Equivariant Indoor Illumination Map Estimation from a Single Image

Yusen Ai[1], Xiaoxue Chen[2], Xin Wu[1], and Hao Zhao[2](✉)

[1] Key Laboratory of Machine Perception(MOE), School of AI,
Peking University, Beijing, China
ysai@pku.edu.cn

[2] Institute for AI Industry Research,
Tsinghua University, Beijing, China
zhaohao@air.tsinghua.edu.cn

Abstract. Thanks to the recent development of inverse rendering, photorealistic re-synthesis of indoor scenes have brought augmented reality closer to reality. All-angle environment illumination map estimation of arbitrary locations, as a fundamental task in this domain, is still challenging to deploy due to the requirement of expensive depth input. As such, we revisit the appealing setting of illumination estimation from a single image, using a cascaded formulation. The first stage predicts faithful depth maps from a single RGB image using a distortion-aware architecture. The second stage applies point cloud convolution operators that are equivariant to SO(3) transformations. These two technical ingredients collaborate closely with each other, because equivariant convolution would be meaningless without distortion-aware depth estimation. Using the public Matterport3D dataset, we demonstrate the effectiveness of our illumination estimation method both quantitatively and qualitatively. Code is available at https://github.com/Aitensa/Img2Illum.

1 Introduction

Understanding the physical properties that can be used to generate an image, which is often referred to as inverse rendering [13,32], is not only a fundamental computer vision problem but also an enabling technique of emerging A/VR applications. If this goal is finally achieved with high accuracy, we can insert any objects into captured photos without human-perceptible artifacts. But this is very challenging as the inverse rendering problem involves many sub-tasks that are difficult on their own. Among them, estimating the lighting condition is an indispensable module.

While there exist other lighting parameterizations, we choose panoramic illumination map [21,30] among alternatives due to its simplicity and expressiveness. Specifically, the task is to infer a panoramic illumination map for a certain pixel in a perspective RGB image as shown in Fig. 1. We first recap two highly related prior works as follows: (1) Neural Illumination [21] uses exactly the same setting as ours, but it's quite complicated, consisting of a geometry estimation module, a differentiable warping module and an HDR reconstruction module. (2) PointAR

L. Fang et al. (Eds.): CICAI 2023, LNAI 14473, pp. 143–155, 2024.
https://doi.org/10.1007/978-981-99-8850-1_12

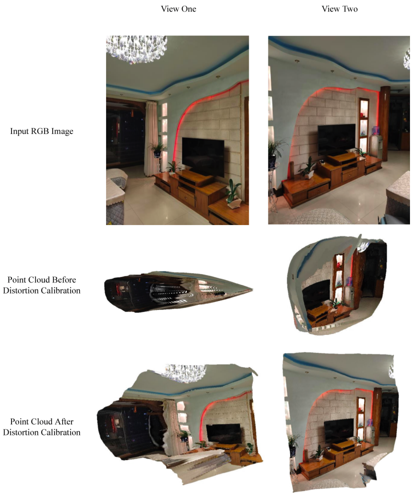
View One
View Two
Input RGB Image
Point Cloud Before Distortion Calibration
Point Cloud After Distortion Calibration

Fig. 1. The task is to infer a panoramic illumination map from a single perspective RGB image and the first step of our method is to infer a point cloud from RGB images. We want the algorithm to be equivariant to SO(3) transformations. For example, for the same point (e.g., a point on the floor) in two viewpoints shown in the left panel, we want the illumination map to be equivariant to viewpoint changes. As such, depth map distortion calibration becomes important because imposing equivariant convolution on point clouds (middle panel) is meaningless.

[30] assumes that an RGB-D image, which can be converted into a point cloud, is available as input. Then a point convolutional network directly extracts spherical harmonics that approximates the illumination map, from input point clouds.

We note that PointAR is not applicable to most cellphones without depth cameras, so we revisit the more generic single-image setting of Neural Illumination. Unfortunately, the network of Neural Illumination involves several dense prediction modules that are also inefficient for deployment. As such, we propose a new cascaded formulation that firstly predicts depth from a single RGB image and then applies a PointAR-like architecture to regress spherical harmonics from the predicted point cloud.

