Rendering-Aware HDR Environment Map Prediction From A Single Image

Jun-Peng Xu[1], Chen-Yu Zuo[2], Fang-Lue Zhang[3], Miao Wang[1,4*]

[1]State Key Laboratory of Virtual Reality Technology and Systems, SCSE, Beihang University, China
[2]Xidian University, China [3]Victoria University of Wellington, New Zealand [4]Peng Cheng Laboratory, China
xujunpeng1998@buaa.edu.cn cyzuo@stu.xidian.edu.cn fanglue.zhang@vuw.ac.nz miaow@buaa.edu.cn

Abstract

High dynamic range (HDR) illumination estimation from a single low dynamic range (LDR) image is a significant task in computer vision, graphics and augmented reality. We present a two-stage deep learning-based method to predict an HDR environment map from a single narrow field-of-view LDR image. We first learn a hybrid parametric representation that sufficiently covers high- and low-frequency illumination components in the environment. Taking the estimated illuminations as the guidance, we build a generative adversarial network to synthesize an HDR environment map that enables realistic rendering effects. We specifically consider the rendering effect by supervising the networks using rendering losses in both stages, on the predicted environment map as well as the hybrid illumination representation. Quantitative and qualitative experiments demonstrate that our approach achieves lower relighting errors for virtual object insertion and is preferred by users compared to state-of-the-art methods.

1 Introduction

Learning and predicting the high dynamic range (HDR) illumination from a low dynamic range (LDR) partial-view image has a broad range of practical applications in augmented reality (AR). As a convenient way, relighting virtual objects using 360° HDR environment maps is able to generate appropriate lighting effects that are consistent with its real-world surroundings. However, since images are formed by radiometric and geometric processes to project real-world 3D scenes to 2D images, involving many factors such as lighting conditions, surface material, scene geometry, and camera parameters, the inverse process to estimate the global environment lighting from a narrow field-of-view (FOV) image is under-constrained. It is even more complicated to infer an HDR panoramic environment map when the image is recorded in LDR with a limited FOV.

Thanks to the emergence of deep neural networks and to the availability of large-scale lighting-related datasets, predicting light from a single image has become achievable by direct generation of environment maps (Gardner et al. 2017; Song and Funkhouser 2019), or regression of lighting parametric models, such as spherical harmonics (SH) (Cheng et al. 2018; Zhao and Guo 2020) and spherical Gaussians (SG) (Li et al. 2019a, 2020). However, parametric models often struggle to accurately and sufficiently estimate full-frequency lighting. As has been demonstrated, SH have limited capacity to interpret high-frequency lighting (Cheng et al. 2018). SG (Zhan et al. 2021) enable high-quality specular reflection and highlights. But the constant ambient lighting term limits the capability of describing low-frequency lighting. Meanwhile, direct predicting an HDR environment map that retains both high- and low-frequency could suffer from scalability issues due to the complexity of the solution space (Gardner et al. 2017). Inspecting the rendering quality by existing methods, there still remains a large space for further exploration in this field.

In this work, we build a novel two-stage framework to predict an HDR environment map from a single LDR image with a limited FOV, as shown in Figure 1. We first learn parameters of a hybrid illumination representation of the environment. Specifically, we estimate SG for main lighting sources via a Transformer, and predict the 2nd order SH for low-frequency lighting using CNNs. The estimated parameters are transformed to a Gaussian lighting map and a diffuse irradiance map via projection functions. In the second stage, a generative adversarial network (GAN) takes the transformed maps as guidance to predict an HDR environment map. We design a panoramic image-based rendering layer that efficiently computes a rendering loss between the predicted and ground truth environment maps. As a result, the generated HDR environment map enables photorealistic rendering results for virtual objects with diverse reflectances. Our main contributions are:

- A novel two-stage deep architecture that enables high-quality HDR environment map prediction, which outperforms the previous approaches.
- Rendering-aware learning schemes to both the lighting parameter and environment map prediction, including a novel differentiable rendering layer to obtain perceptual rendering loss from an environment map.
- Sub-network structures and learning strategies designed to address difficulties in environment map prediction, such as the Transformer-based SG prediction, and the hybrid illumination representation combining SH and SG as complementary components.

[*]Corresponding Author.

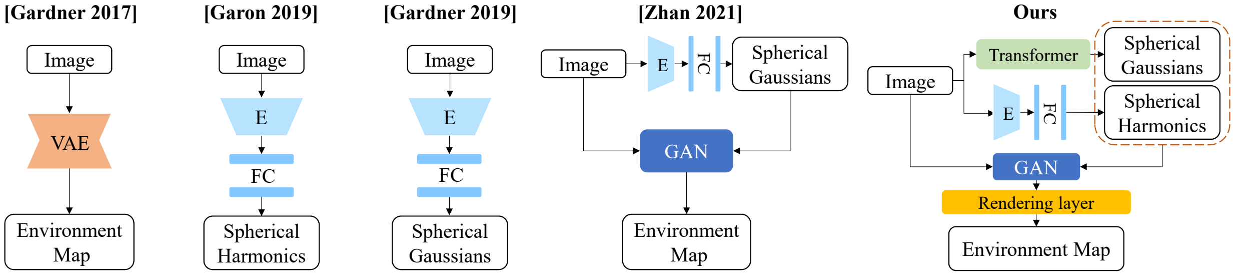

Figure 1: Different light estimation networks. (Gardner et al. 2017) directly generate environment map, (Garon et al. 2019; Gardner et al. 2019) regress SH and SG parameters as their lighting representations respectively. (Zhan et al. 2021) only use SG to guide the environment map generation. Ours takes both high and low frequency into account, utilize a transformer-based network in the regression stage, and adopt a rendering-aware generative module to produce environment maps.

