LightOctree: Lightweight 3D Spatially-Coherent Indoor Lighting Estimation

Xuecan Wang[1*]   Shibang Xiao[1*]   Xiaohui Liang[1,2†]

[1] Beihang University, China

[2] Zhongguancun Laboratory, China

{xuecan_wang, xiaoshibang, liang_xiaohui}@buaa.edu.cn

Abstract

We present a lightweight solution for estimating spatially-coherent indoor lighting from a single RGB image. Previous methods for estimating illumination using volumetric representations have overlooked the sparse distribution of light sources in space, necessitating substantial memory and computational resources for achieving high-quality results. We introduce a unified, voxel octree-based illumination estimation framework to produce 3D spatially-coherent lighting. Additionally, a differentiable voxel octree cone tracing rendering layer is proposed to eliminate regular volumetric representation throughout the entire process and ensure the retention of features across different frequency domains. This reduction significantly decreases spatial usage and required floating-point operations without substantially compromising precision. Experimental results demonstrate that our approach achieves high-quality coherent estimation with minimal cost compared to previous methods.

1. Introduction

In the realm of mixed reality and image editing, achieving visual coherence is a fundamental research concern when integrating virtual objects into real-world images. Lighting is particularly crucial for seamlessly blending virtual and real elements, ensuring consistent shading and shadowing between virtual objects and the surrounding environment. However, estimating the inverse process of global environmental lighting from narrow field-of-view (FOV) images is a challenging task due to the lack of sufficient information. Particularly in indoor environments, the proximity of lights to objects typically leads to significant variations in incident illumination across different positions within the scene. Moreover, when incorporating moving objects into the scene, ensuring lighting coherence during spatial changes is crucial. This necessitates the estimation of a spatially-
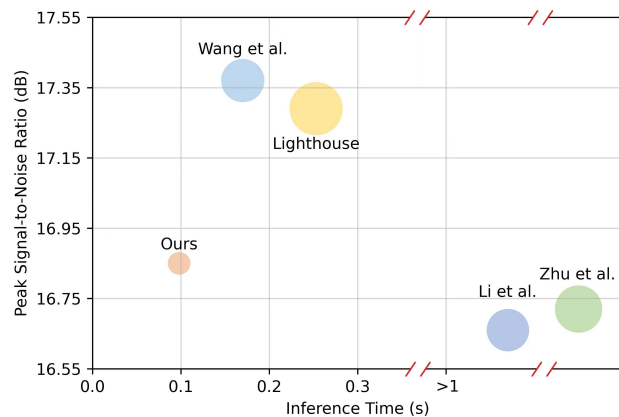
---

*Equal contribution
†Corresponding author

Figure 1. The comparison with state-of-the-art (SOTA) methods shows that our approach achieves relatively good accuracy while requiring lower storage and computational costs.

coherent model of illumination that is 3D spatially-coherent [25]. Given these requirements and application scenarios, it is essential for the entire process of lighting estimation and virtual-real fusion rendering to be efficient and lightweight.

Some of existing work generate 2D illumination map as global lighting estimation. Such as methods using light probes [7, 21, 29], or deep neural networks like 2D CNNs [1, 22] to make predictions in image space. However, these algorithms currently used to estimate global illumination either ignore spatial variation effects and predict a single illumination for the entire scene [2, 3, 8, 28, 30, 34, 35], or estimate spatially-varying illumination by separately predicting the lighting at each 3D position in the scene [20, 23, 37, 38, 42]. Although these approaches may produce impressive results, they cannot guarantee smooth variations in predicted illumination with changes in position, especially when inserting moving virtual objects, which can easily break the illusion. Therefore, recent research has explored illumination estimation methods based on extending the lighting representation to three-dimensional space, typically using 3D volumetric $RGB\alpha$ models [25, 41, 43]. These methods facilitate spatially-coherent lighting estima-

tion, but they also introduce higher-dimensional lighting representations and more complex computations. Such as 3D CNNs, which are typically computationally and memory-intensive. At the same time, they overlook the sparse distribution of lighting in the environment, leading to significant waste of space and computing power, making them unsuitable for performance-constrained applications.

Upon observation, we note that assuming the incident radiation at a certain point comes solely from the surfaces in the scene, a common assumption in existing methods, leads to a sparse distribution of surfaces in three-dimensional space. Therefore, we can use a sparse data structure as a lighting representation instead of a coarse uniform grid. Drawing inspiration from NeRFs[9, 16, 36] and octree-based global illumination researches[5, 18], our approach delves into a lighting representation based on sparse voxel octrees and proposed a lightweight, spatially-coherent global lighting estimation network that accounts for the distribution characteristics of the light field in the scene. We also introduce a differentiable rendering layer based on the voxel octree representation to streamline the framework and better align with the training of the octree-based network. Additionally, we explore a rendering method for virtual object insertion based on a hybrid scene representation using voxel octrees and point clouds. By restricting data storage and calculations to octants, our method incur a memory and computational cost of $O(n^2)$, where n is the voxel resolution per dimension at the finest granularity level. In contrast, utilizing a 3D uniform voxel grid representation solution results in a memory and computational cost of $O(n^3)$. This approach enables low-cost, high-quality coherent augmentations.

In summary, our main contributions are as follows:

- An octree-based framework to achieve virtual-real lighting consistency, optimized for less memory and hardware requirements with minimal loss of accuracy. This framework efficiently generates 3D spatially-coherent lighting interactively and inserts virtual objects in real-time using only a single RGB image of indoor scenes.
- A novel lighting representation base on voxel octree for indoor lighting estimation, which considered the sparsity of light field in 3D space, significantly reducing storage space and computational complexity.
- A lightweight lighting estimation network, featuring a novel multi-scale rendering layer. This network enables end-to-end estimation high-quality incident radiance fields in the form of the voxel octree.

2. Related Work

To achieve coherent augmentations, three primary issues must be addressed: geometric registration, photometric registration, and camera simulation [27]. This paper specifically focuses on photometric registration, which involves the interaction of light between the real world and the augmented

effects. In this section, we will provide a brief review of pertinent prior research endeavors focused on estimating diverse lighting representations from images, leveraging the simplifying assumptions of the light field.

