EMLight: Lighting Estimation via Spherical Distribution Approximation

Fangneng Zhan * [1], **Changgong Zhang** * [2], **Yingchen Yu** [1], **Yuan Chang** [3], **Shijian Lu** † [1], **Feiying Ma** [2], **Xuansong Xie** [2]

[1] Nanyang Technological University
[2] DAMO Academy, Alibaba Group
[3] Beijing University of Posts and Telecommunications
{fnzhan,shijian.lu}@ntu.edu.sg, yingchen001@e.ntu.edu.sg, changyuan@bupt.edu.cn,
{changgong.zcg,feiying.mfy}@alibaba-inc.com, xingtong.xxs@taobao.com

Abstract

Illumination estimation from a single image is critical in 3D rendering and it has been investigated extensively in the computer vision and computer graphic research community. On the other hand, existing works estimate illumination by either regressing light parameters or generating illumination maps that are often hard to optimize or tend to produce inaccurate predictions. We propose Earth Mover's Light (EMLight), an illumination estimation framework that leverages a regression network and a neural projector for accurate illumination estimation. We decompose the illumination map into spherical light distribution, light intensity and the ambient term, and define the illumination estimation as a parameter regression task for the three illumination components. Motivated by the Earth Mover's distance, we design a novel spherical mover's loss that guides to regress light distribution parameters accurately by taking advantage of the subtleties of spherical distribution. Under the guidance of the predicted spherical distribution, light intensity and ambient term, the neural projector synthesizes panoramic illumination maps with realistic light frequency. Extensive experiments show that EMLight achieves accurate illumination estimation and the generated relighting in 3D object embedding exhibits superior plausibility and fidelity as compared with state-of-the-art methods.

Introduction

Lighting estimation aims to recover illumination from a single image with limited field of view. It has a wide range of applications in various computer vision and computer graphics tasks such as high-dynamic-range (HDR) relighting in mixed reality, etc. However, lighting estimation is an under-constrained problem as it aims to recover a 360-degree full-view illumination map from an image with limited field of view. In addition, high-dynamic-range (HDR) illumination is required to be inferred from low-dynamic-range (LDR) observations for the purpose of realistic object relighting.

Lighting estimation has been tackled through direct generation of illumination maps (Gardner et al. 2017; Song and Funkhouser 2019; Srinivasan et al. 2020) or regression of parameters of representative illumination functions such as spherical harmonics function (Cheng et al. 2018; Garon et al. 2019) and spherical Gaussian function (Gardner et al. 2019; Li et al. 2020). However, the functional representation methods struggles to regress accurate frequency information (especially high-frequency information) that often leads to inaccurate shading and shadow effects in relighting (Garon et al. 2019) or require complex optimization steps (Gardner et al. 2019). Directly generating illumination maps can preserve some high-frequency information, but it can hardly recover other information of the light sources such as light directions and sizes (Chen et al. 2020).

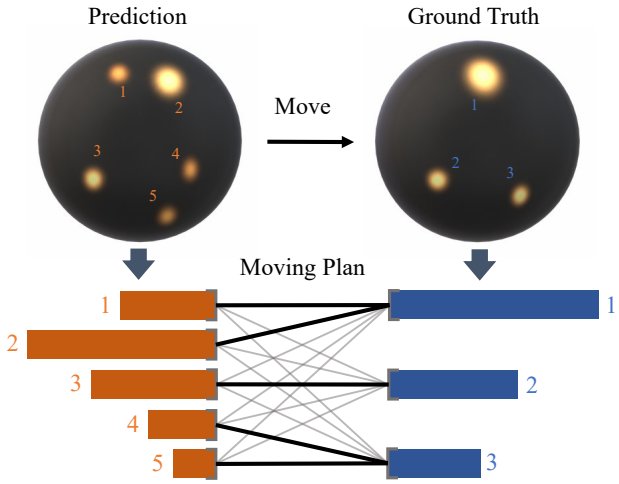

Figure 1: Illustration of our proposed Earth Mover's Light (EMLight): EMLight treats two illumination maps as two discrete spherical distributions. Motivated by the idea of Earth Mover's distances, we design a spherical mover's loss (SML) to measure the distance between two spherical distributions by calculating the minimum distance of moving one distribution to another along the spherical surface. SML aims to find the best *Moving Plan* (with minimum total distance) as illustrated by connections between the two distributions. The thickness of connecting lines denotes the amount of 'Earth' moved between the two points.

*equal contribution
†corresponding author

In this work, we propose **EMLight** (**E**arth **M**over's **Light**), an accurate illumination estimation framework that is capable of locating light sources and recover illumination with realistic frequency simultaneously. EMLight consists of an inter-connected regression network and neural projector, where the regression network predicts illumination parameters accurately and the neural projector leverages the estimated illumination parameters to synthesize illumination maps with realistic frequency information. Instead of regressing illumination parameters separately without considering them as a whole as in many existing works, we formulate the overall scene illumination by a spherical distribution and treat the illumination estimation as the regression of a spherical distribution as illustrated in Fig. 1.

