Multi-view Inverse Rendering for Large-scale Real-world Indoor Scenes

Zhen Li[1] Lingli Wang[1] Mofang Cheng[1] Cihui Pan[1,*] Jiaqi Yang[2,*]
[1]Realsee [2]Northwestern Polytechnical University
yodlee@mail.nwpu.edu.cn,{wanglingli008,chengmofang001,pancihui001}@realsee.com,jqyang@nwpu.edu.cn

Abstract

*We present a efficient multi-view inverse rendering method for large-scale real-world indoor scenes that reconstructs global illumination and physically-reasonable SVBRDFs. Unlike previous representations, where the global illumination of large scenes is simplified as multiple environment maps, we propose a compact representation called **T**exture-**b**ased **L**ighting (TBL). It consists of 3D mesh and HDR textures, and efficiently models direct and infinite-bounce indirect lighting of the entire large scene. Based on TBL, we further propose a hybrid lighting representation with precomputed irradiance, which significantly improves the efficiency and alleviates the rendering noise in the material optimization. To physically disentangle the ambiguity between materials, we propose a three-stage material optimization strategy based on the priors of semantic segmentation and room segmentation. Extensive experiments show that the proposed method outperforms the state-of-the-art quantitatively and qualitatively, and enables physically-reasonable mixed-reality applications such as material editing, editable novel view synthesis and relighting. The project page is at https://lzleejean.github.io/TexIR.*

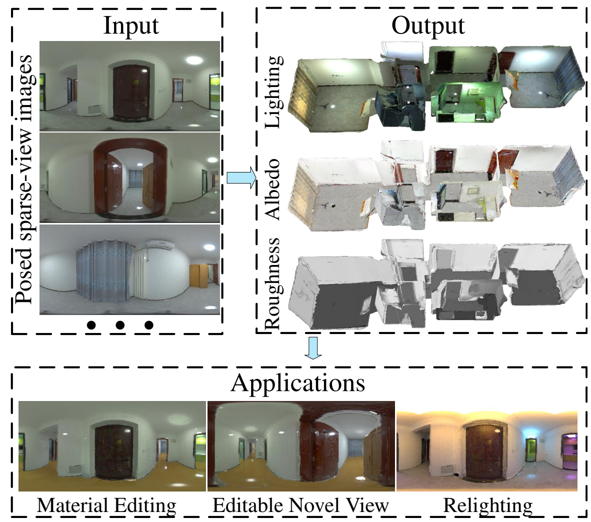

Figure 1. Given a set of posed sparse-view images for a large-scale scene, we reconstruct global illumination and SVBRDFs. The recovered properties are able to produce convincing results for several mixed-reality applications such as material editing, editable novel view synthesis and relighting. Note that we change roughness of all walls, and albedo of all floors. The detailed specular reflectance shows that our method successfully decomposes physically-reasonable SVBRDFs and lighting. Please refer to supplementary videos for more animations.

1. Introduction

Inverse rendering aims to reconstruct geometry, material and illumination of an object or a scene from images. These properties are essential to downstream applications such as scene editing, editable novel view synthesis and relighting. However, decomposing such properties from the images is extremely ill-posed, because different configurations of such properties often lead to similar appearance. With recent advances in differentiable rendering and implicit neural representation, several approaches have achieved significant success on small-scale object-centric scenes with explicit or implicit priors [7,32,34,43,50,51,55,57,58]. However, inverse rendering of large-scale indoor scenes has not been well solved.

There are two main challenges for large-scale indoor scenes. 1) ***Modelling the physically-correct global illumination.*** There are far more complex lighting effects, *e.g.*, inter-reflection and cast shadows, in large-scale indoor scenes than object-centric scenes due to complex occlusions, materials and local light sources. Although the widely-used image-based lighting (IBL) is able to efficiently model direct and indirect illumination, it only represents the lighting of a certain position [13, 17, 18, 42]. The spatial consistency of per-pixel or per-voxel IBL representations [26, 29, 47, 59] is difficult to ensure. Moreover, such incompact representations require large memory. Parameterized lights [16, 27] such as point light, area light and directional light are naturally globally-consistent, but modeling the expensive global light transport will be in-

*Co-corresponding authors.

evitable [1,37,58]. Thus, simple lighting representations applied in previous works are unsuitable in large-scale scenes. 2) ***Disentangling the ambiguity between materials.*** Different configurations of materials often lead to similar appearance, and to add insult to injury, there are an abundance of objects with complex and diverse materials in large-scale scenes. In object-centric scenes, dense views distributed on the hemisphere are helpful for alleviating the ambiguity [14, 22, 32, 35, 55, 56]. However, only sparse views are available in large-scale scenes, which more easily lead to ambiguous predictions [37,51].

In this work, we present TexIR, an efficient inverse rendering method for large-scale indoor scenes. Aforementioned challenges are tackled individually in the following. 1) We model the infinite-bounce global illumination of the entire scene with a novel compact lighting representation, called TBL. The TBL is able to efficiently represent the infinite-bounce global illumination of any position within the large scene. Such a compact and explicit representation provides more physically-accurate and spatially-varying illumination to guide the material estimation. Directly optimizing materials with TBL leads to expensive computation costs caused by high samples of the monte carlo sampling. Therefore, we precompute the irradiance based on our TBL, which significantly accelerates the expensive computation in the material optimization process. 2) To ameliorate the ambiguity between materials, we introduce a segmentation-based three-stage material optimization strategy. Specifically, we optimize a coarse albedo based on Lambertian-assumption in the first stage. In the second stage, we integrate semantics priors to guide the propagation of physically-correct roughness in regions with same semantics. In the last stage, we fine-tune both albedo and roughness based on the priors of semantic segmentation and room segmentation. By leveraging such priors, physically-reasonable albedo and roughness are disentangled globally.

To summarize, the main contributions of our method are as follows:

1. A compact lighting representation for large-scale scenes, where the infinite-bounce global illumination of the entire large scene can be handled efficiently.
2. A segmentation-based material optimization strategy to globally and physically disentangle the ambiguity between albedo and roughness of the entire scene.
3. A hybrid lighting representation based on the proposed TBL and precomputed irradiance to improve the efficiency in the material optimization process.