2 Related Works

Non-Learning Approaches typically require user input, multiple view images, additional scene geometry, or the assumption of known shape and material of objects to alleviate the under-constrained problem. (Karsch et al. 2011) recovers parametric 3D lighting from a single image, yet needs manual annotations to initialize the lighting and get coarse geometric estimation. (Gruber, Richter-Trummer, and Schmalstieg 2012; Zhang, Cohen, and Curless 2016; Monroy, Hudon, and Smolic 2018) require multi-view captures to reconstruct the scene. (Barron and Malik 2013; Maier et al. 2017) add additional depth input to recover spatially-varying SH illumination. (Marschner and Greenberg 1997; Lombardi and Nishino 2015) estimate scene illumination from a 3D surface, with the assumption of a known shape.

Learning-based Light Prediction aims to recover the lights from real-life images. Most popular representations of lights include SH and SG. SH are a convenient lighting representation. They can be estimated from a single object-based image (Li et al. 2018) or two images taken from the front and rear cameras of a smartphone (Cheng et al. 2018). (Garon et al. 2019) uses a two-stream network to regress SH coefficients for the illumination on a local patch of a single image. Comparing with SH, SG restore the high-frequency illumination required to handle specular highlights. (Gardner et al. 2019) represents light sources by position, intensity, color and depth, and convert the above parameters into spherical Gaussian function for illumination estimation. Instead of manually designing the number of light sources like (Gardner et al. 2019), (Li et al. 2019a) uses 128 sampling directions to generate corresponding SG coefficients. (Li et al. 2020) finds SG can recover high frequency lighting much better than SH with fewer parameters, and estimate per-pixel illumination from the input image. Some other works handle illumination in an adversarial way. (Pandey et al. 2021) adds an adversarial loss to remove high-frequency shading effects from the input image. (Liu et al. 2020; Zhan et al. 2020; Zhang, Liang, and Wang 2019) adopt generative adversarial networks to generate shadows without explicit lighting estimation.

Learning-based Environment Map Generation High-quality environment maps are crucial for virtual reality applications (Wang et al. 2020). (Gardner et al. 2017) is the first work for direct environment map generation from a single image with a two-stage training scheme. (Song and Funkhouser 2019) uses a geometry-aware warping module to project the input image into a panoramic map, and then complete the unobserved region. (Srinivasan et al. 2020) proposes a 3D volumetric model based on standard volume rendering to estimate illuminations maps. (Zhao, Chalmers, and Rhee 2021) propose to use dynamic filtering for adaptively learning lighting-related features, and augment the training data in terms of white balance, color temperature and exposure to achieve a better performance. (Zhan et al. 2021) combines SG regression and environment map generation using a two-stage network, where the SG parameters are used as the guidance to generate illumination maps.

The aforementioned works either directly generate environment map or predict illumination parameters such as SH or SG for illumination estimation. In contrast, we combine the advantages of both SH and SG by projecting them into guidance maps, and build a rendering-aware generative network to generate panoramic environment map which contains more details for photorealistic rendering.

3 Method

Our method has two stages, the illumination parameter regression stage and the HDR environment map prediction stage. In the first stage (Section 3.1), we learn to regress a hybrid illumination representation that consists of SG and SH as complementary components, via two parallel regression networks. In the second stage (Section 3.2), we first transform the illumination parameters to Gaussian and irradiance maps indicating main light sources and the low-frequency lighting. We then build a generative adversarial network that takes the LDR partial-view image as input and the illumination maps as the guidance to predict an HDR environment map, with the consideration of the perceptual rendering quality. Methodological differences of our method and representative previous works are illustrated in Figure 1.

3.1 Illumination Parameter Regression

As aforementioned, we propose to use a hybrid parametric representation for illumination estimation, which consists of

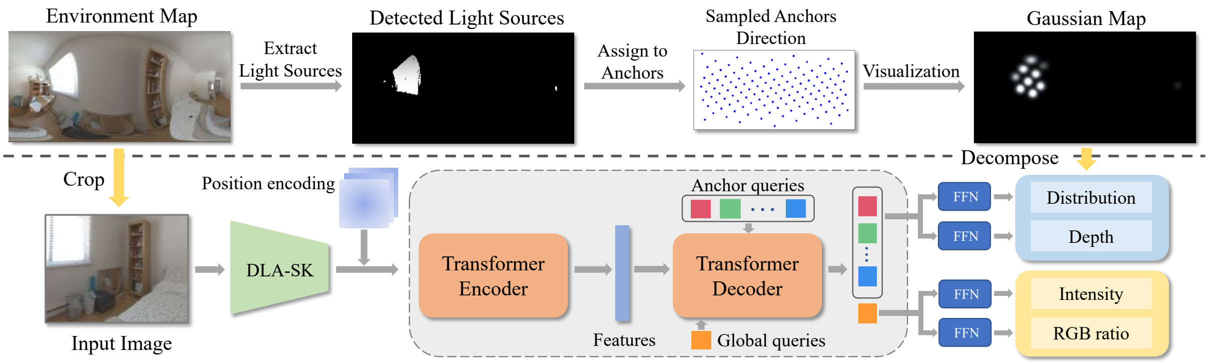

Figure 2: The structure of SG Regression Module: DLA-SK and a transformer based encoder-decoder are used to extract and interpret features. For each anchor, a feedforward network (FFN) estimates its SG parameters: light distribution, depth, global intensity, and RGB ratio.

SG and the 2nd order SH as complementary components. We use a Transformer-based network with novel losses to regress better SG than previous works, and use rendering-aware CNNs to regress SH.

Spherical Gaussian Regression From an input image, our SG regression network employs a CNN backbone to extract features, which are then feed into a Transformer-based encoder-decoder architecture, followed by feed-forward networks (FFN) to interpret features into SG parameters. Inspired by (Zhan et al. 2021), we evenly sample $N = 128$ anchor points on the sphere and assign light sources to neighboring anchor points according to the minimum radian distance. Thus, the parameters to be regressed are distribution, intensity, depth, and RGB ratio associated to each anchor. The whole process is illustrated in Figure 2.