For accurate illumination representation, we decompose the illumination map into *light distribution*, *light intensity*, and *ambient term* that capture the energy distribution of light sources, the overall intensity of light sources, and the average of remaining energy excluding light sources, respectively. As illumination maps are spherical images, we define N anchor points on a unit sphere to model discrete light distributions. The task of illumination prediction is thus converted to the problem of regressing light distribution, light intensity and ambient term. The light intensity and ambient term are scalar values, which can be directly regressed with a naive L2 loss by the regression network. However, directly regressing N discrete values (at N anchor points) of light distribution with a naive L2 loss or cross-entropy loss is undesirable as this does not take advantage of the subtleties of spherical distributions such as the spatial information.

Inspired by the Earth Mover's distance (Rubner, Tomasi, and Guibas 2000) that measures the distance between two distributions, we design a spherical mover's loss to regress light distributions by conducting 'Earth Mover' on the unit sphere as illustrated in Fig. 1. SML evaluates the distance between two spherical distributions by measuring the minimum radian distance required to move one spherical distribution to another along the spherical surface, and the target is to find the optimal moving plan among all possible moves between two distributions. It captures spatial information of spherical distribution, which greatly helps for accurate estimation of spherical light distribution.

Under the guidance of illumination parameters that are predicted by the regression network, the neural projector generates accurate illumination maps with realistic frequency information in an adversarial manner. Different from normal images, the illumination map is a panorama that usually suffers from different levels of spherical distortions at different latitudes. We therefore adopt spherical convolution (Coors, Condurache, and Geiger 2018) for the accurate generation of panoramic illumination maps.

The contribution of this work can be summarized in three aspects. First, we formulate the illumination estimation as the regression of the spherical distribution of illumination. To the best of our knowledge, this is the first work that estimates illumination in this manner. Second, we design a novel spherical mover's loss that takes advantage of the subtleties of spherical illumination distributions. Third, we design a neural projector that employs spherical convolution to synthesize panoramic illumination maps with realistic frequency information through adversarial training.

Related Works

Lighting estimation is a classic challenge in computer vision and computer graphics, and it is critical for realistic relighting in objects insertion and image synthesis (Lalonde, Efros, and Narasimhan 2012; Barron and Malik 2013; Hold-Geoffroy et al. 2017; Murmann et al. 2019; Zhan, Zhu, and Lu 2019b; Zhan, Lu, and Xue 2018; Zhan and Lu 2019; Zhan, Xue, and Lu 2019; Zhan, Huang, and Lu 2019; Zhan, Zhu, and Lu 2019a; Zhan et al. 2020b; Zhan and Zhang 2020; Boss et al. 2020; Xue, Lu, and Zhan 2018; Zhan et al. 2021). Traditional approaches require user intervention or assumptions about the underlying illumination model, scene geometry, etc. For example, Karsch et al. (2011) recover parametric 3D lighting from a single image but requires user annotations for initial lighting and geometry estimates. Zhang, Cohen, and Curless (2016) require a full multi-view 3D reconstruction of scenes. Lombardi and Nishino (2015) estimate illumination from an object of known shape with a low-dimensional model. (Maier et al. 2017) makes use of additional depth information to recover spherical harmonics illumination.

On the other hand, the recent works aim to estimate lighting from images by regressing representation parameters (Cheng et al. 2018; Gardner et al. 2019; Li et al. 2020) or generating illumination maps (Gardner et al. 2017; Song and Funkhouser 2019). Garon et al. (2019) estimate lighting by predicting spherical harmonic (SH) coefficients from a background image and local patch. Gardner et al. (2019) estimate the positions, intensities, and colours of light sources and reconstructs illumination maps with a spherical Gaussian function. On top of it, Li et al. (2019) represent illumination maps with multiple spherical Gaussian functions and regresses the corresponding Gaussian parameters for lighting estimation. Gardner et al. (2017) generate illumination maps directly with a two-steps training strategy. Song and Funkhouser (2019) estimate per-pixel 3D geometry and uses a convolutional network to predict unobserved contents in the environment map. LeGendre et al. (2019) regress HDR lighting from LDR images by comparing the rendered sphere with predicted illumination to the ground truth. Srinivasan et al. (2020) estimate a 3D volumetric RGB model of a scene and uses standard volume rendering to estimate incident illuminations. Given any illumination map, the framework proposed by Sun et al. (2019) is able to achieve relighting on the RGB portrait image taken in an unconstrained environment. Besides, several works (Liu et al. 2020; Zhan et al. 2020a) adopt Generative Adversarial Network to generate shadow without explicitly estimating the illumination map.

The aforementioned works either lose realistic frequency information or produce inaccurate light sources in illumination estimation, the proposed EMLight combines a regression network and a generation network to achieve accurate estimation of illumination with high frequency information.

