PhyIR: Physics-based Inverse Rendering for Panoramic Indoor Images

Zhen Li[1] Lingli Wang[1] Xiang Huang[1] Cihui Pan[1,*] Jiaqi Yang[2,*]
[1]Realsee [2]Northwestern Polytechnical University
yodlee@mail.nwpu.edu.cn,{wanglingli008,huangxiang003,pancihui001}@ke.com,jqyang@nwpu.edu.cn

Abstract

Inverse rendering of complex material such as glossy, metal and mirror material is a long-standing ill-posed problem in this area, which has not been well solved. Previous approaches cannot tackle them well due to simplified BRDF and unsuitable illumination representations. In this paper, we present PhyIR, a neural inverse rendering method with a more completed SVBRDF representation and a physics-based in-network rendering layer, which can handle complex material and incorporate physical constraints by re-rendering realistic and detailed specular reflectance. Our framework estimates geometry, material and Spatially-Coherent (SC) illumination from a single indoor panorama. Due to the lack of panoramic datasets with completed SVBRDF and full-spherical light probes, we introduce an artist-designed dataset named FutureHouse with high-quality geometry, SVBRDF and per-pixel Spatially-Varying (SV) lighting. To ensure the coherence of SV lighting, a novel SC loss is proposed. Extensive experiments on both synthetic and real-world data show that the proposed method outperforms the state-of-the-arts quantitatively and qualitatively, and is able to produce photorealistic results for a number of applications such as dynamic virtual object insertion.

Figure 1. Given an LDR panorama (a), we estimate geometry (b-c), SV illumination and SVBRDFs, including base color (d), roughness (e) and metalness (f). Our physics-based differentiable renderer can produce detailed specular reflectance (g) on complex material. Based on such physical constraint, our predictions are qualified to produce virtual object insertion (h) with realistic lighting effects, *e.g.*, highlight caused by unseen light source in ① and specular reflectance on the cabinet in ②.

1. Introduction

Inverse rendering is a fundamental yet challenging task in computer vision and computer graphics. This task aims to recover geometry, material and illumination from a single image. The above properties play a vital role in emerging applications, such as scene editing and virtual object insertion in mixed reality. All these applications require physically reasonable realism. However, reconstructing physically accurate properties of a scene is very difficult, because inverse rendering is an ill-posed problem. It contains complicated geometry, different types of material and varying local illumination, which will result in complex lighting effects, *e.g.*, specular reflectance on glossy and mirror material, inter-reflection and cast shadows.

There are three main challenges in solving this task physically. 1) ***Complex material is difficult to model.*** Most existing methods assume that all surfaces are Lambertian [2, 21, 27, 30, 34–37, 40, 46, 56] and only produce diffuse reflectance. Some methods handle specular reflectance in an unphysical way, such as neural residual renderer [39], additional specular shading [48] and phong parameters [17]. Although some approaches use a relatively physical BRDF representation [7, 29, 31, 52], complex material, *e.g.*, glossy, metal and mirror material, still cannot be handled well due to limited BRDF. Moreover, since the re-renderer is built on such a limited BRDF, physical constraints cannot be incorporated well. 2) ***Changeable local illumination is difficult to represent.*** The illumination of indoor scene is spatially-varying (SV) because of occlusion and non-uniform light distribution [16], and is also spatially-coherent (SC) due to coherent variability. Most approaches fail to ensure coher-

*Co-corresponding authors. The project page is at https://lzleejean.github.io/PhyIR ence [16, 29, 36, 41, 55, 56], which leads to flickering results for dynamic object insertion. Although the projection-based method [27] and uniform volumetric lighting representation [43, 46] are used to alleviate this issue, they are not easily incorporated into a physics-based framework due to non-differentiable or memory-hungry. 3) ***The lack of high-quality datasets containing comprehensive labels.*** Collecting ground truth (GT) labels from real-world images is difficult at scale. Moreover, some properties are quite difficult to measure. Meanwhile, recently used synthetic datasets [28, 54] lack necessary properties, *e.g.*, HDR lighting and essential material; the dataset proposed by OpenRooms [32] consists of SVBRDFs used in InvIndoor [29] and hemisphere lighting.

Motivated by these concerns, we propose PhyIR, an end-to-end neural inverse rendering framework with a more completed SVBRDF representation and a physics-based in-network rendering layer, as shown in Figure 1. We tackle the aforementioned three challenges from the following perspectives. 1) We present a more physical inverse rendering model without Lambertian assumption. It can process specular reflectance well for glossy, metal and even mirror material; it provides physics-based constraints, which can significantly assist the optimization of components. 2) A novel SC loss is proposed to ensure the consistence of neighboring SV light probes, which provides an overall constraint on per-pixel lighting of the whole scene to avoid mutation. 3) With great efforts, we generate a large-scale photorealistic panorama dataset with high-quality depth, normal, per-pixel illumination and comprehensive SVBRDFs, namely, base color, roughness and metalness. Thanks to physics-based rendering, there is a smaller divergence between our artist-designed dataset and the real-world data (as detailed in Sec. 3.1).

In summary, the main contributions of our method are as follows:

1. A physics-based inverse rendering framework that can handle complex material, including metal and mirror material.
2. A spatially-coherent loss to guarantee spatial consistence of neighboring per-pixel illumination.
3. A large-scale photorealistic indoor panorama dataset with high-quality depth, normal, SVBRDFs and per-pixel spatially-varying illumination.