Non Spatially-varying Lighting Estimation. Early research assumed that the incident light field received at any position in the scene is uniform, simplifying the spatial distribution of illumination to a single spherical distribution. In this scenario, the scene's lighting can be represented by a HDR panorama. This conclusion was validated by Debevec et al.[6], who demonstrated that virtual objects can be realistically inserted into real photographs using HDR environment maps obtained through multiple exposures of a chrome probe. Subsequent methods [1–3, 8, 22, 28, 30, 34, 35] have illustrated that deep learning techniques can estimate HDR environment maps from a single LDR photograph. These methods offer the advantage of significantly reducing the dimensionality of the output space of the lighting estimation algorithm, thereby enabling the design of complex network models to accomplish more detailed image generation tasks. However, a single environment map is insufficient for compositing multiple, large, or moving virtual objects into the captured scene[25], particularly in indoor environments where light sources and other scene content may be in close proximity to the insertion positions of the objects.

Spatially-varying Lighting Estimation. To overcome the limitations of single environment map methods, recent research has delved into spatially-varying lighting estimation algorithms for indoor scenes. Numerous studies have predicted spatially-varying lighting in images by estimating per-pixel spherical lobes [10, 26, 38–40] (spherical harmonics/Gaussians) or individual environment maps for each pixel [20, 23] in the input image. However, these methods overlook the depth factor and can only estimate the lighting for the surface corresponding to any given pixel in the image. Consequently, the estimated lighting lacks continuity in three-dimensional space, as highlighted by Srinivasan et al.[25] and Karsch et al.[15]. In the most recent work[25, 41, 43], higher-order voxelized lighting representations were utilized for estimation, and a 3D neural network was employed to achieve spatially-coherent lighting estimation. Nevertheless, the use of 3D networks and 3D lighting representations significantly increases the storage and computational complexity of the algorithm, hindering its application in AR. Additionally, in general indoor scenes where special objects like participating media are not considered, the incident radiance at a point can be simplified to only come from the surfaces in the scene. Therefore, using the 3D voxel grid structure of existing methods to represent lighting would result in unnecessary storage and computational losses. Our approach explores the use of voxel octree as a compressed representation for lighting/scene and designs a
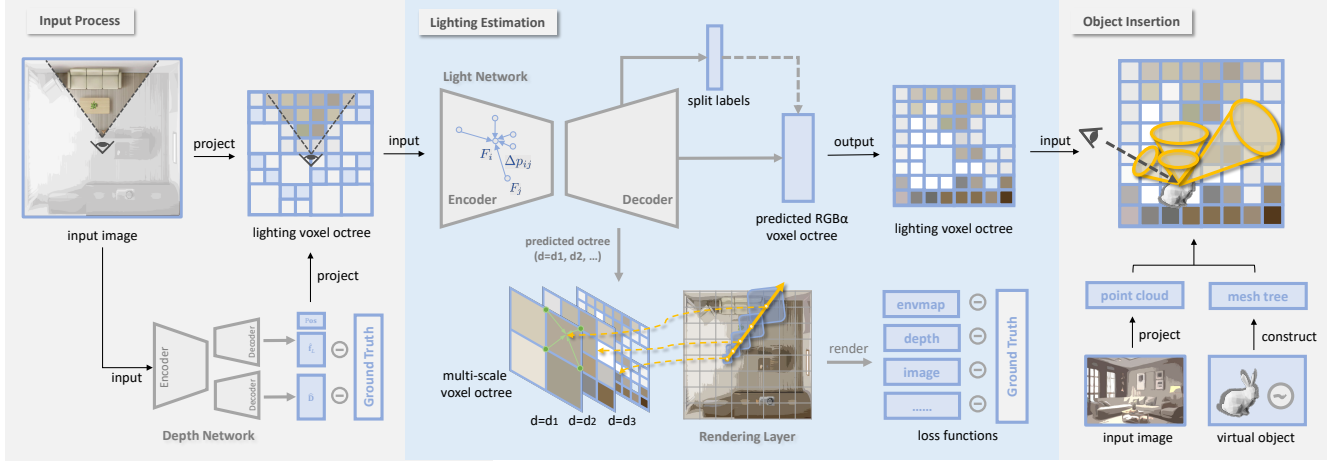
Figure 2. **Overall structure of our framework.** The process starts with the Input Process stage, where the direct prediction module estimates depth and extracts global illumination features from the input image. The depth values and features are combined to create a point cloud, which is used to build a 3D voxel octree scene representation. Next, the Lighting Estimation stage uses a U-Net structure to predict the lighting from the constructed octree. The Object Insertion stage then combines the original RGB information, predicted depth values, predicted lighting voxel octree, and user-specified mesh data to render an image with consistent lighting for the inserted virtuals object.

lightweight and spatially-coherent global lighting estimation network. By restricting network and rendering to an octree, and combining them with a hierarchical feature fusion rendering layer design, we achieve low-cost full-frequency global lighting estimation from a single RGB image.

3. Overview

Given a single LDR RGB image in narrow FOV, we aim to estimate a 3D voxel octree lighting representation of the scene with minimal storage and computational cost, and integrate the estimation results with information about virtual objects. Our pipeline can be divided into three stages, as shown in Fig.2, including two neural network and one object rendering module. In the follow sections, we will illustrate how our voxel octree is involved throughout the entire workflow to reduce the cost of storage and computation.

In Sec.4, we will describe our lightweight illumination estimation method and corresponding training strategy, covering the input process and the lighting estimation stages. In Sec. 5, we will showcase several experiments conducted to validate our designs. Finally, Sec.6 will provide a brief summary and outlook of our approach.