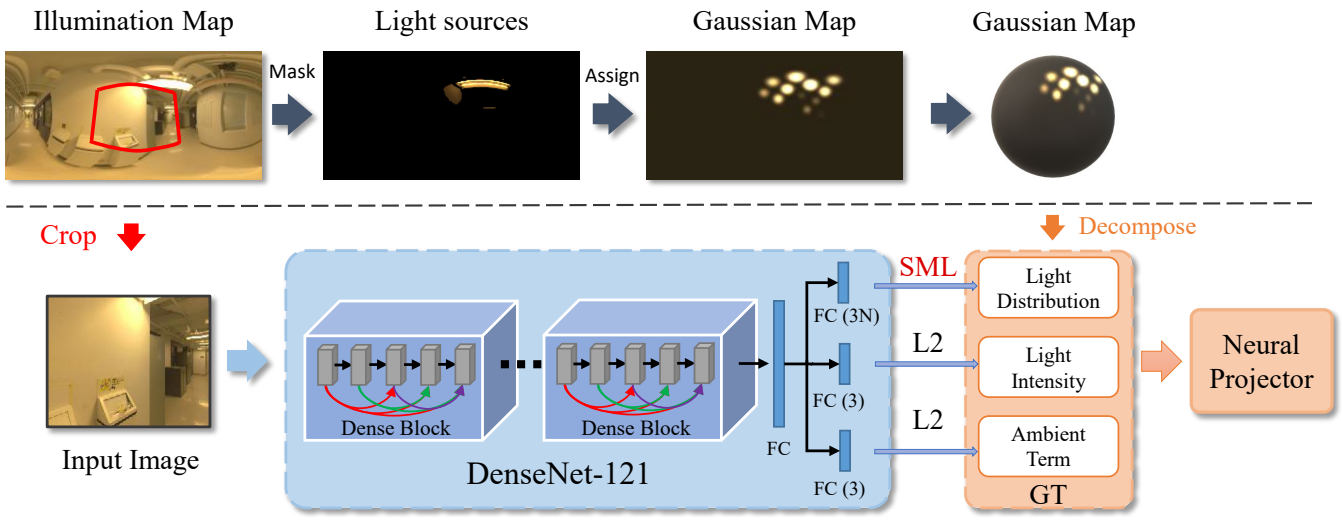

Figure 2: Illumination decomposition and estimation in EMLight: The upper and lower graphs illustrate the illumination map decomposition and the structure of the light parameter regression network, respectively. Given an *Illumination Map*, we first derive the *Light sources* region via thresholding and then assign light source pixels to N anchor points as illustrated in *Gaussian Map* (visualized by spherical Gaussian function). We decompose the illumination map into light distribution, light intensity and ambient term and use the decomposition as ground truth for regression network training. The regression network employs a local region (as highlighted by the red box) as the input and three fully-connected layers (FC) with output size of $3N$, 3 and 3 (RGB images have 3 channels) to regress the light distribution, light intensity and ambient term, respectively. The estimated illumination parameters are fed to the *Neural Projector* for illumination map generation.

Proposed Method

EMLight consists of two sequential modules including a regression network and a neural projector as illustrated in Figs. 2 and 3. The illumination parameters estimated by the regression network will guide the neural projector to generate illumination maps accurately.

Regression Network

The structure of the regression network is shown in Fig. 2. The regression network aims to estimate three set of our decomposed illumination parameters including ***light distribution*** P, ***light intensity*** I and ***ambient term*** A, which will be explained as below. For clarity, we take one channel of RGB images as an example in the following description. We first separate light sources from illumination maps since light sources in scenes are most critical in illumination prediction. Following (Gardner et al. 2019), we separate the light source by taking the 5% pixels that have the highest values within the HDR illumination map. The ***Light intensity*** I can then be determined by the summation of all the pixels within the light sources region, and the the ***ambient term*** A is further determined by the averaged pixel value within the remaining region (excluding light sources). We then employ Vogel's method (Vogel 1979) to generate N (N=128 by default in this work) uniformly distributed anchor points on a unit sphere. The value of pixels within the light source region will be assigned to the anchor point with minimum radian distance. The value of each anchor point is the summation of all assigned pixel values. The value of all anchor points will be normalized by the intensity I to ensure their summation equals one, so that the N anchor points form a standard discrete distribution on a unit sphere as denoted by ***light distribution*** P.

Three branches as shown in Fig. 2 are adopted to regress the three sets of parameters respectively. For the light intensity I and ambient term A, a naive L2 loss can be adopted for the regression. But for the light distribution P which are localized on a sphere, a naive L2 loss cannot effectively utilize the spatial information of spherical distribution and the property of standard distribution (the summation of all anchor point values equals one). We take advantage of the subtleties of spherical distribution and propose a novel spherical mover's loss to regress light distribution.

Spherical Mover's Loss

A naive method to predict the discrete spherical distribution is using L2 loss or cross-entropy loss to regress the values of N anchor points, but this naive method often introduce various problems. Firstly, L2 loss only regresses each anchor point separately and cannot effectively evaluate the discrepancy between two sets of anchor points (two distributions). Secondly, both L2 loss and cross-entropy loss cannot effectively utilize the spatial information of the discrete light distribution which is localized on the spherical surface.

