SGformer: Boosting Transformers for Indoor Lighting Estimation from a Single Image

Junhong Zhao, Bing Xue, Mengjie Zhang
Centre for Data Science and Artificial Intelligence & School of Engineering and Computer Science
Victoria University of Wellington, New Zealand
j.zhao@vuw.ac.nz, {bing.xue,mengjie.zhang }@ecs.vuw.ac.nz

Abstract

Predicting lighting from standard images can effectively circumvent the need for resource-intensive High Dynamic Range (HDR) lighting acquisition. However, this task is often ill-posed and challenging, particularly within indoor scenes, due to the intricacy and ambiguity inherent in various indoor illumination sources. We propose an innovative transformer-based method called SGformer for lighting estimation through the modeling of Spherical Gaussian (SG) distributions—a compact yet expressive lighting representation. Diverging from previous approaches, we explore underlying local and global dependencies in lighting features, which are crucial for reliable lighting estimation. Additionally, we investigate the structural relationships spanning various resolutions of SG distributions, ranging from sparse to dense, aiming to enhance structural consistency and curtail potential stochastic noise stemming from independent SG component regressions. By harnessing the synergy of local-global lighting representation learning and incorporating consistency constraints from various SG resolutions, the proposed method yields more accurate lighting prediction results, which allow for more realistic lighting effects in object relighting and composition. ***The code for the implementation of our work will be publicly available online.***

Keywords: Lighting estimation, transformer, Spherical Gaussian, augmented reality

1. Introduction

In today's world of gaming, digital effects, and the surging popularity of augmented and mixed reality (AR/MR) applications, there is a growing demand for more realistic lighting. Achieving this realism is essential to ensure consistent shading and shadow alignment between virtual and real objects. Traditionally, capturing a scene's lighting involves using light probes or omnidirectional 360° capturing devices. However, the use of specialized devices is time-consuming and often cost-prohibitive, which barriers their widespread use. To overcome these limitations, recent advancements in deep learning techniques [15, 17, 7] and the availability of extensive lighting-related datasets have prompted the development of methods that predict global illuminations from standard partial field-of-view images, providing a more accessible and cost-effective approach to approximate lighting [34, 21, 28].

The challenge of estimating indoor scene lighting, where lighting source quantities, distribution, and intrinsic properties may vary significantly between scenes, is widely recognized. Various deep learning methods tackle this challenge by generating indoor lighting conditions using different lighting representations, including environment maps or regressing lighting parameters like Spherical Harmonics (SH) and Spherical Gaussian (SG). In contrast to environment maps, which provide dense pixel-level representations, parametric lighting models offer condensed lighting representation focusing on the distribution of key lighting sources, making them favored for real-time rendering and relighting applications [25, 36, 30]. Among these parametric lighting models, the SG model is notable for its compactness and efficiency [14, 8]. It excels in capturing intricate high-frequency lighting details, enabling robust rendering of specular reflections and highlights in images, and has gained considerable attention recently [1, 32, 33].

Gardner et al. [8] introduced a set of SG parameters representing light sources, considering their direction, position, color, and size. However, directly regressing such light source-dependent SG parameters often leads to unstable model training and inference due to the unconstrained lighting source quantities and floating lighting positions, thereby limiting the accuracy of the predictions. Alternatively, Li et al.[14] and Zhan et al.[33] employed a different SG representation where multiple SG components are evenly distributed over a unit sphere [27]. In this representation, each SG component encodes the local light direction and intensity, as well as the ambient lighting. This Gaussian map representation effectively enhances inference stability and enables more effective optimization. Subse-

quent works [32, 1, 31] have built upon this SG map representation to predict lighting conditions for direct object rendering [14] or as a concise prior for improved environment map prediction [32, 31]. However, while Gaussian map predictions have shown promise, increasing the number of Gaussian components for better high-frequency information approximation often results in noisy predictions with more missing or superfluous Gaussian components.

In this paper, we introduce an innovative deep architecture aimed at enhancing Gaussian map predictions for improved indoor lighting estimation. Specifically, we leverage the Conformer architecture [18] to effectively extract lighting features from low dynamic range (LDR) input images by modeling both local and global lighting features, as well as their intricate relationships. Given that the input information is considerably limited compared to the target panorama, which is typically less than 10% of the full scene in a standard image [13], a comprehensive understanding of both local lighting cues (e.g., small specular highlights) and global lighting cues (e.g., ambient lighting, shadows) is essential for enhancing the network's ability to infer reliable lighting conditions. Furthermore, to improve the structural distribution of predicted Spherical Gaussian components, we propose a multi-head transformer decoder structure, accompanied by a distribution consistency loss across multi-resolution SG distributions for better lighting structure learning. These innovations effectively mitigate potential noise in all spectra of SG map predictions and enhance the overall spatial structure in predicted lighting. Our experimental results demonstrate that our boosted transformer-based framework effectively enhances Spherical Gaussian map predictions, leading to more realistic object rendering and more accurate guidance for environment map predictions. Our contributions are summarized as follows:

- We propose SGformer, a novel transformer-based network that combines a Conformer encoder with a multi-head transformer decoder to enhance SG predictions.
- We design a novel SG consistency loss to improve lighting structure predictions by exploring the spatial relationships across different SG resolution levels. To the best of our knowledge, this is a pioneering work to harness multiple SG resolutions for lighting prediction.
- Our SGformer model can effectively serve as a tool to enhance environment map generation and enable more realistic object rendering.

2. Related work

Lighting Representations: Extensive research efforts have been dedicated to devising methods for representing environmental lighting conditions. One widely adopted representation is the environment map [19, 20], which characterizes lighting using dense 2D images. Typically, an environment map is derived from the projection of a high-dynamic-range (HDR) spherical image, employing techniques such as equirectangular projection or cube mapping. Environment maps are extensively employed in image-based rendering pipelines. However, their high dimensionality poses significant challenges when predicting individual pixels. Additionally, the non-uniform sampling on a spherical surface often introduces distortions or irregular shapes in the image, differentiating them from traditional images and presenting estimation challenges.

Alternatively, lighting parametric models provide compact representations commonly used as prior lighting information for real-time rendering. Various parametric models, such as SG [29, 26] and SH [11, 2] lighting models, have been introduced. The SG model characterizes environmental lighting using Gaussian lobes, each defined by several parameters like size, central direction, and fatness/sharpness. Utilizing more SG components/functions often leads to a more precise description of the lighting conditions. In comparison to SH models with a predefined set of orthogonal polynomial functions[16], SG provides greater flexibility in configuring the shape, number, and distribution of basis functions. It effectively mitigates potential ring artifacts introduced by high-order SH functions while approximating full-frequency lighting. With a similar number of parameters, SG excels in capturing specular reflections and highlights. In this paper, we primarily concentrate on estimating parametric lighting from a single standard image. We make the first attempt at investigating the spatial relationships among spherical Gaussian components across various SG resolutions.