This new formulation is conceptually simple but successfully pushing it to the state-of-the-art performance level needs specific designs. The **first** design is introducing equivariant point convolution of [3]. The illumination map of the same point (e.g., a point on the floor in the scene shown in Fig. 1) should be equivariant to SO(3) transformation like the viewpoint changes in two rows of Fig. 1. To clarify, since the predicted point cloud needs to be re-centered to the point of interest as PointAR does, we only need to concern about SO(3) equivariance instead of SE(3) equivariance. The **second** design is introducing the distortion calibration technique proposed in [26]. It is widely known that single-view depth estimation is troubled by incorrect depth scale and bias. As shown in the middle panel of Fig. 1, without (scale/bias) distortion calibration the point clouds generated from two viewpoints are completely different. In this case, SO(3) equivariance becomes meaningless so using calibrated piont clouds (Fig. 1 right panel) is the right choice. To summarize, in this study:

- We propose a new framework that estimates panoramic illumination maps from a single RGB image, which cascades a depth estimation network and a network that estimates spherical harmonics from predicted point clouds.
- We introduce SO(3) equivariant point convolution and depth calibration into the framework. Although they are existing techniques, we are the first to show their collaboration and significant impact on illumination estimation.
- We benchmark on the large-scale public dataset Matterport3D, achieving state-of-the-art results. Through ablations, we demonstrate the impact of newly introduced modules. Codes are publicly available.

2 Related Work

Lighting Estimation has been a long-standing challenge in computer vision, and is critical for real-world AR applications like realistic relighting and object replacement. A direct way of capturing the illumination of an environment is to use a physical probe [4]. This process, though accurate, can be expensive and unsuitable for lighting estimation of different locations. Another line of works estimates illumination as a sub-task of inverse rendering [16,32], whose goal is to jointly estimate intrinsic properties of the scene, e.g. geometry, reflectance, and lighting from the input image. Classical methods formulate inverse rendering as an energy optimization problem with heuristic priors [13]. With the rapid development of deep neural networks, we could also learn generalizable models directly from single images in a data-driven fashion. Existing works estimate environment lighting in simplified problem settings, such as outdoor scenes [12,29] and objects [1,17]. In this work, we focus on more complex indoor environments, where the spatially-varying effects are not negligible.

For indoor scenes, Karsch et al. [13] recover parametric 3D lighting from a single image assuming known geometry. Gardner et al. [10] propose to learn the location and intensity of light sources in an end-to-end manner. However, their models don't handle spatially-varying lighting, i.e., different locations within the scene can have different lighting. [9] improves it by representing lighting as a set of discrete 3D lights with geometric and photometric parameters. Song et al. [21] decompose illumination prediction into several simpler differentiable sub-tasks, but suffer from spatial instability. Lighthouse [22] further proposes a multi-scale volumetric lighting representation. Wang et al. [24] leverage a holistic inverse rendering framework to guarantee physically correct HDR lighting prediction. Li et al. [16] use 360° panoramic images to obtain high-definition spatially-varying lighting. Zhan et al. [28] solve illumination estimation via spherical distribution approximation.

Meanwhile, to enable real-time AR applications on modern mobile devices, [11,30] use the spherical harmonics (SH) lighting model for fast estimation. In this work, we predict both the SH coefficients and the irradiance map.

Equivariance is a promising property of feature representation. Compared to invariance, equivariance maintains the influence of different transformations, ensuring stable and reasonable performance. In the field of computer vision, since neural networks are often sensitive to rotation transformations, a large body of work has been proposed for rotation equivariance. Existing techniques could be roughly divided into spectral and non-spectral methods.

Spectral methods usually design intrinsically rotation-equivariant basis functions [23], and develop special network architectures with these basis functions [5,14,20]. Tensor-field based networks [7,8,23] implement convolutional kernels in the spherical harmonics domain to make the features equivariant to rotations and translations. However, spherical harmonics leads to high space and time complexity. Deng et al. [5] propose a general framework built on vector activations to enable SO(3)-equivariance. Due to the nature of linear combination, [5] fails to conduct flexible vector transformations. Luo et al. [19] introduce orientations for each point to achieve equivariance based on graph neural network schemes in a fully end-to-end manner. As for non-spectral methods [3,6,15,27], they discretize the rotation group and construct a set of kernels for the group equivariant computation. EPN [3] introduces a tractable approximation to $SE(3)$ group equivariant convolution on point clouds. [27] further transfers this framework to object-level equivariance for 3D detection. Du et al. [6] proposes to construct $SE(3)$ equivariant graph neural networks with complete local frames, approximating the geometric quantities efficiently.