Inspired by the Earth Mover's distance which measures the discrepancy between two distributions, we propose a novel spherical mover's loss (SML) to measure the discrepancy between two discrete spherical distributions. To derive the SML, we define two discrete distributions with N points on the sphere as denoted by U and V. Intuitively, SML can

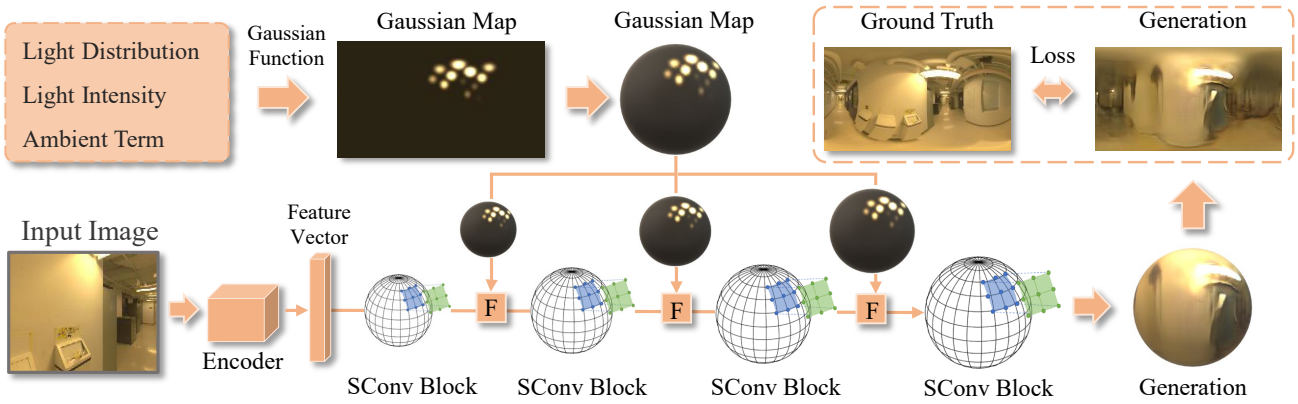

Figure 3: The structure of neural projector: *SConv Block* denote spherical convolution block, *F* denotes feature fusion blocks, and *Gaussian Map* is acquired through spherical Gaussian mapping according to the predicted parameters. The *Input Image* is fed to an *Encoder* to produce a feature vector for the ensuing spherical generation. The *Gaussian Map* is fused with the multi-scale spherical generation to synthesize the final illumination map.

be treated as the minimum amount of work required to transform U into V, where the work is measured by multiplying the amount of distribution to be moved and the distance to be moved. Then a transportation plan (or moving plan) matrix T with size of (N, N) can be defined, where each entry T_{ij} in T represents the the amount of probability moved between point U_i and point V_j. Besides, a cost matrix C with size of (N, N) is also defined where each entry C_{ij} in C gives the distance of moving U_i to V_j. Specifically, the distance between a point U_i and another point V_j is measured by their radian distance along the unit sphere. As the N anchor points on the sphere surface are pre-defined by the Vogel's method (Vogel 1979), the cost matrix C can be easily pre-computed as a constant matrix in training. With the defined transportation plan matrix T and cost matrix C, SML can be formulated as follows:

$$\begin{aligned} L_{sml} &= \min_{T}(\sum_{i=1}^{N}\sum_{j=1}^{N} C_{ij}T_{ij}) = \min_{T}\langle C, T\rangle \\ subject\ to \quad & T\cdot\vec{1} = U, \quad T^{\top}\cdot\vec{1} = V \end{aligned} \tag{1}$$

To solve this problem in a differentiable way, an entropic regularization term $H(T)$ is introduced as defined by $H(T) = -\sum_{i=1}^{N}\sum_{j=1}^{N} T_{ij}\log T_{ij}$. Then the original problem can be formulated as below:

$$\begin{aligned} L_{sml} &= \min_{T}\langle C, T\rangle - \epsilon H(T) \\ subject\ to \quad & T\cdot\vec{1} = U, \quad T^{\top}\cdot\vec{1} = V \end{aligned} \tag{2}$$

where ϵ is the regularization coefficients which denote the smoothness of the transportation plan matrix T. In our model, ϵ is set to 0.0001 empirically. The regularized form of the problem can be solved efficiently by Sinkhorn iteration (Cuturi 2013) in a differentiable way.

Using SML for the regression of spherical distribution has two clear advantages. First, it makes the regression sensitive to the global geometry, thus effectively penalizing predicted activation that is far away from the ground truth distribution. Second, SML is smooth in training which enables stable optimization which is beneficial to the under-constraint problem in illumination prediction.

Neural Projector

With the predicted light distribution, light intensity and ambient term, we propose a Neural Projector to formulate the synthesis of illumination map as a conditional image generation task with paired data as illustrated in Fig. 3. To synthesize realistic frequency information in the illumination map, the neural projector is trained in an adversarial manner. The input to the neural projector includes the predicted illumination parameters and the input image. Firstly, we map the parameters into a Gaussian map through spherical Gaussian function (Gardner et al. 2019) as shown below:

$$M = \sum_{i=1}^{N} v_i * exp\frac{d_i * u - 1}{s} + A \tag{3}$$

where M is the gaussian map, N is the number of anchor points, v_i denotes the RGB value of a anchor points which is the product of light distribution on this anchor point and light intensity (namely $v_i = P_i * I$), d_i is the direction of an anchor point (pre-defined by Vogel's method (Vogel 1979)), u is a unit vector giving a direction on the sphere, s is the angular size (selecting 0.0025 empirically), A is the ambient term. Then the constructed Gaussian map will be treated as a guidance (or a condition) for the following generation.

