Time-conditioned Illumination for Inverse Rendering of Outdoor Scenes

Xiaoxue Chen[1,2]
chenxx21@mails.tsinghua.edu.cn
Hao Zhao[1]
zhaohao@air.tsinghua.edu.cn
Guyue Zhou[1]
zhouguyue@air.tsinghua.edu.cn
Ya-Qin Zhang[1]
zhangyaqin@air.tsinghua.edu.cn

[1] Institute for AI Industry Research (AIR), Tsinghua University, Beijing, China

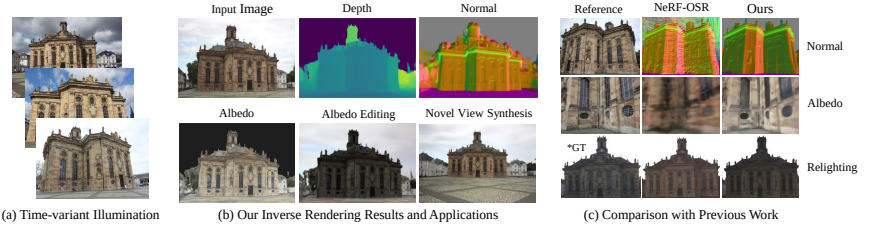
[2] Department of Computer Science and Technology, Tsinghua University, Beijing, China

Abstract

Inverse rendering has been a long-standing problem in computer graphics and vision, with the objective of decomposing images into intrinsic scene properties including geometry, illumination, and material. While conventional works mainly focus on object-level or indoor scenes, addressing the inverse rendering problem for outdoor scenes poses more challenges, which arise from the complex geometry and time-variant nature of outdoor illumination due to changing sun position and atmosphere condition. In this paper, we present a novel inverse rendering framework tailored for outdoor scenes characterized by varying illumination. Specifically, we first disentangle the underlying geometry from appearance based on the neural radiance fields and incorporate monocular geometric cues to resolve the complexity of geometry. In addition, we introduce a time-dependent field to model the time-variant illumination from the sky dome and parameterize the material properties with the microfacet Bidirectional Reflectance Distribution Function (BRDF). Finally, we propose a differentiable re-rendering module that integrates all the decomposed properties to generate new renderings. Experiments demonstrate that our novel inverse rendering framework yields high-quality reconstruction of scenes' geometry, material, and illumination, and outperforms previous SOTA methods in the task of novel view synthesis for outdoor scenes. Moreover, this framework facilitates various scene editing applications including material editing, object removal, and relighting.

1 Introduction

Inverse rendering is one of the fundamental tasks in the domains of computer graphics and vision, which aims at reconstructing 3D intrinsic properties from its corresponding 2D images, and the properties typically encompass geometry [25, 45], materials (e.g., albedo, roughness) [26, 64], and lighting [39, 40]. In computer graphics, rendering[1, 20] means generating photo-realistic images from known properties of a scene through rasterization or ray-tracing, which is a process of projecting from a high-dimensional scene space to a



(a) Time-variant Illumination (b) Our Inverse Rendering Results and Applications (c) Comparison with Previous Work

Figure 1: (a) Illustration of time-variant illumination of an outdoor scene. (b) Our inverse rendering framework decomposes intrinsic scene properties including depth, normal, and albedo. (c) Comparison results with NeRF-OSR[37].

two-dimensional image space. Conversely, inverse rendering involves the intricate process of elevating the two-dimensional image space to a high-dimensional parameter space, which is evidently ill-posed and challenging. Nonetheless, addressing this challenge is vital, as inverse rendering enables the reconstruction of high-fidelity 3D scenes from captured imagery. Furthermore, it allows for adding or removing objects and manipulating scene properties, such as user-specified lighting (relighting) and material editing that benefit a series of downstream applications including augmented reality and virtual reality.

Based on the target scenarios, inverse rendering problems can be categorized into three types: single-object scenes[19, 59], indoor scenes[63, 64], and outdoor scenes[37]. The principal distinctions among them revolve around the diversity of objects and different lighting conditions. Obviously, outdoor scenes have more categories and numbers of objects. Besides, in single-object scenes, the illumination is typically simplified through directional lighting or a static environmental map, and the lighting of indoor scenes commonly comes from fixed light sources such as lamps or windows. By contrast, outdoor illumination mainly comes from the sky dome, where the presence of intricate and dynamic weather conditions poses challenges for accurate illumination modeling, and the appearance of the scene also varies with the illumination, as illustrated in Fig. 1 (a). It's difficult to capture multi-view images under consistent lighting conditions, as this requires setting up cameras at various locations simultaneously for image acquisition. Even the exposures of cameras need to be the same. Therefore, it is necessary to explore how to learn static intrinsic properties like geometry and material from multi-view images captured under varying illuminations.

Recently, neural radiance fields (NeRFs) [31] have revolutionized the field of image-based view synthesis, and they represent a 3D scene as a continuous radiance field. Despite showing promising results in geometry reconstruction [25, 44], object-level decomposition [19, 59] and indoor scene decomposition [63], the problem of reconstructing intrinsic properties of dynamic outdoor scenes with NeRF remains largely unexplored. Therefore, we apply them to the outdoor scene. Specifically, we introduce a NeRF-based framework that parses geometry, material, and lighting and rerenders novel views for outdoor scenes with variant illumination from a set of posed images.