4. Lightweight Lighting Estimation

4.1. Depth Network

The Depth Network is intended to provide initial predictions of depth and global features including lighting information when given a single image as input. Adopting a similar approach to [41], we employ DenseNet121 [13] as the backbone for the 2D CNN encoder. The depth decoder uti-

lizes skip connections to generate the final predicted depth $\hat{D} \in \mathbb{R}^{3 \times H \times W}$. Moreover, drawing inspiration from the work in [43], we incorporate an additional module subsequent to the initial encoder to extract global lighting information, which is ultimately output as a feature vector $\hat{\mathbf{f}_L} \in \mathbb{R}^C$. These features are then amalgamated with the input of the Light Network to facilitate the prediction process.

4.2. Light Network

The Light Network is to predict lighting octree from RGB-D values and global lighting features. To achieve lightweight lighting estimation, a compact lighting representation supporting fast, parallel network computation and rendering is essential. Thus, we devised a lighting voxel octree, a 3D volumetric lighting representation organized using an octree. The octree efficiently represents the surface radiance exiting from the entire scene, encompassing both visible surfaces and those outside the FoV. However, introducing the octree has presented significant challenges for network design, particularly for the ill-posed regression task of lighting estimation based on a single image, which previous neural network models using octree representation could not address. To tackle this issue, we integrated the hierarchical structure of the octree with a rendering layer based on differentiable cone tracing, facilitating the fusion of multi-scale lighting features. This enables us to maximize the network's receptive field while ensuring its preference for inductive lighting estimation at minimal cost. Subsequently, we developed a lightweight lighting estimation network based on octree graph network operators. The following section provides a comprehensive description of the model design and
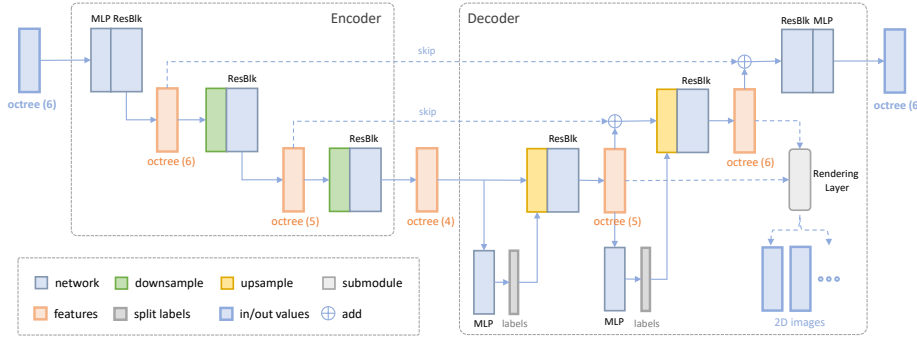
Figure 3. **Light Network.** A lightweight U-Net architecture constructed with graph O-CNN operators. These modules work in tandem to forecast octree subdivisions, field values, and establish correspondence with ground truth through the Rendering Layer.To provide a clearer illustration of the structure, here presents a network with a full depth of 4 and depth of 6, while the actual depth is 7 (equivalent to voxel resolution of 128^3).

training specifics for this segment.

4.2.1 Lighting Presentation

Lighting Voxel Octree. We assumes that the lighting in the scene emanates solely from the object surfaces, including light sources and other objects, by simplifying the image formation process into a surface rendering model. Based on this characteristic, a 3D voxel octree structure is adopted to tightly store the scene and lighting. The voxel octree, developed as a sparse lighting representation, has eight child nodes for each non-empty non-leaf node. Each non-empty node stores its corresponding shuffle key, split label, and features. For illumination, the feature information stored in each node is the RGBα value, representing the radiance emitted from that position. Meanwhile, each empty node only stores the shuffle key and label. This structure is compatible with the data structure used in [33] and is utilized to build and design the lighting prediction network using basic modules such as octree-based neural modules proposed by [31–33].

Octree Construction. When provided with depth values and corresponding features (*i.e.* RGB values) , and the camera intrinsic matrix, a 3D point cloud of thus region can be obtained through projection transformation, denoted as $\mathcal{P} = \{(\mathbf{p}_1, \mathbf{f}_1), (\mathbf{p}_2, \mathbf{f}_2), \dots, (\mathbf{p}_N, \mathbf{f}_N)\}$. Here, $\mathbf{p}_i$ represents the position information of point i, which is the 3D coordinate (x_i, y_i, z_i) in the camera coordinate system, and $\mathbf{f}_i$ represents the feature information of the point, *i.e.* the color information (r_i, g_i, b_i) corresponding to the RGB color value of the point. To construct the octree of the input 3D point cloud data, the point cloud is initially proportionally scaled to an axis-aligned unit bounding box. Subsequently, the 3D voxel grid data undergoes recursive subdivision in breadth-first order until the predefined octree depth d is reached. At each step, all non-empty octree nodes which is occupied at the current depth l are traversed, and they are subdivided into eight child octree nodes at the next depth $l + 1$.

4.2.2 Network Architecture

Given an RGB image and pixel-wise depth as input, we start by constructing a voxel octree representing the visible area. Subsequently, we incorporate the global illumination feature information $\mathbf{f}_L$ and positional data, which is stochastically generated based on the input depth distribution. This leads to an update in the octree structure and features. Following this, utilizing the voxel octree structure and a series of neural modules based on the octree data structure [31–33], we design a network model for light estimation task, as illustrated in Fig.3. This model employs the U-Net [4] structure:

$$\hat{L} = h_{LN}(I, D, \mathbf{f}_L; \Theta_{LN}) \tag{1}$$

ResBlock(n, c) in Fig.3 represents a stack of n ResNet blocks [11], each comprising two graph convolutions with a channel number of c. Moreover, DownSample(c) and UpSample(c) refer to graph downsampling and upsampling operations based on shared fully connected layers, with both input and output channels set to c. At the end of each stage in the decoder, the extracted features are input into the PredictionModule to predict the subdivision of octree nodes. This module consists of an MLP with two fully connected layers. Additionally, the extracted features are passed to the RegressionModule to predict the field values corresponding to the octree nodes in that layer. Subsequently, features stored in different layers of the voxel octree, along with camera poses, are sent to the RenderingLayer to render novel view images and calculate losses with ground truths.