2. Related Work

Image-based lighting. Debevec *et al*. [13] first introduced the high dynamic range (HDR) light probe as a omni-directional lighting representation of a certain position, called IBL, which plays a vital role on inverse rendering and lighting estimation tasks in the computer vision. Barron and Malik [3] fitted the spherical-harmonic (SH) illumination, a parameterized representation of IBL, from a single image. Then they extended this single SH illumination into a set of SH illumination [2]. Zhou *et al*. [59] predicted the per-pixel SH lighting from a single image. Further, the per-pixel spherical-gaussian (SG) lighting [26] and the per-pixel light probe [29] are applied. Wang *et al*. [47] proposed a per-voxel SG lighting in the 3D indoor scene. The spatial consistency of illumination is unable to ensure with above lighting representations, especially for 3D large-scale scenes. Although light probes are able to project consistent light probes in others positions with known geometry, the efficiency in the optimization process is limited [25]. Our proposed TBL not only reserves the advantage of IBL, modelling infinite-bounce global illumination efficiently, but also is naturally globally-consistent.

Parametric lighting. Parametric lights, such as point lights, spot lights, area lights and directional lights, are classical lighting representations to define the illumination of a scene in computer graphics. Most methods only use one of the above lighting to represent the illumination of a scene. Nestmeyer *et al*. [36] predicted a directional light from a face image. Junxuan Li and Hongdong Li [24] used different directional lights as their illumination setting. Several approaches [6, 32, 43, 55] model the illumination of a static object with point lights. Li *et al*. [27] predicted area lights and SG directional lights for a single indoor image. Zhang *et al*. [54] optimized a more complex configuration via vertex-based lighting, including point lights, line lights, area lights and a environment map, but this method require user inputs to label where is the light source. Our TBL leverages the HDR texture of the entire scene to represent the illumination at any position within this 3D scene without any user input. Moreover, the TBL is compact and naturally globally-consistent.

Implicit lighting. With great advances in implicit representation, researchers began to use implicit neural networks to represent scenes. In particular, NeRF [33] has shown impressive results on scene representation. It leverages a neural density field and a neural radiance field to model a static small-scale scene. Zhang *et al*. [58] obtained the incident radiance of a certain position from a pre-trained outgoing radiance field [52]. To go a step further, the neural incident radiance field (NIRF), which represents incident radiance from any direction at any position, have shown great potential for inverse rendering [51] and novel view synthesis [45]. However, without constraints for such powerful implicit representation, the ambiguity between material and lighting is difficult to disentangle. Our explicit TBL is capable of eliminating this ambiguity in a interpretable manner.

Figure 2. **Overview of our inverse rendering pipeline.** Given sparse calibrated HDR images for a large-scale scene, we reconstruct the geometry and HDR textures as our lighting representation. PBR material textures of the scene, including albedo and roughness, are optimized by differentiable rendering (DR). The ambiguity between materials is disentangled by the semantics prior and the room segmentation prior. Gradient flows in Green Background.

Material estimation. The deep neural network has achieved significant success in many vision tasks with large-scale real-world datasets, such as object detection, semantic segmentation and depth estimation. Unfortunately, collecting the annotation of materials in real-world is difficult at scale. IIW [4] only labels sparse pairwise reflectance comparison for a single image. Therefore, most methods [8, 26, 28–30, 42, 47, 60] are alternative to use synthetic dataset to disambiguate the ambiguous properties. These approaches training on synthetic datasets struggle to eliminate the inevitable domain gap though several datasets have achieved photo-realistic results [29, 31, 38]. Optimization-based methods [2, 3, 7, 34, 37, 43, 56, 58] have shown impressive results on real-world multi-view images. It is known that the prior of different materials is essential to be leveraged in the optimization process to alleviate the ambiguity between materials. Barron and Malik [3] applied a smoothness term for albedo. Multiple variants of filters are used to smooth material in follow-up works [5, 26, 29]. Zhang *et al*. [57] leveraged learning priors from real-world BRDFs and imposed a spatial smoothness prior in the latent space. A similar idea has been adopted by a recent work [58]. Schmitt *et al*. [40] and Luan *et al*. [32] assumed similar diffuse albedos to have similar specular ones, and applied a non-local bilateral regularizer for specular albedo. Compared to object-centric scenes [7, 34, 43, 55–58], disentangling the ambiguity between materials is more challenging in the large-scale scene due to sparser observations and wider variety of objects and reflectance properties. Zhang *et al*. [54] recovered a constant attribute for each floor, wall and ceiling of an empty room. Most similar to ours, Nimier-David *et al*. [37] and Haefner *et al*. [20] also assumed similar semantic regions to have similar roughness and specularity. The first difference is we optimize the roughness in a soft manner instead optimizing a constant value [20, 37]. The another difference is we leverage the room segmentation prior, which enables us to recover different roughness even for similar semantic regions. Thanks to our efficient global illumination representation and explicit optimization strategy, our method is 20× faster than the differentiable path tracing-based method [37].

3. Methodology

As shown in Fig. 2, given a set of calibrated HDR images of a large-scale indoor scene, our method aims to accurately recover globally-consistent illumination and Spatially-Varying Bidirectional Reflectance Distribution Functions (SVBRDFs), which can be conveniently integrated into graphics pipelines and downstream applications. To this end, we propose TBL to represent the global illumination of large-scale indoor scenes (Sec. 3.1). In order to improve the optimizing efficiency and quality of re-rendered images in the material estimation stage, we adopt a hybrid lighting representation based on our TBL (Sec. 3.2). To reconstruct physically-reasonable SVBRDFs, we present a segmentation-based three-stage material estimation strategy (Sec. 3.3), which can handle ambiguity of materials in complex large-scale indoor scenes very well.

3.1. Texture-based Lighting

We address the issue of how to represent the illumination of a large-scale indoor scene with TBL. The advantages of TBL respectively are the compactness of neural representation, the global illumination of IBL, and the interpretability and spatial consistency of parametric lights.

The proposed TBL, which is a global representation for the entire scene, defines the outgoing radiance for all surface points. We assume that only the diffuse lighting exists in the scene since the diffuse lighting often dominates the scene, similar to radiosity-based methods [48, 49, 53]. Therefore, the outgoing radiance of one point is typically equal to the value of the HDR texture, *i.e.*, the observed HDR radiance of corresponding pixels in input HDR images.