The overall architecture of the generation part is similar to SPADE (Park et al. 2019) as shown in Fig. 3. Instead of sampling a random vector, we encode the input image as a latent feature vector for the adversarial generation. The illumination map is panoramic image and pixels at different latitudes of a panorama are stretched in different scales. As a result, normal convolution suffers from distortions heavily at different latitudes especially around the polar regions of the panoramic image. Previous work SphereNet (Coors, Condurache, and Geiger 2018) builds on regular convolutional filters, which naturally enables the transfer of CNNs

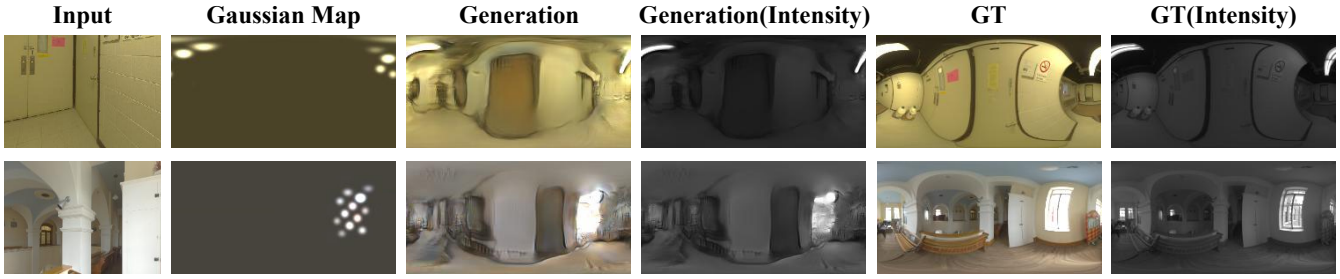

Figure 4: Illustration of EMLight illumination estimation: For the input images in column 1, columns 2 shows the constructed Gaussian maps based on the regressed illumination parameters and columns 3 and 4 show the generated illumination map under the guidance of Gaussian map and the corresponding intensity map, respectively. Columns 5 and 6 show the ground truth of the illumination maps and the corresponding intensity maps, respectively.

Figure 5: The scenes used in evaluations consist of three spheres with different materials including diffuse gray, matte silver and mirror silver.

between different image representations by adapting the sampling locations of the convolution kernels. We therefore adopt spherical convolution (SConv Block) for the generation of panoramic illumination map, effectively reversing distortions and wraps the filters around the sphere. The Gaussian map is then fused with the feature of SConv Block in multiple scales through the spatially-adaptive normalization as described in (Park et al. 2019). The details of generation part is provided in the supplementary file.

The neural projector employs several losses to drive the generation of high-quality illumination maps. We denote the input Gaussian map as x, the ground-truth illumination map as y, and the generated illumination map as x'. To stabilize the training, we introduce a feature matching loss $\mathcal{L}_{feat}$ to match the intermediate features of discriminator between the generated illumination map and ground truth:

$$\mathcal{L}_{feat} = \sum_{l} \lambda_l ||D_l(x, x') - D_l(x, y)||_1 \tag{4}$$

where D_l represents the activation of layer l in the discriminator and λ_l denotes the balanced coefficients. To obtain a similar illumination distribution instead of excessively emphasizing the absolute intensity, a cosine similarity is computed between the generated illumination map and ground truth as below:

$$\mathcal{L}_{cos} = (1 - Cos(x', y)) * \lambda_{cos} \tag{5}$$

where λ_{cos} is the weight of this term. The discriminator adopts the same architecture with Patch-GAN (Isola et al. 2017), thus obtaining the adversarial loss of discriminator D and generator G as denoted by $\mathcal{L}_D$ and $\mathcal{L}_G$. Then the neural projector is optimized following the objective as below:

$$\mathcal{L} = \min_G \max_D (\mathcal{L}_{feat} + \mathcal{L}_{cos} + \mathcal{L}_G + \mathcal{L}_D) \tag{6}$$

As the regression network and neural projector are all differentiable, the whole framework can be optimized end-to-end.

Experiments

Dataset and Experimental Setting

We evaluate EMLight with the Laval Indoor HDR Dataset (Gardner et al. 2017) that consists of 2,100 HDR panoramas taken in a variety of indoor environments. Similar to Gardner et al. (2017), we crop eight images with limited field of views from each panorama which produces 19,556 training pairs as used in our experiments. For each of the 19,556 images, the same image warping operation in Gardner et al. (2017) is applied. In the experiments, we randomly select 200 images as the testing set and the rest for training.

Consistent with Gardner et al. (2019) and Garon et al. (2019), DenseNet121 is used as the backbone in regression network. Detailed network structure of neural projector and the training settings are provided in the supplementary file.