Learning-based Lighting Parameter Regression: Several works focus on regressing lighting parameters from partial-view images using deep learning methods, primarily targeting real-time rendering and relighting applications. Garon et al.[10] and Cheng et al.[6] introduced a deep learning model for predicting scene illumination by regressing spherical harmonic (SH) coefficients. Gardner et al. [8] employ a parameterization scheme where each light source is represented by a single spherical Gaussian (SG) function. They developed a deep learning model to regress key lighting attributes, including light directions, light intensities, and light colors, for each individual light source. EMlight [33] introduced a method for predicting Spherical Gaussian (SG) maps that encompass a fixed number of lighting components, referred to as anchors, which are uniformly distributed on a unit sphere. A spherical mover's loss was introduced to precisely regularize the distribution of SG components within the SG maps. These predictions are utilized as initial lighting structure guidance for the synthesis of panoramic illumination maps. GMlight [32] ex-

tended this work to incorporate depth information. It regularizes the Gaussian map learning in a geometric space using a geometric mover's loss guided by depth, enabling spatially varying lighting estimation. DSGlight introduced a graph-based framework to enhance SG map estimation. It employs a graph convolutional network (GCN) module to refine the color and depth of each SG component at a semantically structural level. Finally, Xu et al. [31] proposed a transformer-based model with self-attention mechanisms to improve contextual modeling of SG distributions, serving as a pre-processing step for further environment map estimation.

Prior methods have mainly concentrated on enhancing the stability of regression models during training [33], and they have attempted to introduce additional regularization techniques to improve the regression of lighting distributions for better preservation of full-frequency information [33, 32]. However, these efforts have primarily centered on improving the regression decoder [1, 31], often neglecting the significance of the feature extraction module, which has a more critical impact on the final results. Additionally, regularization has typically been applied at the highest resolution, to maintain high-frequency details, which can be particularly challenging when working with very limited input information.

To enhance the accuracy of our predictions, we propose improving the extraction of lighting features by considering both local and global lighting characteristics and their interrelationships through the Conformer network [18]. Furthermore, we introduce a multi-head transformer decoder accompanied by an SG consistency loss to regularize SG distributions by using multi-resolution information, spanning from sparse to dense, to promote a more profound understanding of their structural aspects.

Learning-based Environment Map Generation: Several deep learning method have been introduced to estimate environmental maps from standard images [34]. The origins of these studies can be traced back to the pioneering work of Gardner et al.[9], which has since inspired a series of subsequent efforts[4, 22, 24]. These efforts address various critical aspects in this field, including the ability to handle spatial variations in indoor scenes [23, 24], lightweight solutions for mobile applications [13], and the capacity to generalize to a wide range of input variations [35]. Notably, some recent work [33, 32, 31] has explored the use of sparse lighting representations as guidance for generating dense, pixel-wise environment maps. In such cases, SG maps and/or SH diffuse maps serve as instructive priors for directing the generation of lighting sources for the final environment map.

3. Method

We employ the following equation to represent illumination in the form of the SG map [33], denoted as D:

$$D = \sum_{i=1}^{N} (A_i * L_{hdr>I_s}) e^{\alpha*(d_i*u)-1} + L_{hdr<I_s} \quad (1)$$

The original HDR environment map of the scene is partitioned into two components: the light source component ($L_{hdr>I_s}$) and the ambient component ($L_{hdr<I_s}$), based on an intensity threshold ($I_s = I_{max} * 0.05$ in our case), which depends on the maximal pixel value I_{max} of the environment map. The lighting source regions are approximated using multiple SG functions [8], evenly distributed across a sphere, where N represents the number of SG functions or anchor points. d_i represents the direction of anchor point i, predefined using the method proposed by Vogel [27]. The symbol u denotes an arbitrary direction vector on a unit sphere, and α stands for the inverse of angular size, which we set to constant 1. Each light source is associated with neighboring anchor points based on a minimum radial distance criterion. The RGB value of each anchor point A_i is calculated as follows [33]:

$$A_i = \sum_{p_i} L_{p_i,hdr>I_s} \text{where } p_i \in \text{argmin}\{d\{p_{hdr>I_s}, d_i\}\} \quad (2)$$

where p_i represents the collection of pixels nearest to the anchor point with direction d_i.

In the following sections, we will start by analyzing the capability of SG maps with different resolutions in representing lighting features. This analysis will demonstrate the importance of our SG learning framework's design. Following that, we will introduce our network structure and its associated learning scheme.

3.1. Multi-Resolution SG Analysis

The resolution of the SG map, crucial for accurately capturing lighting information from a scene, is determined by the number of SG functions, denoted as N. Figure 1 illustrates examples of scenes and their SG maps at different resolutions. Lower-resolution SG representations, characterized by fewer anchor points, provide a more abstract depiction of lighting conditions. They offer a rough but still informative overview of the primary light source's position and their central intensity. On the other hand, higher-resolution SG representations, with a greater number of anchor points, excel at conveying intricate lighting source details. It effectively captures nuances such as shape, intensity variations, and directional shifts. Across different resolutions, SG representations maintain consistent distribution patterns. As the resolution increases from sparse to dense, a core spatial distribution is preserved, while each single light source can

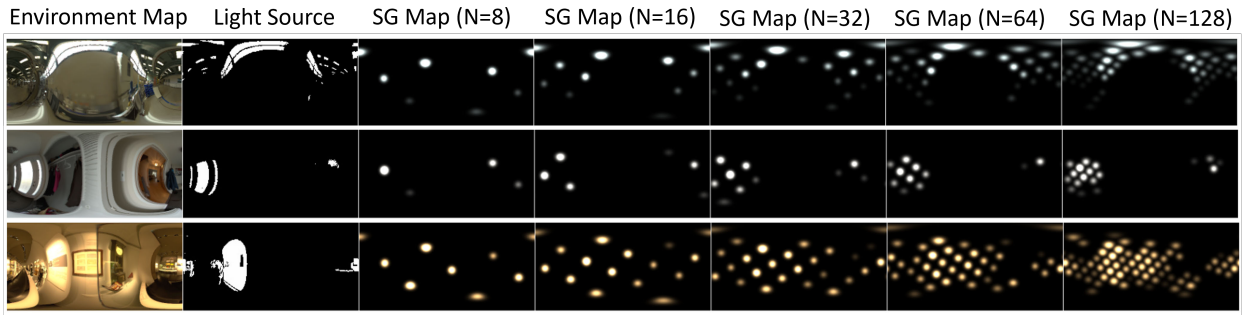

Figure 1. Illustrations of scenes and their SG maps at various resolutions.

SG	Light Source #1		Light Source #2		Light Source #3		Light Source #4		Light Source #5	
	Intensity	Angular	Intensity	Angular	Intensity	Angular	Intensity	Angular	Intensity	Angular
8	12.92	39.76	11.62	37.22	12.00	36.53	11.13	35.74	11.25	35.36
16	13.45	27.77	11.23	26.56	11.35	25.59	10.73	25.66	10.62	24.23
32	13.25	17.47	10.93	18.12	11.27	17.06	10.32	17.05	10.14	16.80
64	13.39	12.05	10.31	12.27	10.95	12.53	9.95	12.18	9.62	12.34
128	12.12	9.51	9.35	10.19	10.86	9.29	9.42	9.27	9.05	9.12

Table 1. The intensity and angular errors of various SG resolutions in representing the primary lighting sources. The table lists the top 5 lighting sources, ranked by their intensity.

encompass multiple anchor points in light shaping. Thus it creates complex semantic relationships between different anchors.

