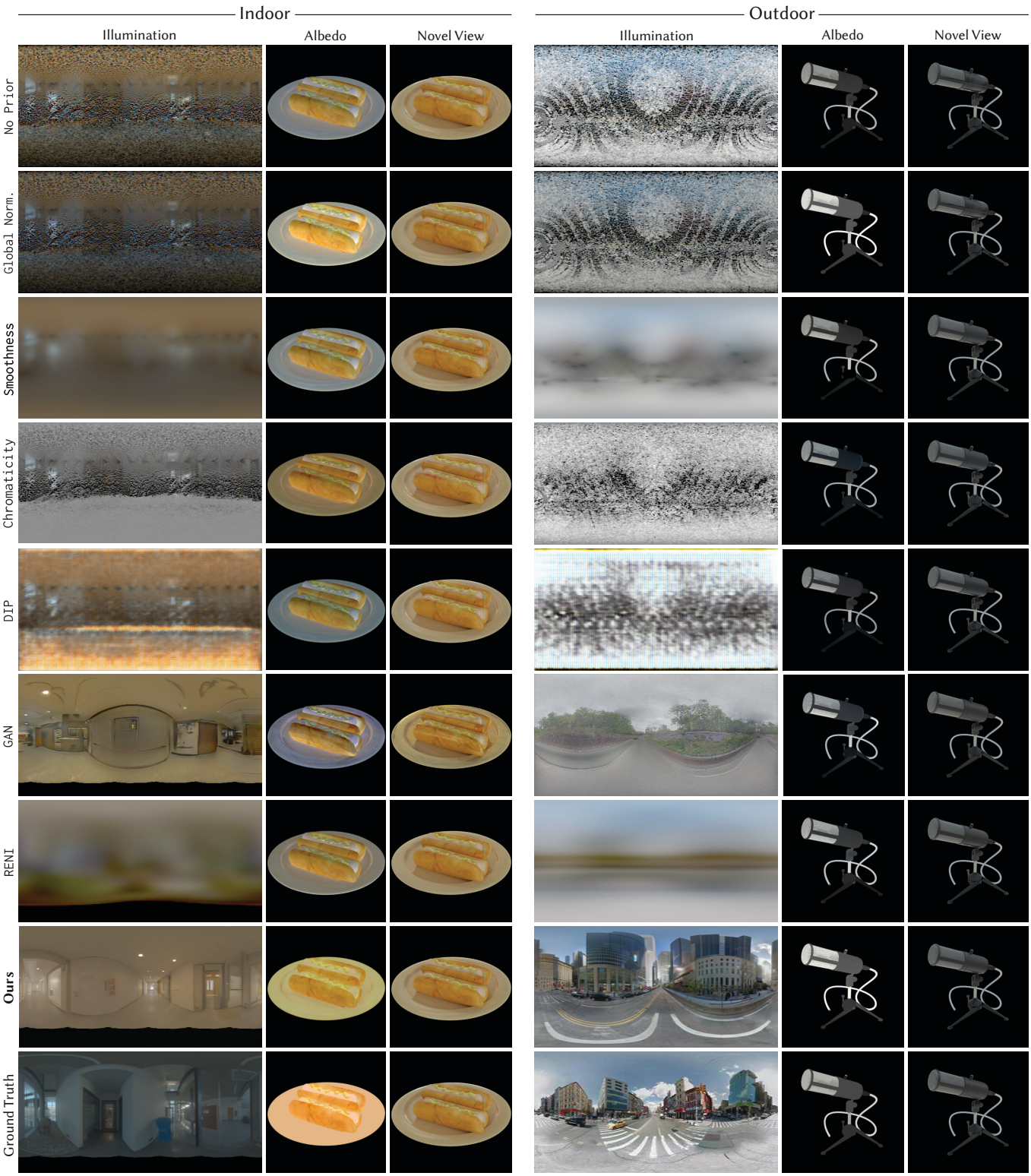

Fig. 5. Scene decompositions using different priors for two synthetic scenes with indoor (left) and outdoor (right) illumination.

Table 2. Quantitative evaluation of **indoor** illumination. For each prior/method, we present statistics on reconstructed *Albedo*, *Illumination*, and novel *View Synthesis*. Our method outperforms all baselines by estimating more accurate albedos and illumination, and consequently synthesizing more accurate novel views. The measures σ^2_{invar} and σ^2_{sample} give an indication of stability and diversity, respectively, but their ground-truth values are unknown.

