Intrinsic Appearance Decomposition Using Point Cloud Representation

Xiaoyan Xing[1], Konrad Groh[2] , Sezer Karaoglu[1], Theo Gevers[1]
[1]UvA-Bosch Delta Lab, University of Amsterdam
[2]Bosch Center of Artificial Intelligence, Robert Bosch GmbH

Abstract

*The aim of intrinsic decomposition is to deduce the albedo and shading components, typically from 2D images. However, this task is ill-posed, necessitating previous methods to rely on imaging assumptions. In contrast to 2D images, point clouds present a promising solution due to their richness as scene representation formats. They inherently align both the geometric and color information of an image, making them valuable to address this challenging problem. Hence, we propose a method, Point Intrinsic Net (**PoInt-Net**), which jointly predicts the albedo, light source direction, and shading by leveraging point cloud representations. Through experiments, we demonstrate the advantages of PoInt-Net, as it outperforms 2D representation methods across multiple metrics and datasets. Moreover, the model exhibits reasonable generalization capabilities for previously unseen objects and scenes.*

1. Introduction

Intrinsic decomposition is a fundamental challenge and ill-posed problem. Given a single 2D image, there exist numerous potential combinations of albedo and shading that can result in that same input image. Previous methods mainly investigate the priors from images, such as human perception [8, 14], context-based priors [4, 5, 20], and geometric priors [1–3, 9, 13, 23]. Despite demonstrating good performance on trained datasets, these methods encounter challenges when the underlying assumptions or learned patterns fail to accurately capture real-world scenarios. [6] proposes a method for learning intrinsic decomposition based on spatial models of albedo and shading. However, there remains a gap between learning from paradigms and human perception.

Recent studies have demonstrated that employing point cloud data representations provide advantageous for low-level vision tasks, such as color constancy [21], synthetic views generation [22], and light field representation [16]. However, the use of point cloud data representation has been overlooked in the context of intrinsic decomposition.

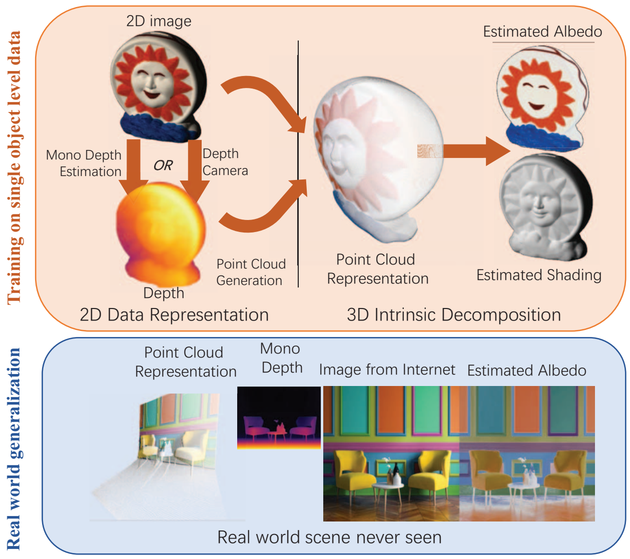

Figure 1: Our approach involves decomposing the intrinsic components of an object/scene by leveraging a point cloud representation of its appearance from a specific viewing angle. Point clouds are generated from $RGB-D$ images, where the depth is obtained by a depth camera (e.g. Lidar or ToF) or is estimated from an RGB image by a monocular depth estimation method such as [19]. Our method is able to generalize well to real-world images taken from unseen shapes/scenes.

Therefore, this paper delves into the exploration of a 3D point cloud representation for intrinsic decomposition. A point-based network (PoInt-Net) is introduced to exploit the 3D structure and appearance of an object or scene, enabling the extraction of surface geometry and intrinsic appearance. By estimating light source direction and surface normals from the input point cloud, the final shading is generated by the shader. The estimated shading is combined with the estimated albedo to reconstruct the input appearance.

Experiments demonstrate that PoInt-Net achieves state-of-the-art shading estimation results on widely used datasets, while maintaining comparable albedo estimation performance. PoInt-Net seamlessly integrates with point clouds generated from estimated or calculated depths (Fig.