With the goal of modeling the intrinsic properties of outdoor scenes, we utilize NeRF as the representation of the scene's appearance and employ the signal distance function (SDF) [55] as the geometric representation, and the SDF values can be integrated into NeRF's volume rendering formulation. We resolve the inherent ambiguity in the decomposition of appearance and geometry by introducing geometric cues as supervision. In addition, we parameterize the material of the scene with a microfacet bidirectional reflectance distribution function (BRDF) and estimate albedo and roughness with an additional branch based

on NeRF's formulation. To account for scene illumination, we consider a HDR environment map as the lighting representation for outdoor scenes and propose a time-dependent illumination field that takes the timestamp and view direction as input, enabling the estimation of varying illumination caused by dynamic weather. The above modules result in a decomposition of the scene into material, geometry, and lighting. Subsequently, the re-rendering of the scene's appearance becomes feasible by employing surface rendering algorithms based on differentiable Monte Carlo ray tracing techniques. Experiment results demonstrate that our framework not only achieves high-quality reconstruction and inverse graphics of outdoor scenes but also facilitates scene editing applications such as relighting and material editing.

2 Related Works

Novel view synthesis involves generating images of unseen viewpoints based on a set of input views. With the advent of deep learning, many learning-based methods [11, 12, 21, 28, 30, 62] emerge and synthesize photorealistic renderings through differentiable rendering pipelines. Recently, NeRF [31] has made significant strides in view synthesis, many works have been dedicated to enhancing its quality and efficiency in novel view synthesis. For instance, some works[8, 49, 54] focus on reducing the number of training views for NeRF, while others aim to achieve anti-aliasing in the rendered images [2, 3, 27]. Besides, a sequence of studies has been undertaken to enhance the representation capabilities of NeRF through the utilization of grid-based architectures [13, 32, 56] and point-based approaches [41, 51]. Furthermore, some works [25, 33, 44, 53] integrate NeRF's volume rendering pipeline with implicit surface representation and achieve both surface reconstruction and volume rendering using a single model. We extend NeRF to outdoor scenes[34, 48] under time-variant lighting and propose a NeRF-based pipeline to learn intrinsic scene properties.

Lighting estimation. In the pioneering work [6], a mirrored sphere is employed as a physical probe to measure the radiance of the surrounding environment. Subsequent studies [7, 36, 47] explore the use of other known objects as light probes. With the development of deep learning, many works learn generalizable models directly from images in a data-driven fashion. For example, [40, 43, 57] estimate the environment lighting from RGB images using end-to-end neural networks. Besides, [39, 61] estimate an HDR environment map and show applications like object insertion. Specific to outdoor illumination, some works [16, 58] apply analytical sky models like Hošek-Wilkie model[17] or the Lalonde-Matthews model[23] to represent the lighting from sky dome. However, these analytical models fail to capture the various illumination conditions for outdoor scenes and ignore the time-variant nature. In this work, we adopt an HDR environment map as the illumination representation and use a time-dependent field to cope with complex and changing outdoor lighting.

Inverse rendering is a long-term goal in computer vision, involving many intrinsic properties to be decomposed, including geometry [18, 50], illumination [14, 40], and materials [9, 29]. Some learning-based approaches [5, 38, 55, 64] employ dense prediction networks to estimate these properties in a data-driven way. These methods typically leverage large-scale datasets[26] for training, however, relying on the ground truth of albedo and geometry poses challenges due to difficulty in acquisition. Additionally, these approaches are typically designed for single-image inputs, making them lack multi-view consistency. The emergence of NeRF [31] has influenced many recent works, such as [4, 19, 46, 59, 60, 63], adopting NeRF as the foundation to address inverse rendering problem. These works are usually limited to single objects[59] or indoor scenes[63] or do not consider the variant appearance of out-

door scenes[52]. [37, 46] addresses varying illumination with per-image learnable lighting parameters, yet their representation of intrinsic properties lacks physical modeling, resulting in inaccurate estimates. To this end, we propose a physically-based inverse rendering framework based on NeRF to resolve the time-variant illumination of outdoor scenes.

3 Method

Our goal is to decompose the intrinsic properties of outdoor scenes from a set of posed images, accounting for their temporal variation due to illumination changes. Our inverse rendering framework is composed of two stages: the first one is to disentangle the underlying geometry from appearance based on NeRF while the second is to optimize the illumination model and material parameters, and a re-rendered image can be generated with the differentiable rendering pipeline. The overall framework is depicted in Fig. 2.

3.1 Geometry Reconstruction

To disentangle the underlying geometry from the scene appearance, we represent the scene geometry as a neural implicit surface and combine it with volume rendering following previous arts [25, 44, 45]. Specifically, we model the geometry using the signed distance function (SDF) denoted as $s(x)$, which is a continuous function that describes the distance to the nearest surface given a 3D coordinate x. Following [32], we parameterize the SDF using a grid-based neural network f_s built on multi-resolution hash encoding:

$$s(x), z(x) = f_s(h(x)), \tag{1}$$

in which $h(\cdot)$ is the multi-resolution hash encoding and $z(x)$ is the geometric feature of point x. While the geometry is constant under lighting variation, the scene's appearance usually changes with the illumination. To model the variant appearance, we employ a NeRF-based approach, parameterized by a neural network f_θ. In light of the time-variant nature of appearance, we enhance the vanilla NeRF by introducing a timestamp denoted as τ as an additional input. More specifically, given a 3D coordinate x and a viewing direction d at a particular timestamp τ_i, we represent the scene appearance as follows:

$$c = f_\theta(z(x), d, e(\tau_i)). \tag{2}$$

c is the RGB color value, and e is a Fourier embeder to encode timestamps into high-frequency signals. Let $r(t) = o + td$ denote a camera ray, the color $C(r)$ is computed as:

$$C(r) = \sum_{i=1}^{N} T_r^i \alpha_r^i c_r^i, \quad T_r^i = \prod_{j=1}^{i-1}(1-\alpha_r^j), \quad \alpha_r^i = 1 - \exp(-\sigma_r^i \delta_r^i), \tag{3}$$

where σ is the volume density, T_r^i and α_r^j denote the transmittance and alpha value of sample point i along ray r respectively, δ_r^i is the distance between neighboring sample points. To make the implicit surface representation compatible with NeRF's volume rendering formulation, we transform the SDF value s into the density value σ following VolSDF [53]. Then the geometric properties including normal $N(r)$ and depth $D(r)$ can also be obtained as:

$$D(r) = \sum_{i=1}^{N} T_r^i \alpha_r^i t_r^i, \quad N(r) = \sum_{i=1}^{N} T_r^i \alpha_r^i n_r^i, \tag{4}$$

the 3D unit normal n is the normalized analytical gradient of the SDF function s. The objec-

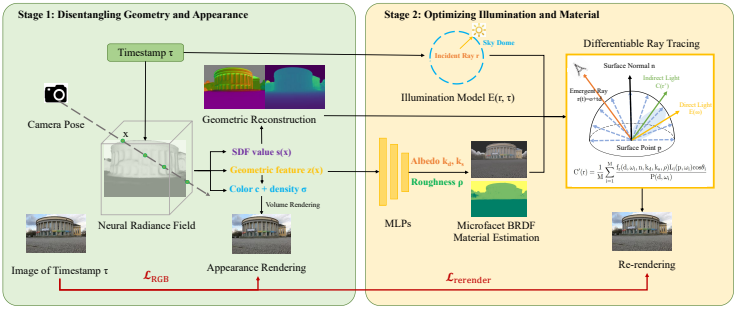


Figure 2: **Overall architecture**. Our inverse rendering framework is composed of two stages: the first one is to disentangle the underlying geometry from appearance based on NeRF while the second is to optimize the illumination model and material parameters, a re-rendered image can be generated with the differentiable rendering pipeline.

tive of the training procedure is defined as:

$$\mathcal{L}_{\text{RGB}} = \frac{1}{|\text{R}|}\sum_{\text{r}\in\text{R}} \|\text{C}(\text{r}) - \hat{\text{C}}(\text{r})\|_2^2, \quad \mathcal{L}_{\text{eikonal}} = \frac{1}{|\text{X}|}\sum_{\text{x}\in\text{X}} (||\nabla \text{S}(\text{x})|| - 1)^2, \tag{5}$$

where R denotes a set of rays, and $\hat{\text{C}}(\text{r})$ is the ground-truth RGB value. $\mathcal{L}_{\text{eikonal}}$ is an eikonal loss to regularize SDF values, where X is the set of 3D points sampled near the surface.

Monocular geometric cues. Disentangling the scene's appearance and underlying geometry from images is an ill-posed problem. Consequently, the incorporation of an additional geometric constraint becomes pivotal in resolving the inherent ambiguity. In our approach, we leverage large vision models to furnish us with geometric priors. While monocular geometric cues from pre-trained models may not be not accurate, they can still serve as a geometric constraint for the reconstruction. In detail, we employ Omnidata[10], a pre-trained model based on the Transformer architecture, to preprocess the multi-view training images and obtain the normal map denoted as $\hat{\text{N}}$. This normal map serves as a supervision to guide the reconstruction of geometry, which is formulated as:

$$\mathcal{L}_{\text{normal}} = \frac{1}{|\text{R}|}\sum_{\text{r}\in\text{R}} \|\text{N}(\text{r}) - \hat{\text{N}}(\text{r})\|_2^2. \tag{6}$$

3.2 Illumination Modeling

In order to facilitate scene editing applications like relighting, it is important to have an editable light source representation that can accommodate various lighting conditions and support realistic rendering of complex scenes. For outdoor scenes, illumination primarily originates from the sky. Hence, we employ High Dynamic Range (HDR) environment maps as the illumination representation for the sky dome. Following the definition of spherical environment mapping, the radiance of the emergent ray r(t)=o+td only depends on its direction d, which can be formulated as a neural illumination field. Furthermore, given the time-variant nature of outdoor illumination, we extend the representation of illumination to a time-dependent field, which is defined as:

$$\text{E}(\text{r}, \tau_\text{i}) = \text{f}_\text{e}(\text{d}, \text{e}(\tau_\text{i})), \tag{7}$$

where E is the corresponding HDR value, given a ray r pointing directly to the sky without being occluded by the foreground of the scene. This allows us to use a set of images captured

at various times and under different lighting conditions for training. Moreover, relighting becomes a straightforward process by simply adjusting the timestamp of the corresponding illumination. Besides, it is essential to model the spatial distribution of light sources, which means we need to distinguish between the foreground and the sky. To achieve this, we incorporate a pre-trained semantic segmentation model [24] to derive a mask for the sky, denoted as $\hat{M}_e$. To acquire the corresponding sky mask at any view, we leverage the NeRF's density to render the foreground probability M_f for a given ray r:

$$M_f(r) = \sum_{i=1}^{N} T_r^i \alpha_r^i \sigma_r^i, \quad M_e(r) = 1 - M_f(r). \tag{8}$$