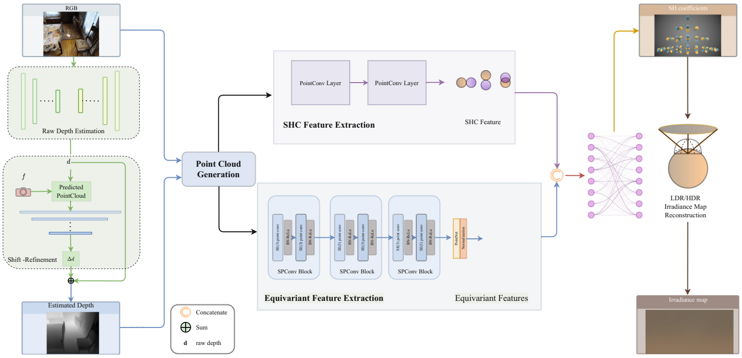

Fig. 2. The overview of our framework. We propose a cascaded illumination estimation formulation of two stages, composed of a point cloud generation module and an equivariant illumination estimation module.

3 Method

3.1 Overall Architecture

Our goal is to estimate illumination from a single perspective RGB image. This is an extremely ill-posed problem since different lighting might lead to the same appearance. Therefore, we choose to leverage geometric priors by predicting depth from the image first and then generating the corresponding point cloud at the rendered position. Following [30], we formulate the illumination estimation as a spherical harmonic coefficients regression problem, and regress spherical harmonics from the predicted point cloud.

However, such a framework still has potential issues. As shown in Fig. 1, the images of the same scenario from various perspectives may lead to different distortion of predicted depth and mislead the limited illumination estimation, especially in indoor scenes, let alone all-angle environment illumination map estimation at arbitrary locations. Moreover, there is a basic fact for image-based estimation: with light sources fixed, the illumination is consistent when the rendered point rotates or the viewpoint changes, which is namely the **equivariance of illumination**. Recently, the depth estimation from a single RGB image has made great progress [26], with a distortion-ware depth estimation paradigm and offers a precise depth estimation. In this case, the equivariance of SO(3) transformations do make sense. The precise distortion-aware depth estimation can not only guarantee the reliability of the combination of RGB and its predicted depth, but also unleash the potential of equivariance in lighting estimation.

To this end, we propose a cascaded network in Fig. 2. It contains two stages, generating point clouds with distortion-aware depth estimation and equivariant illumination estimation from different viewpoints. The model is composed of a point cloud generation module D, an equivariant feature extraction module E, a PointConv model P and a lighting estimation module R. r denotes the rendered point. The formulation of our **equivariant indoor illumination map estimation** from a single image is defined as a mapping:

$$\mathcal{F} : L(R(E(D(S, r)), P(D(S, r)))) \rightarrow I \tag{1}$$

where S is the source image, I is the target illumination map, and L is the transformation from SH coefficients to illumination map. Specifically,

$$L(I_{shc}) = I, \quad R(e_r, S_f) = I_{shc}, \tag{2}$$

$$E(P_r) = e_r, \quad P(P_r) = S_f, \tag{3}$$

$$D(S, r) = Sample(S, g(S) \circ f(S), r) = P_r. \tag{4}$$

P_r is the generated point cloud, I_{shc} is the predicted spherical harmonics, and e_r and S_f are both point features. $g(S)$ and $f(S)$ will be elaborated later.

Point Cloud Generation. In the first stage, we generate the point cloud based on the rendered point leveraging the distortion-aware depth estimation [26], which is formulated as Eq. 4. Firstly, $g(S)$ maps from an RGB image to a raw depth estimation by a Depth Prediction Model (DPM), then $f(S)$ takes the raw depth as input and estimates a refined shift for the raw depth through a Point cloud Module (PCM), $g(S) \circ f(S)$ make refinement shift on raw depth and output a corrected depth image. $Sample(S, g(S) \circ f(S), r)$ make a point cloud P_r centered at the render position r for subsequent process on the input estimated depth and corresponding image, which will be described in Sect. 3.2.