4.2.3 Rendering Layer

Leveraging multi-level voxel octree for illumination representation, we propose a differentiable cone tracing rendering layer. This module efficiently conducts rendering and provides supervision for the network. Building upon the octree-based illumination representation, it rapidly renders panoramic images while performing mip-map sampling based on the distance between octants and camera. It exhibits high sensitivity to nearby light octants, resulting in

high-frequency rendering effects. Simultaneously, it aggregates distant light sources, reducing sampling frequency to enhance speed without compromising the final rendering quality. Furthermore, we derive the differentiable form of this method, enabling it to serve as a gradient-propagating rendering layer within neural networks.

The conventional volume rendering process calculates the light radiation $L(\mathbf{x}, \omega)$ from viewpoint $\mathbf{x}$ in direction ω, factoring in the sampling distance δ_k between sampling points $n-1$ and n. This process can often lead to redundant computations and complicate the efficient integration of feature information across different scales throughout the rendering process. However, our method not only streamlines the rendering process but also skillfully facilitates feature fusion across the outputs of different network layers.

In particular, when intending to render a panoramic image of the target camera pose from the predicted lighting voxel octree, we first generate cones that cover the entire 360-degree sphere based on the predefined rendering resolution and cone angle θ. We consider the distance δ_n of the current sampling point to calculate the position s_n of the next sampling point, leading to larger sampling intervals for points that are farther away, as shown in the equation:

$$s_n = \sum_{k=0}^{n} \delta_k, \quad \delta_n = \begin{cases} c_0, & n = 0 \\ c \cdot s_{n-1} \tan \theta, & n \geq 1 \end{cases} \quad (2)$$

where c controls the rate of increase in the sampling step size. We use octree nodes of depth $d_n = \lceil \log_2 \delta_n / l_0 \rceil$ in the octree structure for sampling, where l_0 denotes the minimum side length of the leaf node in the octree. This significantly reduces the number of necessary samples and accelerates computations, while still maintaining the final rendering quality. Based on these processes, we can formulate the new rendering equation as follows:

$$L(s) = \sum_{n=1}^{N} w_n C_n^{d_n} T_n \left(1 - e^{-\sigma_n^{d_n} \delta_n}\right) \quad (3)$$

where $C_n^{d_n}$ and $\sigma_n^{d_n}$ represents the radiance and density of sampling point n at a depth of d_n. And $w_n = \delta_n / s_n$ is the weight for each sampling point, while $T_n = \exp\left(-\sum_{k=1}^{n-1} \sigma_k^{d_k} \delta_k\right)$ is the transmittance which represents the degree of light attenuation in medium. To ensure the differentiability of the above process, we derive the derivatives of the color and opacity with respect to the forward rendering process. Please see details in supplementary file.

4.3. Loss Function

Our model is trained using synthetic data with ground truth $\{I, D, O_s, \{I_{\text{nv}}, D_{\text{nv}}, P_{\text{nv}}\}_N\}$, where I is the input image, D denotes pixel-wise depth, O_s is a voxel octree constructed from a complete scene point cloud, and $\{I_{\text{nv}}, D_{\text{nv}}, P_{\text{nv}}\}_N$ represent HDR panoramas, depth panoramas, and camera poses
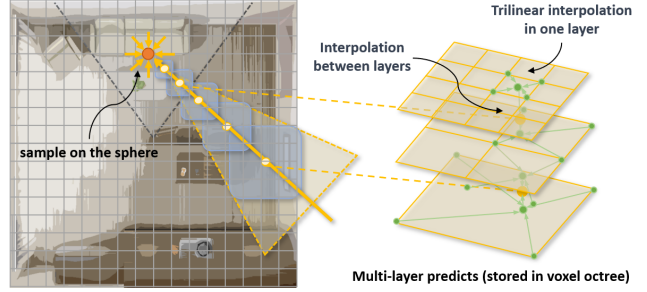

Figure 4. Rendering layer.

for N novel views, respectively. This 3D voxel octree lighting representation allows us to avoid the need for densely rendered spherical lobe lighting ground truth used in prior works [26, 38, 40]. Additionally, this representation ensures that the estimated results maintain angular frequency and sphere coherence, similar to other volumetric representation methods [25, 41, 43], but with less computational overhead.

In the Depth Network, we utilize a log-encoded L2 loss to accommodate the high dynamic range of depth[38] and a scale-invariant L2 loss to promote relative consistency, addressing the inherent scale ambiguity of depth[43]. Inspired by [25, 43], we utilized an adversarial loss[14] with a GAN-like structure to supervise the global lighting decoder, enhancing its ability to capture global features of the scene.

In the Light Network, to regress the ground truth volumetric fields, we utilize an octree constructed from the ground truth point cloud of the entire scene as supervision. We employ binary cross-entropy loss to determine the emptiness of a node, providing the supervision for predicting split labels:

$$\mathcal{L}_{\text{octree}} = \sum_d \frac{1}{N_d} \sum_{o \in O_d} \text{CrossEntropy}\left(o, o_{\text{s}}\right) \quad (4)$$

where d represents the depth of an octree layer, N_d represents the total number of nodes in the d-th layer, O_d represents the predicted octree node status, and o_{gt} is the corresponding ground truth node status.