To assess the accuracy of different SG map resolutions in representing the lighting sources, we conduct a comparison of angular and intensity deviations between the ground-truth lighting sources in the HDR environment map and those extracted from different SG maps. Following the approach of Gardner et al. [8], we detect the ground-truth lighting source regions using a region expansion method applied to the HDR environment map. This process begins with light peaks as initial seeds and incrementally expands until the intensity falls below one-third of the peak value. Subsequently, region merging is carried out based on overlapping regions. To extract lighting sources from SG maps, we aligned anchor points with their closest ground-truth lighting source regions. We ranked the top five lighting source regions based on peak intensity, arranging them from highest to lowest. For each of these regions, we conducted a statistical analysis of all anchor points within that specific region. The anchor point displaying the highest lighting intensity was designated as the extracted light source center. The outcomes, as presented in Table 1, demonstrate that as SG resolution increases, the deviations in intensity and angular accuracy stemming from the lighting representation diminish. This highlights the significance of high-resolution SG maps in achieving precise lighting representation.

The task becomes more challenging when aiming to predict high-resolution SG maps, primarily because it involves dealing with a substantially larger number of parameters. Unlike low-resolution predictions, which focus on estimating average positions and intensities, high-resolution predictions must delve into more intricate details, including the shapes of light sources and subtle intensity variations. Therefore, a more comprehensive set of representative lighting features becomes essential. Additionally, as the number of SG components increases, preserving the structural semantics of numerous discrete points becomes crucial for creating a meaningful representation of lighting sources. Ensuring structural consistency across resolutions can serve as a valuable regularization technique in this context, effectively reducing noise in high-resolution predictions and addressing sparsity issues in low-resolution predictions.

3.2. Network Architecture

Given a partial view image captured in a scene, we introduced SGformer for predicting the lighting conditions in a scene, and Figure 2 presents an overview of the entire architectural structure. We utilize a Conformer encoder [18] to extract lighting features, which are then input into a multi-head transformer-based decoder. Subsequently, a fully connected layer is employed to convert these lighting features into SG lighting parameters for each resolution. This includes estimating the anchor point distributions, along with associated global average lighting intensity and RGB ratios [33, 31]. The network model is trained holistically, and we augment it with a structural consistency loss for additional regularization.

Conformer Encoder: Considering the limited scene information available from the input image and the wide varia-

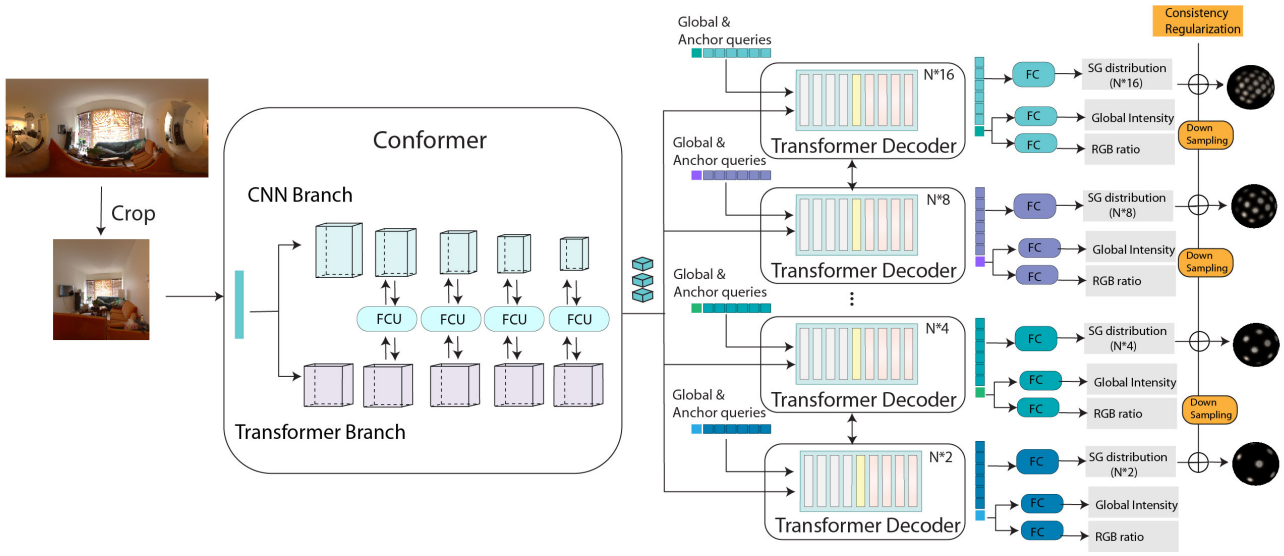

Figure 2. The overall architecture of SGformer takes a standard LDR image as input and produces SG parameters at various resolutions as output. It comprises a Conformer encoder and a multi-head transformer decoder. Additionally, we introduce a novel SG consistency loss to enhance the regularization of the lighting structure learning.

tions in lighting conditions in indoor illumination, learning discriminative lighting-specific features concealed within the input photos is crucial for a meaningful understanding and inference of lighting. Real-world lighting distributions exhibit intricate properties, with lighting features sometimes confined to small areas, like a specular highlight on an object's surface, and at other times extending widely, such as with large window light or shadows. Different local and global lighting features often interact with each other; for instance, a specular highlight can be generated by a light source located far away from the viewer's perspective.

Convolutional networks excel at capturing local features but tend to struggle when it comes to learning long-distance global relationships. On the other hand, the vision transformer is proficient in learning global representations but often overlooks finer local feature details. In our preliminary experiments, we observed that relying solely on a Transformer as the feature extractor is susceptible to inaccuracies in predicting lighting intensity. To tackle this challenge, we get inspiration from the recent trend of combining these two technologies in various visual and non-visual tasks [18, 5, 37, 12]. In the context of light estimation, Xu et al. [31] leveraged the DETR method [3, 38] that blends convolutional layers and Transformers to extract lighting features. While this approach provided some relief to the issue, it still exhibited limitations in effectively modeling the interaction between local and global features within a cascade paradigm.

We propose a new lighting feature extraction module based on Conformer [18] to enhance local and global feature extraction at various resolutions in an interactive manner. It comprises both a CNN branch and a transformer branch, with each layer featuring a dedicatedly designed feature coupling unit (FCU) to manage conflicts and complementarities among lighting features at different levels. When combined with our multi-head decoder design, the Conformer encoder's feature extraction capabilities are bolstered by considering data from different SG resolutions.

Multi-head Transformer-Based Decoder: We propose a new multi-head decoder based on the aforementioned transformer-based DETR [31] method. In contrast to previous lighting prediction methods [31], which primarily focused on a single SG resolution, our approach introduces a multi-head decoder that simultaneously estimates various SG parameters across different resolutions. This architectural approach enhances feature learning, enabling the encoder to generalize to a variety of tasks with differing levels of complexity. The DETR decoder consists of multiple transformer blocks that include self-attention and cross-attention mechanisms, along with a learnable anchor query system for the simultaneous prediction of multiple targets. In our multi-head decoder structure, distinct anchor queries are used to handle different SG embeddings, in addition to a global query for obtaining global embeddings. Subsequent fully connected layers serve as the prediction head, transforming these feature embeddings into the respective SG parameters.