Prior/Method	Albedo			Illumination						View Synthesis		
	MSE↓	σ^2_{sample}	σ^2_{invar}	PSNR↑	SSIM↑	LPIPS↓	FID↓	NIQE↓	σ^2_{sample}	PSNR↑	SSIM↑	LPIPS↓
`No Prior`	0.086	0.0007	0.030	16.3	0.14	0.89	329.6	10.5	0.0017	30.5	0.947	0.027
`Global Norm.`	0.078	0.0006	0.025	20.8	0.38	0.64	304.3	9.4	0.0001	30.5	0.947	0.027
`Smoothness`	0.081	0.0006	0.028	18.8	0.61	0.46	341.3	9.26	0.0001	33.8	0.964	0.024
`Chromaticity`	0.085	0.0002	0.029	18.2	0.41	0.63	344.9	10.5	0.0001	32.0	0.948	0.025
`DIP`	0.123	0.0057	0.038	11.0	0.26	0.56	301.8	43.1	0.0052	32.3	0.943	0.029
`GAN`	0.106	0.0069	0.055	16.9	0.59	0.35	187.1	6.98	0.0072	29.4	0.944	0.029
`RENI`	0.091	0.0215	0.043	15.9	0.49	0.54	371.8	17.0	0.0225	27.8	0.913	0.037
NeRFactor	0.084	0.0011	0.032	7.3	0.31	0.69	453.7	18.9	0.0028	26.4	0.918	0.051
NvDiffRec	0.075	0.0001	0.038	8.5	0.17	0.65	468.3	10.6	0.0049	29.3	0.940	0.049
w/ GT geom.	0.063	0.0001	0.030	10.2	0.20	0.61	415.6	6.84	0.0077	32.7	0.952	0.037
NvDiffRecMC	0.103	0.0001	0.076	14.8	0.27	0.55	322.2	9.98	0.0002	28.1	0.934	0.057
w/ GT geom.	0.082	0.0001	0.049	15.4	0.34	0.62	307.8	6.61	0.0004	29.0	0.939	0.046
Ours	**0.033**	0.0082	0.021	**21.9**	**0.67**	**0.26**	**135.6**	**5.14**	0.0096	**34.0**	**0.967**	**0.022**

Table 3. Quantitative evaluation of **outdoor** illumination. Refer to the caption of Tab. 2 for more details.

Prior/Method	Albedo			Illumination						View Synthesis		
	MSE↓	σ^2_{sample}	σ^2_{invar}	PSNR↑	SSIM↑	LPIPS↓	FID↓	NIQE↓	σ^2_{sample}	PSNR↑	SSIM↑	LPIPS↓
`No Prior`	0.087	0.0010	0.031	13.0	0.12	0.87	379.6	8.88	0.0030	31.6	0.939	0.029
`Global Norm.`	0.080	0.0007	0.027	15.1	0.16	0.83	380.5	8.63	0.0005	31.6	0.939	0.029
`Smoothness`	0.084	0.0005	0.031	14.9	0.49	0.64	376.3	7.25	0.0001	33.1	0.960	0.025
`Chromaticity`	0.088	0.0001	0.032	13.5	0.15	0.81	398.7	8.83	0.0001	31.5	0.931	0.029
`DIP`	0.142	0.0014	0.054	9.7	0.19	0.63	371.5	106.24	0.0071	32.1	0.927	0.037
`GAN`	0.078	0.0051	0.094	15.4	0.41	0.50	163.3	6.26	0.0086	29.7	0.936	0.031
`RENI`	0.108	0.0136	0.090	11.3	0.36	0.76	487.9	27.8	0.0232	30.6	0.931	0.044
NeRFactor	0.103	0.0009	0.035	6.4	0.28	0.86	477.6	21.2	0.0024	25.7	0.912	0.056
NvDiffRec	0.088	0.0001	0.037	8.4	0.09	0.70	448.2	9.67	0.0055	31.0	0.938	0.039
w/ GT geom.	0.049	0.0002	0.032	9.5	0.07	0.63	392.2	6.42	0.0069	32.3	0.950	0.037
NvDiffRecMC	0.185	0.0001	0.082	11.1	0.13	0.68	390.5	8.73	0.0002	29.8	0.926	0.039
w/ GT geom.	0.109	0.0001	0.044	13.6	0.22	0.69	350.4	6.10	0.0002	30.1	0.941	0.038
Ours	**0.031**	0.0069	0.016	**20.0**	**0.58**	**0.30**	**109.1**	**4.36**	0.0108	**34.3**	**0.968**	**0.022**

where $\mathbf{u}$ denotes the directional coordinates in the environment map.

- `Chromaticity`: Observing that illumination is frequently not very colorful, we follow Munkberg et al. [2022] and use

$$\mathcal{L}_{\text{chroma}} = \left\| \sum_{c=1}^{3} \left(\mathbf{x}_0^c - \frac{1}{3} \sum_{c=1}^{3} \mathbf{x}_0^c \right) \right\|,$$

  where $\mathbf{x}_0^c$ denotes the c's color channel of $\mathbf{x}_0$.
- `DIP`: The Deep Image Prior [Ulyanov et al. 2018] represents a signal by optimizing the weights of a CNN with random inputs. This enforces a structural prior, empirically favoring solutions that have natural-image statistics.
- `GAN`: We train a StyleGAN2-ADA [Karras et al. 2020] model on our datasets and optimize over the model's latent space using Pivotal Tuning Inversion [Roich et al. 2022]. As for our method, we train two separate models for indoor and outdoor scenes.
- `RENI`: The recent method of Gardner et al. [2022] uses a rotation-equivariant neural field to train a generative model of environment maps. We follow their method to optimize the latent code of this model. Also for this prior, we train two separate models.