1), providing flexibility and robustness to generalize intrinsic decomposition across a broader spectrum of shapes and scenes. The contributions of the paper are:

- A novel paradigm is proposed to intrinsic decomposition by reformulating the task into a 3D point cloud representation, which explicitly incorporates geometric priors and sparse representations.
- A point-based intrinsic decomposition network is introduced, PoInt-Net, includes explainable subnets for light direction estimation, shading rendering, and albedo reconstruction.
- The proposed method outperforms prior methods across multiple datasets, demonstrating robust generalization capabilities.

2. Point Cloud Intrinsic Representation

2.1. Intrinsic Decomposition

Given a viewing angle, if the surface is Lambertian, the appearance $I_{diffuse}$ formulation can be simplified as:

$$\mathbf{I}_{diffuse} = \int_{\omega_i \in \Omega +} f_r(\omega_i) L_i(\omega_i)(N \cdot \omega_i) d\omega_i, \qquad (1)$$

where ω_i is the lighting angle from the upper hemisphere Ω_+, ω_o is the viewing angle, N is the surface normal, $L_i(x, \omega_i)$ is the position of the lighting angle and its direction, and f_r is the surface reflectance, modeled by a Bidirectional Reflectance Distribution Function (BRDF) [15]. Conventionally, $\frac{\rho_d}{\pi}$ denotes the reflectivity of the surface (albedo), where $f_r(\omega_i) = \frac{\rho_d}{2\pi}$. Therefore, if the illumination is uniform, the intrinsic model is defined by:

$$\mathbf{I}_{diffuse} = \frac{\rho_d}{\pi} \cdot (\mathbf{N} \cdot \mathbf{L}_{in}), \qquad (2)$$

where, $\mathbf{L}_{in}$ represents the visible incident light. The aim of intrinsic decomposition is to disentangle the albedo $\mathbf{A} = \frac{\rho_d}{\pi}$ and shading $\mathbf{S} = (N \cdot L_{in})$ from the appearance $\mathbf{I}_{diffuse}$, where $(\cdot)$ is the dot product.

2.2. Intrinsic Appearance on Point Cloud

According to Eq. 1, the appearance of the object under a given lighting condition is acquired as a RGB image $\mathbf{I} = [\mathbf{I}_r, \mathbf{I}_g, \mathbf{I}_b] \in \mathbb{R}^{U \times V \times 3}$. Additionally, its corresponding depth map is represented by $\mathrm{D} \in \mathbb{R}^{U \times V \times 1}$. The depth information can either be directly obtained by a depth camera, e.g., LiDAR, or can be estimated by a (monocular) depth estimation method based on 2D images, e.g.,[19]. The RGB image and corresponding depth map are transformed into a (colored) point cloud representation, $\mathbf{P} = \{\mathbf{p}_i | i \in 1, \ldots, n\}$. Specifically, each point $\mathbf{p}_i$ is represented as a vector of $[x, y, d, r, g, b]$ values:

$$\mathbf{p}_i = \left(\frac{(u - c_x)d}{f_x}, \frac{(v - c_y)d}{f_y}, d, r, g, b \right), \qquad (3)$$

where, f_x and f_y are the focal lengths, and (c_x, c_y) is the principal point. Given a dataset of M point clouds, $\mathcal{P} = \{\mathbf{P}_1, \mathbf{P}_2, ..., \mathbf{P}_M\}$, its intrinsic components can be defined by: 1) Albedo $\mathcal{A} = \{\mathbf{A}_1, \mathbf{A}_2, ..., \mathbf{A}_M\}$, 2) Shading $\mathcal{S} = \{\mathbf{S}_1, \mathbf{S}_2, ..., \mathbf{S}_M\}$, 3) Surface Normal $\mathcal{N} = \{\mathbf{N}_1, \mathbf{N}_2, ..., \mathbf{N}_M\}$, and 4) Light source position $\mathcal{L} = \{\mathbf{L}_1, \mathbf{L}_2, ..., \mathbf{L}_M\}$.

Albedo contains the invariant (albedo) information. Therefore, a direct point based mapping ($f_\alpha : \mathcal{P} \rightarrow \mathcal{A}$) is employed to decompose the reflectance appearance.

Shading depends on the object geometry, viewing and lighting conditions. Thus, instead of directly learning the shading, a point-light direction net ($f_\theta : \mathcal{P} \rightarrow \mathcal{L}$) is used to estimate the light direction from the point cloud representation. Then, a point-learnable shader ($f_\sigma : \mathcal{L}, \mathcal{N} \rightarrow \mathcal{S}$) is trained to generate the rendering effects based on the surface normals (from input point cloud) and light direction estimation (from point-light direction net).