M_e shows the probability that r directly points to the sky. Then a cross-entropy loss is introduced as:

$$\mathcal{L}_{seg} = \frac{1}{|R|} \sum_{r \in R} \left[\hat{M}_e \log M_e + (1 - \hat{M}_e) \log(1 - M_e) \right]. \tag{9}$$

3.3 Material Estimation

We employ the GGX microfacet BRDF [15] for the representation of spatial-varying materials. The microfacet model serves as a physically-based surface illumination model, facilitating high-fidelity and photorealistic rendering results. In the specific formulation of the BRDF denoted as $f_r(d, \omega, n, k_d, k_s, \rho)$, the variables are as follows: d, ω, and n represent geometric properties including the emergent ray's direction, incident ray's direction, and the surface point's normal, while k_d, k_s, and ρ denote material parameters describing diffuse albedo, specular albedo, and surface roughness respectively.

Given that the material parameters exhibit certain spatial distributions, it is natural to employ a neural field to estimate them. Since the materials are independent of view direction and timestamp, the neural material field can be conditioned solely on geometric features as:

$$k_d(x), k_s(x) = f_m(z(x)), \quad \rho(x) = f_\rho(z(x)). \tag{10}$$

$k_d(x)$, $k_s(x)$ and $\rho(x)$ are the corresponding material prediction given a coordinate x. Consequently, the material estimation for a given ray is expressed as follows:

$$k_d(r) = \sum_{i=1}^{N} T_r^i \alpha_r^i k_{d_r}^i, \quad k_s(r) = \sum_{i=1}^{N} T_r^i \alpha_r^i k_{s_r}^i, \tag{11}$$

$\rho(r)$ can be calculated in the same way. The material of the scene surface is relatively smooth, we use a smoothness loss to supervise the material parameters:

$$\mathcal{L}_{mat} = \frac{1}{|X|} \sum_{x \in X} [\|k_d(x) - k_d(x')\|_2 + \|k_s(x) - k_s(x')\|_2 + \|\rho(x) - \rho(x')\|_2]. \tag{12}$$

where X is the set of sampled points near the surface, $x' = x + \varepsilon$ and ε is a small random uniform 3D perturbation. Moreover, a regularization loss is introduced according to energy conservation law that the sum of diffuse and specular albedo should not exceed 1:

$$\mathcal{L}_{reg} = \frac{1}{|X|} \sum_{x \in X} \mathrm{ReLU}(k_d(x) + k_s(x) - 1). \tag{13}$$

3.4 Rerender with Intrinsic Properties

Building upon the decomposed intrinsic properties encompassing geometry, material, and illumination, our objective is to integrate them into the scene's appearance and regenerate a photorealistic image. This process empowers various image editing applications, such as relighting and material editing. Leveraging the previously estimated BRDF, we can efficiently

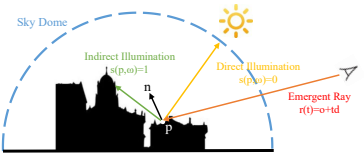


Figure 3: Illustration of outdoor illumination model.

re-render the image using a surface rendering algorithm. For every pixel in the image, the emergent ray r(t) = o + td through the pixel can be computed with camera parameters. For rays not intersecting with scene surfaces, the rerendered color $C'(r)$ equals the sky illumination $E(r)$, τ is omitted for simplicity. For the rays intersecting with the surface, the outgoing radiance at the intersection point is given by the rendering equation which is grounded in the principle of energy conservation:

$$C'(r) = L_o(p,d) = \int_{\Omega^+} f_r(d,\omega,n,k_d,k_s,\rho)L_i(p,\omega)\cos\theta d\omega \tag{14}$$

where p is the intersection point, Ω^+ is the positive hemisphere determined by the surface normal n at point p, ω represents all possible incident ray directions in Ω^+, L_i is the incident radiance from direction ω on point p, and θ is the angle between d and ω.

As illustrated in Fig. 3, the incident illumination L_i of outdoor scenes can be divided into two types: direct illumination and indirect illumination. The former is the lights coming directly from the sky dome while the latter is lights reflected by the scene surfaces. Given the incident ray $r'(t) = p + t\omega$, L_i can be calculated from the sky illumination and the surface radiance as:

$$L_i(p,\omega) = [1 - M_e(r')]C(r') + M_e(r')E(\omega), \tag{15}$$

where $M_e(r') \in [0,1]$ shows the probability that the ray $r'(t) = p + t\omega$ is not occluded by the scene surface, which can be obtained as Eq. 8.