3.3. Loss Functions

SG Consistency Loss: Within the multi-head SG decoder, separate transformer decoders handle the decoding of lighting features for different SG resolutions. Consequently,

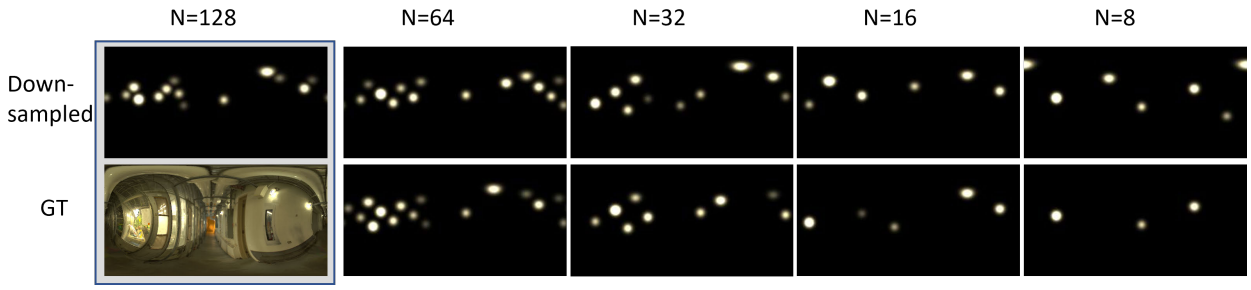

Figure 3. Comparison between downsampled SG maps from higher resolutions (top row) and the corresponding ground truth SG maps (bottom row). The top-left corner shows the corresponding environment map, while the bottom-left corner displays the ground-truth SG map with the highest resolution of N=128.

there is no inherent guarantee of consistency among them. To enhance structural uniformity in predictions and simultaneously optimize both lighting encoding and decoding in a holistic manner, we introduce the SG consistency loss. This loss aligns SG maps across different resolutions via a spherical SG downsampling technique. It calculates the loss between the downsampled SG map and the real SG map of the current resolution, serving as a regularization term for model optimization.

The SG map downsampling is performed by searching for neighboring anchor points on the spherical surface between two adjacent SG map resolutions. We define the downsampling procedure as follows:

$$A_{i,dn} = \max(A_{(j,r_{up})}) \text{ where } j \in ||d_{j,r_{up}} - d_{i,r_c}|| < R), \quad (3)$$

where the downsampled value for anchor point i at the current resolution, represented as $A_{i,dn}$, is determined by selecting the maximum intensity value from all its neighboring anchor points j in the higher resolution that is located within a radian range R of it. In our configuration, we set the resolutions to N=[8, 16, 32, 64, 128], and R is adjusted to [0.65, 0.4, 0.3, 0.3] accordingly to identify these neighboring points. The downsampling process is applied dynamically to the predicted SG parameters during training to compute the consistency loss. Figure 3 provides an example of SG downsampling between different SG resolutions and their comparisons with actual SG maps. The SG consistency loss is calculated by applying both Earth Mover's distance and L_2 norm on the predicted SG A_{pd}, the ground truth A_{gt}, and the downsampled SG A_{dn}, as outlined below:

$$L_{cns} = \beta_1 C_{gm} + \beta_2 C_{l2} \quad (4)$$

where

$$EM(A_i, A_j) = \min_T \left(\sum_{i=1}^{N} \sum_{j=1}^{N} \text{Dist}_{ij} T_{ij} \right) = \min_T < \text{Dist}, T >$$

$$C_{em} = EM(A_{dn}, A_{pd}) - EM(A_{dn} - A_{gt}) + EM(A_{pd} - A_{gt})$$

$$C_{l2} = L_2(A_{dn}, A_{pd}) - L_2(A_{dn} - A_{gt}) + L_2(A_{pd} - A_{gt})$$

Here, $EM(\cdot)$ represents the spherical Earth Mover's distance [31]. It measures the minimum amount of probability required to move points from one distribution to another, taking into account a cost matrix $T_(ij)$ determined by the predefined anchor positions and their radian distances along the sphere ($Dist_(ij)$) [27]. Unlike L2 or cross-entropy metrics, the Earth Mover's distance can effectively leverage spatial information between distributed points when assessing the dissimilarity between two distributions.

Multiple Resolution Loss: To individually supervise the generation of each SG resolution, we propose the following multiple-resolution loss function:

$$L_{sg} = \sum_{i=1}^{l} (\alpha_1 L^i_{em} + \alpha_2 L^i_{l2} + \alpha_3 L^i_{log} + \alpha_4 L^i_{rgb} + \alpha_5 L^i_{IG}) \quad (5)$$

For each resolution level i, L_{em}, L_{l2}, L_{log} represent the loss terms that control the intensity distribution of SG resolution n. L_{log} is a log-transformed version of the root mean square error of the intensity distribution, designed to mitigate extreme values [31]. L_{em} is the Earth Mover's loss. L_{IG} and L_{rgb} are losses related to global intensity and RGB ratios. The weights α from 1 to 5 are empirically set to [10^3, 10^3, 10^{-6}, 10^2, 10^{-1}].Additionally, for high-resolution SG map prediction with $n = 128$, we incorporate the render loss proposed by [31], which proves beneficial for modeling high-frequency lighting features.

During the training process, the multiple resolution loss function is initially employed to supervise each individual branch together. Once the model's training reaches a relatively stable state, the SG consistency loss across different resolutions is then applied to further refine the model holistically.

4. Experiments

We evaluate our proposed method on the Laval Indoor HDR Dataset [9]. Each panoramic image in the dataset is cropped into eight images and tone-mapped into standard

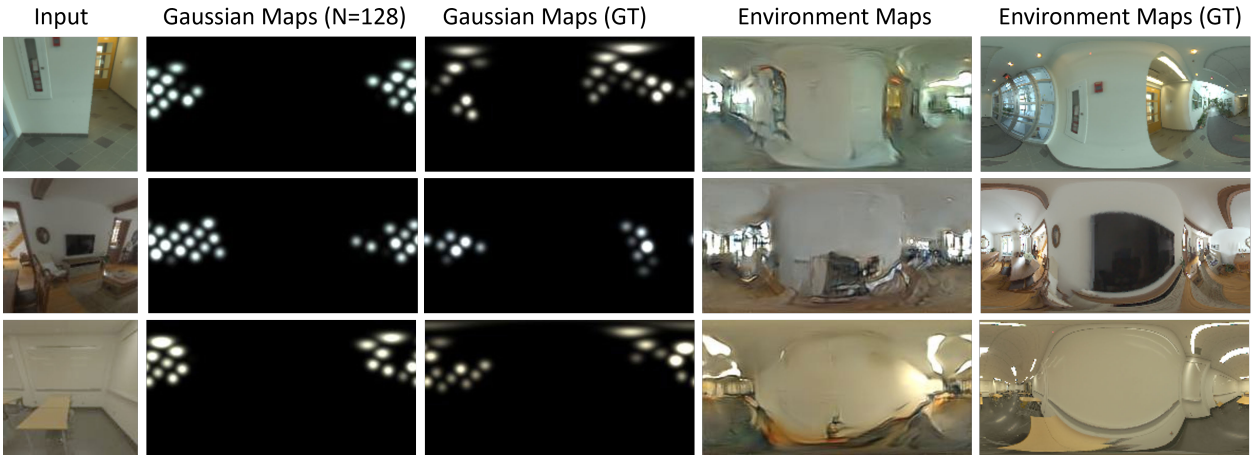

Figure 4. Visualization of SG maps generated by SGformer and the corresponding synthesized environment maps guided by them through a GAN-based environment map neural projector.