3. Point Based Intrinsic Decomposition

3.1. Point Intrinsic Net

The proposed point intrinsic network (PoInt-Net) consists of three modules: 1) the Point Albedo-Net, which is designed to learn the material properties of object surfaces, 2) the Light Direction Estimation Net, dedicated to infer the lighting conditions. The aim is to support the estimation of the (point cloud) albedo, and 3) the Learnable shader, which combines the inferred light direction and surface normals to generate the shading map. The closest approach to ours is [9]. However, this method estimates the normal information directly from the input images and can only be generalized at a (single) object level. Fig. 2 shows the proposed network architecture and the details of the forward connections. Each of the three sub-nets shares a similar design, with only a few differences such as the activation functions. Specifically, all three sub-nets are adopted from [17], and employ Multi-Layer Perceptrons (MLPs) for point-feature extraction and decoding, with the aim of solving the point-to-point relationship.

Point Albedo-Net takes as input a 6-D point cloud containing color information and spatial coordinates, and produces estimates of surface reflectances. To produce scaled output colors, the Rectified Linear Unit (ReLU) is utilized as the activation function.

Light Direction Estimation Net takes the same input as the *Point Albedo-Net*, and predicts the point-wise light directions. The final two layers use the hyperbolic tangent function (Tanh) as activation to ensure that all light directions are estimated.

Surface Normal Calculation computes the surface normal based on the input point cloud including: 1) neighboring point identification and covariance matrix calculation, 2) eigenvector computation of the covariance matrix, and 3)

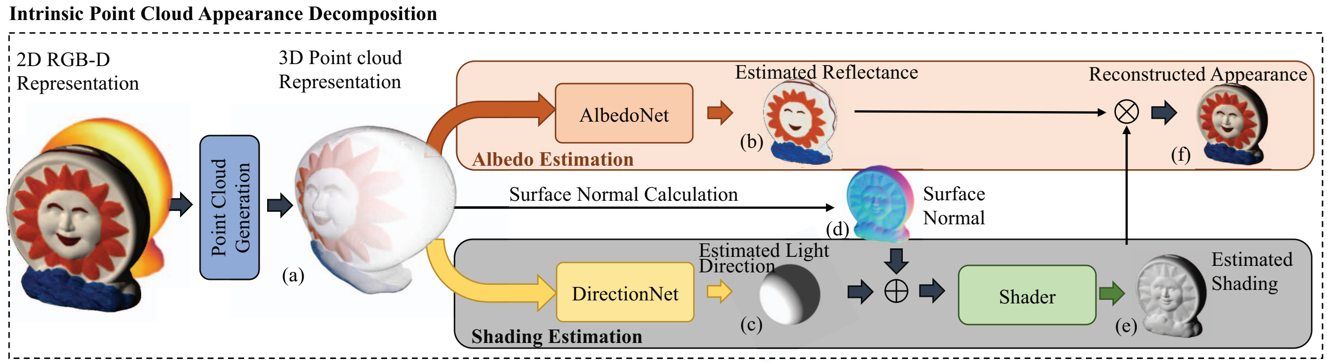

Figure 2: Our proposed framework for intrinsic point cloud decomposition starts by transforming the $RGB{-}D$ representation to a point cloud representation (a). The point cloud representation is used as input to train two separate components: the shading and the albedo estimation. The shading estimation is supported by the DirectionNet (Light Direction Estimation Net), which takes (a) as input and and outputs light direction estimates (c). Surface normals (d) are calculated using local neighborhoods within (a). The Shader (Learnable Shader) then uses the concatenated vectors of (c) and (d) to generate the final shading estimation (e). The albedo estimation is obtained by the AlbedoNet (Point-Albedo Net) which extracts invariant reflectance (b) from (a) based on the Lambertian assumption. Finally, by multiplying (b) and (e), the reconstructed image (f) is generated. Please refer to the supplementary for the detailed architecture.

normal vector selection based on the smallest eigenvalue. To speed up the training process, normal information is pre-computed and used during training.