We initially reconstruct the mesh model of a entire large-scale scene with off-the-shelf classical MVS techniques, *e.g.*, colmap [41]. Finally, we reconstruct the HDR texture based on the input HDR images. Therefore, the global illumination is queried from any direction at any position through the HDR texture, as shown in Fig. 3.

3.2. Hybrid Lighting Representation

According to our proposed TBL, the render equation [21] can be rewritten as follow:

$$L_o(x, \omega_o) = \int_{H^+} f_r(x, \omega_i, \omega_o) Q(x, \omega_i, G, T_{hdr})(\omega_i \cdot n) d\omega_i \tag{1}$$

where H^+ denotes hemisphere; x denotes a surface point; ω_i denotes inverse incident light direction; ω_o denotes view direction; n denotes normal; f_r denotes the BRDF of point x; Q denotes the HDR lighting of the intersection point between the known geometry G and the ray $r(t) = x + t\omega_i$, queried by HDR textures T_{hdr}.

In practice, we calculate the Monte Carlo numerical integration with importance sampling [23]. To decrease the variance, a large sample number seems to be inevitable, which will significantly increase computation cost and memory cost in the optimization process. Inspired by precomputed radiance transfer [19], we precompute irradiance of surface points for the diffuse component. Therefore, the irradiance is queried efficiently without expensive online computation as shown in Fig. 3 (right) [48]. The Eq.1 can be rewritten as:

$$L_o(x, \omega_o) = L_d(x, \omega_o) + L_s(x, \omega_o) \tag{2}$$

where L_d denotes the diffuse component and L_s denotes the specular component. The f_r of L_s would be modified as f_s. The detailed formulation can be found in the supplementary material.

We propose two representations to model precomputed irradiance. One is a **n**eural **ir**radiance **f**ield (NIrF), a shallow Multi-Layer-Perceptrons (MLP). It takes a surface point p as input and outputs the irrdiance of p. Another one is a **ir**radiance **t**exture (IrT), similar to the light map widely used in computer graphics. Based on such hybrid lighting representation consists of precomputed irradiance for the diffuse component and source TBL for the specular component, the rendering noise significantly decrease and the materials are optimized efficiently. Therefore, the diffuse component of Eq. 2 can be modeled as Eq. 3.

$$L_d(x, \omega_o) = f_d(x) Ir(x) \tag{3}$$

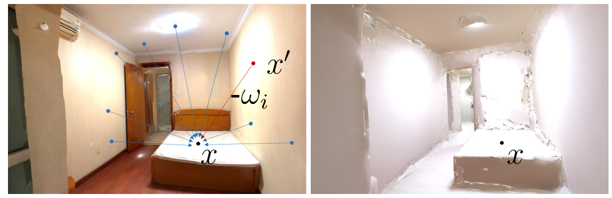

Figure 3. **Visualization of TBL (left) and precomputed irradiance (right).** For any surface point x, the incident radiance from direction $-\omega_i$ can be queried from the HDR texture of the point x', which is the intersection point between the geometry and the ray $r(t) = x + t\omega_i$. The irradiance can be directly queried from the precomputed irradiance of x via NIrF or IrT.

where f_d denotes the diffuse BRDF of point x and Ir denotes the irradiance of point x.

3.3. Segmentation-based Material Estimation

Instead of optimizing neural material [7, 43, 50, 51, 56–58], which is hard to model a large-scale scene with extremely complex materials and is mismatched to the traditional graphics engines, we directly optimize explicit material textures of the geometry.

We use the simplified Disney BRDF model [9] with Spatially-Varying (SV) albedo and SV roughness as parameters. Optimizing explicit material textures straightforwardly leads to inconsistent and unconverged roughness due to sparse observations [58]. We address this problem by leveraging the priors of semantics and room segmentation. The semantic images are predicted by learning-based models [11] and room segmentation is calculated by the occupancy grid [15]. Our segmentation-based strategy has three phases. Details of each stage are described in the following subsections.

Stage I: albedo initialization. We optimize a coarse albedo based on Lambertian assumption instead of initializing the albedo as a constant, which is widely used in object-centric scenes [32]. According to estimated illumination in Sec. 3.2, we can directly calculate the albedo through Eq. 3. However, it recovers over-bright albedo on the highlight regions, which leads to high roughness in the next stage. Therefore, we apply a semantic smoothness constraint to encourage that the albedo/roughness/feature is as close to the mean within the class as possible:

$$\mathcal{L}_{ss} = \sum_c \left| F - \frac{\sum_p F \odot M_{seg}(c)}{\sum_p M_{seg}(c) + \epsilon} \right| \odot M_{seg}(c) \tag{4}$$

where L_{ss} denotes the semantic smoothness loss; c denotes one of the classes of semantics; F denotes the feature image to be smoothed and we use the image-space diffuse albedo A in this stage; p denotes the pixel of F; $\odot$ is an element-wise product; M_{seg} denotes the mask of semantic segmen-

Table 1. **Quantitative comparison on our synthetic dataset.** Our method significantly outperforms the state-of-the-arts in roughness estimation. NeILF∗ [51] denotes source method with their implicit lighting representation.

Method	Albedo			Roughness			Novel view synthesis			Re-rendering		
	PSNR↑	SSIM↑	MSE↓	PSNR↑	SSIM↑	MSE↓	PSNR↑	SSIM↑	MSE↓	PSNR↑	SSIM↑	MSE↓
PhyIR [29]	11.9726	0.6880	0.0635	12.5468	0.7671	0.0556	-	-	-	-	-	-
InvRender [58]	16.9760	0.6305	0.0201	9.1806	0.4787	0.1208	22.2771	0.7826	0.0059	24.2851	0.7834	0.0037
NVDIFFREC [34]	**21.2551**	0.8100	**0.0075**	7.6269	0.1348	0.1727	23.4959	0.9019	0.0045	29.7279	0.9323	0.0011
NeILF∗ [51]	14.2137	0.5184	0.0379	11.5778	0.5974	0.0695	22.3765	0.7598	0.0058	25.1092	0.7654	0.0031
NeILF [51]	17.0707	0.6489	0.0196	11.1654	0.7099	0.0765	22.0703	0.7823	0.0062	24.4710	0.7857	0.0036
Ours	20.4169	**0.8514**	0.0091	**20.2132**	**0.9161**	**0.0095**	**25.0462**	**0.9264**	**0.0031**	**34.2669**	**0.9635**	**0.0004**

tation; ϵ denotes a tiny number. The coarse albedo is optimized by:

$$\mathcal{L}_{albedo} = |I - L_d| + \beta_{ssa}\mathcal{L}_{ss} \tag{5}$$

where β_{ssa} denotes the weight of semantic smoothness loss for albedo, and I denotes the input HDR images.