The quality of unconditional samples of the `GAN` and `RENI` models is evaluated in Tab. 1.

In addition to the priors discussed above, we also compare our method to three state-of-the-art full-blown inverse-rendering systems. First, we consider NvDiffRec [Munkberg et al. 2022] and NvDiffRecMC [Hasselgren et al. 2022]. In addition to the recovery of illumination and reflectance, both methods also optimize for geometry. For a fair differential comparison, we consider two variants of their methods: The full version, and replacing their geometry estimation with fixed ground-truth geometry. In addition, we include a comparison with NeRFactor [Zhang et al. 2021b] as a representative of inverse rendering using a volumetric representation.

Results. We list numerical results of our experiments in Tab. 2 for indoor and Tab. 3 for outdoor illumination conditions. Fig. 4 displays samples obtained from our method. A qualitative comparison against the baseline priors in our setup is provided in Fig. 5, while Fig. 6 illustrates comparisons against the inverse-rendering systems. Our supplemental video shows additional qualitative results and comparisons. We provide runtime evaluations in the supplemental document.

We observe that our method excels in reconstructing environment maps and materials, where the latter tend to be reasonably stable across different illumination conditions. More importantly, though, our environment maps are significantly more *natural* than all baselines and achieve high *diversity* scores, while at the same time resulting in the highest novel-view synthesis *quality*. Fig. 7 illustrates the distribution of methods according to these three core criteria. While the `GAN` prior achieves competitive scores in terms of diversity and naturalness, it fails to properly explain the input data. On the other end of the spectrum, the `Smoothness` prior achieves high-quality novel-view synthesis results (explaining its popularity in the inverse-rendering community), but always converges to the same solution while producing rather unnatural results. NvDiffRec achieves competitive results in terms of novel-view synthesis quality and diversity, but also suffers from rather unrealistic environment maps.

5.4 Qualitative Evaluation

To evaluate our method on real scenes, we consider a set of scenes from the MipNeRF-360 dataset [Barron et al. 2022], using 100 images per scene. We first run NeuS [Wang et al. 2021] on the multi-view images to obtain geometry of the foreground objects as well as a foreground–background separation, and automatically generate the texture map for the mesh extracted from the SDF from Blender's "Smart UV Project" operator [Community 2018]. We then run our optimization, while masking out the background in the input images. Fig. 8 shows the results of this experiment, which illustrate that we can sample high-quality decompositions. The supplemental document and video provide more results.

5.5 Analysis

Here, we investigate several aspects of our method in more detail.

First, we consider illumination reconstruction quality and diversity as a function of material roughness. Intuitively, as the roughness of scene materials is increased towards diffuse appearance, we expect an increasing amount of ambiguity during illumination

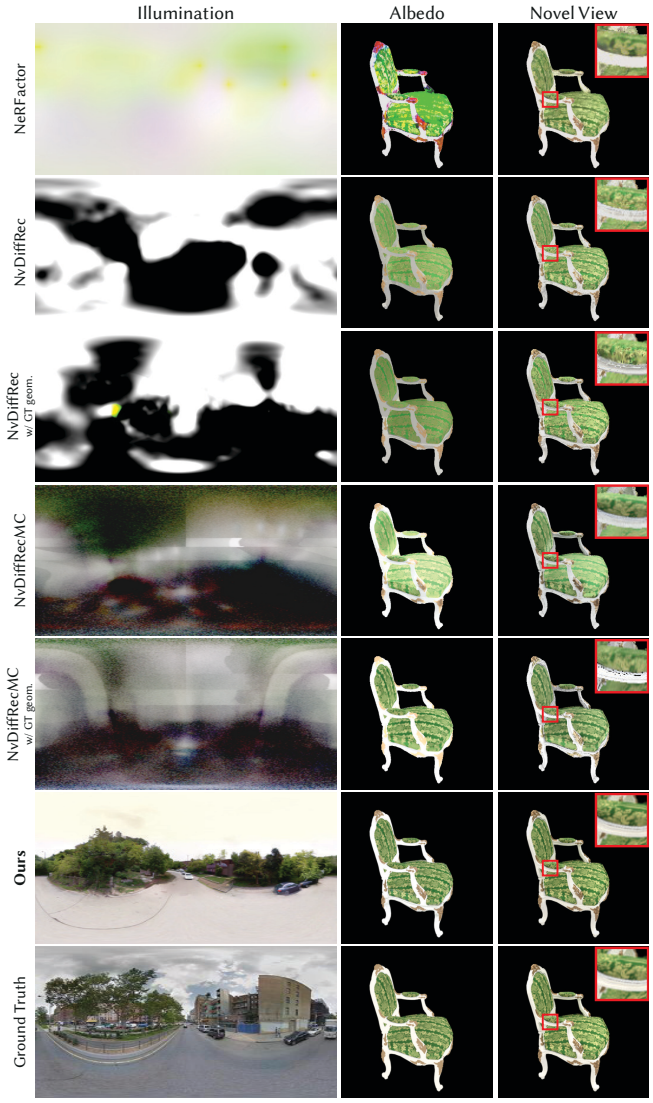

Fig. 6. Comparison with baseline differentiable rendering methods.