2. Related Work

Inverse rendering. Barrow and Tenenbaum [4] first introduced the concept of intrinsic image decomposition, which decomposes image into reflectance and shading. More complex models were subsequently proposed. Barron and Malik [3] proposed an optimization-based method to decompose shape, Lambertian reflectance and single lighting. They estimated SV illumination with RGBD as input in [2]. With great advances in deep learning, researchers began to use neural networks to solve this problem. Janner *et al*. [21] decomposed shading into illumination and normal for object-specific images with self-supervised learning; SfSNet [40] addresses inverse rendering for the human face object. Next some approaches solve indoor scenes [6, 13, 26, 30], but they also only focus on diffuse reflectance. NIR [39] proposes a neural renderer to generate residual appearance, *e.g.*, highlight, but such neural renderer is non-interpretable and not physical; Wei *et al*. [48] decomposed specular shading additionally; Georgoulis *et al*. [17] estimated phong parameters from specular objects. These models are not physical enough. Li *et al*. [31] estimated more physical material called microfacet BRDF from the specific object. InvIndoor [29] is the most similar work to ours, they extended their model proposed in [31] to indoor scenes. However, their method is unable to handle metal material due to limited BRDFs. Furthermore, the physics-based in-network rendering layer in InvIndoor [29] cannot produce detailed specular reflectance. In our work, we leverage more comprehensive BRDFs and an improved physics-based in-network rendering layer to produce detailed specular reflectance on complex material, such as glossy material, metal material and even mirror material.

Lighting estimation. Most existing efforts on lighting estimation [14, 15, 19, 20, 27, 47, 50, 51] only predict a single lighting (always in the center of image) and ignore SV lighting. It will produce unexpected identical results at different locations of an image, especially for indoor scenes. Recent works explore SV lighting [16, 29, 41, 55] by estimating dense and even per-pixel lighting. These methods are good at predicting unobserved light source because they utilize GT local lighting, which can capture all visible illumination at local position. However, they cannot ensure spatial consistency of neighboring illumination due to separate prediction. This inconsistent illumination will produce flickering results for dynamic virtual object insertion. Srinivasan *et al*. [43] proposed a volumetric method to ensure the coherence of neighboring illumination by creating a uniform volumetric grid, but it cannot guarantee physically correct lighting due to missing of HDR. In addition, the projection-based or warp-based method [14, 27, 41] helps the approach ignoring SV lighting to generate SC lighting. However, it needs scene depth as the input, which is not easy to capture in real world.

Our work can ensure spatial consistency of local illumination. We address this problem by designing a spatially-coherent loss for per-pixel illumination representation. It can restrict the variability for neighboring local light probes.

Dataset. Dataset is the basis for learning-based methods.

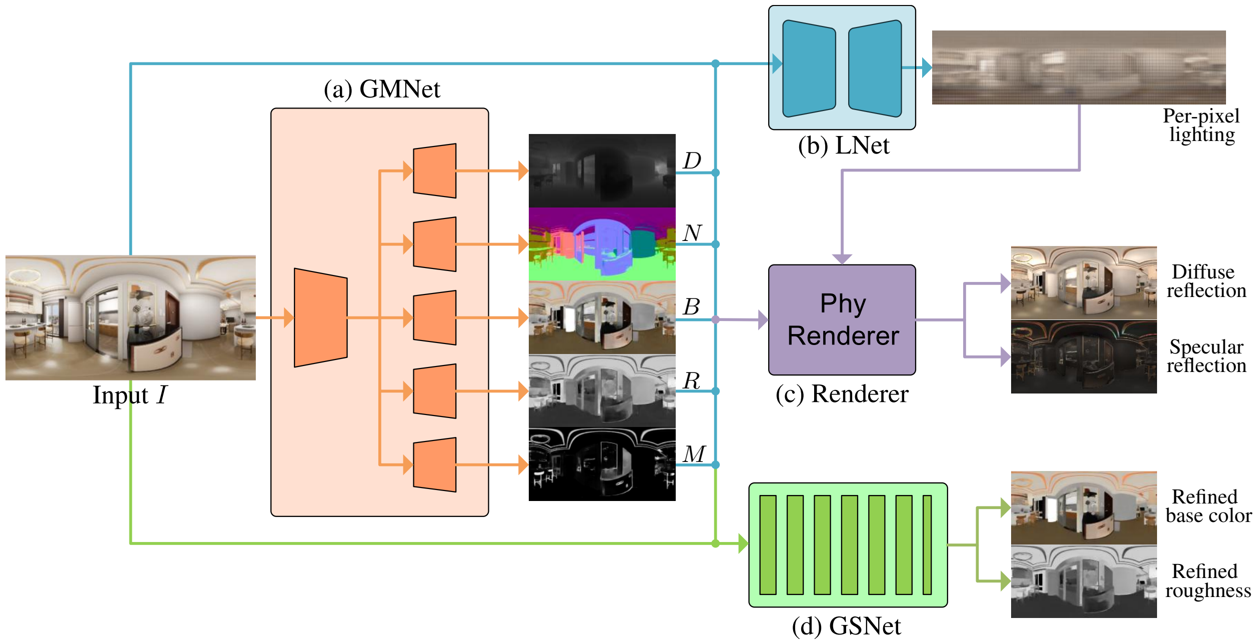

Figure 2. Overview of our physics-based inverse rendering architecture. The framework consists of four modules (a-d). Given an LDR panorama I, geometry and material estimation module (a) first predicts coarse geometry (D, N) and SVBRDF (B, R, M). SC lighting estimation module (b) predicts physically correct per-pixel illumination with physics-based in-network rendering module (c). Finally, trainable guider filter module (d) refines the predicted BRDF.

Current datasets captured in the real-world include scene datasets [1, 9, 11] and lighting datasets [8, 10, 15, 16, 19]. However, these scene datasets do not have essential material and illumination; these lighting datasets lack necessary geometry and material. Virtual datasets play a greater role thanks to controlled rendering. A widely used virtual scene dataset is SUNCG [42], and many methods [16, 30, 39, 53] generate specific training data based on this dataset. Unfortunately, these datasets have unrealistic material (Lambertian or Phong) and lighting configuration [29]. To add insult to injury, these datasets are not available now due to copyright issues. Later, some virtual scene datasets [28, 54] are used for inverse rendering [27, 43, 46], which lack HDR lighting and comprehensive material. Li *et al.* [32] generated a dataset for indoor scenes with HDR lighting and microfacet material, the divergence between rendering data and real-world data still exists due to cheap assets. We build a large-scale photorealistic panorama dataset for indoor scenes, which is generated in Unreal Engine 4 [12] based on professional layout designs and tens of thousands of high-quality models. Our dataset is fully panoramic, therefore it can be used for both perspective image and omnidirectional image tasks. Moreover, we also captured a panoramic HDR illumination dataset from the real-world to evaluate spatially-coherent lighting.