partial-view images as our inputs. To account for the spatially varying indoor lighting within a fixed position, we follow the approach outlined in [4, 35, 9, 31]. This involves applying spatial warping and re-centralization operations to the panoramic images, resulting in the final ground truth. By applying this transformation, the lighting representation is adjusted to better reflect the illumination conditions at the position where the object would be composited, rather than relying solely on the 360° camera's position, which could be situated at varying distances from the composition place. Our training dataset consists of randomly selected 1200 scenes, providing a total of $1200 * 8$ training pairs. The remaining 512 scenes are used for testing. The input crop size is set to (128, 128), and SG parameters in different resolution levels are extracted from the HDR environment map with dimensions (256, 128).

We conduct a comprehensive evaluation, encompassing both qualitative and quantitative assessments, to thoroughly assess the performance of SG predictions and their impact on environmental map generation. Our evaluation includes comparisons with state-of-the-art techniques and in-depth ablation studies. For a quantitative assessment of SG parameter predictions, we utilize the Root Mean Square Error (RMSE), ranked matching error (RME) [31], and L2 error for measuring global intensity and RGB ratio. And to measure the quality of environmental map generation, we employ well-established metrics such as RMSE, scale-invariant RMSE (si-RMSE), and angular lighting error.

4.1. Comparisons

To illustrate the impact of our improved SG predictions on environment map generation and rendering, we utilize a neural projector model proposed by Xu et al.[Xu22] [31] for environment map generation. This model is distinguished among GAN-based neural projectors for its capability to generate high-frequency environments, attributed to its integration of both high-frequency SG and low-frequency SH as lighting priors. We conduct comparisons by generating environment maps based on our SG predictions and contrasting them with state-of-the-art approaches introduced by Gardner et al. [9], LeGendre et al. [13], Chalmers et al. [4], Zhao et al. [35], Zhan et al. [33], and Xu et al. [31], here we denoted as [Gardner17], [LeGendre19], [Chalmers20], [Zhao21], [Zhan21], and [Xu22], respectively. It's noteworthy that the results presented in [Gardner17], [LeGendre19], [Chalmers20], and [Zhao21] are generated by each author using the testing data we provided. In the case of comparisons with [Zhan21], we have generated the results ourselves based on their publicly available code and model. It is important to mention that our method and [Xu22] share the same neural projector and SH generation model, with the difference being that we utilize SGformer to generate SG lighting priors.

Figure 4 displays the predicted SG maps (N=128) generated by SGformer alongside the synthesized environment maps. These environment maps are produced using the predicted SG maps as a prior input to the GAN-based environment map neural projector. Our results indicate that SGformer can generate authentic SG maps that are close to the ground truth. These predictions play a dominant role in shaping the final environment map synthesis, particularly in influencing the lighting structure under the SPADE paradigm within the neural projector module. Precise SG map priors yield the creation of high-fidelity environment maps featuring realistic lighting representations.

In comparative evaluation against various environment map estimation methods, our approach excels in both qualitative and quantitative evaluations. Figure 5 presents the

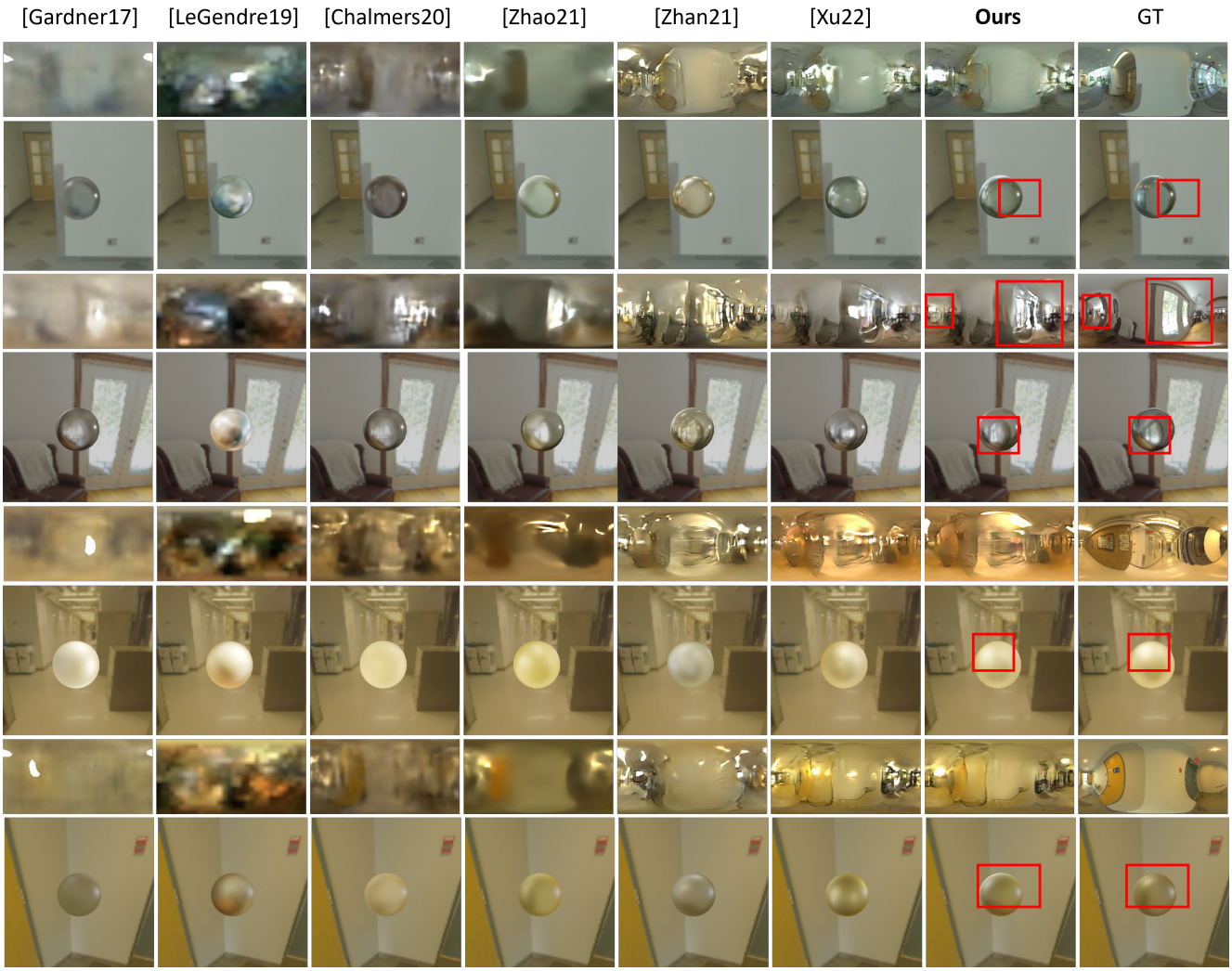

Figure 5. Visual comparison of environment map generation and their rendering results alongside state-of-the-art methods. Four examples are presented, with the upper part displaying the generated environment map and the lower part showing the input crop containing an inserted virtual ball (with roughness levels of 0.2 for rows 2 and 4, and 0.5 for rows 6 and 8) rendered using the predicted environment map in each case. Additional environment map synthesis results can be found in the supplementary material.