Considering that we need to utilize the re-rendering process to optimize the decomposed intrinsic properties, the re-rendering should be differentiable. We now describe how we achieve the differentiable re-rendering based on Eq. 14. Given an emergent ray r(t) = o + td, we can easily trace the intersection point p with surface based on SDF values as $p = o + (\sum_{i=1}^{N} T_r^i \alpha_r^i t_r^i)d$, and the normal of p can be accessed through Eq. 4. Since the definite integration in Eq. 14 is intractable, we use Monte Carlo numerical integration to resolve it. Specifically, given a sampling rate M, we generate M incident rays using importance sampling, and the rays are denoted as $r'_i(t) = p + t\omega_i, i = 1...M$, then Eq. 14 can be reformulated through Monte Carlo integration as:

$$C'(r) = \frac{1}{M}\sum_{i=1}^{M} \frac{f_r(d,\omega_i,n,k_d,k_s,\rho)L_i(p,\omega_i)\cos\theta_i}{P(d,\omega_i)}. \tag{16}$$

$P(d,\omega_i)$ is the probability density function dependent on the estimated material parameters k_d, k_s, ρ. The optimization objective for re-rendering is defined as:

$$\mathcal{L}_{\text{rerender}} = \frac{1}{|R|}\sum_{r\in R} \|C'(r) - \hat{C}(r)\|_2^2. \tag{17}$$

3.5 Training process

Inverse rendering is inherently an ill-posed problem, therefore it's important to impose constraints on the decomposition of intrinsic properties. Given our utilization of monocular

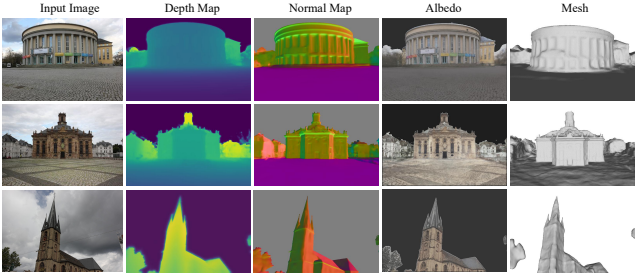


Figure 4: Qualitative results of intrinsic properties on NeRF-OSR dataset.

geometric cues as priors, it is reasonable to reconstruct the scene geometry first and subsequently estimate the material based on the results of geometric decomposition. Therefore, we adopt a two-stage training pipeline. In the first stage, we disentangle the geometry and appearance, employing the following loss function:

$$\mathcal{L}_{\text{stage1}} = \lambda_{\text{RGB}}\mathcal{L}_{\text{RGB}} + \lambda_{\text{normal}}\mathcal{L}_{\text{normal}} + \lambda_{\text{eikonal}}\mathcal{L}_{\text{eikonal}} + \lambda_{\text{seg}}\mathcal{L}_{\text{seg}}, \tag{18}$$

in which $\lambda_{(\cdot)}$ is the corresponding loss weight. In the second stage, we aim to optimize the material estimation results through the self-supervision of re-rendering, and the network parameters trained in the first stage are frozen. The total training objective for this stage is:

$$\mathcal{L}_{\text{stage2}} = \lambda_{\text{mat}}\mathcal{L}_{\text{mat}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}} + \lambda_{\text{rerender}}\mathcal{L}_{\text{rerender}}. \tag{19}$$

4 Experiment

We conduct experiments on the NeRF-OSR dataset [37], a benchmark of several outdoor sites. Each site within the dataset has been extensively photographed from multiple viewpoints and at various times. Specifically, these sites have been captured under diverse weather conditions, resulting in a total of 3,240 viewpoints captured across 110 different recording sessions. The illuminations encompass a range of weather conditions, including both sunny and cloudy days. We choose three representative sites and train a separate model for each of them. Each site is trained with images under different illuminations to account for the variations in lighting conditions. We employ the standard image quality metrics PSNR and SSIM for quantitative evaluations.

4.1 Comparison with SOTA methods

Novel view synthesis .Considering the impact of lighting on the appearance of the scene, we validate the performance of novel view synthesis on an illumination-invariant outdoor dataset, Tanks and Temples (TnT) [22]. Specifically, we use 100 frames from the Barn scene for experiments, with 90 images for training and 10 images for testing. The estimation of camera poses and training for baselines are implemented using Nerfstudio [42]. The quantitative results in Tab. 1 demonstrate that our method achieves better performance on view synthesis when compared to the baseline methods, which can be credited to the utilization of monocular normal cues, offering a valuable geometric prior for complex outdoor scenes.

Geometry reconstruction. Given that the dataset lacks ground truth of intrinsic properties, we present qualitative results for geometry reconstruction with our prediction of depth and normal maps. As depicted in Fig. 4, we can accurately estimate the normal map of the

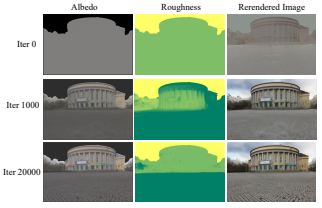


Figure 5: The material optimization process on NeRF-OSR dataset.

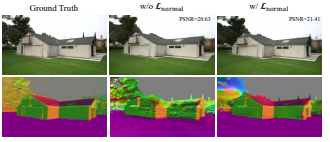


Figure 6: Qualitative comparison for novel view synthesis on TnT dataset with or without $\mathcal{L}_{\text{normal}}$.

Method	PSNR↑	SSIM↑
Nerfacto [42]	16.722	0.647
InstantNGP [32]	19.735	0.770
Ours (w/o $\mathcal{L}_{\text{normal}}$)	19.218	0.732
Ours	**20.031**	**0.775**

Table 1: Quantitative results for novel view synthesis on TnT dataset.