Equivariant Illumination Estimation. In the second stage, we formulate the equivariant illumination map estimation as a composite point cloud-based learning problem L that takes a predicted point cloud P_r as input and outputs a scene-consistent equivariant irradiance map. In this stage, we first simultaneously compute the raw estimate sphere harmonic coefficient (SHC) feature S_f through PointConv module P and extract the structure-aware equivariant feature e_r through the equivariant feature extraction module E, then concatenate both SHC and equivariant feature to acquire a 2nd order SH coefficients I_{shc} through Eq. 2. Finally, $L(I_{shc})$ inputs the SHC and outputs the target illumination map. More details about the modules will be described in Sect. 3.3.

3.2 Point Cloud Generation

In this section, we will describe the details of generating a point cloud recentered on the rendered point from a single RGB image.

Distortion-Aware Depth Estimation. Given an RGB image, there are two modules in order for processing. The DPM module based on PVCNN [18], generates a depth image with unknown scale and shift, and the PCM module takes the distorted point cloud as input and predicts shift refinement to the depth image.

Recentered Point Cloud Generation. Given the distortion-aware depth image Z and camera intrinsic matrix, we can easily transform the depth image into a point cloud P centered on the camera origin through:

$$x = \frac{(u - c_x)z}{f_x}, y = \frac{(u - c_y)z}{f_y}, \tag{5}$$

where u and v are the photo pixel coordinates, z is the corresponding depth value of pixel (u,v), f_x and f_y are the vertical and horizontal camera focal length, c_x and c_y are the photo-optical center, and (x,y,z) is the corresponding point of pixel (u,v).

With the rendered point r, we apply a linear translation T to P and transform the point cloud center to the observation point P_r:

$$\mathrm{P_r} = \mathrm{T(P)} = \mathrm{P} - \mathrm{r}. \tag{6}$$

Unit-Sphere Downsampling. Accounting for the efficiency of reserving the spatial structure, we exert a new technique of sphere sampling – Unit-sphere Downsampling [31]. At downsampling, for each input point cloud P_{input}, we will project the point cloud on a unit surface, and accumulate the area of the uniform surface anchor's coverage. Theoretically, let P_{data} and P_{anchor} correspondingly be the input point cloud's and the uniform surface anchors' distribution, the completeness of observation was measured by the joint entropy $H(P_{data}, P_{anchor})$,

$$H(P_{data}, P_{anchor}) = -\sum_{i \in S} \sum_{j=1}^{i} P(p'_{ij}, p_i) \log_2 [P(p'_{ij}, p_i)] \tag{7}$$

where $P(p'_{ij}, p_i)$ s the joint probability of projecting points into a unit sphere with i anchor point, and S is the set of possible anchors, $\{2^k | 1 \leq k \leq 12\}$ can be accepted. In this work, we set 1280 points as the target for downsampling.

3.3 Equivariant Illumination Estimation

Equivariant Feature Extraction. For this module, we were inspired by recent works on Equivariance [3,15,19,20,27], and based on [3], we put forward the current network. For every RGB and corresponding point cloud input, we use three basic SPConv blocks whose layer channel settings are [32,32],[64,128],[128,128] and a Pointnet with channel [128,64] to output a 64-dimension Equivariant feature, which uses 60 SO(3) bases in this work. As to the other configuration of this module, the $initial\_radius\_ratio = 0.2, sampleing\_ratio = 0.4$.

SHC Feature Extraction. The module is a PointConv-based [25] backbone derived from [30]. It takes downsampled RGB and the corresponding point cloud as input, and outputs a 256-dimensional tensor, which we regard as an SHC feature because it numerically describes SHC. The two PointConvs' channel numbers are set as [64,128] and [128,256] respectively.

Concatenating the above two features, we then use a fully connected layer as the prediction head to get the predicted spherical harmonic coefficients I_{shc} and transform them into the illumination map.

3.4 Loss

To provide an auxiliary constraint on the outputs, we supervise both SH coefficients and the irradiance map, and the total loss function L is defined as:

$$\mathcal{L} = \mathcal{L}_{sh} + \mathcal{L}_{ir} \tag{8}$$

$\mathcal{L}_{sh}$ is the L2 loss of SH coefficients, which is defined in Eq. 9:

$$\mathcal{L}_{sh} = \frac{1}{9} \sum_{c=l}^{3} \sum_{l=0}^{2} \sum_{m=-l}^{l} (i_{l,c}^{m*} - i_{l,c}^{m}) \tag{9}$$

where c is the color channel (RGB), l and m are the degree and order of SH coefficients. $\mathcal{L}_{ir}$ is L2 loss for irradiance map, and defined in Eq. 10.

$$\mathcal{L}_{ir} = \frac{1}{N_{env}} \sum_{p=0}^{N_{env}} (i_p^* - i_p)^2 \tag{10}$$

where N_{env} is the number of pixels of the target image, and i is the value of the corresponding pixel.