Feature Extraction Backbone. Illumination in a scene can be widely distributed, and vary with objects' materials and scene geometry. Thus, a large receptive field is required for SG recovery. We propose a DLA-SK module, an enhanced version of deep aggregation (DLA) (Zhang et al. 2020) over the SKNet backbone (Li et al. 2019b) to allow the network to adaptively adjust the receptive fields.

Transformer Network. Due to the strong spatial relationship between the anchor points and the input image, the self-attention mechanism of Transformer can simulate the pair-wise interactions between anchor points, which suits the SG regression task well. Inspired by (Carion et al. 2020), we use an Transformer encoder-decoder and multiple FFNs to learn characteristic feature embeddings for each anchor. Using self-attentions over these embeddings, the network takes the input image as context, and globally reasons over all the anchor points by considering their pair-wise relations.

Loss Functions. The SG regression network employs several losses for a faithful recovery of high-frequency lighting details in a scene. To robustly extract main light sources from a ground truth HDR environment map, we first convert the environment map from RGB color space to HSV space, then regard those pixels whose brightness are among the top 5% as the main light sources.

Following GMLight that estimates the illumination via geometric distribution approximation (Zhan et al. 2021), we impose a geometric mover's loss (GML) $\mathcal{L}_{\text{gml}}$ and a mean square error loss $\mathcal{L}_{\text{mse}}$ to penalize the overall light intensity distribution differences between the estimated and ground truth illuminations, and use a depth loss $\mathcal{L}_{\text{depth}}$, an intensity loss $\mathcal{L}_{\text{int}}$, and an RGB ratio loss $\mathcal{L}_{\text{rgb}}$ to constrain other perspectives.

In our experiments, we found that simply using above losses leads to evenly distributed lights over the whole panoramic image, where even small intensity values assigned to anchors are not close to zero. Meanwhile, by inspecting the light intensity distribution of typical real scenes, we observed that the anchor points of high intensity values should be sparse and concentrated.

In order to suppress small intensity values while retaining the magnitude of the main light sources, we introduce the logarithmic loss $\mathcal{L}_{\text{log}}$ for light intensity distribution:

$$\mathcal{L}_{\text{log}} = \frac{\sum_{i=1}^{N} \left(ln(I_i + \varepsilon) - ln(I_i^* + \varepsilon)\right)^2}{N}, \tag{1}$$

where I_i^* and I_i are the normalized predicted and ground truth light intensity at the i-th anchor point respectively, $\varepsilon = 10^{-10}$ is a constant value used to prevent the numerical instability.

The addition of the logarithmic loss may cause a loss of total light intensity in the estimated results. To mitigate the problem, we introduce a sum loss that penalizes the overall intensity difference between the predicted and ground truth parameters of all the anchor points:

$$\mathcal{L}_{\text{sum}} = (\sum_{i=1}^{N} I_i - \sum_{i=1}^{N} I_i^*)^2. \tag{2}$$

Therefore, the loss function of the light intensity distribution is defined as:

$$\mathcal{L}_{\text{d}} = \alpha_1 \cdot \mathcal{L}_{\text{gml}} + \alpha_2 \cdot \mathcal{L}_{\text{mse}} + \alpha_3 \cdot \mathcal{L}_{\text{log}} + \alpha_4 \cdot \mathcal{L}_{\text{sum}}, \tag{3}$$

where $\alpha_1 = 10^3, \alpha_2 = 10^3, \alpha_3 = 10^{-6}, \alpha_4 = 2$ are constant weights.

For the remaining SG parameters including global intensity, RGB ratio and depth, we follow GMLight (Zhan et al. 2021) and use mean square error (MSE) losses $\mathcal{L}_{\text{int}}$, $\mathcal{L}_{\text{rgb}}$ and $\mathcal{L}_{\text{depth}}$ as supervisions. Please refer to the supplementary material and (Zhan et al. 2021) for more details.

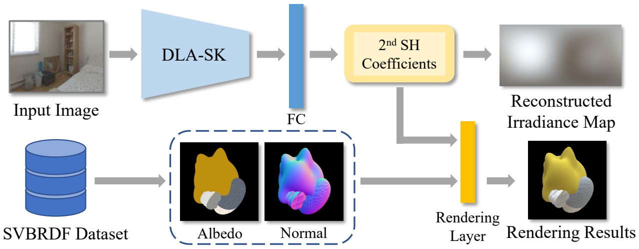

Figure 3: The structure of SH Regression Module: SH module adopts DLA-SK as backbone, and use a fully-connected (FC) layer to regress 2nd SH coefficients.

Spherical Harmonic Regression SG can fit the high-frequency lighting components of the scene very well, but it lacks the ability to represent low-frequency lighting. Since spherical harmonics can be used with restricted bandwidths to simulate low-frequency lighting, we therefore use SH as a complementary illumination representation. We adopt the same backbone architecture of the SG module for the low-frequency SH regression, as shown in Figure 3.

Rendering-Aware Loss Functions. We found that using standard MSE to supervise SH coefficients prediction makes the loss dramatically fluctuates during training. This is mainly because the error from higher order SH coefficients have an effect on the regression of lower order coefficients. Therefore, we propose to use a coarse-to-fine learning strategy. We use larger weights on the lower order coefficients at the beginning, and gradually increase the weights of the higher order coefficients and decrease the weights of lower-order coefficients. We denote the order of SH coefficients as k, ranging from 0 to 2 as we use 2nd SH. We introduce a level loss $\mathcal{L}_{\text{level}}$ which is formulated as follows:

$$\begin{cases}\mathcal{L}^k_{\text{level}} = \frac{1}{3\times(2k+1))}\sum_{c=1}^{3}\sum_{m=-k}^{k}(i^{m*}_{k,c} - i^{m}_{k,c})^2 \\ \mathcal{L}_{\text{level}} = \sum_{k=0}^{2}\lambda_k\mathcal{L}^k_{\text{level}}\end{cases} \tag{4}$$

where $i^{m*}_{k,c}$ and $i^{m}_{k,c}$ are the predicted SH coefficients for the c-th color channel, and λ_k is the weight of k-th order loss.