3. Methodology

Our physics-based inverse rendering method aims to recover geometry, complex SVBRDFs and SC illumination from a single indoor panorama. To address this challenging problem, we design a seperate deep model with physics-based constraints. The framework consists of four modules, namely geometry and material estimation module, SC lighting estimation module, physics-based in-network rendering module and trainable guided filter module.

As shown in Figure 2, the geometry and material estimation network first predicts coarse geometry and BRDF from an input image. Then, all of these predictions and input image are fed into SC lighting estimation network to predict per-pixel illumination. The third module provides the most important physical constraints. Finally, the fast trainable guider filter module refines the BRDF to make it smoother.

3.1. The FutureHouse Synthetic Dataset

Capturing the essential BRDF and illumination of real-world scenes is almost impossible. IIW [5] is captured from real-world scenes, but only sparse labels of pairwise reflectance comparison are available. Otherwise, the captured image is not omnidirectional.

Therefore, there is no alternative to render synthetic datasets. The most influential synthetic dataset named

Table 1. The comparison between previous datasets and our proposed *FutureHouse*. Our high-quality dataset contains comprehensive annotations.

	Layout Type	CAD Model	Geometry Annotation	Material Annotation	Lighting Annotation	Light source Annotation	Panorama
InteriorNet [28]	artist-designed	artist-designed	✓	diffuse	shading	✗	✓
Structure3D [54]	artist-designed	artist-designed	✓	diffuse	shading	✗	✓
OpenRooms [32]	auto-generated	scanned	✓	microfacet	per-pixel HDR envmap	✓	✗
FutureHouse	artist-designed	artist-designed	✓	microfacet	per-pixel HDR envmap	✓	✓

SUNCG [42] contains 45,622 houses with 404,058 rooms and 2644 unique objects. Although the render quality is not ideal, many approaches [16, 30, 39, 53] generate their training data with improved rendering methods. However, these methods with Lambertian assumption are not suitable for complex material. InvIndoor [29] represents material with a physically motivated microfacet BRDF model [24]. This representation can handle common material in real-world. Unfortunately, these SUNCG-based datasets are not available now due to copyright issues. Recently, Li *et al*. [32] generated a large-scale dataset for indoor scene with model scanning and material mapping. However, the divergence between rendering data and real-world data still exists due to cheap assets and limited computational budgets.

In this work, we present a new large-scale photorealistic panoramic dataset named *FutureHouse*, which has the following characteristics. 1) It contains over 70,000 high-quality models with high-resolution meshes and physical material. All models are measured in real world standards. 2) Selected scene layouts are carefully designed by over 100 excellent artists. All of selected layouts are used in real-world display. 3) It contains 28,579 good panoramic views from 1,752 house-scale scenes. Therefore, it can be used for perspective image tasks as well as omnidirectional image tasks. 4) More physical material representation. Most materials are represent by microfacet BRDF modeling metalness, and the rest are represent by special shading models, *e.g.*, cloth material and transmission material. 5) High rendering quality. Benefiting from commercial rendering engine, Unreal engine 4 [12], and powerful deep learning super sampling (DLSS) [33], our renderings have less noise. The comparison of characteristics is shown in Table 1 and more comparisons and examples can be found in supplementary material. Our SVBRDF representation including base color and metalness is capable of producing non-monochrome specular reflectance.

Our data will greatly aid research on multiple topics, such as inverse rendering (as well as its sub-tasks, *e.g.*, depth and normal estimation, material estimation, intrinsic image decomposition and lighting estimation) and robotics. *The FutureHouse dataset will be released once the work is published.*

3.2. Network and Loss

As shown in Figure 2, our network consists of four modules, namely geometry and material estimation module, SC lighting estimation module, physics-based differentiable rendering module and trainable guided filter module. Details of each sub-module are shown in the following.

Geometry and material estimation. The geometry and material estimation module aims to predict coarse results of base color ($\tilde{B}$), roughness ($\tilde{R}$), metalness ($\tilde{M}$), normal ($\tilde{N}$) and depth ($\tilde{D}$) from a single LDR panorama (I). To address this multi-task problem, we use a multi-branch encoder-decoder architecture based on ResNet [18] and Unet [38]. The encoder is RetNet-18, and the decoder consists of five convolutional layers with four skip connections. All five branch decoders have same structure except for the output layer. We use Circular Padding (CirP) [45] to extract 3D space features from panoramas. The GMNet can be modeled as Eq. 1:

$$\tilde{N}, \tilde{D}, \tilde{B}, \tilde{R}, \tilde{M} = \mathbf{GMNet}(I). \tag{1}$$

We use L_2 loss for base color and roughness. For metalness estimation, standard L_2 loss makes the training unstable due to the imbalanced value. Therefore, we propose a re-weighting L_2 loss to prevent falling into a local minima predicting zeros. We define the loss as:

$$L_M = \|M - \tilde{M}\|_2^2 \times (2 - \frac{1}{n}\sum \tilde{M}(m)), \tag{2}$$

where M is the GT metalness, $\tilde{M}$ is predicted metalness and m is the index of pixels classified as metal. For depth, we use the popular BerHu loss [25] as objective. For normal, we define the cosine loss as Eq. 3:

$$L_N = \|1 - N^T\tilde{N}\|_1. \tag{3}$$

Because base color, roughness, normal, metalness are piece-wise smooth, we also add gradient loss for them. The gradient loss is :

$$L_g = \|\nabla X - \nabla\tilde{X}\|_1, \tag{4}$$

where ∇X is the gradient of GT base color, roughness, normal and metalness. The final training loss function of GMNet is:

$$\begin{aligned} L_{GM} &= \beta_A L_A + \beta_R L_R + \beta_M L_M \\ &+ \beta_D L_D + \beta_R L_R + \beta_g L_g. \end{aligned} \tag{5}$$

SC lighting estimation. SV lighting is essential for generating different virtual object insertion results at different locations of a scene. The approximate representation used in previous approches [16, 29, 36, 55] cannot model the whole panoramic environment accurately both in low frequency and high frequency. On the one hand, benefited from the 360° input, we can use the source HDR environment map as our lighting representation to avoid ambiguous predictions caused by limited field of view (LFOV) input. On the other hand, the accurate light probe representation is suitable for our proposed SC loss.