4 Experiment

4.1 Datasets and Preprocessing

We carry out our experiments on the Matterport3D dataset [2], with illumination ground truth generated by Neural Illumination. Matterport3D is a large-scale dataset that contains RGB-D images and panoramic views for indoor scenes. Each RGB-D scene contains undistorted color and depth images (of size 1028 × 1024) of 18 viewpoints. The Neural Illumination dataset [21] was derived from Matterport3D and packed with additional information that associates images and the relationship between images at observation and rendering locations. We derive depth images from Matterport3D as Fig. 3 shows, whose depth images come from the DepthEst module and Unit-Sphere Downsampling module.

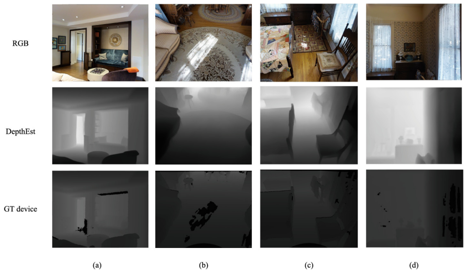

Fig. 3. Depth estimation results. The RGB images are from Matterpord3D, each column in order is RGB, predicted depth image from our DepthEst module and depth image captured from the device.

4.2 Comparison in Quality and Quantity

Quantitative Results. As shown in Table 1, We carried out our experiments based on the Matterport Dataset, we compare our solution with other illumination estimation solutions using L2 loss of SH coefficients, Irradiance map on the test set. For baseline solutions targeted at RGB-D, we keep the same setting in experiments. And we achieved superior performance compared with previous arts. We credit this to our Equivariant feature extraction module which has learned how to adjust to the transforms of the scenes and combined multiple rotation cases' tradeoffs during training, and finally calibrated and enhanced the sources for SHC regression. And the robustness will be seen in qualitative results.

Table 1. Comparison to state-of-the-art networks. Our approach achieved the lowest loss for both spherical harmonics coefficients l2 and irradiance map l2.

Method	SH coefficients l2 Loss	Irradiance map l2 Loss	United l2 Loss
Song et al. [21]	N/A	0.619	N/A
Garon et al. [11]	1.10 (±0.1)	0.63 (±0.03)	N/A
PointAR [30]	0.31 (±0.03)	0.434 (± 0.02)	0.32 (±0.02)
Ours	**0.11** (±0.05)	**0.31** (±0.04)	**0.20** (±0.04)

Also, we compare the complexity of the networks of illumination estimation stage with PointAR [30] in Table 3. Account for the difference in target task between PointAR and Ours, the increments of complexity are acceptable, so that we believe that our model can still be applied on mobile platforms.

Table 2. Comparison of loss coefficients.

Model	α	β	Valid Loss
PointAR [30]	1	10	9.04
	5	10	29.60
	10	1	5.77
Ours	1	10	**0.20**
	5	10	**0.17**
	10	1	**1.36**

Table 3. Comparison of model complexity.

Method	Parameters(M)
PointAR [30]	1.42
Ours	2.42

Qualitative Results. Here we demonstrate the quality of our method. Both PointAR and our method take RGB-D input, use unite loss as supervision for training and train 10 epochs. At test, we input the RGB-PD pairs and obtain illumination estimation results as shown in Fig. 4. It is evident that our method exhibits more detailed results. There are two main reasons for this improvement. Firstly, This optimization by DPM effectively reduces the impact of noise and enhances the level of detail in our results. Secondly, our method leverages equivariance, which enables it to effectively overcome slight perturbations. For further generalization, incorporating additional equivariance bases becomes crucial.

4.3 Ablation Study

In this section, we describe the ablation experiments performed on the model settings, sampling methods and loss function coefficients. Among them, the model setup needs to be especially noted that PointAR [30] is the degenerate model and also the method of this paper, so the ablation experiments of the equivariant module proposed in this paper will be obtained by comparing the method of this paper with PointAR. As displayed in Table 1, compared with PointAR, our

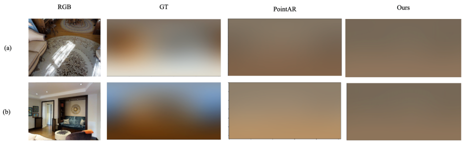

Fig. 4. Comparison of illumination estimation. Each row is a partial image of a scene and the column in order is the RGB image, GT illumination map, results of PointAR and our method.

method attained a comparative result than PointAR, which shows the efficiency of the equivariance in quantity.