Stage II: VHL-based sampling and semantics-based propagation. In multi-view images, only sparse specular cues are observed, which lead to globally inconsistent roughness [58], especially for large-scale scenes. As shown in Baseline of Fig. 9 and NVDIFFREC [34] of Fig. 6, only the roughness of highlight regions is optimized reasonably. Therefore, by leveraging the prior of semantic segmentation, the reasonable roughness of highlight regions would be propagated into the regions with same semantics.

We first render images based on input poses with roughness 0.01 to find **v**irtual **h**igh**l**ight (VHL) regions for each semantic class. Then, we optimize the roughness on these VHL regions according to the *frozen* coarse albedo and illumination. Meanwhile, the reasonable roughness can be propagated into the same semantic segmentation through Eq.6:

$$\mathcal{L}_{sp} = \sum_c \left| R - quantile(R \odot M_{vhl}(c), q) \right| \odot (M_{seg}(c) - M_{vhl}(c)) \tag{6}$$

where L_{sp} denotes the semantics-based propagation loss; R denotes the image-space roughness; $quantile$ denotes the q-th quantiles on VHL regions and we set the q as 0.4 for robustness; M_{vhl} denotes the mask of VHL regions for each class. The roughness can be optimized by:

$$\mathcal{L}_{roughness} = |I - L_o| + \beta_{sp}\mathcal{L}_{sp} \tag{7}$$

Stage III: Segmentation-based fine-tuning. In the last phase, we fine-tune all material textures based on the priors of semantic segmentation and room segmentation. Specifically, we apply a similar smoothness constraint to Eq. 4 and a room smoothness constraint for roughness, which makes the roughness of different rooms smoother in a soft manner. The room smoothness constraint is formulated by Eq. 8:

$$\mathcal{L}_{rs} = \sum_c \left| R - \frac{\sum_p R \odot M_{room}(c)}{\sum_p M_{room}(c) + \epsilon} \right| \odot M_{room}(c) \tag{8}$$

where L_{rs} denotes the room segmentation-based smoothness loss; M_{room} denotes the mask of different room segmentation; c denotes one of the index of each room.

We do not apply any smoothness constraint for albedo. The total loss is defined as:

$$\mathcal{L}_{all} = |I - L_o| + \beta_{ssr}(\mathcal{L}_{ss} + \mathcal{L}_{rs}) \tag{9}$$

where β_{ssr} denotes the weight of segmentation smoothness loss for roughness and we use the image-space roughness R in $\mathcal{L}_{ss}$ in Stage III.

4. Experiments

4.1. Datasets

Synthetic dataset. We create a synthetic scene with diverse material and light sources with a path tracer [26]. We render 24 views for optimization and 14 views as novel views, and render Ground Truth material images for each view. The details of the scene can be found in the supplementary material.

Real dataset. Widely used real datasets of large-scale scenes, *e.g.*, Scannet [12], Matterport3D [10] and Replica [44], lack full-HDR images. Therefore, we collect 10 full-HDR real scenes. For each scene, 10 to 20 full-HDR panoramic images are captured by merging 7 bracketed exposures (from $\frac{1}{25000}$s to $\frac{1}{8}$s).

4.2. Baselines and Metrics

To our best knowledge, there are only a few methods that recover the SVBRDFs from multi-view images for large-scale scenes. We compare with the following inverse rendering approaches: (1) The state-of-the-art single image learning-based method: PhyIR [29]; (2) The state-of-the-art multi-view object-centric neural rendering methods: InvRender [58], NVDIFFREC [34] and NeILF [51]. Please note that these object-centric approaches are unsuitable to

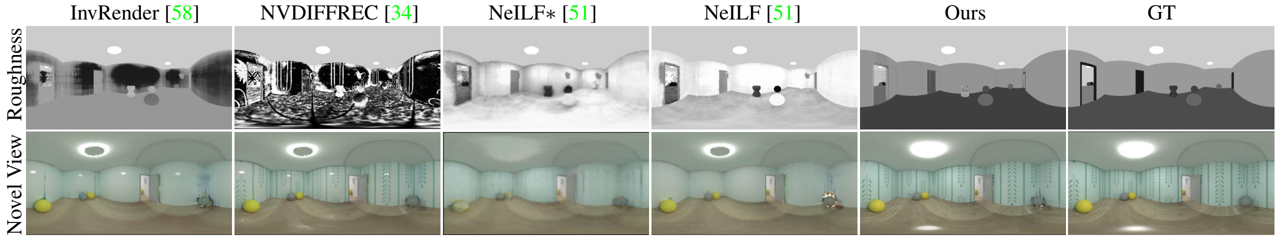

Figure 4. **Qualitative comparison on synthetic dataset.** Our method is able to produce realistic specular reflectance. NeILF∗ [51] denotes source method with their implicit lighting representation.

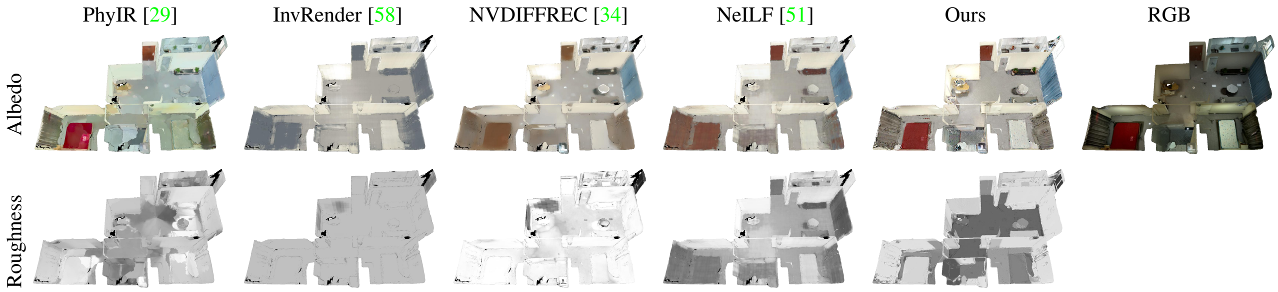

Figure 5. **Qualitative comparison in the 3D view on challenging real dataset.** This sample is Scene 1. Our method reconstructs globally-consistent and physically-reasonable SVBRDFs while other approaches struggle to produce consistent results and disentangle ambiguity of materials. Note that the low roughness (around 0.15 in ours) leads to the strong highlights, which are similar to GT.

evaluate on large-scale scenes due to simple illumination representation except for NeILF. For fair comparisons, we integrate their material optimization strategies with our hybrid lighting representation, which is designed specifically for large-scale real-world scenes. [1]

We use Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM) [46] and Mean Squared Error (MSE) to evaluate the material predictions and the re-rendered images for quantitative comparisons. Moreover, we use Mean Absolute Error (MAE) and SSIM to evaluate the relighting images rendered by different lighting representations.