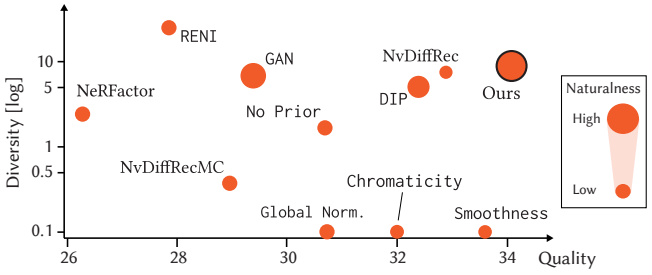

Fig. 7. Overview of different methods in terms of novel-view synthesis quality (PSNR; x-axis), environment map diversity (σ^2_{sample}; y-axis, log-scale), and environment map naturalness (FID; point size). Only our solution achieves high scores across all three aspects.

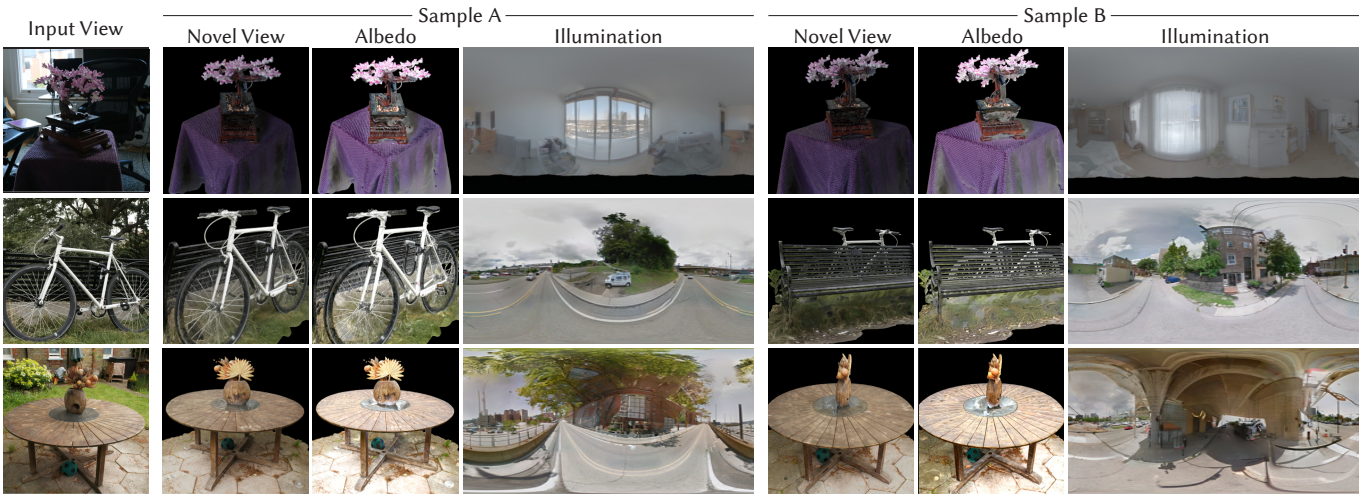

Fig. 8. Qualitative results on real-world datasets. Here we show two samples per scene. Notice that our optimization does only have access to the foreground object, but still manages to synthesize environment maps that match the overall color distribution of the background, while exhibiting a high level of realism.

reconstruction, as the environment map is convolved with a progressively larger BRDF lobe. To systematically investigate the behavior of our method in this regard, we set up a synthetic scene containing a Stanford Dragon to be rendered from three views. We then run a sequence of experiments, where we perform a sweep over the roughness parameter of the dragon's material, with five illumination conditions per roughness setting. In Fig. 9 we investigate the behavior of our method and compare it against the GAN prior. In Fig. 9a we observe a consistently high-quality reconstruction result across all roughness levels, while the GAN prior exhibits overall inferior results, struggling in particular with highly specular materials. Fig. 9b reveals that, in contrast to the GAN prior, the diversity of our samples strongly correlates with roughness, satisfying our expectations. For example, diversity is almost zero when the material is purely specular, which is expected behavior since the input images essentially contain a distorted reflection of the environment map, eliminating uncertainty. In Fig. 10 we provide a corresponding qualitative study.

In Fig. 11 we study the behavior of our method in the presence of highly specular materials: We mask out regions in the ground-truth environment map (top left in Fig. 11) and record which pixels in the image observations are affected by this mask (bottom left in Fig. 11). We then run our method on the observations, but exclude the masked-out pixels from receiving path-tracer gradients. We see that the samples from our method reconstruct the ground-truth environment map quite faithfully while hallucinating plausible content in the masked regions (right in Fig. 11). We further visualize the amount of information that can be extracted from highly specular materials in the supplemental document.