Learnable Shader takes as input the concatenated vectors of the surface normal information (calculated from the input point cloud) and light direction estimation, and outputs the point-wise shading map.

3.2. Joint-learning Strategy

A two-step training strategy is employed to arrive at an end-to-end intrinsic decomposition learning pipeline. First, for shading estimation, the *Light Direction Estimation Net* and *Learnable Shader* are trained using ground-truth light position **L** and shading **S**. Then, for albedo estimation, the parameters in these two sub-nets are preserved and frozen, while the *Point Albedo-Net* is constrained by the ground-truth albedo **A** and the final reconstructed image $\hat{\mathbf{I}}$ (multiplied by the estimated albedo map $\hat{\mathbf{A}}$ and the estimated shading map $\hat{\mathbf{S}}$). During training, the mean square error is used. The loss function[1], for stage one, is:

$$\mathcal{L}_{shading} = \frac{1}{M}\sum^{M}(|\mathbf{L}-\hat{\mathbf{L}}|^2 + |\mathbf{S}-\hat{\mathbf{S}}|^2). \tag{4}$$

For stage two, a series of loss functions are used to constrain the invariant color information. To address reflectance changes, a color cross ratio loss inspired by [7] is used, formulated as follows:

$$\mathcal{L}_{ccr} = |M_{RG} - M_{\hat{R}\hat{G}}| + |M_{RB} - M_{\hat{R}\hat{B}}| + |M_{GB} - M_{\hat{G}\hat{B}}|, \tag{5}$$

where $\{M_{RG}, M_{RB}, M_{GB}\}$, $\{M_{\hat{R}\hat{G}}, M_{\hat{R}\hat{B}}, M_{\hat{G}\hat{B}}\}$ are the cross color ratios from the ground-truth albedo and the estimated albedo respectively. Please refer to supplemental for the details of cross color ratios calculation. Similarly to [5], the gradient difference is considered and is formulated by:

$$\mathcal{L}_{grad} = |\nabla\mathbf{A} - \nabla\hat{\mathbf{A}}|_2^2 \tag{6}$$

Hence, the reconstruction loss is applied to constrain the estimated albedo:

$$\mathcal{L}_{rec} = \frac{1}{M}\sum^{M}(|\mathbf{A}-\hat{\mathbf{A}}|^2 + |\mathbf{I}-\hat{\mathbf{I}}|^2), \tag{7}$$

The final loss function is:

$$\mathcal{L}_{albedo} = \mathcal{L}_{rec} + \mathcal{L}_{grad} + \mathcal{L}_{ccr}, \tag{8}$$

where $\{\hat{\cdot}\}$ represents the estimated values, and M is the number of input point clouds in a mini-batch. Adam [10] is employed as the optimizer.

4. Experiments

We first train PoInt-Net on the ShapeNet-Intrinsic dataset [9], with ground-truth labels for intrinsic images and light positions. Then, the pre-trained parameters are fine-tuned on the MIT-intrinsic dataset[2] [8], with only ground truth labels for intrinsic images. The quantitative and qualitative results are presented across the two datasets. For the numerical results, three common metrics are employed for evaluation: mean square error (MSE), local mean squared error (LMSE), and structural dissimilarity (DSSIM).

[1] For datasets without light direction labels, the loss function only constrains the shading map $\hat{\mathbf{S}}$.

[2] Depth information is available for download at: here

	MSE×10^2			LMSE×10^2	DSSIM×10^2
	A	S	Avg.	Total	Total
CGIntrinsics[11]	3.38	2.96	3.17	6.23	-
Fan *et al*. [5]	3.02	3.15	3.09	7.17	-
Ma *et al*. $*$[14]	2.84	2.62	2.73	5.44	-
USI3D$*$ [12]	1.85	1.08	1.47	4.65	-
Ours (w/o. shader)	0.48	0.57	0.53	1.15	4.93
Ours	**0.46**	**0.38**	**0.42**	**1.00**	4.15

* Unsupervised methods but finetune on the dataset.

Table 1: Results and ablation study for ShapeNet-Intrinsic.