4.3. Comparisons

Evaluation on synthetic dataset. As shown in Tab. 1 and Fig. 4. Our method significantly outperforms the state-of-the-arts in roughness estimation and our roughness is able to produce physically-reasonable specular reflectance. Moveover, NeILF [51] with our hybrid lighting representation successfully disentangles the ambiguity between material and lighting, compared to their implicit representation.

Evaluation on real dataset. More importantly, we conduct the experiment on our challenging real dataset containing complex materials and illumination. The quantitative comparison in Tab. 2 shows our approach outperforms previous methods. Although these methods have approximate re-rendering error, only our proposed approach disentangles globally-consistent and physically-reasonable materials. We show the qualitative comparison in the 3D view in Fig. 5 and the 2D image views in Fig. 6. PhyIR [29] suffers from poor generalization performance due to a large domain gap and fails in globally-consistent predictions. InvRender [58], NVDIFFREC [34] and NeILF [51] produce blur predictions with artifacts and struggle to disentangle correct materials. Although NVDIFFREC [34] reaches similar performance to our method in Tab. 2, it fails to disentangle the ambiguity between albedo and roughness, *e.g.*, highlights in the specular component are recovered incorrectly as diffuse albedo.

Table 2. **Quantitative comparison of re-rendered images on our real dataset.**

Method	PSNR↑	SSIM↑	MSE↓
InvRender [58]	21.9993	0.7668	0.0065
NVDIFFREC [34]	23.7464	0.8389	0.0044
NeILF [51]	21.9260	0.7687	0.0066
Ours	**24.6093**	**0.8623**	**0.0035**

4.4. Ablation studies

To showcase the effectiveness of proposed lighting representation and material optimization strategy, we ablate the TBL, hybrid lighting representation, albedo initialization in Stage I, VHL-based sampling and semantics-based propagation for roughness estimation in Stage II, and

[1] we tried to compare with optimization-based methods [20, 37], but failed to get available results due to the lack of available source code.

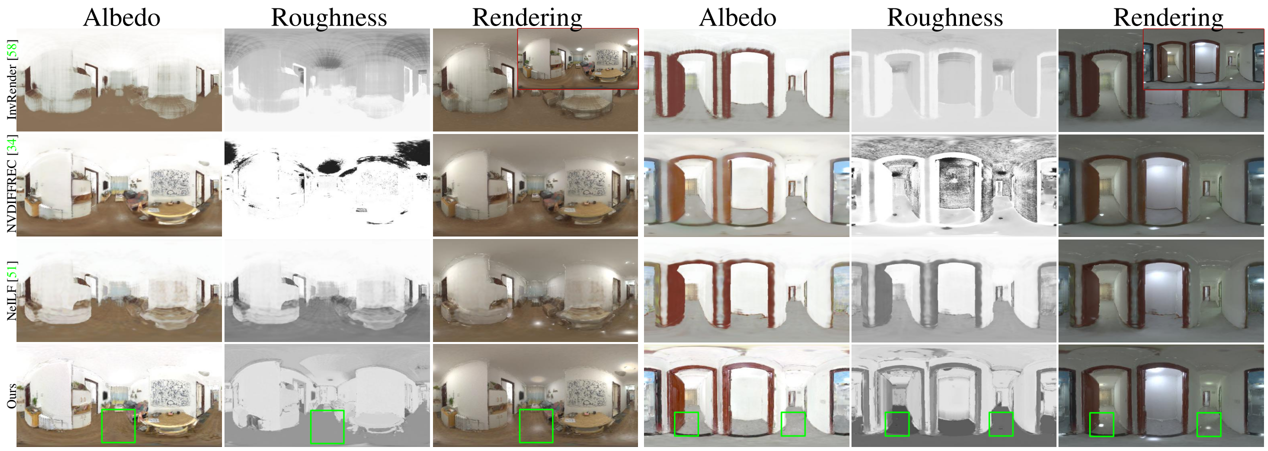

Figure 6. **Qualitative comparison in the image view on challenging real dataset.** From left to right: Scene 8 and Scene 9. Red denotes the Ground Truth image. Our physically-reasonable materials are able to render similar appearance to GT. Note that Invrender [58] and NeILF [51] do not produce correct highlights, and NVDIFFREC [34] fails to distinguish the ambiguity between albedo and roughness.

Table 3. **Ablation study of roughness estimation on synthetic dataset.**

Method	PSNR↑	SSIM↑	MSE↓
Baseline	7.8012	0.52680	0.1659
w/o Stage I	10.3561	0.7044	0.0921
w/o Stage II	7.9627	0.5570	0.1599
w/o Stage III	17.4177	0.8347	0.0181
Ours	**20.2132**	**0.9161**	**0.0095**

Table 4. **Quantitative comparison of relighting spheres.** We use 5th order SH coefficients and 12 SG lobes for the comparison.

Lighting	Diffuse		Matte Sliver		Mirror Sliver	
	MAE↓	SSIM↑	MAE↓	SSIM↑	MAE↓	SSIM↑
SH	0.0602	0.9982	0.0811	0.9977	0.1083	0.9801
SG	0.0027	**0.9995**	0.0054	0.9992	0.0348	0.9815
TBL	**0.0021**	0.9994	**0.0028**	**0.9992**	**0.0055**	**0.9984**

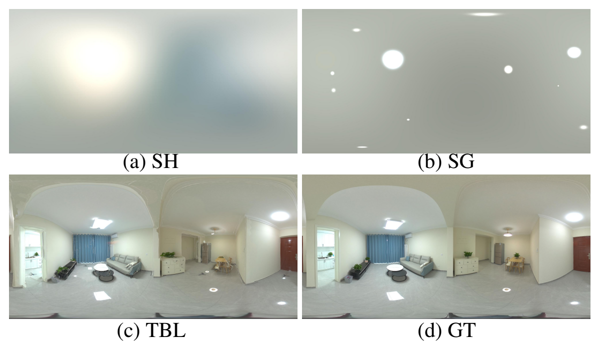

Figure 7. **Comparison of different lighting representations.** The result of TBL is reprojected from 3D mesh. The proposed TBL exhibits high fidelity both in low-frequency and high-frequency.