Just using SH coefficients difference in the loss function is not ideal, as rotating SH with a small angle will cause a large difference in MSE, but it may only introduce a slight change on its rendering result. Therefore, we employ additional losses to drive high-quality SH coefficients regression. We use the method proposed by (Ramamoorthi and Hanrahan 2001) to reconstruct a diffuse irradiance map P_{sh}. We define the MSE loss for the reconstructed irradiance map as $\mathcal{L}_{\text{diff}}$:

$$\mathcal{L}_{\text{diff}} = \left\|\hat{P}_{\text{sh}} - P_{\text{sh}}\right\|_2^2 \tag{5}$$

where $\hat{P}_{\text{sh}}$ and P_{sh} are reconstructed by predicted and ground truth SH coefficients respectively. We follow the method of (Ramamoorthi and Hanrahan 2001), and propose a differentiable rendering layer using SH coefficients to render objects from an SVBRDF dataset (Li et al. 2018), as shown in Figure 3. The SH render loss is defined as:

$$\mathcal{L}^{\text{sh}}_{\text{render}} = \left\|\mathcal{R}_{\text{sh}}(A, N, \hat{S}) - \mathcal{R}_{\text{sh}}(A, N, S)\right\|_2^2 \tag{6}$$

where A and N are the diffuse albedo and normal texture of the object respectively. $\hat{S}$ (S) are the predicted (w.r.t. ground truth) SH coefficients, and $\mathcal{R}_{\text{sh}}(\cdot)$ takes these variables to get the final rendering result. The overall loss of SH regression is:

$$\mathcal{L} = \lambda_l\mathcal{L}_{\text{level}} + \lambda_d\mathcal{L}_{\text{diff}} + \lambda_r\mathcal{L}^{\text{sh}}_{\text{render}}, \tag{7}$$

where $\lambda_l = 10$, $\lambda_d = 1$, $\lambda_r = 10^3$ are constants.

3.2 Environment Map Generation

In the second stage, we first convert the SG and SH parameters into guidance maps, which are then fed into the generative network along with the partial view image. We utilize several spherical ResNet blocks to generate the equirectangular representation of an environment map. A rendering layer is employed to calculate the rendering loss to improve the fidelity of the lighting effect generated by the predicted HDR map. Figure 4 illustrates the pipeline of the environment map generation process.

Generative Network Compared to the parametric representation of illumination, environment map contains more lighting information. However, directly regressing an environment map from a single image (Gardner et al. 2017) tends to have insufficient generalizability. In contrast, training the network in an adversarial manner (Zhan et al. 2021; Sajjadi, Scholkopf, and Hirsch 2017; Pandey et al. 2021) could enable the network to generate more realistic predictions. We thus leverage a GAN-based network to synthesize an environment map from a single image with the guidance of regressed SG and SH, which contain specular and diffuse information respectively.

The predicted SG parameters from the first stage include the light intensity distribution *D*, light intensity *I* and RGB ratio *R*. These parameters can be converted to a panoramic Gaussian map P_{sg} using the spherical Gaussian function proposed in (Gardner et al. 2019). The panoramic Gaussian map P_{sg} and the diffuse irradiance map P_{sh} reconstructed by the SH regression model are the illumination guidance for the generative network.

We use an architecture similar to SPADE (Park et al. 2019). Unlike SPADE which samples a random vector as input, we use an encoder to transform the input image into a panoramic feature map with an aspect radio of 2:1. Inspired by the work of (Zhan et al. 2021), which fuses the Gaussian maps (similar to our P_{sg}) of different scales through spatially adaptive modulation in the generation process, we concatenate P_{sg} and P_{sh} as the condition for the generation, as illustrated in Figure 4. Compare to just using P_{sg} (Zhan et al. 2021) that lacks the ability to handle low-frequency lighting, our condition contains more illumination details.

The equirectangular representation of environment maps suffers from heavy distortions and oversampling in the polar regions, resulting in unsatisfactory results in top/bottom regions if we treat them using normal 2D convolutional neural networks. To address this issue, we use spherical convolution kernels proposed in SphereNet (Coors, Condurache, and Geiger 2018) that can adapt to the equirectangular distortion when sampling the neighbors for convolution. We build several ResNet blocks to generate the final equirectangular im-

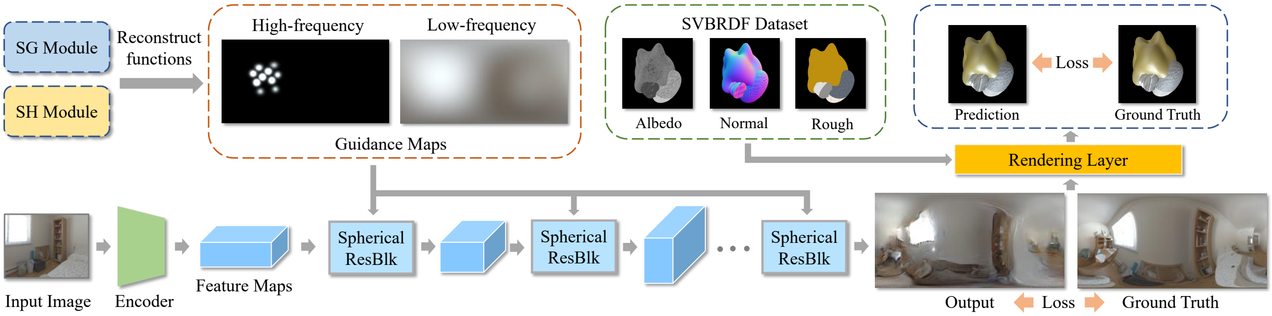

Figure 4: The structure of the generative network: we use spherical convolutions in SPADE ResBlk (Park et al. 2019) in Encoder. The conditional maps are fused into a multi-stage generation process. Finally, a differentiable rendering layer is used to calculate the rendering error of the generated environment map.

age where all the convolutional operations are performed by the spherical kernels.