We supervise the prediction of the illumination $\hat{L}$ using the image intensity loss of the HDR panoramic image I_{nv}, as the pixel values in images reflect the corresponding HDR radiance along the camera rays. To achieve this, for each pixel p, we calculate the camera ray starting from the camera center $\mathbf{c}$ along the direction $\mathbf{r}$ using the camera pose P_{nv} and camera intrinsic parameters. Subsequently, we render the environment map $\hat{I}_{\text{nv},p}$ of the new view using the predicted illumination $\{\hat{L}_d\}$ with the assistance of the rendering layer:

$$\hat{I}_{\text{nv},p} = \text{RenderingLayer}(\mathbf{c}, \mathbf{r}, \{\hat{L}_d\}) \quad (5)$$

If only the alpha channel is rendered, a depth panorama $\hat{D}_{\text{nv}}$ for each camera pose can also be obtained using a similar

method. We use log-encoded L2 loss to enforce consistency between the rendered new view $\hat{I}_{nv}$ and the ground truth I_{nv}:

$$\mathcal{L}_{\text{light}} = \lambda_{\text{li}} \left\| \log(I_{nv} + 1) - \log(\hat{I}_{nv} + 1) \right\|_2^2 \\ + \lambda_{\text{ld}} \left\| \log(D_{nv} + 1) - \log(\hat{D}_{nv} + 1) \right\|_2^2 \quad (6)$$

An additional means of supervising the illumination is by ensuring consistency with the visible scene. So we employ the same log-encoded L2 loss $\mathcal{L}_{\text{visible}}$, which quantifies the disparity between the rendered results $\hat{I}_{\text{visible}}, \hat{I}_{\text{depth}}$ from predicted lighting voxel octree and ground truth.

4.4. Training Details

Our model is designed for end-to-end training, employing a progressive training scheme. Initially, a single limited FOV image is used as input for the Depth Network to predict scene depth, supervised by paired depth and environment map ground truth. Subsequently, a 3D voxel octree is constructed as a new input, enabling the Light Network to predict a completed lighting voxel octree, supervised by paired HDR environment maps from novel views. Detailed information on dataset and training will follow in subsequent sections

Dataset Construction We train our model using photo-realistic renderings of indoor scenes from the FutureHouse synthetic dataset[19]. This dataset contains artist-designed indoor panoramas with high-quality geometry and HDR environment maps, from which we extract photographs to obtain input/output pairs for training. Notably, we opted for panorama datasets over the InteriorNet dataset utilized by [25] and [43], as well as the OpenRoom dataset used by [38] and [39]. This is because the low dynamic range of images in InteriorNet and the insufficient resolution of the panoramic data in OpenRoom, which render these datasets less suitable for lighting estimation tasks. Although the InteriorVerse proposed by [41] is a better option, the environment map and other information of this dataset were unavailable at the time of writing. Fortunately, the HDR panoramas and related geometry information provided by FutureHouse allow us to construct the training data that aligns with our requirements (see the supplementary PDF for additional details). Based on the 28,579 panoramic views from 1,752 house-scale scenes provided by FutureHouse, we construct and select 113,232 pairs of data. We use 90% (1,570) of the scenes to train our model and reserve 10% (180) for evaluation.

Additional Details Our network is implemented in PyTorch [24], while the differentiable rendering layer is implemented using Taichi[12]. Model training is conducted on a NVIDIA GeForce RTX 3090 GPU using the Adam optimizer[17] with a batch size of 1. Given that the input image corresponds to a small proportion of the overall scene, we employ a step-by-step training scheme to ensure stable

training. Initially, the Depth Network is trained separately to provide reasonable output values for the subsequent training of the Light Network, which is then trained based on real depth data and projected global features. Subsequently, both components of the network are jointly trained to address the discrepancy in the input data. This approach also bolsters the robustness of the primary Light Network across a wide spectrum of application scenarios, irrespective of whether depth is inferred or directly captured by sensors.

5. Experiments

We conduct a comprehensive evaluation of our unified lighting estimation framework by comparing it with existing methods, both qualitatively and quantitatively. Through this evaluation, we demonstrate the effectiveness and efficiency of our approach. Additionally, we evaluate our method against prior techniques in terms of lighting estimation and showcase its application in virtual object insertion. The results emphasize the effectiveness of our method in generating high-quality insertion outcomes at a lower cost.

5.1. Experiment Settings

Datasets. We train our network on the data constructed from the FutureHouse dataset, which was introduced in Sec.4.4. For efficiency validation experiments, we measured inference time and the VRAM usage on a TITAN V for 480x320 input image and 256x128 output environment map resolutions. For quantitative experiments, we compared the accuracy of lighting prediction using the PSNR metric on the InteriorNet dataset. Additionally, inspired by [19], we compared a new metric for measuring the coherence of neighboring illumination on pairs built from the FutureHouse dataset. For qualitative comparison experiments (perhaps the most effective lighting evaluation), we compared with prior works by visualizing object insertion results at a given location on the Laval Indoor dataset[22]. Following [22, 23, 37], we also designed a user study to validate that our spatially-coherent lighting estimation and object rendering method can achieve more photorealistic effects compared to other methods.

Baselines. In our evaluation, we compared our method with the current state-of-the-art techniques in single RGB indoor photograph-based light estimation, considering both quantitative and qualitative aspects. For the efficiency comparison experiment, we compared [38] and several algorithms that also utilize three-dimensional neural networks to predict three-dimensional lighting representations, including [25, 41, 43]. For the quantitative experiment, we compared the light estimation models including [3, 22, 23, 25, 28, 38, 43]. For the qualitative experiment and user study experiment, we compared [22, 25, 28, 38, 43].

Virtual Object Insertion Setup. To ensure a fair comparison across as many methods as possible, we uniformly adopt