We perform an ablation analysis of our seamless environment map generation process (Sec. 4.4) in Fig. 12. We see that our method is effective in removing seams in the samples.

We are further interested in validating the choice of DPS as our solver. In Fig. 13 we compare our approach against a best-effort result based on score distillation sampling (SDS) [Poole et al. 2022].

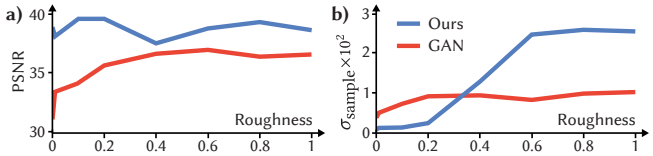

Fig. 9. View synthesis quality (a) and environment map diversity (b) as a function of material roughness. Our approach consistently outperforms the GAN prior in terms of quality and exhibits diversity only for ambiguous scenes.

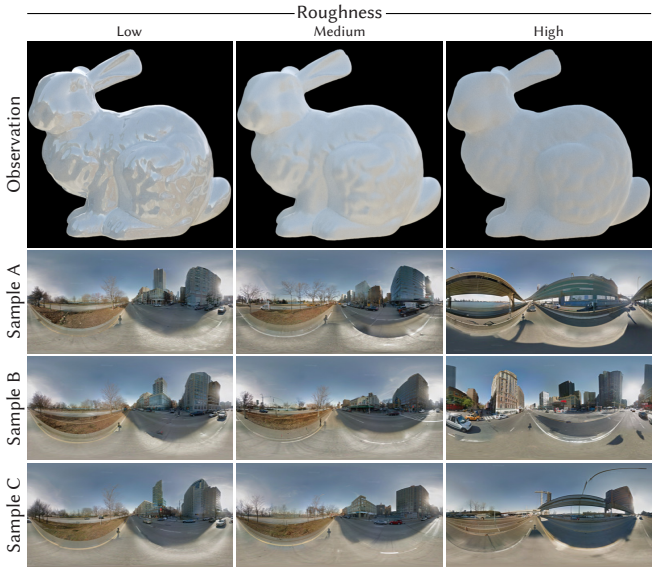

Fig. 10. Qualitative results with three levels of material roughness. We show three illumination samples from our method for each roughness level. While all samples are very similar for the highly specular material, their diversity increases as the material approaches diffuse appearance.

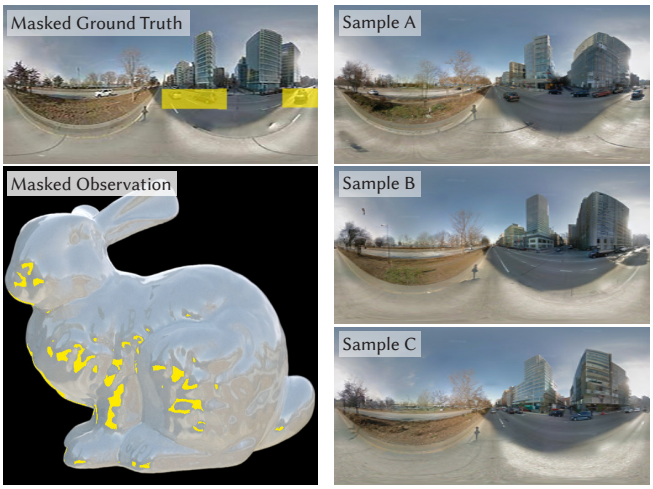

Fig. 11. Environment map sampling from masked observations on specular materials. We mask out cars in the ground-truth environment map (top left) and mark all corresponding pixels in the observations (bottom left). During training, these pixels are ignored. Our samples (right) match the observable areas while generating distinct details in the unseen parts.

Fig. 12. Ablation: Seamless environment map generation.

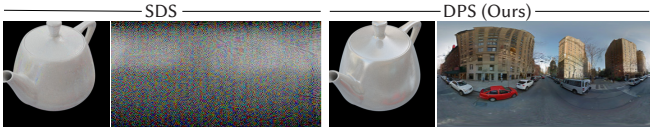

Fig. 13. Ablation: Training with a score distillation sampling (SDS) loss instead of our DPS-based formulation. For each configuration, we show a reconstructed view (left) alongside an environment map sample (right).

In each time step, we add a loss term that moves the current state of the optimized environment map closer to higher-density regions learned by the pre-trained diffusion model. We first add a random amount of noise to the current estimate of the environment map. The loss function then computes the difference between the estimated and ground truth noise. As proposed by Poole et al. [2022], we do not backpropagate through the diffusion model itself. We observe that the SDS results are consistently inferior to the ones from our DPS-based solution, failing to produce realistic samples.

5.6 Applications

Our scene decompositions can be used for a variety of applications. In Fig. 1 we show an example of material editing and relighting. In this case, geometry was reconstructed using multiple views, but input to our system is a single view of the almost diffuse object, resulting in substantial illumination ambiguity. We see that changing the original material to a translucent one produces a plausible appearance, while the `No Prior` or `Smoothness` solutions result in noise and loss of overall structure, respectively.