Table 4. Ablation study of the sampling method in different measurements.

Method	SHC coefficients Loss	Irradiance Map Loss	Unite Loss
PointAR(w/o)	**0.417**	1.698	0.45
PointAR	0.914	**0.009**	**0.32**
Ours(w/o)	0.210	1.02	0.396
Ours	**0.11**	**0.012**	**0.204**

In Table 4, we conducted the comparative test on the use of sphere sample, which is regarded as more geometry information. Obviously, both PointAR and our method with the sphere sampling achieve better results, for instance, the SH coefficients loss decreased from 0.21 to 0.11 with sphere sampling.

As shown in Table 2, in our proposed method, the final result varies from the coefficients of the loss function. We separately conduct experiments on different parameter settings, and measures on united valid loss. It needs to be mentioned that the results are different from the Table 1. Then we can arrive at that $\alpha = 5, \beta = 10$ is a relatively optimal setting.

4.4 Application on AR

End-to-End Generation. As shown in Fig. 5, we build a pipeline for capturing, uploading and attaining the irradiance map in Fig. 5. It demonstrates the feasibility of the proposed framework.

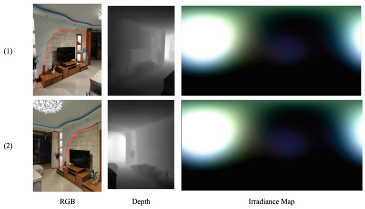

Fig. 5. Demonstration from a snapshot RGB Image to its corresponding Irradiance map.

Rendering. As shown in Fig. 6, the rabbit is rendered at the user's preferred location. Although the rendering result seems reasonable in human's observation, it still lacks some details.

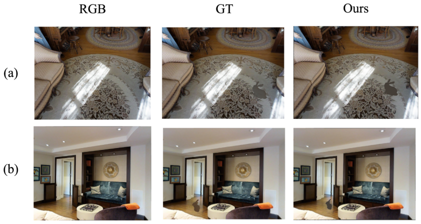

Fig. 6. AR applications. Our method is tested in indoor scenes.

5 Conclusion

In this study, we propose a novel cascaded approach for estimating illumination from a single RGB image. The first stage utilizes a distortion-aware architecture to accurately predict depth maps. In the second stage, a PointAR-like architecture is employed to regress spherical harmonics from the predicted point clouds. To ensure equivariance of the illumination map under SO(3) transformation, we introduce equivariant point convolution for estimating spherical harmonics coefficients based on distortion-aware single-frame depth estimation. These two techniques work closely together as distortion calibration is crucial to generate consistent point clouds from single images captured from different viewpoints. Experimental results demonstrate that our method achieves state-of-the-art (SOTA) performance on the large-scale Matterport3D dataset, highlighting the effectiveness of the cascaded formulation and equivariance in illumination modeling.