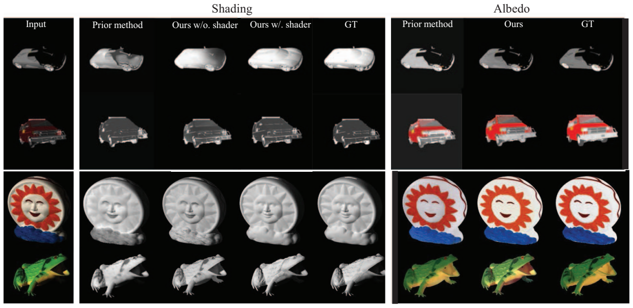

Figure 3: Qualitative results on the ShapeNet-Intrinsic (top 2 rows) & MIT-intrinsic (bottom 2 rows) datasets.

ShapeNet-Intrinsic. We compare our approach to the latest open source methods [5, 11, 12, 14]. Table 1 demonstrate that our approach outperforms existing methods by a large margin for all three metrics. This superior performance attributes to PoInt-Net's ability to accurately predict intrinsic properties by capturing and leveraging complex relationships among them, resulting in more robust and reliable estimations. The qualitative results are presented in Fig. 3, using USI3D [12] (fine-tuned version) as a reference. PoInt-Net's shading map, based on light direction and surface normal, accurately separates shading from the composite image. Consequently, the output images display realistic and consistent shading even when the surface color is ambient. The ablation results further highlight the importance of the shader net; without it, the network cannot accurately estimate shading. Overall, the results underline PoInt-Net's effectiveness and robustness in producing high-quality, visually appealing outputs that accurately capture the objects' intrinsic properties.

MIT-intrinsic. The quantitative results are presented in Table 2, showcasing PoInt-Net's superior performance in producing state-of-the-art shading results on the MIT-intrinsic dataset across all metrics. Notably, our method achieves the best LMSE and ranks second in albedo output performance, in terms of MSE and DSSIM metrics. It is worth noting that [4] utilizes an additional input to enhance albedo estimation. Fig. 3 provides a visualization of our results, where our method surpasses others in accurately identifying the spots on the frog's back—a detail missed by PIE-Net [4]. The high-quality shading results stem from PoInt-Net's use of light direction estimation and surface normal calculation, underscoring its ability to render intricate details.

	MSE×10^2		LMSE×10^2		DSSIM×10^2	
	A	S	A	S	A	S
SIRFS [1]	1.47	1.83	4.16	1.68	12.38	9.85
Ma *et al*. $*$ [14]	3.13	2.07	1.16	0.95	-	-
Janner *et al*. [9]	3.36	1.95	2.10	1.03	-	-
CGIntrinsics [11]	1.67	1.27	3.19	2.21	12.87	13.76
USI3D$*$ [12]	1.57	1.35	1.46	2.31	-	-
FFI-Net [18]	1.11	0.93	2.91	3.19	10.14	11.39
PIE-Net [4]	**0.28**	0.35	1.36	1.83	**3.40**	4.93
Ours	0.89	**0.34**	**0.97**	**0.37**	4.39	**3.02**

* Unsupervised methods but finetune on the dataset.

Table 2: Results for MIT Intrinsic.

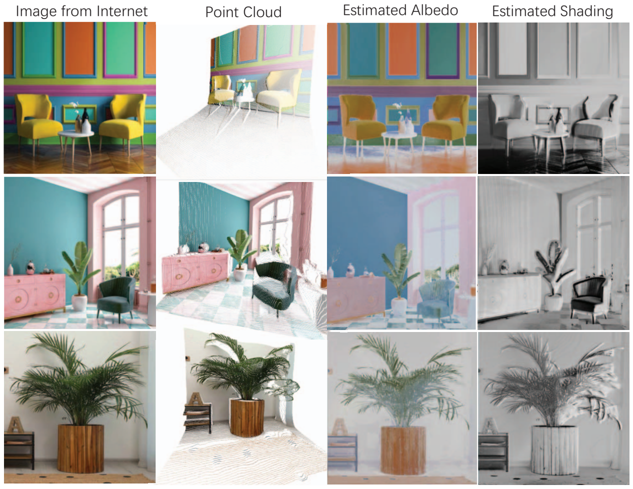

Figure 4: Real-world intrinsic estimated by PoInt-Net.

Real-world generalization. Fig. 4 shows the robust generalization of PoInt-Net on real-world images. The point clouds are generated based on the estimated depth maps from [19], accordingly. Although, PoInt-Net is trained on the single object level dataset. It can still accurately estimate the surface reflectance and shading, for single objects and complex scenes.