Our SC lighting network takes the LDR panorama $I \in \mathbb{R}^{3\times H\times W}$, predicted geometry and material as input ($\tilde{N}$, $\tilde{B} \in \mathbb{R}^{3\times H\times W}$, $\tilde{D}$, $\tilde{R}$, $\tilde{M} \in \mathbb{R}^{1\times H\times W}$). The architecture is similar to InvIndoor [29], a UNet-based network. It predicts per-pixel light probes $L \in \mathbb{R}^{3\times(H\times h)\times(W\times w)}$. The LNet can be modeled as Eq. 6:

$$\tilde{L} = \mathbf{LNet}(I, \tilde{N}, \tilde{D}, \tilde{B}, \tilde{R}, \tilde{M}). \tag{6}$$

We use log-scale loss for HDR light probe due to its high dynamic range.

$$L_L = 0.5 \times (1 - \mathit{SSIM}(L_{log} \odot M_{mask}, \tilde{L}_{log} \odot M_{mask})), \tag{7}$$

where M_{mask} is the mask of object regions except light source and transmission object, L_{log} is the log-scale lighting and $\odot$ is an element-wise product.

Previous methods [16, 29, 55] with the per-pixel illumination fail to consider the coherence of neighboring illumination, thus these approaches produce spatial flickering results in virtual object insertion. Unlike InvIndoor [29] using a hemispherical light representation, we use a full-spherical light representation. This representation allows us to add SC constraints on neighboring light probes. We propose a novel SC loss to impose constraints related to the 3D position on predicted light probes:

$$L_{SC} = \frac{1}{N}\sum(|\mathbf{Warp}(\tilde{L}) - \tilde{L}| \odot e^{\alpha\|\nabla\tilde{D}\|_1}), \tag{8}$$

where $\mathbf{Warp}$ is a projection operator; $\nabla\tilde{D}$ is the gradient of predicted depth. The $\exp$ function re-weights the loss of neighboring light probes according to the gradient of depth. We use α = -5.0 in our model. The $\mathbf{Warp}$ operator is similar with the method proposed by Gardner *et al*. [15], which calculates the panorama of any 3D position from source panorama by projecting and sampling. Our operator is parallel and differentiable, thus it can be easily integrated for training LNet.

Physics-based in-network rendering module. It is known that the re-rendering module is essential to rectify all predictions in inverse rendering. However, previous methods [27, 29, 39, 46, 55] are not able to rectify components in a physically meaningful way causing unreasonable predictions. Therefore, we propose a more physical differentiable in-network rendering module with microfacet BRDF modeling metal material, which can physically re-render realistic reflectance even on complex material. We define our physical rendering function as:

$$\begin{aligned}\tilde{I} = f_d \int_{H^+} L_i(\omega_i)(\omega_i \cdot n)d\omega_i \\ + \int_{H^+} f_s L_i(\omega_i)(\omega_i \cdot n)d\omega_i,\end{aligned} \tag{9}$$

where H^+ denotes hemisphere; L_i denotes illumination; ω_i denotes light direction; n denotes normal; f_d denotes diffuse BRDF and f_s denotes specular BRDF. Detailed formulation is provided in supplementary material. To compute detailed specular reflectance, even perfect reflectance on mirror material, and radiance integral with image-based lighting, we calculate Monte Carlo numerical integration with importance sampling [24] according to:

$$\int_{H^+} f_r L_i(\omega_i)(\omega_i \cdot n)d\omega_i \approx \frac{1}{N}\sum_{k=1}^{N}\frac{f_r L_i(l_k)(l_k \cdot n)}{p(l_k, v)}, \tag{10}$$

where $f_r = f_d + f_s$, p is probability density function and v denotes view direction. We use $N = 512$ for the diffuse component and $N = 256$ for the specular component.

We apply the importance sampling method to decrease variance, which allows us only cover the important direction according to known BRDF of surface.

The proposed physics-based in-network rendering module will be incorporated into the training of LNet. The physical constraints are added by re-rendering loss:

$$L_{render} = \|I - \tilde{I}\|_2^2. \tag{11}$$

Therefore, the final loss of LNet is:

$$L_{LNet} = \beta_L L_L + \beta_{SC} L_{SC} + \beta_{render} L_{render}. \tag{12}$$

Fast trainable guided filter. Due to piece-wise smooth of base color, roughness and normal, several learning-based methods has been proposed [27, 29, 55] to refine them. Inspired by [49], we train a CNN with the guided filter named GSNet on half-resolution components and upsample learned parameters to filter source-resolution components. Thus, our trainable guided solver can be trained efficiently.

4. Experiments

In this section, we verify the validity of the proposed refine module, SC loss and physics-based rendering module from different sub-tasks including material estimation, lighting estimation and geometry estimation. Experiments are deployed on several benchmarking datasets with both synthetic and real ones, together with comparisons to the state-of-the-arts. Specifically, we compare with InvIndoor [29] in all three sub-tasks due to the similar SVBRDF representation and per-pixel illumination representation; we

Table 2. Quantitative comparison of base color, normal, roughness, metalness and re-rendered image on *FutureHouse*. MSE metric for BRDFs and re-rendered image, Mean Angular Error for normal. † Because the refine process of LRG360 [27] is very time-consuming, the result of LRG360 [27] is calculated by theirfound coarse albedo.