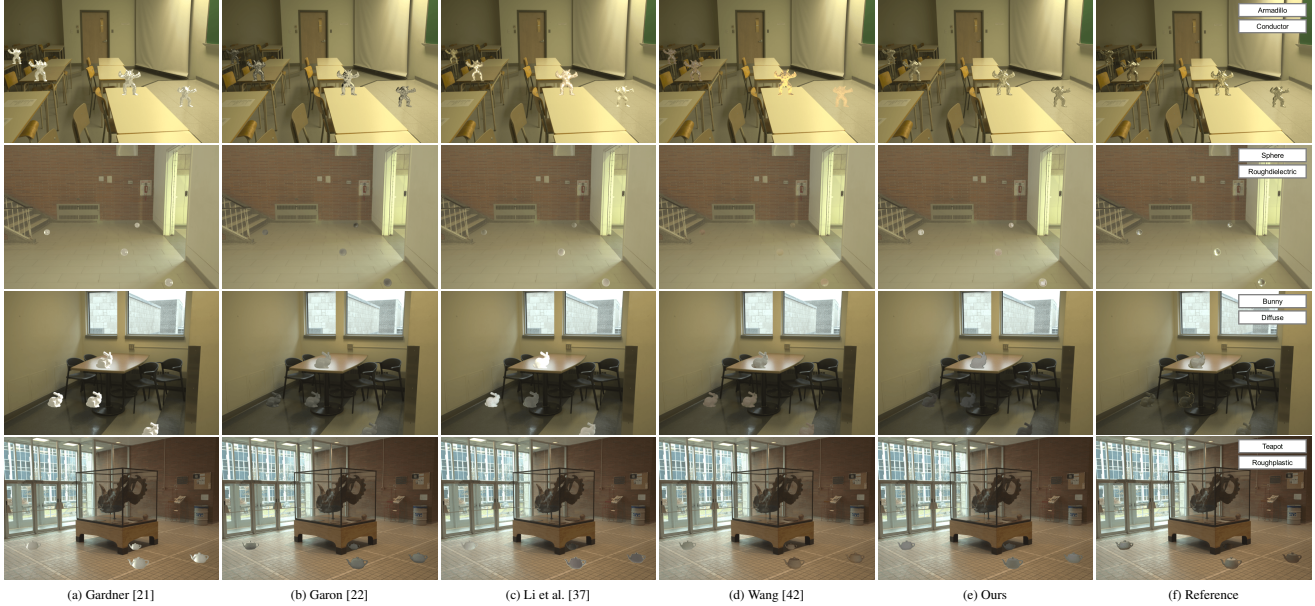| (a) Gardner [21] | (b) Garon [22] | (c) Li et al. [37] | (d) Wang [42] | (e) Ours | (f) Reference |

Figure 5. Qualitative evaluation on virtual object insertion.

Image-Based Lighting (IBL) for rendering. This involves illuminating the virtual objects using environment maps output by different methods (or environment maps converted from the output). Furthermore, for our method, we have implemented an efficient, high-quality fusion rendering approach that combines point clouds, mesh, and lighting voxel octree storage structures. The background scene is organized using a point cloud for simple occlusion detection; virtual objects are represented by a mesh grid, with triangular faces managed using a multi-level regular grid via SNode[12] to accelerate the object intersection process; for each shading point on the object, the radiance is computed using the cone tracing method introduced earlier, sampling from our predicted lighting voxel octree. For further details, please refer to the supplementary file.

5.2. Comparisons to Baseline Methods

Quantitative evaluation on InteriorNet dataset. We compared the lighting prediction performance of our method with baseline methods on the InteriorNet synthetic indoor dataset, and the results are shown in the Table.1. Our method outperforms both Gardner et al. [22] and NIR [28] significantly, as it can better model spatially-varying lighting. Compared to spatially-varying method such as DeepLight [3], Garon et al. [23] and Li et al.[25], our method still be better. Our method is slightly inferior to Lighthouse[25], Wang et al.[43] in terms of accuracy. It is worth noting that Lighthouse[25] uses stereo pairs as input, which provides more information about depth and visible surfaces compared to monocular images. On the other hand, Wang et al.[43] is based on complex inverse rendering frameworks, which

provide more prior knowledge about materials and geometry. Fig.1 illustrates our comparative experiment on efficiency, demonstrating that our method estimates spatially-coherent lighting at interactive frame rates with minimal cost.

Method	PSNR(dB)↑
Gardner et al. [22]	13.42
NIR[28]	16.59
DeepLight [3]	13.36
Garon et al. [23]	13.21
Li et al.[38]	16.66
Lighthouse[25]	17.29
Wang et al.[43]	**17.37**
Ours	16.82

Table 1. Quantitative evaluation on InteriorNet dataset.

Quantitative evaluation on FutureHouse dataset. We further evaluated the spatial consistency of estimated lighting variations on the FutureHouse dataset, as shown in Table 2. We created test samples from this dataset, each comprising two input images and N ground truth environment maps, which represent lighting at a randomly chosen 3D location within the camera frustum of the input image. For the same input scene, we estimated a set of environment maps $\tilde{L}$ at N positions. Drawing inspiration from the SC loss proposed in PhyIR[19], we devised a new metric to quantitatively assess the spatially-coherent quality of estimations from different methods and computed:

$$L_{SC} = \frac{1}{N} \sum \left(|\tilde{L} - L| \odot e^{\alpha \|\nabla D\|_1} \right), \quad (7)$$

where $\nabla\tilde{D}$ is the gradient of the panorama depth. The exponential function re-weights the metrics of neighboring light probes based on the depth gradient. It can be seen that compared to using single or pixel-wise environment map estimation algorithms such as Gardner et al.[22] and Li et al.[38], our algorithm shows significant improvement in this metric, thanks to the 3D lighting voxel octree and rendering lay we designed. Compared to methods that also use 3D lighting representation like Lighthouse[25], our algorithm achieves similar accuracy.

Qualitative evaluation on virtual object insertion. We compared the lighting estimation and virtual object insertion results with baselines in Fig.5 on the Laval Indoor spatially-varying HDR dataset[23]. Methods [22, 28] use a single low-resolution environment map, which cannot handle spatially-varying effects and only recovers low frequency lighting, resulting in severe artifacts. [23, 38] employs 2D or 3D spatially-varying spherical lobes, which can produce spatially-varying lighting, but the local lighting is still low frequency spherical lobes and cannot account for angular high-frequency details. These methods fail when inserting highly specular objects. Methods [25, 43] use a volumetric $RGB\alpha$ lighting representation, allowing for 3D spatially-varying lighting. Our algorithm achieves similar results with lower computational cost and even better object insertion effects in some cases. Moreover, compared to the differential rendering method based on IBL lighting, the virtual objects rendered using our proposed method have more realistic reflections of the real scene.