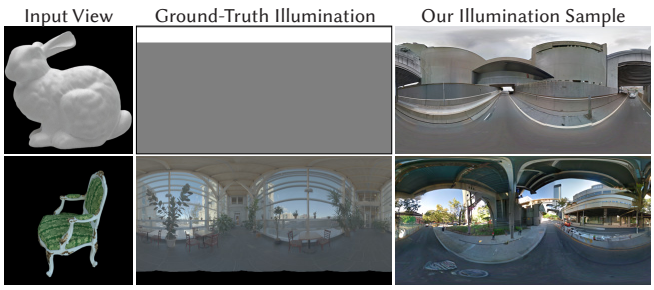

Fig. 14. Limitation: If the illumination of the scene does not follow the data distribution used to train our diffusion model, our samples can be far from the ground truth, while still explaining the input observations.

6 DISCUSSION AND CONCLUSION

We have presented a novel approach that addresses the ambiguity problem in inverse rendering. At the core of our method is a denoising diffusion probabilistic model trained on natural environment maps. On the technical level, we have extended diffusion posterior sampling towards an optimizable measurement operator, which allows for the simultaneous reconstruction of illumination and materials. We have shown that our recovered environment maps have high quality and high diversity while exhibiting an unprecedented level of naturalness.

Our approach is not free from limitations, providing plenty opportunity for future work. We did not aim at devising a full inverse-rendering system, as we assume geometry to be given. This has allowed us to study the problem of material–illumination ambiguities in depth. A natural next step is the inclusion of geometry estimation into our pipeline.

Importantly, we can only expect reasonable results if the distribution of environment maps captured by our generative model matches the task. Fig. 14 illustrates an example of an intentionally created mismatch. In the top row, we use a simple synthetic environment map to illuminate a bunny and run our method using the outdoor model. In the bottom row, we illuminate an armchair using indoor illumination and, again, run our method using the outdoor model. We see that, in these cases, the recovered environment maps are far from the ground truth, but nevertheless capture the overall illumination condition well.

Current diffusion models are notoriously inefficient when it comes to generating samples. We inherit this limitation, which makes our (unoptimized) implementation roughly 5x slower than a vanilla path-tracing-based solution.

We hope to inspire future work on generative inverse rendering, balancing the intricate triangle between accuracy, naturalness, and ambiguity.

Acknowledgements.

This work was supported by the ERC Consolidator Grant 4DReply (770784) and the Lise Meitner Postdoctoral Fellowship.

REFERENCES

Brian DO Anderson. 1982. Reverse-time diffusion equation models. *Stochastic Processes and their Applications* 12, 3 (1982), 313–326.

Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. 2022. Mip-NeRF 360: Unbounded Anti-Aliased Neural Radiance Fields. *CVPR* (2022).

Ronen Basri and David W Jacobs. 2003. Lambertian reflectance and linear subspaces. *IEEE transactions on pattern analysis and machine intelligence* 25, 2 (2003), 218–233.

Yochai Blau and Tomer Michaeli. 2018. The perception-distortion tradeoff. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6228–6237.

Brent Burley. 2015. Extending the Disney BRDF to a BSDF with integrated subsurface scattering. *Physically Based Shading in Theory and Practice'SIGGRAPH Course* (2015).

Brent Burley and Walt Disney Animation Studios. 2012. Physically-based shading at disney. In *Acm Siggraph*, Vol. 2012. vol. 2012, 1–7.

Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. 2021. Ilvr: Conditioning method for denoising diffusion probabilistic models. *arXiv preprint arXiv:2108.02938* (2021).

Hyungjin Chung, Jeongsol Kim, Sehui Kim, and Jong Chul Ye. 2022a. Parallel Diffusion Models of Operator and Image for Blind Inverse Problems. *arXiv preprint arXiv:2211.10656* (2022).

Hyungjin Chung, Jeongsol Kim, Michael T Mccann, Marc L Klasky, and Jong Chul Ye. 2022b. Diffusion posterior sampling for general noisy inverse problems. *arXiv preprint arXiv:2209.14687* (2022).

Hyungjin Chung, Byeongsu Sim, and Jong Chul Ye. 2022c. Come-closer-diffuse-faster: Accelerating conditional diffusion models for inverse problems through stochastic contraction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12413–12422.

Blender Online Community. 2018. Blender - a 3D modelling and rendering package. http://www.blender.org

Mohammad Reza Karimi Dastjerdi, Yannick Hold-Geoffroy, Jonathan Eisenmann, Siavash Khodadadeh, and Jean-François Lalonde. 2022. Guided Co-Modulated GAN for 360° Field of View Extrapolation. In *2022 International Conference on 3D Vision (3DV)*. IEEE, 475–485.

Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems* 34 (2021), 8780–8794.

Ron O Dror, Alan S Willsky, and Edward H Adelson. 2004. Statistical characterization of real-world illumination. *Journal of Vision* 4, 9 (2004), 11–11.