	Base color	Normal	Roughness	Metalness	Re-render
InvIndoor [29]	0.1093	63.73°	0.0868	N.A.	0.0108
LRG360 [27]	0.0968†	11.40°	N.A.	N.A.	-
Ours	**0.0090**	**10.26°**	**0.0187**	**0.0113**	**0.0061**

Table 3. Quantitative comparison of material and geometry between LRG360 [27] and our method on test data provided by LRG360 [27]. MSE metric ($\times 10^{-2}$) for albedo and Mean Angular Error for normal.

	LRG360 [27]		Ours	
	Coarse	Refine	Coarse	Refine
Albedo	5.574	2.600	2.260	2.165
Normal	16.5°	N.A.	15.1°	15.2°

Table 4. Ablation study of material estimation and normal estimation. MSE metric ($\times 10^{-2}$) for BRDFs.

	Base color	Normal	Roughness	Metalness
Baseline	0.955	10.20°	2.037	1.147
+CirP	0.940	10.12°	1.934	1.119
+CirP+Joint	0.926	10.17°	1.928	1.133
+CirP+GSNet	0.902	10.26°	1.872	N.A.

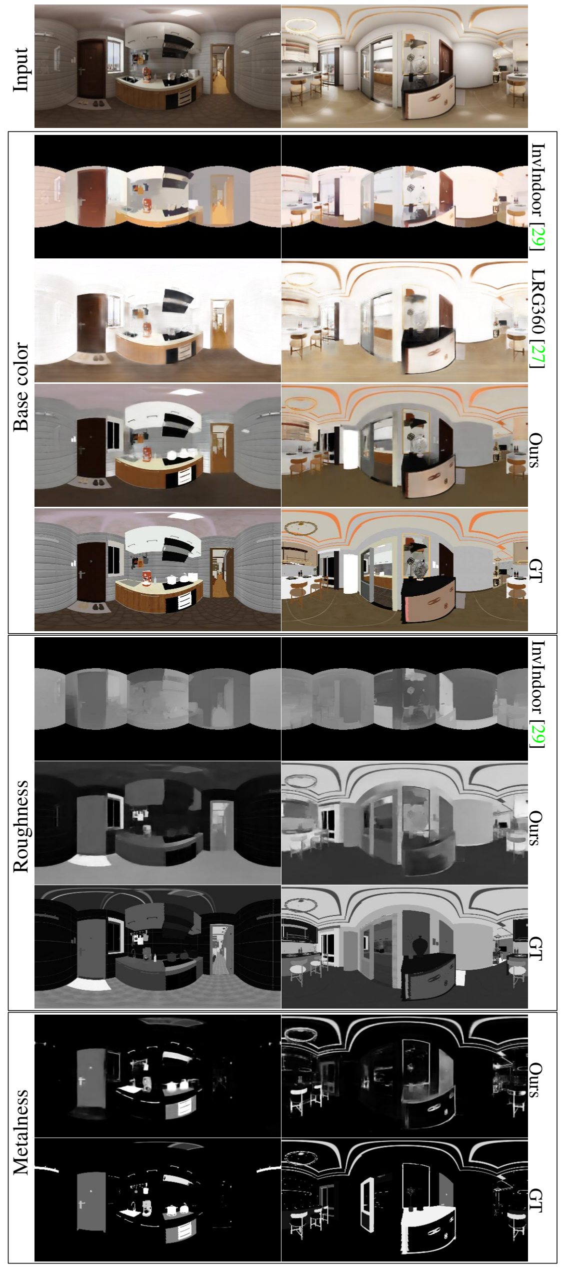

Figure 3. Qualitative comparison of material estimation on *FutureHouse*.

also compare with a panoramic intrinsic image decomposition method, LRG360 [27], in material estimation and geometry estimation; some panoramic methods [22,44,57] are considered for comparison in depth estimation[1].

4.1. Material estimation

We compare two methods [27, 29] on our unseen test set of *FutureHouse* and test data provided by LRG360 [27]. Because InvIndoor [29] takes a single perspective image as input, we only compute four horizontal maps of panorama for a fair comparison, following LRG360 [27].

As shown in Table 2, InvIndoor [29] has a larger mean angular error for normal due to the lack of global features of the entire panorama in their LFOV input. By the virtue of panorama input and depth input, LRG360 [27] predicts a relatively accurate normal map. However, the base color cannot be estimated accurately due to the limited BRDF representation, only Lambertian BRDF. With the more physical and completed representation of material, our method significantly outperforms LRG360 [27]. As shown in Figure 3, our method predicts high-quality and comprehensive components.

The quantitative results on test data proposed by LRG360 [27] are shown in Table 3. Our method outperforms LRG360 [27] both in albedo estimation and normal estimation tasks. As shown in Figure 4, our method produces more detailed predictions even on out-of-distribution (OOD) data. The test-time optimization proposed in LRG360 [27] generates more smooth predictions, but we found that it brings some lighting effects back, *e.g.*, highlight on the stool and shadow on the floor in Figure 4.

[1]we tried to compare with several inverse rendering approaches [43, 46, 55], but failed to receive available results after e-mail query.

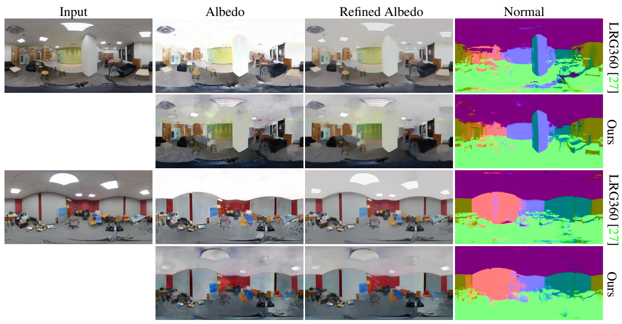

Figure 4. Qualitative comparison on real data provided by LRG360 [27]. Although LRG360 with depth as input, our method predicts more detailed geometry.