Method	PSNR↑	SSIM↑
End-to-end	14.533	0.412
Two-stage	**20.438**	**0.623**
Time-invariant	23.201	0.724
Time-variant	**24.435**	**0.734**

Table 2: Ablation study on the training process and the time-variant illumination.

scene, which is credited to the monocular geometric cues. Additionally, despite the absence of depth supervision, we can successfully estimate reasonable depth values, demonstrating that we have a comprehensive understanding of geometry. Moreover, we also generate explicit meshes of the scenes through marching cubes based on the SDF values. These meshes exhibit sufficient geometric details to serve as 3D assets in downstream applications.

Material and illumination estimation. As shown in Fig. 4, we achieve reasonable material estimation for various scenes. We further present the optimization results during the second stage in Fig. 5. As depicted in the figures, with an increasing number of optimization iterations, the albedo and roughness map gradually converge to reasonable estimates, for example, the texture on the ground and letters on the signboard. Simultaneously, the re-rendered images and corresponding illuminations progressively become photorealistic and increasingly resemble the ground truth image. These results verify the effectiveness of our method in successfully estimating the material and illumination for outdoor scenes.

4.2 Ablation Studies

Monocular geometric cues. To assess the effectiveness of normal supervision, we conducted an ablation study of $\mathcal{L}_{\text{normal}}$ on the TnT dataset, and the results are shown in Tab. 1. Introducing $\mathcal{L}_{\text{normal}}$ leads to a notable increase in performance, as the supervision serves as a geometric prior to resolving the inherent ambiguity for disentanglement of geometry and appearance. Furthermore, Fig. 6 offers qualitative comparisons that highlight the impact of utilizing geometric cues. With normal supervision, our model learns a superior geometric representation and excels in synthesizing the scene's appearance with more details.

Two-stage training. We explore the impact of two-stage training, and the results are presented in Tab. 2. End-to-end training indicates that the entire network is trained simultaneously with the objective function defined as $\mathcal{L}_{\text{end2end}} = \mathcal{L}_{\text{stage1}} + \mathcal{L}_{\text{stage2}}$. It's observed that both PSNR and SSIM exhibit a significant decrease compared to two-stage training. Because all the variables need to be optimized simultaneously, and the material parameters lack con-

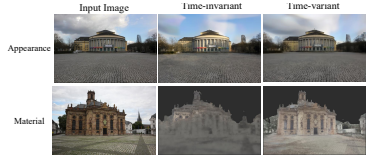


Figure 7: Comparison for time-variant model on NeRF-OSR dataset.

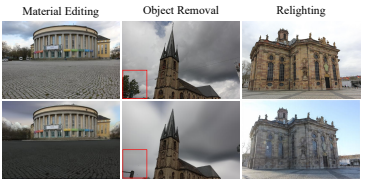


Figure 8: Results for scene editing.

Method	PSNR↑	SSIM↑
NeRF-OSR [37]	16.474	0.478
Ours	**17.091**	**0.497**

Table 3: Quantitative results for relighting on NeRF-OSR dataset.

straints, making the optimization highly ill-posed. In contrast, in the two-stage training, the optimization for material and illumination is conditioned on the frozen results of geometric reconstruction, which provides a more stable and convergent optimization process.

Time-variant illumination. We remove the timestamp input from both the NeRF and the illumination model. Tab. 2 presents the quantitative results for appearance rendering, highlighting the benefits of incorporating a time-variant input for learning dynamic appearance. As depicted in Fig. 7, the absence of the timestamp input results in the model's inability to accurately capture the changes in appearance caused by dynamic illumination. Moreover, it fails to learn the variant illumination, which leads to the learned albedo being blurred and covered by some unreasonable shading. This underscores the crucial role of our time-variant illumination model in representing the dynamic nature of outdoor lighting.

4.3 Scene Editing Applications

We conduct experiments on a set of scene-editing tasks, and the qualitative results are shown in Fig. 8. The first row is the inputs while the second row is the edited results. It's worth noting that all the renderings are synthesized using our differentiable renderer rather than rendering engines. For relighting, we relight the scene with an unseen environment map, which is implemented by mapping the illumination value of a ray to the RGB value of the environment map. Tab. 3 provide results for relighting a typical site with an environment map, and our framework achieves superior performance compared to NeRF-OSR.

5 Conclusion

In this work, we introduce a novel inverse rendering framework that targets reconstructing outdoor scenes under varying illumination. We disentangle geometry from appearance based on NeRF and introduce monocular geometric cues for geometric prior. Besides, we model the illumination with a time-dependent field and parameterize the material properties with the microfacet BRDF. A differentiable re-rendering pipeline is proposed to generate new renderings from the intrinsic properties. We evaluate the effectiveness of our framework on an outdoor dataset and demonstrate that it outperforms existing methods on novel view synthesis. Additionally, we provide qualitative results on a set of image editing tasks, which can serve as a hint for potential AR/VR applications.

References

[1] Tomas Akenine-Moller, Eric Haines, and Naty Hoffman. *Real-time rendering*. AK Peters/crc Press, 2019.

[2] 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.

[3] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5470–5479, 2022.

[4] Xiaoxue Chen, Junchen Liu, Hao Zhao, Guyue Zhou, and Ya-Qin Zhang. Nerrf: 3d reconstruction and view synthesis for transparent and specular objects with neural refractive-reflective fields. *arXiv preprint arXiv:2309.13039*, 2023.