visual results, showcasing both the generated environment maps and their rendering effects on a virtual object (virtual ball). It is observed that our method outperforms competing approaches in several aspects, including lighting distribution, color tones, and intensity variations. While methods proposed by [Gardner17], [LeGendre19], [Chalmers20], and [Zhao21] have gradually improved the approximation of lighting distributions, their generated environment maps often lack detail and clarity when rendering objects with low-roughness materials.

The introduction of GAN loss, as seen in [Zhan21] and [Xu22], enhances the fidelity of generated environment maps. However, these methods still struggle with accurate lighting distribution due to limitations in feature encoding and lighting decoding capabilities. Artifacts is apparent in specular and texture reflections on composite object surfaces with roughness=0.2 (Figure 5, rows 2 and 4, columns 5 and 6). Moreover, discrepancies are observed in color tones, intensity, and the highlight areas on object surfaces with roughness=0.5 (Figure 5, rows 6 and 8, columns 5 and 6). Our results demonstrate a distinct advantage in lighting predictions, compelling in the creation of the most authentic environment map and realistic object rendering effects. The overall structure of the environment map is improved. The composition of virtual objects within the scene is seamlessly integrated, with preferable detail and accurate highlight distribution.

These findings are consistent with the quantitative out-

Method	RMSE↓	si-RMSE↓	Angular Error↓
[Gardner17]	0.528	0.184	37.5
[LeGendre19]	0.422	0.180	31.2
[Chalmers20]	0.356	0.167	30.23
[Zhao21]	0.303	0.159	28.2
[Zhan21]	0.256	0.149	27.1
[Xu22]	0.203	0.141	25.3
Ours	**0.181**	**0.135**	**23.2**
Ours vs. [Xu22]	**10.84%**	**4.26%**	**8.30%**

Table 2. Quantitative comparisons of the quality of estimated environment maps, as evaluated through RMSE, si-RMSE, and Angular error metrics.

comes presented in Table 2, where we have compiled the average RMSE, si-RMSE, and angular error metrics [8] in comparison to prior works. The results indicate the advancements achieved by our approach in terms of both image quality and lighting direction within the generated environment maps. Our method achieved lower RMSE, si-RMSE, and angular error values associated with the main light source compared with others. It's worth noting that our method, as well as [Xu22], utilizes the same environment map generator, but our unique strength lies in our ability to produce more accurate environmental details and precise lighting directions. These improvements are primarily attributed to the advanced SG predictions generated by SGformer, which is essential in the paradigm of lighting predictions.

4.2. Ablation Study

To understand the individual contributions of each component within SGformer towards SG predictions, we conducted the ablations study that encompasses both encoder variants and decoder structures alongside the consistency loss design. Both quantitative and qualitative evaluations have been performed. The quantitative assessments involve evaluating RMSE, L1 error, and the Ranked Matching Error (referred to as RME) [31]. Additionally, we analyze the L2 shift in intensity and RGB ratio values within the SG parameters. It's worth noting that, in contrast to the approach taken by Xu et al. [31], our RME calculations are specifically concentrated on the first half of the sorted anchor points. This adjustment allows us to place more emphasis on evaluating the primary lighting aspects.

4.2.1 Encoder Variants

To investigate the impact of different encoders on the extraction of lighting features and their subsequent influence on lighting predictions, we conducted an ablation study comparing the performance of DenseNet and DETR when used as the encoder alongside Conformer. For consistency, we employed the same DETR decoder with anchor points set at $N = 128$ for all encoder variants to generate the SG parameters.

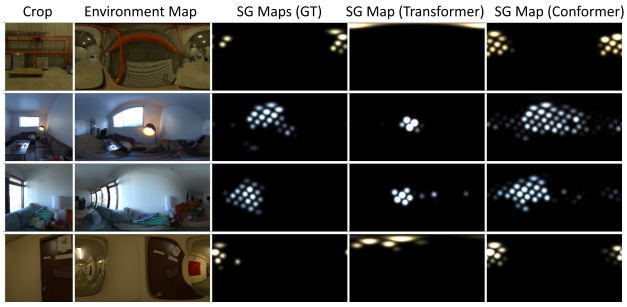

Figure 6. Comparison of SG maps with anchor number $N = 128$ using DLA-SK and Conformer encoders. The same transformer-based decoder is employed for testing.

Encoder	RMSE↓	L1↓	RME↓	Intensity↓	RGB↓
DenseNet	0.063	3.624	0.425	33.513	0.614
DERT (DLA-SK)	0.053	3.385	0.314	31.715	0.532
Conformer	**0.046**	**2.835**	**0.233**	**28.178**	**0.304**

Table 3. Comparison of SG predictions using different encoders.

DenseNet encoder is CNN-based and functioned as the primary feature encoder in previous works by Zhan et al.[33] and[32]. DETR encoders is a mixer of CNN and Transformer, introduced by Xu et al. [31] in their deep learning architecture for lighting predictions. In this configuration, a particularly designed CNN module, DLA-SK, is used for local feature extraction, while a Transformer module is for global feature extraction. However, this concatenated structure overlooks the nuanced interplay between local and global features. In contrast, the Conformer we proposed to use adopts a concurrent structure that enables a more interactive combination of local lighting features from the CNN branch and long-context global lighting features from the Transformer branch. This approach better addresses the conflict and mutual enhancement of these features. Both the output of the CNN branch and the Transformer branch can be used as the lighting features. Here we used CNN output and fed them into the following transformer-based decoder to estimate the SG parameter.