Rendering Layer We propose a rendering layer to imitate the image formation process for learning lighting features of great importance, as shown in Figure 4. We adopt the physically based microfacet BRDF model in (Karis and Games 2013), and use a SVBRDF dataset comprised of diffuse albedo, specular roughness, and surface normals from (Li et al. 2018) to achieve differentiable rendering.

More particularly, for each pixel p, we sample light directions over the upper hemisphere using the corresponding normals as the up vectors. The diffuse part I_d is computed as follows:

$$I_d = \sum_{p=1}^{P} \frac{A \cdot L(\mathcal{P}, l_p) \cdot N \cdot l_p}{\pi \cdot f(p)}, \tag{8}$$

where A and N are the diffuse albedo and normal respectively. l_p is the sampled light direction, and $f(p)$ denotes its probability density function. $\mathcal{P}$ is the generated environment map, $L(\mathcal{P}, l_p)$ takes both $\mathcal{P}$ and l_p to get the sampled lighting. The specular part I_s is computed as follows:

$$I_s = \sum_{p=1}^{P} \frac{f_s(v, l_p, N, R) \cdot L(\mathcal{P}, l_p) \cdot N \cdot l_p}{f(p)}, \tag{9}$$

where v is the viewing orientation, R is the specular roughness, and $f_s(\cdot)$ is the specular BRDF components. The final rendering results can be illustrated as $\mathcal{R}(\mathcal{P}) = I_d + \lambda_s I_s$, where $\lambda_s = 10$ is the weight factor of specular term.

Environment map contains enormously more parameters than SG and SH, which takes a longer rendering time. As all parameters are known except $L(\mathcal{P}, l_p)$, we precompute the product of these parameters to speed up the training process. Furthermore, we calculate the corresponding position of each sampled light in $\mathcal{P}$, so that we can directly access the result of $L(\mathcal{P}, l_p)$.

Losses We employ several loss terms to supervise the generation of environment maps. As multi-scale discriminator can largely stabilize the training of conditional GANs, we adopt a feature matching loss $\mathcal{L}_{\text{feat}}$ as in (Wang et al. 2018) to match the intermediate features of the discriminator between the generated environment map generated at different stages and the ground truth. The MSE loss only penalizes the network based on the magnitude of errors, regardless of the impact of different image content. Instead, we use a cosine similarity loss $\mathcal{L}_{\text{cos}}$ and a perceptual loss $\mathcal{L}_{\text{vgg}}$ based on VGG-19 (Simonyan and Zisserman 2014) on the generated environment map. Our render loss is defined as the MSE loss between the rendered images $\mathcal{R}(\hat{\mathcal{P}})$ and $\mathcal{R}(\mathcal{P})$ using predicted environment map and the ground truths, respectively:

$$\mathcal{L}_{\text{render}} = \lambda_r \left\| \mathcal{R}(\hat{\mathcal{P}}) - \mathcal{R}(\mathcal{P}) \right\|_2^2, \tag{10}$$

where $\lambda_r = 10^2$ is the weight of the render loss. The discriminator uses the same architecture and adversarial loss $\mathcal{L}_{\text{adv}}$ as Patch-GAN (Isola et al. 2017). Thus, we use the following combined loss function to optimize the generative network:

$$\mathcal{L} = \underset{G}{min}\,\underset{D}{max}(\mathcal{L}_{\text{feat}} + \mathcal{L}_{\text{cos}} + \mathcal{L}_{\text{vgg}} + \mathcal{L}_{\text{render}} + \mathcal{L}_{\text{adv}}). \tag{11}$$

The generative network is trained in an adversarial manner, with all weights included in each loss. For more details, please refer to the supplementary material.

4 Experiments

We conduct both qualitative and quantitative comparisons with the state-of-the-art indoor illumination estimation methods, and we also conduct ablation studies to demonstrate the importance of each component of our method.

The Laval Indoor HDR Dataset (Gardner et al. 2017) is used for evaluation. We crop eight images from each panorama as input, and use the same warping operation in (Gardner et al. 2017) to get ground truth environment map, thus producing 19,557 training pairs. We randomly select 300 images as the test set, and the rest are used for training. The virtual object models used for calculating rendering errors are from (Li et al. 2018) with SVBRDF textures. Quantitative metrics include widely used RMSE, scale-invariant RMSE (si-RMSE) (Grosse et al. 2009) and RGB Angular Error (LeGendre et al. 2019). In addition, the fidelity of the relighting results generated by different methods are subjectively evaluated by a user study. Results of (Gardner et al. 2017, 2019; Chalmers et al. 2020; Zhao, Chalmers, and Rhee 2021) were provided by the authors, and the method of (Zhan et al. 2021) was implemented by ourselves.

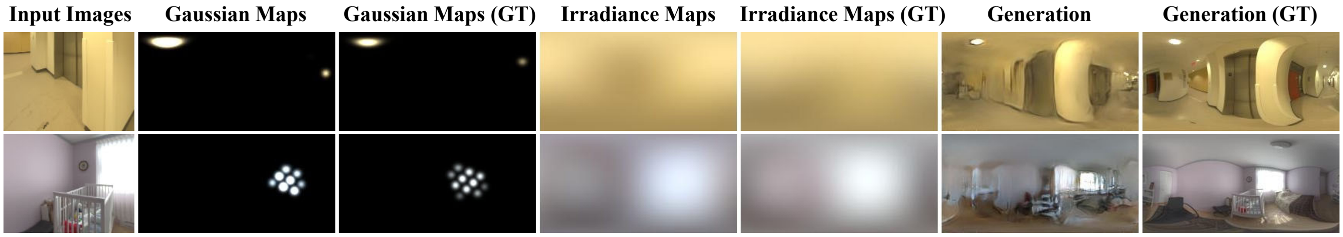

Figure 5: Prediction results of our two-stage light estimation network. We show the predicted Gaussian maps, irradiance maps, and the predicted final environment maps and their corresponding ground truths.