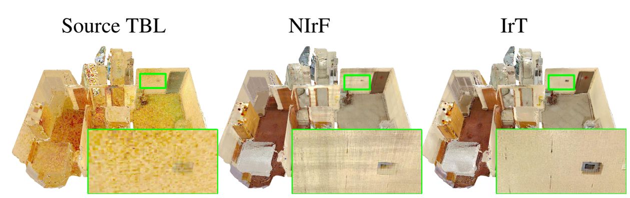

Figure 8. **Ablation study of hybrid lighting representation.** This sample is Scene 11. Note that we decrease the resolution of input images and samples in Source TBL for keeping almost same time cost.

segmentation-based fine-tuning in Stage III.

Effectiveness of TBL. We compare the proposed TBL to SH lighting and SG lighting widely used in previous methods [18, 26, 47, 59]. As shown in Fig. 7, our TBL exhibits high fidelity in high-frequency features. Moreover, we evaluate the relighting error of three re-rendering virtual spheres rendered by different lighting representations in Tab. 4. Except for accuracy, the TBL only costs around 20 MB storage to represent illumination while the dense grid-based VSG lighting [47] costs around 1 GB storage and the sparse grid-based SH lighting, Plenoxels [39], costs around 750 MB storage. Therefore, our TBL achieves improved accuracy while being compact in storage.

Effectiveness of Hybrid Lighting Representation. We compare the hybrid lighting representation described in Sec. 3.2 to source TBL. As shown in Fig. 8, Without hybrid lighting representation, the albedo leads to noise and converges slowly. With precomputed irradiance, we can use high resolution inputs to recover detailed materials, and significantly accelerate the optimization process. The IrT produces more detailed and artifacts-free albedo, compared to the NIrF. Furthermore, we also compare to implicit lighting in Tab. 1 and Fig. 4. NeILF [51] with their implicit lighting

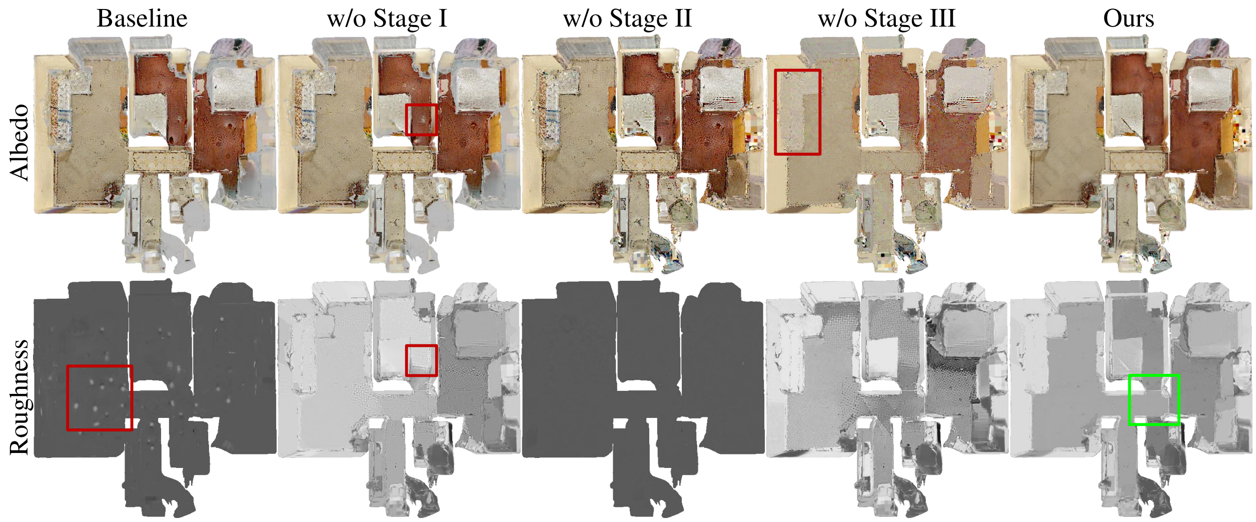

Figure 9. **Ablation study of our material optimization strategy in the 3D mesh view on challenging real dataset.** This sample is Scene 11. In baseline, we jointly optimize albedo and roughness.

fails in disentangling the ambiguity between materials and lighting, *e.g.*, The lighting effects are incorrectly recovered as materials in Fig. 4.

Effectiveness of the Three Stage Strategy. The results are shown in Tab. 3 and Fig. 9. The roughness of baseline fails to converge and only the highlight regions are updated. Without albedo initialization in Stage I, albedo in highlight regions is over-bright and leads to incorrect roughness. VHL-based sampling and semantics-based propagation in Stage II is crucial to recover the reasonable roughness of areas where highlights are not observed. Segmentation-based fine-tuning in Stage III produces detailed albedo, makes final roughness smoother and prevent the wrong propagation of roughness between different materials.

4.5. Applications

Our final output is a triangle mesh with PBR material textures, which is compatible with standard graphic engines and 3D modeling tools. We demonstrate in Fig. 1 that the proposed approach is able to produce convincing results on material editing, editable novel view synthesis and relighting. Moreover, we show several results of editable novel view synthesis in Fig. 10. Note that the view-dependent specular highlights reasonably change as view changes. See the supplementary material for more results.

5. Conclusion

In this paper, we propose a novel inverse rendering framework that recovers globally-consistent lighting and materials from posed sparse-view images and geometry for large-scale scenes. Our texture-based lighting, which not only represents infinite-bounce global illumination but also is compact and globally-consistent, is suitable for modelling illumination of large-scale scenes. Our material optimization strategy leveraging semantics and room segmentation priors is able to reconstruct physically-reasonable and globally-consistent PBR materials. Such a triangle mesh with material textures is compatible with common graphic engines, which benefits several downstream applications such as material editing, editable novel view synthesis and relighting.