Bernhard Egger, Sandro Schönborn, Andreas Schneider, Adam Kortylewski, Andreas Morel-Forster, Clemens Blumer, and Thomas Vetter. 2018. Occlusion-aware 3d morphable models and an illumination prior for face image analysis. *International Journal of Computer Vision* 126 (2018), 1269–1287.

James AD Gardner, Bernhard Egger, and William AP Smith. 2022. Rotation-Equivariant Conditional Spherical Neural Fields for Learning a Natural Illumination Prior. *arXiv preprint arXiv:2206.03858* (2022).

Marc-André Gardner, Kalyan Sunkavalli, Ersin Yumer, Xiaohui Shen, Emiliano Gambaretto, Christian Gagné, and Jean-François Lalonde. 2017. Learning to predict indoor illumination from a single image. *arXiv preprint arXiv:1704.00090* (2017).

Param Hanji, Rafal Mantiuk, Gabriel Eilertsen, Saghi Hajisharif, and Jonas Unger. 2022. Comparison of single image HDR reconstruction methods—the caveats of quality assessment. In *ACM SIGGRAPH 2022 Conference Proceedings*. 1–8.

Jon Hasselgren, Nikolai Hofmann, and Jacob Munkberg. 2022. Shape, light & material decomposition from images using monte carlo rendering and denoising. *arXiv preprint arXiv:2206.03380* (2022).

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs trained by a two time-scale update rule converge to a local nash equilibrium. *NeurIPS* 30 (2017).

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems* 33 (2020), 6840–6851.

Aapo Hyvärinen and Peter Dayan. 2005. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research* 6, 4 (2005).

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

Wenzel Jakob, Sébastien Speierer, Nicolas Roussel, and Delio Vicini. 2022a. Dr.Jit: A Just-In-Time Compiler for Differentiable Rendering. *Transactions on Graphics (Proceedings of SIGGRAPH)* 41, 4 (July 2022). https://doi.org/10.1145/3528223.3530099

Haian Jin, Isabella Liu, Peijia Xu, Xiaoshuai Zhang, Songfang Han, Sai Bi, Xiaowei Zhou, Zexiang Xu, and Hao Su. 2023. TensoIR: Tensorial Inverse Rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 165–174.

James T Kajiya. 1986. The rendering equation. In *Proceedings of the 13th annual conference on Computer graphics and interactive techniques*. 143–150.

Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. 2020. Training generative adversarial networks with limited data. *Advances in neural information processing systems* 33 (2020), 12104–12114.

Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada. 2018. Neural 3d mesh renderer. In *CVPR*. 3907–3916.

Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. 2022. Denoising diffusion restoration models. *arXiv preprint arXiv:2201.11793* (2022).

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *ICLR*.

Samuli Laine, Janne Hellsten, Tero Karras, Yeongho Seol, Jaakko Lehtinen, and Timo Aila. 2020. Modular primitives for high-performance differentiable rendering. *ACM Transactions on Graphics (TOG)* 39, 6 (2020), 1–14.

Jaakko Lehtinen, Jacob Munkberg, Jon Hasselgren, Samuli Laine, Tero Karras, Miika Aittala, and Timo Aila. 2018. Noise2Noise: Learning image restoration without clean data. *arXiv preprint arXiv:1803.04189* (2018).

Tzu-Mao Li, Miika Aittala, Frédo Durand, and Jaakko Lehtinen. 2018. Differentiable monte carlo ray tracing through edge sampling. *ACM Transactions on Graphics (TOG)* 37, 6 (2018), 1–11.

Shichen Liu, Weikai Chen, Tianye Li, and Hao Li. 2019. Soft rasterizer: Differentiable rendering for unsupervised single-view mesh reconstruction. *arXiv preprint arXiv:1901.05567* (2019).

Matthew M Loper and Michael J Black. 2014. OpenDR: An approximate differentiable renderer. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VII 13*. Springer, 154–169.

Guillaume Loubet, Nicolas Holzschuch, and Wenzel Jakob. 2019. Reparameterizing discontinuous integrands for differentiable rendering. *ACM Transactions on Graphics (TOG)* 38, 6 (2019), 1–14.

Linjie Lyu, Marc Habermann, Lingjie Liu, Ayush Tewari, Christian Theobalt, et al. 2021. Efficient and differentiable shadow computation for inverse problems. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 13107–13116.

Linjie Lyu, Ayush Tewari, Thomas Leimkühler, Marc Habermann, and Christian Theobalt. 2022. Neural Radiance Transfer Fields for Relightable Novel-view Synthesis with Global Illumination. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVII*. Springer, 153–169.

David McAllester. 2023. On the Mathematics of Diffusion Models. *arXiv preprint arXiv:2301.11108* (2023).

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Commun. ACM* 65, 1 (2021), 99–106.