[5] Xiaoxue Chen, Yuhang Zheng, Yupeng Zheng, Qiang Zhou, Hao Zhao, Guyue Zhou, and Ya-Qin Zhang. Dpf: Learning dense prediction fields with weak supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15347–15357, 2023.

[6] Paul Debevec. Rendering synthetic objects into real scenes: Bridging traditional and image-based graphics with global illumination and high dynamic range photography. In *Acm siggraph 2008 classes*, pages 1–10. 2008.

[7] Paul Debevec, Paul Graham, Jay Busch, and Mark Bolas. A single-shot light probe. In *ACM SIGGRAPH 2012 Talks*, pages 1–1. 2012.

[8] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised nerf: Fewer views and faster training for free. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12882–12891, 2022.

[9] Timothy D Dorney, Richard G Baraniuk, and Daniel M Mittleman. Material parameter estimation with terahertz time-domain spectroscopy. *JOSA A*, 18(7):1562–1571, 2001.

[10] Ainaz Eftekhar, Alexander Sax, Jitendra Malik, and Amir Zamir. Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10786–10796, 2021.

[11] John Flynn, Ivan Neulander, James Philbin, and Noah Snavely. Deepstereo: Learning to predict new views from the world's imagery. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5515–5524, 2016.

[12] John Flynn, Michael Broxton, Paul Debevec, Matthew DuVall, Graham Fyffe, Ryan Overbeck, Noah Snavely, and Richard Tucker. Deepview: View synthesis with learned gradient descent. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2367–2376, 2019.

[13] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5501–5510, 2022.

[14] Marc-André Gardner, Yannick Hold-Geoffroy, Kalyan Sunkavalli, Christian Gagné, and Jean-François Lalonde. Deep parametric indoor lighting estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7175–7183, 2019.

[15] Roy Hall. Illumination and color in computer-generated imagery. 1989.

[16] Yannick Hold-Geoffroy, Kalyan Sunkavalli, Sunil Hadap, Emiliano Gambaretto, and Jean-François Lalonde. Deep outdoor illumination estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7312–7321, 2017.

[17] Lukas Hosek and Alexander Wilkie. An analytic model for full spectral sky-dome radiance. *ACM Transactions on Graphics (TOG)*, 31(4):1–9, 2012.

[18] Mengqi Ji, Juergen Gall, Haitian Zheng, Yebin Liu, and Lu Fang. Surfacenet: An end-to-end 3d neural network for multiview stereopsis. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2307–2315, 2017.

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

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

[21] Nima Khademi Kalantari, Ting-Chun Wang, and Ravi Ramamoorthi. Learning-based view synthesis for light field cameras. *ACM Transactions on Graphics (TOG)*, 35(6): 1–10, 2016.

[22] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics*, 36 (4), 2017.

[23] Jean-François Lalonde and Iain Matthews. Lighting estimation in outdoor image collections. In *2014 2nd international conference on 3D vision*, volume 1, pages 131–138. IEEE, 2014.

[24] John Lambert, Zhuang Liu, Ozan Sener, James Hays, and Vladlen Koltun. Mseg: A composite dataset for multi-domain semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2879–2888, 2020.

[25] Zhaoshuo Li, Thomas Müller, Alex Evans, Russell H Taylor, Mathias Unberath, Ming-Yu Liu, and Chen-Hsuan Lin. Neuralangelo: High-fidelity neural surface reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8456–8465, 2023.

[26] Zhengqin Li, Ting-Wei Yu, Shen Sang, Sarah Wang, Meng Song, Yuhan Liu, Yu-Ying Yeh, Rui Zhu, Nitesh Gundavarapu, Jia Shi, et al. Openrooms: An open framework for photorealistic indoor scene datasets. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7190–7199, 2021.

[27] Junchen Liu, Wenbo Hu, Zhuo Yang, Jianteng Chen, Guoliang Wang, Xiaoxue Chen, Yantong Cai, Huan-ang Gao, and Hao Zhao. Rip-nerf: Anti-aliasing radiance fields with ripmap-encoded platonic solids. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024.

[28] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. Neural volumes: Learning dynamic renderable volumes from images. *arXiv preprint arXiv:1906.07751*, 2019.

[29] Abhimitra Meka, Maxim Maximov, Michael Zollhoefer, Avishek Chatterjee, Hans-Peter Seidel, Christian Richardt, and Christian Theobalt. Lime: Live intrinsic material estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6315–6324, 2018.

[30] Ben Mildenhall, Pratul P Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (TOG)*, 38(4):1–14, 2019.

[31] 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.

[32] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)*, 41(4):1–15, 2022.

[33] Michael Oechsle, Songyou Peng, and Andreas Geiger. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5589–5599, 2021.

[34] Aadi Palnitkar, Rashmi Kapu, Xiaomin Lin, Cheng Liu, Nare Karapetyan, and Yiannis Aloimonos. Chatsim: Underwater simulation with natural language prompting. In *OCEANS 2023-MTS/IEEE US Gulf Coast*, pages 1–7. IEEE, 2023.

[35] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019.

[36] Jeong Joon Park, Aleksander Holynski, and Steven M Seitz. Seeing the world in a bag of chips. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1417–1427, 2020.