User study on virtual object insertion. Similar to previous studies[22, 23, 37], we conducted a pairwise comparison user study to assess the accuracy of our light estimation method compared to ground truth. However, in order to better compare the consistency of our method's estimation performance at different locations, we used videos instead of images as in previous studies. Participants were presented with pairs of videos containing moving virtual objects, specifically two Stanford bunnies with different materials (diffuse and metallic), composited into photo sequences. These videos were illuminated using our method and one of the state-of-the-art techniques. Participants were asked to select the image that most closely resembled the reference (ground truth), with comparisons made only between the left and right images (the middle image was provided as a reference only). Each participant provided three pairwise comparisons for each video, comparing our method against three state-of-the-art methods[22, 25, 38]. A total of 20 participants (age range: 16 to 38, 5 females, 15 males) took part in the study. Statistical analysis, including user responses and p-values (represented as p), was conducted using a binomial test. A summary of the statistical results, along with the corresponding p-values, is presented in Tab.2 and Fig.6. The results

indicate that our method was preferred by the majority of participants over the prior techniques across all evaluated properties, with statistical significance (p<0.001).

Ours vs. -	SC Metric↓	Response↑	p-value↓
Gardner et al.[22]	0.291	75.2%vs.24.8%	<0.001
Li et al.[38]	0.274	68.5%vs.31.5%	<0.001
Lighthouse [25]	0.243	65.2%vs.34.8%	<0.001
Ours	**0.226**	-	-

Table 2. Statistical results of quantitative evaluation for spatially-coherent metrics and user study on FutureHouse dataset.

Ablation evaluation on the FutureHouse dataset. A deeper octree or more complex models may potentially augment the PSNR of our method's predictions, but such enhancements necessitate trade-offs. We have undertaken additional ablation studies, exploring varying octree depths and more sophisticated lighting parametric models (*i.e.*, Spherical Gaussian [43]). These studies, as delineated in Table 3, provide partial corroboration of our perspective.
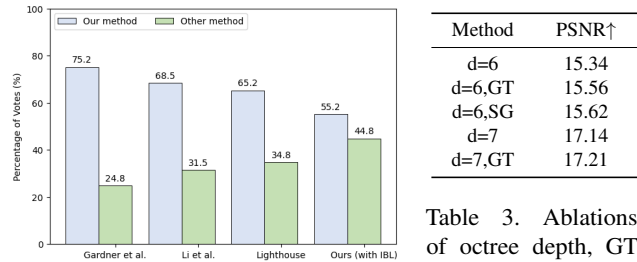

Figure 6. Pairwise user study results.

Method	PSNR↑
d=6	15.34
d=6,GT	15.56
d=6,SG	15.62
d=7	17.14
d=7,GT	17.21

Table 3. Ablations of octree depth, GT depth map and SG on Futurehouse dataset.

6. Conclusion

In conclusion, our study introduces a novel framework for efficient and high-quality estimation of spatially-coherent indoor lighting from a single image. Leveraging voxel octree and a lightweight lighting estimation network with a multi-scale rendering layer, our approach significantly reduces memory and computational resources while ensuring accurate and 3D spatially-coherent lighting estimation. These advancements present promising practical applications for more realistic AR/MR experiences. Regrettably, our approach currently falls short of SOTAs in metrics such as PSNR. Future work will aim to incorporate advanced generative models to improve our method. Another potential direction is to introduce joint estimation tasks like inverse rendering to strike a better balance between performance and accuracy, fully capitalizing on the octree-based structure.

References

[1] 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. 1, 2

[2] Andrew Chalmers, Junhong Zhao, Daniel Medeiros, and Taehyun Rhee. Reconstructing reflection maps using a stacked-cnn for mixed reality rendering. *IEEE Transactions on Visualization and Computer Graphics*, page 4073–4084, 2021. 1

[3] Chloe LeGendre, Wan-Chun Ma, Graham Fyffe, John Flynn, Laurent Charbonnel, Jay Busch, and Paul E. Debevec. Deeplight: Learning illumination for unconstrained mobile mixed reality. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 5918–5928. Computer Vision Foundation / IEEE, 2019. 1, 2, 6, 7

[4] Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: Learning dense volumetric segmentation from sparse annotation. *CoRR*, abs/1606.06650, 2016. 4

[5] Cyril Crassin, Fabrice Neyret, Miguel Sainz, Simon Green, Elmar Eisemann, and MLeal Llaguno. Interactive indirect illumination using voxel cone tracing. *Computer Graphics Forum*, page 1921–1930, 2011. 2

[6] Paul Debevec. Rendering synthetic objects into real scenes: bridging traditional and image-based graphics with global illumination and high dynamic range photography. In *ACM SIGGRAPH 2008 classes*, 2008. 2

[7] Paul Debevec, Paul Graham, Jay Busch, and Mark Bolas. A single-shot light probe. In *ACM SIGGRAPH 2012 Posters on - SIGGRAPH '12*, 2012. 1

[8] Fangneng Zhan, Changgong Zhang, Yingchen Yu, Yuan Chang, Shijian Lu, Feiying Ma, and Xuansong Xie. Emlight: Lighting estimation via spherical distribution approximation. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 3287–3295. AAAI Press, 2021. 1, 2

[9] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[10] Marc-André Gardner, Yannick Hold-Geoffroy, Kalyan Sunkavalli, Christian Gagné, and Jean-François Lalonde. Deep parametric indoor lighting estimation. 2

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

[12] Yuanming Hu, Tzu-Mao Li, Luke Anderson, Jonathan Ragan-Kelley, and Frédo Durand. Taichi: a language for high-performance computation on spatially sparse data structures. *ACM Transactions on Graphics (TOG)*, 38(6):201, 2019. 6, 7

[13] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 3

[14] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 5967–5976. IEEE Computer Society, 2017. 5

[15] Kevin Karsch, Kalyan Sunkavalli, Sunil Hadap, Nathan Carr, Hailin Jin, Rafael Fonte, Michael Sittig, and David Forsyth. Automatic scene inference for 3d object compositing. *ACM Transactions on Graphics*, 33(3):1–15, 2014. 2

[16] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. 2023. 2

[17] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, San Diega, CA, USA, 2015. 6