References

1. Boss, M., Jampani, V., Kim, K., Lensch, H., Kautz, J.: Two-shot spatially-varying BRDF and shape estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3982–3991 (2020)
2. Chang, A., et al.: Matterport3D: learning from RGB-D data in indoor environments. arXiv preprint arXiv:1709.06158 (2017)
3. Chen, H., Liu, S., Chen, W., Li, H., Hill, R.: Equivariant point network for 3D point cloud analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14514–14523 (2021)
4. Debevec, P.: Rendering synthetic objects into real scenes: bridging traditional and image-based graphics with global illumination and high dynamic range photography. In: ACM SIGGRAPH 2008 Classes, pp. 1–10 (2008)
5. Deng, C., Litany, O., Duan, Y., Poulenard, A., Tagliasacchi, A., Guibas, L.J.: Vector neurons: a general framework for so (3)-equivariant networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 12200–12209 (2021)
6. Du, W., et al.: Se (3) equivariant graph neural networks with complete local frames. In: International Conference on Machine Learning, pp. 5583–5608. PMLR (2022)
7. Esteves, C., Allen-Blanchette, C., Makadia, A., Daniilidis, K.: Learning SO(3) equivariant representations with spherical CNNs. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11217, pp. 54–70. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01261-8_4
8. Fuchs, F., Worrall, D., Fischer, V., Welling, M.: Se (3)-transformers: 3D roto-translation equivariant attention networks. In: Advances in Neural Information Processing Systems, vol. 33, pp. 1970–1981 (2020)
9. Gardner, M.A., Hold-Geoffroy, Y., Sunkavalli, K., Gagné, C., Lalonde, J.F.: Deep parametric indoor lighting estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7175–7183 (2019)
10. Gardner, M.A., et al.: Learning to predict indoor illumination from a single image. arXiv preprint arXiv:1704.00090 (2017)
11. Garon, M., Sunkavalli, K., Hadap, S., Carr, N., Lalonde, J.F.: Fast spatially-varying indoor lighting estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6908–6917 (2019)
12. Hold-Geoffroy, Y., Athawale, A., Lalonde, J.F.: Deep sky modeling for single image outdoor lighting estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6927–6935 (2019)
13. Karsch, K., Hedau, V., Forsyth, D., Hoiem, D.: Rendering synthetic objects into legacy photographs. ACM Trans. Graph. (TOG) **30**(6), 1–12 (2011)
14. Keriven, N., Peyré, G.: Universal invariant and equivariant graph neural networks. In: Advances in Neural Information Processing Systems, vol. 32 (2019)
15. Li, J., Bi, Y., Lee, G.H.: Discrete rotation equivariance for point cloud recognition. In: 2019 International Conference on Robotics and Automation (ICRA), pp. 7269–7275. IEEE (2019)
16. Li, J., Li, H., Matsushita, Y.: Lighting, reflectance and geometry estimation from 360 panoramic stereo. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10586–10595. IEEE (2021)
17. Li, Z., Xu, Z., Ramamoorthi, R., Sunkavalli, K., Chandraker, M.: Learning to reconstruct shape and spatially-varying reflectance from a single image. ACM Trans. Graph. (TOG) **37**(6), 1–11 (2018)

18. Liu, Z., Tang, H., Lin, Y., Han, S.: Point-voxel CNN for efficient 3D deep learning. In: Advances in Neural Information Processing Systems, vol. 32 (2019)
19. Luo, S., et al.: Equivariant point cloud analysis via learning orientations for message passing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18932–18941 (2022)
20. Shen, W., Zhang, B., Huang, S., Wei, Z., Zhang, Q.: 3D-rotation-equivariant quaternion neural networks. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12365, pp. 531–547. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58565-5_32
21. Song, S., Funkhouser, T.: Neural illumination: lighting prediction for indoor environments. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6918–6926 (2019)
22. Srinivasan, P.P., Mildenhall, B., Tancik, M., Barron, J.T., Tucker, R., Snavely, N.: Lighthouse: predicting lighting volumes for spatially-coherent illumination. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8080–8089 (2020)
23. Thomas, N., et al.: Tensor field networks: rotation-and translation-equivariant neural networks for 3D point clouds. arXiv preprint arXiv:1802.08219 (2018)
24. Wang, Z., Philion, J., Fidler, S., Kautz, J.: Learning indoor inverse rendering with 3D spatially-varying lighting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 12538–12547 (2021)
25. Wu, W., Qi, Z., Fuxin, L.: Pointconv: deep convolutional networks on 3D point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9621–9630 (2019)
26. Yin, W., et al.: Learning to recover 3D scene shape from a single image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 204–213 (2021)
27. Yu, H.X., Wu, J., Yi, L.: Rotationally equivariant 3D object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1456–1464 (2022)
28. Zhan, F., et al.: Emlight: lighting estimation via spherical distribution approximation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 3287–3295 (2021)
29. Zhang, J., Sunkavalli, K., Hold-Geoffroy, Y., Hadap, S., Eisenman, J., Lalonde, J.F.: All-weather deep outdoor lighting estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10158–10166 (2019)
30. Zhao, Y., Guo, T.: PointAR: efficient lighting estimation for mobile augmented reality. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12368, pp. 678–693. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58592-1_40
31. Zhao, Y., Guo, T.: Xihe: a 3D vision-based lighting estimation framework for mobile augmented reality. In: The 19th ACM International Conference on Mobile Systems, Applications, and Services (2021)
32. Zhu, R., Li, Z., Matai, J., Porikli, F., Chandraker, M.: Irisformer: dense vision transformers for single-image inverse rendering in indoor scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2822–2831 (2022)