Furthermore, we evaluate the re-rendering error with InvIndoor [29] on *FutureHouse* in Table 2. Our method dramatically outperforms InvIndoor [29]. As shown in Figure 5, our method can re-render complex lighting effects while InvIndoor [29] losses these details due to the limited SVBRDF representation and unsuitable sampling. We verify the validity of CirP, joint training of the whole pipeline and refine module, i.e., GSNet, in Table 4. The joint training improves the overall performance of all components with physical constraints. The performance of depth with joint training can be found in the supplement.

4.2. Lighting estimation

Due to the lack of panoramic data with spatially-coherent local lighting, we captured a real panoramic dataset with high-resolution (8K) spatially-coherent illumination. In the following, we use this dataset to compare with InvIndoor [29] quantitatively and qualitatively.

Spatially-Coherent illumination dataset: All panoramas are captured by a Insta360 pro 2 camera with six fisheye lens. For HDR information, the scene is captured by merging seven exposures (shutter speed from $\frac{1}{8000}$ seconds to $\frac{1}{2}$ seconds) with f2.0 aperture. We first capture a center HDR panorama as input. For each input, we select several local positions at this center panorama to put camera. Especially, there is a position for putting a slideway to capture SC illumination (more details in the supplementary). For each scene, we fix the camera direction by the compass to ensure that all local light probes are aligned with center input. In total, a real panoramic dataset including 7 indoor scenes and 72 local high-resolution HDR light probes is captured.

The metric of lighting estimation is the relighting error of the virtual sphere rendered by predicted or GT light probes. The results are reported in Table 5. Each method renders three spheres with different material, including pure diffuse, matte sliver and mirror sliver. The relighting error of diffuse sphere measures the dynamic range of predicted illumination and the relighting error of mirror sliver measures the detail of illumination. Detailed parameters of sphere are shown in the supplementary. The qualitative results are shown in Figure 6. Our method can estimate more consistent lighting compared to InvIndoor [29]. We ablate the SC loss, the re-render loss and the joint training in Table 6. The result suggests that our SC loss provides meaningful constraints of 3D coherence the proposed re-rendering loss ensures that our predicted illumination is correct physically and the joint training achieves the best overall performance.

Table 5. Quantitative comparison (relighting errors) between InvIndoor [29] and our approach on the spatially-coherent illumination dataset.

	Diffuse		Matte Sliver		Mirror Sliver	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
InvIndoor [29]	0.0975	0.1153	0.1440	0.1807	0.2407	0.2869
Ours	**0.0645**	**0.0789**	**0.0858**	**0.1190**	**0.1117**	**0.1449**

Table 6. Evaluation of SC loss, re-render loss and joint training on *FutureHouse*.

	SSIM↑	Re-render Error (MSE)↓
L_L	0.6150	0.0583
$L_L + L_{SC}$	0.6169	0.0714
$L_L + L_{SC} + L_{render}$	0.6124	0.0060
$L_L + L_{SC} + L_{render} + joint$	0.6169	0.0061

4.3. Depth and normal estimation

We conduct this experiment on a widely used panoramic dataset, 3D60 [23, 57]. It consists of two realistic datasets and a synthetic dataset, i.e., Matterport3D [9], 2D3D-S [1], and SUNCG [42]. We compare with the recent state-of-arts of panoramic depth estimation [22, 44, 57] on the 3D60 dataset. The results are shown in Table 7. Although our model is inferior to UniFuse [22], our parameters are less than half of it. The competitive performance shows that our proposed method is able to effectively estimate accurate geometry, which is helpful in realistic mixed reality applications.

More importantly, we compare with previous panoramic inverse rendering method, LRG360 [27], for normal estimation on *FutureHouse*, test data provided by LRG360 [27] and 3D60. As shown in Table 2 and Table 3, our method achieves state-of-the-art performance without depth as input, whereas LRG360 [27] requires RGBD as input. In Table 8, our method also significantly outperforms LRG360 [27] (w/ pred depth). The performance of LRG360 [27] greatly depends on the quality of depth. However, the high-quality depth is not always available. The Mean Angular Error of our prediction on 3D60 is similar to that on *FutureHouse* and test data provided by LRG360 [27], which indicates the good generalization capability of our model.

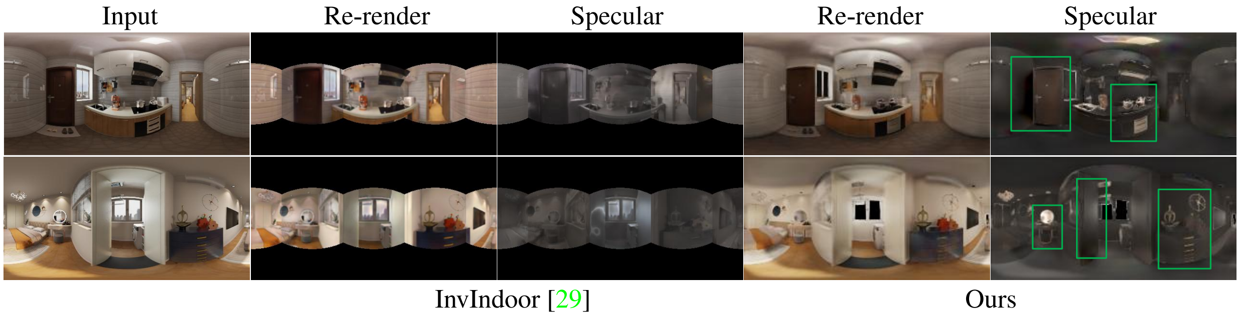

Figure 5. Qualitative comparison of re-rendered images. With a more physical SVBRDF model and the physics-based differentiable renderer, our approach reproduces realistic lighting effects, especially non-monochrome specular reflectance on complex material, *e.g*., glossy wall and metal kettle.

Table 7. Quantitative comparison of depth on 3D60 Dataset [23, 57]. The performance evaluated on standard metrics are shown in below. The results of OminiDepth [57], BiFuse [44] and UniFuse [22] are taken from UniFuse [22].