[18] Samuli Laine and Tero Karras. Efficient sparse voxel octrees. In *Proceedings of the 2010 ACM SIGGRAPH symposium on Interactive 3D Graphics and Games*, pages 55–63, 2010. 2

[19] Zhen Li, Lingli Wang, Xiang Huang, Cihui Pan, and Jiaqi Yang. Phyir: Physics-based inverse rendering for panoramic indoor images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 12703–12713. IEEE, 2022. 6, 7

[20] Celong Liu, Lingyu Wang, Zhong Li, Shuxue Quan, and Yi Xu. Real-time lighting estimation for augmented reality via differentiable screen-space rendering. *IEEE Transactions on Visualization and Computer Graphics*, page 2132–2145, 2023. 1, 2

[21] Joakim Löw, Anders Ynnerman, Per Larsson, and Jonas Unger. Hdr light probe sequence resampling for realtime incident light field rendering. In *Proceedings of the 25th Spring Conference on Computer Graphics*, 2009. 1

[22] Marc-André Gardner, Kalyan Sunkavalli, Ersin Yumer, Xiaohui Shen, Emiliano Gambaretto, Christian Gagné, and Jean-François Lalonde. Learning to predict indoor illumination from a single image. *ACM Transactions on Graphics*, 36(6): 176:1–176:14, 2017. 1, 2, 6, 7, 8

[23] Mathieu Garon, Kalyan Sunkavalli, Sunil Hadap, Nathan Carr, and Jean-François Lalonde. Fast spatially-varying indoor lighting estimation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6908–6917. Computer Vision Foundation / IEEE, 2019. 1, 2, 6, 7, 8

[24] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. 6

[25] Pratul P. Srinivasan, Ben Mildenhall, Matthew Tancik, Jonathan T. Barron, Richard Tucker, and Noah Snavely. Lighthouse: Predicting lighting volumes for spatially-coherent illumination. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 8077–8086. Computer Vision Foundation / IEEE, 2020. 1, 2, 5, 6, 7, 8

[26] Rui Zhu, Zhengqin Li, Janarbek Matai, Fatih Porikli, and Manmohan Chandraker. Irisformer: Dense vision transformers for single-image inverse rendering in indoor scenes. *CoRR*, abs/2206.08423, 2022. 2, 5

[27] Dieter Schmalstieg and Tobias Hollerer. Augmented reality: Principles and practice. In *2017 IEEE Virtual Reality (VR)*, 2017. 2

[28] 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 *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 8597–8606. IEEE, 2019. 1, 2, 6, 7, 8

[29] Jonas Unger, Stefan Gustavson, and Anders Ynnerman. Spatially varying image based lighting by light probe sequences: Capture, processing and rendering. *The Visual Computer*, 23 (7):453–465, 2007. 1

[30] Guangcong Wang, Yinuo Yang, Chen Change Loy, and Ziwei Liu. Stylelight: Hdr panorama generation for lighting estimation and editing. In *European Conference on Computer Vision (ECCV)*, 2022. 1, 2

[31] Peng-Shuai Wang, Yang Liu, Yuxiao Guo, Chun-Yu Sun, and Xin Tong. O-cnn: octree-based convolutional neural networks for 3d shape analysis. *ACM Transactions on Graphics*, 2017. 4

[32] Peng-Shuai Wang, Yang Liu, and Xin Tong. Deep octree-based cnns with output-guided skip connections for 3d shape and scene completion. *Cornell University - arXiv,Cornell University - arXiv*, 2020.

[33] Peng-Shuai Wang, Yang Liu, and Xin Tong. Dual octree graph networks for learning adaptive volumetric shape representations. *ACM Transactions on Graphics*, page 1–15, 2022. 4

[34] Jun-Peng Xu, Chenyu Zuo, Fang-Lue Zhang, and Miao Wang. Rendering-aware hdr environment map prediction from a single image. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(3):2857–2865, 2022. 1, 2

[35] Ye Yu and William A. P. Smith. Inverserendernet: Learning single image inverse rendering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 3155–3164. Computer Vision Foundation / IEEE, 2019. 1, 2

[36] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. Plenoctrees for real-time rendering of neural radiance fields. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2

[37] Junhong Zhao, Andrew Chalmers, and Taehyun Rhee. Adaptive light estimation using dynamic filtering for diverse lighting conditions. *IEEE transactions on visualization and computer graphics*, 27(11):4097–4106, 2021. 1, 6, 8

[38] 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 *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 2472–2481. Computer Vision Foundation / IEEE, 2020. 1, 2, 5, 6, 7, 8

[39] Zhengqin Li, Ting-Wei Yu, Shen Sang, Sarah Wang, Meng Song, Yuhan Liu, Yu-Ying Yeh, Rui Zhu, Nitesh B. 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 *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 7190–7199. Computer Vision Foundation / IEEE, 2021. 6

[40] Zhengqin Li, Jia Shi, Sai Bi, Rui Zhu, Kalyan Sunkavalli, Milos Hasan, Zexiang Xu, Ravi Ramamoorthi, and Manmohan Chandraker. Physically-based editing of indoor scene lighting from a single image. *CoRR*, abs/2205.09343, 2022. 2, 5

[41] Jingsen Zhu, Fujun Luan, Yuchi Huo, Zihao Lin, Zhihua Zhong, Dianbing Xi, Jiaxiang Zheng, Rui Tang, Hujun Bao, and Rui Wang. Learning-based inverse rendering of complex indoor scenes with differentiable monte carlo raytracing. In *SIGGRAPH Asia*, 2022. 1, 2, 3, 5, 6

[42] Rui Zhu, Zhengqin Li, Janarbek Matai, Fatih Porikli, and Manmohan Chandraker. Irisformer: Dense vision transformers for single-image inverse rendering in indoor scenes. 2022. 1

[43] Zian Wang, Jonah Philion, Sanja Fidler, and Jan Kautz. Learning indoor inverse rendering with 3d spatially-varying lighting. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 12518–12527. IEEE, 2021. 1, 2, 3, 5, 6, 7, 8