Figure 6 visually illustrates the significant advantages of the Conformer over the DERT encoder in inferring lighting cues and generating highly accurate lighting tones, as exemplified in rows 2 and 3, where bluish lighting predictions can be observed as expected in Conformer results. Furthermore, this distinction becomes evident in challenging scenarios, such as row 4, where only minimal lighting cues, the faint lighting reflections on the door and wall, are present in the input images. Quantitative results further validate the Conformer's effectiveness (Table. 3), as it consistently improves across all metrics. This indicates the essential role of the encoder in deducing lighting cues and emphasizes the Conformer's capability in capturing both local and global

	Baseline (DERT (DLA-SK))				Conformer				Conformer+Multihead				Conformer+Multihead+Consistency			
	RMSE↓	RME↓	I↓	RGB↓	RMSE↓	RME↓	I↓	RGB↓	RMSE↓	RME↓	I↓	RGB↓	RMSE↓	RME↓	I↓	RGB↓
SG8	0.320	0.353	32.581	0.523	0.230	0.263	31.109	0.599	0.201	0.168	28.321	0.568	**0.190**	**0.143**	**27.654**	**0.546**
SG16	0.164	0.362	33.309	0.323	0.159	0.243	31.282	0.243	0.147	0.183	30.103	0.177	**0.1221**	**0.177**	**28.876**	**0.183**
SG32	0.224	0.309	31.252	0.738	0.152	0.258	28.424	0.688	0.102	0.218	27.705	0.627	**0.032**	**0.183**	**25.236**	**0.587**
SG64	0.047	0.292	34.683	0.834	0.039	0.251	32.603	0.702	0.035	0.216	29.238	0.652	**0.022**	**0.197**	**27.607**	**0.579**
SG128	0.053	0.314	34.612	0.532	0.046	0.233	32.211	0.304	0.039	0.182	31.688	0.223	**0.031**	**0.175**	**28.187**	**0.195**

Table 4. Ablation study on SG regression across different SG resolutions.

lighting features while enhancing their synergy for superior lighting predictions.

4.2.2 Regression on Multi-Resolution SG

To progressively investigate the impact of each key component within SGformer, including encoder, decoder structure, and loss function design within SGformer, we conducted ablation studies using the models created under three distinct setups: 1) Train our backbone network with Conformer encoder and a single decoder separately for each SG resolution (referred to as "Separate" in Figure 7, the 4th row); 2) Training the network with Conformer encoder and multiple decoders (as illustrated in Figure 2, referred to as "Multihead" in Figure 7, the 3rd row), and 3) Training with Conformer encoder and multiple decoders, augmented by our proposed consistency loss (referred to as "Ours" in Figure 7, the 2nd row). These setups were evaluated across a spectrum of SG resolutions, ranging from anchor points N set at 8 to 128.

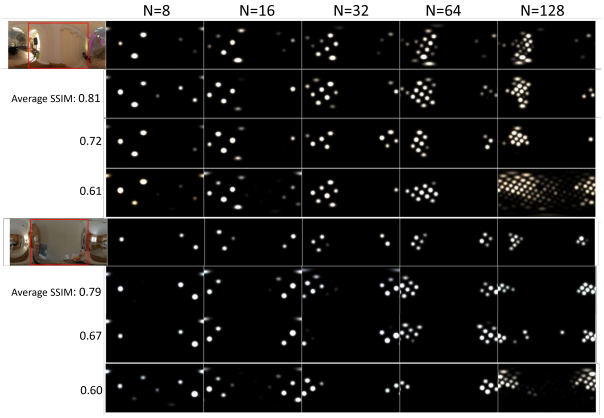

Figure 7. An ablation study on SG regression was conducted across various resolutions. Two exemplars are provided, where in each case, the top row exhibits the ground truth SGs, and the second row presents our results. The third row showcases the results of the 'Multihead' model. The fourth row displays the results of the 'Separate' model. For additional results, please refer to the supplementary material.

As illustrated in Figure 7, the results obtained from our model that encompasses all setup variants exhibit the closest alignment of anchor distributions with the ground truth across all SG resolutions (as seen in both "Ours" rows). In contrast, the separately trained model tends to exhibit random shifting or fading of anchors across different SG resolutions due to the isolation in their learning. Additionally, it struggles to precisely predict lighting shapes and locate lighting sources in high-resolution SG prediction (observed in the "Separate" row, "N=128" columns). The incorporation of a multi-head decoder, which utilizes a common encoder and facilitates joint learning for diverse SG regression tasks, enhances the consistency of anchor distributions across various SG resolutions. The introduction of a consistency loss further regulates the variations in anchor distributions spanning the spectrum of SG resolutions, resulting in improved lighting shape (as seen in the "Ours" row, "N=128" column) and spatial distributions. We assessed the mean similarity between the estimated SG maps and their corresponding ground truth by employing the Structural Similarity Index (SSIM) metric. We can see that the estimated SG maps exhibit the highest similarity to the ground truth compared to other methods.

The quantitative results are provided in Table 4, encompassing various metrics including RMSE, RME, and L2 shift of Intensity and RGB ratio values. The baseline network, introduced by Xu et al. [31] with a DETR encoder, is trained across different SG resolutions and serves as the reference point. The outcomes illustrate that the introduction of the Conformer encoder results in notable improvements over the baseline, particularly on RMSE, RME, and global intensity metrics. The integration of multi-head decoding and consistency loss further enhances the model's capabilities, resulting in consistent improvements across all metrics. This enhancement is particularly evident in the reduction of RME, indicating optimized main lighting source distributions.

4.3. Discussions

Our proposed method aims to estimate environmental lighting from a standard image, avoiding the need for direct panoramic image capture with expensive devices to obtain global illumination. While our supervised deep-learning method requires ground truth data during the training stage, once the model is trained, it can be applied to any arbitrary standard image captured by lightweight cameras, such as mobile phone cameras. The training ground truth in Laval Dataset has been publicly available, and we make full use of

it in a one-off manner, consistent with the approach taken in most developments of machine/deep learning applications.

5. Conclusions and future work

In summary, this paper introduced an advanced deep transformer-based approach to enhance indoor lighting estimation performance from single standard images. Our novel network architecture combines a Conformer model for global and local lighting feature extraction with a multi-resolution transformer-based decoder for simultaneous SG parameter predictions across various resolutions. We are the first to explore the interplay of spatial distributions across multiple SG resolutions and utilize this to enhance the spatial distribution of lighting sources. To improve lighting structure modeling, we introduced an SG consistency loss designed to ensure consistency in spatial distributions across different SG resolutions. Our comprehensive experiments have demonstrated significant improvements in lighting estimation, enhancing predictions of lighting source shapes, color tones, and lighting directions. As a powerful tool, SGformer effectively enhances the realism of environment map estimation, providing precise guidance for highly realistic environment map synthesis and realizing seamless object rendering. Looking ahead, our future research will focus on advancing methods to gain a deeper understanding of the visual context and semantics within scenes, with the goal of achieving even more accurate and context-aware illumination predictions.