	MAE	Abs Rel	RMSE	RMSElog	$\delta_1 \uparrow$	$\delta_2 \uparrow$	$\delta_3 \uparrow$
OmniDepth [57]	-	0.0702	0.2911	0.0725	0.9574	0.9933	0.9979
BiFuse [44]	0.1143	0.0615	0.2440	0.0428	0.9699	0.9927	0.9969
UniFuse [22]	0.0996	0.0466	0.1968	0.0315	0.9835	0.9965	0.9987
Ours (w/ finetune)	0.1236	0.0575	0.2367	0.0382	0.9656	0.9938	0.9982

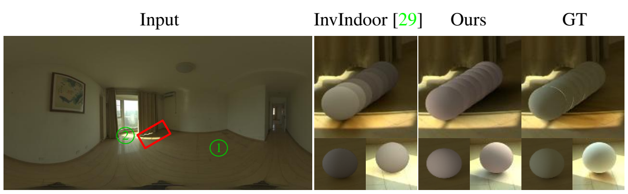

Figure 6. Qualitative comparison of virtual object insertion on captured SC illumination dataset. Our method produces more consistent results while InvIndoor [29] generates flicking results.

Table 8. Quantitative comparison of normal between LRG360 [27] and our method on 3D60 Dataset [23, 57].

	Mean Angular Error
LRG360 [27](w/ pred depth)	28.017°
LRG360 [27](w/ GT depth)	6.957°
Ours (w/o finetune)	12.353°

5. Conclusion

In this paper, we proposed a physics-based inverse rendering framework that recovers geometry, material and SV lighting from a single panorama. Our more completed SVBRDF representation can handle complex material such as glossy, metal and even mirror material, which have been overlooked in previous approaches. With detailed non-monochrome specular reflectance on complex material rendered by our physics-based in-network rendering module, more physical constraints are incorporated. Experimental results verified that our model outperforms prior works for material, lighting and geometry estimation. In the future work, we consider extending this physics-based architecture to additional illumination representations.

Acknowledgements

We thank Zhengqin Li, Tong Rao, Yan Wu for constructive discussions. This work is supported in part by the National Natural Science Foundation of China (NFSC) (No. 62002295) and Shaanxi Provincial Key R&D Program (No. 2021KWZ-03).

References

[1] Iro Armeni, Sasha Sax, Amir R Zamir, and Silvio Savarese. Joint 2d-3d-semantic data for indoor scene understanding. *arXiv preprint arXiv:1702.01105*, 2017. 3, 7

[2] Jonathan T Barron and Jitendra Malik. Intrinsic scene properties from a single rgb-d image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 17–24, 2013. 1, 2

[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

[4] H. G. Barrow and J. M. Tenenbaum. Recovering intrinsic scene characteristics from images. In A. Hanson and E. Riseman, editors, *Computer Vision Systems*, pages 3–26. Academic Press, 1978. 2

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

[6] 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 - Experimental Ideas Implementations*. The Eurographics Association, 2018. 2

[7] Mark Boss, Varun Jampani, Kihwan Kim, Hendrik P.A. Lensch, and Jan Kautz. Two-shot spatially-varying brdf and shape estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

[8] Dan A. Calian, Jean-François Lalonde, Paulo Gotardo, Tomas Simon, Iain Matthews, and Kenny Mitchell. From Faces to Outdoor Light Probes. *Computer Graphics Forum*, 2018. 3

[9] 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 (3DV)*, 2017. 3, 7

[10] Dachuan Cheng, Jian Shi, Yanyun Chen, Xiaoming Deng, and Xiaopeng Zhang. Learning Scene Illumination by Pairwise Photos from Rear and Front Mobile Cameras. *Computer Graphics Forum*, 2018. 3

[11] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017. 3

[12] Epic Games. Unreal engine. 3, 4

[13] Qingnan Fan, Jiaolong Yang, Gang Hua, Baoquan Chen, and David Wipf. Revisiting deep intrinsic image decompositions. 2018. 2

[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 International Conference on Computer Vision*, pages 7175–7183, 2019. 2

[15] 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 (SIGGRAPH Asia)*, 9(4), 2017. 2, 3, 5

[16] Mathieu Garon, Kalyan Sunkavalli, Sunil Hadap, Nathan Carr, and Jean-Francois Lalonde. Fast spatially-varying indoor lighting estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1, 2, 3, 4, 5

[17] S. Georgoulis, K. Rematas, T. Ritschel, E. Gavves, M. Fritz, L. Van Gool, and T. Tuytelaars. Reflectance and natural illumination from single-material specular objects using deep learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(8):1932–1947, Aug 2018. 1, 2

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 4

[19] Yannick Hold-Geoffroy, Akshaya Athawale, and Jean-François Lalonde. Deep sky modeling for single image outdoor lighting estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6927–6935, 2019. 2, 3

[20] 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. 2

[21] Michael Janner, Jiajun Wu, Tejas Kulkarni, Ilker Yildirim, and Joshua B Tenenbaum. Self-Supervised Intrinsic Image Decomposition. In *Advances In Neural Information Processing Systems*, 2017. 1, 2

[22] Hualie Jiang, Zhe Sheng, Siyu Zhu, Zilong Dong, and Rui Huang. Unifuse: Unidirectional fusion for 360° panorama depth estimation. *IEEE Robotics and Automation Letters*, 2021. 6, 7, 8

[23] Antonis Karakottas, Nikolaos Zioulis, Stamatis Samaras, Dimitrios Ataloglou, Vasileios Gkitsas, Dimitrios Zarpalas, and Petros Daras. 360 surface regression with a hyper-sphere loss. In *International Conference on 3D Vision*, September 2019. 7, 8

[24] Brian Karis and Epic Games. Real shading in unreal engine 4, 2013. 4, 5

[25] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *3D Vision (3DV), 2016 Fourth International Conference on*, pages 239–248. IEEE, 2016. 4