Figure 10. **Editable novel view synthesis.** In Scene 8, we edit the albedo of the wall, and edit the roughness of the floor. In Scene 9, we edit the albedo of the floor and the wall. Our method produces convincing results (see the lighting effects in the floor and wall).

Acknowledgements

We thank reviewers for their constructive comments, and thank Kunlong Li and Wei Li for generating annotations. This work is supported in part by the National Natural Science Foundation of China (NFSC) (No. 62002295) and Shanxi Provincial Key R&D Program (No. 2021KWZ-03).

References

[1] Dejan Azinovic, Tzu-Mao Li, Anton Kaplanyan, and Matthias Nießner. Inverse path tracing for joint material and lighting estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*, 2019. 2

[2] Jonathan T Barron and Jitendra Malik. Intrinsic scene properties from a single rgb-d image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*, pages 17–24, 2013. 2, 3

[3] Jonathan T. Barron and Jitendra Malik. Shape, illumination, and reflectance from shading. *IEEE Transactions on Pattern Analysis and Machine Intelligence.*, 37(8):1670–1687, 2015. 2, 3

[4] Sean Bell, Kavita Bala, and Noah Snavely. Intrinsic images in the wild. *ACM Transactions on Graphics.*, 33(4), 2014. 3

[5] Sai Bi, Nima Khademi Kalantari, and Ravi Ramamoorthi. Deep Hybrid Real and Synthetic Training for Intrinsic Decomposition. In Wenzel Jakob and Toshiya Hachisuka, editors, *Eurographics Symposium on Rendering.* The Eurographics Association, 2018. 3

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

[7] Mark Boss, Raphael Braun, Varun Jampani, Jonathan T. Barron, Ce Liu, and Hendrik P.A. Lensch. Nerd: Neural reflectance decomposition from image collections. In *Proceedings of the IEEE International Conference on Computer Vision.*, 2021. 1, 3, 4

[8] Mark Boss, Varun Jampani, Kihwan Kim, Hendrik P.A. Lensch, and Jan Kautz. Two-shot spatially-varying brdf and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*, 2020. 3

[9] Brent Burley. Physically-based shading at disney. 2012. 4

[10] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *International Conference on 3D Vision.*, 2017. 5

[11] MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation, 2020. 4

[12] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*, 2017. 5

[13] 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*, page 32, 2008. 1, 2

[14] Valentin Deschaintre, Miika Aittala, Frédo Durand, George Drettakis, and Adrien Bousseau. Flexible svbrdf capture with a multi-image deep network. *Computer Graphics Forum.*, 38(4), July 2019. 2

[15] Alberto Elfes. Using occupancy grids for mobile robot perception and navigation. *Computer*, 22:46–57, 1989. 4

[16] Marc-André Gardner, Yannick Hold-Geoffroy, Kalyan Sunkavalli, Christian Gagné, and Jean-François Lalonde. Deep parametric indoor lighting estimation. In *Proceedings of the IEEE International Conference on Computer Vision.*, pages 7175–7183, 2019. 1

[17] Marc-André Gardner, Kalyan Sunkavalli, Ersin Yumer, Xiaohui Shen, Emiliano Gambaretto, Christian Gagné, and Jean-François Lalonde. Learning to predict indoor illumination from a single image. *ACM Transactions on Graphics. (Proceedings of SIGGRAPH Asia.)*, 9(4), 2017. 1

[18] Mathieu Garon, Kalyan Sunkavalli, Sunil Hadap, Nathan Carr, and Jean-Francois Lalonde. Fast spatially-varying indoor lighting estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*, June 2019. 1, 7

[19] Gene Greger, Peter Shirley, Philip M. Hubbard, and Donald P. Greenberg. The irradiance volume. *IEEE Computer Graphics and Applications*, 18:32–43, 1998. 4

[20] B. Haefner, S. Green, A. Oursland, D. Andersen, M. Goesele, D. Cremers, R. Newcombe, and T. Whelan. Recovering real-world reflectance properties and shading from HDR imagery. In *International Conference on 3D Vision.*, London, UK, December 2021. 3, 6

[21] James T Kajiya. The rendering equation. In *ACM Transactions on Graphics.*, volume 20, pages 143–150, 1986. 4

[22] Kaizhang Kang, Cihui Xie, Chengan He, Mingqi Yi, Minyi Gu, Zimin Chen, Kun Zhou, and Hongzhi Wu. Learning efficient illumination multiplexing for joint capture of reflectance and shape. *ACM Transactions on Graphics.*, 38(6), Nov. 2019. 2

[23] Brian Karis and Epic Games. Real shading in unreal engine 4, 2013. 4

[24] Junxuan Li and Hongdong Li. Neural reflectance for shape recovery with shadow handling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*, pages 16221–16230, 2022. 2

[25] Junxuan Li, Hongdong Li, and Yasuyuki Matsushita. Lighting, reflectance and geometry estimation from 360° panoramic stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*, 2021. 2

[26] Zhengqin Li, Mohammad Shafiei, Ravi Ramamoorthi, Kalyan Sunkavalli, and Manmohan Chandraker. Inverse rendering for complex indoor scenes: Shape, spatially-varying lighting and svbrdf from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*, pages 2475–2484, 2020. 1, 2, 3, 5, 7

[27] Zhengqin Li, Jia Shi, Sai Bi, Rui Zhu, Kalyan Sunkavalli, Miloš Hašan, Zexiang Xu, Ravi Ramamoorthi, and Manmohan Chandraker. Physically-based editing of indoor scene lighting from a single image. In *Proceedings of the European Conference on Computer Vision.*, 2022. 1, 2

[28] Zhengqi Li and Noah Snavely. Cgintrinsics: Better intrinsic image decomposition through physically-based rendering. In *Proceedings of the European Conference on Computer Vision.*, 2018. 3

[29] Zhen Li, Lingli Wang, Xiang Huang, Cihui Pan, and Jiaqi. Yang. Phyir: Physics-based inverse rendering for panoramic indoor images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*, 2022. 1, 2, 3, 5, 6

[30] Zhengqin Li, Zexiang Xu, Ravi Ramamoorthi, Kalyan Sunkavalli, and Manmohan Chandraker. Learning to reconstruct shape and spatially-varying reflectance from a single image. In *ACM Transactions on Graphics. (Proceedings of SIGGRAPH Asia.)*, page 269, 2018. 3