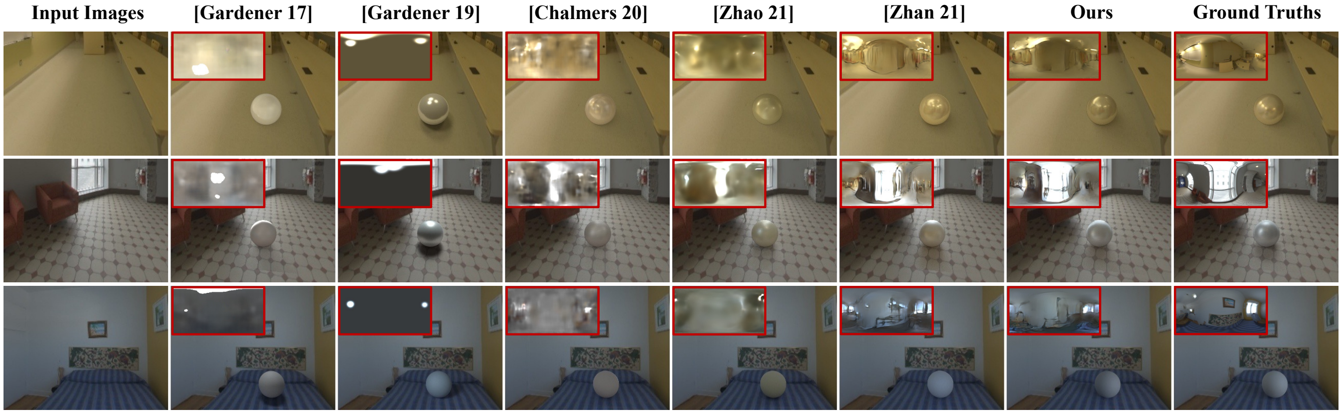

Figure 6: Visual comparison with state-of-the-art methods: three different roughness spheres (roughness = 0.2, 0.4, 0.8, from row 1 to row 3) are inserted into scenes using the predicted environment maps shown at top-left of each rendered image.

4.1 Qualitative Evaluation

Our SG regression module predicts authentic light distribution close to the ground truth, and the SH regression module estimates accurate low-frequency lighting. As a result, our method generates high-fidelity environment maps, as shown in Figure 5. We show in Figure 6 the qualitative comparison results with state-of-the-art light estimation methods by inserting virtual spheres with different roughness . As can be seen, (Gardner et al. 2017) struggles to determine the direction of light sources, resulting in wrong direction of shadow. (Gardner et al. 2019) uses simplified spherical Gaussian model, lacking of the details of the scene, thus produces unrealistic rendering with low roughness material. Without the illumination guidance, (Chalmers et al. 2020; Zhao, Chalmers, and Rhee 2021) directly predict blurry environment maps and result in unrealistic specular reflections. Our method can predicts more accurate light directions and low-frequency lighting than (Zhan et al. 2021), with the Transformer-based SG regression module and the guidance of irradiance map, which produces plausible and realistic renderings with accurate shades and shadows. More results are provided in our supplementary material.

4.2 Quantitative Evaluation and User Study

We compare our framework with four state-of-the-art methods that directly generate environment maps (Gardner et al. 2017; Chalmers et al. 2020; Zhao, Chalmers, and Rhee 2021) or estimate representative illumination functions (Gardner et al. 2019). We render 300 objects with SVBRDF textures using the environment maps predicted from the test set, and show relighting errors of each method in Table 1. We also quantitatively evaluate the predicted HDR environment maps as reported in Table 2. As a result, our method outperforms the competing methods under RMSE, si-RMSE and angular error metrics.

For perceptual evaluation, we conduct a user study using images with inserted virtual objects, rendered by ground truth and the results of each compared method in 20 randomly selected test scenes. For each method, we invited 25 subjects to select the more realistic picture between a pair of results, rendered using either the ground truth environment map or predicted one by that method. As shown in Table 1, 39.6% of the choices agree that our results are more realistic than the ground truth, which is the best among all the methods.

We observe that (Gardner et al. 2019) fails to recover low-frequency illumination, resulting in inaccurate shading and shadows measured by si-RMSE. Meanwhile, this simplified lighting parametric model can not restore the details of the scene except for main light sources , which are vital to render objects with high specular reflections, thus gets low scores in the user study. In (Gardner et al. 2017; Chalmers et al. 2020; Zhao, Chalmers, and Rhee 2021), direct generation of environment maps without any guidance tends to overfit to the training data and generate blurry results. In contrast, our method regresses SG coefficients for main light sources, SH

Table 1: Quantitative comparisons on relighting errors. We report the RMSE, si-RMSE, and Angular Error for each method. The last column is the preference rate in our user study, where higher means more confusions with ground truth.

Methods	RMSE	si-RMSE	Angular Error	User Study
[Gardner 17]	0.0873	0.0703	12.24	27.0%
[Gardner 19]	0.0741	0.0603	10.55	20.2%
[Chalmers 20]	0.0811	0.0583	10.31	32.2%
[Zhao 21]	0.0821	0.0615	10.78	30.4%
Ours	**0.0494**	**0.0420**	**7.35**	**39.6**%

Table 2: Quantitative comparisons on the quality of predicted HDR environment maps.

Methods	RMSE	si-RMSE	Angular Error
[Gardner 17]	0.4196	0.1645	30.67
[Gardner 19]	0.2117	0.2220	37.78
[Chalmers 20]	0.3054	0.1539	30.23
[Zhao 21]	0.2842	0.1588	31.05
Ours	**0.1877**	**0.1411**	**27.20**

Table 3: Ablation study on SG regression. Ranked matching error (RME) of light intensity distribution between regressed and ground truth SG parameters.