Piotr Mirowski, Andras Banki-Horvath, Keith Anderson, Denis Teplyashin, Karl Moritz Hermann, Mateusz Malinowski, Matthew Koichi Grimes, Karen Simonyan, Koray Kavukcuoglu, Andrew Zisserman, et al. 2019. The streetlearn environment and dataset. *arXiv preprint arXiv:1903.01292* (2019).

Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. 2013. Making a "Completely Blind" Image Quality Analyzer. *IEEE Signal Processing Letters* 20, 3 (2013), 209–212. https://doi.org/10.1109/LSP.2012.2227726

Jacob Munkberg, Jon Hasselgren, Tianchang Shen, Jun Gao, Wenzheng Chen, Alex Evans, Thomas Müller, and Sanja Fidler. 2022. Extracting Triangular 3D Models, Materials, and Lighting From Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8280–8290.

Alexander Quinn Nichol and Prafulla Dhariwal. 2021. Improved denoising diffusion probabilistic models. In *ICML*. PMLR, 8162–8171.

Merlin Nimier-David, Delio Vicini, Tizian Zeltner, and Wenzel Jakob. 2019. Mitsuba 2: A retargetable forward and inverse renderer. *ACM Transactions on Graphics (TOG)* 38, 6 (2019), 1–17.

Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. 2019. Deepsdf: Learning continuous signed distance functions for shape representation. In *CVPR*. 165–174.

Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. 2022. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988* (2022).

Ravi Ramamoorthi and Pat Hanrahan. 2001. An efficient representation for irradiance environment maps. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*. 497–500.

Danilo Rezende and Shakir Mohamed. 2015. Variational inference with normalizing flows. In *International conference on machine learning*. PMLR, 1530–1538.

Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. 2022. Pivotal tuning for latent-based editing of real images. *ACM Transactions on graphics (TOG)* 42, 1 (2022), 1–13.

Marcel Santana Santos, Ren Tsang, and Nima Khademi Kalantari. 2020. Single Image HDR Reconstruction Using a CNN with Masked Features and Perceptual Loss. *ACM Transactions on Graphics* 39, 4 (7 2020). https://doi.org/10.1145/3386569.3392403

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*. PMLR, 2256–2265.

Jiaming Song, Arash Vahdat, Morteza Mardani, and Jan Kautz. 2023. Pseudoinverse-guided diffusion models for inverse problems. In *ICLR*.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 2020. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456* (2020).

Pratul P. Srinivasan, Boyang Deng, Xiuming Zhang, Matthew Tancik, Ben Mildenhall, and Jonathan T. Barron. 2021. NeRV: Neural Reflectance and Visibility Fields for Relighting and View Synthesis. In *CVPR*.

Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. 2018. Deep image prior. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 9446–9454.

Delio Vicini, Sébastien Speierer, and Wenzel Jakob. 2021. Path Replay Backpropagation: Differentiating Light Paths using Constant Memory and Linear Time. *Transactions on Graphics (Proceedings of SIGGRAPH)* 40, 4 (Aug. 2021), 108:1–108:14. https://doi.org/10.1145/3450626.3459804

Pascal Vincent. 2011. A connection between score matching and denoising autoencoders. *Neural computation* 23, 7 (2011), 1661–1674.

Guangcong Wang, Yinuo Yang, Chen Change Loy, and Ziwei Liu. 2022. Stylelight: Hdr panorama generation for lighting estimation and editing. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XV*. Springer, 477–492.

Jiaping Wang, Peiran Ren, Minmin Gong, John Snyder, and Baining Guo. 2009. All-frequency rendering of dynamic, spatially-varying reflectance. In *ACM SIGGRAPH Asia 2009 papers*. 1–10.

Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. 2021. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689* (2021).

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 13, 4 (2004), 600–612.

Haoqian Wu, Zhipeng Hu, Lincheng Li, Yongqiang Zhang, Changjie Fan, and Xin Yu. 2023. NeFII: Inverse Rendering for Reflectance Decomposition with Near-Field Indirect Illumination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4295–4304.

Ye Yu and William AP Smith. 2021. Outdoor inverse rendering from a single image using multiview self-supervision. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 7 (2021), 3659–3675.

Fangneng Zhan, Changgong Zhang, Yingchen Yu, Yuan Chang, Shijian Lu, Feiying Ma, and Xuansong Xie. 2021. Emlight: Lighting estimation via spherical distribution approximation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 3287–3295.

Kai Zhang, Fujun Luan, Qianqian Wang, Kavita Bala, and Noah Snavely. 2021a. Physg: Inverse rendering with spherical gaussians for physics-based material editing and relighting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5453–5462.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *CVPR*.

Xiuming Zhang, Pratul P Srinivasan, Boyang Deng, Paul Debevec, William T Freeman, and Jonathan T Barron. 2021b. Nerfactor: Neural factorization of shape and reflectance under an unknown illumination. *ACM Transactions on Graphics (TOG)* 40, 6 (2021), 1–18.

Yuanqing Zhang, Jiaming Sun, Xingyi He, Huan Fu, Rongfei Jia, and Xiaowei Zhou. 2022. Modeling indirect illumination for inverse rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18643–18652.