[31] Zhengqin Li, Ting-Wei Yu, Shen Sang, Sarah Wang, Meng Song, Yuhan Liu, Yu-Ying Yeh, Rui Zhu, Nitesh Gundavarapu, Jia Shi, Sai Bi, Hong-Xing Yu, Zexiang Xu, Kalyan Sunkavalli, Milos Hasan, Ravi Ramamoorthi, and Manmohan Chandraker. 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, June 2021. 3

[32] Fujun Luan, Shuang Zhao, Kavita Bala, and Zhao Dong. Unified shape and svbrdf recovery using differentiable monte carlo rendering. *Computer Graphics Forum.*, 40, 2021. 1, 2, 3, 4

[33] 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. In *Proceedings of the European Conference on Computer Vision.*, 2020. 2

[34] Jacob Munkberg, Jon Hasselgren, Tianchang Shen, Jun Gao, Wenzheng Chen, Alex Evans, Thomas Mueller, and Sanja Fidler. Extracting triangular 3d models, materials, and lighting from images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*, 2022. 1, 3, 5, 6, 7

[35] Giljoo Nam, Joo Ho Lee, Diego Gutierrez, and Min H. Kim. Practical svbrdf acquisition of 3d objects with unstructured flash photography. *ACM Transactions on Graphics. (Proceedings of SIGGRAPH Asia.)*, 37(6):267:1–12, 2018. 2

[36] Thomas Nestmeyer, Jean-FranÃ§ois Lalonde, Iain Matthews, and Andreas M Lehrmann. Learning physics-guided face relighting under directional light. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*, 2020. 2

[37] Merlin Nimier-David, Zhao Dong, Wenzel Jakob, and Anton Kaplanyan. Material and Lighting Reconstruction for Complex Indoor Scenes with Texture-space Differentiable Rendering. In *Eurographics Symposium on Rendering.*, 2021. 2, 3, 6

[38] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M. Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *Proceedings of the IEEE International Conference on Computer Vision.*, 2021. 3

[39] Sara Fridovich-Keil and 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.*, 2022. 7

[40] Carolin Schmitt, Simon Donne, Gernot Riegler, Vladlen Koltun, and Andreas Geiger. On joint estimation of pose, geometry and svbrdf from a handheld scanner. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*, June 2020. 3

[41] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *Proceedings of the European Conference on Computer Vision.*, 2016. 4

[42] 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 International Conference on Computer Vision.*, 2019. 1, 3

[43] Pratul P. Srinivasan, Boyang Deng, Xiuming Zhang, Matthew Tancik, Ben Mildenhall, and Jonathan T. Barron. Nerv: Neural reflectance and visibility fields for relighting and view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*, 2021. 1, 2, 3, 4

[44] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, Brian Budge, Yajie Yan, Xiaqing Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesus Briales, Tyler Gillingham, Elias Mueggler, Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke M. Strasdat, Renzo De Nardi, Michael Goesele, Steven Lovegrove, and Richard Newcombe. The Replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 5

[45] Huan Wang, Jian Ren, Zeng Huang, Kyle Olszewski, Menglei Chai, Yun Fu, and Sergey Tulyakov. R2l: Distilling neural radiance field to neural light field for efficient novel view synthesis. In *Proceedings of the European Conference on Computer Vision.*, 2022. 2

[46] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing.*, 13(4):600–612, 2004. 6

[47] Zian Wang, Jonah Philion, Sanja Fidler, and Jan Kautz. Learning indoor inverse rendering with 3d spatially-varying lighting. In *Proceedings of the IEEE International Conference on Computer Vision.*, 2021. 1, 2, 3, 7

[48] Gregory J. Ward, Francis M. Rubinstein, and Robert D. Clear. A ray tracing solution for diffuse interreflection. *Computer Graphics*, 22(4), 1988. 4

[49] Daniel N. Wood, Daniel I. Azuma, Ken Aldinger, Brian Curless, Tom Duchamp, David H. Salesin, and Werner Stuetzle. Surface light fields for 3d photography. SIGGRAPH '00, page 287–296, 2000. 4

[50] Wenqi Yang, Guanying Chen, Chaofeng Chen, Zhenfang Chen, and Kwan-Yee K. Wong. Ps-nerf: Neural inverse rendering for multi-view photometric stereo. In *Proceedings of the European Conference on Computer Vision.*, 2022. 1, 4

[51] Yao Yao, Jingyang Zhang, Jingbo Liu, Yihang Qu, Tian Fang, David McKinnon, Yanghai Tsin, and Long Quan. Neilf: Neural incident light field for physically-based mate-

rial estimation. In *Proceedings of the European Conference on Computer Vision.*, 2022. 1, 2, 4, 5, 6, 7

[52] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. *Advances in Neural Information Processing Systems.*, 33, 2020. 2

[53] Yizhou Yu, Paul Debevec, Jitendra Malik, and Tim Hawkins. Inverse global illumination: Recovering reflectance models of real scenes from photographs. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '99, page 215–224, 1999. 4

[54] Edward Zhang, Michael F. Cohen, and Brian Curless. Emptying, refurnishing, and relighting indoor spaces. *ACM Transactions on Graphics. (Proceedings of SIGGRAPH Asia.)*, 35(6), 2016. 2, 3

[55] 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.*, 2022. 1, 2, 3

[56] 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.*, 2021. 2, 3, 4

[57] Xiuming Zhang, Pratul P Srinivasan, Boyang Deng, Paul Debevec, William T Freeman, and Jonathan T Barron. NeRFactor: Neural Factorization of Shape and Reflectance Under an Unknown Illumination. *ACM Transactions on Graphics.*, 2021. 1, 3, 4

[58] Yuanqing Zhang, Jiaming Sun, Xingyi He, Huan Fu, Rongfei Jia, and Xiaowei Zhou. Modeling indirect illumination for inverse rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*, 2022. 1, 2, 3, 4, 5, 6, 7

[59] Hao Zhou, Xiang Yu, and David W. Jacobs. Glosh: Global-local spherical harmonics for intrinsic image decomposition. In *Proceedings of the IEEE International Conference on Computer Vision.*, October 2019. 1, 2, 7

[60] 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, June 2022. 3