Network	Loss	RME
DLA-SK	GMLight	1.385
DLA-SK	GMLight +$\mathcal{L}_{\text{log}}$	1.120
DLA-SK	GMLight+$\mathcal{L}_{\text{log}}$ + $\mathcal{L}_{\text{sum}}$	1.233
DLA-SK+Transformer	GMLight	1.041
DLA-SK+Transformer	GMLight+$\mathcal{L}_{\text{log}}$	0.639
DLA-SK+Transformer	GMLight+$\mathcal{L}_{\text{log}}$ + $\mathcal{L}_{\text{sum}}$	**0.363**

coefficients for low-frequency lighting, then uses a generative network to predict high-quality environment maps under the guidance of the estimated lighting parameters.

4.3 Ablation Study

SG Regression We evaluate the SG regression module structure and the design of loss function by examining the estimated light intensity distribution in Table 3. We use the ranked matching error (RME) to evaluate the light intensity distribution of all anchors, defined as $\sum_{k=1}^{N} |I_{r_k} - I^*_{r^*_k}| \times dis(r_k, r^*_k)$, where r_k the k-th anchor point in I sorted by corresponding intensity value in descending order, and r^*_k is similarly defined in I^*. $dis(r_k, r^*_k)$ represents the spatial distance between the two anchor points r_k and r^*_k. Please refer to the supplementary material for the validity of RME. As a result, our method using DLA-SK and Transformer as network components with full set of losses performed better than other alternative solutions.

Table 4: Ablation study on SH regression. We report the RMSE, si-RMSE and angular error of rendering results between the predicted and ground truth SH illuminations.

	RMSE	si-RMSE	Angular Error
$\mathcal{L}_{\text{mse}}$	0.0461	0.0370	5.90
$\mathcal{L}_{\text{diff}}$	0.0452	0.0354	5.66
$\mathcal{L}_{\text{mse}}$ + $\mathcal{L}_{\text{diff}}$	0.0440	0.0341	5.46
$\mathcal{L}_{\text{level}}$ + $\mathcal{L}_{\text{diff}}$	0.0437	0.0337	5.41
$\mathcal{L}_{\text{level}}$ + $\mathcal{L}_{\text{diff}}$ + $\mathcal{L}^{\text{sh}}_{\text{render}}$	**0.0271**	**0.0224**	**3.56**

Table 5: Ablation study on environment map generation. GM and IM denote the Gaussian map and irradiance map respectively, R_d denote the rendering layer.

	RMSE	si-RMSE	Angular Error
GM	0.0526	0.0441	7.68
GM+IM	0.0508	0.0431	7.49
GM+IM+R_d	**0.0494**	**0.0420**	**7.35**

SH Regression We evaluate the impact of the rendering layer and the weighting schemes on different orders of SH coefficients. We use the predicted SH coefficients to render virtual object models and report the relighting errors in Table 4. Higher order SH coefficients get larger $\mathcal{L}_{\text{mse}}$ at the beginning of training, it effects the convergence of low-order coefficients, resulting in a low accuracy. $\mathcal{L}_{\text{level}}$ sets lower weights on higher order coefficients and avoids the above issue. Due to the sparsity of SH, large difference in coefficients may have little impact on rendering. By introducing $\mathcal{L}_{\text{diff}}$ that densely measures the irradiance map reconstruction error, we further improves the accuracy. Even better result is obtained by simulating the rendering process and supervising the network with $\mathcal{L}_{\text{render}}$.

Environment Map Generation We evaluate the effectiveness of the Gaussian map, irradiance map and the rendering layer for environment map generation in the second stage, using the same relighting metrics. As shown in Table 5, adding irradiance maps generated by the SH regression module can bring more accurate ambient lighting information, thus achieves better prediction. Besides, the novel rendering layer and the render loss introduce stronger supervision on network training, demonstrating that image formation-based rendering supervision can improve the accuracy of prediction.

5 Conclusion

In this paper, we present a novel two-stage lighting estimation pipeline that enables high-quality HDR environment map prediction from a partial view LDR image. We regress spherical Gaussians and spherical harmonics in the regression stage, to represent main light sources and low-frequency lighting respectively. In the generation stage, we use the estimated lighting parameters as the guidance for environment map generation. We apply rendering-aware learning schemes to both stages, which promotes the networks learning from the process of image formation. Quantitative and qualitative experiments show merits of our method. In the future, we will explore the usage of scene contexts and semantics for HDR environment map prediction.

Acknowledgments

We sincerely thank the reviewers for their insightful comments and suggestions that helped improve the paper. This work is supported by the National Natural Science Foundation of China (Project Number: 61932003 and 61902012), Young Elite Scientist Sponsorship Program by China Association for Schience and Technology (No. YESS20200096). Fang-Lue Zhang is supported by the Marsden Fund Council managed by Royal Society of New Zealand (No. MFP-20-VUW-180).

References

Barron, J. T.; and Malik, J. 2013. Intrinsic scene properties from a single rgb-d image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 17–24.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-End Object Detection with Transformers.

Chalmers, A.; Zhao, J.; Medeiros, D.; and Rhee, T. 2020. Reconstructing reflection maps using a stacked-cnn for mixed reality rendering. *IEEE Transactions on Visualization and Computer Graphics*.

Cheng, D.; Shi, J.; Chen, Y.; Deng, X.; and Zhang, X. 2018. Learning scene illumination by pairwise photos from rear and front mobile cameras. In *Computer Graphics Forum*, volume 37, 213–221. Wiley Online Library.

Coors, B.; Condurache, A. P.; and Geiger, A. 2018. Spherenet: Learning spherical representations for detection and classification in omnidirectional images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 518–533.

Gardner, M.-A.; Hold-Geoffroy, Y.; Sunkavalli, K.; Gagné, C.; and Lalonde, J.-F. 2019. Deep parametric indoor lighting estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7175–7183.

Gardner, M.-A.; Sunkavalli, K.; Yumer, E.; Shen, X.; Gambaretto, E.; Gagné, C.; and Lalonde, J.-F. 2017. Learning to predict indoor illumination from a single image. *arXiv preprint arXiv:1704.00090*.