Evaluation Method and Metric

Similar to the evaluation setting in DeepLight (LeGendre et al. 2019), our scenes for evaluations include three spheres with different materials: gray diffuse, matte silver and mirror as illustrated in Fig. 5. The performance is evaluated by comparing the scene images rendered (by Blender (Hess 2010)) with predicted illumination maps and ground truth. The evaluation metrics include Root mean square error (RMSE) and scale-invariant RMSE (si-RMSE) that focus on the estimated light intensity and light directions (or shadings), respectively. Both metrics have been widely adopted in the evaluation of illumination prediction. In addition, we also adopt the per-pixel linear RGB angular error (LeGendre et al. 2019) and Amazon Mechanical Turk (AMT) which performs crowdsourcing user study for subjective assessment of empirical realism of rendered images. In the experiments, each compared model predicts 200 illumination maps on the testing set for quantitative evaluation. For qualitative evaluation, we design 25 scenes for 3D insertion and render them with the predicted illumination maps.

Metrics	Gardner et al. (2017)			Gardner et al. (2019)			Li et al. (2019)			Garon et al. (2019)			EMLight		
	D	S	M	D	S	M	D	S	M	D	S	M	D	S	M
RMSE	0.146	0.173	0.202	0.084	0.112	0.147	0.203	0.218	0.257	0.181	0.207	0.249	**0.062**	**0.071**	**0.089**
si-RMSE	0.142	0.151	0.174	0.073	0.093	0.119	0.193	0.212	0.243	0.177	0.196	0.221	**0.043**	**0.054**	**0.078**
Angular Error	8.12°	8.37°	8.81°	6.82°	7.15°	7.22°	9.37°	9.51°	9.81°	9.12°	9.32°	9.49°	**6.43°**	**6.61°**	**6.95°**
AMT	28.0%	23.0%	20.5%	33.5%	28.0%	24.5%	25.0%	21.5%	17.5%	27.0%	22.5%	19.0%	**40.0%**	**35.5%**	**31.0%**

Table 1: Comparison of EMLight with several state-of-the-art lighting estimation methods: The evaluation metrics include the widely used RMSE, si-RMSE, Angular Error and AMT. D, S, M denote a diffuse, a matte silver and a mirror material of the rendered objects, respectively.

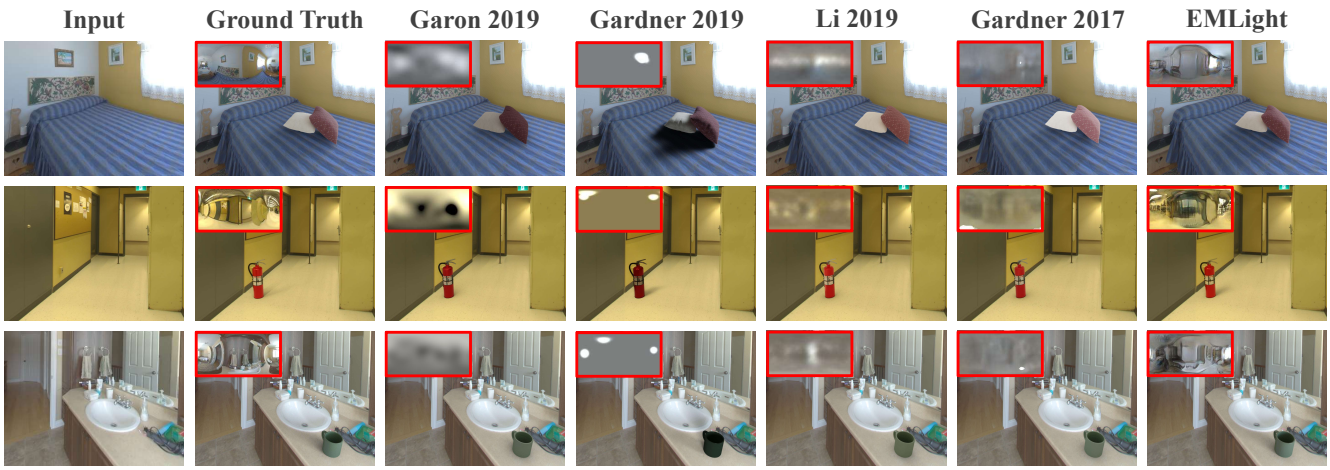

Figure 6: Visual comparison of EMLight with state-of-the-art lighting estimation methods: With the illumination maps predicted by different methods (at top-left corner of each rendered image), the rendered virtual objects demonstrate different light intensity, color, shadow and shade.

Quantitative Evaluation

We compare EMLight with several state-of-the-art methods that estimate illumination maps directly (Gardner et al. 2017) or estimate representative illumination parameters (Garon et al. 2019; Li et al. 2019; Gardner et al. 2019). For each compared method, we render 200 images of the testing scene (three spheres with diffuse, matte silver, mirror silver materials) by using the illumination maps predicted from the testing set. Table 1 shows experimental results, where *D*, *S* and *M* denote diffuse, matte silver and mirror material of the objects to be rendered, respectively. AMT user study is conducted by showing two images rendered by the ground truth and each compared methods in Table 1 to 20 users who will pick more realistic image. The score is the percentage of rendered images (200 images in total) that is deemed as more realistic than the ground-truth rendering.