5. Conclusion

We introduced a novel method, Point Intrinsic Network (PoInt-Net), utilizing point intrinsic representation for 3D intrinsic appearance decomposition. PoInt-Net effectively decomposes light direction, surface reflectance, and shading maps by leveraging the unique properties of point clouds. Experiments showed that our method outperforms 2D representation methods and exhibits reasonable generalization capabilities. In line with the short paper theme, the proposed framework presents a work in progress. The future work plan is to extend the work to non-Lambertain surfaces. Other directions include the impact of different quality depths on estimate results and network pruning.

References

[1] Jonathan T Barron and Jitendra Malik. Shape, illumination, and reflectance from shading. *IEEE TPAMI*, 37(8):1670–1687, 2015.

[2] Anil S Baslamisli, Partha Das, Hoang-An Le, Sezer Karaoglu, and Theo Gevers. Shadingnet: Image intrinsics by fine-grained shading decomposition. *IJCV*, 129(8):2445–2473, 2021.

[3] Qifeng Chen and Vladlen Koltun. A simple model for intrinsic image decomposition with depth cues. In *ICCV*, 2013.

[4] Partha Das, Sezer Karaoglu, and Theo Gevers. Pie-net: Photometric invariant edge guided network for intrinsic image decomposition. In *CVPR*, 2022.

[5] Qingnan Fan, Jiaolong Yang, Gang Hua, Baoquan Chen, and David Wipf. Revisiting deep intrinsic image decompositions. In *CVPR*, 2018.

[6] David Forsyth and Jason J Rock. Intrinsic image decomposition using paradigms. *IEEE TPAMI*, 44(11):7624–7637, 2021.

[7] Theo Gevers and Arnold WM Smeulders. Color-based object recognition. *Pattern recognition*, 32(3):453–464, 1999.

[8] Roger Grosse, Micah K Johnson, Edward H Adelson, and William T Freeman. Ground truth dataset and baseline evaluations for intrinsic image algorithms. In *ICCV*, 2009.

[9] Michael Janner, Jiajun Wu, Tejas D Kulkarni, Ilker Yildirim, and Josh Tenenbaum. Self-supervised intrinsic image decomposition. In *NIPS*, 2017.

[10] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[11] Zhengqi Li and Noah Snavely. Learning intrinsic image decomposition from watching the world. *CVPR*, 2018.

[12] Yunfei Liu, Yu Li, Shaodi You, and Feng Lu. Unsupervised learning for intrinsic image decomposition from a single image. In *CVPR*, 2020.

[13] Jundan Luo, Zhaoyang Huang, Yijin Li, Xiaowei Zhou, Guofeng Zhang, and Hujun Bao. Niid-net: adapting surface normal knowledge for intrinsic image decomposition in indoor scenes. *IEEE TVCG*, 26(12):3434–3445, 2020.

[14] Wei-Chiu Ma, Hang Chu, Bolei Zhou, Raquel Urtasun, and Antonio Torralba. Single image intrinsic decomposition without a single intrinsic image. In *ECCV*, 2018.

[15] Fred E Nicodemus. Directional reflectance and emissivity of an opaque surface. *Applied optics*, 4(7):767–775, 1965.

[16] Julian Ost, Issam Laradji, Alejandro Newell, Yuval Bahat, and Felix Heide. Neural point light fields. In *CVPR*, pages 18419–18429, 2022.

[17] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *CVPR*, pages 652–660, 2017.

[18] Yanlin Qian, Miaojing Shi, Joni-Kristian Kamarainen, and Jiri Matas. Fast fourier intrinsic network. In *WACV*, 2021.

[19] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE TPAMI*, 44(3), 2022.

[20] Li Shen, Chuohao Yeo, and Binh-Son Hua. Intrinsic image decomposition using a sparse representation of reflectance. *IEEE TPAMI*, 35(12):2904–2915, 2013.

[21] Xiaoyan Xing, Yanlin Qian, Sibo Feng, Yuhan Dong, and Jiří Matas. Point cloud color constancy. In *CVPR*, 2022.

[22] Qiangeng Xu, Zexiang Xu, Julien Philip, Sai Bi, Zhixin Shu, Kalyan Sunkavalli, and Ulrich Neumann. Point-nerf: Point-based neural radiance fields. In *CVPR*, 2022.

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