Garon, M.; Sunkavalli, K.; Hadap, S.; Carr, N.; and Lalonde, J.-F. 2019. Fast spatially-varying indoor lighting estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6908–6917.

Grosse, R.; Johnson, M. K.; Adelson, E. H.; and Freeman, W. T. 2009. Ground truth dataset and baseline evaluations for intrinsic image algorithms. In *2009 IEEE 12th International Conference on Computer Vision*, 2335–2342. IEEE.

Gruber, L.; Richter-Trummer, T.; and Schmalstieg, D. 2012. Real-time photometric registration from arbitrary geometry. In *2012 IEEE international symposium on mixed and augmented reality (ISMAR)*, 119–128. IEEE.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1125–1134.

Karis, B.; and Games, E. 2013. Real shading in unreal engine 4. *Proc. Physically Based Shading Theory Practice*, 4(3).

Karsch, K.; Hedau, V.; Forsyth, D.; and Hoiem, D. 2011. Rendering synthetic objects into legacy photographs. *ACM Transactions on Graphics (TOG)*, 30(6): 1–12.

LeGendre, C.; Ma, W.-C.; Fyffe, G.; Flynn, J.; Charbonnel, L.; Busch, J.; and Debevec, P. 2019. DeepLight: Learning Illumination for Unconstrained Mobile Mixed Reality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Li, M.; Guo, J.; Cui, X.; Pan, R.; Guo, Y.; Wang, C.; Yu, P.; and Pan, F. 2019a. Deep spherical Gaussian illumination estimation for indoor scene. In *Proceedings of the ACM Multimedia Asia*, 1–6.

Li, X.; Wang, W.; Hu, X.; and Yang, J. 2019b. Selective Kernel Networks. In *CVPR*, 510–519. Computer Vision Foundation / IEEE.

Li, Z.; Shafiei, M.; Ramamoorthi, R.; Sunkavalli, K.; and Chandraker, M. 2020. 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*, 2475–2484.

Li, Z.; Xu, Z.; Ramamoorthi, R.; Sunkavalli, K.; and Chandraker, M. 2018. Learning to reconstruct shape and spatially-varying reflectance from a single image. *ACM Transactions on Graphics (TOG)*, 37(6): 1–11.

Liu, D.; Long, C.; Zhang, H.; Yu, H.; Dong, X.; and Xiao, C. 2020. ARShadowGAN: Shadow generative adversarial network for augmented reality in single light scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8139–8148.

Lombardi, S.; and Nishino, K. 2015. Reflectance and illumination recovery in the wild. *IEEE transactions on pattern analysis and machine intelligence*, 38(1): 129–141.

Maier, R.; Kim, K.; Cremers, D.; Kautz, J.; and Nießner, M. 2017. 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*, 3114–3122.

Marschner, S. R.; and Greenberg, D. P. 1997. Inverse lighting for photography. In *Color and Imaging Conference*, volume 1997, 262–265. Society for Imaging Science and Technology.

Monroy, R.; Hudon, M.; and Smolic, A. 2018. Dynamic environment mapping for augmented reality applications on mobile devices. *arXiv preprint arXiv:1809.08134*.

Pandey, R.; Escolano, S. O.; Legendre, C.; Haene, C.; Bouaziz, S.; Rhemann, C.; Debevec, P.; and Fanello, S. 2021. Total relighting: learning to relight portraits for background replacement. *ACM Transactions on Graphics (TOG)*, 40(4): 1–21.

Park, T.; Liu, M.-Y.; Wang, T.-C.; and Zhu, J.-Y. 2019. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2337–2346.

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

Sajjadi, M. S.; Scholkopf, B.; and Hirsch, M. 2017. Enhancenet: Single image super-resolution through automated texture synthesis. In *Proceedings of the IEEE International Conference on Computer Vision*, 4491–4500.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Song, S.; and Funkhouser, T. 2019. Neural illumination: Lighting prediction for indoor environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6918–6926.

Srinivasan, P. P.; Mildenhall, B.; Tancik, M.; Barron, J. T.; Tucker, R.; and Snavely, N. 2020. Lighthouse: Predicting lighting volumes for spatially-coherent illumination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8080–8089.

Wang, M.; Lyu, X.-Q.; Li, Y.-J.; and Zhang, F.-L. 2020. VR content creation and exploration with deep learning: A survey. *Computational Visual Media*, 6(1): 3–28.

Wang, T.-C.; Liu, M.-Y.; Zhu, J.-Y.; Tao, A.; Kautz, J.; and Catanzaro, B. 2018. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8798–8807.

Zhan, F.; Lu, S.; Zhang, C.; Ma, F.; and Xie, X. 2020. Adversarial image composition with auxiliary illumination. In *Proceedings of the Asian Conference on Computer Vision*.

Zhan, F.; Yu, Y.; Wu, R.; Zhang, C.; Lu, S.; Shao, L.; Ma, F.; and Xie, X. 2021. Gmlight: Lighting estimation via geometric distribution approximation. *arXiv preprint arXiv:2102.10244*.

Zhang, E.; Cohen, M. F.; and Curless, B. 2016. Emptying, refurnishing, and relighting indoor spaces. *ACM Transactions on Graphics (TOG)*, 35(6): 1–14.

Zhang, S.; Liang, R.; and Wang, M. 2019. ShadowGAN: Shadow synthesis for virtual objects with conditional adversarial networks. *Computational Visual Media*, 5(1): 105–115.

Zhang, Y.; Wang, C.; Wang, X.; Zeng, W.; and Liu, W. 2020. FairMOT: On the Fairness of Detection and Re-Identification in Multiple Object Tracking.

Zhao, J.; Chalmers, A.; and Rhee, T. 2021. Adaptive Light Estimation using Dynamic Filtering for Diverse Lighting Conditions. *IEEE Transactions on Visualization and Computer Graphics*, 27(11): 4097–4106.

Zhao, Y.; and Guo, T. 2020. Pointar: Efficient lighting estimation for mobile augmented reality. In *European Conference on Computer Vision*, 678–693. Springer.