References

[1] J. Bai, J. Guo, C. Wang, Z. Chen, Z. He, S. Yang, P. Yu, Y. Zhang, and Y. Guo. Deep graph learning for spatially-varying indoor lighting prediction. *Science China Information Sciences*, 66(3):132106, 2023. 1, 2, 3

[2] R. Basri and D. W. Jacobs. Lambertian reflectance and linear subspaces. *IEEE transactions on pattern analysis and machine intelligence*, 25(2):218–233, 2003. 2

[3] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 5

[4] A. Chalmers, J. Zhao, D. Medeiros, and T. Rhee. Reconstructing reflection maps using a stacked-CNN for mixed reality rendering. *IEEE Transactions on Visualization and Computer Graphics*, 2020. 3, 7

[5] Q. Chen, Q. Wu, J. Wang, Q. Hu, T. Hu, E. Ding, J. Cheng, and J. Wang. Mixformer: Mixing features across windows and dimensions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5249–5259, 2022. 5

[6] D. Cheng, J. Shi, Y. Chen, X. Deng, and X. Zhang. Learning scene illumination by pairwise photos from rear and front mobile cameras. In *Computer Graphics Forum*, volume 37, pages 213–221. Wiley Online Library, 2018. 2

[7] F. Einabadi, J.-Y. Guillemaut, and A. Hilton. Deep neural models for illumination estimation and relighting: A survey. In *Computer Graphics Forum*, volume 40, pages 315–331. Wiley Online Library, 2021. 1

[8] M.-A. Gardner, Y. Hold-Geoffroy, K. Sunkavalli, C. Gagné, and J.-F. Lalonde. Deep parametric indoor lighting estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7175–7183, 2019. 1, 2, 3, 4, 9

[9] M.-A. Gardner, K. Sunkavalli, E. Yumer, X. Shen, E. Gambaretto, C. Gagné, and J.-F. Lalonde. Learning to predict indoor illumination from a single image. *ACM Transactions on Graphics (TOG)*, 9(4), 2017. 3, 6, 7

[10] M. Garon, K. Sunkavalli, S. Hadap, N. Carr, and J.-F. Lalonde. Fast spatially-varying indoor lighting estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6908–6917, 2019. 2

[11] R. Green. Spherical harmonic lighting: The gritty details. In *Archives of the game developers conference*, volume 56, page 4, 2003. 2

[12] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, et al. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*, 2020. 5

[13] C. LeGendre, W.-C. Ma, G. Fyffe, J. Flynn, L. Charbonnel, J. Busch, and P. Debevec. DeepLight: Learning illumination for unconstrained mobile mixed reality. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5918–5928, 2019. 2, 3, 7

[14] M. Li, J. Guo, X. Cui, R. Pan, Y. Guo, C. Wang, P. Yu, and F. Pan. Deep spherical gaussian illumination estimation for indoor scene. In *Proceedings of the ACM Multimedia Asia*, pages 1–6. 2019. 1, 2

[15] N. Li, L. Ma, G. Yu, B. Xue, M. Zhang, and Y. Jin. Survey on evolutionary deep learning: Principles, algorithms, applications, and open issues. *ACM Computing Surveys*, 56(2):1–34, 2023. 1

[16] Z. Li, M. Shafiei, R. Ramamoorthi, K. Sunkavalli, and M. Chandraker. Inverse rendering for complex indoor scenes: Shape, spatially-varying lighting and svbrdf from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2475–2484, 2020. 2

[17] Y. Liu, Y. Sun, B. Xue, M. Zhang, G. G. Yen, and K. C. Tan. A survey on evolutionary neural architecture search. *IEEE transactions on neural networks and learning systems*, 2021. 1

[18] Z. Peng, W. Huang, S. Gu, L. Xie, Y. Wang, J. Jiao, and Q. Ye. Conformer: Local features coupling global representations for visual recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 367–376, 2021. 2, 3, 4, 5

[19] R. Ramamoorthi and P. Hanrahan. An efficient representation for irradiance environment maps. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 497–500, 2001. 2

[20] R. Ramamoorthi and P. Hanrahan. Frequency space environment map rendering. In *Proceedings of the 29th annual conference on Computer graphics and interactive techniques*, pages 517–526, 2002. 2

[21] Z. Shi, X. Lin, and Y. Song. An attention-embedded gan for svbrdf recovery from a single image. *Computational Visual Media*, 9(3):551–561, 2023. 1

[22] G. Somanath and D. Kurz. Hdr environment map estimation for real-time augmented reality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11298–11306, 2021. 3

[23] S. Song and T. Funkhouser. Neural Illumination: Lighting prediction for indoor environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6918–6926, 2019. 3

[24] P. P. Srinivasan, B. Mildenhall, M. Tancik, J. T. Barron, R. Tucker, and N. Snavely. Lighthouse: Predicting lighting volumes for spatially-coherent illumination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8080–8089, 2020. 3

[25] C. Suppan, A. Chalmers, J. Zhao, and T. Rhee. Neural screen space rendering of direct illumination. 2021. 1

[26] Y.-T. Tsai and Z.-C. Shih. All-frequency precomputed radiance transfer using spherical radial basis functions and clustered tensor approximation. *ACM Transactions on graphics (TOG)*, 25(3):967–976, 2006. 2

[27] H. Vogel. A better way to construct the sunflower head. *Mathematical biosciences*, 44(3-4):179–189, 1979. 1, 3, 6

[28] G. Wang, Y. Yang, C. C. Loy, and Z. Liu. Stylelight: Hdr panorama generation for lighting estimation and editing. In *European Conference on Computer Vision*, pages 477–492. Springer, 2022. 1

[29] J. Wang, P. Ren, M. Gong, J. Snyder, and B. Guo. All-frequency rendering of dynamic, spatially-varying reflectance. In *ACM SIGGRAPH Asia 2009 papers*, pages 1–10. 2009. 2

[30] J. Weir, J. Zhao, A. Chalmers, and T. Rhee. Deep portrait delighting. In *European Conference on Computer Vision*, pages 423–439. Springer, 2022. 1

[31] J.-P. Xu, C. Zuo, F.-L. Zhang, and M. Wang. Rendering-aware hdr environment map prediction from a single image. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2857–2865, 2022. 2, 3, 4, 5, 6, 7, 9, 10

[32] F. Zhan, Y. Yu, C. Zhang, R. Wu, W. Hu, S. Lu, F. Ma, X. Xie, and L. Shao. Gmlight: Lighting estimation via geometric distribution approximation. *IEEE Transactions on Image Processing*, 31:2268–2278, 2022. 1, 2, 3, 9

[33] F. Zhan, C. Zhang, Y. Yu, Y. Chang, S. Lu, F. Ma, and X. Xie. Emlight: Lighting estimation via spherical distribution approximation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3287–3295, 2021. 1, 2, 3, 4, 7, 9

[34] F. Zhang, J. Zhao, Y. Zhang, and S. Zollmann. A survey on 360 images and videos in mixed reality: Algorithms and applications. *Journal of Computer Science and Technology*, 38(3):473–491, 2023. 1, 3

[35] J. Zhao, A. Chalmers, and T. Rhee. Adaptive light estimation using dynamic filtering for diverse lighting conditions. *IEEE Transactions on Visualization and Computer Graphics*, 27(11):4097–4106, 2021. 3, 7

[36] J. Zhao, C. J. Parry, R. dos Anjos, C. Anslow, and T. Rhee. De-lighting human images using region-specific data augmentation. In *2020 35th International Conference on Image and Vision Computing New Zealand (IVCNZ)*. IEEE, 2023. 1

[37] D. Zhou, Y. Shi, B. Kang, W. Yu, Z. Jiang, Y. Li, X. Jin, Q. Hou, and J. Feng. Refiner: Refining self-attention for vision transformers. *arXiv preprint arXiv:2106.03714*, 2021. 5

[38] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 5