We can observe that EMLight outperforms all compared methods under different evaluation metrics and materials consistently, largely attributed to the accurate generation of illumination under the guidance of the Gaussian map. Gardner et al. (2017) predict illumination maps directly but the direct generation without any guidance tends to over-fit training data and lead to sub-optimal generalization due to the unconstrained nature of illumination estimation from a single image. Gardner et al. (2019) regress spherical Gaussian parameters of light sources which tends to lose useful frequency information and generate inaccurate shading and shadow as measured by si-RMSE. Li et al. (2019) adopt spherical Gaussian functions to reconstruct the full illumination map in the spatial domain which often fails to recover high-frequency illumination. Garon et al. (2019) recover lighting by regressing spherical harmonic coefficients while it struggles to regress accurate light directions and recover high-frequency information. Though a masked L2 loss is employed in Garon et al. (2019) for preserving high-frequency information, it does not solve the problem essentially as illustrated in Fig. 6. As a comparison, EMLight estimates illumination parameters accurately by regressing light distribution under a spherical mover's loss. Under the guidance of estimated parameters, the neural projector generates accurate and high-fidelity illumination maps with realistic frequency information through adversarial training.

Qualitative Evaluation

We visualize our predicted Gaussian maps, generated illumination maps, and the corresponding intensity maps in Fig. 4. As Fig. 4 shows, our regression network predicts light distribution accurately as shown in *Gaussian Map*. The neural projector generates accurate and realistic HDR illumination maps as shown in *Generation*. To further verify the quality

Models	RMSE			si-RMSE			Angular Error			AMT		
	D	S	M	D	S	M	D	S	M	D	S	M
EMLight (SG+L2)	0.204	0.213	0.238	0.188	0.203	0.229	9.18°	9.42°	9.73°	26.0%	22.5%	18.0%
EMLight (SD+L2)	0.133	0.161	0.178	0.117	0.132	0.161	7.60°	7.88°	8.12°	30.5%	25.5%	22.0%
EMLight (SD+SML)	0.080	0.103	0.117	0.072	0.087	0.106	6.78°	6.98°	7.12°	34.0%	31.5%	26.0%
EMLight (SD+SML+NP)	**0.062**	**0.071**	**0.089**	**0.043**	**0.054**	**0.078**	**6.43°**	**6.61°**	**6.95°**	**40.0%**	**35.5%**	**31.0%**

Table 2: Ablation study of the proposed EMLight: SG and SD denote the spherical Gaussian representation and our spherical distribution representation of the illumination map. L2 and SML denote using the L2 and spherical mover's loss to regress the representation parameters. NP denotes the proposed neural projector.

Models	RMSE	si-RMSE
Anchor=64	0.091	0.075
Anchor=196	0.076	**0.055**
Cross-Entropy Loss	0.102	0.082
Normal Convolution	0.086	0.071
EMLight*	**0.074**	0.058

Table 3: Ablation studies over anchor points, loss functions and convolution types: EMLight* denotes the standard EMLight with 128 anchor points, spherical mover's loss (SML), and spherical convolution. We create four EMLight variants by setting the number of anchor points to 64 and 196, replacing SML with cross-entropy loss, and replacing spherical convolution with normal convolution.

of generated HDR illumination, we visualize the intensity maps of the illumination maps.

We compare EMLight with four state-of-the-art light estimation methods qualitatively. Fig. 6 shows rendered images with the predicted illumination maps (highlighted by red boxes). We can observe that EMLight predicts realistic illumination maps with plausible light sources and produces realistic rendering with clear and accurate shade and shadows. As a comparison, direct generation (Gardner et al. 2017) struggles to identify the direction of light sources as there is no guidance for the generation. Illumination maps by Gardner et al. (Gardner et al. 2019) are over-simplified with a limited number of light sources, and the simplification loses accurate frequency information which results in unrealistic shadow and shading in rendering. Garon et al. (Garon et al. 2019) and Li et al. (Li et al. 2019) regress illumination parameters but are often constrained by the order of representative functions (spherical harmonic and spherical Gaussian). As a result, they predict illumination of low frequency and produce renderings with very weak shade and shadow in rendering as illustrated in Fig. 6.

Ablation Study

We develop several EMLight variants as listed in Table 2 to evaluate the effectiveness of proposed methods. The variants include 1) the baseline *EMLight (SG+L2)* that regresses spherical Gaussian representative parameters with L2 loss as in (Gardner et al. 2017); 2) the *EMLight (SD+L2)* that regresses the spherical distribution of illumination with L2 loss; 3) the (*EMLight (SD+SML)*) that regresses the spherical distribution of illumination with SML; and 4) the standard model *EMLight (SD+SML+NP)*. Similar to the setting in *Quantitative Evaluation*, we apply all variant models to render 200 images of the testing scene. As Table 2 shows, (*EMLight (SD+L2)*) outperforms *EMLight (SG+L2)* clearly, demonstrating the superiority of the spherical distribution representation of illumination. *EMLight(SD+SML)* also produces better estimation than *EMLight(SD+L2)*, validating the effectiveness of SML. *EMLight (SD+SML+NP)* achieves the best estimation, demonstrating that the neural projector promotes the performance of illumination prediction significantly.