[26] Louis Lettry, Kenneth Vanhoey, and Luc Van Gool. Unsupervised Deep Single-Image Intrinsic Decomposition using Illumination-Varying Image Sequences. *Computer Graphics Forum (Proceedings of Pacific Graphics)*, 37(10), October 2018. 2

[27] Junxuan Li, Hongdong Li, and Yasuyuki Matsushita. Lighting, reflectance and geometry estimation from 360° panoramic stereo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 1, 2, 3, 5, 6, 7, 8

[28] Wenbin Li, Sajad Saeedi, John McCormac, Ronald Clark, Dimos Tzoumanikas, Qing Ye, Yuzhong Huang, Rui Tang, and Stefan Leutenegger. Interiornet: Mega-scale multi-sensor photo-realistic indoor scenes dataset. In *British Machine Vision Conference (BMVC)*, 2018. 2, 3, 4

[29] 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, 4, 5, 6, 7, 8

[30] Zhengqi Li and Noah Snavely. Cgintrinsics: Better intrinsic image decomposition through physically-based rendering. In *European Conference on Computer Vision (ECCV)*, 2018. 1, 2, 3, 4

[31] Zhengqin Li, Zexiang Xu, Ravi Ramamoorthi, Kalyan Sunkavalli, and Manmohan Chandraker. Learning to reconstruct shape and spatially-varying reflectance from a single

image. In *SIGGRAPH Asia 2018 Technical Papers*, page 269. ACM, 2018. 1, 2

[32] 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 (CVPR)*, pages 7190–7199, June 2021. 2, 3, 4

[33] Shiqiu Liu. Dlss 2.0 - image reconstruction for real-time rendering with deep learning, 2020. 4

[34] Yunfei Liu, Yu Li, Shaodi You, and Feng Lu. Unsupervised learning for intrinsic image decomposition from a single image. In *CVPR*, 2020. 1

[35] Stephen Lombardi and Ko Nishino. Reflectance and illumination recovery in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38:1–1, 12 2015. 1

[36] Robert Maier, Kihwan Kim, Daniel Cremers, Jan Kautz, and Matthias Nießner. Intrinsic3d: High-quality 3d reconstruction by joint appearance and geometry optimization with spatially-varying lighting. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3114–3122, 2017. 1, 2, 5

[37] Konstantinos Rematas, Tobias Ritschel, Mario Fritz, Efstratios Gavves, and Tinne Tuytelaars. Deep reflectance maps. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 1

[38] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241. Springer International Publishing, 2015. 4

[39] 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 *International Conference on Computer Vision (ICCV)*, 2019. 1, 2, 3, 4, 5

[40] Soumyadip Sengupta, Angjoo Kanazawa, Carlos D. Castillo, and David W. Jacobs. Sfsnet: Learning shape, refectance and illuminance of faces in the wild. In *Computer Vision and Pattern Regognition (CVPR)*, 2018. 1, 2

[41] Shuran Song and Thomas Funkhouser. Neural illumination: Lighting prediction for indoor environments. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2

[42] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. *Proceedings of 30th IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 3, 4, 7

[43] Pratul P Srinivasan, Ben Mildenhall, Matthew Tancik, Jonathan T Barron, Richard Tucker, and Noah Snavely. Lighthouse: Predicting lighting volumes for spatially-coherent illumination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8080–8089, 2020. 2, 3, 6

[44] Fu-En Wang, Yu-Hsuan Yeh, Min Sun, Wei-Chen Chiu, and Yi-Hsuan Tsai. Bifuse: Monocular 360 depth estimation via bi-projection fusion. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 6, 7, 8

[45] Tsun-Hsuan Wang, Hung-Jui Huang, Juan-Ting Lin, Chan-Wei Hu, Kuo-Hao Zeng, and Min Sun. Omnidirectional cnn for visual place recognition and navigation. *arXiv preprint arXiv:1803.04228*, 2018. 4

[46] Zian Wang, Jonah Philion, Sanja Fidler, and Jan Kautz. Learning indoor inverse rendering with 3d spatially-varying lighting. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2021. 1, 2, 3, 5, 6

[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. 2

[48] Xin Wei, Guojun Chen, Yue Dong, Stephen Lin, and Xin Tong. Object-based illumination estimation with rendering-aware neural networks. *arXiv preprint arXiv:2008.02514*, 2020. 1, 2

[49] Huikai Wu, Shuai Zheng, Junge Zhang, and Kaiqi Huang. Fast end-to-end trainable guided filter. In *CVPR*, 2018. 5

[50] Fangneng Zhan, Changgong Zhang, Yingchen Yu, Yuan Chang, Shijian Lu, Feiying Ma, and Xuansong Xie. Emlight: Lighting estimation via spherical distribution approximation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021. 2

[51] Jinsong Zhang, Kalyan Sunkavalli, Yannick Hold-Geoffroy, Sunil Hadap, Jonathan Eisenman, and Jean-Francois Lalonde. All-weather deep outdoor lighting estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2

[52] 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 *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1

[53] Yinda Zhang, Shuran Song, Ersin Yumer, Manolis Savva, Joon-Young Lee, Hailin Jin, and Thomas Funkhouser. Physically-based rendering for indoor scene understanding using convolutional neural networks. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 3, 4

[54] Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. Structured3d: A large photo-realistic dataset for structured 3d modeling. In *Proceedings of The European Conference on Computer Vision (ECCV)*, 2020. 2, 3, 4

[55] Hao Zhou, Xiang Yu, and David W. Jacobs. Glosh: Global-local spherical harmonics for intrinsic image decomposition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 2, 5, 6

[56] Yongjie Zhu, Yinda Zhang, Si Li, and Boxin Shi. Spatially-varying outdoor lighting estimation from intrinsics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12834–12842, June 2021. 1, 2

[57] Nikolaos Zioulis, Antonis Karakottas, Dimitrios Zarpalas, and Petros Daras. Omnidepth: Dense depth estimation for indoors spherical panoramas. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 448–465, 2018. 6, 7, 8