[37] Viktor Rudnev, Mohamed Elgharib, William Smith, Lingjie Liu, Vladislav Golyanik, and Christian Theobalt. Nerf for outdoor scene relighting. In *European Conference on Computer Vision*, pages 615–631. Springer, 2022.

[38] Soumyadip Sengupta, Jinwei Gu, Kihwan Kim, Guilin Liu, David W Jacobs, and Jan Kautz. Neural inverse rendering of an indoor scene from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8598–8607, 2019.

[39] Gowri Somanath and Daniel Kurz. Hdr environment map estimation for real-time augmented reality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11298–11306, 2021.

[40] Shuran Song and Thomas Funkhouser. Neural illumination: Lighting prediction for indoor environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6918–6926, 2019.

[41] Xiaowei Song, Jv Zheng, Shiran Yuan, Huan-ang Gao, Jingwei Zhao, Xiang He, Weihao Gu, and Hao Zhao. Sa-gs: Scale-adaptive gaussian splatting for training-free anti-aliasing. *arXiv preprint arXiv:2403.19615*, 2024.

[42] Matthew Tancik, Ethan Weber, Evonne Ng, Ruilong Li, Brent Yi, Terrance Wang, Alexander Kristoffersen, Jake Austin, Kamyar Salahi, Abhik Ahuja, et al. Nerfstudio: A modular framework for neural radiance field development. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–12, 2023.

[43] Guangcong Wang, Yinuo Yang, Chen Change Loy, and Ziwei Liu. 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*, pages 477–492. Springer, 2022.

[44] Jiepeng Wang, Peng Wang, Xiaoxiao Long, Christian Theobalt, Taku Komura, Lingjie Liu, and Wenping Wang. Neuris: Neural reconstruction of indoor scenes using normal priors. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXII*, pages 139–155. Springer, 2022.

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

[46] Zian Wang, Tianchang Shen, Jun Gao, Shengyu Huang, Jacob Munkberg, Jon Hasselgren, Zan Gojcic, Wenzheng Chen, and Sanja Fidler. Neural fields meet explicit geometric representations for inverse rendering of urban scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8370–8380, 2023.

[47] Henrique Weber, Donald Prévost, and Jean-François Lalonde. Learning to estimate indoor lighting from 3d objects. In *2018 International Conference on 3D Vision (3DV)*, pages 199–207. IEEE, 2018.

[48] Zirui Wu, Tianyu Liu, Liyi Luo, Zhide Zhong, Jianteng Chen, Hongmin Xiao, Chao Hou, Haozhe Lou, Yuantao Chen, Runyi Yang, et al. Mars: An instance-aware, modular and realistic simulator for autonomous driving. In *CAAI International Conference on Artificial Intelligence*, pages 3–15. Springer, 2023.

[49] Dejia Xu, Yifan Jiang, Peihao Wang, Zhiwen Fan, Humphrey Shi, and Zhangyang Wang. Sinnerf: Training neural radiance fields on complex scenes from a single image. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII*, pages 736–753. Springer, 2022.

[50] Lan Xu, Zhuo Su, Lei Han, Tao Yu, Yebin Liu, and Lu Fang. Unstructuredfusion: Realtime 4d geometry and texture reconstruction using commercial rgbd cameras. *IEEE transactions on pattern analysis and machine intelligence*, 42(10):2508–2522, 2019.

[51] Qiangeng Xu, Zexiang Xu, Julien Philip, Sai Bi, Zhixin Shu, Kalyan Sunkavalli, and Ulrich Neumann. Point-nerf: Point-based neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5438–5448, 2022.

[52] Siqi Yang, Xuanning Cui, Yongjie Zhu, Jiajun Tang, Si Li, Zhaofei Yu, and Boxin Shi. Complementary intrinsics from neural radiance fields and cnns for outdoor scene relighting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16600–16609, 2023.

[53] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. *Advances in Neural Information Processing Systems*, 34:4805–4815, 2021.

[54] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4578–4587, 2021.

[55] Ye Yu and William AP Smith. Inverserendernet: Learning single image inverse rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3155–3164, 2019.

[56] Shiran Yuan and Hao Zhao. Slimmerf: Slimmable radiance fields. In *2024 International Conference on 3D Vision (3DV)*, pages 64–74. IEEE, 2024.

[57] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12830–12839, 2021.

[58] Jinsong Zhang, Kalyan Sunkavalli, Yannick Hold-Geoffroy, Sunil Hadap, Jonathan Eisenman, and Jean-François Lalonde. All-weather deep outdoor lighting estimation. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 10158–10166, 2019.

[59] Kai Zhang, Fujun Luan, Qianqian Wang, Kavita Bala, and Noah Snavely. 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*, pages 5453–5462, 2021.

[60] Kai Zhang, Fujun Luan, Zhengqi Li, and Noah Snavely. Iron: Inverse rendering by optimizing neural sdfs and materials from photometric images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5565–5574, 2022.

[61] Yiqin Zhao and Tian Guo. Pointar: Efficient lighting estimation for mobile augmented reality. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII 16*, pages 678–693. Springer, 2020.

[62] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018.

[63] Jingsen Zhu, Yuchi Huo, Qi Ye, Fujun Luan, Jifan Li, Dianbing Xi, Lisha Wang, Rui Tang, Wei Hua, Hujun Bao, et al. I2-sdf: Intrinsic indoor scene reconstruction and editing via raytracing in neural sdfs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12489–12498, 2023.

[64] 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.