We also benchmark spherical mover's loss (SML) with the widely adopted Cross-Entropy Loss for distribution regression, compare the spherical convolution with normal convolution, and study how anchor points affect the light estimation as shown in Table 3. We followed the experimental setting as in Table 2 and measure the averaged RMSE and si-RMSE on three materials. As Table 3 show, SML outperforms cross-entropy loss clearly as SML captures spatial information of spherical distributions effectively. In addition, spherical convolution performs better than normal convolution consistently in panoramic image generation. Further, the prediction performance drops slightly when 64 instead of 128 anchor points are used, and increasing anchor points to 196 doesn't improve the performance obviously. We conjecture that the larger number of parameters with 196 anchor points affects the regression accuracy negatively.

Conclusions

This paper presents EMLight, a lighting estimation framework that formulates the illumination prediction as a light distribution regression problem over a spherical surface. A spherical mover's loss is proposed to achieve the effective regression of spherical light distribution. To generate illumination maps with realistic frequency information, we introduce a novel neural projector with spherical convolution that generates panoramic illumination maps through adversarial training. Quantitative and qualitative experiments show that EMLight is capable of predicting illumination accurately from a single indoor image. We will continue to investigate illumination estimation from the perspective of spherical distributions in our future works.

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.

Boss, M.; Jampani, V.; Kim, K.; Lensch, H.; and Kautz, J. 2020. Two-shot spatially-varying brdf and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3982–3991.

Chen, Z.; Chen, A.; Zhang, G.; Wang, C.; Ji, Y.; Kutulakos, K. N.; and Yu, J. 2020. A neural rendering framework for free-viewpoint relighting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5599–5610.

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.

Cuturi, M. 2013. Sinkhorn distances: lightspeed computation of optimal transport. In *NIPS*, volume 2, 4.

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.

Hess, R. 2010. *Blender Foundations: The Essential Guide to Learning Blender 2.6*. Focal Press. ISBN 0240814304, 9780240814308.

Hold-Geoffroy, Y.; Sunkavalli, K.; Hadap, S.; Gambaretto, E.; and Lalonde, J.-F. 2017. Deep outdoor illumination estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7312–7321.

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.

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.

Lalonde, J.-F.; Efros, A. A.; and Narasimhan, S. G. 2012. Estimating the natural illumination conditions from a single outdoor image. *International Journal of Computer Vision* 98(2): 123–145.

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*, 5918–5928.

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

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.

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.

Murmann, L.; Gharbi, M.; Aittala, M.; and Durand, F. 2019. A dataset of multi-illumination images in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4080–4089.

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.

Rubner, Y.; Tomasi, C.; and Guibas, L. J. 2000. The earth mover's distance as a metric for image retrieval. *International journal of computer vision* 40(2): 99–121.

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.

Sun, T.; Barron, J. T.; Tsai, Y.-T.; Xu, Z.; Yu, X.; Fyffe, G.; Rhemann, C.; Busch, J.; Debevec, P. E.; and Ramamoorthi, R. 2019. Single image portrait relighting. *ACM Trans. Graph.* 38(4): 79–1.

Vogel, H. 1979. A better way to construct the sunflower head. *Mathematical biosciences* 44(3-4): 179–189.

Xue, C.; Lu, S.; and Zhan, F. 2018. Accurate scene text detection through border semantics awareness and bootstrapping. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 355–372.

Zhan, F.; Huang, J.; and Lu, S. 2019. Adaptive composition gan towards realistic image synthesis. *arXiv preprint arXiv:1905.04693* .

Zhan, F.; and Lu, S. 2019. Esir: End-to-end scene text recognition via iterative image rectification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2059–2068.

Zhan, F.; Lu, S.; and Xue, C. 2018. Verisimilar image synthesis for accurate detection and recognition of texts in scenes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 249–266.

Zhan, F.; Lu, S.; Zhang, C.; Ma, F.; and Xie, X. 2020a. Adversarial Image Composition with Auxiliary Illumination. In *Proceedings of the Asian Conference on Computer Vision*.

Zhan, F.; Lu, S.; Zhang, C.; Ma, F.; and Xie, X. 2020b. Towards realistic 3d embedding via view alignment. *arXiv preprint arXiv:2007.07066* .

Zhan, F.; Xue, C.; and Lu, S. 2019. GA-DAN: Geometry-aware domain adaptation network for scene text detection and recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, 9105–9115.

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* .

Zhan, F.; and Zhang, C. 2020. Spatial-Aware GAN for Unsupervised Person Re-identification. *Proceedings of the International Conference on Pattern Recognition* .

Zhan, F.; Zhu, H.; and Lu, S. 2019a. Scene text synthesis for efficient and effective deep network training. *arXiv preprint arXiv:1901.09193* .

Zhan, F.; Zhu, H.; and Lu, S. 2019b. Spatial fusion gan for image synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3653–3662.

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